



HAL
open science

Information Theory with Kernel Methods

Francis Bach

► **To cite this version:**

| Francis Bach. Information Theory with Kernel Methods. 2022. hal-03577992v2

HAL Id: hal-03577992

<https://hal.science/hal-03577992v2>

Preprint submitted on 26 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Information Theory with Kernel Methods

Francis Bach
Inria, Ecole Normale Supérieure
PSL Research University
francis.bach@inria.fr

August 26, 2022

Abstract

We consider the analysis of probability distributions through their associated covariance operators from reproducing kernel Hilbert spaces. We show that the von Neumann entropy and relative entropy of these operators are intimately related to the usual notions of Shannon entropy and relative entropy, and share many of their properties. They come together with efficient estimation algorithms from various oracles on the probability distributions. We also consider product spaces and show that for tensor product kernels, we can define notions of mutual information and joint entropies, which can then characterize independence perfectly, but only partially conditional independence. We finally show how these new notions of relative entropy lead to new upper-bounds on log partition functions, that can be used together with convex optimization within variational inference methods, providing a new family of probabilistic inference methods.

1 Introduction

Characterizing and studying probability distributions through moments has a long history. For distributions supported in a vector space, only an infinite number of moments can characterize the distribution. Beyond traditional polynomial moments, kernel methods based on reproducing kernel Hilbert spaces (RKHS) [48, 10] have emerged as a natural tool for studying the interactions between moments and other properties of the underlying distributions (such as independence or conditional independence), as well as providing algorithmic tools to estimate these moments from data.

A natural way of using kernel methods is to consider a feature map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ from the underlying space \mathcal{X} and a specific Hilbert space \mathcal{H} , and to consider the *mean element* [53], also referred to as the *mean embedding*, for a probability distribution p , defined as

$$\mu_p = \int_{\mathcal{X}} \varphi(x) dp(x), \tag{1}$$

which is an element of the Hilbert space \mathcal{H} . The key idea behind kernel methods is to study properties of \mathcal{H} and its interaction with \mathcal{X} only through the kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined as

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle.$$

Under classical universality conditions on the kernel function k [53, 34], the mapping $p \mapsto \mu_p$ is injective (and thus characterizes the distribution p), and can be estimated from independent and identically distributed (i.i.d.) samples efficiently, with convergence rates proportional to $1/\sqrt{n}$ where n is the number of observations.

Representing a probability distribution p through $\mu_p \in \mathcal{H}$ defines explicitly a metric between distributions as $(p, q) \mapsto \|\mu_p - \mu_q\|$ (for the \mathcal{H} -norm), which has been extensively used in machine learning and data science [35], in particular for measuring the dependence between two random variables, leading to algorithms for independent component analysis [17, 6] and for feature selection [51], or as a model fitting criterion when estimating parameters of probability distributions [13].

A classical drawback of the underlying geometry is the lack of straightforward connection with classical information theory tools. For example for discrete data, this leads to Euclidean norms between probability mass functions, which is rarely seen as an appropriate geometry for the simplex.

In this paper, we consider the second-order moment, which we call the *covariance operator*¹

$$\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x),$$

which is an operator from \mathcal{H} to \mathcal{H} , defined through $\langle f, \Sigma_p g \rangle = \int \langle f, \varphi(x) \rangle \langle g, \varphi(x) \rangle dp(x)$, self-adjoint and positive semi-definite.²

As we show in this paper, it shares similar nice properties with the mean element in terms of universality, estimation from finite samples, and general applicability to all sets \mathcal{X} where positive definite kernels can be defined (e.g., structured discrete objects). Furthermore, owing to the many tools from quantum information theory, and in particular the von Neumann entropy $\text{tr}[\Sigma_p \log \Sigma_p]$ and its associated divergence (relative entropy), we define these notions and demonstrate their many nice properties and applications, as well as explicit quantitative links with regular Shannon entropies (in fact, we can see the regular notion of entropy as the limit of our kernel entropies when the bandwidth of the kernel goes to zero, and for discrete data, they are equivalent).

The paper is organized as follows:

- We review in Section 2 properties of covariance operators and reproducing kernel Hilbert spaces.
- In Section 3, we review some of the key mathematical results on quantum relative entropies, that we will leverage in further sections.
- In Section 4, we define our notions of kernel entropy and Kullback-Leibler divergence (relative entropy), and give their main properties, which are shared with the classical notions, such as convexity, or additivity for independent variables. In particular our notions of relative entropies will always be lower bounds on the Shannon relative entropy (in this paper, we use interchangeably the terms “relative entropy” and “Kullback-Leibler (KL) divergence”).

¹Note that covariance operators are typically “centered”, that is, $\mu_p \mu_p^*$ is subtracted.

²In this paper, we use the notation uu^* for an element $u \in \mathcal{H}$ to denote the operator $uu^* : \mathcal{H} \rightarrow \mathcal{H}$ such that $uu^*(f) = \langle f, u \rangle u$. The usual notation $u \otimes u$ will be used for tensor products in Section 6.

- In Section 5, we show how we can leverage existing work in kernel methods to estimate these quantities from finite samples, with a convergence rate proportional to $1/\sqrt{n}$ from n observations, and without the need for regularization. We also propose statistically more efficient estimators if integrals of kernel functions under p can be computed.
- In Section 6, we consider product spaces $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_d$, and show that for tensor product kernels, we can define notions of mutual information and joint entropies. We can then characterize independence perfectly, but only partially conditional independence, and extend the submodularity property of the regular entropy.
- In Section 7, we show how these new notions of relative entropy lead to new upper-bounds on log partition functions, that can be used together with convex optimization within variational inference methods [58], providing a new family of probabilistic inference methods. Illustrative applications to $[0, 1]$ and $\{-1, 1\}^d$ are explicitly developed.

2 Covariance operators

In this section, we review the notions of covariance operators, starting with reproducing kernel Hilbert spaces. For more details, see [10, 48, 49].

2.1 Kernels and RKHSs

In this paper we consider a compact set \mathcal{X} and probability distributions on \mathcal{X} (which we assume to be Borel probability measures), as well as a kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, such that

(A1) k is a continuous positive definite kernel on the compact set \mathcal{X} , with $k(x, x) = 1$ for all $x \in \mathcal{X}$.

Positive definite kernel functions are functions for which all matrices of pairwise kernel evaluations are positive semi-definite. They are known to define a Hilbert space \mathcal{H} (with dot product $\langle \cdot, \cdot \rangle$) of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, called a reproducing kernel Hilbert space (RKHS), as well as a map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ such that, for all $x, y \in \mathcal{X}$,

$$(a) \quad f(x) = \langle f, \varphi(x) \rangle, \quad (b) \quad k(\cdot, x) = \varphi(x), \quad \text{and} \quad (c) \quad k(x, y) = \langle \varphi(x), \varphi(y) \rangle.$$

A useful representation of functions in \mathcal{H} is through linear combinations of kernel functions, that is, for any $\alpha \in \mathbb{R}^n$ and arbitrary elements x_1, \dots, x_n of \mathcal{X} , the function $f = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ is in \mathcal{H} and its norm is equal to $\sum_{i,j=1}^n \alpha_i \alpha_j k(x_i, x_j)$. It turns out that all functions in \mathcal{H} can be generated as appropriate limits of such finite linear combinations.

Note that from $k(x, x) = 1$, we get $\|\varphi(x)\| = 1$ for all $x \in \mathcal{X}$. Note that most of the developments in this paper do not require further structure on the set \mathcal{X} , except when characterizing regularity of density functions is needed in Section 4.2, where we will assume the existence of a distance.

Universal and characteristic kernels. We note that our covariance operator Σ_p happens to be a mean embedding for the feature map $x \mapsto \varphi(x)\varphi(x)^*$, from \mathcal{X} to the space of operators from \mathcal{H} to \mathcal{H} . Equipped with the Hilbert-Schmidt norm, we obtain a Hilbert space which happens to be isomorphic³ to the RKHS obtained from the kernel $(x, y) \mapsto k(x, y)^2$.

Therefore, injectivity of the map $p \mapsto \Sigma_p$ can be obtained from sufficient conditions for injectivity of the mean element map, a property for the kernel referred to as *characteristic* [52]. As shown in [26, Prop. 5], this is equivalent for the associated RKHS to be dense in $L_2(q)$ for all probability measures q on \mathcal{X} , a sufficient condition being here (because of our compactness assumption) that the RKHS is dense in the set of continuous functions equipped with the uniform norm, a property referred to as *universality* [54]. Note that if **(A1)** is satisfied, then the universality of k implies the universality of k^2 .

We will see in Section 2.3 below that many kernels are indeed universal, and thus the covariance operator will be characteristic of the corresponding probability distribution.

2.2 Covariance operators

Given a probability distribution p on \mathcal{X} , we consider the operator $\Sigma_p : \mathcal{H} \rightarrow \mathcal{H}$, defined as

$$\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x). \quad (2)$$

For any $f, g \in \mathcal{H}$, we have, by definition: $\langle g, \Sigma_p f \rangle = \mathbb{E}_{X \sim p}[\langle f, \varphi(X) \rangle \langle g, \varphi(X) \rangle] = \mathbb{E}_{X \sim p}[f(X)g(X)]$.

Proposition 1 (Properties of covariance operators) *Assume (A1).*

- (a) *For any probability distribution p on \mathcal{X} , the operator $\Sigma_p : \mathcal{H} \rightarrow \mathcal{H}$ defined in Eq. (2), is self-adjoint, positive-semidefinite and has unit trace.*
- (b) *If the kernel k^2 is universal, then the mapping $p \mapsto \Sigma_p$ is injective from probability distributions to self-adjoint positive-semidefinite, unit trace operators.*
- (c) *If p has full support in \mathcal{X} , then the operator Σ_p has a trivial null space.*
- (d) *Given the mean element defined in Eq. (1), we have $\Sigma_p \succcurlyeq \mu_p \otimes \mu_p$.*

Proof (a) The operator is well defined and has unit trace because $\text{tr}[\varphi(x)\varphi(x)^*] = \|\varphi(x)\|^2 = 1$ for all $x \in \mathcal{X}$.

(b) If $\Sigma_p = \Sigma_q$, then for all $y \in \mathcal{X}$, $\int_{\mathcal{X}} k(x, y)^2 [dp(x) - dq(x)] = 0$. The representation of functions in the RKHS associated with k^2 as linear functions of $x \mapsto k(x, y)^2$, and the universality of k^2 implies that $\int_{\mathcal{X}} f(x)[dp(x) - dq(x)] = 0$ for all continuous functions f , hence, $p = q$ and the injectivity of the map.

Property (c) is a simple consequence of $\langle g, \Sigma_p g \rangle = \mathbb{E}_{X \sim p}[g(X)^2]$, which is valid for all $g \in \mathcal{H}$. If this expectation is equal to zero, and since elements of \mathcal{H} are continuous functions, we must have $g = 0$ as soon as p has full support.

Property (d) is a consequence of the non-negativity of the variance of $f(X)$, for f in \mathcal{H} . ■

³We can build the simple mapping $M \mapsto \langle M, \varphi(x)\varphi(x)^* \rangle$ obtained from $fg^* \mapsto \langle fg^*, \varphi(x)\varphi(x)^* \rangle = f(x)g(x)$.

Since the operators are trace-class, they have a discrete spectrum [41, 16], composed of a summable sequence of strictly positive eigenvalues tending to zero, and potentially zero (if p does not have full support in \mathcal{X}). The spectral decay is a key quantity for the study of kernel methods (see [21] and references therein), as used in Section 5 to estimate entropies from data.

Throughout this paper, we will consider *spectral functions* of trace-class self-adjoint operators A , such as $\log A$, $\exp(A)$, or more generally $f(A)$, for a function $f : \mathbb{R} \rightarrow \mathbb{R}$. These are defined through the eigenvalue decomposition by keeping eigenvectors and replacing eigenvalues λ by $f(\lambda)$ (for polynomials, we then recover the usual matrix polynomials).

Uniform distribution and symmetric sets. In order to define the notion of kernel entropy from kernel relative entropy, we will need to use a specific “base” distribution τ , that will play the role of the uniform distribution when this notion is classically defined (e.g., \mathcal{X} is a finite set or a subset of \mathbb{R}^d , like in all examples in Section 2.3).

For this base distribution τ on \mathcal{X} , we use the notation (instead of Σ_τ):

$$\Sigma = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* d\tau(x).$$

In this paper, some concepts will be easier for *symmetric sets*, that is, sets where there exists a set \mathcal{T} of bijective transformations $t : \mathcal{X} \rightarrow \mathcal{X}$, such that for all $x, y \in \mathcal{X}$, $\exists t \in \mathcal{T}$ such that $t(x) = y$. We will then assume that the kernel k and the base distribution τ are compatible with this notion of symmetry, that is, $\tau(A) = \tau(t(A))$ for any Borel set A and $t \in \mathcal{T}$, and $k(x, y) = k(t(x), t(y))$ for all transformations $t \in \mathcal{T}$.

Then the covariance operator for the base distribution will also exhibit symmetries. More precisely, all eigensubspaces will be invariant by symmetry, that is, if $e_1, \dots, e_m \in \mathcal{H}$ is an orthonormal basis of an eigensubspace corresponding to an eigenvalue λ , then⁴ $\sum_{i=1}^m e_i(t(x))^2 = \sum_{i=1}^m e_i(x)^2$ for all transformations $t \in \mathcal{T}$ and $x \in \mathcal{X}$. This implies that the function $x \mapsto \sum_{i=1}^m e_i(t(x))^2$ is constant, equal to its expectation under τ (which is itself equal to $m\lambda$).

As a consequence, for a spectral function $\Sigma \mapsto f(\Sigma)$ based on a real-valued function $f : \mathbb{R} \rightarrow \mathbb{R}$, the dot-product $\langle \varphi(x), f(\Sigma)\varphi(x) \rangle$ can be expanded as the sum of the contributions coming from all eigensubspaces, with the one associated to the eigenvalue λ leading to $f(\lambda) \sum_{i=1}^m e_i(x)^2$ being independent of x (and equal to $m\lambda f(\lambda)$). This leads notably to $\langle \varphi(x), f(\Sigma)\varphi(x) \rangle = \text{tr}[\Sigma f(\Sigma)]$ for all real-valued functions f defined on \mathbb{R} , and to $\text{tr}[\Sigma f(\Sigma)] = \text{tr}[\Sigma_p f(\Sigma)]$ for any probability distributions p .

Surjectivity. By definition, every covariance operator is in the closure \mathcal{M} of the *convex hull* of all $\varphi(x)\varphi(x)^*$, for $x \in \mathcal{X}$. We will also need the closure \mathcal{A} of the *span* of these vectors. By definition, the mapping $p \mapsto \Sigma_p$ is surjective from probability distributions to \mathcal{M} . Moreover, if the kernel k is universal,

⁴This is a simple consequence of $(e_1 \circ t, \dots, e_m \circ t)$ being another orthonormal basis of the same eigensubspace and thus equal to an orthogonal matrix times (e_1, \dots, e_m) , so that $\sum_{i=1}^m e_i e_i^* = \sum_{i=1}^m (e_i \circ t)(e_i \circ t)^*$, leading to $\langle \varphi(x), \sum_{i=1}^m e_i e_i^* \varphi(x) \rangle$ constant on \mathcal{X} , and thus implying the desired result.

then⁵

$$\mathcal{M} = \{A \in \mathcal{A}, A \succcurlyeq 0, \text{tr } A = 1\}. \quad (3)$$

For orthonormal embeddings, where $k(x, y) = 0$ if $x \neq y$, \mathcal{M} is (in the appropriate basis) the set of non-negative diagonal operators, while \mathcal{A} is the set of diagonal operators (in the same basis).

Note that in most cases \mathcal{A} is strictly included in the set of trace class operators from \mathcal{H} to \mathcal{H} , which is equivalent to the existence of bounded self-adjoint operators $M : \mathcal{H} \rightarrow \mathcal{H}$ such that $\forall x \in \mathcal{X}, \langle \varphi(x), M\varphi(x) \rangle = 0$. Since $\|\varphi(x)\|^2 = 1$ for all $x \in \mathcal{X}$, this is equivalent to the existence of a positive bounded operator M so that $\forall x \in \mathcal{X}, \langle \varphi(x), M\varphi(x) \rangle = 1$, which, with an eigenvalue decomposition leads to a finite or countable family of functions $(f_i)_{i \in I}$ in \mathcal{H} such that $\sum_{i \in I} f_i(x)^2 = 1$, that is, a “partition of unity”, which are well known to exist in many spaces (see, e.g., [9, Chapter 3] or [29, Section 1.4]).

2.3 Examples

Finite set. If $\mathcal{X} = \{x_1, \dots, x_m\}$ is a finite set of cardinality m , then the Gram matrix $K \in \mathbb{R}^{m \times m}$ defined as $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$ completely characterizes the kernel notions up to rotation. If $K = I$ (orthonormal embedding), then there exist m orthonormal vectors u_1, \dots, u_m such that $\Sigma_p = \sum_{i=1}^m \mathbb{P}(X = i) u_i u_i^*$, and thus, all covariance matrices commute and therefore share the same eigenbasis, and we will recover exactly the regular notions of discrete entropy.

We get a symmetric set if K is invariant by permutations, that is, $K = \alpha I + (1 - \alpha)11^\top$, and as soon as K is invertible, a universal kernel.⁶

Polynomial kernels. When \mathcal{X} is a subset of \mathbb{R}^d , then we can consider kernels of the form $k(x, y) = (\alpha + \beta x^\top y)^r$, which correspond to $\varphi(x)$ composed of all monomials of order up to r , and the covariance operator is thus composed of traditional moments. Here, we can have $k(x, x) = 1$ for all $x \in \mathcal{X}$ only if \mathcal{X} is included in a centered sphere. Alternatively, we may only impose $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, which still preserves many of our results in later sections.

Translation-invariant kernels on a compact subset of \mathbb{R}^d . These are the usual kernels $k(x, y)$ of the form $k(x, y) = q(x - y)$ where q is a function with strictly positive Fourier transform, which are all universal [54]. Classical examples are the Gaussian kernels $k(x, y) = \exp(-\|x - y\|_2^2 / \sigma^2)$, where \mathcal{H} is a space of infinitely differentiable functions, or the exponential kernel $k(x, y) = \exp(-\|x - y\|_2 / \sigma)$, where \mathcal{H} is the Sobolev space of functions with square-integrable derivatives up to order $(d + 1)/2$. However, they typically do not respect the symmetry of the underlying set (for example for spheres, like below).

Torus. If $\mathcal{X} = [0, 1]$, we can consider kernels of the form $k(x, y) = k(x - y)$ where $k : \mathbb{R} \rightarrow \mathbb{R}$ is 1-periodic and has a strictly positive Fourier series (note that we use the same notation for the kernel and the underlying function). We then get a universal kernel and a symmetric set (invariant by periodic

⁵For the non trivial inclusion, we consider $A = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$ with $\int_{\mathcal{X}} dp(x) = 1$, such that $A \succcurlyeq 0$. We have for all $f \in \mathcal{H}, \int_{\mathcal{X}} f(x)^2 dp(x) \geq 0$. If k is universal, then all positive continuous functions g are limits of functions f^2 , so that $\int_{\mathcal{X}} g(x) dp(x) \geq 0$, which implies that p is a probability measure.

⁶Note that K invertible implies that the Hadamard (i.e., pointwise) product $K \circ K$ is invertible, but not vice-versa.

translation); moreover, the base distribution will be the usual uniform distribution. For example, we will often consider:

$$k(x) = \frac{(1 - e^{-\sigma})^2}{1 + e^{-2\sigma} - 2e^{-\sigma} \cos 2\pi x} = \frac{1}{1 + \frac{\sin^2 \pi x}{\sinh^2(\sigma/2)}},$$

for which the Fourier series is $\hat{k}(\omega) = \int_0^1 e^{-2i\pi\omega x} k(x) dx = \tanh(\frac{\sigma}{2}) e^{-\sigma|\omega|}$ for $\omega \in \mathbb{Z}$.

This extends naturally to $[0, 1]^d$, by considering tensor products, that is, $k(x, y) = \prod_{i=1}^d k(x_i - y_i)$, which for example leads to $\hat{k}(\omega) = \tanh^d(\frac{\sigma}{2}) e^{-\sigma\|\omega\|_1}$ for $\omega \in \mathbb{Z}^d$ in the example above. We will also need an explicit formula for $\sum_{\omega \in \mathbb{Z}} \hat{k}(\omega) \log \hat{k}(\omega) = \log \tanh \frac{\sigma}{2} - \frac{\sigma}{\sinh \sigma}$.

Spheres. We can also consider \mathcal{X} as the unit ℓ_2 -sphere in dimension d , with kernels that are invariant by rotation, leading also to symmetric sets. Kernels are then defined through their expansion in spherical harmonics [50], with links with neural networks [4].

We could also imagine extensions to Stiefel or Grassman manifolds [23], and of course to all sets \mathcal{X} where positive definite kernels can be defined (such as graphs, trees, permutations, sets, etc.) [49].

3 Review of quantum information theoretic quantities

We consider here complex-valued Hilbert spaces, but all results will apply to operators on real Hilbert spaces. In quantum mechanics (see, e.g., [30]), physical states are represented as elements of a Hilbert space \mathcal{H} , and a physical system is characterized by a convex combination of orthogonal projections on these states, which is often called a density matrix. Our covariance operators defined as $\Sigma_p = \int_{\mathcal{X}} \varphi(x) \varphi(x)^* dp(x)$ can thus be seen as “density operators”. It turns out that in quantum information theory (see, e.g., [59]), information-theoretic quantities are defined based on these density operators. In this section, we review these notions from a mathematical point of view.

