A generalisation of the maximum entropy principle for curved statistical manifolds

Pablo A. Morales[1] and Fernando E. Rosas[2, 3, 4]

[1] *Research Division, Araya Inc., Tokyo 107-6019, Japan*[*]
[2] *Data Science Institute, Imperial College London, London SW7 2AZ, UK*
[3] *Centre for Psychedelic Research, Department of Brain Science, Imperial College London, London SW7 2DD, UK*
[4] *Centre for Complexity Science, Imperial College London, London SW7 2AZ, UK*

The maximum entropy principle (MEP) is one of the most prominent methods to investigate and model complex systems. Despite its popularity, the standard form of the MEP can only generate Boltzmann-Gibbs distributions, which are ill-suited for many scenarios of interest. As a principled approach to extend the reach of the MEP, this paper revisits its foundations in information geometry and shows how the geometry of curved statistical manifolds naturally leads to a generalisation of the MEP based on the Rényi entropy. By establishing a bridge between non-Euclidean geometry and the MEP, our proposal sets a solid foundation for the numerous applications of the Rényi entropy, and enables a range of novel methods for complex systems analysis.

I. INTRODUCTION

The progressive unveiling of the intricate connections that exists between information theory and statistical mechanics has allowed fundamental advances on our understanding of complex systems [1]. One of the most important methods resulting from those discoveries is the *maximum entropy principle* (MEP), which unifies multiple results and procedures under a single heuristic that operationalises Occam's razor [2, 3]. From a pragmatic perspective, the MEP can be understood as a modeling framework that is particularly well-suited for building statistical descriptions of a broad class of systems in contexts of incomplete knowledge [4]. The high versatility of the MEP has allowed it to find applications in a wide range scenarios, including the analysis of DNA motifs of transcription factor binding sites [5], covariations in protein families and amino acid contact prediction [6, 7], diversity of antibody repertoires in the immune system [8, 9], coordinated firing patterns of neural populations [10–13], collective behavior of bird flocks and mice [14–16], the abundance and distribution of species in ecological niches [17, 18], and patterns of behavior in various complex human endeavours [19, 20].

The efficacy of the MEP rests on Shannon's entropy, which acts as as an estimate of "uncertainty" that guides the modeling procedure. Colloquially, the MEP generates the statistical model that is less structured while being consistent with the available knowledge, building on the available knowledge but nothing else. However, the functional form of the Shannon entropy greatly restricts the range outputs that the MEP can offer. In particular, standard applications of the MEP can only generate Boltzmann-Gibbs distributions, which are unsuitable to describe complex systems displaying long-range correlations or other effects related to different types of statistics [21–24]. This important limitation have triggered various efforts to generalise the MEP by means of leveraging generalisations of Shannon's entropy, resulting in a rich array of proposals (see e.g. [25–29]). However, we argue that plugging a generalised entropy into the MEP framework inevitably leads to an adhoc procedure whose value is fundamentally hindered by the heuristic nature of the MEP itself.

An alternative approach to extend the MEP is to consider it not as a stand-alone principle, but as a consequence of deeper mathematical laws. One route to do this — that we follow in this paper — is to regard the MEP as a direct consequence of the geometry of statistical manifolds [30, Sec.III-D]. In effect, by leveraging the structure of dual orthogonal projections allowed by the flat geometry associated with the Kullback–Leibler divergence [31, 32], the seminal work of Amari established how the standard MEP naturally emerges when considering hierarchical "foliations" of the manifold. This perspective not only sets the MEP on a firm mathematical bases, but further endows it with sophisticated tools from information geometry — which can be used e.g. to disentangle the relevance of interactions of different orders within the system [33–35].

In this paper we show how the geometry of curved statistical manifolds naturally leads to an extension of the MEP based on the Rényi entropy. In contrast to flat cases, the geometrical structure of curved statistical manifolds disrupts the standard construction of orthogonal projections based on Legendre-dual coordinates, making the analysis of foliations highly non-trivial. Nonetheless, by leveraging the rich literature on curved statistical manifolds [32, 36–40], the framework put forward in this paper reveals how the geometry established by the Rényi divergence is suitable for establishing hierarchical foliations that, in turn, lead to a generalisation of the MEP.

The results presented in this paper serve to emphasise the special place that the Rényi entropy has among other generalised entropies — at least from the perspective of the MEP. Furthermore, it provides a solid mathematical foundation for the plethora of existent applications based

[*] pablo_morales@araya.org

on the Rényi entropy (see e.g. Refs. [41–44]). Furthermore, the novel connection established between information geometry and this generalised MEP opens the door for fertile explorations combining non-Euclidean geometry methods and statistical analyses, which may lead to new insights and techniques to further deepen our understanding of complex systems.

The rest of this article is structured as follows. First, Section II provides a brief introduction to information geometry, emphasising concepts that are key to our proposal. Then, Section III develops the analysis of foliations in curved statistical manifolds, and Section IV establishes its relationship with a maximum Rényi entropy principle. Finally, Section V discusses the implications of our findings and summarises our main conclusions.

