HAL open science

Sum-of-Squares Relaxations for Information Theory and Variational Inference

Francis Bach

▶ To cite this version:

Francis Bach. Sum-of-Squares Relaxations for Information Theory and Variational Inference. 2023. hal-03703475v3

HAL Id: hal-03703475
https://hal.science/hal-03703475v3

Preprint submitted on 15 Sep 2023

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.

Sum-of-Squares Relaxations for Information Theory and Variational Inference

Francis Bach

September 15, 2023

Abstract

We consider extensions of the Shannon relative entropy, referred to as f-divergences. Three classical related computational problems are typically associated with these divergences: (a) estimation from moments, (b) computing normalizing integrals, and (c) variational inference in probabilistic models. These problems are related to one another through convex duality, and for all them, there are many applications throughout data science, and we aim for computationally tractable approximation algorithms that preserve properties of the original problem such as potential convexity or monotonicity. In order to achieve this, we derive a sequence of convex relaxations for computing these divergences from non-centered covariance matrices associated with a given feature vector: starting from the typically non-tractable optimal lower-bound, we consider an additional relaxation based on "sums-of-squares", which is is now computable in polynomial time as a semidefinite program. We also provide computationally more efficient relaxations based on spectral information divergences from quantum information theory. For all of the tasks above, beyond proposing new relaxations, we derive tractable convex optimization algorithms, and we present illustrations on multivariate trigonometric polynomials and functions on the Boolean hypercube.

1 Introduction

Tools from information theory are ubiquitous in data science. Starting with the notion of Shannon entropy, other notions have emerged, in particular f-divergences [19, 1], which are defined as

$$D(p\|q) = \int_{\mathcal{X}} f\Big(\frac{dp}{dq}(x)\Big) dq(x), \tag{1}$$

where p and q are two finite positive measures on an arbitrary set $\mathcal{X}$, $\frac{dp}{dq}$ is the density of p with respect to q, and $f : \mathbb{R}_+^* \to \mathbb{R}$ is a convex function.[1] A classical example is $f(t) = t \log t - t + 1$, where $D(p\|q)$ is the usual Kullback-Leibler divergence, associated with Shannon information theory [17], which we will use as a running example.

These divergences have been used in many areas in machine learning, signal processing or statistics, such as within message passing and variational inference [41], concentration inequalities [11], PAC-Bayes analysis [50], independent component analysis [14], information theory [51], differential privacy [42], design of surrogate losses for classification [45], and optimization [7]. We review f-divergences and their basic properties in Section 2, see [38, 37, 56] for a more complete treatment.

[1] Note that our notation $D(p\|q)$ ignores the dependence in f.

Two classical related computational problems are typically associated with f-divergences, which have to be estimated or optimized in some way, a task that can become difficult in multivariate settings. For all them, there are many applications throughout data science, and we aim for computationally tractable algorithms that preserve properties of the original problem (such as potential convexity or monotonicity).

(1) **Estimation of divergences from moments:** Given some function T from $\mathcal{X}$ to some vector space, the goal is to estimate $D(p\|q)$ defined in Eq. (1) only from the knowledge of the integrals $\int_{\mathcal{X}} T(x)dp(x)$ and $\int_{\mathcal{X}} T(x)dq(x)$. Our aim in this paper is to estimate $D(p\|q)$ from below, and to obtain the largest possible lower bound. We focus on particular functions T of the form $T(x) = \varphi(x)\varphi(x)^*$, where $\varphi : \mathcal{X} \to \mathbb{C}^d$ is some complex-valued feature map, and where M^* denotes the conjugate transpose of the matrix M. Thus, in our particular situation, T takes values in the set $\mathbb{H}_d^+$ of positive semi-definite Hermitian matrices of size $d \times d$. This choice of the feature map T as a rank-one Hermitian matrix is not a limitation in many instances, such as with polynomials (as monomials can be arranged in Hankel matrices) and is key to our methodological developments.

For this particular form of moments as non-centered covariance matrices, we first provide in Section 5 a characterization of the tightest such lower bound. This formulation involves the maximization over $\mathcal{X}$ of quadratic forms in φ, that is, functions of the form $x \mapsto \varphi(x)^* M \varphi(x)$, where $M \in \mathbb{H}_d$ (the set of Hermitian matrices of size $d \times d$).

Our first contribution is to replace the exact maximization of such quadratic forms of $\varphi(x)$ by "sum-of-squares" relaxations, that is, relaxations based on semi-definite programming and the representation of non-negative functions as positive-semidefinite quadratic forms in $\varphi(x)$ [34, 49] (see review in Section 4). This relaxation is developed in Section 6 and allows to bring to bear the well-developed area of sum-of-squares optimization with its computational tools and extensive analyses. We also provide in Section 7 a further relaxation which is based on information divergences from quantum information theory (which are reviewed in Section 3).

Note that a related interesting task is to estimate estimation divergences directly from samples [46, 54]. We could use our algorithms with increasingly large feature vectors and use empirical estimates, but a detailed analysis is left for future research.

(2) **Variational inference in probabilistic models:** One classical inference task in probabilistic modeling (see [62, 43] and references therein) is to compute moments of some distributions from which we know the density. In our context of f-divergences, we consider a density (with respect to some positive measure q) proportional to $(f^*)'(h(x) - \rho)$, where f^* is the Fenchel conjugate of f, $h : \mathcal{X} \to \mathbb{R}$ is an arbitrary function, and $\rho \in \mathbb{R}$ is a normalizing constant making sure that we obtain a probability distribution. As shown in Section 8.1, this density happens to be exactly the maximizer in

$$c_q(h) = \sup_{p \text{ probability measure on } \mathcal{X}} \int_{\mathcal{X}} h(x)dp(x) - D(p\|q).$$

The optimal quantity $c_q(h)$ is referred to as the f-partition-function, and for $f(t) = t\log t - t + 1$, we recover the usual log-partition function, and densities proportional to $e^{h(x)}$.

When we restrict h to be a quadratic form in $\varphi(x)$, that is of the form $\varphi(x)^* H \varphi(x)$ for some $H \in \mathbb{H}_d$, then, (a) we can replace $D(p\|q)$ by the lower-bound we just defined above, and obtain a computable *upper*-bound of $c_q(h)$, and (b) the gradient with respect to H of the f-partition function ends up being exactly the moment of $T(x) = \varphi(x)\varphi(x)^* \in \mathbb{H}_d$ for the desired distribution. This relaxation is presented in Section 9, and can be extended to the task of computing integrals of the form $\int_{\mathcal{X}} f^*(h(x))dq(x)$ (see Appendix D).

For well-chosen feature vectors, e.g., polynomials on $\{-1,1\}^d$, log-densities that can be expressed as quadratic forms cover a wide set of Markov random field models in statistical modeling [36] and image processing [10]. For these models, exact moment estimation and log-partition estimations are key computational tasks that are intractable, even in moderate dimensions, hence the need for approximations. The formulations presented in this paper follow a line of work based on tractable convex relaxations, typically based on linear programming, with few examples using the more powerful semi-definite programming framework that we further develop (see [62] for a thorough introduction).

Contributions. In this paper, we first derive a sequence of three convex formulations of f-divergences based on covariance matrices. Starting from the typically non-tractable optimal lower-bound, we consider an additional relaxation based on "sums-of-squares", which is now computable in polynomial time as a semidefinite program, as well as further computationally more efficient relaxations based on spectral information divergences from quantum information theory. For all of the tasks above, beyond proposing new relaxations, we derive tractable algorithms based on convex optimization, and we present illustrations on multivariate trigonometric polynomials and functions on the Boolean hypercube. We then extend these bounds by duality to lower bounds on partition functions.

Since these contributions involve three traditionally separate domains, we start by a review of those, that is, f-divergences and associated variational formulations in Section 2 (where we propose a new one more adapted to our purposes), quantum information divergences in Section 3, and finally sum-of-squares relaxations in Section 4.

2 Review of f-divergences

We consider f-divergences, with $f : \mathbb{R}_+^* \to \mathbb{R}_+$ is a convex function, where $\mathbb{R}_+^*$ denotes the set of strictly positive real numbers. We assume that f is strictly convex and differentiable, so that the Fenchel conjugate f^* is differentiable and non-decreasing, with $(f^*)'(u) \geqslant 0$ for u in the domain of f^*. Moreover, we assume that $f(1) = 0$, and thus 1 is the minimizer of f, leading to $f'(1) = 0$ and $(f^*)'(0) = 1$. Moreover, we then have $f^*(0) = 0$. Our running example is $f(t) = t \log t - t + 1$ with $f^*(u) = e^u - 1$ (see more examples below).

On the set $\mathcal{X}$ (which we only assume to be equipped with a topology, and compact), we consider several sets of finite Borel measures: $\mathcal{M}_+(\mathcal{X})$ the set of finite *positive* measures on $\mathcal{X}$, $\mathcal{M}(\mathcal{X})$ the set of finite *signed* measures on $\mathcal{X}$, and $\mathcal{P}(\mathcal{X})$ the set of *probability* measures on $\mathcal{X}$ (that is, finite positive measures in $\mathcal{M}_+(\mathcal{X})$ that integrate to one).

For two finite positive measures p, q in $\mathcal{M}_+(\mathcal{X})$, we can define

$$D(p\|q) = \int_{\mathcal{X}} f\Big(\frac{dp}{dq}(x)\Big) dq(x),$$

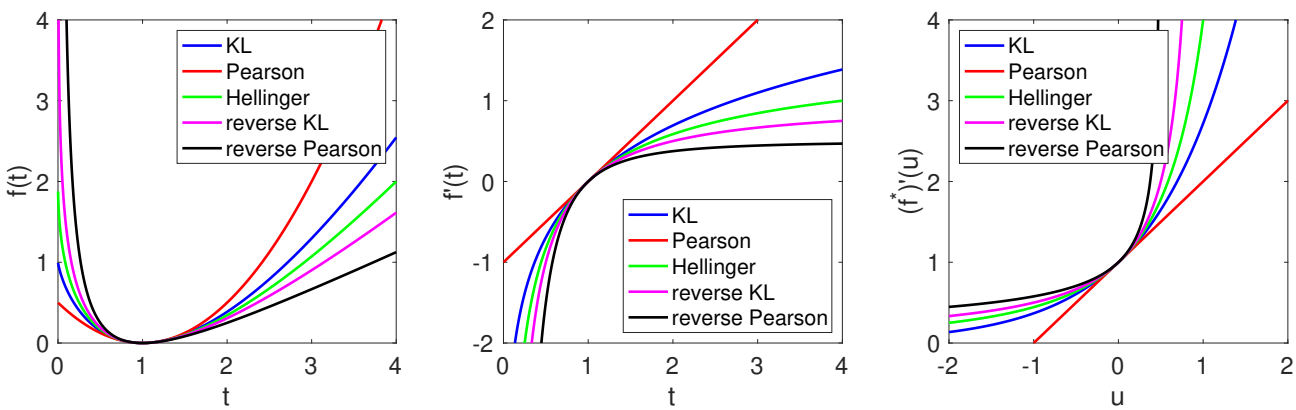
for all non-negative measures (possibly non normalized), assuming that the density $\frac{dp}{dq}(x)$ exists for all $x \in \mathcal{X}$ and that the integral is finite. We now review several properties and examples, see [19, 56, 50] for more results.

Classical properties. Given our assumption that 1 is a global minimizer of f, $f(1) = 0$, and f is strictly convex, we have $D(p\|q) \geqslant 0$ with equality if and only if $p = q$. Moreover, $D(p\|q)$ is jointly convex in p and q.

Examples. We have the following classical examples, with the usual "reversion" of f-divergences: if we define $g(t) = tf(1/t)$, swapping p and q in $D(p\|q)$ is equivalent to replacing f by g (for α-divergences below, this corresponds to replacing α by $1-\alpha$). All of the approximations that we consider in this paper will satisfy this reversibility: swapping p and q (and later moment matrices A and B) is equivalent to replacing f by g.

Note that the total variation case, where $f(t) = |t-1|$ is excluded from most developments because it is neither differentiable nor strictly convex (nor operator convex, as defined in Section 3.1), but many results (except the quantum ones) would apply as well. We normalize all functions f so that $f''(1) = 1$. See table and plots below.

Divergence	$f(t)$	$f^*(u)$	$(f^*)'(u)$
α-Rényi	$\frac{1}{\alpha(\alpha-1)}\big[t^\alpha - \alpha t + (\alpha-1)\big]$	$\frac{1}{\alpha}\big[-1+(1+(\alpha-1)u)^{\alpha/(\alpha-1)}\big]$	$(1+(\alpha-1)u)^{1/(\alpha-1)}$
Kullback-Leibler, $\alpha=1$	$t\log t - t + 1$	$e^u - 1$	e^u
Reverse KL, $\alpha=0$	$-\log t + t - 1$	$-\log(1-u)$	$\frac{1}{1-u}$
squared Hellinger, $\alpha=\frac{1}{2}$	$2(\sqrt{t}-1)^2$	$\frac{u}{1-u/2}$	$\frac{1}{(1-u/2)^2}$
Pearson χ^2, $\alpha=2$	$\frac{1}{2}(t-1)^2$	$\frac{1}{2}(u+1)_+^2 - \frac{1}{2}$	$(u+1)_+$
Reverse Pearson, $\alpha=-1$	$\frac{1}{2}\big(\frac{1}{t}+t\big) - 1$	$1-\sqrt{1-2u}$	$\frac{1}{\sqrt{1-2u}}$
Le Cam	$\frac{(t-1)^2}{t+1}$	$2-u-2\sqrt{1-2u}$	$\frac{2}{\sqrt{1-2u}} - 1$
Jensen-Shannon	$2t\log\frac{2t}{t+1} + 2\log\frac{2}{t+1}$	$-2\log(2-e^{u/2})$	$\frac{1}{2\exp(-u/2)-1}$



2.1 Variational representations

The f-divergence has a variational representation obtained from the Fenchel conjugate of perspective functions [53]. Indeed, the function $(p,q) \mapsto qf\big(\frac{p}{q}\big)$ defined on $\mathbb{R}_+ \times \mathbb{R}_+^*$ is referred to as the *perspective*

function of f, is convex, and has the variational representation for $p \in \mathbb{R}_+$ and $q \in \mathbb{R}_+^*$:

$$qf\Big(\frac{p}{q}\Big) = \sup_{v,w \in \mathbb{R}} \; vp + wq \;\text{ such that }\; \forall r \geqslant 0,\; rv + w \leqslant f(r), \tag{2}$$

where for p and $q \in \mathbb{R}_+^*$, the maximizer w is $\inf_{r \geqslant 0} f(r) - rv = -f^*(v)$, and thus the optimal value of v is the supremum of $vp - qf^*(v)$, leading to optimal values $v^* = f'\big(\frac{p}{q}\big)$, and $w^* = -f^*(v^*) = f\big(\frac{p}{q}\big) - \frac{p}{q} f'\big(\frac{p}{q}\big)$. Depending on the behavior of f at 0 and $+\infty$, we can extend the perspective function to $\mathbb{R}_+ \times \mathbb{R}_+$.

Following [40], for $p, q \in \mathcal{M}_+(\mathcal{X})$, applying Eq. (2) to densities, this leads to a variational representation of $D(p\|q)$ as the supremum of linear functions of the measures p and q (with functions v and w that are measurable and bounded):

$$D(p\|q) = \sup_{v,w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} w(x) dq(x) \;\text{ such that }\; \forall x \in \mathcal{X},\; \forall r \geqslant 0,\; rv(x) + w(x) \leqslant f(r). \tag{3}$$

The optimal functions w and v are such that $v(x) = f'\big(\frac{dp}{dq}(x)\big)$, and $w(x) = f\big(\frac{dp}{dq}(x)\big) - \frac{dp}{dq}(x) f'\big(\frac{dp}{dq}(x)\big) = -f^*(v(x))$. We can also use the function g defined above through $g(t) = tf(1/t)$, to get $w(t) = g'\big(\frac{dq}{dp}(x)\big)$.