Entropy. For a compact positive self-adjoint operator A on the complex Hilbert space \mathcal{H} with finite trace, we can define the negative entropy as [2]:

$$\text{tr} [A \log A] = \sum_{\lambda \in \Lambda(A)} \lambda \log \lambda,$$

where $\Lambda(A)$ is the set of eigenvalues of A (with the convention that $0 \log 0 = 0$). It may not be finite if the sequence of eigenvalues is not decreasing sufficiently fast (note that this is unlikely since they are summable). As defined, $\text{tr}[A \log A]$ is always less than $(\text{tr} A) \log(\text{tr} A)$ and can be equal to $-\infty$.

Kullback-Leibler (KL) divergence (relative entropy). It is defined as

$$D(A||B) = \text{tr} [A(\log A - \log B)],$$

for any two positive Hermitian operators with finite trace. It can only be finite if the null space of B is included in the null space of A , that is, if $A = B^{1/2}MB^{1/2}$ for a certain bounded operator M , and we then have $\text{tr } A - \text{tr } B \leq D(A\|B) \leq (\text{tr } A) \log \|M\|_{\text{op}}$.

The quantity $D(A\|B) - \text{tr } A + \text{tr } B$ is always non-negative as the Bregman divergence [15] associated with the convex function $A \mapsto \text{tr}[A \log A]$, and may not always be finite. The following properties are classical and are proved from first principles in Appendix A.

Proposition 2 (Properties of quantum entropy and relative entropy) *For Hermitian positive operators A, B , we have:*

- (a) *The mappings $A \mapsto \text{tr}[A \log A]$ and $(A, B) \mapsto D(A\|B)$ are convex.*
- (b) *We have $D(A\|B) \geq \text{tr } A - \text{tr } B$, with equality if and only if $A = B$.*
- (c) *If $A_i, B_i, i = 1, \dots, n$, are Hermitian operators and $\lambda_i > 0$ such that $\sum_{i=1}^n \lambda_i = 1$, then*

$$D\left(\sum_{i=1}^n \lambda_i A_i \middle\| \sum_{i=1}^n \lambda_i B_i\right) \leq \sum_{i=1}^n \lambda_i D(A_i\|B_i),$$

with equality if and only if $\log B_i - \log A_i$ does not depend on i . Moreover, if there is equality, there exists an operator M such that for all i , $A_i = B_i M$.

- (d) *Monotonicity of quantum operations: Given operators $C_i : \mathcal{H} \rightarrow \mathcal{K}$ and $A, B : \mathcal{H} \rightarrow \mathcal{H}$, $i = 1, \dots, n$, such that $\sum_{i=1}^n C_i^* C_i = I$, then*

$$D\left(\sum_{i=1}^n C_i A C_i^* \middle\| \sum_{i=1}^n C_i B C_i^*\right) \leq D(A\|B).$$

The last property is classical in quantum information theory as the mapping $A \mapsto \sum_{i=1}^n C_i A C_i^*$ is usually referred to as a “quantum operation”. Note that quantum operations are usually defined as “completely positive” trace-preserving maps, and that owing to Stinespring representation theorem (see [12, Theorem 3.1.2]), they are essentially all of that form.

One particular application of monotonicity is to consider n positive self-adjoint operators $D_i : \mathcal{H} \rightarrow \mathcal{H}$, $i = 1, \dots, n$, such that $\sum_{i=1}^n D_i = I$, and consider $\mu, \nu \in \mathbb{R}^n$ defined through $\mu_i = \text{tr}(D_i A)$ and $\nu_i = \text{tr}(D_i B)$ for A, B a positive self-adjoint operator with unit trace, then μ and ν are on the simplex, and the monotonicity leads to (see detailed proof in Appendix A.2):

$$\sum_{i=1}^n \mu_i \log \frac{\mu_i}{\nu_i} \leq D(A\|B). \tag{4}$$

Further results. In Appendix A.1, we provide finer results beyond expectation with respect to a probability measure with finite support, as well as equality cases.

Integral representation. For the KL divergence, we have, by direct integration [1]:

$$D(A\|B) = \text{tr} [A(\log A - \log B)] = - \int_0^{+\infty} \left(\text{tr} [A(A + \lambda I)^{-1}] - \text{tr} [A(B + \lambda I)^{-1}] \right) d\lambda, \quad (5)$$

which will be used in Section 5.2 for the estimation from a finite sample. As shown in Appendix A, we also have another integral representation formula, that helps in proving additional convexity results.

4 Kernel entropy and Kullback-Leibler divergence

We define the kernel Kullback-Leibler (KL) divergence, or kernel relative entropy, between two probability distribution p and q as

$$D(\Sigma_p\|\Sigma_q) = \text{tr} [\Sigma_p(\log \Sigma_p - \log \Sigma_q)]. \quad (6)$$

In this section, we list some of its properties, which are similar to the properties of $(p, q) \mapsto \|\mu_p - \mu_q\|$, but with the added benefit that it has a direct link with Shannon entropy, which is useful in itself when a precise information measure is needed (e.g., for differential privacy [22]), or when used within variational inference (see Section 7).

First, the kernel relative entropy is finite under general conditions.

Proposition 3 (Finiteness of kernel KL divergence) *Assume (A1). If p is absolutely continuous with respect to q , with $\|\frac{dp}{dq}\|_\infty \leq \alpha$, then both the regular and kernel KL divergences are between zero and $\log \alpha$.*

Proof We have then $\Sigma_p \preceq \alpha \Sigma_q$, leading to (note that we must have $\alpha \geq 1$):

$$D(\Sigma_p\|\Sigma_q) = \text{tr} [\Sigma_p(\log \Sigma_p - \log \Sigma_q)] \leq \text{tr} [\Sigma_p(\log(\alpha \Sigma_q) - \log \Sigma_q)] = (\log \alpha) \text{tr} \Sigma_p = \log \alpha.$$

The positivity follows from Prop. 2, property (b). ■

This implies that if p has a bounded density with respect to the base measure τ on \mathcal{X} , then the relative entropy $D(\Sigma_p\|\Sigma)$ is well-defined, and always non-negative (recall that Σ is the covariance operator for the base probability measure τ on \mathcal{X}).

We can then define the kernel entropy as

$$\begin{aligned} H(\Sigma_p) &= -D(\Sigma_p\|\Sigma) - \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle \\ &= -\text{tr} [\Sigma_p \log \Sigma_p] + \text{tr} [\Sigma_p \log \Sigma] - \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle. \end{aligned} \quad (7)$$

Note that we add terms on top of $-\text{tr} [\Sigma_p \log \Sigma_p]$, one linear term in Σ_p , and one constant term, to extend properties of the regular Shannon discrete entropy in all cases. However, everything simplifies in some cases. Indeed, when there are some invariances that are shared by the kernel and the base distribution τ (that is, a symmetric set as defined in Section 2.2), $x \mapsto \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle$ is constant on \mathcal{X} , and thus equal to its expectation under p , which implies $\min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle = \text{tr} [\Sigma_p \log \Sigma]$, and then the definition above simplifies to $H(\Sigma_p) = -\text{tr} [\Sigma_p \log \Sigma_p]$.

The kernel information quantities have natural properties mimicking the traditional quantities, as well as a bound by the regular KL divergence.

Proposition 4 (Properties of kernel entropy and relative entropy) *Assume (A1).*

- (a) *The kernel KL divergence defined in Eq. (6) is non-negative, and equal to 0 for $p = q$. If the kernel k^2 is universal, it is zero if and only if $p = q$.*
- (b) *The kernel entropy defined in Eq. (7) is always between 0 and $-\min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle$ (which is equal to $-\text{tr} \Sigma \log \Sigma$ for symmetric sets). If $p = \tau$ (base measure), it is equal to the upper-bound, while for Dirac measures with a point mass at any minimizers of $x \mapsto \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle$, it is equal to 0 (for symmetric sets, this is all of \mathcal{X}). If the kernel k^2 is universal, these conditions are also necessary.*
- (c) *The mapping $(p, q) \mapsto D(\Sigma_p \| \Sigma_q)$ is convex, and thus the mapping $p \mapsto H(\Sigma_p)$ is concave.*
- (d) *For all probability measures p and q , $D(\Sigma_p \| \Sigma_q) \leq D(p \| q)$. Moreover, the differential entropy with respect to the base measure τ , that is, $-\int_{\mathcal{X}} \log \frac{dp}{d\tau}(x) dp(x)$, is less than $H(\Sigma_p) + \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle$.*
- (e) *For all probability measures p and q , $D(\Sigma_p \| \Sigma_q) \geq \frac{1}{2} \|\Sigma_p - \Sigma_q\|_*^2 \geq \frac{1}{2} \|\Sigma_p - \Sigma_q\|_{\text{HS}}^2$, where $\|\cdot\|_*$ denotes the nuclear norm and $\|\cdot\|_{\text{HS}}$ the Hilbert-Schmidt norm.*

Proof Property (a) is the consequence of the non-negativity of the relative entropy when the two arguments have the same trace (here equal to 1), and injectivity of the map $p \mapsto \Sigma_p$ when k^2 is universal.

For property (b), the upper-bound results follow similarly. For the non-negativity, we notice that $H(\Sigma_p)$ defined in Eq. (7) is the sum of two non-negative terms $-\text{tr}[\Sigma_p \log \Sigma_p]$ (negative because all eigenvalues of Σ_p are less than one) and $\text{tr}[\Sigma_p \log \Sigma] - \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma) \varphi(x) \rangle$. It is equal to zero if and only the two terms are equal to zero. For the first term, this is equivalent to Σ_p being rank one, while for the second one, this is equivalent to p being supported in the set of minimizers. The sufficient condition for being zero thus follows; for the necessary condition, if k^2 is universal, $k(x, x')^2 < 1$ if $x \neq x'$, and thus Σ_p being rank one implies that p is a Dirac measure.

Property (c) is a consequence of the convexity of the map $(A, B) \mapsto \text{tr}[A(\log A - \log B)]$ and of the linearity of the mapping $p \mapsto \Sigma_p$. Property (d) is obtained from

$$\begin{aligned}
 D(\Sigma_p \| \Sigma_q) &= D\left(\int_{\mathcal{X}} \varphi(x) \varphi(x)^* dp(x) \left\| \int_{\mathcal{X}} \frac{dq}{dp}(x) \varphi(x) \varphi(x)^* dp(x)\right.\right) \\
 &\leq \int_{\mathcal{X}} D\left(\varphi(x) \varphi(x)^* \left\| \frac{dq}{dp}(x) \varphi(x) \varphi(x)^*\right.\right) dp(x) \\
 &= \int_{\mathcal{X}} \|\varphi(x)\|^2 D\left(1 \left\| \frac{dq}{dp}(x)\right.\right) dp(x) = \int_{\mathcal{X}} \log\left(\frac{dp}{dq}(x)\right) dp(x) = D(p \| q),
 \end{aligned}$$

using joint convexity of the quantum relative entropy, and the fact that $\|\varphi(x)\| = 1$ for all $x \in \mathcal{X}$. Note that we can only have equality if for all $x \neq y$, $k(x, y) = 0$, that is, we have an orthonormal embedding, which is only possible for discrete \mathcal{X} . Note also that the upper-bound remains valid as soon as $k(x, x) \leq 1$ for all $x \in \mathcal{X}$.

Property (e) is simply a consequence of Pinsker inequality for the quantum relative entropy (see [62]). ■

Note that while the relative entropy results do not depend on the choice of the base measure, the notion of kernel entropy depends on this choice (however, for the symmetric sets presented in Section 2.3, the natural choice of the base measure is the uniform measure).

As detailed in Appendix A.4, the properties outlined in Prop. 4, in particular the lower-bound on the relative entropy, is valid for other “ f -divergences” [19, 55], such as the squared Hellinger distance, or the Pearson χ^2 -divergence.

4.1 Lower bound on kernel relative entropy

Beyond the upper-bound by the Shannon relative entropy (property (d) in Prop. 4), we can also get a lower bound on the kernel KL divergence, by defining for all $y \in \mathcal{X}$, $D(y) = \Sigma^{-1/2}(\varphi(y)\varphi(y)^*)\Sigma^{-1/2}$, which is a positive self-adjoint operator, so that $\int_{\mathcal{X}} D(y)d\tau(y) = I$. We can then apply the monotonicity of the relative entropy with operators $D(y)^{1/2}$, $y \in \mathcal{X}$, that is, Eq. (4) extended to general expectations. We have

$$\text{tr}[D(y)\Sigma_p] = \int_{\mathcal{X}} \langle \varphi(x), \Sigma^{-1/2}\varphi(y) \rangle^2 dp(x) = \int_{\mathcal{X}} h(x, y) dp(x),$$

defining the function $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ as $h(x, y) = \langle \varphi(x), \Sigma^{-1/2}\varphi(y) \rangle^2$, which is non negative and such that

$$\forall y \in \mathcal{X}, \int_{\mathcal{X}} h(x, y) d\tau(x) = \langle \varphi(y), \Sigma^{-1/2} \left(\int_{\mathcal{X}} \varphi(x)\varphi(x)^* d\tau(x) \right) \Sigma^{-1/2} \varphi(y) \rangle = \langle \varphi(y), \varphi(y) \rangle = 1.$$

The function h can thus be seen as a smoothing kernel⁷. Thus, by monotonicity of the relative entropy (Appendix A.2), we get:

$$D(\tilde{p}||\tilde{q}) \leq D(\Sigma_p||\Sigma_q),$$

where $\tilde{p}(y) = \text{tr}[D(y)\Sigma_p] = \int_{\mathcal{X}} h(x, y) dp(x)$ is a smoothed version of p (with the same definition for \tilde{q}). Note that we have directly $D(\tilde{p}||\tilde{q}) \leq D(p||q)$ from standard Markov chain arguments in information theory⁸ [18, Section 2.9]. Overall, we have the following sequence of inequalities:

$$D(\tilde{p}||\tilde{q}) \leq D(\Sigma_p||\Sigma_q) \leq D(p||q). \tag{8}$$

These can lead to quantitative bounds between $D(\Sigma_p||\Sigma_q)$ and $D(p||q)$, in particular when the smoothing function h is putting most of its mass on pairs (x, y) where x is close to y . This is made explicit below for distributions on a metric space and distributions with Lipschitz-continuous densities.

4.2 Small-width asymptotics for metric spaces

We now provide a bound between $D(\tilde{p}||\tilde{q})$ and $D(p||q)$ in Eq. (8), and thus between $D(\Sigma_p||\Sigma_q)$ and $D(p||q)$, when the set \mathcal{X} is equipped with a distance d , used to characterize the regularity of density functions.⁹

⁷Note here that the smoothing property of h corresponds to the pointwise non-negativity, but that h is also a positive definite kernel.

⁸The distributions \tilde{p} and \tilde{q} are obtained by the same transition kernel, respectively from the distributions p and q .

⁹This section is not crucial to understand the rest of the paper.

This corresponds to bounding the difference in the classical data processing inequality for the Shannon entropy, using regularity properties of the densities of p and q . The following proposition is shown in Appendix B.

Proposition 5 (Joint bound on relative entropy) *We assume that p, q have strictly positive Lipschitz-continuous densities (denoted p and q) with respect to the base measure τ . Using definitions from Section 4.1, we have:*

$$0 \leq D(p||q) - D(\tilde{p}||\tilde{q}) \leq E(p, q) \times \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} h(x, y) d(x, y)^2 dy, \quad (9)$$

with $E(p, q) = 2C_{p/q}^2 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} + 4 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} (C_p + C_q)^2 \sup_{x \in \mathcal{X}} \frac{p(x)}{q(x)} (1 + C_p^2 \text{diam}(\mathcal{X})^2)$, where C_p, C_q and $C_{p/q}$ are Lipschitz constants satisfying for all $x, y \in \mathcal{X}$, $|\frac{q(x)}{q(y)} - 1| \leq d(x, y)C_q$, $|\frac{p(x)}{p(y)} - 1| \leq d(x, y)C_p$, and $|\frac{p(y)}{q(y)} - \frac{p(x)}{q(x)}| \leq d(x, y)C_{p/q}$ for some distance d on \mathcal{X} .

Thus, for Lipschitz-continuous densities, the approximation of $D(p||q)$ by $D(\Sigma_p||\Sigma_q)$ is controlled by a kernel dependent quantity, that we now study in special cases.

Discrete orthonormal embeddings. For these kernels where $k(x, y) = 1_{x=y}$, then we have $\Sigma = I$ and thus $h(x, y) = 1_{x=y}$; therefore the quantity $\sup_{x \in \mathcal{X}} \int_{\mathcal{X}} h(x, y) d(x, y)^2 dy$ is always zero, which is natural here since in this situation, kernel and Shannon entropies are equal.

Torus. We have by definition $h(x, y) = \langle \varphi(x), \Sigma^{-1/2} \varphi(y) \rangle^2$. For a kernel on the torus $[0, 1]$ of the form $k(x, y) = k(x - y)$ where k has Fourier series \hat{k} , then we have:

$$h(x, y) = \left(\sum_{\omega \in \mathbb{Z}} \hat{k}(\omega)^{1/2} e^{2i\pi\omega(x-y)} \right)^2,$$

and for the particular choice of $\hat{k}(\omega) \propto e^{-\sigma|\omega|}$, we have: $h(x, y) = \frac{\tanh \frac{\sigma}{2}}{\tanh^2 \frac{\sigma}{4}} \left(1 + \frac{\sin^2 \pi(x-y)}{\sinh^2(\sigma/4)} \right)^{-2}$, and for $d(x, y) = |\sin \pi(x-y)|$, then, using exact integration, we get:

$$\int_{y-1/2}^{y+1/2} h(x, y) d(x, y)^2 dx = \frac{\tanh \frac{\sigma}{2}}{\tanh^2 \frac{\sigma}{4}} \frac{\sinh^3(\sigma/4)}{2(1 + \sinh^2(\sigma/4))^{3/2}} \leq \frac{\sigma^2}{16}.$$

For the d -dimensional torus, and, $d(x, y)^2 = \sum_{j=1}^d \sin^2 \pi(x_j - y_j)$, we get the bound $\frac{d\sigma^2}{16}$, and hence a nice scaling in dimension (non-exponential). Overall, in this situation, we thus get an approximation of entropy from Eq. (9) up to $O(\sigma^2)$, where σ is the width of the kernel. Thus, if our goal is to estimate the true relative entropy, and not only have a lower bound, we need to have σ going to zero, which corresponds to RKHSs which are bigger and bigger. See discussion in Section 5.3 when estimating quantities from a finite sample.

Note that this extends to all translation invariant kernels on $[0, 1]^d$ (and not only for $\hat{k}(\omega) \propto e^{-\sigma\|\omega\|_1}$), and more generally to translation-invariant kernels on compact spaces.

Alternative proof through leverage scores. While we provided a proof of the $O(\sigma^2)$ approximation of the entropy through quantum data processing inequalities, the relationship between the regular entropy and the kernel entropy can also be studied through the integral representation in Eq. (5). We indeed have, using Eq. (5):

$$\begin{aligned}
D(\Sigma_p \parallel \Sigma_q) &= \int_0^{+\infty} \text{tr} \left[\Sigma_p ((\Sigma_q + \lambda I)^{-1} - (\Sigma_p + \lambda I)^{-1}) \right] d\lambda \\
&= \int_0^{+\infty} \int_{\mathcal{X}} \langle \varphi(x), ((\Sigma_q + \lambda I)^{-1} - (\Sigma_p + \lambda I)^{-1}) \varphi(x) \rangle dp(x) d\lambda \\
&= \int_0^{+\infty} \int_{\mathcal{X}} \left(\langle \varphi(x), (\Sigma_q + \lambda I)^{-1} \varphi(x) \rangle - \langle \varphi(x), (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle \right) dp(x) d\lambda.
\end{aligned}$$

We now consider that both p and q have densities with respect to the base measure τ (which we also denote p and q). The main idea, using and extending results from [38], is that for all x and λ , and p smooth enough, the quantity $\langle \varphi(x), (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle$, often referred to as a ‘‘leverage score’’ can be approximated as $\langle \varphi(x), (p(x)\Sigma + \lambda I)^{-1} \varphi(x) \rangle$.

We thus consider the potential approximation:

$$\begin{aligned}
&\int_0^{+\infty} \int_{\mathcal{X}} \left[\langle \varphi(x), (q(x)\Sigma + \lambda I)^{-1} \varphi(x) \rangle - \langle \varphi(x), (p(x)\Sigma + \lambda I)^{-1} \varphi(x) \rangle \right] p(x) d\tau(x) d\lambda \\
&= \int_0^{+\infty} \int_{\mathcal{X}} \left(\text{tr} \left[\Sigma(q(x)\Sigma + \lambda I)^{-1} \right] - \text{tr} \left[\Sigma(p(x)\Sigma + \lambda I)^{-1} \right] \right) p(x) d\tau(x) d\lambda \\
&= \int_{\mathcal{X}} \text{tr} \left[\Sigma(\log p(x)\Sigma - \log q(x)\Sigma) \right] p(x) d\tau(x) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} d\tau(x),
\end{aligned}$$

which is exactly the traditional KL divergence $D(p \parallel q)$. In order to show a bound similar to Eq. (9), a non-asymptotic extension of results from [38] could be carried out.