II. PRELIMINARIES

A. The Dual Structure of Statistical Manifolds

Our exposition is focused on statistical manifolds $\mathscr{M}$, whose elements are probability distributions $p_\xi(x)$ with $x \in \boldsymbol{\chi}$ and $\xi \in \mathbb{R}^d$. The geometry of such statistical manifolds is determined by two structures: a metric tensor g_p, and a torsion-free affine connection pair (∇, ∇^*) that are dual with respect to g_p. Intuitively, g_p defines norms and angles between tangent vectors and, in turn, establishes curve length and the *shortest* curves. On the other hand, the affine connection establishes contravariant derivatives of vector fields establishing the notion of parallel transportation between neighbouring tangent spaces, which defines what is a *straight* curve.

Traditional Riemannian geometry is build on the assumption that the shortest and the straightest curves coincide, which led to the study of metric-compatible (Levi-Civita) connections — pivotal to the development of the theory of general relativity. However, modern approaches motivated in information geometry [45] and gravitational theories [46, 47] consider more general cases, where the metric and connections are independent from one another. In such geometries, the parallel transport operator $\Pi : T_p\mathscr{M} \to T_q\mathscr{M}$ and its dual Π^* [48] (induced by ∇ and ∇^*, respectively) might differ. The departure of ∇ and ∇^* from self-duality can be shown to be proportional to Chentsov's tensor, which allows for a single degree of freedom traditionally denoted by $\alpha \in \mathbb{R}$ [45]. Put simply, α captures the degree of asymmetry between short and straight curves, with $\alpha = 0$ corresponding to metric-compatible connections where $\nabla = \nabla^*$.

An important property of the geometry of a statistical manifold $(\mathscr{M}, g, \nabla, \nabla^*)$ is its curvature, which can be of two types: the (Riemann-Christoffel) metric curvature, or the curvature associated to the connection. Both quantities capture the distortion induced by parallel transport over closed curves — the former with respect to the Levi-Civita connection, and the latter with respect to ∇ and ∇^*. In the sequel we use the term "curvature" to refer exclusively to the latter type.

B. Establishing geometric structures via divergences

A convenient way to establish a geometry on a statistical manifold is via *divergence maps* [49]. Divergences are smooth, distance-like mappings for the form $\mathcal{D} : \mathscr{M} \times \mathscr{M} \to \mathbb{R}$, which satisfy $\mathcal{D}(p||q) \geq 0$ and vanish only when $p = q$ [50]. We use the shorthand notation $\mathcal{D}[\xi;\xi'] := \mathcal{D}(p_\xi||q_{\xi'})$ when expressing $\mathcal{D}$ under a parametrisation of $\mathscr{M}$ in terms of coordinates $\xi = (\xi^1, \ldots, \xi^n)$ [30]; divergences in this form are often called "contrast functions" (see Ref. [51, Sec. 11]).

Let us see how one can naturally build a metric from a contrast function [49, Sec. 4]. A metric $g(\xi)$ can be built from the second-order expansion of the divergence $\mathcal{D}$ as

$$g_{ij}(\xi) = \left\langle \partial_{\xi^i}, \partial_{\xi^j} \right\rangle = -\partial_{\xi^i,\xi'^j}\mathcal{D}[\xi;\xi']\big|_{\xi=\xi'} , \tag{1}$$

which is positive-definite due to the non-negativity of $\mathcal{D}$. This construction leads to the *Fisher's metric*, which is the unique metric that emerges from a broad class of divergences [49, Th. 5], with this being this closely related Chentsov's theorem [52–55]. Analogously, connections (or equivalently Christoffel symbols) emerge at the third order expansion of the divergence as follows:

$$\Gamma_{ijk}(\xi) = \left\langle \nabla_{\partial_{\xi^i}} \partial_{\xi^j}, \partial_{\xi^k} \right\rangle = - \ \partial_{i,j}\partial_{k'}\mathcal{D}[\xi;\xi']|_{\xi=\xi'}, \tag{2a}$$

$$\Gamma^*_{ijk}(\xi) = \left\langle \nabla^*_{\partial_{\xi^i}} \partial_{\xi^j}, \partial_{\xi^k} \right\rangle = - \ \partial_k\partial_{i',j'}\mathcal{D}[\xi;\xi']|_{\xi=\xi'}, \tag{2b}$$

where the shorthand notation $\partial_{\xi^i} = \partial_i$ and $\partial_{\xi'^i} = \partial_{i'}$ has been adopted for brevity. In summary, Fisher's metric is insensible the choice of divergence but the resulting connections are, and therefore the effects of a particular $\mathcal{D}$ manifest only at third-order. Interestingly, this construction relating the metric and connections with the second and third derivatives of a scalar potential bears a striking resemblance to Kähler structures on complex manifolds, which can be built through further constraints and are applicable to a range of inference problems [56, 57].

The approach of building geometries based on divergences does not lack generality, as it has been shown that any geometry can be expressed by an appropriate divergence [58, 59]. Of the various types of divergences explored in the literature (c.f. [60] and references within), two classes are particularly important: *f-divergences* of the form

$$\mathcal{D}_f[\xi;\xi'] = \int_{\boldsymbol{\chi}} p_\xi(x) f\left(\frac{p_\xi(x)}{q_{\xi'}(x)}\right) d\mu(x) \tag{3}$$

for $f(x)$ convex with $f(1) = 0$, and *Bregman divergences* of the form

$$\mathcal{D}_\phi[\xi;\xi'] = (\xi - \xi') \cdot \mathrm{D}\phi(\xi') - \big(\phi(\xi) - \phi(\xi')\big) \tag{4}$$

$$= \xi \cdot \eta' - \phi(\xi) - \psi(\eta') \tag{5}$$

for $\phi(\xi)$ a concave function [61], with $\mathrm{D}\phi = (\partial\phi/\partial\xi_1, \ldots, \partial\phi/\partial\xi_d)$ denoting the gradient of ϕ, $\psi(\eta) = \min_\xi \big(\eta\cdot\xi - \phi(\xi)\big)$ is the Fenchel–Legendre concave conjugate of ϕ, and η the dual coordinates of ξ such that

$$\xi = \mathrm{D}\psi(\eta) \quad \text{and} \quad \eta = \mathrm{D}\phi(\xi)\ . \tag{6}$$