Note that in this representation, the non-negativity of the measures p and q is automatically satisfied (the value of the optimization problem in Eq. (3) is infinite otherwise). Optimizing with respect to $w(x)$ in closed form as above then leads to the representation of $D(p\|q)$ from [13, 46] as the supremum with respect to $v : \mathcal{X} \to \mathbb{R}$ of $\int_{\mathcal{X}} v(x) dp(x) - \int_{\mathcal{X}} f^*(v(x)) dq(x)$.

New variational unconstrained formulation. In this paper, we will need a novel *unconstrained* variational formulation similar to Eq. (3). To this effect, we introduce the function $F : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$F(v, w) = \sup_{r \geqslant 0} \frac{rv + w - f(r)}{r + 1}. \tag{4}$$

The function F is convex as a supremum of affine functions and given our assumption on f that $f(t) \geqslant f(1) = 0$ for all $t > 0$, we have $\frac{v+w}{2} \leqslant F(v, w) \leqslant \max\{v, w\}$ (which also shows it has full domain). In addition all subgradients of F are in the simplex in $\mathbb{R}^2$. Moreover, for any constant $u \in \mathbb{R}$, $F(v-u, w-u) = F(v, w) - u$, and $F(w, v) \leqslant 0$ if and only if for all $r \geqslant 0$, $rv + w - f(r) \leqslant 0$.

Thus, starting from Eq. (2), we have, for $p, q \in \mathbb{R}_+^*$:

$$\begin{aligned}
qf\Big(\frac{p}{q}\Big) &= \sup_{v,w \in \mathbb{R}} \; vp + wq \;\text{ such that } F(v, w) \leqslant 0 \\
&= \sup_{u,v,w \in \mathbb{R}} \; (v-u)p + (w-u)q \;\text{ such that } F(v-u, w-u) \leqslant 0, \text{ by adding an extra variable } u, \\
&= \sup_{u,v,w \in \mathbb{R}} \; vp + wq - (p+q)u \;\text{ such that } F(v, w) \leqslant u, \text{ using properties of } F, \\
&= \sup_{v,w \in \mathbb{R}} \; vp + wq - (p+q)F(v, w) \text{ since the optimal } u \text{ is } F(v, w).
\end{aligned} \tag{5}$$

Using the same technique as above and replacing $w(x)$ and $v(x)$ by $w(x) - u$ and $v(x) - u$) in Eq. (3) , we can optimize with respect to u and thus obtain: as

$$D(p\|q) = \sup_{v,w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} w(x) dq(x) - \Big(\int_{\mathcal{X}} dp(x) + \int_{\mathcal{X}} dq(x) \Big) \sup_{x \in \mathcal{X}} F\big(v(x), w(x)\big). \tag{6}$$

Eq. (6) above will be crucial when estimating f-divergences of partition functions because it leads to *unconstrained* optimization problems, while Eq. (3) was constrained.

Computing F. In our algorithms in later sections, we will need to compute F for any $(v, w) \in \mathbb{R}^2$, that is, solve for $r \geqslant 0$ in Eq. (4). This is a one-dimensional root-finding problem for which Newton method can be used with quadratic convergence, and thus with few iterations. See Appendix A for details for the function $f(t) = t \log t - t + 1$.

Variational formulations as infimum. Through convex duality, we can derive variational formulations of $D(p\|q)$ as minimization problems rather than maximization problems like in Eq. (3) and Eq. (6). Since this is not crucial to our developments, this is presented in Appendix E for all our formulations.

3 Review of quantum information theory

In order to define quantum information divergences, we first need to introduce operator convexity.

3.1 Operator convexity

All the examples of convex functions proposed in Section 2 also happen to be "operator convex", meaning that for two positive semi-definite Hermitian matrices A, B, and any $\lambda \in [0, 1]$,

$$f(\lambda A + (1-\lambda)B) \preccurlyeq \lambda f(A) + (1-\lambda) f(B),$$

where $\preccurlyeq$ defines the Löwner order between Hermitian matrices ($A \preccurlyeq B$ if and only if $B - A$ is positive semi-definite), and $f(A)$ is the spectral function defined as $f(A) = \sum_{i=1}^d f(\lambda_i) u_i u_i^*$ when $A = \sum_{i=1}^d \lambda_i u_i u_i^*$ is an eigenvalue decomposition of A.

A classical necessary and sufficient condition for f being operator-convex is the existence of a representation of f as, for f additionally satisfying $f(1) = f'(1) = 0$:

$$f(t) = \beta(t-1)^2 + (t-1)^2 \int_0^{+\infty} \frac{1}{\lambda + t} d\nu(\lambda), \tag{7}$$

for some $\beta \in \mathbb{R}_+$ and a positive measure ν on $\mathbb{R}_+$ [9]. When the function f is extendable to an analytic function on $\mathbb{C}$, then the measure ν can be obtained from the Stieltjes inversion formula [63], as the limit of the measure with density $\frac{1}{\pi}\mathrm{Im}\Big(\frac{f(-\lambda - it)}{(\lambda + it + 1)^2}\Big) - \beta$ when $t \to 0^+$. In Appendix B, we provide this decomposition for the examples from the beginning of Section 2.

Operator convexity is crucial for the quantum information divergences that we now consider.

3.2 Quantum information divergences

We consider two Hermitian positive semi-definite matrices A and B in $\mathbb{H}_d^+$. If A and B commute, then they are jointly diagonalizable, and we can naturally define a divergence as

$$\sum_{i=1}^d \lambda_i(B) f\Big(\frac{\lambda_i(A)}{\lambda_i(B)}\Big),$$

where $\lambda_i(A)$ and $\lambda_i(B)$ are the corresponding non-negative eigenvalues of A and B (with the same eigenvectors). When A and B do not commute, there are several notions of f-information divergences that reduce to the formula above when matrices commute [60]. Among the several candidates from quantum information theory [40, 23, 29], two are particularly interesting in our context.

The so-called *maximal divergence* is equal to

$$\tilde{D}^{\mathrm{QT}}_{\max}(A\|B) = \operatorname{tr}\left[B^{1/2} f(B^{-1/2}AB^{-1/2})B^{1/2}\right] = \operatorname{tr}\left[B f(B^{-1/2}AB^{-1/2})\right],$$

while the *standard divergence* is equal to:

$$\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B) = \operatorname{vec}(B^{1/2})^* f(A\otimes B^{-1})\operatorname{vec}(B^{1/2}), \tag{8}$$

with the usual Kronecker product notation between matrices and $\operatorname{vec}(M)$ the column vector obtained by stacking the columns of M [26]. It is equal to

$$\sum_{i,j=1}^{d} \lambda_i f\left(\frac{\mu_j}{\lambda_i}\right)|u_i^* v_j|^2,$$

where $A = \sum_{j=1}^d \mu_j v_j v_j^*$ and $B = \sum_{i=1}^d \lambda_i u_i u_i^*$ are eigenvalue decompositions of A and B. Both are jointly convex and equal to zero if and only if $A = B$.

An important feature of these divergences is that they can both be computed in closed form from spectral decompositions. This will give a strong computational advantage for the relaxations that are based on these.

Examples of the standard divergence. We have the following classical examples below, with simpler formulas than Eq. (8), where we recover the von Neumann relative entropy and classical matrix formulations of the Rényi entropies.

Divergence	$f(t)$	$\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B)$
α-Rényi	$\frac{1}{\alpha(\alpha-1)}\left[t^\alpha - \alpha t + (\alpha-1)\right]$	$\frac{1}{\alpha(\alpha-1)}\left[\operatorname{tr}[B^{1-\alpha}A^\alpha] - \alpha\operatorname{tr}[A] + (\alpha-1)\operatorname{tr}[B]\right]$
Kullback-Leibler, $\alpha = 1$	$t\log t - t + 1$	$\operatorname{tr}\left[A\log A - A\log B\right]$
squared Hellinger, $\alpha = \frac{1}{2}$	$2(\sqrt{t}-1)^2$	$2\operatorname{tr} A + 2\operatorname{tr} B - 4\operatorname{tr}\left[A^{1/2}B^{1/2}\right]$
Pearson χ^2, $\alpha = 2$	$\frac{1}{2}(t-1)^2$	$\frac{1}{2}\operatorname{tr}\left[B^{-1}(B-A)^2\right]$

From the representation of operator convex functions in Eq. (7), we can infer properties of these divergences from the particular example $f(t) = \frac{(t-1)^2}{\lambda+t}$ for $\lambda > 0$, for which we have

$$\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B) = \operatorname{vec}(A-B)^*(A\otimes I + \lambda\cdot B\otimes I)^{-1}\operatorname{vec}(A-B),$$

and

$$\tilde{D}^{\mathrm{QT}}_{\max}(A\|B) = \operatorname{tr}\left[(A-B)(A+\lambda B)^{-1}(A-B)\right].$$

This shows immediately that the two quantum divergences are jointly convex in A and B. A less direct property is that for all A and B (see proof in [29, Prop. 4.1]):

$$\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B) \leqslant \tilde{D}^{\mathrm{QT}}_{\max}(A\|B).$$

Thus, in our context of lower bounds on the regular f-divergence $D(p\|q)$, we get a tighter result with $\tilde{D}^{\mathrm{QT}}_{\max}$, and a strict improvement over [4] which uses $\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}$ in the same context (but only for the Kullback-Leibler divergence). Moreover, the key property outlined by [4] that $\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B) \leqslant D(p\|q)$ for $A = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$ and $B = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dq(x)$ as soon as for all $x \in \mathcal{X}$, $\|\varphi(x)\|^2 \leqslant 1$, is preserved for $\tilde{D}^{\mathrm{QT}}_{\max}$. In fact, in this paper, we derive a sequence of lower bounds that are all improvements on [4]. See Table 1 in Section 11 for a summary.

While the Fenchel conjugate of $\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B)$ with respect to A can be computed in closed form in most cases, this is not the case for $\tilde{D}^{\mathrm{QT}}_{\max}(A\|B)$. Thus, some of the algorithms from [4] cannot be extended and we need to derive new ones based on the unconstrained variational formulation presented in Section 2.1.

Special case of von Neumann relative entropy. When $f(t) = t\log t - t + 1$, for the standard divergence $\tilde{D}^{\mathrm{QT}}_{\mathrm{standard}}(A\|B)$, we get $\operatorname{tr}\big[A\log A - A\log B\big]$, which is the Bregman divergence associated with the von Neumann entropy $A \mapsto \operatorname{tr}[A\log A]$. Note that this is different from seeing that A and B are covariance matrices, and considering the Kullback-Leibler between zero-mean Gaussian distributions with these covariance matrices (which would lead to $\frac{1}{2}\operatorname{tr}[AB^{-1}] - \frac{1}{2}\log\det[AB^{-1}] - \frac{d}{2}$). For an approach linking semi-definite programming and Gaussian entropies, see [31].

4 Review of sum-of-squares relaxations

In this section, we assume that φ is continuous on $\mathcal{X}$, and thus bounded since $\mathcal{X}$ is assumed compact. To make the results simpler, we assume that features are normalized to unit norm, that is, $\forall x \in \mathcal{X}$, $\|\varphi(x)\| = 1$, for the standard Hermitian norm. We consider the task of computing

$$\Gamma(M) = \max_{x\in\mathcal{X}} \ \varphi(x)^* M \varphi(x), \tag{9}$$

for some matrix $M \in \mathbb{H}_d$. Since φ is bounded, Γ is a positively homogeneous everywhere finite convex function on $\mathbb{H}_d$. We now introduce necessary tools and notations for presenting sum-of-squares (SOS) relaxations [34, 49].

Let $\mathcal{K}$ be the closure of the convex hull of all $\varphi(x)\varphi(x)^*$, $x \in \mathcal{X}$, $\mathcal{C}$ the closure of its conic hull, and $\mathcal{V}$ its linear span. By construction, we have $\mathcal{K} \subset \mathcal{C} \subset \mathcal{V}$, and, since we have assumed that for all $x \in \mathcal{X}$, $\|\varphi(x)\|$=1, we have:

$$\Sigma \in \mathcal{K} \quad \Leftrightarrow \quad \Sigma \in \mathcal{C} \text{ and } \operatorname{tr}[\Sigma] = 1.$$

We make the extra assumption that $\mathcal{V}$ contains a positive definite matrix (this will most often be the identity matrix in examples in Section 4.2).

By definition of Γ in Eq. (9), and since maximizing linear functions of $\varphi(x)\varphi(x)^*$ with respect to $x \in \mathcal{X}$ leads to the same value as maximizing over its convex hull, we get:

$$\Gamma(M) = \max_{x\in\mathcal{X}} \ \ \varphi(x)^* M \varphi(x) = \max_{\Sigma\in\mathcal{K}} \ \ \operatorname{tr}[\Sigma M],$$

that is, the function Γ is the support function of the convex set $\mathcal{K}$. Moreover, using our notations for finite measures from Section 2, we have $\mathcal{V} = \big\{ \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x),\ p \in \mathcal{M}(\mathcal{X})\big\}$, $\mathcal{C} = \big\{ \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x),\ p \in \mathcal{M}_+(\mathcal{X})\big\}$ and $\mathcal{K} = \big\{ \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x),\ p \in \mathcal{P}(\mathcal{X})\big\}$.

4.1 Outer approximations of convex hulls

In order to obtain an *upper-bound* on $\Gamma(M)$ defined in Eq. (9), we will look for *outer* approximations of the set $\mathcal{K}$. By construction, the convex hull is included in the affine hull, that is, if $\Sigma \in \mathcal{K}$, then $\mathrm{tr}[\Sigma] = 1$ and $\Sigma \in \mathcal{V}$. The extra condition we will use in this paper follows [34, 49], and is simply that Σ is positive semi-definite, which is a direct consequence of $\varphi(x)\varphi(x)^* \in \mathbb{H}_d^+$ for all $x \in \mathcal{X}$.

We thus consider outer approximations of $\mathcal{K} = \mathcal{C} \cap \{\Sigma,\ \mathrm{tr}[\Sigma] = 1\}$, through the outer approximation of $\mathcal{C}$ as $\widehat{\mathcal{C}} = \mathcal{V} \cap \mathbb{H}_d^+$, which corresponds to $\widehat{\mathcal{K}} = \mathcal{V} \cap \mathbb{H}_d^+ \cap \{\Sigma,\ \mathrm{tr}[\Sigma] = 1\}$, with $\mathbb{H}_d^+$ the set of PSD Hermitian matrices. This leads to our approximation of $\Gamma(M)$ as:

$$\widehat{\Gamma}(M) \quad = \quad \max_{\Sigma \in \widehat{\mathcal{K}}} \ \mathrm{tr}[\Sigma M] \quad = \quad \max_{\Sigma \in \mathbb{H}_d} \ \mathrm{tr}[\Sigma M] \quad \text{such that} \quad \mathrm{tr}[\Sigma] = 1,\ \Sigma \in \mathcal{V},\ \text{and}\ \Sigma \succcurlyeq 0, \tag{10}$$

which satisfies $\Gamma(M) \leqslant \widehat{\Gamma}(M)$ for all $M \in \mathbb{H}_d$.