5 Estimation from finite sample

Given that the covariance operator is an infinite-dimensional operator, naively we would need eigenvalue decompositions of such operators to compute the kernel information quantities, which is computationally hard. Thus finite-dimensional estimation algorithms are needed. In this section, we provide algorithms for estimating the kernel entropy $H(\Sigma_p)$ from independent and identically distributed (i.i.d.) samples from p (as required for model fitting [13] or for independent component analysis [35]), as well as from certain kernel integrals. Estimators from other oracles on p or q , as well as estimators for $D(\Sigma_p \parallel \Sigma_q)$ could be derived and analyzed similarly.¹⁰

¹⁰With potentially extra terms due to the term $\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_q]$.

5.1 Estimators

Given $\Sigma_p = \int_X \varphi(x)\varphi(x)^* dp(x)$ and x_1, \dots, x_n sampled i.i.d. from p , we consider the natural estimator

$$\hat{\Sigma}_p = \frac{1}{n} \sum_{i=1}^n \varphi(x_i)\varphi(x_i)^*. \quad (10)$$

The natural estimate for $\text{tr} [\Sigma_p \log \Sigma_p]$ is $\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p]$, that can be computed from the kernel matrix $K \in \mathbb{R}^{n \times n}$ defined as $K_{ij} = k(x_i, x_j)$.

Proposition 6 (entropy for empirical covariance estimators) *With $\hat{\Sigma}_p$ defined in Eq. (10), and the kernel matrix defined above, we have*

$$\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] = \text{tr} \left[\frac{1}{n} K \log \left(\frac{1}{n} K \right) \right]. \quad (11)$$

Proof The non-zero eigenvectors of $\hat{\Sigma}_p$ belong to the image space of $\hat{\Sigma}_p$ and are thus linear combinations $f = \sum_{j=1}^n \alpha_j \varphi(x_j)$ for $\alpha \in \mathbb{R}^n$. Then $\hat{\Sigma}_p f = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \alpha_j [\varphi(x_i)\varphi(x_i)^*] \varphi(x_j) = \frac{1}{n} \sum_{i=1}^n (K\alpha)_i \varphi(x_i)$. Thus, if $K\alpha = n\lambda\alpha$, $\hat{\Sigma}_p f = \lambda f$, and if $\hat{\Sigma}_p f = \lambda f$ with $\lambda \neq 0$ and $f \neq 0$ (which implies $K\alpha \neq 0$), then $\sum_{i=1}^n [(K\alpha)_i - n\lambda\alpha_i] \varphi(x_i) = 0$, which implies $K^2 = n\lambda K\alpha$ and then $K\alpha = n\lambda\alpha$ since $K\alpha \neq 0$. Thus, the non-zero eigenvalues of $\hat{\Sigma}_p$ are exactly the ones of $\frac{1}{n}K$, and we thus get $\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] = \text{tr} \left[\frac{1}{n} K \log \left(\frac{1}{n} K \right) \right]$. ■

More generally, for any approximation of Σ_p as $\sum_{i=1}^n \eta_i \varphi(x_i)\varphi(x_i)^*$, with positive weights η_i that sum to one (for example obtained from a finite grid and not sampling), we would get

$$\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] = \text{tr} \left[\text{Diag}(\eta)^{1/2} K \text{Diag}(\eta)^{1/2} \log \left(\text{Diag}(\eta)^{1/2} K \text{Diag}(\eta)^{1/2} \right) \right].$$

Running-time complexity. In order to compute the eigenvalue decomposition of K above (needed to compute the matrix logarithm), we need a running time of $O(n^3)$. This can be reduced by using column sampling techniques [14], also known as “Nyström method” in this context [60]: by projecting the kernel matrix to the span of m of its columns, it leads to a running time in $O(m^2n)$, with some approximation that could be controlled explicitly [43].

5.2 Analysis

In order to quantify the difference between $\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p]$ and $\text{tr} [\Sigma_p \log \Sigma_p]$, we could use regular perturbation arguments for eigenvalues of self-adjoint operators [32]. However, the function $t \mapsto t \log t$ has diverging derivatives around zero, which prevents their direct use. Instead, given the integral representation in Eq. (5), we see that it will be sufficient to estimate quantities of the form $\text{tr} [(\Sigma_p + \lambda I)^{-1} \Sigma_p]$, which are usually referred so as “degrees of freedom” in kernel methods [21]. The natural estimator is

$\text{tr} [(\hat{\Sigma}_p + \lambda I)^{-1} \hat{\Sigma}_p]$ and its performance has already been thoroughly studied (see [45]), and an important quantity there is the maximal leverage score for the base distribution $\sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle$, which could be replaced by the the average one (which is then equal to the degrees of freedom) if “leverage score sampling” is used (see [21] for more details). For simplicity, we do not consider leverage score sampling, noting that for symmetric sets, it does not change anything.

The next proposition shows that these estimation results for fixed λ can be integrated over λ to get an estimation bound for the entropy (see proof in Appendix C.4), *without* the need for extra regularization, leading to an estimation in $O(1/\sqrt{n})$.

Proposition 7 *Assume (A1) and that p has a density with respect to the base measure which is greater than $\alpha < 1$. Assume that $c = \int_0^{+\infty} \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle^2 d\lambda$ is finite. Given i.i.d. samples x_1, \dots, x_n from p , and the estimator defined by Eq. (11), we have:*

$$\mathbb{E} \left[\left| \text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] - \text{tr} [\Sigma_p \log \Sigma_p] \right| \right] \leq \frac{1 + c(8 \log n)^2}{n\alpha} + \frac{17}{\sqrt{n}}(2\sqrt{c} + \log n).$$

Note that as a consequence of Jensen’s inequality, we always have $\mathbb{E}(\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p]) \geq \text{tr} [\Sigma_p \log \Sigma_p]$, as can be seen in simulations (see Figure 1).

5.3 Examples

We consider the d -dimensional torus $[0, 1]^d$ with a translation-invariant kernel $k(x, y) = k(x - y)$ with $\hat{k}(\omega) = \hat{k}(0)e^{-\sigma \|\omega\|_1}$. We have by symmetry $\langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle = \text{tr} [\Sigma(\Sigma + \lambda I)^{-1}] = \sum_{\omega \in \mathbb{Z}^d} \frac{\hat{k}(\omega)}{\hat{k}(\omega) + \lambda}$.

As shown in Appendix C.6, we have the bound $\text{tr} [\Sigma(\Sigma + \lambda I)^{-1}] \leq e^{-\sigma d/2} \frac{d!}{\sinh^d(\sigma/2)} \left[1 + \left(\log \frac{\tanh^d \frac{\sigma}{2}}{\lambda} \right)^d \right]$.

Thus, the constant c in Prop. 7 above can be upper bounded by a constant times σ^{-d} (see Appendix C.6 for details). We thus get an overall estimation rate for the kernel entropy proportional to $\sigma^{-d/2} \sqrt{n}$.

When balancing with the $O(\sigma^2)$ estimation of the entropy for Lipschitz-continuous strictly positive densities in Section 4.2, we can take $\sigma \propto n^{-1/(d+4)}$ (lower bandwidth with more observations), leading to a rate of $n^{-2/(d+4)}$, which is to be compared to the optimal rate equal to $n^{-4/(d+4)}$ for entropy estimation for densities with the same regularity [28]. Although our primary goal was not to estimate entropies, we get a rate which is the square root of the optimal rate.

Illustrative experiments. We consider $\mathcal{X} = [0, 1]$ and $p(x) = 4|x - \frac{1}{2}|$, with negative differential entropy

$\int_0^1 p(x) \log p(x) dx = \log 2 - \frac{1}{2} \approx 0.1931$. We consider estimation from various numbers of i.i.d. observations. See left plot in Figure 1, where we compare two values of σ . Note that the estimation is empirically here in expectation from above, as expected because of Jensen’s inequality (as opposed to the right plot).

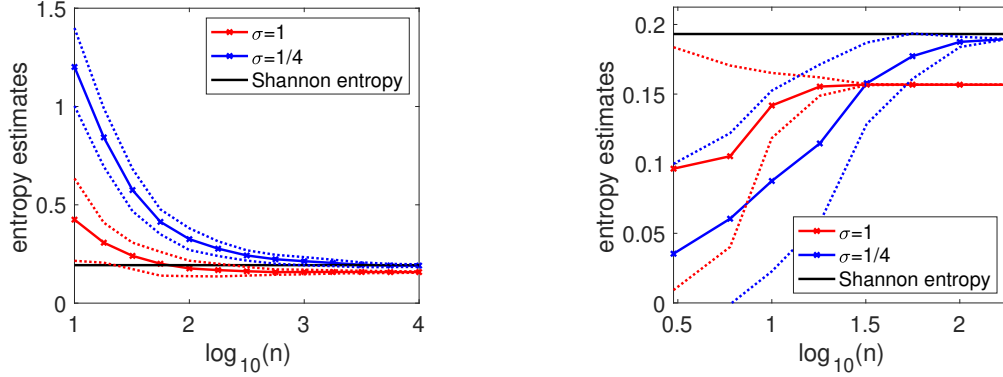


Figure 1: Negative entropy estimation from 20 replications (mean and standard deviation) for two values of the kernel bandwidth σ , as n increases. Left: from i.i.d. samples (method from Section 5.1). Right: from kernel integrals evaluated as uniform samples (method from Section 5.4). As expected we estimate the kernel entropy (limit when n tends to infinity) from above on the left plot, and from below on the right plot, with a faster convergence rate on the right plot.

5.4 Estimation from other oracles

If we have access to other oracles on p and q , we can consider other estimators.¹¹ For example, if we can estimate $\int_{\mathcal{X}} k(x, y)k(x, z)dp(x)$ and $\int_{\mathcal{X}} k(x, y)k(x, z)dq(x)$ for any $y, z \in \mathcal{X}$, we can simply sample n points x_1, \dots, x_n from the base measure on \mathcal{X} , and, given the projection operator $\hat{\Pi}$ on the span of $\varphi(x_1), \dots, \varphi(x_n)$, consider

$$\hat{\Sigma}_p = \hat{\Pi}\Sigma_p\hat{\Pi} \text{ and } \hat{\Sigma}_q = \hat{\Pi}\Sigma_q\hat{\Pi}, \quad (12)$$

and the estimator $D(\hat{\Sigma}_p\|\hat{\Sigma}_q)$ which is now a pointwise lower bound¹² on $D(\Sigma_p\|\Sigma_q)$ (and thus on $D(p\|q)$), which can be beneficial in variational inference, as presented in Section 7.

Noting that $\langle \varphi(x_i), \Sigma_p \varphi(x_j) \rangle = \int_{\mathcal{X}} k(x_i, x)k(x, x_j)dp(x)$ (and similarly for Σ_q), and that K is the matrix of dot-products of $\varphi(x_1), \dots, \varphi(x_n)$, the orthogonal projection may be done as

$$D(\hat{\Sigma}_p\|\hat{\Sigma}_q) = D\left(K^{-1/2} \int_{\mathcal{X}} k(X, x)k(x, X)dp(x)K^{-1/2} \parallel K^{-1/2} \int_{\mathcal{X}} k(X, x)k(x, X)dq(x)K^{-1/2}\right),$$

where $k(X, x) \in \mathbb{R}^n$ is the vector with components $k(x_i, x)$, $i = 1, \dots, n$, and $K \in \mathbb{R}^{n \times n}$ the kernel matrix, typically with some added regularization for numerical stability, and incomplete Cholesky decomposition so as to minimize the number of kernel evaluations (see, e.g., [7] and references therein), with an overall running time complexity of $O(n^3)$. A key property is that it always underestimates the kernel relative entropy (and hence the Shannon relative entropy). The following proposition shows that it provides a better estimate (better dependence in n) than with i.i.d. samples (see proof in Appendix C.5).

Proposition 8 *Assume (A1) and that p has a density with respect to the base measure which is bounded. Given uniform i.i.d. samples x_1, \dots, x_n from the base distribution, and the estimator defined by Eq. (12),*

¹¹This section is not crucial to understand the rest of the paper.

¹²Since $\hat{\Pi}\hat{\Pi}^* + (I - \hat{\Pi})(I - \hat{\Pi})^* = I$, the monotonicity of relative entropy (Prop. 14 in Appendix A.2) leads to $D(\Sigma_p\|\Sigma_q) \geq D(\hat{\Pi}\Sigma_p\hat{\Pi} + (I - \hat{\Pi})\Sigma_p(I - \hat{\Pi})\|\hat{\Pi}\Sigma_q\hat{\Pi} + (I - \hat{\Pi})\Sigma_q(I - \hat{\Pi})) = D(\hat{\Pi}\Sigma_p\hat{\Pi}\|\hat{\Pi}\Sigma_q\hat{\Pi}) + D((I - \hat{\Pi})\Sigma_p(I - \hat{\Pi})\|(I - \hat{\Pi})\Sigma_q(I - \hat{\Pi}))$, which is greater than $D(\hat{\Pi}\Sigma_p\hat{\Pi}\|\hat{\Pi}\Sigma_q\hat{\Pi})$.

we have, if for all $\lambda > 0$, $\sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma + \lambda I)^{-1} \varphi(x) \rangle \leq B \log(C/\lambda)^d$, for some constants B, C :

$$0 \leq \text{tr} [\hat{\Pi} \Sigma_p \hat{\Pi} \log(\hat{\Pi} \Sigma_p \hat{\Pi})] - \text{tr} [\Sigma_p \log \Sigma_p] = O\left(\left(\log(BC)\right)^d (\sqrt{BC} + B) \exp\left(-\frac{1}{4}(Bn)^{1/(d+1)}\right)\right).$$

The proposition above applies to the torus situation in $[0, 1]^d$ and the kernel with exponential Fourier series, with now an improved dependence in n and in d , since in that case $B \sim C \sim \sigma^d$. That is, to reach precision ε , we need $n \propto \sigma^{-d} \left(\log \frac{\sigma^d}{\varepsilon}\right)^{d+1}$ instead of $n \sim \sigma^{-d} \varepsilon^{-2}$. Other decays for the maximal degrees of freedom are considered in Appendix C.5.

Illustrative experiments. We consider as above $p(x) = 4|x - \frac{1}{2}|$, and consider estimation from various numbers of i.i.d. observations for the projection. The required integrals are computed by numerical integration. See right plot in Figure 1, where we compare two values of σ . Note that the estimation is as expected here from below, and with a faster rate of convergence than for the estimation method from Section 5.1.

6 Multivariate probabilistic modelling

Information-theoretic quantities are particularly useful at characterizing statistical dependence between random variables. This is for example the basis of independent component analysis [17], and can be used for feature selection [51]. We explore how the new notions of entropy can be used in this goal.

6.1 Joint entropy, marginal entropy and mutual information

We assume that we have a product set $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, and a feature space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$, corresponding to the kernel $k(x, y) = k_1(x_1, y_1)k_2(x_2, y_2)$, and feature map $\varphi(x) = \varphi_1(x_1) \otimes \varphi_2(x_2)$. We also consider a base distribution on $\mathcal{X}_1 \times \mathcal{X}_2$ which is the product of the base distributions. The tensor product is here defined as the span of all functions $(x_1, x_2) \mapsto f_1(x_1)f_2(x_2)$ for $f_i \in \mathcal{H}_i$, $i = 1, 2$, with the norm $\|f_1\| \cdot \|f_2\|$. Given a joint probability $p_{X_1 X_2}$ and its marginal probability distributions p_{X_1} and p_{X_2} , we can define the covariance operators $\Sigma_{p_{X_1 X_2}}$ on $\mathcal{H}_1 \otimes \mathcal{H}_2$, $\Sigma_{p_{X_1}}$ on \mathcal{H}_1 , and $\Sigma_{p_{X_2}}$ on \mathcal{H}_2 , as well as Σ_{12} , Σ_1 and Σ_2 the covariance operators of the base distributions on $\mathcal{X}_1 \times \mathcal{X}_2$, \mathcal{X}_1 and \mathcal{X}_2 :

We then define the marginal and joint entropies as:

$$\begin{aligned} H(\Sigma_{p_{X_1}}) &= -\text{tr} [\Sigma_{p_{X_1}} \log \Sigma_{p_{X_1}}] + \text{tr} [\Sigma_{p_{X_1}} \log \Sigma_1] - \min_{x_1 \in \mathcal{X}_1} \langle \varphi_1(x_1), (\log \Sigma_1) \varphi_1(x_1) \rangle \\ H(\Sigma_{p_{X_2}}) &= -\text{tr} [\Sigma_{p_{X_2}} \log \Sigma_{p_{X_2}}] + \text{tr} [\Sigma_{p_{X_2}} \log \Sigma_2] - \min_{x_2 \in \mathcal{X}_2} \langle \varphi_2(x_2), (\log \Sigma_2) \varphi_2(x_2) \rangle \\ H(\Sigma_{p_{X_1 X_2}}) &= -\text{tr} [\Sigma_{p_{X_1 X_2}} \log \Sigma_{p_{X_1 X_2}}] + \text{tr} [\Sigma_{p_{X_1 X_2}} \log \Sigma_{12}] - \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma_{12}) \varphi(x) \rangle. \end{aligned}$$

The product distribution $p_{X_1} p_{X_2}$ (with independent components with same marginal distributions as X), which is different in general from the joint distribution $p_{X_1 X_2}$, has covariance operator $\Sigma_{p_{X_1} p_{X_2}} = \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}}$, and its entropy is:

$$H(\Sigma_{p_{X_1} p_{X_2}}) = -\text{tr} [\Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}} \log \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}}] + \text{tr} [\Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}} \log \Sigma_{12}] - \min_{x \in \mathcal{X}} \langle \varphi(x), (\log \Sigma_{12}) \varphi(x) \rangle.$$

We have $\log \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}} = \log \Sigma_{p_{X_1}} \otimes I + \log I \otimes \Sigma_{p_{X_2}}$. Moreover, because the base distribution is separable by assumption, we have $\langle \varphi(x), (\log \Sigma_{12})\varphi(x) \rangle = \langle \varphi_1(x_1), (\log \Sigma_1)\varphi_1(x_1) \rangle + \langle \varphi_2(x_2), (\log \Sigma_2)\varphi_2(x_2) \rangle$, thus, we exactly have:

$$H(\Sigma_{p_{X_1}p_{X_2}}) = H(\Sigma_{p_{X_1}}) + H(\Sigma_{p_{X_2}}).$$

We can therefore define the kernel mutual information (not to be confused with the regular Shannon mutual information) as

$$I(X_1, X_2) = D(\Sigma_{p_{X_1}X_2} \| \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}}).$$

From the developments above, it is equal to

$$I(X_1, X_2) = H(\Sigma_{p_{X_1}}) + H(\Sigma_{p_{X_2}}) - H(\Sigma_{p_{X_1}X_2}),$$

mimicking the traditional equality for regular entropies. Moreover, it is equal to zero if and only if the variables are independent.

Proposition 9 (Characterization of independence) *Assume (A1). Then $I(X_1, X_2)$ is a lower bound on the Shannon mutual information, and $I(X_1, X_2) = 0$ if and only if $\Sigma_{p_{X_1}X_2} = \Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}}$. If the kernel k^2 is universal, this is equivalent to the random variables X_1 and X_2 being independent.*

Proof The first statement is straightforward because of the interpretation as a relative entropy. The second statement is simply the consequence of the fact that $\Sigma_{p_{X_1}} \otimes \Sigma_{p_{X_2}}$ is the covariance operator of the product distribution $p_{X_1}p_{X_2}$. ■

We can make the following observations:

- The extension to mutual information between more than two random variables is straightforward by considering tensor products between all marginals, with the same characterization property of independence.
- Our use of covariance operators is related to the “kernel generalized variance” from [6], which considers a joint covariance operator $\begin{pmatrix} \mathbb{E}[\varphi_1(X_1)\varphi_1(X_1)^*] & \mathbb{E}[\varphi_1(X_1)\varphi_2(X_2)^*] \\ \mathbb{E}[\varphi_2(X_2)\varphi_1(X_1)^*] & \mathbb{E}[\varphi_2(X_2)\varphi_2(X_2)^*] \end{pmatrix} = \mathbb{E} \left[\begin{pmatrix} \varphi_1(X_1) \\ \varphi_2(X_2) \end{pmatrix} \begin{pmatrix} \varphi_1(X_1) \\ \varphi_2(X_2) \end{pmatrix}^* \right]$, and compute the mutual information for a Gaussian vector with this covariance operator. The kernel generalized variance can be generalized to more than two random variables, but then the independence characterization property is not satisfied. Moreover, the relationship with the usual mutual information is not as straightforward.
- The extension to conditional entropies is not straightforward. We could define $\Sigma_{p_{X_2|X_1=x_1}}$ in the natural way, but then we do not have the joint entropy being the sum of the marginal and conditional entropy (otherwise, because of axiomatic characterizations of entropy [20], we would get back exactly the Shannon entropy). There is also a potential for exponential families and conditional models akin to generalized models, which remains to be explored.

6.2 Data processing inequality

We have a form of data processing inequality, which is classical in quantum information theory (see, e.g., [40]), which we extend here.

Proposition 10 (Data processing inequality) *Assume (A1). We have*

$$D(\Sigma_{p_{X_1 X_2}} \parallel \Sigma_{q_{X_1 X_2}}) \geq D(\Sigma_{p_{X_1}} \parallel \Sigma_{q_{X_1}}),$$

with equality implying (when the kernel k is universal) that the conditional distribution of X_2 given X_1 is the same for p and q , but the reverse implication is not true in general.