Each of these types of divergences have important properties from an information geometry perspective: f-divergences are monotonic with respect to coarse-grainings of the domain of events $\boldsymbol{\chi}$, while Bregman divergences enable dual structures that set the basis for orthogonal projections [62].

As mentioned above, the deviation of a given connection ∇ from its corresponding metric-compatible (i.e. Levi-Civita) counterpart can be measured by αT, where T corresponds to the invariant *Amari-Chensov* tensor [63, 64] and $\alpha\in\mathbb{R}$ is a free parameter. The invariance of T implies that the value of α entirely determines the connection, and the corresponding geometry can be obtained from a divergence of the form

$$\mathcal{D}_\alpha(p||q) = \frac{4}{1-\alpha^2}\int_{\boldsymbol{\chi}}\Big\{1 - p^{\frac{1-\alpha}{2}}q^{\frac{1+\alpha}{2}}\Big\}\,d\mu(x)\ , \tag{7}$$

which is known as *α-divergence*. As important particular cases, if $\alpha = 0$ then $\mathcal{D}_\alpha$ becomes the square of Hellinger's distance, and if $\alpha = 1$ then it gives the well-known Kullback-Leibler

$$\mathcal{D}_{\mathrm{KL}}(p||q) = \int_{\boldsymbol{\chi}} p(x)\log\left(\frac{p(x)}{q(x)}\right)d\mu(x)\ . \tag{8}$$

It is worth noting that geometrical structures are invariant under certain types of transformations. For example, consider a divergence $\tilde{\mathcal{D}}$ given by $\tilde{\mathcal{D}}[\xi;\xi'] := F(\mathcal{D}[\xi;\xi'])$, with F a monotone and differentiable function satisfying $F(0) = 0$ [65]. Then, it can be shown using Eqs. (1) and (2) that the metric and connections induced by $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are related as follows:

$$\tilde{g} = F'(0)\,g, \quad \tilde{\Gamma} = F'(0)\,\Gamma, \quad \tilde{\Gamma}^* = F'(0)\,\Gamma^*\ . \tag{9}$$

Therefore, $\tilde{\mathcal{D}}$ gives rise to exactly the same geometrical structure when $F'(0) = 1$, and a scaled version otherwise. More general transformations between divergences and their corresponding geometries are discussed in Section II D.

C. A Pythagorean relationship in curved spaces via the Rényi divergence

The connection induced by the KL divergence and its natural coordinates is flat (i.e. $\Gamma_{ijk}(\xi) = \Gamma^*_{ijk}(\xi) = 0$). However, this does not hold for α-divergences when $\alpha\neq 1$, which retain the same Fisher's metric but induce a connection with constant sectional curvature $\omega = (1-\alpha^2)/4$ over the whole manifold [39, Theorem 7]. This results into a spherical (S^n) geometry for $\alpha\in(0,1)$, or an hyperbolic (H^n) geometry for $\alpha\notin(0,1)$.

A non-zero curvature affects the relationship between geodesics [66]: if the "α-geodesic" joining p and q is orthogonal (with respect to the Fisher metric) to the one joining q and from r, then

$$\begin{aligned}\mathcal{D}_\alpha(p||r) =& \mathcal{D}_\alpha(p||q) + \mathcal{D}_\alpha(q||r)\\ &- \frac{1-\alpha^2}{4}\mathcal{D}_\alpha(p||q)\mathcal{D}_\alpha(q||r)\ ,\end{aligned} \tag{10}$$

resulting in a deviation from the standard "Pythagorean relationship" that is observed for the case of $\alpha = 1$ [31]. However, one can rewrite Eq. (10) as

$$1 - \omega\mathcal{D}_\alpha(p||r) = \big(1-\omega\mathcal{D}_\alpha(p||q)\big)\big(1-\omega\mathcal{D}_\alpha(q||r)\big), \tag{11}$$

which describes the relationship between angles on the sphere or hyperbolic space — depending on the sign of ω [31]. Interestingly, Eq. (11) suggests that a divergence of the form

$$\mathscr{D}_\gamma(p||q) := \frac{1}{\gamma}\log\big(1+\gamma(1+\gamma)\mathcal{D}_\alpha(p||q)\big) \tag{12}$$

$$= \frac{1}{\gamma}\log\int_{\boldsymbol{\chi}} p(x)^{\gamma+1}q(x)^{-\gamma}d\mu(x) \tag{13}$$

with $\alpha = -1-2\gamma$ would recover the "Pythagorean relationship." In fact, $\mathscr{D}_\gamma$ can be recognised as the well-known Rényi divergence of order $\gamma-1$ [39, 45], noting that we follow Ref. [67] in adopting a shifted indexing.