These relaxations are often referred to as "sum-of-squares" (SOS) relaxations, because of the following dual interpretation.[2] Introducing Lagrange multipliers, $c \in \mathbb{R}$ for the constraint $\mathrm{tr}[\Sigma] = 1$, $Y \in \mathcal{V}^\perp$ for $\Sigma \in \mathcal{V}$, and $B \succcurlyeq 0$ for $\Sigma \succcurlyeq 0$, we get, using strong duality (which holds by Slater's condition since Eq. (12) has a strictly feasible point):

$$\begin{aligned}
\widehat{\Gamma}(M) \quad &= \quad \sup_{\Sigma \in \mathbb{H}_d} \ \inf_{c \in \mathbb{R},\ Y \in \mathcal{V}^\perp,\ B \succcurlyeq 0} \ \mathrm{tr}[\Sigma M] + c(1 - \mathrm{tr}[\Sigma]) + \mathrm{tr}[Y\Sigma] + \mathrm{tr}[B\Sigma] \\
&= \quad \inf_{c \in \mathbb{R},\ Y \in \mathcal{V}^\perp,\ B \succcurlyeq 0} \ \sup_{\Sigma \in \mathbb{H}_d} \ \mathrm{tr}[\Sigma M] + c(1 - \mathrm{tr}[\Sigma]) + \mathrm{tr}[Y\Sigma] + \mathrm{tr}[B\Sigma] \qquad (11) \\
&= \quad \inf_{c \in \mathbb{R},\ Y \in \mathcal{V}^\perp,\ B \succcurlyeq 0} \ c \quad \text{such that} \quad M = cI - Y - B \qquad (12) \\
&= \quad \inf_{c \in \mathbb{R},\ B \succcurlyeq 0} \ c \quad \text{such that} \quad \forall x \in \mathcal{X},\ c - \varphi(x)^* M \varphi(x) = \varphi(x)^* B \varphi(x).
\end{aligned}$$

This can be interpreted as finding the lowest upper-bound c on the function $x \mapsto \varphi(x)^* M \varphi(x)$ by relaxing the non-negativity of $c - \varphi(x)^* M \varphi(x)$ by the existence of $B \succcurlyeq 0$ such that $c - \varphi(x)^* M \varphi(x) = \varphi(x)^* B \varphi(x)$ (which is indeed non-negative). Finally, using the eigendecomposition of B, $\varphi(x)^* B \varphi(x)$ can be written as a sum of square functions, hence the denomination. Note that it is common to add extra conic constraints to further restrict $\widehat{\mathcal{C}}^*$, often leading to hierarchies of relaxations (see examples below and [34]), which makes the relaxations tighter and tighter. This corresponds to adding to φ another feature vector φ^+, and see a quadratic form in φ as a quadratic form in $\tilde{\varphi} = \binom{\varphi}{\varphi^+}$, which leads to a tighter relaxation.

In terms of computational complexity, because Slater's condition is satisfied, interior-point methods have a polynomial number of iterations, each based on polynomial-time numerical linear algebra algorithms [44].

Spectral relaxation. Because of our unit norm assumption for the features, we can maximize with respect to c and B in Eq. (12) and reformulate Eq. (12) as the problem of minimizing $\lambda_{\max}(M + Y)$ over $Y \in \mathcal{V}^\perp$. Simply taking $Y = 0$ corresponds to computing the largest eigenvalue of M. This corresponds to relaxing $\mathcal{K}$ to $\{\Sigma \succcurlyeq 0,\ \mathrm{tr}[\Sigma] = 1\}$.

Throughout the paper, we will often use the following statements based on dual cones (using that the dual of the intersection of closed cones is the sum of their duals when their relative interiors intersect [53,

[2] Note that using an outer approximation in Eq. (10) leads to a relaxation *per se*, while in the dual view presented here, we obtain a *strengthening* due to replacing non-negative functions by the smaller set of sums-of-squares. Throughout the paper, we will use the term "relaxation" in all cases.

Corollary 16.4.2], and the assumption that $\mathcal{V}$ contains a positive definite matrix):

$$\begin{array}{rcl} \Gamma(M) \leqslant t & \Leftrightarrow & \Gamma(tI - M) \leqslant 0 \quad \Leftrightarrow \quad tI - M \in \mathcal{C}^* \\ \widehat{\Gamma}(M) \leqslant t & \Leftrightarrow & \widehat{\Gamma}(tI - M) \leqslant 0 \quad \Leftrightarrow \quad tI - M \in \widehat{\mathcal{C}}^* = \mathbb{H}_d^+ + \mathcal{V}^\perp \\ & \Leftrightarrow & \exists Y \in \mathcal{V}^\perp, \ \lambda_{\max}(M + Y) \leqslant t. \end{array} \tag{13}$$

4.2 Examples

Finite set with injective embedding. If $\mathcal{X}$ is finite and the Gram matrix of all features for all values of $\mathcal{X}$ is invertible, then the SOS relaxation is tight. Indeed, assuming (potentially after applying an invertible linear transformation to φ) that $\varphi(x)^*\varphi(y) = 1_{y=x}$, $\mathcal{K}$ is the set of diagonal matrices with a diagonal belonging to the simplex.

Affine functions on the Euclidean unit sphere. We consider $\mathcal{X}$ the unit sphere in $\mathbb{R}^{d-1}$, with $\varphi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ x \end{pmatrix} \in \mathbb{R}^d$. Then $\mathcal{V}$ is the set of matrices $\begin{pmatrix} \alpha & x^\top \\ x & X \end{pmatrix}$ such that $\operatorname{tr}(X) = \alpha$. This is another situation where the sum-of-squares relaxation is tight [27].

Trigonometric polynomials on $[0, 1]$. We consider $\mathcal{X} = [-1, 1]$ and $\varphi(x) \in \mathbb{C}^{2r+1}$, with $\varphi(x)_\omega = e^{2i\pi\omega x}/\sqrt{2r+1}$ for $\omega \in \{-r, \dots, r\}$. Then $(\varphi(x)\varphi(x)^*)_{\omega\omega'} = e^{2i\pi(\omega-\omega')x}/(2r+1)$, and thus $\mathcal{V}$ is the set of Hermitian Toeplitz matrices. It turns out that the sum-of-squares relaxation is tight, see [59, Theorem 1.2.1] and [21].

Trigonometric polynomials on $[0, 1]^n$. We consider $\mathcal{X} = [-1, 1]^n$ and $\varphi(x)_\omega = e^{2i\pi\omega^\top x}/(2r+1)^{n/2} \in \mathbb{C}$ for ω in a certain set $\Omega \subset \mathbb{Z}^n$, typically $\Omega = \{\omega \in \mathbb{Z}^n, \ \|\omega\|_\infty \leqslant r\}$. We then have $(\varphi(x)\varphi(x)^*)_{\omega\omega'} = e^{2i\pi(\omega-\omega')^\top x}/(2r+1)^n$, which depends only on $\omega - \omega'$, which defines a set of linear constraints defining $\mathcal{V}$. The relaxation is then not tight, but by embedding Ω in a larger set, we can make the relaxation as tight as desired (see [21]), while for $n = 2$, it will be tight after a certain (unknown) degree [57, Corollary 3.4].

Polynomials on $[-1, 1]$. In order to tackle polynomials, we could simply consider $\varphi(x)$ composed of monomials, but this would not lead to a normalized feature map. Rather, following [6, Section 2.2], given a classical polynomial P, we consider the trigonometric polynomial $x \mapsto f(x) = P(\cos 2\pi x)$, which we can represent with a normalized feature map. This extends to polynomials on $[-1, 1]^n$.

Polynomials on the Euclidean hypersphere. With $\mathcal{X} = \{x \in \mathbb{R}^{n+1}, x^\top x = 1\}$, we can consider all harmonic polynomials, as described by [22]. By only considering functions depending on the first n variables, this allows to consider $\mathcal{X}$ the Euclidean unit ball.

Boolean hypercube. We consider $\mathcal{X} = \{-1, 1\}^n$ with feature vectors composed of Boolean Fourier components of increasing orders [47]. This corresponds to features $\varphi_A(x) = \prod_{i \in A} x_i \in \{-1, 1\}$, where A is a subset of $\{1, \dots, n\}$. Moreover, given two sets A and B, we have $\varphi_A(x)\varphi_B(x) = \varphi_{A \Delta B}(x)$, where $A \Delta B$ is the symmetric difference between A and B.

If we consider a set $\mathcal{A}$ of subsets of $\{1, \ldots, n\}$, then, the element indexed (A, B) of $\varphi(x)\varphi(x)^*$ only depends on the symmetric difference $A \Delta B = (A \backslash B) \cup (B \backslash A)$, and this leads to a set of linear constraints defining $\mathcal{V}$. The relaxation is not tight in general, but if we see our moment matrix as a submatrix obtained from a sufficiently larger set of subsets, then we obtain a tight formulation (see [33, 35, 58] and references therein).

5 Exact lower bounds based on moments

We consider the optimal lower bound on $D(p\|q)$ given the integrals $\Sigma_p = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$ and $\Sigma_q = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dq(x)$,[3] that is, given two Hermitian matrices A and B:

$$D^{\mathrm{OPT}}(A\|B) = \inf_{p,q \in \mathcal{M}_+(\mathcal{X})} D(p\|q) \ \text{ such that } \ \Sigma_p = A \text{ and } \Sigma_q = B. \tag{14}$$

Note that we do not assume that p and q integrates to one. By construction, $D^{\mathrm{OPT}}(\Sigma_p\|\Sigma_q) \leqslant D(p\|q)$. Moreover, we have some immediate properties for the function of (A, B) defined in Eq. (14), which preserves similar properties of $D(p\|q)$. For other potential properties such as used within multivariate probabilistic modeling, see [4]:

- If A or B is not in $\mathcal{C}$ (the closure of the convex hull of all $\varphi(x)\varphi(x)^*$), then the optimization problem in Eq. (14) is infeasible as no (p, q) can be found to satisfy $\Sigma_p = A$ and $\Sigma_q = B$, and, following the standard convention in convex analysis [53], we set the value of $D^{\mathrm{OPT}}(A\|B)$ to infinity.
- If $\mathrm{tr}[A] = \mathrm{tr}[B] = 1$, since φ is normalized to unit norm, then the optimal measures p and q in Eq. (14) are probability measures.
- The function $(A, B) \mapsto D^{\mathrm{OPT}}(A\|B)$ is *jointly* convex in A and B, as the optimal value of a jointly convex problem in A, B, p, q.
- If φ is replaced by $T\varphi$ for an injective linear map T (which changes also Σ_p and Σ_q), the quantity $D^{\mathrm{OPT}}(\Sigma_q\|\Sigma_q)$ is unchanged.[4]
- If φ is replaced by $T\varphi$ for a (potentially non-injective) linear map T, the quantity $D^{\mathrm{OPT}}(\Sigma_q\|\Sigma_q)$ is reduced. Therefore, to have tighter lower-bounds $D^{\mathrm{OPT}}(\Sigma_q\|\Sigma_q)$ on $D(p\|q)$, we need to use high-dimensional features. In other words, for all p, q, the approximation is typically tight, that is, $D(p\|q)$ close to $D^{\mathrm{OPT}}(\Sigma_p\|\Sigma_q)$ only if the feature $\varphi : \mathcal{X} \to \mathbb{C}^d$ is rich enough. For approximation capabilities when the feature size grows to infinity, and the use of positive definite kernel methods, see [4]. In this paper, all feature vectors will have a fixed dimension.

Variational representation. We have, using the representation of the f-divergence $D(p\|q)$ from Eq. (3), and strong convex duality for an infinite-dimensional optimization problem with linear constraints [39, Section 8.6], with Lagrange multipliers $M, N \in \mathbb{H}_d$, for the finite-dimensional equality constraints $A =$

[3] Note that Σ_p and Σ_q do depend on φ, but we omit the dependence in the notation.

[4] In this section, we do not need to impose that $\|T\varphi(x)\| = 1$ for all $x \in \mathcal{X}$, but we will in subsequent sections.

$\int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)$ and $B = \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dq(x)$:

$$
\begin{aligned}
& D^{\mathrm{OPT}}(A\|B) \\
= & \inf_{p,q\in\mathcal{M}(\mathcal{X})} \sup_{M,N\in\mathbb{H}_d,\ v,w:\mathcal{X}\to\mathbb{R}} \operatorname{tr}\Big[M\Big(A - \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dp(x)\Big)\Big] + \operatorname{tr}\Big[N\Big(B - \int_{\mathcal{X}} \varphi(x)\varphi(x)^* dq(x)\Big)\Big] \\
& \qquad + \int_{\mathcal{X}} v(x)dp(x) + \int_{\mathcal{X}} w(x)dq(x) \text{ such that } \forall x\in\mathcal{X}, \forall r \geqslant 0,\ rv(x) + w(x) \leqslant f(r) \\
= & \sup_{M,N\in\mathbb{H}_d,\ v,w:\mathcal{X}\to\mathbb{R}} \inf_{p,q\in\mathcal{M}(\mathcal{X})} \operatorname{tr}[MA] + \operatorname{tr}[NB] + \int_{\mathcal{X}} \big(v(x) - \varphi(x)^* M\varphi(x)\big)dp(x) \\
& \qquad + \int_{\mathcal{X}} \big(w(x) - \varphi(x)^* N\varphi(x)\big)dq(x) \text{ such that } \forall x\in\mathcal{X}, \forall r \geqslant 0,\ rv(x) + w(x) \leqslant f(r) \\
& \qquad\qquad \text{by swapping inf and sup,} \\
= & \sup_{M,N\in\mathbb{H}_d} \operatorname{tr}[MA] + \operatorname{tr}[NB] \text{ such that } \forall x\in\mathcal{X}, \forall r\geqslant 0,\ r\varphi(x)^* M\varphi(x) + \varphi(x)^* N\varphi(x) \leqslant f(r) \qquad (15) \\
& \qquad \text{since taking the infimum with respect to } p \text{ and } q \text{ leads to explicit expressions for } v \text{ and } w, \\
= & \sup_{M,N\in\mathbb{H}_d} \operatorname{tr}[MA] + \operatorname{tr}[NB] \text{ such that } \forall r\geqslant 0,\ \Gamma(rM+N) \leqslant f(r), \text{ by definition of } \Gamma. \qquad (16)
\end{aligned}
$$

Note that above, we can restrict M and N to be in $\mathcal{V}$ by orthogonal projection on $\mathcal{V}$, as any component along $\mathcal{V}^\perp$ does not change the optimization problem. The representation above shows that being able to compute D^{OPT} requires the computability of Γ defined in Eq. (9), that is, maximizing quadratic forms in $\varphi(x)$, which is exactly what SOS methods presented in Section 4 are tailored to approximate, and that will be used in Section 6 below.

Algorithms to compute $D^{\mathrm{OPT}}(A\|B)$. This tightest lower bound can only be computed precisely, if we can compute Γ arbitrarily precisely. This is typically only easily possible without brute force enumeration with sum-of-squares relaxations which are asymptotically tight (thus using hierarchies in dimensions larger than one). In our experiments where we compare all bounds, we consider the case of uni-dimensional trigonometric polynomials, for which our simplest relaxation is already tight. Computable lower bounds are considered in Section 6 (based on SOS relaxations) and Section 7 (based on quantum information divergences).

Alternative derivation. While we gave a definition of $D^{\mathrm{OPT}}(A\|B)$ in terms of smallest possible $D(p\|q)$ given $\Sigma_p = A$ and $\Sigma_q = B$, the formulation in Eq. (15) shows that it is equivalent to taking the variational formulation of f-divergences in Eq. (3), and only allowing quadratic forms in φ for the functions w and v, that is, $w(x) = \varphi(x)^* M\varphi(x)$ and $v(x) = \varphi(x)^* N\varphi(x)$.

Relaxations. What makes the computation of $D^{\mathrm{OPT}}(A\|B)$ difficult is the need to deal with the constraints $\Gamma(rM+N) \leqslant f(r)$ in Eq. (16). In the next two sections, we will explore two successive relaxations:

- The first one in Section 6 corresponds to using the constraints $\widehat{\Gamma}(rM+N) \leqslant f(r)$, which is thus using the SOS relaxation for the optimization problem. This is equivalent to the existence of $Y^{(r)} \in \mathcal{V}^\perp$ (one for each r) such that $rM + N + Y^{(r)} \preccurlyeq f(r)I$. Equivalently, this is using $\widehat{\mathcal{C}}$ instead of $\mathcal{C}$, and will lead to the lower-bound $D^{\mathrm{SOS}}(A\|B)$.

- The second one in Section 7 adds a further relaxation by having a unique Y independent of r, in a specific form that leads to a direct link with spectral quantum divergence from Section 3.2. This will lead to the lower-bound $D^{\mathrm{QT}}(A\|B)$.