Proof The marginal operators $\Sigma_{p_{X_1}}$ and $\Sigma_{q_{X_1}}$ are obtained by taking “partial traces” which are quantum operations as defined in Section 3, hence the inequality. Following [46], we have inequality if and only if $\log \Sigma_{p_{X_1 X_2}} - \log \Sigma_{q_{X_1 X_2}} = \log(\Sigma_{p_{X_1}} \otimes I) - \log(\Sigma_{q_{X_1}} \otimes I)$, which is unfortunately not always satisfied when conditional distributions are equal. However, in case of equality, the extra condition (c) from Prop. 2, becomes $\Sigma_{p_{X_1 X_2}}(\Sigma_{q_{X_1}} \otimes I) = \Sigma_{q_{X_1 X_2}}(\Sigma_{p_{X_1}} \otimes I)$, and can be leveraged. Indeed, for all test functions f_1, g_1, f_2, g_2 in \mathcal{H} , it leads to:

$$\begin{aligned} & \int_{X_1 \times X_2} \int_{X_1} f_1(x_1) f_2(x_2) g_1(y_1) g_2(y_2) k_1(x_1, y_1) dp_{X_1 X_2}(x_1, x_2) dq_{X_1}(y_1) \\ &= \int_{X_1 \times X_2} \int_{X_1} f_1(x_1) f_2(x_2) g_1(y_1) g_2(y_2) k_1(x_1, y_1) dq_{X_1 X_2}(x_1, x_2) dp_1(y_1). \end{aligned}$$

Considering f_1 and g_1 tending to Dirac functions at \bar{x}_1 (which is possible since the kernel is universal) we get $\int_{X_2} f_2(x_2) g_2(y_2) dp_{X_1 X_2}(\bar{x}_1, x_2) dq_{X_1}(\bar{x}_1) = \int_{X_2} f_2(x_2) g_2(y_2) dq_{X_1 X_2}(\bar{x}_1, x_2) dp_1(\bar{x}_1)$, which implies that the conditional distributions of x_2 given x_1 is the same for p and q . \blacksquare

6.3 Submodularity and conditional independence

We can now consider three variables X_1, X_2, X_3 , and use strong sub-additivity of the quantum relative entropy [46, 40].

Proposition 11 (Submodularity) *Assume (A1). We have:*

$$H(\Sigma_{p_{X_1 X_2 X_3}}) - H(\Sigma_{p_{X_1 X_2}}) - H(\Sigma_{p_{X_2 X_3}}) + H(\Sigma_{p_{X_2}}) \leq 0,$$

with equality (for universal kernels) implying that X_1 and X_3 are independent given X_2 , but in general the reverse implication is not true.

Proof As done by [40], we can simply apply the data processing result (Prop. 10) to $\tilde{X}_1 = (X_1, X_2)$, and $\tilde{X}_2 = X_3$, with the distributions $\tilde{p}_{\tilde{X}_1 \tilde{X}_2} = p_{X_1 X_2 X_3}$ and $\tilde{q}_{\tilde{X}_1 \tilde{X}_2} = p_{X_1} p_{X_2 X_3}$. \blacksquare

As a classical consequence, the entropy function is submodular like the regular entropy (see, e.g., [3, Section 6.5]).

Note that we do not obtain a necessary and sufficient characterization of conditional independence, which can be obtained with other tools based on covariance operators [27].

7 Convex duality, log-partition functions and variational inference

Given that our kernel notions of relative entropies are lower bounds on the regular notions, by the traditional convex duality results between maximum entropy and log-partition functions (see, e.g., [58]), we should obtain upper-bounds on log-partition functions. In this section, we show how such bounds can be obtained and computed. We then show that such bounds can also be optimized with respect to the positive definite kernel.

This leads to a new family of variational inference methods, which can then be used in various probabilistic inference tasks [58], and in particular in Bayesian inference [42].

In this section, we will sometimes only assume that for our kernels satisfy (on top of being positive definite) $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, noting that the upper-bound result on the regular Shannon relative entropy is still valid in this case.

7.1 Convex duality between operators

We have for self-adjoint operators (regardless of the link with input spaces and probabilities):

$$\begin{aligned} \sup_{A \succ 0, \operatorname{tr} A = 1} \operatorname{tr}(AM) - D(A\|B) &= \sup_{A \succ 0, \operatorname{tr} A = 1} \operatorname{tr}(A(M + \log B)) - \operatorname{tr}[A \log A] \\ &= \log \operatorname{tr} \exp(M + \log B), \end{aligned}$$

with the optimal operator A equal to $A = \frac{1}{\operatorname{tr} \exp(M + \log B)} \exp(M + \log B)$. This implies the representation

$$D(A\|B) = \sup_M \operatorname{tr} AM - \log \operatorname{tr} \exp(M + \log B),$$

with the constraints that $A \succ 0$ and $\operatorname{tr} A = 1$ are automatically satisfied (if $\operatorname{tr} A \neq 1$, then by replacing M by $M + tI$, we can make the quantity go to infinity).

We also have (when the constraint for unit trace is removed):

$$\begin{aligned} \sup_{A \succ 0} \operatorname{tr}(AM) - D(A\|B) &= \sup_{A \succ 0} \operatorname{tr}(A(M + \log B)) - \operatorname{tr}[A \log A] \\ &= \operatorname{tr} \exp(M + \log B - I), \end{aligned}$$

with an optimal $A = \frac{1}{e} \exp(M + \log B)$. This implies the representation

$$D(A\|B) = \sup_M \operatorname{tr} A + \operatorname{tr} AM - \operatorname{tr} \exp(M + \log B),$$

with the constraint that $A \succ 0$ is automatically satisfied (but not $\operatorname{tr} A = 1$), with the optimal M equal to $\log A - \log B$.

7.2 Bounds on the log-partition function

We can now apply the duality relationships above to covariance operators. Given a bounded function $f : \mathcal{X} \rightarrow \mathbb{R}$ and a distribution q on \mathcal{X} , our goal is to obtain upper-bounds on the log-partition function

$\log \int_{\mathcal{X}} e^{f(x)} dq(x)$ which is equal to, by convex duality:

$$\log \int_{\mathcal{X}} e^{f(x)} dq(x) = \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - D(p||q).$$

In our situation, our operators are subject to belong to the subspace \mathcal{A} , equal to the span of all $\varphi(x)\varphi(x)^*$, for $x \in \mathcal{X}$, which leads to slight adaptations. We do not necessarily assume that the kernels are universal, so that the hull of all $\varphi(x)\varphi(x)^*$ may not be equal to the intersection of \mathcal{A} and positive operators with unit trace.

Isotropic kernels. If $k(x, x) = 1$, for all $x \in \mathcal{X}$, then $\text{tr } \Sigma_p = 1$ is equivalent to $\int_{\mathcal{X}} dp(x) = 1$. We define for a bounded function $f : \mathcal{X} \rightarrow \mathbb{R}$,

$$a(f) = \sup_{p \text{ probability measure}} \int_{\mathcal{X}} f(x) dp(x) - D(\Sigma_p || \Sigma_q).$$

Using convex duality tools from Section 7.1, we have:

$$\begin{aligned} a(f) &= \sup_{p \text{ non-negative measure}} \inf_M \int_{\mathcal{X}} f(x) dp(x) - \text{tr } \Sigma_p M + \log \text{tr } \exp(M + \log \Sigma_q) \\ &= \inf_M \log \text{tr } \exp(M + \log \Sigma_q) \text{ such that } \forall x \in \mathcal{X}, f(x) \leq \langle \varphi(x), M \varphi(x) \rangle. \end{aligned}$$

Note that we can get rid of the constraint $\int_{\mathcal{X}} dp(x) = 1$ because it is equivalent to $\text{tr } \Sigma_p = 1$.

A relaxation is to relax p not to be a non-negative measure (and use the positivity of Σ_p as the only constraint), leading to

$$b(f) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - D(\Sigma_p || \Sigma_q). \quad (13)$$

By construction, we have $a(f) \leq b(f)$, and by convex duality,

$$\begin{aligned} b(f) &= \sup_{p \text{ measure}} \inf_M \int_{\mathcal{X}} f(x) dp(x) - \text{tr } \Sigma_p M + \log \text{tr } \exp(M + \log \Sigma_q) \\ &= \inf_M \log \text{tr } \exp(M + \log \Sigma_q) \text{ such that } \forall x \in \mathcal{X}, f(x) = \langle \varphi(x), M \varphi(x) \rangle. \end{aligned} \quad (14)$$

Note that that if the problem above is non-feasible because φ is too small (so that f cannot be represented as a quadratic form in φ), then we get a value $+\infty$.

If we can write (usually non-uniquely), $f(x) = \langle \varphi(x), F \varphi(x) \rangle$ for some self-adjoint operator F , then we have another representation as:

$$b(f) = \inf_{N \in \mathcal{A}^\perp} \log \text{tr } \exp(F + N + \log \Sigma_q).$$

Since $D(\Sigma_p || \Sigma_q) \leq D(p || q)$, we have:

$$b(f) \geq a(f) \geq \sup_{p \text{ probability}} \int_{\mathcal{X}} f(x) dp(x) - D(p || q) = \log \int_{\mathcal{X}} e^{f(x)} dq(x).$$

As expected, we thus obtain upper bounds on the usual log-partition function.

Note that we could also get lower bounds using tools from Section 4.1, in particular for functions f that can be written as $f(x) = \int_{\mathcal{X}} h(x, y) g(y) d\tau(y)$.

Non isotropic kernels. If we only assume that $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, then $b(f)$ defined in Eq. (13) and Eq. (14) is not any more an upper-bound but we can define instead

$$\begin{aligned}
b^{(c)}(f) &= \sup_{p \text{ measure, } \int_{\mathcal{X}} dp(x)=1, \int_{\mathcal{X}} k(x,x)dp(x) \leq 1} \inf_M \int_{\mathcal{X}} f(x)dp(x) - \text{tr } \Sigma_p M - \text{tr } \Sigma_p + \text{tr exp}(M + \log \Sigma_q) \\
&= \sup_{p \text{ measure}} \inf_{M, c, b \geq 0} \int_{\mathcal{X}} f(x)dp(x) - \text{tr } \Sigma_p M - \text{tr } \Sigma_p + \text{tr exp}(M + \log \Sigma_q) \\
&\quad - c \left(\int_{\mathcal{X}} dp(x) - 1 \right) - b \left(\int_{\mathcal{X}} k(x, x) dp(x) - 1 \right) \\
&= \inf_{M, c, b \geq 0} c + b + \text{tr exp}(M + \log \Sigma_q) \text{ such that } \forall x \in \mathcal{X}, f(x) = c + \langle \varphi(x), (M + I + bI)\varphi(x) \rangle \\
&= \inf_{M, c} c + \inf_{b \geq 0} (b + e^{-b} \text{tr exp}(M + \log \Sigma_q)) \text{ such that } \forall x \in \mathcal{X}, f(x) = c + \langle \varphi(x), (M + I)\varphi(x) \rangle \\
&= \inf_{M, c} c + 1 + \widetilde{\log} \text{tr exp}(M + \log \Sigma_q) \text{ such that } \forall x \in \mathcal{X}, f(x) = c + \langle \varphi(x), (M + I)\varphi(x) \rangle \\
&= \inf_{M, c} c + 1 + \widetilde{\log} \left(\frac{1}{e} \text{tr exp}(M + \log \Sigma_q) \right) \text{ such that } \forall x \in \mathcal{X}, f(x) = c + \langle \varphi(x), M\varphi(x) \rangle,
\end{aligned}$$

with $\widetilde{\log}(a) = \log a$ if $a > 1$, and $a - 1$ otherwise.

If f can be represented through the operator F , and the constant function through the operator U , then we get

$$b^{(c)}(f) = \inf_{M, c} c + 1 + \widetilde{\log} \left(\frac{1}{e} \text{tr exp}(M + \log \Sigma_q) \right) \text{ such that } F - cU - M \in \mathcal{A}^\perp.$$

We can check that if $U = I$, then the constraints in the definition of $b^{(c)}(f)$ are equivalent to $\text{tr } \Sigma_p$, and we exactly recover the expression of $b(f)$ in Eq. (14).

Properties of upper-bounds. The definition of upper-bounds on the log-partition function naturally leads to several questions and applications:

- Can these upper-bounds be tight? In situations like in Section 4.2 where the entropy lower-bound can be made tight (for example when the kernel bandwidth tends to zero), we should also get tight upper-bounds on the log-partition functions, that is, upper-bounds that tend to the classical log-partition function when the bandwidth goes to zero. We leave for future work a detailed study of these approximations.
- Can these upper-bounds be estimated efficiently from simple oracles on the function f ? We provide efficient algorithms in Section 7.4.
- Can these upper-bounds be minimized? We show in Section 7.5 that these upper-bounds happen to be convex in the kernel, so minimizing with respect to the kernel is possible.

7.3 Relationship with optimization

When adding a temperature parameter $\varepsilon > 0$, we can extend the traditional link between optimization and log-partition functions. This corresponds essentially to considering a potential $\frac{1}{\varepsilon}f$, or defining

$$b_\varepsilon(f) = \varepsilon b\left(\frac{1}{\varepsilon}f\right) = \sup_{p \text{ measure}} \int_{\mathcal{X}} f(x) dp(x) - \varepsilon D(\Sigma_q \| \Sigma_q).$$

By convex duality, we get:

$$b_\varepsilon(f) = \inf_M \varepsilon \log \text{tr} \exp\left(\frac{1}{\varepsilon}M + \log \Sigma_q\right) \text{ such that } \forall x \in \mathcal{X}, f(x) = \langle \varphi(x), M\varphi(x) \rangle.$$

When ε tends to zero, then $b_\varepsilon(f)$ converges to

$$\inf_M \lambda_{\max}(M) \text{ such that } \forall x \in \mathcal{X}, f(x) = \langle \varphi(x), M\varphi(x) \rangle.$$

Given that $\|\varphi(x)\| = 1$ for all $x \in \mathcal{X}$, by writing $M = cI - A$ for A positive, this is equal to

$$\inf_{c \in \mathbb{R}, A \succ 0} c \text{ such that } \forall x \in \mathcal{X}, f(x) = c - \langle \varphi(x), A\varphi(x) \rangle,$$

which is exactly the optimization formulation of [44] based on “kernel sums-of-squares”. Following the traditional relationship between log-partition functions and maxima, we can thus consider our bound on log-partition functions as a smoothed version on the maximum. Therefore, some of the approximation techniques developed for the optimization formulation [44] can be extended to our set-up as well. Moreover, as common in convex optimization, the entropy could be used for smoothing and accelerated optimization algorithms [37].

7.4 Computable bounds

In order to compute or approximate $b(f)$, which depends on the function $f : \mathcal{X} \rightarrow \mathbb{R}$ and the feature map $\varphi : \mathcal{X} \rightarrow \mathcal{H}$, we will consider a finite-dimensional approximation $\tilde{\varphi} : \mathcal{X} \rightarrow \tilde{\mathcal{H}}$ where $\tilde{\mathcal{H}}$ will be a finite-dimensional subspace of \mathcal{H} . All of our approximations will always be upper-bounds on the true log-partition function.

We will always need an efficient representation of $\tilde{\mathcal{A}}$ the span of all $\tilde{\varphi}(x)\tilde{\varphi}(x)^*$, for $x \in \mathcal{X}$, and sometimes of $\tilde{\mathcal{M}}$ the hull of all $\tilde{\varphi}(x)\tilde{\varphi}(x)^*$ for $x \in \mathcal{X}$. See examples in Section 7.6 and Section 7.7 for the torus and the hypercube.

In order to estimate the relative entropy $D(\Sigma_p \| \Sigma_q)$, we will use $D(\tilde{\Sigma}_p \| \tilde{\Sigma}_q)$ instead. If the kernel $(x, y) \mapsto k(x, y) - \tilde{k}(x, y)$ is positive definite, then this is a lower bound, and thus the bounds on the regular quantities are preserved.

In order to compute $D(\tilde{\Sigma}_p \| \tilde{\Sigma}_q)$, we assume we can compute explicitly $\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$ and $\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dq(x)$. If $\tilde{\varphi}$ is obtained from projection on a span of some $\varphi(x_1), \dots, \varphi(x_n)$ like in Section 5.4, then this can be done if we can compute expectations of $k(x, y)k(x, z)$ for any $y, z \in \mathcal{X}$. We can also build set-specific approximate feature maps (as done in Section 7.6 and Section 7.7).

With an explicit representation of f . If we have an explicit (non-unique) representation of f as $f(x) = \langle \tilde{\varphi}(x), \tilde{F}\tilde{\varphi}(x) \rangle$. We then have if $\langle \tilde{\varphi}(x), \tilde{\varphi}(x) \rangle = 1$ for all $x \in \mathcal{X}$ (a similar approximation can be obtained for $b^{(e)}(f)$ for non-isotropic approximations):

$$\begin{aligned} \tilde{b}(f) &= \sup_{\tilde{C} \in \mathcal{A}, \tilde{C} \succ 0, \text{tr } \tilde{C} = 1} \text{tr}[C\tilde{F}] - D(C\|\tilde{\Sigma}_q) \\ &= \inf_{\tilde{M}} \log \text{tr} \exp(\tilde{M} + \log \tilde{\Sigma}_q) + \sup_{\tilde{C} \in \mathcal{A}} \text{tr}[\tilde{C}(F - \tilde{M})] \\ &= \inf_{\tilde{M}} \log \text{tr} \exp(\tilde{M} + \log \tilde{\Sigma}_q) \text{ such that } \tilde{M} - \tilde{F} \in \mathcal{A}^\perp, \end{aligned}$$

which is a finite-dimensional convex optimization problem. Since the function $\tilde{M} \mapsto \log \text{tr} \exp(\tilde{M} + \log \tilde{\Sigma}_q)$ is smooth, the optimization problem above can be approximately solved by projected gradient descent (with unit step-size), with iteration:

$$\tilde{M} \leftarrow \Pi_{\tilde{F} + \mathcal{A}^\perp} \left(\tilde{M} - \frac{\exp(\tilde{M} + \log \tilde{\Sigma}_q)}{\text{tr} \exp(\tilde{M} + \log \tilde{\Sigma}_q)} \right).$$

We then obtain the (approximated) optimal \tilde{C} as $\tilde{C} = \frac{\exp(\tilde{M} + \log \tilde{\Sigma}_q)}{\text{tr} \exp(\tilde{M} + \log \tilde{\Sigma}_q)}$. It can be accelerated using extrapolation [8]. Note that for any feasible \tilde{M} we obtain a lower bound on $\tilde{b}(f)$.

Without an explicit representation of f . If an explicit representation of f as $f(x) = \langle \tilde{\varphi}(x), \tilde{F}\tilde{\varphi}(x) \rangle$ is not available, we can build an approximate representation of f such that $\|f - \langle \tilde{\varphi}(\cdot), \tilde{F}\tilde{\varphi}(\cdot) \rangle\|_\infty$ is as small as possible, and then use the previous algorithm, with this extra approximation factor. On the torus, we could also use the same technique as [61] based on approximating the L_∞ -norm with Fourier transforms.

Illustrative experiments. We consider here $\mathcal{X} = [0, 1]$, and the kernel with Fourier transforms proportional to $e^{-\sigma|\omega|}$. We consider the function $f(x) = \cos(2\pi x)$, and approximate $\log(\int_0^1 e^{f(x)} dx) \approx 0.2359$. We could either use the finite feature map from Section 5.4 based on random projections (which is a non-isotropic approximation). For simplicity, we consider $\tilde{\varphi}(x)_\omega = \hat{k}(\omega)e^{2i\pi\omega x}$, only for $\omega \in \{-r, \dots, r\}$. We have $\|\tilde{\varphi}(x)\|^2 = \sum_{\omega=-r}^r \tanh(\frac{\sigma}{2})e^{-\sigma|\omega|} = \tanh(\frac{\sigma}{2}) \left(-1 + 2\frac{1-e^{-\sigma(r+1)}}{1-e^{-\sigma}} \right) = 1 - \frac{e^{-\sigma(r+1/2)}}{\cosh(\frac{\sigma}{2})}$, and the set \mathcal{A} is exactly the set of Toeplitz matrices. See Figure 2 for an illustration with several values of σ .

7.5 Kernel learning

The key property we will leverage is the *concavity* of $D(\Sigma_p\|\Sigma_q)$ with respect to the kernel. This allows to maximize efficiently with respect to the kernel to get as close to the Shannon relative entropy $D(p\|q)$ as possible.

Concavity with respect to the kernel. Given two distributions p and q , then $D(\Sigma_p\|\Sigma_q)$ depends on the kernel function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. This happens to be a *concave* function of k (which is only assumed to be positive definite with no bounds on $k(x, x)$). Indeed, if k_1 and k_2 are two kernels with feature maps $\varphi_1 : \mathcal{X} \rightarrow \mathcal{H}_1$ and $\varphi_2 : \mathcal{X} \rightarrow \mathcal{H}_2$, with the associated covariance operators $\Sigma_p^{(1)}, \Sigma_p^{(2)}, \Sigma_q^{(1)}, \Sigma_q^{(2)}$,

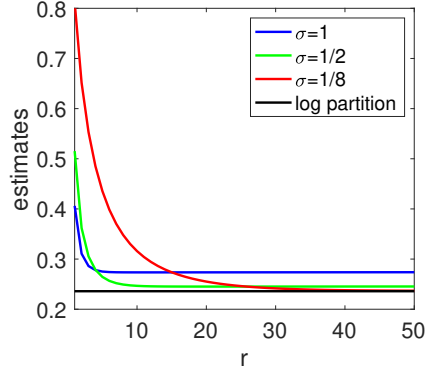


Figure 2: Log-partition estimation with different values of σ , as a function of r , the number of considered frequencies.

and $\eta_1, \eta_2 \geq 0$ such that $\eta_1 + \eta_2 = 1$, then with $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ and $\varphi(x) = (\sqrt{\eta_1}\varphi_1(x), \sqrt{\eta_2}\varphi_2(x))$, using $E_1 = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ and $E_2 = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}$, using the monotonicity of relative entropy for quantum operations, we have:

$$\eta_1 D(\Sigma_p^{(1)} \|\Sigma_q^{(1)}) + \eta_2 D(\Sigma_p^{(2)} \|\Sigma_q^{(2)}) = D(E_1 \Sigma_p E_1 + E_2 \Sigma_p E_2 \|\ E_1 \Sigma_q E_1 + E_2 \Sigma_q E_2) \leq D(\Sigma_p \|\Sigma_q),$$

which shows the concavity.