The Rényi divergence is an f-divergence with $f(x) = x^\gamma$ but it is not a Bregman divergence; however, one can re-cast it as a "Bregman-like" divergence [39]. To see this, let's consider $\tilde{p}_\xi\in\mathscr{M}$ to be a deformed exponential family distribution of the form (see Appendix A)

$$\tilde{p}_\xi(x) = \big(1+\gamma\xi\cdot h(x)\big)^{-\frac{1}{\gamma}}e^{-\varphi_\gamma(\xi)}\ , \tag{14}$$

where $h(x)\in\mathbb{R}^d$ is a vector of sufficient statistics of x and φ_γ is a normalising potential given by

$$\varphi_\gamma(\xi) := -\log\int_{\boldsymbol{\chi}}(1+\gamma\xi\cdot h(x))^{-\frac{1}{\gamma}}d\mu(x)\ . \tag{15}$$

Note that Eq. (14) gives a standard exponential family distribution when $\gamma\to 0$. By defining $\mathscr{D}_\gamma[\xi;\xi'] := \mathscr{D}_\gamma(\tilde{p}_\xi||\tilde{p}_{\xi'})$ to be the corresponding contrast function of the Rényi divergence, then one can show that [39, Th.13]

$$\mathscr{D}_\gamma[\xi;\xi'] = \frac{1}{\gamma}\log(1+\gamma\xi\cdot\eta') - \varphi_\gamma(\xi) - \psi_\gamma(\eta')\ , \tag{16}$$

which resembles Eq. (5) but with the factor $\xi\cdot\eta$ replaced by a logarithm. Above,

$$\psi_\gamma(\eta) := \min_\xi\Big\{\log(1+\gamma\xi\cdot\eta) - \varphi_\gamma(\xi)\Big\} \tag{17}$$

is a generalisation of the Fenchel–Legendre transform of φ_γ, which has conjugate coordinates established by

$$\eta = \frac{1}{1+\gamma\xi\cdot\mathrm{D}\varphi_\gamma(\xi)}\mathrm{D}\varphi_\gamma(\xi)\ , \tag{18a}$$

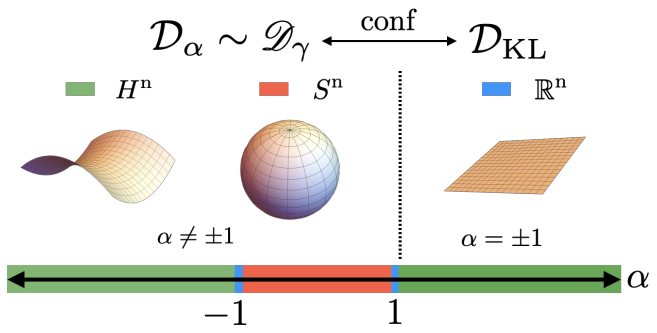

FIG. 1. An schematic diagram depicting the three classes of geometrical structures that arise from their α-value. The curved (i.e. $\alpha \neq \pm 1$) geometries are characterized by the α- and Rényi's divergence, both of which are conformally-projectively related to the KL divergence — which in turn generates a flat geometry.

$$\xi = \frac{1}{1+\gamma\xi\cdot \mathrm{D}\psi_\gamma(\eta)}\mathrm{D}\psi_\gamma(\eta)\ , \tag{18b}$$

with $\mathrm{D}\varphi$ denoting the Euclidean gradient of φ. Finally, it is worth noting that

$$\mathrm{D}\varphi_\gamma(\xi) = \mathbb{E}_\xi\left\{\frac{h(X)}{1+\gamma\xi\cdot h(X)}\right\} =: \mathbb{E}_\xi\{Z_\xi(h)\}\ , \tag{19}$$

where X is a random variable that follows the distribution $p_\xi(x)$, $h(X)$ denotes the sufficient statistics of X, and $Z_\xi(h)$ is defined implicitly as the quantity within the curly brackets. Hence these generalised Fenchel-Legendre dual coordinates can be alternatively expressed as

$$\eta = \frac{1}{1+\gamma\xi\cdot \mathbb{E}_\xi\{Z_\xi(h)\}}\mathbb{E}_\xi\{Z_\xi(h)\}\ . \tag{20}$$

For the case of $\gamma = 0$, Eq. (20) reduces to the well-known relationship given by $\eta = \mathbb{E}_\xi\{h(X)\}$, (see Appendix B for further comments).

D. Conformal-projective classes

Conformal transformations are operations over geometric structures that are angle-preserving, amounting to (pseudo) rotations and dilation of the points in the manifold. Technically, a conformal transformation on $\mathscr{M}$ is defined as an invertible map $\omega : \mathscr{M} \to \mathscr{M}$ such that the induced metric by the pull-back map $\omega_* : T_{\omega(p)}\mathscr{M} \to T_p\mathscr{M}$ is related to the original metric up to a scaling factor $\lambda(p) : \mathscr{M} \to \mathbb{R}$ such that

$$g_p(\omega_*(X), \omega_*(Y)) = \lambda(\omega(p)) g_{\omega(p)}(X, Y) \tag{21}$$

for all $X, Y \in T_{\omega(p)}\mathscr{M}$. Correspondingly, two metrics g and $\tilde{g}$ are said to be *conformally equivalent* if they can be linked via a conformal factor λ as in Eq. (21).