6 Relaxed f-divergence based on SOS

We consider replacing Γ in the optimal relaxation $D^{\mathrm{OPT}}(A\|B)$ in Eq. (16) by its approximation $\widehat{\Gamma}$ based on sums-of-squares, as defined in Section 4. This leads to, using that $\widehat{\mathcal{C}} = \mathbb{H}_d^+ \cap \mathcal{V}$, and thus $\widehat{\mathcal{C}}^* = \mathbb{H}_d^+ + \mathcal{V}^\perp$, from Eq. (13):

$$\begin{aligned} D^{\mathrm{SOS}}(A\|B) &= \sup_{M,N\in\mathbb{H}_d} \ \mathrm{tr}[AM]+\mathrm{tr}[BN] \ \text{ such that } \ \forall r\geqslant 0,\ \widehat{\Gamma}(rM+N)\leqslant f(r) \qquad (17)\\ &= \sup_{M,N\in\mathbb{H}_d} \ \mathrm{tr}[AM]+\mathrm{tr}[BN] \ \text{ such that } \ \forall r\geqslant 0,\ f(r)I - rM - N \in \widehat{\mathcal{C}}^* = \mathbb{H}_d^+ + \mathcal{V}^\perp. \end{aligned}$$

Since $\Gamma \leqslant \widehat{\Gamma}$, we have by construction $D^{\mathrm{SOS}} \leqslant D^{\mathrm{OPT}}$. Most of the properties of D^{OPT} are preserved, such as convexity, invariance by invertible transforms. In terms of domain, it is now finite if only if $A, B \in \widehat{\mathcal{C}} = \mathbb{H}_d^+ \cap \mathcal{V}$ (rather than $\mathcal{C}$).

Variational unconstrained formulation. In order to derive an algorithm to estimate $D^{\mathrm{SOS}}(A\|B)$, we will use the same unconstrained variational formulation as Eq. (5),[5] and introduce the function:

$$\begin{aligned} \widehat{G}(M,N) &= \sup_{\Sigma\in\widehat{\mathcal{K}}} F\big(\mathrm{tr}[M\Sigma],\mathrm{tr}[N\Sigma]\big) = \sup_{\Sigma\in\widehat{\mathcal{K}}}\sup_{r\geqslant 0} \frac{r\,\mathrm{tr}[M\Sigma]+\mathrm{tr}[N\Sigma]-f(r)}{r+1}\\ &= \sup_{r\geqslant 0} \frac{\widehat{\Gamma}(rM+N)-f(r)}{r+1} = \sup_{s\in[0,1]} \widehat{\Gamma}(sM+(1-s)N)-(1-s)f\big(\frac{s}{1-s}\big), \end{aligned}$$

by the change of variable $s = r/(r+1) \Leftrightarrow r = s/(1-s)$. We can therefore use the same reasoning as the one leading to Eq. (6), now based on $\widehat{G}(M-uI, N-uI) = \widehat{G}(M,N) - u$, and obtain from Eq. (17) the unconstrained variational formulation:

$$\begin{aligned} D^{\mathrm{SOS}}(A\|B) &= \sup_{M,N\in\mathbb{H}_d} \ \mathrm{tr}[AM]+\mathrm{tr}[BN] \ \text{ such that } \ \widehat{G}(M,N) \leqslant 0\\ &= \sup_{M,N\in\mathbb{H}_d} \ \mathrm{tr}[AM]+\mathrm{tr}[BN] - (\mathrm{tr}[A]+\mathrm{tr}[B])\widehat{G}(M,N). \qquad (18) \end{aligned}$$

Estimation algorithm based on Kelley's method. Given the unconstrained formulation above, we are faced with the maximization of a convex function over a vector space. Assuming that subgradients of $\widehat{G}$ can be computed, we can use Kelley's method [32], which is constructing a sequence of piecewise-affine lower-bounds on $\widehat{G}(M,N)$ obtained from subgradients and use them to find the next candidate maximizer for $M, N \in \mathbb{H}_d$. Note that we can restrict the search for $M - M_0$ and $N - N_0$ in $\mathcal{V}$ (for any well-chosen M_0, N_0 typically from the computationally more efficient relaxations derived in Section 7) since any component along $\mathcal{V}^\perp$ has no impact on the objective function. Since we do not know a priori an upper bound on $\|N - N_0\|$ and $\|M - M_0\|$, we restrict the minimization to bounded sets, and increase the bound if the boundedness constraints are active.

[5] A similar formulation can be derived for D^{OPT} using Γ instead of $\widehat{\Gamma}$.

Computing subgradients. We need to find maximizers defining the convex function $\widehat{G}(M,N)$ above, that is, finding maximizer s and Σ in $\sup_{\Sigma\in\widehat{\mathcal{K}}}\sup_{s\in[0,1]}\operatorname{tr}[\Sigma(sM+(1-s)N)]-(1-s)f\big(\frac{s}{1-s}\big)$. Given s, this is equivalent to solving one SOS maximization problem. We use the Chambolle-Pock algorithm [15] for the primal-dual formulation in Eq. (11), with a fixed maximal number of iterations. Since we end up solving many such problems for values of s which are close, we use warm starts and simple predictor-corrector steps so that the maximal number of iterations is rarely achieved.

To find an approximate joint maximizer (s,Σ), we first take a fine grid $s\in[0,1]$, and compute the value of $\widehat{\Gamma}(sM+(1-s)N)-(1-s)f\big(\frac{s}{1-s}\big)$ for each element of this grid. From the largest elements, we launch alternating optimization (alternating optimizing over s and over Σ), which is converging quickly. The number of steps can be controlled if needed, but we simply take a grid of step less than 10^{-3}.

Using hierarchies. In order to get tighter approximations to D^{OPT}, like in classical SOS optimization, we can embed the feature φ into a larger feature vector $\tilde{\varphi}=\binom{\varphi}{\varphi^+}$, where φ^+ is an additional feature map. This defines the expression $D^{\mathrm{SOS}}_{\tilde{\varphi}}(\tilde{A}\|\tilde{B})$, for moment matrices $\tilde{A}$ and $\tilde{B}$, where we add the dependence on the feature map to makes the difference explicit. We can then the consider a tighter bound by minimizing $D^{\mathrm{SOS}}_{\tilde{\varphi}}(\tilde{A}\|\tilde{B})$ with respect to $\tilde{A}$ and $\tilde{B}$ that have their upper-left blocks equal to A and B.

In this paper, we focus on the computation of upper-bounds of the partition function. The study of the approximation capabilities, when the feature vector grows, is left for future work.

7 Relaxed f-divergence based on quantum information theory

In the previous section, we relaxed Γ into $\widehat{\Gamma}$, which corresponds to using SOS relaxations for the optimization of quadratic forms. This was equivalent to replacing $\mathcal{C}$ by $\widehat{\mathcal{C}}$. This can be further relaxed into the spectral relaxation where we replace $\widehat{\Gamma}$ by the spectral relaxation $\lambda_{\max}$ (which is larger), which is equivalent to replacing $\widehat{\mathcal{C}}$ by the PSD cone $\mathbb{H}_d^+$ (which is larger).

Starting from Eq. (17), this thus leads to:

$$\sup_{M,N\in\mathbb{H}_d}\ \operatorname{tr}[MA]+\operatorname{tr}[NB]\ \text{ such that }\ \forall r\geqslant 0,\ rM+N\preccurlyeq f(r)I. \tag{19}$$

The following lemma, taken from [40, Section 9.1] and with a proof shown in Appendix C, shows that we obtain exactly the "maximal" quantum information divergence $\tilde{D}^{\mathrm{QT}}_{\max}(A\|B)$ defined in Section 3.2.

Lemma 1 *[40, Section 9.1] Assume $A,B\succcurlyeq 0$ and f is operator-convex. The maximal value of Eq. (19) is equal to* $\operatorname{tr}\big[Bf\big(B^{-1/2}AB^{-1/2}\big)\big]$ *(see proof for minimizer).*

The simplest spectral relaxation thus satisfies $\tilde{D}^{\mathrm{QT}}_{\max}(A\|B)\leqslant D^{\mathrm{SOS}}(A\|B)$. Note that this is already an improvement on [4], which considers the larger "standard" quantum divergence.

In the relaxation $\tilde{D}^{\mathrm{QT}}_{\max}$, we keep the joint convexity, but we lose the partial invariance by invertible linear transform that both D^{OPT} and D^{SOS} had. This can be remedied by defining D^{QT} as the largest possible value of $\tilde{D}^{\mathrm{QT}}_{\max}$ once these invariances are taken into account. This link will be explored in Section 7.1, and for now we define D^{QT} as follows:

$$D^{\mathrm{QT}}(A\|B)=\sup_{M,N,V\in\mathbb{H}_d}\operatorname{tr}[MA]+\operatorname{tr}[NB]\text{ such that }\forall r\geqslant 0, rM+N\preccurlyeq f(r)V,\ V\succcurlyeq 0,\ V-I\in\mathcal{V}^{\perp}, \tag{20}$$

where the only difference with Eq. (19) is the introduction of the matrix V which can be different from I. Comparing to Eq. (17), we get that it is between $\tilde{D}^{\mathrm{QT}}_{\max}(A\|B)$ and $D^{\mathrm{SOS}}(A\|B)$, because the constraint $rM + N \preccurlyeq f(r)V$ in Eq. (20) implies $f(r)I - rM - N \in \mathbb{H}_d^+ + \mathcal{V}^\perp$ (and we thus get a feasible point for Eq. (17)).

It turns out that we can also optimize in closed form with respect to M, N leading to a similar expression for $D^{\mathrm{QT}}(A\|B)$ in Eq. (20). The following lemma is proved in Appendix C, and is a slight modification of Lemma 1.

Lemma 2 *Assume $A, B \succcurlyeq 0$ and f is operator-convex. The maximal value of Eq. (20) is equal to:*

$$D^{\mathrm{QT}}(A\|B) = \sup_{V \succcurlyeq 0,\ V - I \in \mathcal{V}^\perp} \operatorname{tr}\left[V B^{1/2} f\big(B^{-1/2} A B^{-1/2}\big) B^{1/2}\right]. \tag{21}$$

We can thus obtain the value of $D^{\mathrm{QT}}(A\|B)$ by a convex optimization problem that does not involve any spectral function, is finite-dimensional, and can be solved using any semi-definite programming solver.

Properties of $D^{\mathrm{QT}}(A\|B)$. We start by deriving an alternative formulation. Defining the matrix $Q = B^{1/2} f\big(B^{-1/2} A B^{-1/2}\big) B^{1/2} \succcurlyeq 0$, we get:

$$\begin{aligned}
D^{\mathrm{QT}}(A\|B) &= \max_V \ \operatorname{tr}[QV] \text{ such that } V \succcurlyeq 0 \text{ and } \forall x \in \mathcal{X},\ \varphi(x)^* V \varphi(x) = 1 \\
&= \max_V \ \operatorname{tr}[QV] \text{ such that } V \succcurlyeq 0 \text{ and } V - I \in \mathcal{V}^\perp \\
&= \min_{\Sigma \in \mathcal{V}} \ \operatorname{tr}[\Sigma] \text{ such that } \Sigma \succcurlyeq Q, \text{ by Lagrange duality,} \\
&= \min_{\Sigma \in \mathcal{V}} \ \operatorname{tr}[\Sigma] \text{ such that } \Sigma \succcurlyeq B^{1/2} f\big(B^{-1/2} A B^{-1/2}\big) B^{1/2}.
\end{aligned}$$

Note that like D^{OPT} and D^{SOS}, D^{QT} is jointly convex, and $D^{\mathrm{QT}}(\Sigma_p\|\Sigma_q)$ is invariant by invertible linear transform of φ (which would not be the case without optimizing with respect to V) (see Section 7.1). Moreover, $D^{\mathrm{QT}}(A\|B)$ is finite only if A and B are positive semi-definite (as opposed to be also in $\mathcal{V}$ for $D^{\mathrm{SOS}}(A\|B)$).

In terms of algorithms to approximate $D^{\mathrm{QT}}(A\|B)$, we can use interior-point methods to solve Eq. (21) as this is not a computational bottleneck.

Unconstrained formulation. In order to compute the lower-bound on the f-divergence D^{QT} the formulations above are the most appropriate. However, when dealing with variational inference in Section 9.2, we will the need the following unconstrained formulation, akin to Eq. (18), and obtained by replacing $\widehat{\Gamma}$ by $\lambda_{\max}$:

$$\begin{aligned}
D^{\mathrm{QT}}(A\|B) \quad = \quad & \sup_{M,N,V \in \mathbb{H}_d} \operatorname{tr}[MA] + \operatorname{tr}[NB] - (\operatorname{tr}[A] + \operatorname{tr}[B]) \sup_{r \geqslant 0} \frac{\lambda_{\max}(rM + N - f(r)V)}{r+1} \\
& \text{such that } \ V \succcurlyeq 0,\ V - I \in \mathcal{V}^\perp.
\end{aligned} \tag{22}$$

7.1 Link with quantum information theory and metric learning

Given a fixed feature map $\varphi : \mathcal{X} \to \mathbb{C}^d$, we consider an invertible matrix $T \in \mathbb{C}^{d\times d}$, such that $V = T^*T$ is such that $\varphi(x)^*V\varphi(x) = 1$ for all $x \in \mathcal{X}$. This corresponds to *invertible* matrices V such that $V \succcurlyeq 0$ and $V - I \in \mathcal{V}^\perp$. Writing $\tilde{A} = TAT^*$, and $\tilde{B} = TBT^*$, we get, using the *maximal divergence* defined in Section 3.2:

$$\tilde{D}^{\mathrm{QT}}_{\max}(TAT^*\|TBT^*) = \tilde{D}^{\mathrm{QT}}_{\max}(\tilde{A}\|\tilde{B}) = \mathrm{tr}(\tilde{B}^{1/2} f(\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2})\tilde{B}^{1/2}).$$

Since $\tilde{B}^{1/2}$ and $TB^{1/2}$ are two square roots of $\tilde{B}$, there exists a unitary matrix R such that $\tilde{B}^{1/2} = TB^{1/2}R$. We then get $\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2} = R^*B^{-1/2}AB^{-1/2}R$, leading to $f(\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2}) = R^*f(B^{-1/2}AB^{-1/2})R$, which in turn leads to

$$\begin{aligned}\tilde{D}^{\mathrm{QT}}_{\max}(\tilde{A}\|\tilde{B}) &= \mathrm{tr}(\tilde{B}^{1/2} f(\tilde{B}^{-1/2}\tilde{A}\tilde{B}^{-1/2})\tilde{B}^{1/2}) = \mathrm{tr}\left[TB^{1/2}f(B^{-1/2}AB^{-1/2})B^{1/2}T^*\right] \\ &= \mathrm{tr}\left[T^*TB^{1/2}f(B^{-1/2}AB^{-1/2})B^{1/2}\right] = \mathrm{tr}\left[VB^{1/2}f(B^{-1/2}AB^{-1/2})B^{1/2}\right],\end{aligned}$$

which is exactly the objective function maximized to define $D^{\mathrm{QT}}(A\|B)$ in Eq. (21). Thus, by optimizing over all matrices T, and thus with respect to all matrices V, we have:

$$D^{\mathrm{QT}}(A\|B) = \sup_{T\in\mathbb{C}^{d\times d},\ \forall x\in\mathcal{X}, \|T\varphi(x)\|=1} \tilde{D}^{\mathrm{QT}}_{\max}(TAT^*\|TBT^*).$$

As observed in [4] for $\tilde{D}^{QT}_{\mathrm{standard}}$, we have $\tilde{D}^{\mathrm{QT}}_{\max}(TAT^*\|TBT^*) \leqslant D(p\|q)$ for any $p, q \in \mathcal{M}_+(\mathcal{X})$ such that $\Sigma_p = A$ and $\Sigma_q = B$, as soon as $T^*T \in I + \mathcal{V}^\perp$. Our new relaxation is thus equivalent to estimating the best feature vector in a linear model defined by φ. A simple consequence is that, while $\tilde{D}^{\mathrm{QT}}_{\max}$ is not invariant by invertible linear transforms, D^{QT} is (just like D^{OPT} and D^{SOS}). Note finally, that the use of $\tilde{D}^{QT}_{\mathrm{standard}}$ instead of $\tilde{D}^{QT}_{\max}$, as done in [4] for the particular case of the KL divergence, leads to a weaker relaxation and a more complex optimization problem in V (concave maximization instead of linear maximization).