Moreover, the relative entropy is monotonic, that is, if $k_1 - k_2$ is positive definite, $D(\Sigma_p^{(2)} \|\Sigma_q^{(2)}) \leq D(\Sigma_p^{(1)} \|\Sigma_q^{(1)})$. Therefore, when optimizing over the kernel, we only need to look at maximal elements with respect to the positive definite order.

Linear parametrization of kernels. Given a finite-dimensional feature map $\psi : \mathcal{X} \rightarrow \mathbb{R}^d$, we look at kernels of the form

$$k(x, y) = \psi(x)^\top \Lambda \psi(y),$$

where $\Lambda \in \mathbb{R}^{d \times d}$ is a symmetric positive semidefinite matrix. In order to have the upper-bound on entropy, we consider the set \mathcal{L} of matrices Λ such that for all $x \in \mathcal{X}$, $\psi(x)^\top \Lambda \psi(x) \leq 1$. This is a finite-dimensional convex set, for which we need to know *inner* approximations (to preserve directions of bounds). We then have that

$$D(\Sigma_p \|\Sigma_q) = D\left(\int_{\mathcal{X}} \Lambda^{1/2} \psi(x) \psi(x)^\top \Lambda^{1/2} dp(x) \ \Big\| \ \int_{\mathcal{X}} \Lambda^{1/2} \psi(x) \psi(x)^\top \Lambda^{1/2} dq(x)\right)$$

is convex in Λ (see a direct proof in Appendix A.3). As mentioned earlier, by monotonicity, we only need to consider maximal elements of \mathcal{L} .

Relative entropy maximization. Given p and q , from which we can compute $C_p = \mathbb{E}_{X \sim p}[\psi(X)\psi(X)^\top]$ and $C_q = \mathbb{E}_{X \sim q}[\psi(X)\psi(X)^\top]$, we maximize $D(\Lambda^{1/2} C_p \Lambda^{1/2} \|\Lambda^{1/2} C_q \Lambda^{1/2})$ with respect to $\Lambda \in \mathcal{L}$.

We can decompose it as the difference of two convex functions as:

$$D(\Lambda^{1/2} C_p \Lambda^{1/2} \|\Lambda^{1/2} C_q \Lambda^{1/2}) = \text{tr} [C_p^{1/2} \Lambda C_p^{1/2} \log(C_p^{1/2} \Lambda C_p^{1/2})] - \text{tr} [\Lambda^{1/2} C_p \Lambda^{1/2} \log(\Lambda^{1/2} C_q \Lambda^{1/2})].$$

This naturally leads to an optimization algorithm where we lower-bound the first convex function by its tangent and then maximize this affine function minus the other convex functions. This is a parameter-free globally and monotonically convergent algorithm.

Log-partition minimization. Given q and $f : \mathcal{X} \rightarrow \mathbb{R}$, we can try to minimize the lower bound on the log-partition function instead. We assume for simplicity that we know a representation of f as $f(x) = \psi(x)^\top G \psi(x)$, that the span \mathcal{B} of all $\psi(x)\psi(x)^\top$ for $x \in \mathcal{X}$ is manageable, and that (for simplicity), $\psi(x)^\top \Lambda \psi(x) = 1$ for all $\Lambda \in \mathcal{L}$.

We thus need to minimize with respect to $\Lambda \in \mathcal{L}$ the following function (which is convex in Λ as the supremum of convex functions):

$$\begin{aligned} & \sup_{C \in \mathcal{B}, C \succ 0, \text{tr } C = 1} \text{tr}(CG) - D(\Lambda^{1/2} C \Lambda^{1/2} \| \Lambda^{1/2} C_q \Lambda^{1/2}) \\ &= \inf_M \sup_{C \in \mathcal{B}} \text{tr } C(G - \Lambda^{1/2} M \Lambda^{1/2}) + \log \text{tr} \exp(M + \log(\Lambda^{1/2} C_q \Lambda^{1/2})) \\ &= \inf_M \log \text{tr} \exp(M + \log(\Lambda^{1/2} C_q \Lambda^{1/2})) \text{ such that } G - \Lambda^{1/2} M \Lambda^{1/2} \in \mathcal{B}^\perp. \end{aligned}$$

The optimal C and M can be obtained by accelerated projected gradient descent. We can then minimize with respect to Λ by updating Λ like for the entropy maximization, once M (and thus C) have been estimated.

We now provide two illustrating examples, on the torus and the hypercube.

7.6 Example 1: Torus $[0, 1]$

We consider $\mathcal{X} = [0, 1]$, and assume that we have $m = 2r + 1$ frequencies $\omega \in \{-r, \dots, +r\}$, and $\psi_\omega(x) = e^{2i\pi\omega x}$. Then we have $(C_p)_{\omega\omega'} = \hat{p}(\omega - \omega')$ and $(C_q)_{\omega\omega'} = \hat{q}(\omega - \omega')$, and the constraint to be in \mathcal{B} is that C_p and C_q are Toeplitz matrices (note that this requires to know $\hat{p}(\omega)$ and $\hat{q}(\omega)$ for $\omega \in \{-2r, \dots, 2r\}$).

If f has band-limited Fourier series \hat{f} (zero outside $\{-2r, \dots, 2r\}$), then we have $f(x) = \langle \psi(x), G \psi(x) \rangle$, with any matrix G such that $\sum_{\omega - \omega' = \delta} G_{\omega\omega'} = \hat{f}(\delta)$ for all $\delta \in \{-2r, \dots, 2r\}$. We then have $\text{tr}[C_p G] = \sum_{\omega, \omega' = -r}^r \hat{p}(\omega - \omega')^* G_{\omega\omega'} = \sum_{\delta = -2r}^{2r} \hat{p}(\delta)^* \hat{f}(\delta)$.

Moreover, we consider $\Lambda = \text{Diag}(\eta)$ with η in the simplex, so that, for $\hat{q}(\omega) = 1_{\omega \neq 0}$ (for the uniform distribution), we have

$$D(\Lambda^{1/2} C_p \Lambda^{1/2} \| \Lambda^{1/2} C_q \Lambda^{1/2}) = \text{tr} [C_p^{1/2} \Lambda C_p^{1/2} \log(C_p^{1/2} \Lambda C_p^{1/2})] - \sum_{\omega = -r}^r \eta(\omega) \log \eta(\omega).$$

Relative entropy maximization. We can maximize with respect to η , using the (here convergent by convexity) the majorization-minimization approach, e.g., by using the affine Taylor expansion of the first term, and minimizing $\text{tr} [C_p^{1/2} \text{Diag}(\eta_+) C_p^{1/2} \log(C_p^{1/2} \text{Diag}(\eta_+) C_p^{1/2})] - \sum_{\omega = -r}^r \eta_+(\omega) \log \eta_+(\omega)$ with respect to η^+ is in the simplex, thus leading to $\log \eta^+$ equal to a constant plus $\text{diag} [C_p^{1/2} \log(C_p^{1/2} \text{Diag}(\eta) C_p^{1/2}) C_p^{1/2}]$. This makes an algorithm in closed form to maximize $D(\Sigma_p \| \Sigma_q)$ with respect to η .

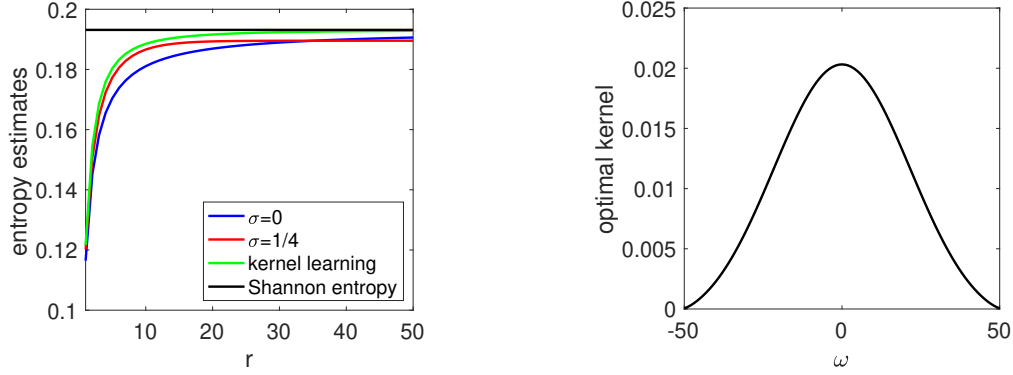


Figure 3: Kernel learning for entropy estimation on the torus $[0, 1]$. Left: Entropy estimates for increasing order r . Right: optimal kernel for $r = 50$.

This is illustrated in Figure 3, where we compare various lower bounds on the negative entropies, either with a fixed kernel, or with kernel learning, showing the benefits of estimating $\hat{k}(\omega)$.

Log-partition minimization. In order to compute (and then minimize) the upper-bound on the log-partition function, we can write

$$\begin{aligned}
 b(f) &= \sup_{\hat{p}} \sum_{\delta=-2r}^{2r} \hat{p}(\delta)^* \hat{f}(\delta) - \text{tr} [C_p^{1/2} \text{Diag}(\eta) C_p^{1/2} \log(C_p^{1/2} \text{Diag}(\eta) C_p^{1/2})] + \sum_{\omega=-r}^r \eta(\omega) \log \eta(\omega) \\
 &= \inf_M \log \text{tr} \exp(M) + \sum_{\omega=-r}^r \eta(\omega) \log \eta(\omega) \text{ s. t. } \sum_{\omega-\omega'=\delta} \eta(\omega)^{1/2} \eta(\omega')^{1/2} M_{\omega\omega'} = \hat{f}(\delta), \forall \delta \in \{-2r, \dots, 2r\}.
 \end{aligned}$$

Note here that by the classical Caratheodory interpolation theorem (see, e.g., [24, Chapter 2]), this is exactly $a(f)$ which is computed.

It can be minimized by (accelerated) projected gradient descent in M , since the function $M \mapsto \log \text{tr} \exp(M)$ is 1-smooth. We simply need to be able to project on the affine subspace, which can be done by solving a linear system.¹³

In order to update the kernel η , we use the exact same update as for maximizing the entropy. This is illustrated in Figure 4, where we show benefit of estimating the kernel rather than using a uniform one.

7.7 Example 2: Hypercube

We consider $\mathcal{X} = \{-1, 1\}^d$, and $\varphi(x) = \text{Diag}(\eta)^{1/2} \begin{pmatrix} x \\ 1 \end{pmatrix} \in \mathbb{R}^{d+1}$, for η in the simplex in \mathbb{R}^{d+1} . Here \mathcal{B} is exactly the set of all self-adjoint operators with diagonal equal to η . To compute the entropy, we consider the set \mathcal{C} of positive matrices with unit diagonal, corresponding to $\mathbb{E}[\begin{pmatrix} x \\ 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}^\top]$.

¹³By Lagrangian duality, the minimizer of $\frac{1}{2} \|M - Z\|_F^2$ such that $\forall \delta \in \{-r, \dots, r\}, \text{tr}(MA_\delta) = \hat{f}(\delta)$ obtained as $Z + \sum_{\delta} \lambda_{\delta} A_{\delta}$ with λ maximizing $\sum_{\delta=-r}^r \lambda_{\delta} (\hat{f}_{\delta} - \text{tr} Z A_{\delta}) - \frac{1}{2} \sum_{\delta, \delta'=-r}^r \lambda_{\delta} \lambda_{\delta'} \text{tr} A_{\delta} A_{\delta'}$.

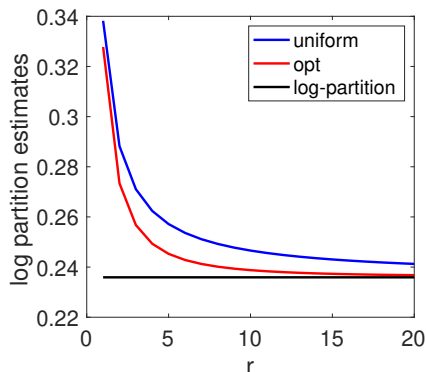


Figure 4: Kernel learning for log partition estimation on the torus $[0, 1]$ with increasing order r .

To maximize the relative entropy with respect to the uniform distribution, we simply maximize

$$\text{tr} [C(p)^{1/2} \text{Diag}(\eta)C(p)^{1/2} \log(C(p)^{1/2} \text{Diag}(\eta)C(p)^{1/2})] - \sum_{j=1}^{d+1} \eta_j \log \eta_j,$$

which is a lower bound on $\sum_{x \in \{-1, 1\}^d} p(x) \log p(x) + d \log 2$. Thus, we get an upperbound on the entropy

$$- \sum_{x \in \{-1, 1\}^d} p(x) \log p(x) \leq d \log 2 - \text{tr} [\text{Diag}(\eta)^{1/2} C(p) \text{Diag}(\eta)^{1/2} \log(\text{Diag}(\eta)^{1/2} C(p) \text{Diag}(\eta)^{1/2})] + \sum_{j=1}^{d+1} \eta_j \log \eta_j.$$

With η constant equal to $1/(d+1)$, we get the upper bound: $d \log 2 - \frac{1}{d+1} \text{tr} C(p) \log C(p)$, which happens to be tight for $d = 1$.

This is to be compared to the upper-bound of [31], which is $\frac{1}{2} \log \det \left(C(p) + \frac{1}{3} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \right) + \frac{d}{2} \log \frac{\pi e}{2}$. In Figure 5, we compare these bounds on matrices C corresponding to independent components with mean vectors uniformly distributed, showing the potential benefits of the new upperbound on entropy.

8 Conclusion

In this paper, we presented properties of the von Neumann entropy and relative entropy of covariance operators obtained from reproducing kernel Hilbert spaces. These notions are intimately related to the usual notions of Shannon entropy and relative entropy, share many of their properties, and can be applied to any type of data where positive definite kernels can be defined. They also come with additional computational benefits, with several estimation algorithms with theoretical guarantees. We have also highlighted several properties in terms of multivariate probabilistic modeling or variational inference, but several interesting avenues for future research are worth exploring. First, exploring further approximation guarantees for log-partition function estimation can probably lead to new improved algorithms for variational inference [58] in continuous and discrete domains. Second, we could extend to Rényi entropies [36, 25] with similar developments (along the lines of Appendix A.4). Finally, the optimization with respect to the probability measure, which is at the core of log-partition estimation, can be extended to other optimization problems

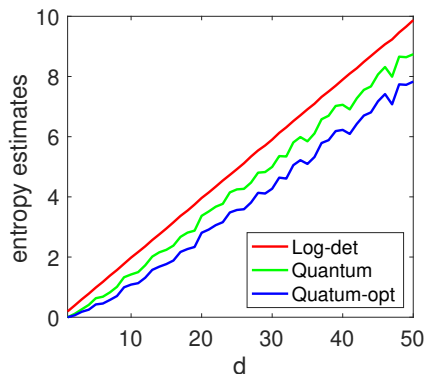


Figure 5: Comparison of entropy bounds with increasing dimensions (the log-determinant bound from [31], our bound based on quantum entropy with uniform η and the optimized bound). We plot the bounds for independent distributions with random probabilities for each component. We subtract the true entropy and compare upper-bounds (with zero meaning that the bounds are tight).

on probability measures, such as already done with similar tools for optimal transport [57] or optimal control [11].

Acknowledgements

The author thanks Alessandro Rudi and Loucas Pillaud-Vivien for fruitful discussions related to this work. This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support the European Research Council (grant SEQUOIA 724063). The comments and suggestions of the anonymous reviewers were greatly appreciated.

A Proof of convexity results on operators

In this section, we provide simple direct proofs of convexity results that we use in this paper. All of these proofs are obtained from the literature, and in particular [33, 47, 46].

A.1 Joint convexity of the relative entropy

We consider a function $f : \mathbb{R}_+^* \rightarrow \mathbb{R}$, which is operator-convex, that is, for which the function $A \mapsto f(A)$ is convex as an operator-valued function, i.e., $f\left(\sum_{i=1}^n \lambda_i A_i\right) \preceq \sum_{i=1}^n \lambda_i f(A_i)$. These functions can exactly be represented as: $f(t) = \alpha + \beta(t - 1) + \gamma(t - 1)^2 + \delta \frac{(t - 1)^2}{t} + \int_0^\infty \frac{(t - 1)^2}{\lambda + t} d\nu(\lambda)$, where $\gamma, \delta > 0$ and ν is a positive measure with finite mass on \mathbb{R}_+ [33]. We have $f(1) = \alpha$. The interesting case for this paper

is $f(t) = -\log(t) = 1 - t + \int_0^\infty \frac{(t-1)^2}{\lambda+t} \frac{1}{(1+\lambda)^2} d\lambda$, but there are others, such as, for $p \in (0, 1]$:

$$-t^p = -1 - p(t-1) + \frac{\sin p\pi}{\pi} \int_0^{+\infty} \frac{(t-1)^2}{\lambda+t} \frac{\lambda^p}{(1+\lambda)^2} d\lambda.$$

Given A and B two Hermitian operators, we denote by \mathcal{R}_A the right multiplication by A and \mathcal{L}_A the left multiplication by A . We have, since \mathcal{L}_B and \mathcal{R}_A commute, and $\log \mathcal{L}_B = \mathcal{L}_{\log B}$ and $\log \mathcal{R}_A = \mathcal{R}_{\log A}$:

$$\begin{aligned} D(A\|B) &= \operatorname{tr} A^{1/2}(\log A)A^{1/2} - \operatorname{tr} A^{1/2}(\log B)A^{1/2} = \operatorname{tr} [(\log \mathcal{R}_A)(A^{1/2})A^{1/2}] - \operatorname{tr} [A^{1/2}(\log \mathcal{L}_B)(A^{1/2})] \\ &= \langle A^{1/2}, [\log \mathcal{R}_A - \log \mathcal{L}_B]A^{1/2} \rangle = \langle A^{1/2}, -\log [\mathcal{L}_B \mathcal{R}_A^{-1}]A^{1/2} \rangle, \end{aligned}$$

where $\langle M, N \rangle = \operatorname{tr}(M^*N)$ denotes the usual dot-products between operators. Following [47], we thus have:

$$\begin{aligned} D(A\|B) &= \langle A^{1/2}, -\log [\mathcal{L}_B \mathcal{R}_A^{-1}]A^{1/2} \rangle \\ &\quad + \int_0^\infty \langle (\mathcal{L}_B \mathcal{R}_A^{-1} - I)A^{1/2}, (\mathcal{L}_B \mathcal{R}_A^{-1} + \lambda I)^{-1}(\mathcal{L}_B \mathcal{R}_A^{-1} - I)A^{1/2} \rangle \frac{1}{(1+\lambda)^2} d\lambda \\ &= \operatorname{tr} A - \operatorname{tr} B + \int_0^\infty \langle \mathcal{R}_A^{-1/2}(B-A), (\mathcal{L}_B \mathcal{R}_A^{-1} + tI)^{-1} \mathcal{R}_A^{-1/2}(B-A) \rangle \frac{1}{(1+\lambda)^2} d\lambda \\ &= \operatorname{tr} A - \operatorname{tr} B + \int_0^\infty \langle (B-A), (\mathcal{L}_B + \lambda \mathcal{R}_A)^{-1}(B-A) \rangle \frac{1}{(1+\lambda)^2} d\lambda. \end{aligned}$$

Note we can ensure the finiteness of $D(A\|B)$ as soon as $B^{-1/2}AB^{-1/2}$ is bounded. Given that the mapping $(u, M) \mapsto \langle u, M^{-1}u \rangle$ is convex, we obtain the joint convexity of the relative entropy. It can also be shown more directly as shown below.