Due to their non-Riemannian geometry, geometric transformations on statistical manifolds that are "structure-preserving" are not fully specified by their effect on the metric, but also need to characterise its effect on the connections — which may diverge from metric-dependence via Chentsov's tensor. This characterisation can be done by relaying on the notion of *projectively equivalence*: two connections ∇ and $\bar{\nabla}$ are said to be *projectively equivalent* if there exists a 1-form $\nu = a_i(\xi)\mathrm{d}\xi^i$ that satisfies

$$\Gamma^k_{ij}(\xi) = \tilde{\Gamma}^k_{ij}(\xi) + a_i(\xi)\delta^k_j + a_j(\xi)\delta^k_i\ , \tag{22}$$

with δ^j_i the Kronecker delta [68].

A convenient way to put these notions together and build conformal-projective transformations is by considering transformations over divergences. Two divergences $\mathcal{D}$ and $\tilde{\mathcal{D}}$ are said to belong to the same *conformal-projective class* if two conditions are met: (i) their induced metrics are conformally equivalent, and (ii) their induced connections are projectively equivalent. It can be shown that two divergences belong to the same conformally-projective class if and only if they satisfy

$$\tilde{\mathcal{D}}[\xi;\xi'] = \lambda(\xi)\mathcal{D}[\xi;\xi']\ , \tag{23}$$

with λ being the *conformal-projective factor* [69].

Let us now study the relationship between the geometries induced by $\mathscr{D}_\gamma$, $\mathcal{D}_\alpha$, and $\mathcal{D}_{\mathrm{KL}}$. By considering the inverse of Eq. (12), one finds that the function

$$F(x) = \frac{e^{\gamma x}-1}{(1+\gamma)\gamma}\ , \tag{24}$$

establishes the diffeomorphism $F(\mathscr{D}_\gamma[\xi;\xi']) = \mathcal{D}_\alpha[\xi;\xi']$, which reveals that the Rényi divergence and α-divergences generate exactly the same geometry (as described by Eqs. (9)). Building on this fact, and leveraging the Legendre-like form of the Rényi entropy shown in Eq. (16), a direct calculation shows that the action of F on $\mathscr{D}_\gamma$ can be expressed as a Bregman divergence $\mathcal{D}_\phi$ scaled by a conformal-projective factor [70, Th. 1]:

$$F(\mathscr{D}_\gamma[\xi;\xi']) = \kappa(\xi)\mathcal{D}_\phi[\xi;\xi']\ . \tag{25}$$

Above, ϕ is a scalar potential given by $\phi(\xi) = e^{\gamma\varphi_0(\xi)}$ with $\varphi_0(\xi)$ as given in Eq. (15), and the conformal-projective factor κ has the form

$$\kappa(\xi) = -\frac{1}{\gamma\phi(\xi)}\ . \tag{26}$$

Moreover, please note that $\mathcal{D}_\phi$ describes a dually-flat geometry, belonging to the same equivalent class as the KL divergence. Thus, these results together establishes that Rényi's $\mathscr{D}_\gamma$, $\mathcal{D}_\alpha$, and $\mathcal{D}_{\mathrm{KL}}$ belong to the same conformal-projective equivalence class.

To conclude, let us present a derivation of the functional form of $\kappa(\xi)$ used in Eq. (25) following Ref. [70]. The metric induced by $\mathscr{D}_\gamma[\xi;\xi']$ is given by

$$\tilde{g}_{ij}(\xi) := -\,\partial_{i,j'}\mathscr{D}_\gamma[\xi;\xi']|_{\xi=\xi'} = \kappa(\xi)\partial_{ij}\phi(\xi)\ , \tag{27}$$

and hence $\tilde{g}_{ij}(\xi) = \kappa(\xi) g_{ij}(\xi)$. Furthermore, its induced connection and metric curvature can be found to be

$$\tilde{\Gamma}_{ij}^{\ k}(\xi) = \frac{\partial_i \kappa(\xi)}{\kappa(\xi)} \delta_j^k + \frac{\partial_j \kappa(\xi)}{\kappa(\xi)} \delta_i^k \ , \tag{28a}$$

$$\tilde{R}_{ijk}^{\ \ \ l}(\xi) = \kappa(\xi) \left(\partial_{jk} \frac{1}{\kappa(\xi)} \delta_i^l - \partial_{ik} \frac{1}{\kappa(\xi)} \delta_j^l \right) \ . \tag{28b}$$

Hence, by introducing the 1-form $\nu = \mathrm{d} \log \kappa(\xi)$, one can identify the affine connection induced by $\tilde{\Gamma}_{ij}^{\ k}(\xi)$ as being projectively flat. This 1-form — or equivalently, the conformal factor $\kappa(\xi)$ — can be derived from the Riemann curvature tensor, which for spaces of constant sectional curvature takes the form $R_{ijk}^{\ \ \ l} = K(g_{jk}\delta_i^l - g_{ik}\delta_j^l)$, with $K \in \mathbb{R}$ corresponding to its scalar curvature. As mentioned in Section II C, the geometry induced by the α-divergence has curvature $\omega = (1-\alpha^2)/4$ throughout the whole manifold, and hence its Riemann tensor can be rewritten as

$$R_{ijk}^{\ \ \ l} = \frac{1+\alpha}{2} (\tilde{g}_{jk}\delta_i^l - \tilde{g}_{ik}\delta_j^l) \ , \tag{29}$$

where a factor $\frac{1-\alpha}{2} = \gamma + 1$ from ω has been absorbed by the metric [71]. Moreover, using the fact that the Riemann tensor is left unchanged by the conformal-projective transformation (i.e. $\tilde{R}_{ijk}^{\ \ \ l} = R_{ijk}^{\ \ \ l}$), and recognising that $K = -\gamma$, one can use Eqs. (27), (28b) and (29) to show that

$$\frac{1}{\kappa(\xi)} = -\gamma\phi(\xi) + \sum_i a_i \xi^i + b \ , \tag{30}$$

for some $a_i, b \in \mathbb{R}$. Finally, as the linear terms can be absorbed in the definition of ϕ, Eq. (30) leads to the expression for $\kappa(\xi)$ as shown above.