8 Variational inference with f-divergences

Now that we have explored convex lower-bounds for the f-divergence, we can explore the Fenchel-dual equivalent, that is, convex upper-bounds on f-partition functions, and the natural link with variational inference. We start with a review of probabilistic concepts without any feature map in Section 8.1 and then with a feature map in Section 8.2, before exploring relaxations in subsequent sections.

8.1 f-partition function

Given a function $h : \mathcal{X} \to \mathbb{R}$, and q a fixed positive measure not necessarily summing to one (that is, in $\mathcal{M}_+(\mathcal{X})$), we can define the "$f$-partition function" as the Fenchel dual with respect to a probability measure p of $D(p\|q)$, that is:

$$c_q(h) = \sup_{p\in\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} h(x)dp(x) - D(p\|q). \tag{23}$$

Using the variational formulation in Eq. (3), we can optimize with respect to w to obtain

$$c_q(h) = \sup_{p\in\mathcal{M}(\mathcal{X})} \inf_{v:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} h(x)dp(x) - \int_{\mathcal{X}} v(x)dp(x) + \int_{\mathcal{X}} f^*(v(x))dq(x) \text{ such that } \int_{\mathcal{X}} dp(x) = 1,$$

where it is sufficient to consider $p \in \mathcal{M}(\mathcal{X})$ rather than $p \in \mathcal{M}_+(\mathcal{X})$ because the infimum is only finite for non-negative measures. We can then introduce a Lagrange multiplier ρ for the single equality constraint $\int_{\mathcal{X}} dp(x) = 1$, and using strong duality to swap infimum and supremum [39, Section 8.6], we get (see [50] for similar derivations in the context of PAC-Bayes analysis):

$$\begin{aligned} c_q(h) &= \sup_{p \in \mathcal{M}(\mathcal{X})} \inf_{\rho \in \mathbb{R},\ v : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} h(x) dp(x) - \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} f^*(v(x)) dq(x) - \rho \Big(\int_{\mathcal{X}} dp(x) - 1 \Big) \\ &= \inf_{\rho \in \mathbb{R},\ v : \mathcal{X} \to \mathbb{R}} \sup_{p \in \mathcal{M}(\mathcal{X})} \int_{\mathcal{X}} h(x) dp(x) - \int_{\mathcal{X}} v(x) dp(x) + \int_{\mathcal{X}} f^*(v(x)) dq(x) - \rho \Big(\int_{\mathcal{X}} dp(x) - 1 \Big). \end{aligned}$$

We can then optimize in closed form with respect to p, which leads to $v = h - \rho$, and then with respect to v, which leads to

$$c_q(h) = \inf_{\rho \in \mathbb{R}} \ \rho + \int_{\mathcal{X}} f^*(h(x) - \rho) dq(x). \tag{24}$$

The optimality condition for ρ is that $\int_{\mathcal{X}} (f^*)'(h(x) - \rho) dq(x) = 1$. Moreover, the set of functions $h : \mathcal{X} \to \mathbb{R}$ such that $c_q(h)$ is finite is a convex set.

This means that we can define a probability distribution with density $(f^*)'(h(x) - \rho)$ with respect to q, which we denote $p(x|h)$. For $f(t) = t \log t - t + 1$, where $(f^*)'(u) = e^u$, we recover classical exponential families (see [62, 43] and references therein), and $c_q(h) = \log \big(\int_{\mathcal{X}} e^{h(x)} dq(x) \big) + 1 - \int_{\mathcal{X}} q(x)$, which is the traditional log-partition function, and Eq. (23) is often referred to as the Donsker-Varadhan inequality [20].

Variational formulation. We can use the variational formulation in Eq. (3), but without first optimizing w out, still using strong duality, and adding a Lagrange multiplier ρ for the constraint $\int_{\mathcal{X}} dp(x) = 1$:

$$\begin{aligned} c_q(h) &= \sup_{p \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} h(x) dp(x) - D(p \| q) \\ &= \sup_{p \in \mathcal{M}(\mathcal{X})} \inf_{v, w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} h(x) dp(x) - \int_{\mathcal{X}} v(x) dp(x) - \int_{\mathcal{X}} w(x) dq(x) \text{ such that } \int_{\mathcal{X}} dp(x) = 1 \\ &\qquad\qquad \text{such that } \forall x \in \mathcal{X}, \forall r \geqslant 0,\ r v(x) + w(x) \leqslant f(r) \\ &= \sup_{p \in \mathcal{M}(\mathcal{X})} \inf_{\rho,\ v, w : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} h(x) dp(x) - \int_{\mathcal{X}} v(x) dp(x) - \int_{\mathcal{X}} w(x) dq(x) - \rho \Big(\int_{\mathcal{X}} dp(x) - 1 \Big) \\ &\qquad\qquad \text{such that } \forall x \in \mathcal{X}, \forall r \geqslant 0,\ r v(x) + w(x) \leqslant f(r) \\ &= \inf_{\rho,\ w : \mathcal{X} \to \mathbb{R}} \ \rho - \int_{\mathcal{X}} w(x) dq(x) \ \text{ such that } \forall x \in \mathcal{X}, \forall r \geqslant 0,\ r h(x) + w(x) \leqslant f(r) + \rho r, \end{aligned} \tag{25}$$

since the optimization with respect to $p \in \mathcal{M}(\mathcal{X})$ leads to $v = h$. Like in earlier variational formulations, we can replace the constrained formulation by an unconstrained one using the function F defined in Eq. (4) in Section 2.1. By replacing w by $w - \rho$, and optimizing with respect to ρ in Eq. (25), we get:

$$\begin{aligned} c_q(h) &= \inf_{w : \mathcal{X} \to \mathbb{R}} \Big(1 + \int_{\mathcal{X}} dq(x) \Big) \sup_{x \in \mathcal{X}} \sup_{r \geqslant 0} \frac{r h(x) + w(x) - f(r)}{r + 1} - \int_{\mathcal{X}} w(x) dq(x) \\ &= \inf_{w : \mathcal{X} \to \mathbb{R}} \Big(1 + \int_{\mathcal{X}} dq(x) \Big) \sup_{x \in \mathcal{X}} F(h(x), w(x)) - \int_{\mathcal{X}} w(x) dq(x), \end{aligned} \tag{26}$$

where the optimal value of ρ is obtained as $\rho = \sup_{x \in \mathcal{X}} \sup_{r \geqslant 0} \frac{r h(x) + w(x) - f(r)}{r+1} = \sup_{x \in \mathcal{X}} F(h(x), w(x))$. In variational inference, we will need to obtain the probability distribution p in Eq. (23) from w. Here, it simply has density $(f^*)'(h(x) - \rho)$ with respect to q.

8.2 Variational inference with feature maps

In the previous section, we have considered probability densities and partition functions for potentials that were allowed to take any functional form in $x \in \mathcal{X}$. We now specialize to potentials that are quadratic forms in $\varphi(x)$.

In this section, we thus extend the notion of exponential families, which is classical for $f(t) = t \log t - t + 1$, to all f-divergences. These are also called "q-exponential families" for α-divergences [2].

f-family of probability distributions. Following Section 8.1, given the matrix feature map $x \mapsto \varphi(x)\varphi(x)^* \in \mathbb{H}_d$, we define the distribution $p(\cdot|H)$ with density with respect to q of the form $(f^*)'\big(\varphi(x)^* H \varphi(x) - \rho\big)$ for a certain Hermitian matrix $H \in \mathbb{H}_d$, and with the normalizing constant $\rho = \rho(H) \in \mathbb{R}$ that makes the density sum to one. We can then define

$$C_q(H) = \sup_{p \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \varphi(x)^* H \varphi(x) dp(x) - D(p\|q) = c_q\big(\varphi(\cdot)^* H \varphi(\cdot)\big), \tag{27}$$

with the optimal probability distribution p (which is unique since we have assumed that f is strictly convex) exactly being the density $p(\cdot|H)$ defined above. The set of $H \in \mathbb{H}_d$ such that $C_q(H)$ is finite is convex.

From the representation in Eq. (27) as a maximum of affine functions, we obtain that the gradient $C_q'(H)$ is equal to $\int_{\mathcal{X}} p(x|H)\varphi(x)\varphi(x)^* dq(x)$ as $p(\cdot|H)$ is the maximizer in Eq. (27), that is, $C_q'(H)$ is exactly the expectation of $\varphi(x)\varphi(x)^*$ under $p(\cdot|H)$. Thus, a classical task in variational inference is to compute $C_q'(H)$ [62]. For example, for the traditional Ising model, where $\mathcal{X} = \{-1, 1\}^n$ and $\varphi(x) = (x^\top, 1)^\top$, $C_q'(H)$ is a matrix composed of the expectations of $xx^\top$ and x.

We can then define the Fenchel conjugate C_q^* of C_q as:

$$C_q^*(\Sigma) = \sup_{H \in \mathbb{H}_d} \operatorname{tr}[H\Sigma] - C_q(H).$$

The domain of C_q^* is then exactly the set of attainable moments (denoted $\mathcal{K}$ in Section 4), and the moment $\Sigma(H) = C_q'(H)$ is exactly the maximizer in

$$\sup_{\Sigma \in \mathbb{H}_d} \operatorname{tr}[H\Sigma] - C_q^*(\Sigma).$$

Note that in the future approximations of C_q^* or C_q, there is both an approximation of the value *and* potentially of the domain.

Estimation. Given some data $x_1, \dots, x_n \in \mathcal{X}$, we can form the empirical moment $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^*$, and estimate $H \in \mathbb{H}_d$ by minimizing $D(p\|q)$ such that $\Sigma_p = \widehat{\Sigma}$. For $f(t) = t \log t - t + 1$, this is exactly maximum entropy estimation, which is classsicaly equivalent to finding the exponential family distributions with feature $x \mapsto \varphi(x)\varphi^*(x)$ and matching moment. This happens to be true for all f-divergences, that is, the optimal distribution p is exactly $p = p(\cdot|H)$ for H maximizing $\operatorname{tr}[H\widehat{\Sigma}] - C_q(H)$, and with matching moments. Note however that the formulation as the minimum (right) Kullback-Leibler divergence does not readily generalize beyond the Shannon entropy.

9 Relaxed f-partition function

Given the sequence of lower bounds on the f-divergence $D^{\mathrm{OPT}} \geqslant D^{\mathrm{SOS}} \geqslant D^{\mathrm{QT}} \geqslant \widetilde{D}^{\mathrm{QT}}_{\max} \geqslant \widetilde{D}^{\mathrm{QT}}_{\mathrm{standard}}$, we get a sequence of upper-bounds for $C_q(H)$ defined in Eq. (27) for any Hermitian matrix $H \in \mathbb{H}_d$. Since q (which is now assumed to sum to one) only appears through $B = \Sigma_q$, we define our upper-bound as the maximal potential $C_q(H)$ for all distributions q such that $\Sigma_q = B$, that is,

$$C_B^{\mathrm{OPT}}(H) = \sup_{q \in \mathcal{P}(\mathcal{X})} C_q(H) \text{ such that } \Sigma_q = B.$$

We can now use the definition of $C_q(H)$ from Eq. (27) to get:

$$\begin{aligned} C_B^{\mathrm{OPT}}(H) &= \sup_{p,q \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \varphi(x)^* H \varphi(x) dp(x) - D(p\|q) \text{ such that } \Sigma_q = B \\ &= \sup_{A \in \mathcal{C},\ \mathrm{tr}[A]=1} \ \sup_{p,q \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \varphi(x)^* H \varphi(x) dp(x) - D(p\|q) \text{ such that } \Sigma_q = B \text{ and } \Sigma_p = A \\ &= \sup_{A \in \mathbb{H}_d} \ \mathrm{tr}[AH] - D^{\mathrm{OPT}}(A\|B) \ \text{ such that } \ \mathrm{tr}[A] = 1, \text{ by definition of } D^{\mathrm{OPT}}. \end{aligned} \tag{28}$$

Note that the constraints that $A \in \mathcal{V}$ and $A \succcurlyeq 0$ are implied by the finiteness of $D^{\mathrm{OPT}}(A\|B)$ (this will not be the case for the other relaxations).

We use the variational representation of $D^{\mathrm{OPT}}(A\|B)$ in Eq. (16), to get, using the Lagrange multiplier ρ for the constraint $\mathrm{tr}[A] = 1$:

$$\begin{aligned} & C_B^{\mathrm{OPT}}(H) \\ = & \sup_{A \in \mathbb{H}_d} \ \inf_{M,N \in \mathbb{H}_d} \ \mathrm{tr}[AH] - \mathrm{tr}[MA] - \mathrm{tr}[NB] \ \text{ such that } \ \forall r \geqslant 0,\ \Gamma(rM + N) \leqslant f(r) \text{ and } \mathrm{tr}[A] = 1 \\ = & \sup_{A \in \mathbb{H}_d} \ \inf_{M,N \in \mathbb{H}_d,\ \rho \in \mathbb{R}} \mathrm{tr}[AH] - \mathrm{tr}[MA] - \mathrm{tr}[NB] - \rho\big(\mathrm{tr}[A] - 1\big) \ \text{ such that } \ \forall r \geqslant 0,\ \Gamma(rM + N) \leqslant f(r). \end{aligned}$$

The convex set $\{(M,N) \in \mathbb{H}_d \times \mathbb{H}_d,\ \forall r \geqslant 0, \Gamma(rM + N) \leqslant f(r)\}$ has non empty interior since $(0, -I)$ is in the interior; thus strong duality holds [39, Section 8.6], and we can swap infimum and supremum. Taking the supremum with respect to A leads to the equality constraint $H - M - \rho I = 0$, and thus

$$C_B^{\mathrm{OPT}}(H) \ = \inf_{N \in \mathbb{H}_d,\ \rho \in \mathbb{R}} \ \rho - \mathrm{tr}[NB] \text{ such that } \forall r \geqslant 0,\ \Gamma(rH + N) \leqslant f(r) + \rho r. \tag{29}$$

By replacing N by $\bar{N} - \rho I$, and then optimizing in closed form with respect to ρ, we get:

$$\begin{aligned} C_B^{\mathrm{OPT}}(H) &= \inf_{\bar{N} \in \mathbb{H}_d,\ \rho \in \mathbb{R}} \ 2\rho - \mathrm{tr}[\bar{N}B] \text{ such that } \forall r \geqslant 0,\ \Gamma(rH + \bar{N}) \leqslant f(r) + \rho(r+1) \\ &= \inf_{\bar{N} \in \mathbb{H}_d} \ 2 \sup_{r \geqslant 0} \frac{\Gamma(rH + \bar{N}) - f(r)}{r+1} - \mathrm{tr}[\bar{N}B] = \inf_{N \in \mathbb{H}_d} 2G(H, \bar{N}) - \mathrm{tr}[\bar{N}B]. \end{aligned} \tag{30}$$

This corresponds to the formulation in Eq. (25). This can now be solved with Kelley's method, like described in Section 6.

Recovering the optimal moment matrix $\Sigma_p = A$. Given the solution $\bar{N}$ of Eq. (30), we get the optimal ρ and N of Eq. (29) as $\rho = \sup_{r \geqslant 0} \frac{\Gamma(rH+\bar{N})-f(r)}{r+1}$ and $N = \bar{N} - \rho I$, and $M = H - \rho I$. Optimality conditions for Eq. (30) lead to $B = \int_{\mathcal{X}\times\mathbb{R}_+} \frac{1}{r+1}\varphi(x)\varphi(x)^* d\nu(x,r)$, where ν is a probability distribution supported on the maximizers of $\frac{\varphi(x)^*(rM+N)\varphi(x)-f(r)}{r+1}$. Since $\mathrm{tr}[B] = 1$, we have $\int_{\mathcal{X}\times\mathbb{R}_+} \frac{1}{r+1} d\nu(x,r) = 1$. The optimal A is $A = \int_{\mathcal{X}\times\mathbb{R}_+} \frac{r}{r+1}\varphi(x)\varphi(x)^* d\nu(x,r)$. We then have: $\mathrm{tr}[AM]+\mathrm{tr}[NB] = \int_{\mathcal{X}\times\mathbb{R}_+} \frac{f(r)}{r+1} d\nu(x,r) = D^{\mathrm{OPT}}(A\|B)$.