Proposition 12 *Given positive invertible self-adjoint operators A_i, B_i , then*

$$D\left(\sum_{i=1}^n \lambda_i A_i \middle\| \sum_{i=1}^n \lambda_i B_i\right) \leq \sum_{i=1}^n \lambda_i D(A_i\|B_i),$$

with equality if and only if for all j such that $\lambda_j > 0$,

$$\log B_j - \log A_j = \log \left(\sum_{i=1}^n \lambda_i B_i \right) - \log \left(\sum_{i=1}^n \lambda_i A_i \right).$$

Proof Following [46, Theorem 7], we have: $D(A_i\|B_i) - \operatorname{tr} A_i(\log A - \log B) = \operatorname{tr} A_i(\log A_i - \log A + \log B - \log B_i) = \operatorname{tr} A_i - \operatorname{tr} \exp(\log A - \log B + \log B_i) + B(A_i\|\exp(\log A - \log B + \log B_i))$, where $B(\cdot\|\cdot)$ is the Bregman divergence (always non-negative), defined as $B(A\|B) = D(A\|B) - \operatorname{tr} A + \operatorname{tr} B$. Thus

$$\begin{aligned} &\sum_{i=1}^n \lambda_i D(A_i\|B_i) - D(A\|B) \\ &= \sum_{i=1}^n \lambda_i \left[D(A_i\|B_i) - \operatorname{tr} A_i(\log A - \log B) \right] \\ &= \operatorname{tr} A - \sum_{i=1}^n \lambda_i \operatorname{tr} \exp(\log A - \log B + \log B_i) + \sum_{i=1}^n \lambda_i B(A_i\|\exp(\log A - \log B + \log B_i)), \end{aligned}$$

which is greater than zero since the function $C \mapsto \text{tr} \exp(K + \log C)$ is concave. There is equality if and only for all i such that $\lambda_i > 0$, $A_i = \exp(\log A - \log B + \log B_i)$, because of the Bregman divergences being equal to zero. \blacksquare

More generally, we get for random operators A and B , $\mathbb{E}[D(A\|B)] \geq D(\mathbb{E}A\|\mathbb{E}B)$, with equality if and only if almost surely, $\log A - \log B = \log \mathbb{E}A - \log \mathbb{E}B$. Note that the concavity of the function $B \mapsto \text{tr} \exp(M + \log B)$, which is crucial in deriving concentration inequalities for matrices and operators [56], is simply obtained by partial Fenchel dualization of the relative entropy.

We can also provide more refined relationships at equality, following [39].

Proposition 13 *Given positive invertible self-adjoint operators A_i, B_i , and strictly positive real numbers λ_i that sum to 1, then*

$$D\left(\sum_{i=1}^n \lambda_i A_i \parallel \sum_{i=1}^n \lambda_i B_i\right) = \sum_{i=1}^n \lambda_i D(A_i \parallel B_i),$$

implies that for all i , $A_i B_i^{-1} = AB^{-1}$, where $A = \sum_{i=1}^n \lambda_i A_i$ and $B = \sum_{i=1}^n \lambda_i B_i$.

Proof Given the representation above, and the fact that $(u, M) \mapsto \langle u, M^{-1}u \rangle$ is convex, we get the joint convexity of $(A, B) \mapsto D(A\|B)$ is convex. In order to study the equality sign, we can use the lemma below, and for each $\lambda > 0$, $[\mathcal{L}_{B_i} + \lambda \mathcal{R}_{A_i}]^{-1}(B_i - A_i) = [\mathcal{L}_B + \lambda \mathcal{R}_A]^{-1}(B - A)$. By letting λ go to zero, we get the result.

Note that following [39], we can also consider that

$$(\mathcal{L}_{B_i} + \lambda \mathcal{R}_{A_i})^{-1}(B_i - A_i) = (\mathcal{L}_{B_i} \mathcal{R}_{A_i}^{-1} + \lambda I)^{-1}[\mathcal{L}_{B_i} \mathcal{R}_{A_i}^{-1} - I]I \text{ is independent of } i,$$

and is equal to $(\mathcal{L}_B \mathcal{R}_A^{-1} + \lambda I)^{-1}[\mathcal{L}_B \mathcal{R}_A^{-1} - I]I$. Thus by using representations of a function $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ as $f(t) = (t-1) \int_0^\infty \frac{d\nu(\lambda)}{t+\lambda}$ for a certain signed measure $d\nu$, we can deduce that $f[\mathcal{L}_{B_i} \mathcal{R}_{A_i}^{-1}]I$ is independent of i , and equal to $f[\mathcal{L}_B \mathcal{R}_A^{-1}]I$. For example. with $f(t) = \log t$, we get $\log B_i - \log A_i$ independent of i . For example. with $f(t) = t - 1$, we get $B_i A_i^{-1}$ independent of i . For $f(t) = t^z$, $B_i^z A_i^{-z}$ independent of i , for any complex z . \blacksquare

Lemma 1 *If M_1, \dots, M_n are invertible and all λ_i are strictly positive and sum to 1, then*

$$\sum_{i=1}^n \lambda_i u_i^\top M_i^{-1} u_i = \left(\sum_{i=1}^n \lambda_i u_i\right) \left(\sum_{i=1}^n \lambda_i M_i\right)^{-1} \left(\sum_{i=1}^n \lambda_i u_i\right)$$

if and only if $M_i^{-1} u_i$ is independent of i , and equal to $M^{-1} u$ where $M = \sum_{i=1}^n \lambda_i M_i$ and $u = \sum_{i=1}^n \lambda_i u_i$.

Proof We consider the function $f(u, M) = u^\top M^{-1} u$, with Taylor expansion

$$f(u+\delta, M+\Delta) = f(u, M) + 2\delta^\top M^{-1} u - u^\top M^{-1} \Delta M^{-1} u + 2(\delta - \Delta M^{-1} u)^\top M^{-1} (\delta - \Delta M^{-1} u) + O(\|\Delta\|^3 + \|\delta\|^3).$$

We now use the following result on the remainder in Jensen's inequality (which itself is a direct consequence of the Taylor expansion with integral remainder taken at X , while expanded at $\mathbb{E}X$):

$$\mathbb{E}f(X) - f(\mathbb{E}X) = \int_0^1 \mathbb{E} \left[(X - \mathbb{E}X)^\top f''(tX + (1-t)\mathbb{E}X) (X - \mathbb{E}X) \right] dt. \quad (15)$$

In our particular situation, by the invertibility assumption, the matrix M is invertible almost surely, and having equality in Jensen's inequality imposes to have $(X - \mathbb{E}X)^\top f''(tX + (1-t)\mathbb{E}X)(X - \mathbb{E}X) = 0$ almost surely in t and X . Thus, at $t = 0$ and $X = (u_i, M_i)$, we get: $u_i - u - (M_i - M)M^{-1}u = 0$, which is exactly $M_i^{-1}u_i = M^{-1}u$. \blacksquare

A.2 Monotonicity of the relative entropy

Proposition 14 *Given operators $C_i, i = 1, \dots, n$, such that $\sum_{i=1}^n C_i^* C_i = I$, then*

$$D\left(\sum_{i=1}^n C_i A C_i^* \parallel \sum_{i=1}^n C_i B C_i^*\right) \leq D(A \parallel B).$$

Proof We have, with $\omega = \exp(2i\pi/n)$, using the joint convexity of the relative entropy and the expression $\frac{1}{n} \sum_{k=1}^n \omega^{k(i-j)} = 1_{i=j}$:

$$\begin{aligned} D\left(\sum_{i=1}^n C_i A C_i^* \parallel \sum_{i=1}^n C_i B C_i^*\right) &\leq \sum_{i=1}^n D(C_i A C_i^* \parallel C_i B C_i^*) = D\left(\sum_{i=1}^n (e_i e_i^\top) \otimes E_i A E_i^* \parallel \sum_{i=1}^n (e_i e_i^\top) \otimes C_i B C_i^*\right) \\ &= D\left(\frac{1}{n} \sum_{k=1}^n \omega^{k(i-j)} \sum_{i,j=1}^n (e_i e_j^\top) \otimes C_i A C_j^* \parallel \frac{1}{n} \sum_{k=1}^n \omega^{k(i-j)} \sum_{i,j=1}^n (e_i e_j^\top) \otimes C_i B C_j^*\right). \end{aligned}$$

Consider the matrices M and N defined by blocks $M_{ij} = C_i A C_j^*$ and $N_{ij} = C_i B C_j^*$, as well as, the matrix U_k defined by blocks $(U_k)_{ij} = 1_{i=j} \omega^{ki} I$, we get from the equation above:

$$\begin{aligned} D\left(\sum_{i=1}^n C_i A C_i^* \parallel \sum_{i=1}^n C_i B C_i^*\right) &\leq D\left(\frac{1}{n} \sum_{k=1}^n U_k M U_k^* \parallel \frac{1}{n} \sum_{k=1}^n U_k N U_k^*\right) \\ &\leq \frac{1}{n} \sum_{k=1}^n D\left(U_k M U_k^* \parallel U_k N U_k^*\right) = D(M \parallel N) \end{aligned}$$

because each U_k is unitary. Moreover, we can write $M = V A V^*$ and $N = V B V^*$, with a an operator V defined by a column of blocks $C_i, i = 1, \dots, n$, so that $V^* V = \sum_{i=1}^n C_i^* C_i = I$. Thus $D(M \parallel N) = D(A \parallel B)$, which shows the desired bound.

Using propositions above, there is equality if and only: $\log(C_i B C_i^*) - \log(C_i A C_i^*)$ independent of i and $\log[U_k N U_k^*] - \log[U_k M U_k^*] = U_k(\log N - \log M) U_k^*$ independent of k , that is $\log N - \log M$ block diagonal. Overall, we get equality if $\log N - \log M$ is block diagonal with equal blocks.

We can also obtain necessary conditions for equality, that is, $(C_i B C_i^*)(C_i A C_i^*)^{-1}$ independent of i , and $U_k N U_k^* [U_k N U_k^*]^{-1} = U_k N M^{-1} U_k^*$ independent of k , that is, $N M^{-1}$ block-diagonal with equal blocks. \blacksquare

This extends to a random operator C , such that $\mathbb{E}[C^* C] = I$, for which we get:

$$D\left(\mathbb{E}[C A C^*] \parallel \mathbb{E}[C B C^*]\right) \leq D(A \parallel B).$$

A particular example is obtained from n Hermitian positive operators D_1, \dots, D_n such that $\sum_{i=1}^n D_i = I$, and we get, with $\mu_i = \text{tr}(D_i A)$ and $\nu_i = \text{tr}(D_i B)$:

$$\sum_{i=1}^n \mu_i \log \frac{\mu_i}{\nu_i} \leq D(A\|B).$$

This corresponds to, if $D_i = \sum_k \lambda_{ik} u_{ik} \otimes u_{ik}$, to $C_{ik} = \lambda_{ik}^{1/2} e_i u_{ik}^*$, so that

$$\sum_{i,k} E_i A E_i^* = \sum_{i,k} \lambda_{ik} e_i u_{ik}^* M u_{ik} e_i^* = \sum_{i=1}^n (\text{tr } A D_i) e_i e_i^*.$$

A.3 Concavity of relative entropy

In this section, we consider two positive self-adjoint operators, M and N such that $M \preceq \alpha N$ for some α . We consider the function:

$$f : \Lambda \rightarrow D(\Lambda^{1/2} M \Lambda^{1/2} \| \Lambda^{1/2} N \Lambda^{1/2}).$$

It happens to be convex in (M, N) . It is also concave in Λ in the set of positive bounded operators. This is due to the representation, valid for invertible operators:

$$\begin{aligned} & f(\Lambda) - \text{tr } M \Lambda + \text{tr } N \Lambda \\ &= \int_0^\infty \langle \Lambda^{1/2} (N - M) \Lambda^{1/2}, (\mathcal{L}_{\Lambda^{1/2} N \Lambda^{1/2}} + \lambda \mathcal{R}_{\Lambda^{1/2} M \Lambda^{1/2}})^{-1} \Lambda^{1/2} (N - M) \Lambda^{1/2} \rangle \frac{1}{(1 + \lambda)^2} d\lambda \\ &= \int_0^\infty \langle (N - M), (\mathcal{L}_N \mathcal{R}_\Lambda^{-1} + \lambda \mathcal{L}_\Lambda^{-1} \mathcal{R}_M)^{-1} (N - M) \rangle \frac{1}{(1 + \lambda)^2} d\lambda. \end{aligned}$$

Since the mapping $(A, B) \mapsto (A^{-1} + B^{-1})^{-1} = B - B(A + B)^{-1}B$ is matrix-concave [1, Corollary 1.1], we get the desired concavity.

A.4 Extension to other divergences

As mentioned in [47], both the joint convexity and the monotonicity of the relative entropy $D(A\|B)$ in Prop. 12 and Prop. 14 can be extended to most operator-convex functions on \mathbb{R}_+^* , and

$$D_f(A\|B) = \langle A^{1/2}, f[\mathcal{L}_B \mathcal{R}_A^{-1}] A^{1/2} \rangle,$$

with the relative entropy being $D_{-\log}$. Among other interesting cases, we have:

- $f(t) = (t - 1)^2$, with $D_f(A\|B) = \text{tr} [(B - A)^2 A^{-1}]$.
- $f(t) = \frac{(t - 1)^2}{t}$, with $D_f(A\|B) = \text{tr} [(B - A)^2 B^{-1}]$.
- $f(t) = \frac{(t - 1)^2}{1 + t}$, with $D_f(A\|B) = \langle A^{1/2}, (\mathcal{L}_B + \mathcal{R}_A)^{-1} A^{1/2} \rangle$.

- $f(t) = 1 - \sqrt{t}$, with $D_f(A\|B) = \text{tr}(A) - \langle A^{1/2}, [\mathcal{L}_B \mathcal{R}_A^{-1}]^{1/2} A^{1/2} \rangle$.

Specialized to covariance operators, we have, with $A = \Sigma_p$ and $B = \Sigma_q$, and following the same reasoning as in the proof of Prop. 4:

$$\begin{aligned} D_f(\Sigma_p\|\Sigma_q) &\leq \int_{\mathcal{X}} D_f\left(\varphi(x)\varphi(x)^* \left\| \frac{dq}{dp}(x)\varphi(x)\varphi(x)^*\right.\right) dp(x) \\ &= \int_{\mathcal{X}} f\left(\frac{dq}{dp}(x)\right) dp(x) = D_f(p\|q), \end{aligned}$$

which is the *reverse* f -divergence associated with the function f [19, 55]. We thus immediately get lower bounds on:

- the square Hellinger distance $D_f(p\|q) = \int_{\mathcal{X}} \left[1 - \left(\frac{dq}{dp}(x)\right)^{1/2}\right] dp(x) = \frac{1}{2} \int_{\mathcal{X}} \left[\left(\frac{dq}{d\tau}(x)\right)^{1/2} - \left(\frac{dp}{d\tau}(x)\right)^{1/2}\right]^2 d\tau(x)$ for $f(t) = 1 - \sqrt{t}$,
- the reverse Pearson χ^2 -divergence $D_f(p\|q) = \int_{\mathcal{X}} \left[1 - \frac{dq}{dp}(x)\right]^2 dp(x) = \frac{1}{2} \int_{\mathcal{X}} \left[\frac{dq}{d\tau}(x) - \frac{dp}{d\tau}(x)\right]^2 \frac{d\tau(x)}{dp/d\tau(x)}$ for $f(t) = (t - 1)^2$,
- the Pearson χ^2 -divergence $D_f(p\|q) = \int_{\mathcal{X}} \left[1 - \frac{dq}{dp}(x)\right]^2 \frac{dp(x)}{dq/dp(x)} = \frac{1}{2} \int_{\mathcal{X}} \left[\frac{dq}{d\tau}(x) - \frac{dp}{d\tau}(x)\right]^2 \frac{d\tau(x)}{dq/d\tau(x)}$ for $f(t) = \frac{(t - 1)^2}{t}$,
- the Vincze-Le Cam distance $D_f(p\|q) = \int_{\mathcal{X}} \left[1 - \frac{dq}{dp}(x)\right]^2 \frac{dp(x)}{1 + dq/dp(x)}$ equal to $\frac{1}{2} \int_{\mathcal{X}} \left[\frac{dq}{d\tau}(x) - \frac{dp}{d\tau}(x)\right]^2 \frac{d\tau(x)}{dq/d\tau(x) + dp/d\tau(x)}$ for $f(t) = \frac{(t - 1)^2}{1 + t}$.

B Proof of lower bound (Prop. 5)

Proof Starting from $\tilde{p}(y) = \int_{\mathcal{X}} h(x, y) dp(x)$ and $\tilde{q}(y) = \int_{\mathcal{X}} h(x, y) dq(x)$, we can provide an exact expression in the usual data processing inequality for the regular KL divergence [18], with two joint distributions $\tilde{p}(x, y) = p(x)h(x, y)$ and $\tilde{q}(x, y) = q(x)h(x, y)$, so that, with $f(u) = u \log u$:

$$\begin{aligned} D(p\|q) &= D(\tilde{p}\|\tilde{q}) = \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{q}(x, y) f\left(\frac{\tilde{p}(x, y)}{\tilde{q}(x, y)}\right) d\tau(x) d\tau(y) \\ &\geq \int_{\mathcal{X}} \tilde{q}(y) f\left(\int_{\mathcal{X}} \tilde{q}(x|y) \frac{\tilde{p}(x, y)}{\tilde{q}(x, y)} d\tau(x)\right) d\tau(y) \text{ by Jensen's inequality,} \\ &= \int_{\mathcal{X}} \tilde{q}(y) f\left(\frac{\tilde{p}(y)}{\tilde{q}(y)}\right) d\tau(y) = D(\tilde{p}\|\tilde{q}). \end{aligned}$$

Using Eq. (15), we thus get

$$D(p\|q) - D(\tilde{p}\|\tilde{q}) = \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{q}(y) h(x, y) \left(\frac{p(x)}{q(x)} - \frac{\tilde{p}(y)}{\tilde{q}(y)}\right)^2 \left(\int_0^1 \frac{dt}{t \frac{p(x)}{q(x)} + (1-t) \frac{\tilde{p}(y)}{\tilde{q}(y)}}\right) d\tau(x) d\tau(y).$$

Since for all $x \in \mathcal{X}$, $t \frac{p(x)}{q(x)} + (1-t) \frac{\tilde{p}(y)}{\tilde{q}(y)} \geq \inf_{x \in \mathcal{X}} \frac{p(x)}{q(x)}$, we can then bound as

$$\begin{aligned}
D(p||q) - D(\tilde{p}||\tilde{q}) &\leq \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} \int_{\mathcal{X}} \int_{\mathcal{Y}} \tilde{q}(y) h(x, y) \left(\frac{p(x)}{q(x)} - \frac{\tilde{p}(y)}{\tilde{q}(y)} \right)^2 d\tau(x) d\tau(y) \\
&= \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{q}(y) h(x, y) \left(\frac{p(x)}{q(x)} - \frac{p(y)}{q(y)} + \frac{p(y)}{q(y)} - \frac{\tilde{p}(y)}{\tilde{q}(y)} \right)^2 d\tau(x) d\tau(y) \\
&\leq 2 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} \int_{\mathcal{X}} \int_{\mathcal{X}} \tilde{q}(y) h(x, y) \left(\frac{p(x)}{q(x)} - \frac{p(y)}{q(y)} \right)^2 d\tau(x) d\tau(y) \\
&\quad + 2 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} \int_{\mathcal{X}} \tilde{q}(y) \left(\frac{p(y)}{q(y)} - \frac{\tilde{p}(y)}{\tilde{q}(y)} \right)^2 d\tau(y).
\end{aligned}$$

We assume Lipschitz-continuity for p and q in the following form: for all x, y , $|\frac{q(x)}{q(y)} - 1| \leq d(x, y)C_q$, $|\frac{p(x)}{p(y)} - 1| \leq d(x, y)C_p$, and $|\frac{p(y)}{q(y)} - \frac{p(x)}{q(x)}| \leq d(x, y)C_{p/q}$ for some distance d on \mathcal{X} . Then the first term in the bound above is less than $2C_{p/q}^2 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} \sup_{x \in \mathcal{X}} \int_{\mathcal{X}} h(x, y) d(x, y)^2 d\tau(y)$. For the second term, we write:

$$\begin{aligned}
\left| \frac{p(y)}{q(y)} - \frac{\tilde{p}(y)}{\tilde{q}(y)} \right| &\leq \frac{p(y)|\tilde{q}(y) - q(y)| + q(y)|\tilde{p}(y) - p(y)|}{q(y)\tilde{q}(y)} \\
&\leq \frac{p(y) \int_{\mathcal{X}} |q(y) - q(x)| h(x, y) d\tau(x) + q(y) \int_{\mathcal{X}} |p(y) - p(x)| h(x, y) d\tau(x)}{q(y)\tilde{q}(y)} \\
&\leq \frac{C_q p(y) q(y) \int_{\mathcal{X}} d(x, y) h(x, y) d\tau(x) + C_p q(y) p(y) \int_{\mathcal{X}} d(x, y) h(x, y) d\tau(x)}{q(y)\tilde{q}(y)} \\
&\leq \frac{(C_p + C_q) p(y)}{\tilde{q}(y)} \int_{\mathcal{X}} d(x, y) h(x, y) d\tau(x) \leq \frac{(C_p + C_q) p(y)}{\tilde{q}(y)} \left(\int_{\mathcal{X}} d(x, y)^2 h(x, y) d\tau(x) \right)^{1/2}.
\end{aligned}$$

Moreover

$$\begin{aligned}
\int_{\mathcal{X}} \frac{p(y)^2}{\tilde{q}(y)} d\tau(y) &\leq \sup_{x \in \mathcal{X}} \frac{p(x)}{q(x)} \int_{\mathcal{X}} \frac{2\tilde{p}(y)^2 + 2(\tilde{p}(y) - p(y))^2}{\tilde{p}(y)} d\tau(y) \\
&\leq 2 \sup_{x \in \mathcal{X}} \frac{p(x)}{q(x)} \left(1 + \int_{\mathcal{X}} \frac{(\tilde{p}(y) - p(y))^2}{\tilde{p}(y)} d\tau(y) \right) \leq 2 \sup_{x \in \mathcal{X}} \frac{p(x)}{q(x)} (1 + C_p^2 \text{diam}(\mathcal{X})^2).
\end{aligned}$$

Overall, we get $D(p||q) - D(\tilde{p}||\tilde{q}) \leq E(p, q) \times \sup_{x \in \mathcal{X}} \int h(x, y) d(x, y)^2 dy$, with $E(p, q) = 2C_{p/q}^2 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} + 4 \sup_{x \in \mathcal{X}} \frac{q(x)}{p(x)} (C_p + C_q)^2 \sup_{x \in \mathcal{X}} \frac{p(x)}{q(x)} (1 + C_p^2 \text{diam}(\mathcal{X})^2)$. ■

C Convergence of estimation algorithms

In this Appendix, we give convergence proofs for the estimation algorithms from Section 5, starting with concentration inequalities for operators and bounds on degrees of freedom.