III. ORTHOGONAL FOLIATIONS IN CURVED STATISTICAL MANIFOLDS

This section presents the study of orthogonal foliations in curved statistical manifolds. For simplicity of the exposition, the rest of the paper focuses on multivariate distributions of n binary random variables — i.e. distributions of the form $p(x)$ where $x = (x_1, \dots, x_n)$ with $x_i \in \{0,1\}$, and hence $\boldsymbol{\chi} = \{0,1\}^n$.

A. Orthogonal foliations on flat-projective spaces

Let us consider a parametrisation ν of the manifold $\mathscr{M}$. Then, for a given $p_\nu \in \mathscr{M}$ we define

$$\tilde{\mathbf{M}}_k\{p_\nu\} := \{q_{\nu'} \in \mathscr{M} | \nu'_i = \nu_i \ \forall i = 1, \dots, k\} \ , \tag{31}$$

which establishes a nested structure on the manifold of the form

$$\{p\} = \tilde{\mathbf{M}}_n\{p\} \subset \tilde{\mathbf{M}}_{n-1}\{p\} \subset \cdots \subset \tilde{\mathbf{M}}_0\{p\} = \mathscr{M} \ . \tag{32}$$

The parametrisation p_ν also induces a natural basis for the cotangent space at each $p \in \mathscr{M}$, which we denote by $\partial_{\nu_i} \in T_p^*\mathscr{M}$. To study the geometry of this basis, let's consider the functional form of $\mathscr{D}_\gamma$ induced by ν, which is given by $\mathscr{D}_\gamma[\nu;\nu'] := \mathscr{D}_\gamma(p_\nu || p_{\nu'})$. Then, the inner product between the basis elements ∂_{ν_i} can be calculated as

$$\langle \partial_{\nu_i}, \partial_{\nu'_j} \rangle = -\partial_{\nu_i, \nu'_j} \mathscr{D}_\gamma[\nu;\nu']\big|_{\nu'=\nu} = \tilde{g}^{ij}(\nu) \ . \tag{33}$$

The properties of $\mathscr{D}_\gamma$ guarantees that $\tilde{g}^{ij}(\nu)$ is positive-definite, and hence it has a well-defined inverse for each ν which we denote by $r^{ij}(\nu) := \big(g^{-1}(\nu)\big)^{ij}$. By denoting as θ the primal coordinates with respect to r, one can then define

$$\tilde{\mathbf{E}}_k := \{p_\theta \in \mathscr{M} | \theta_j = \theta_j^u, \ \forall j > k\}, \tag{34}$$

where θ^u denote the θ-coordinates of the uniform distribution u. It is direct to verify that

$$\{u\} = \tilde{\mathbf{E}}_0 \subset \tilde{\mathbf{E}}_1 \subset \cdots \subset \tilde{\mathbf{E}}_n = \mathscr{M} \ . \tag{35}$$

Interestingly, $\tilde{\mathbf{E}}_k$ grows with k while $\tilde{\mathbf{M}}_k$ shrinks such that for each k their combined dimensions sum up to n — being enough to account for the dimensionality of $\mathscr{M}$. Furthermore, due to the fact that these complementary dimensions are orthogonal, this implies that their intersection cannot be empty.

We summarise these ideas in the following definition.

Definition 1. *For a given parametrisation ν of $\mathscr{M}$ for which $\tilde{\boldsymbol{E}}_k$ exists, then the* orthogonal foliation *of $\mathscr{M}$ associated to p_ν is the collection of sets $\{\tilde{\boldsymbol{M}}_k\{p_\nu\}, \tilde{\boldsymbol{E}}_k\}$.*

Please note that the bases of $T_p\mathscr{M}$ and $T_p^*\mathscr{M}$ determined by the generalised Fenchel-Legendre dual coordinates established by Eqs. (18a) and (18b) are not orthogonal under the inner product related to the scalar potential φ and its conjugate if $\gamma > 0$, as discussed in Appendix C. Therefore, the standard relationship between geometric duality and Fenchel-Legendre duality that holds for $\gamma = 0$ is broken in curved statistical manifolds. Nonetheless, projective-flatness allows for the metric induced by $\mathscr{D}_\gamma$ to be expressible in coordinates where the bases are manifestly orthogonal up to a conformal-projective factor, so that $\langle \partial_{\xi_i}, \partial_{\eta^j} \rangle = \kappa(\theta)\delta_i^j$ with $\kappa(\theta)$ as defined in Eq. (26). Then, θ and its Fenchel-Legendre conjugate established by Eq. (6) define a set of conformal-projective coordinates.

Crucially, orthogonal foliations satisfy a Pythagorean property, as shown by the following lemma.