Tightness. In this paper, we focus on the computation of upper-bounds of the partition function. The study of the approximation capabilities when the feature vector grows is left for future work. In particular, it would be interesting to compare to other convex upper-bounds on the log partition functions such as the "tree-reweighted representation" framework [62].

We now consider computable relaxations, first based on SOS in Section 9.1, then on quantum information divergences in Section 9.2.

9.1 Sum-of-squares relaxation

We get the SOS relaxation where Γ is replaced by $\widehat{\Gamma}$ in Eq. (29) and Eq. (30):

$$\begin{aligned} C_B^{\mathrm{SOS}}(H) &= \inf_{N\in\mathbb{H}_d,\ \rho\in\mathbb{R}} \ \rho - \mathrm{tr}[NB] \text{ such that } \forall r \geqslant 0,\ (f(r)+\rho r)I - rH - N \in \widehat{\mathfrak{C}}^* \\ &= \inf_{\bar{N}\in\mathbb{H}_d} \ 2\sup_{r\geqslant 0} \frac{\widehat{\Gamma}(rH+\bar{N})-f(r)}{r+1} - \mathrm{tr}[\bar{N}B] = \inf_{\bar{N}\in\mathbb{H}_d} 2\widehat{G}(H,\bar{N}) - \mathrm{tr}[\bar{N}B]. \end{aligned}$$

This is now approximable in polynomial time and is an upper bound on $C^{\mathrm{OPT}}(H)$.

To obtain the optimal A from an optimal N, optimality conditions lead to a probability measure ν on $\widehat{\mathcal{K}} \times \mathbb{R}_+$, corresponding to the maximizers of $\sup_{r\geqslant 0}\sup_{\Sigma\in\widehat{\mathcal{K}}} \frac{\mathrm{tr}[\Sigma(rH+\bar{N})]-f(r)}{r+1}$ and $A = \int_{\mathcal{X}\times\mathbb{R}_+} \frac{r}{r+1}\Sigma d\nu(\Sigma,r)$.

Algorithms. We can use the same technique as for computing $D^{\mathrm{SOS}}(A\|B)$ and use Kelley's method. It can be initialized by considering the spectral relaxation detailed below.

9.2 Quantum relaxation

We now consider the quantum relaxation instead of the sum-of-squares relaxation, for $H \in \mathbb{H}_d$ (the constraint that $A \succcurlyeq 0$ is automatically satisfied but not the one that $A \in \mathcal{V}$), using convex duality:

$$C_B^{\mathrm{QT}}(H) = \sup_{A\in\mathcal{V}} \ \mathrm{tr}[AH] - D^{\mathrm{QT}}(A\|B) \ \text{ such that } \ \mathrm{tr}[A] = 1.$$

In order to derive estimation algorithms, we also propose a formulation based on the unconstrained formulation of D^{QT} from Eq. (22), and introducing a Lagrange multipler $Y \in \mathcal{V}^\perp$ for the constraint $A \in \mathcal{V}$, as (assuming $\mathrm{tr}[B] = 1$):

$$C_B^{\mathrm{QT}}(H) = \inf_{\bar{N},V\in\mathbb{H}_d, Y\in\mathcal{V}^\perp} -\mathrm{tr}[\bar{N}B] + 2\sup_{r\geqslant 0} \frac{\lambda_{\max}(rH+rY+\bar{N}-f(r)V)}{r+1} \text{ such that } \ V \succcurlyeq 0,\ V - I \in \mathcal{V}^\perp. \tag{31}$$

We can recover the optimal A (which may not be in $\mathcal{V}$) like for the other relaxations.

Spectral relaxation. In order to obtain approximate closed-form expressions for initialization, we could consider the spectral relaxation with $V = I$, and A not constrained to be in $\mathcal{V}$, leading to:

$$\sup_{A \in \mathbb{H}_d^+} \operatorname{tr}[AH] - \operatorname{tr}\left[Bf(B^{-1/2}AB^{-1/2})\right] \text{ such that } \operatorname{tr}[A] = 1,$$

which cannot be solved in closed form in general. We can also consider the standard quantum relaxation, which does not lead to a closed form expression, except for the function $f(t) = t \log t - t + 1$, where we need to solve

$$\sup_{A \in \mathbb{H}_d^+} \operatorname{tr}[AH] - \operatorname{tr}[A \log A - A \log B - A + B] \text{ such that } \operatorname{tr}[A] = 1, \tag{32}$$

which is equal to $\log \operatorname{tr} \exp(H + \log B)$, with $A = \frac{\exp(H + \log B)}{\operatorname{tr} \exp(H + \log B)}$. The resulting function of B is concave, its gradient is an initializer for $-N$, and it can be computed from the Jacobians of the exponential and logarithm maps. The last expression is not invariant to the addition of an element of $\mathcal{V}^\perp$ to H (while it should). Following [6, Appendix B], we can make it invariant by projecting H onto $\mathcal{V}$.

10 Experiments

In this section,[6] we illustrate our various relaxations and algorithms presented in earlier sections. We illustrate our results with the function $f(t) = t \log t - t + 1$, and thus the estimation of relative Shannon entropies and log-partition functions. We focus on two particular examples, $[0,1]^n$ with trigonometric polynomials, and $\{-1,1\}^n$ with regular polynomials, as they correspond to generic multivariate continuous and discrete situations. In the discrete case of $\{-1,1\}^n$, exact computations have an exponential complexity in n, which is unavoidable [61]. In contrast, the approximation methods provided in this paper have polynomial time in n, though with an exponent that grows fast with the maximal cardinality size in the feature φ. The continuous case creates extra difficulties depending on the order of polynomials that we consider (with the need to use high-order quadrature formulas [24] to compute quantities exactly).

We first consider in Section 10.1 computing relative entropies and compare there all relaxations, from the more costly ones based on sums-of-squares (SOS) to quantum-based ones (QT) which rely on faster spectral computations. Given that the quantum-based ones have a very similar performance a much lower computational cost, we only consider these in Section 10.2 where we compute log-partition functions.

10.1 Computing relative entropies

We consider three classical situations, trigonometric polynomials on $[0,1]$, where the SOS relaxation is tight (and thus $D^{\mathrm{OPT}} = D^{\mathrm{SOS}}$), as well as trigonometric polynomials on $[0,1]^n$, and polynomials on $\{-1,1\}^n$.

Trigonometric polynomials on $\mathcal{X} = [0,1]$. We consider q the uniform distribution on $[0,1]$ (with density 1 with respect to the Lebesgue measure), and p with density $\frac{8}{\pi}\sqrt{x(1-x)}$. We have the following

[6] Matlab code to reproduce all experiments can be downloaded from www.di.ens.fr/~fbach/fdiv_quantum_var.zip.

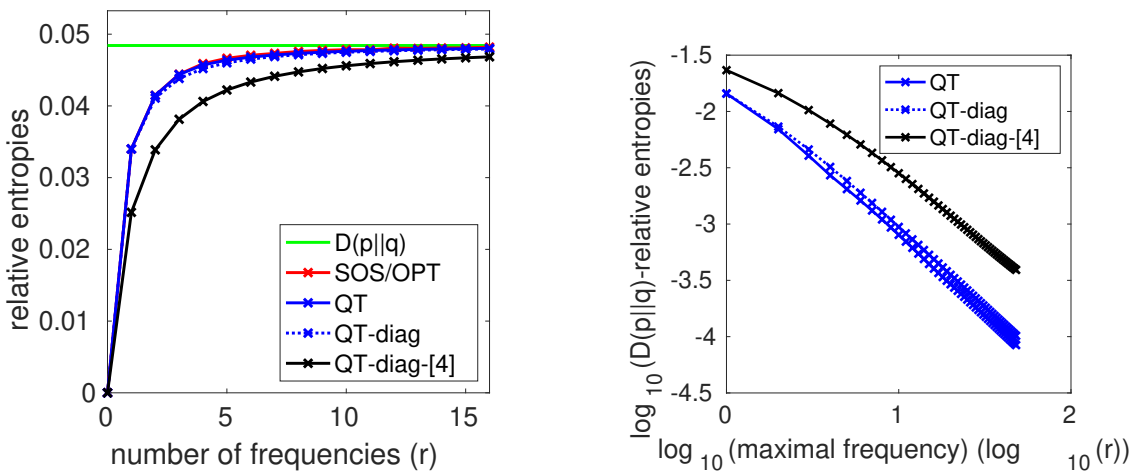


Figure 1: Comparison of relative entropy estimates for several numbers of frequencies, showing the effect of varying r. Left: $D(p\|q)$ is the exact value, with moment-based approximations OPT/SOS (here equal), and the ones based on quantum information divergences, with full metric learning (QT) or diagonal metric learning (QT-diag), with also the use of the weaker quantum divergences proposed by [4]. In several cases the red curve associated with OPT/SOS bounds is not visible as it coincides with the blue curve of QT bounds. Right: same experiment without OPT/SOS, with deviation to f-divergence reported in a log-log plot.

moments:

$$\begin{aligned}\int_0^1 e^{2i\pi\omega x} dq(x) &= 1 \text{ if } \omega = 0, \text{ and } 0 \text{ otherwise},\\ \int_0^1 e^{2i\pi\omega x} dp(x) &= \frac{2(-1)^\omega}{\omega\pi} J_1(\omega\pi) \text{ for } \omega \neq 0,\end{aligned}$$

where J_1 is the Bessel function of the first kind, as well as the relative entropy $D(p\|q) \approx 0.0484$, which can be approximated with high precision with quadrature formulas [24].

We consider $\omega \in \{-r, \ldots, r\}$, and compute the various bounds: OPT, SOS (which are equal), and QT (together with a version only optimized over diagonal V), and the old version of QT from [4] (where we only learn diagonal matrices V).

We see in the left plot of Figure 1 that the optimal/SOS bound is numerically identical to the full quantum bound, and close to the one with diagonal V, but with a strong improvement over the bound from [4]. For the SOS relaxations, results are obtained using Kelley's method described in Section 6.

In the right plot of Figure 1, we only compute the spectral relaxations, for significantly larger r, showing that, as r grows, we get a tighter approximation of $D(p\|q)$ for all methods.

Trigonometric polynomial on $\mathcal{X} = [0,1]^n$. We consider x_1 uniform on $[0,1]$ and $x_{i+1} = \lfloor x_i + \eta_{i+1} \rfloor$, where η_{i+1} is uniform on $[-\rho/2, \rho/2]$. When $\rho = 0$, all x_i's are equal almost surely, while when $\rho = 1$, all x_i's are independent and uniform (we use $\rho = 3/10$ in our simulations).

We can then compute the Kullback-Leibler divergence to the uniform distribution by noticing that the sequence (x_i) forms a Markov chain, so that (using classical entropy decomposition results for tree-structured

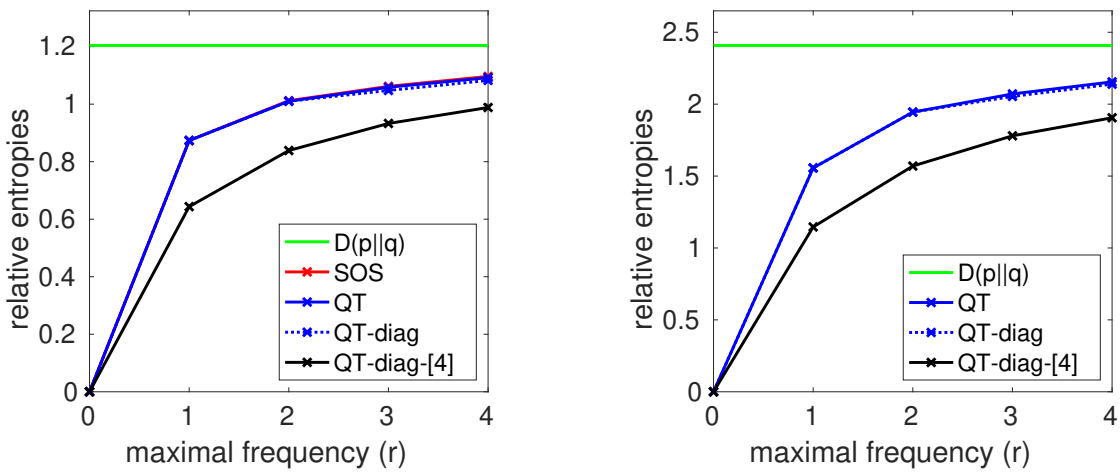


Figure 2: Comparison of relative entropy estimates for several numbers of maximal marginal frequency r, for $n = 2$ (left) and $n = 3$ (right). $D(p\|q)$ is the exact value, with moment SOS-based approximations (only plotted in the left plot, and for which we only compute the cost function of Eq. (18) with the optimal quantum solution from Eq. (20), which is close to optimal for Eq. (18)), and the ones based on quantum information divergences, with full metric learning (QT) or diagonal metric learning (QT-diag), with also the use of the weaker quantum divergences done by [4].

graphical models [62]):

$$\begin{aligned} D(p\|q) &= \int_{[0,1]^n} p(x)\log p(x)dx \\ &= \sum_{i=1}^{n-1}\int_{[0,1]^2} p(x_i,x_{i+1})\log p(x_i,x_{i+1})dx_i dx_{i+1} - \sum_{i=2}^{n-1}\int_{[0,1]} p(x_i)\log p(x_i)dx_i = (n-1)\log\frac{1}{\rho}. \end{aligned}$$

We can also get all Fourier moments by introducing the $n\times(n-1)$ $\{0,1\}$-valued matrix M such that $x_i = (M\eta)_i + x_1$ for $i\in\{2,\dots,n\}$. We then have

$$\mathbb{E}[e^{2i\pi\omega^\top x}] = 1_{\omega^\top 1=0}\cdot\prod_{k=1}^{n-1}\frac{\sin\left[(M^\top\omega)_k\pi\rho\right]}{(M^\top\omega)_k\pi\rho}.$$

In order to estimate entropies, we consider $\|\omega\|_\infty \leqslant r$. See Figure 2, where we can draw similar conclusions as for $n = 1$.

Polynomials on $\mathcal{X} = \{-1,1\}^n$ We consider the task of estimating entropies from moments on a simple example, where we consider x_1 uniform on $\{-1,1\}$ and $x_{i+1} = x_i\eta_{i+1}$, where $\eta_{i+1}\in\{-1,1\}$ is independent and equal to 1 with probability $1-\rho/2$, and -1 otherwise. When $\rho = 0$, all x_i's are equal almost surely, while when $\rho = 1$, all x_i's are independent and uniform (we use $\rho = 1/2$ in our simulations).