C.1 Concentration of sums of random self-adjoint operators

We recall in this Appendix classical results on the concentration of covariance operators.

Lemma 2 (Concentration of covariance operators [45, 56]) *Let φ be mapping $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ between a measurable space \mathcal{X} and a separable Hilbert space \mathcal{H} , such that for all $x \in \mathcal{X}$, $\|\varphi(x)\| \leq R$. Given $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$, assume that $\lambda_{\max}(\Sigma_p) \leq \mu$. Given $x_1, \dots, x_n \in \mathcal{X}$ n i.i.d. samples from p , then, for all $u > 0$,*

$$\mathbb{P}\left(\lambda_{\max}\left(\Sigma_p - \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)\right) > u\right) \leq \frac{\text{tr} \Sigma_p}{\mu} \left(1 + \frac{3\mu^2}{u^4 n^2} (R^2 + \frac{u}{3})^2\right) \exp\left(-\frac{nu^2}{2\mu(R^2 + \frac{u}{3})}\right).$$

We will apply the lemma above to $\tilde{\varphi}(x) = (\Sigma_p + \lambda I)^{-1/2} \varphi(x)$, with $\Sigma_p = \int_{\mathcal{X}} \varphi(x) \otimes \varphi(x) dp(x)$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$, so that, $R^2 = \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle = \text{df}^{\max}(\lambda)$, and $\text{tr} \Sigma_p$ is now $\text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] = \text{df}(\lambda)$, and $\mu = 1$, leading to:

$$\mathbb{P}\left(\lambda_{\max}\left((\Sigma_p + \lambda I)^{-1/2} (\Sigma_p - \hat{\Sigma}_p) (\Sigma_p + \lambda I)^{-1/2}\right) > u\right) \leq \text{df}(\lambda) \left(1 + \frac{3(\text{df}^{\max}(\lambda) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\lambda) + 2u/3}\right).$$

C.2 Degrees of freedom estimation

Proposition 15 (Estimation of degrees of freedom) *Let $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ a mapping between a measurable space \mathcal{X} and a separable Hilbert space \mathcal{H} . Let $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$, with p a probability distribution and $x_1, \dots, x_n \in \mathcal{X}$ n i.i.d. samples from p , and $\hat{\Sigma}_p = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$. Then,*

$$\mathbb{E} \left| \text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] - \text{tr} [\hat{\Sigma}_p (\hat{\Sigma}_p + \lambda I)^{-1}] \right| \leq \frac{d}{\sqrt{n}} + 4 \frac{d^2}{n} + 16dn \exp\left(\frac{-9n}{32d + 8}\right),$$

where $\text{df}(\lambda) = \text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}]$, and $\text{df}^{\max}(\lambda) = \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle$, which is assumed to satisfy $d = \text{df}^{\max}(\lambda) \leq n$.

Proof We follow closely the proof of [45], which starts with:

$$\begin{aligned} A &= \text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] - \text{tr} [\hat{\Sigma}_p (\hat{\Sigma}_p + \lambda I)^{-1}] \\ &= \lambda \text{tr} [(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}] + \lambda \text{tr} [(\Sigma_p + \lambda I)^{-1/2} B (I - B)^{-1} B (\Sigma_p + \lambda I)^{-1/2}], \end{aligned}$$

leading to $|A| \leq \lambda \left| \text{tr} [(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}] \right| + \|B\|_{\text{HS}}^2 \frac{1}{1 - \lambda_{\max}(B)}$, as soon as $\lambda_{\max}(B) < 1$, for $B = (\Sigma_p + \lambda I)^{-1/2} (\Sigma_p - \hat{\Sigma}_p) (\Sigma_p + \lambda I)^{-1/2}$.

Almost surely, $|A| \leq \max\{\text{df}(\Lambda), n\}$. Thus, for $u < 1$,

$$\begin{aligned} |A| &\leq 1_{\lambda_{\max}(B) > u} \max\{\text{df}(\Lambda), n\} + 1_{\lambda_{\max}(B) \leq u} \left(\lambda \left| \text{tr} [(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}] \right| + \|B\|_{\text{HS}}^2 \frac{1}{1 - \lambda_{\max}(B)} \right) \\ &\leq 1_{\lambda_{\max}(B) > u} \max\{\text{df}(\Lambda), n\} + \|B\|_{\text{HS}}^2 \frac{1}{1 - u} + \lambda \left| \text{tr} [(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}] \right|. \end{aligned}$$

Thus,

$$\mathbb{E}|A| \leq \mathbb{P}(\lambda_{\max}(B) > u) \cdot \max\{\text{df}(\Lambda), n\} + \frac{1}{1-u} \mathbb{E}[\|B\|_{\text{HS}}^2] + \mathbb{E}[\lambda |\text{tr}[(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}]|].$$

We have:

$$\mathbb{P}(\lambda_{\max}(B) > u) \leq \text{df}(\lambda) \left(1 + \frac{3(\text{df}^{\max}(\lambda) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\lambda) + 2u/3}\right),$$

and

$$\begin{aligned} \mathbb{E}[\|B\|_{\text{HS}}^2] &= \frac{1}{n} \mathbb{E} \left[\text{tr} \left((\Sigma_p + \lambda I)^{-1} (\Sigma_p - \varphi(x_i) \otimes \varphi(x_i)) (\Sigma_p + \lambda I)^{-1} (\Sigma_p - \varphi(x_i) \otimes \varphi(x_i)) \right) \right] \\ &\leq \frac{1}{n} \mathbb{E} \left[\text{tr} \left((\Sigma_p + \lambda I)^{-1} \varphi(x_i) \otimes \varphi(x_i) (\Sigma_p + \lambda I)^{-1} \varphi(x_i) \otimes \varphi(x_i) \right) \right] \leq \frac{1}{n} \text{df}(\lambda) \text{df}^{\max}(\lambda). \end{aligned}$$

Moreover

$$\begin{aligned} \mathbb{E}[\lambda |\text{tr}[(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}]|] &\leq \lambda \sqrt{\mathbb{E}[\|\text{tr}[(\Sigma_p + \lambda I)^{-1/2} B (\Sigma_p + \lambda I)^{-1/2}]\|^2]} \\ &\leq \lambda \sqrt{\frac{1}{n} \mathbb{E}[\langle \varphi(x), (\Sigma_p + \lambda I)^{-1} (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle^2]} \\ &\leq \sqrt{\frac{1}{n} \mathbb{E}[\langle \varphi(x), (\Sigma_p + \lambda I)^{-1} \varphi(x) \rangle^2]} \leq \sqrt{\frac{1}{n} \text{df}^{\max}(\lambda) \text{df}(\lambda)}. \end{aligned}$$

This leads to

$$\begin{aligned} \mathbb{E}|A| &\leq \sqrt{\frac{1}{n} \text{df}^{\max}(\lambda) \text{df}(\lambda)} + \frac{1}{1-u} \frac{1}{n} \text{df}(\lambda) \text{df}^{\max}(\lambda) \\ &\quad + \max\{\text{df}(\lambda), n\} \text{df}(\lambda) \left(1 + \frac{3(\text{df}^{\max}(\lambda) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\lambda) + 2u/3}\right). \end{aligned}$$

With $u = 3/4$, we get:

$$\begin{aligned} \mathbb{E}|A| &\leq \sqrt{\frac{1}{n} \text{df}^{\max}(\lambda) \text{df}(\lambda)} + 4 \frac{1}{n} \text{df}(\lambda) \text{df}^{\max}(\lambda) \\ &\quad + \max\{\text{df}(\lambda), n\} \text{df}(\lambda) \left(1 + \frac{3(\text{df}^{\max}(\lambda) + 1/4)^2}{(3/4)^4 n^2}\right) \exp\left(\frac{-n(3/4)^2}{2\text{df}^{\max}(\lambda) + 1/2}\right). \end{aligned}$$

Using $\text{df}(\lambda) \leq \text{df}^{\max}(\lambda) = d$, we have the bound:

$$\mathbb{E}|A| \leq \frac{d}{\sqrt{n}} + 4 \frac{d^2}{n} + d \max\{d, n\} \left(1 + \frac{16(4d+1)^2}{27n^2}\right) \exp\left(\frac{-9n}{8(4d+1)}\right).$$

If $d \leq n$, we further get:

$$\mathbb{E}|A| \leq \frac{d}{\sqrt{n}} + 4 \frac{d^2}{n} + dn \left(1 + \frac{16(4d+1)^2}{27n^2}\right) \exp\left(\frac{-9n}{32d+8}\right) \leq \frac{d}{\sqrt{n}} + 4 \frac{d^2}{n} + 16dn \exp\left(\frac{-9n}{32d+8}\right).$$

■

C.3 Degrees of freedom estimation (from projections)

Proposition 16 (Estimation of degrees of freedom (from projections)) *Let $\varphi : \mathcal{X} \rightarrow \mathcal{H}$ a mapping between a measurable space \mathcal{X} and a separable Hilbert space \mathcal{H} . Let $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$, with p a probability distribution and $x_1, \dots, x_n \in \mathcal{X}$ n i.i.d. samples from q such that $\forall x \in \mathcal{X}, q(x) \geq \alpha p(x)$, and $\hat{\Pi}$ the orthogonal projection on the span of all $\varphi(x_1), \dots, \varphi(x_n)$. Let $\text{df}^{\max}(\lambda) = \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma_q + \lambda I)^{-1} \varphi(x) \rangle$. Then*

$$\begin{aligned} \mathbb{E} \left| \text{tr}[\Sigma_p(\Sigma_p + \lambda I)^{-1}] - \text{tr}[\hat{\Pi}\Sigma_p\hat{\Pi}(\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda I)^{-1}] \right| \\ \leq 16\text{df}(\alpha\lambda)\text{df}(\mu) \exp\left(\frac{-9n}{32\text{df}^{\max}(\mu) + 8}\right) + 8\mu \text{tr}[(\Sigma_q + \alpha\lambda I)^{-1}\Sigma_q(\Sigma_q + \mu I)^{-1}], \end{aligned}$$

if $\text{df}^{\max}(\mu) = \sup_{x \in \mathcal{X}} \langle \varphi(x), (\Sigma_q + \mu I)^{-1} \varphi(x) \rangle \leq n$.

Proof We follow a more direct strategy as the previous proof, following estimation results for random feature expansions and column sampling [43, 5].

$$\begin{aligned} A &= \text{tr}[\Sigma_p(\Sigma_p + \lambda I)^{-1}] - \text{tr}[\hat{\Pi}\Sigma_p\hat{\Pi}(\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda I)^{-1}] \\ &= \lambda \text{tr}[(\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda I)^{-1} - (\Sigma_p + \lambda I)^{-1}] \\ &= \lambda \text{tr}[(\Sigma_p + \lambda I)^{-1}(\Sigma_p - \hat{\Pi}\Sigma_p\hat{\Pi})(\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda I)^{-1}] \text{ using standard expansions of the matrix inverse,} \\ &= \lambda \text{tr}[(\Sigma_p + \lambda I)^{-1}(\Sigma_p - \hat{\Pi}\Sigma_p\hat{\Pi})(\Sigma_p + \lambda I)^{-1}] \\ &\quad + \lambda \text{tr}[(\Sigma_p + \lambda I)^{-1}(\Sigma_p - \hat{\Pi}\Sigma_p\hat{\Pi})(\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda I)^{-1}(\Sigma_p - \hat{\Pi}\Sigma_p\hat{\Pi})(\Sigma_p + \lambda I)^{-1}]. \end{aligned}$$

Using $\hat{\Pi}^2 = \hat{\Pi}$, we can further bound

$$\begin{aligned} A &\leq \lambda \text{tr}[(\Sigma_p + \lambda I)^{-1}(\Sigma_p - \hat{\Pi}\Sigma_p\hat{\Pi})(\Sigma_p + \lambda I)^{-1}] \\ &\quad + \lambda \text{tr}[(\Sigma_p + \lambda I)^{-1}[(I - \hat{\Pi})\Sigma_p(I - \hat{\Pi}) + \lambda^{-1}\Sigma_p(I - \hat{\Pi})\Sigma_p](\Sigma_p + \lambda I)^{-1}] \text{ using } \hat{\Pi}^2 = \hat{\Pi}, \\ &= \text{tr}[(\Sigma_p + \lambda I)^{-2}(\lambda\Sigma_p - \lambda\hat{\Pi}\Sigma_p\hat{\Pi} + \lambda\Sigma_p - \lambda\Sigma_p\hat{\Pi} - \lambda\hat{\Pi}\Sigma_p + \lambda\hat{\Pi}\Sigma_p\hat{\Pi} + \Sigma_p(I - \hat{\Pi})\Sigma_p)] \\ &= 2\lambda \text{tr}[(\Sigma_p + \lambda I)^{-2}\Sigma_p(I - \hat{\Pi})] + \text{tr}[(\Sigma_p + \lambda I)^{-2}\Sigma_p^2(I - \hat{\Pi})] \\ &\leq 2 \text{tr}[(\Sigma_p + \lambda I)^{-1}\Sigma_p(I - \hat{\Pi})]. \end{aligned}$$

We can thus use, with $\hat{\Sigma}_q = \frac{1}{n} \sum_{i=1}^n \varphi(x_i)\varphi(x_i)^*$,

$$I - \hat{\Pi} = \lim_{\mu \rightarrow 0} I - \hat{\Sigma}_q(\hat{\Sigma}_q + \mu I)^{-1} = \lim_{\mu \rightarrow 0} \mu(\hat{\Sigma}_q + \mu I)^{-1} \preceq \mu(\hat{\Sigma}_q + \mu I)^{-1} \preceq \frac{1}{1-u} \mu(\Sigma_q + \mu I)^{-1},$$

if we have $(\Sigma_q + \mu I)^{-1/2}(\Sigma_q - \hat{\Sigma}_q)(\Sigma_q + \mu I)^{-1/2} \preceq uI$. This leads to, using Lemma 2,

$$A \leq \frac{2\mu}{1-u} \text{tr}[(\Sigma_p + \lambda I)^{-1}\Sigma_p(\Sigma_q + \mu I)^{-1}],$$

with probability greater than $1 - \text{df}(\mu) \left(1 + \frac{3(\text{df}^{\max}(\mu) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\mu) + 2u/3}\right)$. Therefore, we get, with $u = 3/4$:

$$\begin{aligned}
0 \leq \mathbb{E}A &\leq \text{tr}(\Sigma_q + \alpha\lambda I)^{-1} \Sigma_q \cdot \text{df}(\mu) \left(1 + \frac{3(\text{df}^{\max}(\mu) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\mu) + 2u/3}\right) \\
&\quad + \frac{2\mu}{1-u} \text{tr} [(\Sigma_q + \alpha\lambda I)^{-1} \Sigma_q (\Sigma_q + \mu I)^{-1}] \\
&\leq \text{df}(\alpha\lambda) \text{df}(\mu) \left(1 + \frac{3(\text{df}^{\max}(\mu) + u/3)^2}{u^4 n^2}\right) \exp\left(\frac{-nu^2}{2\text{df}^{\max}(\mu) + 2u/3}\right) \\
&\quad + \frac{2\mu}{1-u} \text{tr} [(\Sigma_q + \alpha\lambda I)^{-1} \Sigma_q (\Sigma_q + \mu I)^{-1}] \\
&\leq \text{df}(\alpha\lambda) \text{df}(\mu) \left(1 + \frac{3(\text{df}^{\max}(\mu) + 1/4)^2}{(3/4)^4 n^2}\right) \exp\left(\frac{-n(3/4)^2}{2\text{df}^{\max}(\mu) + 1/2}\right) \\
&\quad + 8\mu \text{tr} [(\Sigma_q + \alpha\lambda I)^{-1} \Sigma_q (\Sigma_q + \mu I)^{-1}] \text{ with } u = 3/4.
\end{aligned}$$

Using the same reasoning as in the proof of Prop. 15, the probabilistic term is upper-bounded by $16\text{df}(\alpha\lambda)\text{df}(\mu) \exp\left(\frac{-9n}{32\text{df}^{\max}(\mu) + 8}\right)$ if $\text{df}^{\max}(\mu) \leq n$. ■

C.4 Proof of Proposition 7

Proof We assume that $p(x) \geq q(x)/\alpha$ for all $x \in \mathcal{X}$. Given the integral representation of the entropy, we have

$$\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] - \text{tr} [\Sigma_p \log \Sigma_p] = \int_0^{+\infty} (\text{tr} [\Sigma_p (\Sigma_p + \alpha\lambda I)^{-1}] - \text{tr} [\hat{\Sigma}_p (\hat{\Sigma}_p + \alpha\lambda I)^{-1}]) \alpha d\lambda,$$

it turns out it is possible to truncate large values of λ . Indeed,

$$\begin{aligned}
&\int_{\lambda_1}^{+\infty} (\text{tr} [\Sigma_p (\Sigma_p + \alpha\lambda I)^{-1}] - \text{tr} [\hat{\Sigma}_p (\hat{\Sigma}_p + \alpha\lambda I)^{-1}]) \alpha d\lambda \\
&= \text{tr} \hat{\Sigma}_p \log(\hat{\Sigma}_p + \alpha\lambda_1 I) - \text{tr} \Sigma_p \log(\Sigma_p + \alpha\lambda_1 I) \\
&\leq \text{tr} \hat{\Sigma}_p \log(I + \alpha\lambda_1 I) - \text{tr} \Sigma_p \log(\alpha\lambda_1 I) = \log(1 + \alpha\lambda_1) - \log(\alpha\lambda_1) \leq \frac{1}{\alpha\lambda_1},
\end{aligned}$$

with a similar bound for the opposite.

Thus, we have the bound, with $\lambda_1 \geq \lambda_0$:

$$\mathbb{E}A = |\text{tr} [\hat{\Sigma}_p \log \hat{\Sigma}_p] - \text{tr} [\Sigma_p \log \Sigma_p]| \leq \frac{1}{\alpha\lambda_1} + \int_0^{\lambda_0} \mathbb{E} |\text{df}(\alpha\lambda) - \hat{\text{df}}(\alpha\lambda)| \alpha d\lambda + \int_{\lambda_0}^{\lambda_1} \mathbb{E} |\text{df}(\alpha\lambda) - \hat{\text{df}}(\alpha\lambda)| \alpha d\lambda,$$

where $\text{df}(\alpha\lambda) = \text{tr} [\Sigma_p(\Sigma_p + \alpha\lambda I)^{-1}]$ and $\widehat{\text{df}}(\alpha\lambda) = \text{tr} [\widehat{\Sigma}_p(\widehat{\Sigma}_p + \alpha\lambda I)^{-1}]$ are the usual degrees of freedom for Σ_p . We have

$$\text{df}(\alpha\lambda) \leq \text{df}^{\max}(\alpha\lambda) \leq \sup_{x \in \mathcal{X}} \langle \varphi(x), (\alpha\Sigma + \alpha\lambda I)^{-1} \varphi(x) \rangle = \frac{1}{\alpha} C(\lambda),$$

where $C(\lambda) = \text{tr}[\Sigma(\Sigma + \lambda I)^{-1}]$. Thus, we get:

$$A \leq \frac{1}{\alpha\lambda_1} + n\alpha\lambda_0 + \int_0^{\lambda_0} C(\lambda)d\lambda + \int_{\lambda_0}^{\lambda_1} \mathbb{E}|\text{df}(\alpha\lambda) - \widehat{\text{df}}(\alpha\lambda)|\alpha d\lambda.$$

For $\lambda \geq \lambda_0$, we have $\text{df}(\lambda) \leq \text{df}^{\max}(\lambda) = d \leq \frac{1}{\alpha} C(\lambda_0) = d_0$, which we assume to be less than n .