Lemma 1. *Given an orthogonal foliation $\{\tilde{\boldsymbol{M}}_k\{p\}, \tilde{\boldsymbol{E}}_k\}$, if $p \in \tilde{\boldsymbol{M}}_k\{p\}$, $r \in \tilde{\boldsymbol{E}}_k$, and $q \in \tilde{\boldsymbol{M}}_k\{p\} \cap \tilde{\boldsymbol{E}}_k$ then*

$$\mathscr{D}_\gamma(r||p) = \mathscr{D}_\gamma(q||p) + \mathscr{D}_\gamma(r||q) \ . \tag{36}$$

Proof. See Appendix C. □

It is important to note that while building orthogonal coordinates is a relatively simple construction, these don't necessarily generally guarantee a Pythagorean relationship. As a matter of fact, although the equivalence between Rényi's and α-divergences ensures that both divergences induce the same geometry, only Rényi's exhibits a correspondence between orthogonality on the metric and a Pythagorean relationship on the divergence (see Section II C). To illustrate these ideas, let us consider a particular construction where we take $\tilde{\mathbf{M}}_k$ as the set of probabilities distributions with fixed expectation values, denoted by η, and come up with its orthogonal complement. From ϕ as the potential encoding these change of coordinates, we define its conjugate potential $\bar{\psi} = \min_\xi(\xi \cdot \eta - \phi(\xi))$. In this way, the primal coordinates $\bar{\xi}$ orthogonal to η follow from $\mathrm{D}(\xi \cdot \eta - \phi(\xi))$, that is,

$$\bar{\xi}^i = \mathbb{E}_\xi\{h^i(x)\} - \frac{1}{\gamma\kappa(\xi)}(\mathrm{D}\log\kappa(\xi))^i \ , \tag{37}$$

where the first term in the right hand side follows from $\eta^i = \mathbb{E}_\xi\{h^i(x)\}$. The primal coordinates $\bar{\xi}^i$, allows to construct an orthogonal complement to $\tilde{\mathbf{M}}_k$, and from (A1) one finds that

$$\bar{\mathbf{E}}_k(c_{k^+}) = \{p_{\bar{\xi}}(x) \in \mathscr{M} \mid \bar{\xi}_{k^+} = c_{k^+}\}. \tag{38}$$

B. Higher-order hierarchical decomposition

Using a orthogonal foliation, we now introduce the notion of hierarchical decomposition on curved statistical manifolds.

Definition 2. *The k-th order γ-projection of $p \in \mathscr{M}$ under the orthogonal foliation $\{\tilde{\boldsymbol{M}}_k\{p\}, \tilde{\boldsymbol{E}}_k\}$ is*

$$\tilde{p}^{(k)} := \operatorname*{arg\,min}_{q \in \tilde{\boldsymbol{E}}_k} \mathscr{D}_\gamma(p\,;q) = \operatorname*{arg\,min}_{q \in \tilde{\boldsymbol{E}}_k} \mathcal{D}_\alpha(p\,;q) \ . \tag{39}$$

Above, the minimum under $\mathscr{D}_\gamma$ and $\mathcal{D}_\alpha$ is the same, as both divergences are related by a monotonous function as shows by Eq. (12). An useful property of the orthogonal foliation is that it enables a useful characterisation of $\tilde{p}^{(k)}$ for $k > 0$, as shown in the next Lemma.

Lemma 2. *The k-th order γ-projection of $p \in \mathscr{M}$ satisfies $\{\tilde{p}^{(k)}\} = \tilde{\boldsymbol{E}}_k \cap \tilde{\boldsymbol{M}}_k\{p\}$.*

Proof. Consider $q \in \tilde{\mathbf{E}}_k \cap \tilde{\mathbf{M}}_k\{p\}$. It is direct to verify that $p, q \in \tilde{\mathbf{M}}_k\{p\}$ and $q, \tilde{p}^{(k)} \in \tilde{\mathbf{E}}_k$. Then, Lemma 1 implies that

$$\mathscr{D}_\gamma(p||\tilde{p}^{(k)}) = \mathscr{D}_\gamma(p||q) + \mathscr{D}_\gamma(q||\tilde{p}^{(k)}) \geq \mathscr{D}_\gamma(p||q) \ . \tag{40}$$

Additionally, Eq. (39) and the fact that $q \in \tilde{\mathbf{E}}_k$ imply that $\mathscr{D}_\gamma(p||q) \geq \mathscr{D}_\gamma(p||\tilde{p}^{(k)})$, which together with Eq. (40) show that $\mathscr{D}_\gamma(p||\tilde{p}^{(k)}) = \mathscr{D}_\gamma(p||q)$. This, combined again with Eq. (40), implies in turn that $\mathscr{D}_\gamma(q||\tilde{p}^{(k)}) = 0$, which can only be satisfied if $q = \tilde{p}^{(k)}$. □

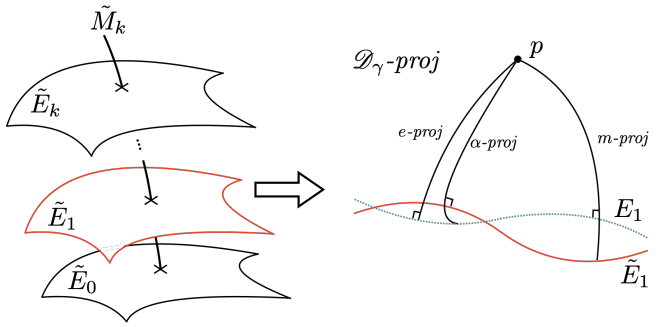

FIG. 2. *(left)* Orthogonal foliation of manifold $\mathscr{M}$. *(right)* Projections onto $\mathbf{E}_1$ leaf (associated with $\alpha = 1$) and its deformation $\tilde{\mathbf{E}}_1$ related to $\alpha \neq \pm 1$.