We can then compute the Kullback-Leibler divergence to the uniform distribution in the same way as for data in $[-1,1]^n$, leading to $D(p\|q) = (n-1)\left[(1-\frac{\rho}{2})\log(2-\rho)+\frac{\rho}{2}\log\rho\right]$. We can also get all Fourier moments as $\mathbb{E}\left[\prod_{i\in A}x_i\right] = (1-(-1)^{|A|})\prod_{i=1}^{n-1}(1-\rho)^{1-(-1)^{(M^\top 1_A)_i}}$. In order to estimate entropies, we consider subsets

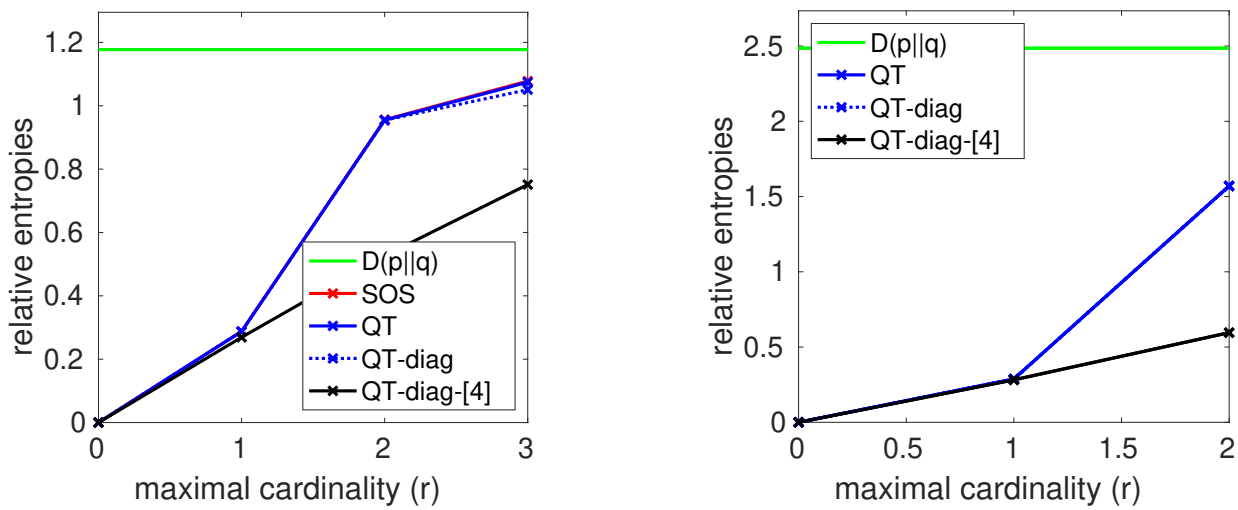


Figure 3: Comparison of relative entropy estimates for several numbers of frequencies for all subsets of cardinality less than r, for $n = 10$ and $n = 20$. $D(p\|q)$ is the exact value, with moment SOS-based approximations, and the ones based on quantum information divergences, with full metric learning (QT) or diagonal metric learning (QT-diag), with also the use of the weaker quantum divergences done by [4].

of cardinality less then r. See Figure 3, for $n = 10$ and $n = 20$, where we can draw similar conclusions as for $n = 1$.

10.2 Computing log-partition functions

We now compare algorithms to upper-bound log-partition functions, by only focusing on the more efficient quantum relaxations. We do so for trigonometric polynomials on $[0, 1]$.

Log-partition functions on $\mathcal{X} = [0, 1]$. We consider $h(x) = \cos(4\pi x)$, with $\log \int_0^1 e^{\cos(4\pi x)} dx = \log I_0(1) \approx 0.2359$, where we use the same feature map $\varphi : [0, 1] \to \mathbb{C}^{2r+1}$ as before, which enables us to write $h(x) = \varphi(x)^* H \varphi(x)$ for some Hermitian matrix H. The matrix H is not unique, and, following [6, Appendix B], we consider the spectral relaxation from Eq. (32), with H orthogonally projected on the set $\mathcal{V}$ of Toeplitz matrices. This spectral relaxation was considered in [4] but in a positive definite kernel context where the projection onto Toeplitz matrices is not applicable. Such a projection is crucial to obtain a meaningful result.

We then compute the approximation C_B^{QT} by using Kelley's method to solve Eq. (31), and we also report results without the optimization with respect to V, with an almost identical curve (showing that the benefits of metric learning are here marginal). While for small r, the new quantum relaxation improves over the bound adapted from [4], it does not for larger r.

11 Conclusion

In this paper, we have proposed to combine tools from information theory, both classical such as f-divergences, and more recent, such as quantum information divergences, with sum-of-squares optimiza-

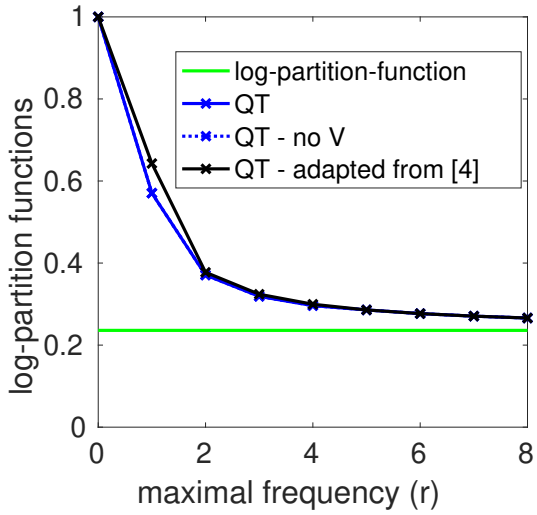


Figure 4: Comparison of log-partition function estimates for several numbers of frequencies for $n = 1$. See text for details.

D^{OPT}	Section 5	Not computable
D^{SOS}	Section 6	Computable with SOS
D^{QT}	Section 7	Computable with spectral method + single SDP
$\widetilde{D}^{\mathrm{QT}}_{\mathrm{max}}$	Section 7	Computable with spectral method
$\widetilde{D}^{\mathrm{QT}}_{\mathrm{standard}}$	[4]	Computable with spectral method

Table 1: Summary of relaxations ordered by strength (function values and domains), from stronger (top) to weaker (bottom).

tion. This leads to several relaxations of f-divergences based on sum-of-squares relaxations or quantum information divergences, together with efficient estimation algorithms for the tasks of divergence estimation from moments and the computation of log-partition functions. These relaxations are summarized in Table 1. While the relaxation based on sums-of-squares (D^{SOS}) is strictly superior, it is only mildly so in our experiments compared to the one based on quantum divergences (D^{QT}), while being more costly to compute. This thus highlights the benefits of the quantum relaxation.

This quantum information relaxation takes its roots in earlier work [4] with significant improvements: (a) the use of a tighter quantum divergence (the "maximal" one rather than the "standard" one), (b) the introduction of the optimal lower bound, (c) taking into account the particular geometries of feature vectors using sum-of-square techniques that improve over spectral relaxations, and (d) the proposal of generic optimization algorithms. Several avenues are worth exploring: (a) check if the new notion of relative entropy with maximal divergence preserves properties from [4], in particular its use in probabilistic modelling and within graphical models, (b) potentially extend the positive definite kernel motivation that allows infinite-dimensional moments, along the lines of [25] which explored this connection for Renyi entropies, (c) obtain convergence rates for entropies and log-partition function estimation to go with our encouraging empirical results, (d) develop algorithms to deal with larger scale problems using approximation techniques from kernel methods [12, 55].

Acknowledgements

The author would like to thank Omar Fawzi for discussions related to quantum information divergences, Adrien Taylor and Justin Carpentier for discussions on optimization algorithms, as well as David Holzmüller for providing clarifying comments. The comments of the anonymous reviewers were greatly appreciated. We acknowledge support from the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), as well as from the European Research Council (grant SEQUOIA 724063).

A Newton method for computing F

In order to compute $F(v,w) = \sup_{r\geqslant 0} \frac{rv+w-f(r)}{r+1}$ in Section 2.1, we assume $(v,w) \in \mathbb{R}^2$ fixed and define $\varphi(r) = \frac{rv+w-f(r)}{r+1}$, which is twice differentiable, and so that, $\varphi'(r) = \frac{1}{(r+1)^2}\big[v - w - \big((r+1)f'(r) - f(r)\big)\big]$. Since the function $r \mapsto (r+1)f'(r) - f(r)$ has derivative $r \mapsto (r+1)f''(r)$, it is strictly positive. Thus the derivative φ' thus has at most one zero, and we aim to solve the equation $(r+1)f'(r) - f(r) = v - w$.

We now focus on the special case $f(t) = t\log t - t + 1$, where $\psi(r) = (r+1)f'(r) - f(r) = r - 1 + \log r$, and has full range on $\mathbb{R}_+$ so that the equation above has a unique solution. In order to compute its solution, for $v - w \geqslant 0$, we iterate Newton's method [24, Section 4.8] $r \leftarrow r + \frac{v-w-\psi(r)}{\psi'(r)} = \frac{2-\log(r)+v-w}{1+1/r}$ for 5 iterations to reach machine precision, while for $v - w \leqslant 0$ we iterate Newton's method on the logarithm of r $\log r \leftarrow \log r + \frac{v-w-\psi(r)}{r\psi'(r)} = \frac{1+v-w+(\log r-1)r}{1+r}$ for 5 iterations.

B Decomposition of operator-convex functions

We have the following particular cases from Section 2.

- α-divergences: $f(t) = \frac{1}{\alpha(\alpha-1)}\big[t^\alpha - \alpha t + (\alpha-1)\big] = \frac{1}{\alpha}\frac{\sin(\alpha-1)\pi}{(\alpha-1)\pi}(t-1)^2\int_0^{+\infty}\frac{1}{t+\lambda}\frac{\lambda^\alpha d\lambda}{(1+\lambda)^2}$ for $\alpha \in (-1,2)$. Other representations exist for $\alpha = -1$ and $\alpha = 2$ (see below), but other cases are not operator-convex.
- KL divergence ($\alpha = 1$): $f(t) = t\log t - t + 1 = \int_0^{+\infty}\frac{(t-1)^2}{t+\lambda}\frac{\lambda d\lambda}{(\lambda+1)^2}$.
- Rerverse KL divergence ($\alpha = 0$): $f(t) = -\log t + t - 1 = \int_0^{+\infty}\frac{(t-1)^2}{t+\lambda}\frac{d\lambda}{(\lambda+1)^2}$.
- Pearson χ^2 divergence ($\alpha = 2$): $f(t) = \frac{1}{2}(t-1)^2$ is operator convex.
- Reverse pearson χ^2 divergence ($\alpha = -1$): $f(t) = \frac{1}{2}\big(\frac{1}{t} + t\big) - 1 = \frac{1}{2}\frac{(t-1)^2}{t}$ is operator convex, with $d\nu(\lambda)$ proportional to a Dirac at $\lambda = 0$.
- Le Cam distance: $f(t) = \frac{(t-1)^2}{t+1}$ is operator convex with $d\nu(\lambda)$ proportional to a Dirac at $\lambda = 1$.
- Jensen-Shannon divergence: $f(t) = 2t\log\frac{2t}{t+1} + 2\log\frac{2}{t+1} = 2t\log t - 2(t+1)\log(t+1) + 2(t+1)\log 2 = 2t\log t - 4\frac{t+1}{2}\log\frac{t+1}{2}$ is operator convex, as it can be written $f(t) = 2(t-1)^2\int_0^{+\infty}\big(\frac{1}{t+\lambda} - \frac{1}{t+1+2\lambda}\big)\frac{\lambda d\lambda}{(1+\lambda)^2}$, which leads to $f(t) = 2(t-1)^2\int_0^{+\infty}\frac{1}{t+\lambda}\frac{\lambda d\lambda}{(1+\lambda)^2} - 2(t-1)^2\int_1^{+\infty}\frac{1}{t+\lambda}\frac{(\lambda-1)d\lambda}{(1+\lambda)^2}$.

C Proofs of Lemma 1 and Lemma 2

In this section, we prove Lemma 1 from Section 7, taken from [40, Section 9.1] and shown in here for completeness and the expression of the maximizers. We also prove the extension Lemma 2.

We start by the dual formulation:

$$\begin{aligned} & \sup_{M,N\in\mathbb{H}_d} \ \operatorname{tr}[MA]+\operatorname{tr}[NB] \ \text{ such that } \ \forall r\geqslant 0,\ rM+N\preccurlyeq f(r)I \\ = & \inf_{\Lambda\ \mathbb{H}_d^+-\text{valued measure on } \mathbb{R}_+} \int_0^{+\infty} f(r)\operatorname{tr}\big[d\Lambda(r)\big] \ \text{ such that } \ \int_0^{+\infty} d\Lambda(r)=B \text{ and } \int_0^{+\infty} r d\Lambda(r)=A, \end{aligned}$$

which can be solved in closed form.

Indeed, given the eigendecomposition $B^{-1/2}AB^{-1/2}=\sum_{i=1}^d \lambda_i u_i u_i^*$, we consider $\Lambda=\sum_{i=1}^d B^{1/2}u_iu_i^*B^{1/2}\delta_{\lambda_i}$, where δ_{λ_i} is the Dirac measure at λ_i, so that we get a feasible measure Λ, and an objective equal to $\sum_{i=1}^d f(\lambda_i)\operatorname{tr}\big[B^{1/2}VB^{1/2}u_iu_i^*\big] = \operatorname{tr}\big[B^{1/2}VB^{1/2}f\big(B^{-1/2}AB^{-1/2}\big)\big]$. Thus the infimum is less than $\operatorname{tr}\big[B^{1/2}VB^{1/2}f\big(B^{-1/2}AB^{-1/2}\big)\big]$.

The other direction is a direct consequence of the operator Jensen's inequality [28]: for any feasible measure Λ approached by an empirical measure $\sum_{i=1}^m M_i\delta_{r_i}$, with $M_i\succcurlyeq 0$, we have $\sum_{i=1}^n (M_i^{1/2}B^{-1/2})^*(M_i^{1/2}B^{-1/2})=I$, and thus

$$\begin{aligned} \int_{\mathbb{R}_+} f(r)d\Lambda(r) &= B^{1/2}\Big(\sum_{i=1}^m (M_i^{1/2}B^{-1/2})^* f(r_iI)(M_i^{1/2}B^{-1/2})\Big)B^{1/2} \\ &\succcurlyeq B^{1/2}f\Big(\sum_{i=1}^m (M_i^{1/2}B^{-1/2})^*(r_iI)(M_i^{1/2}B^{-1/2})\Big)B^{1/2} = B^{1/2}f\big(B^{-1/2}AB^{-1/2}\big)B^{1/2}. \end{aligned}$$

The lower bound follows by letting the number m of Diracs go to infinity to tightly approximate any feasible matrix Λ.

In order to obtain the minimizers M and N, we simply notice that they are the gradients of the function $(A,B)\mapsto \operatorname{tr}\big[Bf\big(B^{-1/2}AB^{-1/2}\big)\big]$ with respect to A and B. We thus get, using gradients of spectral functions

$$M^* = \sum_{i,j=1}^d \frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j} u_i^\top B u_j \cdot B^{-1/2}u_iu_j^*B^{-1/2},$$

with the convention that for $\lambda_i=\lambda_j$, $\frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j}=f'(\lambda_i)$. To obtain N^*, we simply consider the identity $\operatorname{tr}\big[Bf\big(B^{-1/2}AB^{-1/2}\big)\big]=\operatorname{tr}\big[Ag\big(A^{-1/2}BA^{-1/2}\big)\big]$, for $g(t)=tf(1/t)$, and use the corresponding formula.

Proof of Lemma 2. We simply need to change slightly the proof of the lemma above, by replacing $\int_0^{+\infty} f(r)\operatorname{tr}\big[d\Lambda(r)\big]$ by $\int_0^{+\infty} f(r)\operatorname{tr}\big[Vd\Lambda(r)\big]$ with all inequalities preserved because $V\succcurlyeq 0$.

D Computing integrals

A third task is related to f-divergences beyond computing the divergences themselves and the associated log-partition functions. In this section, we consider the task of computing $\int_{\mathcal{X}} f^*(h(x))dq(x)$, where q is a finite positive measure on $\mathcal{X}$, f^* is the Fenchel conjugate of f, and $h : \mathcal{X} \to \mathbb{R}$ an arbitrary function (such that the integral is finite). The difference with computing log-partition functions in Eq. (24) is minor and we thus extend only a few results from Section 8 and Section 9, most without proofs as they follow the same lines as results for f-partition functions.

For $f(t) = t \log t - t + 1$, we have $f^*(u) = e^u - 1$, and we there aim at estimating integrals of exponential functions, a classical task in probabilistic modelling (see [62, 43] and references therein), which up to a logarithm is the same as computing the log-partition function; however, they are different for other functions f.

This computational task can be classically related to f-divergences by Fenchel duality as we have:

$$\int_{\mathcal{X}} f^*(h(x))dq(x) = \sup_{p \text{ positive measure on } \mathcal{X}} \int_{\mathcal{X}} h(x)dp(x) - D(p\|q),$$

where the only difference with Eq. (23) is that p is not assumed to sum to one. Below, we show that for functions $h(x)$ which are quadratic forms in $\varphi(x)$, we can replace $D(p\|q)$ by the lower-bound we just defined above, and obtain a computable upper bound of the integral.