Thus

$$16dn \exp\left(\frac{-9n}{32d+8}\right) \leq 16dn \exp\left(\frac{-9n}{32d_0+8}\right),$$

and we get the bound, using $C(\lambda) \leq \frac{1}{\lambda}$:

$$\begin{aligned} A &\leq \frac{1}{\alpha\lambda_1} + n\alpha\lambda_0 + \int_0^{\lambda_0} C(\lambda)d\lambda + \int_{\lambda_0}^{\lambda_1} \left(\frac{d}{\sqrt{n}} + 4\frac{d^2}{n} + 16dn \exp\left(\frac{-9n}{32d_0+8}\right)\right)\alpha d\lambda \\ &\leq \frac{1}{\alpha\lambda_1} + n\alpha\lambda_0 + \int_0^{\lambda_0} C(\lambda)d\lambda + \left(\frac{1}{\sqrt{n}} + 16n \exp\left(\frac{-9n}{32d_0+8}\right)\right) \int_{\lambda_0}^{\lambda_1} C(\lambda)d\lambda + \frac{4}{n\alpha} \int_{\lambda_0}^{\lambda_1} C(\lambda)^2 d\lambda. \end{aligned}$$

We take λ_0 such that $d_0 = \frac{n}{\log n}$, with a bound:

$$\begin{aligned} A &\leq \frac{1}{\alpha\lambda_1} + n\alpha\lambda_0 + \int_0^{\lambda_0} C(\lambda)d\lambda + \frac{17}{\sqrt{n}} \int_{\lambda_0}^{\lambda_1} C(\lambda)d\lambda + \frac{4}{n\alpha} \int_{\lambda_0}^{\lambda_1} C(\lambda)^2 d\lambda \\ &\leq \frac{1}{\alpha\lambda_1} + n\alpha\lambda_0 + \int_0^{\lambda_0} C(\lambda)d\lambda + \frac{17}{\sqrt{n}} \int_{\lambda_0}^{\lambda_1} C(\lambda)d\lambda + \frac{4c}{n\alpha}. \end{aligned}$$

The condition on C , implies that $C(\lambda)^2 \frac{\lambda}{2} \leq \int_{\lambda/2}^{\lambda} C(\lambda')^2 d\lambda' \leq c$, leading to $C(\lambda) \leq \sqrt{\frac{2c}{\lambda}}$, leading to $\frac{n}{\log n} \alpha \leq \sqrt{\frac{2c}{\lambda_0}}$, and thus $\lambda_0 \leq 2c \frac{(\log n)^2}{n^2 \alpha^2}$. Then

$$A \leq \frac{1}{\alpha\lambda_1} + 2c \frac{(\log n)^2}{n\alpha} + 2c \frac{\log n}{n\alpha} + \frac{17}{\sqrt{n}} \int_{\lambda_0}^{\lambda_1} C(\lambda)d\lambda + \frac{4c}{n\alpha}.$$

We take $\lambda_1 = \lambda_0 + n$ so that $\int_{\lambda_0}^{\lambda_1} C(\lambda)d\lambda \leq \int_0^{\lambda_0+n} C(\lambda)d\lambda + \log \frac{\lambda_0+n}{\lambda_0+1} \leq 2\sqrt{c}\sqrt{1+\lambda_0} + \log n$. Then, overall, we get:

$$A \leq \frac{1+8c(\log n)^2}{n\alpha} + \frac{17}{\sqrt{n}} \left(2\sqrt{c} + 2\sqrt{2c} \frac{\log n}{n\alpha} + \log n\right) \leq \frac{1+c(8\log n)^2}{n\alpha} + \frac{17}{\sqrt{n}} (2\sqrt{c} + \log n),$$

which is the desired bound. ■

C.5 Proof of Proposition 8

Proof We can combine with the proof of Proposition 7, to get:

$$A = \text{tr} [\Pi \Sigma_p \Pi \log \Pi \Sigma_p \Pi] - \text{tr} [\Sigma_p \log \Sigma_p] = \int_0^{+\infty} (\text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] - \text{tr} [\Pi \Sigma_p \Pi (\Pi \Sigma_p \Pi + \lambda I)^{-1}]) d\lambda,$$

which is always non-negative and such that, for $\mu = \alpha \varepsilon \lambda$, if $\text{df}^{\max}(\alpha \varepsilon \lambda_0) \leq n$:

$$\begin{aligned} A &\leq \int_0^{\lambda_0} \text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] d\lambda + \int_{\lambda_0}^{\lambda_1} (\text{tr} [\Sigma_p (\Sigma_p + \lambda I)^{-1}] - \text{tr} [\Pi \Sigma_p \Pi (\Pi \Sigma_p \Pi + \lambda I)^{-1}]) d\lambda + \frac{1}{\lambda_1} \\ &\leq \int_0^{\lambda_0} \text{df}(\alpha \lambda) d\lambda \\ &+ \int_{\lambda_0}^{\lambda_1} \left[16 \text{df}(\alpha \lambda) \text{df}(\alpha \varepsilon \lambda) \exp\left(\frac{-9n}{32 \text{df}^{\max}(\alpha \varepsilon \lambda_0) + 8}\right) + 8 \alpha \varepsilon \lambda \text{tr} [(\Sigma_q + \alpha \lambda I)^{-1} \Sigma_q (\Sigma_q + \alpha \varepsilon \lambda I)^{-1}] \right] d\lambda + \frac{1}{\lambda_1} \\ &\leq \frac{1}{\alpha} \int_0^{\alpha \lambda_0} \text{df}(\lambda) d\lambda + \frac{16}{\alpha} \frac{n}{\beta_n} \exp\left(\frac{-9}{32 \beta_n^{-1} + 8/n}\right) \int_{\alpha \lambda_0}^{\alpha \lambda_1} \text{df}(\lambda) d\lambda + \frac{1}{\lambda_1} \\ &\quad + \int_{\alpha \lambda_0}^{\alpha \lambda_1} 8 \varepsilon \lambda \text{tr} [(\Sigma_q + \lambda I)^{-1} \Sigma_q (\Sigma_q + \varepsilon \lambda I)^{-1}] d\lambda \\ &\leq \frac{1}{\alpha} \int_0^{\alpha \lambda_0} \text{df}(\lambda) d\lambda + \frac{16}{\alpha} \frac{n}{\beta_n} \exp\left(\frac{-9}{32 \beta_n^{-1} + 8/n}\right) \int_0^{\alpha \lambda_1} \text{df}(\lambda) d\lambda + \frac{1}{\lambda_1} \\ &\quad + \int_0^{\alpha \lambda_1} 8 \varepsilon \lambda \text{tr} [(\Sigma_q + \lambda I)^{-1} \Sigma_q (\Sigma_q + \varepsilon \lambda I)^{-1}] d\lambda, \text{ extending integration limits,} \end{aligned}$$

with $\text{df}^{\max}(\alpha \varepsilon \lambda_0) = \frac{n}{\beta_n}$.

We also have

$$\int_0^{\lambda'} \text{df}(\lambda) d\lambda = \sum_{\sigma \in \Lambda(\Sigma_q)} \int_0^{\lambda'} \frac{\sigma}{\sigma + \lambda} d\lambda = \sum_{\sigma \in \Lambda(\Sigma_q)} \sigma \log(1 + \lambda' \sigma^{-1}),$$

and

$$\begin{aligned} \int_0^{\lambda'} \varepsilon \lambda \text{tr} [(\Sigma_q + \lambda I)^{-1} \Sigma_q (\Sigma_q + \varepsilon \lambda I)^{-1}] d\lambda &= \sum_{\sigma \in \Lambda(\Sigma_q)} \int_0^{\lambda'} \frac{\sigma}{\sigma + \lambda} \frac{\varepsilon \lambda}{\sigma + \varepsilon \lambda} d\lambda \\ &= \sum_{\sigma \in \Lambda(\Sigma_q)} \frac{\varepsilon \sigma}{1 - \varepsilon} \int_0^{\lambda'} \left[\frac{1}{\sigma + \varepsilon \lambda} - \frac{1}{\sigma + \lambda} \right] d\lambda \\ &= \sum_{\sigma \in \Lambda(\Sigma_q)} \frac{\varepsilon \sigma}{1 - \varepsilon} \left[\frac{1}{\varepsilon} \log \frac{\sigma + \varepsilon \lambda'}{\sigma} - \log \frac{\sigma + \lambda'}{\sigma} \right] \\ &\leq \sum_{\sigma \in \Lambda(\Sigma_q)} \frac{\sigma}{1 - \varepsilon} \log(1 + \lambda' \varepsilon \sigma^{-1}) = \frac{1}{1 - \varepsilon} \int_0^{\varepsilon \lambda'} \text{df}(\lambda) d\lambda. \end{aligned}$$

Thus, with $\varepsilon = \frac{1}{\alpha \lambda_0} \text{df}^{-1}(n/\beta_n)$,

$$A \leq \frac{1}{\alpha} \int_0^{\alpha \lambda_0} \text{df}(\lambda) d\lambda + \frac{16}{\alpha} \frac{n}{\beta_n} \exp\left(\frac{-9}{32 \beta_n^{-1} + 8/n}\right) \int_0^{\alpha \lambda_1} \text{df}(\lambda) d\lambda + \frac{8}{1 - \varepsilon} \int_0^{\varepsilon \alpha \lambda_1} \text{df}(\lambda) d\lambda + \frac{1}{\lambda_1}.$$

Polynomial decays. If $\text{df}(\lambda) \leq B\lambda^{-1/s}$, then we have: $\frac{n}{\beta_n} \leq B(\alpha\lambda_o\varepsilon)^{-1/s}$, and thus $\varepsilon \leq \frac{1}{\alpha\lambda_o} \left(\frac{B\beta_n}{n}\right)^s$.

Moreover, $\int_0^{\alpha\lambda_o} \text{df}(\lambda)d\lambda \leq \frac{B}{1-1/s}(\alpha\lambda_o)^{1-1/s}$. Thus, with $\varepsilon < 1/2$,

$$A \leq \frac{B\alpha^{-1/s}}{1-1/s}(\lambda_o)^{1-1/s} + \frac{16}{\alpha} \frac{n}{\beta_n} \exp\left(\frac{-9}{32\beta_n^{-1} + 8/n}\right) \frac{B}{1-1/s}(\alpha\lambda_1)^{1-1/s} + \frac{8}{1-\varepsilon} \frac{B}{1-1/s}(\varepsilon\alpha\lambda_1)^{1-1/s} + \frac{1}{\lambda_1}.$$

Now turning to asymptotic expansions, we get

$$A = O\left(B\lambda_o^{1-1/s} + \frac{n}{\beta_n} \exp(-\beta_n/4)B\lambda_1^{1-1/s} + B(\lambda_1/\lambda_o)^{1-1/s} \left(\frac{B\beta_n}{n}\right)^{s-1} + \frac{1}{\lambda_1}\right).$$

We can take $\beta_n \propto \log n$ so that the term $\exp(-\beta_n/4)$ is negligible.

We can choose λ_o such that $B\lambda_o^{1-1/s} \sim B(\lambda_1/\lambda_o)^{1-1/s} \left(\frac{B\log n}{n}\right)^{s-1}$, that is, $\lambda_o \propto B^{s/2}\lambda_1^{1/2} \left(\frac{\log n}{n}\right)^{s/2}$, and then $\lambda_1^{(3s-1)/(2s)} B^{(s+1)/2} \left(\frac{\log n}{n}\right)^{(s-1)/2}$, finally leading to $A = O\left(B^{s(s+1)/(3s-1)} \left(\frac{\log n}{n}\right)^{s(s-1)/(3s-1)}\right)$.

Exponential decays. If $\text{df}(\lambda) \leq B\log(C/\lambda)^d$, then we have: $\frac{n}{\beta_n} \leq B\log(C/(\varepsilon\alpha\lambda_o))^d$, and thus $\varepsilon \leq \frac{C}{\lambda_o\alpha} \exp(-(Bn/\beta_n)^{1/d})$. Moreover $\int_0^{\alpha\lambda_o} \text{df}(\lambda)d\lambda \sim \alpha\lambda_o B\log(C/\lambda_o\alpha)^d$. Thus, asymptotically,

$$A = O\left(\lambda_o B\log(C/\lambda_o)^d + \frac{n}{\beta_n} \exp(-\beta_n/4)\lambda_1 B(\log(C/\lambda_1))^d + \lambda_1 \varepsilon B\log(C/\varepsilon\lambda_1)^d + \frac{1}{\lambda_1}\right).$$

With $\lambda_o \propto C^{1/2}\lambda_1^{1/2} \exp(-\frac{1}{2}(Bn/\beta_n)^{1/d})$, the first and third terms lead to

$$BC^{1/2}\lambda_1^{1/2} \exp(-\frac{1}{2}(Bn/\beta_n)^{1/d}) \log(C/\lambda_o)^d = BC^{1/2}\lambda_1^{1/2} \exp(-\frac{1}{2}(Bn/\beta_n)^{1/d}) \left(\log C + \log \frac{1}{\lambda_1} + (Bn/\beta_n)^{1/d}\right)^d.$$

With $\lambda_1 \propto \frac{1}{B} \exp(\beta_n/8)$, the second term is of order $\frac{n}{\beta_n} \exp(-\beta_n/8) \left(\log C + \log \frac{1}{\lambda_1}\right)^d$. We can then choose $\beta_n \propto (Bn)^{1/(d+1)}$, to get the overall bound $A = (\log(BC))^d (\sqrt{BC} + B) \exp\left(-\frac{1}{4}(Bn)^{1/(d+1)}\right)$. \blacksquare

C.6 Computation of degrees of freedom

We have, when σ tends to zero, with $\nu = \frac{\lambda}{\hat{k}(0)} = \frac{\lambda}{\tanh^d \frac{\sigma}{2}}$,

$$\begin{aligned}
\sum_{\omega \in \mathbb{Z}^d} \frac{1}{1 + \nu \exp(\sigma \|\omega\|_1)} &= \sum_{k \geq 0} \frac{1}{1 + \nu \exp(\sigma k)} |\{\omega \in \mathbb{Z}^d, \|\omega\|_1 = k\}| \leq \sum_{k \geq 0} \frac{1}{1 + \nu \exp(\sigma k)} 2^d k^{d-1} \\
&\leq \sum_{k \geq \frac{1}{\sigma} \log \frac{1}{\nu}} \frac{1}{\nu \exp(\sigma k)} 2^d k^{d-1} + \sum_{k \leq \frac{1}{\sigma} \log \frac{1}{\nu}} 2^d k^{d-1} \\
&\leq 2^d \frac{(d-1)!}{(1-e^{-\sigma})^{-d}} \left[1 + \frac{1}{2} \left(\log \frac{1}{\nu} \right)^{d-1} \right] + 2^d \left(\frac{1}{\sigma} \log \frac{1}{\nu} \right)^d \\
&\leq 2^d \frac{d!}{(1-e^{-\sigma})^{-d}} \left[1 + \left(\log \frac{1}{\nu} \right)^d \right] \\
&\leq e^{-\sigma d/2} \frac{d!}{\sinh^d(\sigma/2)} \left[1 + \left(\log \frac{1}{\lambda} + d \log \tanh \frac{\sigma}{2} \right)^d \right],
\end{aligned}$$

where we have used that, for $\nu \geq e$,

$$(1 - e^{-\sigma}) \sum_{k \geq \frac{1}{\sigma} \log \frac{1}{\nu}} \exp(-\sigma k) k^\alpha \leq \left(\frac{1}{\sigma} \log \frac{1}{\nu} \right)^\alpha \nu + \alpha \sum_{k \geq \frac{1}{\sigma} \log \frac{1}{\nu}} \exp(-\sigma k) k^{\alpha-1}$$

and $\sum_{k \geq \frac{1}{\sigma} \log \frac{1}{\nu}} \exp(-\sigma k) \leq \frac{\nu}{1 - e^{-\sigma}}$, which leads to, with $u_\alpha = \sum_{k \geq \frac{1}{\sigma} \log \frac{1}{\nu}} \exp(-\sigma k) k^\alpha$:

$$\frac{1}{\alpha!} (1 - e^{-\sigma})^{\alpha+1} u_\alpha \leq \frac{1}{\alpha!} (1 - e^{-\sigma})^{\alpha-1} \nu \left(\log \frac{1}{\nu} \right)^\alpha + \frac{1}{(\alpha-1)!} (1 - e^{-\sigma})^\alpha u_{\alpha-1},$$

and thus $u_\alpha \leq \frac{\alpha! \nu}{(1 - e^{-\sigma})^{-\alpha-1}} \left[1 + \frac{1}{2} \left(\log \frac{1}{\nu} \right)^\alpha \right]$.

In order to bound the constant c in Prop. 7, we can split the integral into. We have for $\beta = \tanh^d \frac{\sigma}{2} > 0$,

$$\begin{aligned}
c &\leq \int_\beta^\infty \frac{1}{\lambda^2} d\lambda + \int_0^\beta e^{-\sigma d} \frac{d!^2}{\sinh^{2d}(\sigma/2)} \left[2 + 2 \left(\log \frac{\tanh^d \frac{\sigma}{2}}{\lambda} \right)^{2d} \right] d\lambda \\
&\leq \frac{1}{\beta} + \beta \int_0^1 e^{-\sigma d} \frac{d!^2}{\sinh^{2d}(\sigma/2)} \left[2 + 2 \left(\log \frac{1}{\lambda} \right)^{2d} \right] d\lambda = O(\sigma^{-d})
\end{aligned}$$

when σ goes to zero.

References

- [1] Tsuyoshi Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and its Applications*, 26:203–241, 1979.
- [2] Huzihiro Araki and Elliott H. Lieb. Entropy inequalities. In *Inequalities*, pages 47–57. Springer, 2002.

- [3] Francis Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.
- [4] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.
- [5] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(1):714–751, 2017.
- [6] Francis Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3(Jul):1–48, 2002.
- [7] Francis Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the International Conference on Machine Learning*, pages 33–40, 2005.
- [8] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [9] Marcel Berger and Bernard Gostiaux. *Differential Geometry: Manifolds, Curves, and Surfaces: Manifolds, Curves, and Surfaces*, volume 115. Springer Science & Business Media, 2012.
- [10] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.
- [11] Eloïse Berthier, Justin Carpentier, Alessandro Rudi, and Francis Bach. Infinite-dimensional sums-of-squares for optimal control. Technical Report 2110.07396, arXiv, 2021.
- [12] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2009.
- [13] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.
- [14] Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Symposium on Discrete algorithms*, pages 968–977, 2009.
- [15] Lev 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.
- [16] Haïm Brezis. *Analyse Fonctionnelle*. Masson, Paris, France, 1980.
- [17] Jean-François Cardoso. Dependence, correlation and Gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003.
- [18] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1999.
- [19] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.
- [20] Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

- [21] Ernesto De Vito, Lorenzo Rosasco, and Alessandro Rudi. Regularization: From inverse problems to large-scale machine learning. In *Harmonic and Applied Analysis*, pages 245–296. Springer, 2021.
- [22] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.
- [23] Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.
- [24] Ciprian Foias and Arthur E. Frazho. *The Commutant Lifting Approach to Interpolation Problems*. Springer, 1990.
- [25] Rupert L. Frank and Elliott H. Lieb. Monotonicity of a relative Rényi entropy. *Journal of Mathematical Physics*, 54(12):122201, 2013.
- [26] Kenji Fukumizu, Francis Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *Annals of Statistics*, 37(4):1871–1905, 2009.
- [27] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, pages 489–496, 2007.
- [28] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48(6):3228–3250, 2020.
- [29] Lars Hörmander. *The Analysis of Linear Partial Differential Operators I: Distribution Theory and Fourier Analysis*. Springer, 1984.
- [30] Chris J. Isham. *Lectures on Quantum Theory: Mathematical and Structural Foundations*. Allied Publishers, 2001.
- [31] Michael I. Jordan and Martin J. Wainwright. Semidefinite relaxations for approximate inference on graphs with cycles. *Advances in Neural Information Processing Systems*, 16, 2003.
- [32] Tosio Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, 1966.
- [33] Andrew Lesniewski and Mary Beth Ruskai. Monotone Riemannian metrics and relative entropy on noncommutative probability spaces. *Journal of Mathematical Physics*, 40(11):5702–5724, 1999.
- [34] Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.
- [35] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trend in Machine Learning*, 10(1-2):1–141, 2017.
- [36] Martin Müller-Lennert, Frédéric Dupuis, Oleg Szehr, Serge Fehr, and Marco Tomamichel. On quantum Rényi entropies: A new generalization and some properties. *Journal of Mathematical Physics*, 54(12):122203, 2013.

- [37] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.
- [38] Edouard Pauwels, Francis Bach, and Jean-Philippe Vert. Relating leverage scores and density using regularized Christoffel functions. In *Advances in Neural Information Processing Systems*, pages 1663–1672, 2018.
- [39] Dénes Petz. Sufficient subalgebras and the relative entropy of states of a von Neumann algebra. *Communications in Mathematical Physics*, 105(1):123–131, 1986.
- [40] Dénes Petz. Monotonicity of quantum relative entropy revisited. *Reviews in Mathematical Physics*, 15(01):79–91, 2003.
- [41] Michael Reed and Barry Simon. *Functional analysis*. Methods of Modern Mathematical Physics. Academic Press, 1980.
- [42] Christian P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.
- [43] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015.
- [44] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. Technical Report 2012.11978, arXiv, 2020.
- [45] Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.
- [46] Mary Beth Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *Journal of Mathematical Physics*, 43(9):4358–4375, 2002.
- [47] Mary Beth Ruskai. Another short and elementary proof of strong subadditivity of quantum entropy. *Reports on Mathematical Physics*, 60(1):1–12, 2007.
- [48] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels*. MIT Press, 2001.
- [49] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
- [50] Alex J. Smola, Zoltan L. Ovari, and Robert C. Williamson. Regularization with dot-product kernels. *Advances in Neural Information Processing Systems*, pages 308–314, 2001.
- [51] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.
- [52] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *Annual Conference on Learning Theory (COLT)*, pages 111–122, 2008.
- [53] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

- [54] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2(Nov):67–93, 2001.
- [55] Flemming Topsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.
- [56] Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.
- [57] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173, 2021.
- [58] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008.
- [59] Mark M. Wilde. *Quantum Information Theory*. Cambridge University Press, 2013.
- [60] Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. *Advances in Neural Information Processing Systems*, 13, 2000.
- [61] Blake E. Woodworth, Francis Bach, and Alessandro Rudi. Non-convex optimization with certificates and fast rates through kernel sums of squares. In *Annual Conference on Learning Theory (COLT)*, 2022.
- [62] Yao-Liang Yu. The strong convexity of von Neumann’s entropy. *Unpublished note*, 2013. <http://www.cs.cmu.edu/~yaoliang/mynotes/sc.pdf>.