We also have the representation corresponding to Eq. (25), that will be useful later:

$$\int_{\mathcal{X}} f^*(h(x))dq(x) = \inf_{w:\mathcal{X}\to\mathbb{R}} -\int_{\mathcal{X}} w(x)dq(x) \text{ such that } \forall x \in \mathcal{X}, \forall r \geqslant 0,\ rh(x) + w(x) \leqslant f(r). \tag{33}$$

Related work. There exist many ways of estimating integrals, in particular in compact sets in small dimensions, where various quadrature rules, such as the trapezoidal or Simpson's rule, can be applied to compute integrals based on function evaluations, with well-defined convergence rates [18]. In higher dimensions, still based on function evaluations, Bayes-Hermite quadrature rules [48], and the related kernel quadrature rules [16, 5] come with precise convergence rates linking approximation error and number of function evaluations [3]. An alternative in our context is Monte-Carlo integration from samples from q [52], with convergence rate in $O(1/\sqrt{n})$ from n function evaluations.

In this paper, we follow [8] and consider computing integrals given a specific knowledge of the integrand, here of the form $f^*(h(x))$, where h is a known quadratic form in a feature vector $\varphi(x)$. While we also use a sum-of-squares approach as in [8], we rely on different tools (link with f-divergences and partition functions rather than integration by parts).

Relaxations. In order to compute integrals, we simply use the same technique but without the constraint that measures sum to one, that is, without the constraint that $\operatorname{tr}[A] = 1$. Starting from Eq. (33), we get, with $B = \Sigma_q$:

$$\begin{aligned}
\widetilde{C}_q(H) = \int_{\mathcal{X}} f^*\big(\varphi(x)^* H \varphi(x)\big) dq(x) &= \sup_{p \in \mathcal{M}_+(\mathcal{X})} \int_{\mathcal{X}} \varphi(x)^* H \varphi(x) dp(x) - D(p\|q) \\
&\leqslant \sup_{A \in \mathcal{C}} \operatorname{tr}[HA] - D^{\mathrm{OPT}}(A\|B) = \widetilde{C}_q^{\mathrm{OPT}}(H) \\
&= \inf_{N \in \mathbb{H}_d} -\operatorname{tr}[NB] \text{ such that } \forall r \geqslant 0,\ \Gamma(rH + N) \leqslant f(r).
\end{aligned}$$

Note that we only have an inequality here because we are not optimizing over q. We then get two computable relaxations by considering $D^{\mathrm{SOS}}(A\|B)$ and $D^{\mathrm{QT}}(A\|B)$ instead of $D^{\mathrm{OPT}}(A\|B)$, with the respective formulations:

$$\begin{aligned}
\widetilde{C}_q^{\mathrm{SOS}}(H) &= \inf_{N\in\mathbb{H}_d} \; -\operatorname{tr}[NB] \text{ such that } \forall r \geqslant 0,\; f(r)U - rH - N \in \widehat{\mathcal{C}}^* \\
\widetilde{C}_q^{\mathrm{QT}}(H) &= \sup_{A\in\mathbb{H}_d} \inf_{V\succcurlyeq 0,\; V-I\in\mathcal{V}^\perp} \operatorname{tr}[AH] - \operatorname{tr}\big[B^{1/2}VB^{1/2}f\big(B^{-1/2}AB^{-1/2}\big)\big].
\end{aligned}$$

Dual formulations and algorithms can then easily be derived.

E Dual formulations

In this appendix we present dual variational formulations to most of the formulations proposed in the main paper. We only consider the relaxations of f-divergences. Formulations for f-partition functions can be derived similarly.

f-divergences. We can consider the Lagrangian dual of Eq. (3), by introducing a Lagrange multiplier λ for the infinite-dimensional constraint $\forall x \in \mathcal{X}, \forall r \geqslant 0,\; rv(x) + w(x) \leqslant f(r)$ in the form of a positive finite measure λ on $\mathcal{X} \times \mathbb{R}_+$ [30]. We then obtain, using strong duality for equality constraints [39, Section 8.6]:

$$\begin{aligned}
D(p\|q) &= \inf_{\lambda\in\mathcal{M}_+(\mathcal{X}\times\mathbb{R}_+)} \sup_{v,w:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} v(x)dp(x) + \int_{\mathcal{X}} w(x)dq(x) + \int_{\mathcal{X}}\int_{\mathbb{R}_+} \big[f(r) - rv(x) - w(x)\big]d\lambda(x,r) \\
&= \inf_{\lambda\in\mathcal{M}_+(\mathcal{X}\times\mathbb{R}_+)} \int_{\mathcal{X}}\int_{\mathbb{R}_+} f(r)d\lambda(x,r) \qquad (34) \\
&\qquad\qquad \text{such that } \int_{\mathbb{R}_+} d\lambda(\cdot,r) = dq(\cdot) \text{ and } \int_{\mathbb{R}_+} r d\lambda(\cdot,r) = dp(\cdot),
\end{aligned}$$

with the two constraints resulting from the maximization with respect to v and w.

Optimal relaxation of f-divergences (D^{OPT}). We can also formulate Eq. (15) as a minimization problem; we we can use Lagrangian duality, akin to Eq. (34). This requires to introduce a Lagrange multiplier for the constraint $\forall r \geqslant 0,\; \Gamma(rM+N) \leqslant f(r)$, which is equivalent to, $\forall r \geqslant 0,\; f(r) - rM - N \in \widehat{\mathcal{C}}^*$, which leads to a $\mathcal{C}$-valued finite measure on $\mathbb{R}_+$ [30], to get:

$$\begin{aligned}
& D^{\mathrm{OPT}}(A\|B) \\
&= \inf_{\Lambda\ \mathcal{C}-\text{valued measure on } \mathbb{R}_+} \sup_{M,N\in\mathbb{H}_d} \operatorname{tr}[AM] + \operatorname{tr}[BN] + \int_0^{+\infty} \operatorname{tr}\big[d\Lambda(r)(f(r) - rM - N)\big] \\
&= \inf_{\Lambda\ \mathcal{C}-\text{valued measure on } \mathbb{R}_+} \int_0^{+\infty} f(r)\operatorname{tr}\big[d\Lambda(r)\big] \text{ such that } \int_0^{+\infty} d\Lambda(r) = B \text{ and } \int_0^{+\infty} r d\Lambda(r) = A. \qquad (35)
\end{aligned}$$

SOS relaxation of f-divergences (D^{SOS}). We can also get a formulation for Eq. (17), akin to Eq. (34) and Eq. (35). This requires to introduce a Lagrange multiplier for the constraint $\forall r \geqslant 0,\; f(r)U - rM - N \in$

$\widehat{\mathcal{C}}^*$, which is a $\widehat{\mathcal{C}}$-valued finite measure on $\mathbb{R}_+$ [30], to get:

$$
\begin{aligned}
& D^{\mathrm{SOS}}(A\|B) \\
= & \inf_{\Lambda\ \widehat{\mathcal{C}}-\text{valued measure on } \mathbb{R}_+} \sup_{M,N\in\mathbb{H}_d} \operatorname{tr}[AM] + \operatorname{tr}[BN] + \int_0^{+\infty} \operatorname{tr}\left[d\Lambda(r)(f(r) - rM - N)\right] \\
= & \inf_{\Lambda\ \widehat{\mathcal{C}}-\text{valued measure on } \mathbb{R}_+} \int_0^{+\infty} f(r) \operatorname{tr}\left[d\Lambda(r)\right] \text{ such that } \int_0^{+\infty} d\Lambda(r) = B \text{ and } \int_0^{+\infty} r d\Lambda(r) = A, \qquad (36)
\end{aligned}
$$

as maximizing out $M, N \in \mathbb{H}_d$ introduces linear constraints.

Quantum relaxations of f-divergences (D^{SOS}). Eq. (21) also has a primal-dual form as

$$
\begin{aligned}
D^{\mathrm{QT}}(A\|B) \quad = \quad & \sup_{V\succcurlyeq 0,\ V-I\in\mathcal{V}^\perp} \ \inf_{\Lambda\ \mathbb{H}_d^+-\text{valued measure on } \mathbb{R}_+} \int_0^{+\infty} f(r) \operatorname{tr}\left[d\Lambda(r)V\right] \qquad (37) \\
& \text{such that } \int_0^{+\infty} d\Lambda(r) = B \text{ and } \int_0^{+\infty} r d\Lambda(r) = A.
\end{aligned}
$$

We also have a formulation akin to Eq. (36), that is, adding a Lagrange multiplier $\Sigma \in \mathcal{V}$ in Eq. (37) for the constraint $I - V \in \mathcal{V}^\perp$:

$$
\begin{aligned}
D^{\mathrm{QT}}(A\|B) \quad = \quad & \inf_{\Sigma\in\mathcal{V},\ \Lambda\ \mathbb{H}_d^+-\text{valued measure on } \mathbb{R}_+} \int_0^{+\infty} f(r) \operatorname{tr}\left[d\Lambda(r)\right] + \operatorname{tr}[\Sigma] \\
& \text{such that } \int_0^{+\infty} d\Lambda(r) = B, \int_0^{+\infty} r d\Lambda(r) = A, \text{ and } \int_0^{+\infty} f(r) d\Lambda(r) \preccurlyeq \Sigma,
\end{aligned}
$$

which shows the additional relaxation compared to $D^{\mathrm{SOS}}(A\|B)$, for which Λ is a measure (almost everywhere) valued in $\mathcal{V}$, while here it is only in $\mathbb{H}_d^+$.

References

[1] Syed Mumtaz Ali and Samuel D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966. (cited on page 1)

[2] Shun-ichi Amari and Atsumi Ohara. Geometry of q-exponential family of probability distributions. *Entropy*, 13(6):1170–1185, 2011. (cited on page 18)

[3] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(1):714–751, 2017. (cited on page 28)

[4] Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 2022. (cited on pages 8, 11, 14, 16, 22, 23, 24, and 25)

[5] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *International Conference on Machine Learning*, pages 1355–1362, 2012. (cited on page 28)

[6] Francis Bach and Alessandro Rudi. Exponential convergence of sum-of-squares hierarchies for trigonometric polynomials. *SIAM Journal on Optimization*, 33(3):2137–2159, 2023. (cited on pages 10, 21, and 24)

[7] Aharon Ben-Tal and Marc Teboulle. Penalty functions and duality in stochastic programming via φ-divergence functionals. *Mathematics of Operations Research*, 12(2):224–240, 1987. (cited on page 1)

[8] Dimitris Bertsimas, Xuan Vinh Doan, and Jean-Bernard Lasserre. Approximating integrals of multivariate exponentials: A moment approach. *Operations Research Letters*, 36(2):205–210, 2008. (cited on page 28)

[9] Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013. (cited on page 6)

[10] Andrew Blake, Pushmeet Kohli, and Carsten Rother. *Markov random fields for vision and image processing*. MIT Press, 2011. (cited on page 3)

[11] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. (cited on page 1)

[12] 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. (cited on page 25)

[13] Michel Broniatowski and Amor Keziou. Minimization of φ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006. (cited on page 5)

[14] Jean-François Cardoso. Dependence, correlation and Gaussianity in independent component analysis. *Journal of Machine Learning Research*, 4:1177–1203, 2003. (cited on page 1)

[15] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011. (cited on page 14)

[16] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence*, pages 109–116, 2010. (cited on page 28)

[17] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1999. (cited on page 1)

[18] David Cruz-Uribe and C. J. Neugebauer. Sharp error bounds for the trapezoidal rule and Simpson's rule. *Journal of Inequalities in Pure and Applied Mathematics*, 3(4), 2002. (cited on page 28)

[19] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967. (cited on pages 1 and 3)

[20] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461, 1976. (cited on page 17)

[21] Bogdan Dumitrescu. *Positive Trigonometric Polynomials and Signal Processing Applications*, volume 103. Springer, 2007. (cited on page 10)

[22] Kun Fang and Hamza Fawzi. The sum-of-squares hierarchy on the sphere and applications in quantum information theory. *Mathematical Programming*, 190(1):331–360, 2021. (cited on page 10)

[23] Hamza Fawzi and Omar Fawzi. Defining quantum divergences via convex optimization. *Quantum*, 5:387, 2021. (cited on page 7)

[24] Walter Gautschi. *Numerical Analysis*. Springer Science & Business Media, 2011. (cited on pages 21, 22, and 26)

[25] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C. Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014. (cited on page 25)

[26] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996. (cited on page 7)

[27] William W. Hager. Minimizing a quadratic over a sphere. *SIAM Journal on Optimization*, 12(1):188–208, 2001. (cited on page 10)

[28] Frank Hansen and Gert Kjærgård Pedersen. Jensen's inequality for operators and Löwner's theorem. *Mathematische Annalen*, 258(3):229–241, 1982. (cited on page 27)

[29] Fumio Hiai and Milán Mosonyi. Different quantum f-divergences and the reversibility of quantum operations. *Reviews in Mathematical Physics*, 29(07):1750023, 2017. (cited on page 7)

[30] Johannes Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer, 2020. (cited on pages 29 and 30)

[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. (cited on page 8)

[32] James E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960. (cited on page 13)

[33] Jean-Bernard Lasserre. An explicit exact SDP relaxation for nonlinear 0–1 programs. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 293–303. Springer, 2001. (cited on page 11)

[34] Jean-Bernard Lasserre. *Moments, Positive Polynomials and their Applications*, volume 1. World Scientific, 2010. (cited on pages 2, 8, and 9)

[35] Monique Laurent. A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre relaxations for 0–1 programming. *Mathematics of Operations Research*, 28(3):470–496, 2003. (cited on page 11)

[36] Steffen L. Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996. (cited on page 3)

[37] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. (cited on page 1)

[38] Friedrich Liese and Igor Vajda. f-divergences: sufficiency, deficiency and testing of hypotheses. *Advances in Inequalities from Probability Theory and Statistics*, pages 131–173, 2008. (cited on page 1)

[39] David G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997. (cited on pages 11, 17, 19, and 29)

[40] Keiji Matsumoto. A new quantum version of f-divergence. In *Nagoya Winter Workshop: Reality and Measurement in Algebraic Quantum Theory*, pages 229–273. Springer, 2015. (cited on pages 5, 7, 14, and 27)

[41] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research Ltd, 2005. (cited on page 1)

[42] Ilya Mironov. Rényi differential privacy. In *Computer Security Foundations Symposium*, pages 263–275, 2017. (cited on page 1)

[43] Kevin P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2012. (cited on pages 2, 17, and 28)

[44] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point Polynomial Algorithms in Convex Programming*. SIAM, 1994. (cited on page 9)

[45] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. On surrogate loss functions and f-divergences. *The Annals of Statistics*, 37(2):876–904, 2009. (cited on page 1)

[46] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. (cited on pages 2 and 5)

[47] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014. (cited on page 10)

[48] Anthony O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991. (cited on page 28)

[49] Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003. (cited on pages 2, 8, and 9)

[50] Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f-divergences. Technical Report 2202.05568, arXiv, 2022. (cited on pages 1, 3, and 17)

[51] Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023. (cited on page 1)

[52] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*, volume 2. Springer, 1999. (cited on page 28)

[53] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 2015. (cited on pages 4, 10, and 11)

[54] Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O. Tolstikhin. Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on page 2)

[55] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, 28, 2015. (cited on page 25)

[56] Igal Sason. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018. (cited on pages 1 and 3)

[57] Claus Scheiderer. Sums of squares on real algebraic surfaces. *Manuscripta Mathematica*, 119:395–410, 2006. (cited on page 10)

[58] Lucas Slot and Monique Laurent. Sum-of-squares hierarchies for binary polynomial optimization. *Mathematical Programming*, pages 1–40, 2022. (cited on page 11)

[59] Gabor Szegö. *Orthogonal Polynomials*. American Mathematical Society Colloquium Publications, 1975. (cited on page 10)

[60] Marco Tomamichel. *Quantum Information Processing with Finite Resources: Mathematical Foundations*, volume 5. Springer, 2015. (cited on page 7)

[61] Leslie G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201, 1979. (cited on page 21)

[62] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008. (cited on pages 2, 3, 17, 18, 20, 23, and 28)

[63] David V. Widder. The Stieltjes transform. *Transactions of the American Mathematical Society*, 43(1):7–60, 1938. (cited on page 6)