Following Ref. [30], let us consider the mixed coordinates $\nu_k = (\eta_{k^-}; \xi_{k^+})$. Then, due to the duality of η and ξ, one can verify that $\tilde{p}^{(k)}$ satisfy the mixed coordinates $\tilde{\nu}_k = (\eta_{k^-}; 0)$, where η_{k^-} are the constraints of order up to k of p. Interestingly, note that $u = \tilde{\mathbf{E}}_0(0)$ is equal to the uniform distribution u for all $p \in \mathscr{M}$ and all γ.

With these definitions at hand, we can prove the following result.

Theorem 1. *For a given $p \in \mathscr{M}$, the collection of the γ-projections $\tilde{p}^{(n-1)}, \ldots, u$ satisfy*

$$\mathscr{D}_\gamma(p||u) = \sum_{k=1}^{n} \mathscr{D}_\gamma(\tilde{p}^{(k)}||\tilde{p}^{(k-1)}) \ . \tag{41}$$

Proof. Let's start noting that both $\tilde{p}^{(n-1)}$ and u belong to $\tilde{\mathbf{E}}_{n-1}$, while both p and $\tilde{p}^{(n-1)}$ belong to $\tilde{\mathbf{M}}_{n-1}$ due to Lemma 2. Therefore, Lemma 1 implies that

$$\mathscr{D}_\gamma(p||u) = \mathscr{D}_\gamma(p||\tilde{p}^{(n-1)}) + \mathscr{D}_\gamma(\tilde{p}^{(n-1)}||u) \ . \tag{42}$$

The rest of the proof can be done following a similar rationale recursively on $\mathscr{D}_\gamma(\tilde{p}^{(n-1)}||u)$.

□

To better understand the deformation of the layers induced by γ, it is beneficial to consider the mean-field theory approach presented in Ref. [72]. Let's consider a classic Ising model for which two layers suffice to describe the system, and focus in its projection to E_1. In [72] the m and e projections denote the solution and naive approximations, respectively, which are both orthogonal. Moreover, the α-projection draws the trajectory of solutions in between. In the current picture, however, the submanifolds are deformed in such a way that the α-projection becomes orthogonal with $\alpha = \pm 1$, which are left as fixed points (see Figure 2).

IV. GENERALISING THE MAXIMUM ENTROPY PRINCIPLE

A. Rényi's entropy and related quantities

Consider a manifold of distributions whose support allows a flat distribution. Then, the α-*negentropy* of p is defined as

$$\mathcal{N}_\gamma(p) := \Lambda - H_\gamma(p) \ , \tag{43}$$

with $H_\gamma = \Lambda$ being the Rényi entropy of the uniform distributions, which corresponds to $\log|\boldsymbol{\chi}|$ for finite $\boldsymbol{\chi}$ or $\log n$ in the continuum, and

$$H_\gamma(p) = \frac{-1}{\gamma}\log\int_{\boldsymbol{\chi}} p(x;\xi)^{\gamma+1} d\mu(x) \tag{44}$$

being the well-known Rényi entropy. This definition recovers the standard Shannon entropy and negentropy in the case $\gamma = 0$ [73].

Another quantity of interest is the γ-*Total Correlation*, defined as

$$\mathrm{TC}_\gamma(\boldsymbol{X}^n) = \sum_{i=0}^{n} H_\gamma(X^i) - H_\gamma(\boldsymbol{X}^n), \tag{45}$$

where $\boldsymbol{X}^n := (X_1, \ldots, X_n)$ is a random vector that distributes according to $p_\xi(X = x)$ with $x = (x_1, \ldots, x_n)$. This is a generalisation of the well-known Total Correlation for Shannon's entropy (also known as Multi-information [74]), which is a generalisation of Shannon's mutual information for the case of 3 or more variables [75]. In particular, if $n = 2$ then the total correlation gives a Rényi's mutual information.

B. A hierarchical decomposition of Rényi's entropy

With a hierarchical decomposition $p, p^{(n-1)}, \ldots, u$ at hand, we are now poised to address the problem of entropy decomposition based on the relevance of each order.

Lemma 3. *Consider a the γ-projections of $p \in \mathscr{M}$ under an orthogonal foliation $\{\tilde{\boldsymbol{M}}_k\{p\}, \tilde{\boldsymbol{E}}_k\}$ such that $\tilde{\boldsymbol{E}}_0 = \{u\}$ with u the uniform distribution. Then, the following holds for $l < k$:*

$$\mathscr{D}_\gamma(\tilde{p}^{(k)}||\tilde{p}^{(l)}) = H_\gamma(\tilde{p}^{(l)}) - H_\gamma(\tilde{p}^{(k)}) \ . \tag{46}$$

Proof. A direct application of Eq.(41) shows that

$$\mathscr{D}_\gamma(\tilde{p}^{(k)}||u) = \mathscr{D}_\gamma(\tilde{p}^{(k)}||\tilde{p}^{(l)}) + \mathscr{D}_\gamma(\tilde{p}^{(l)}||u) \ . \tag{47}$$

Then, the desired result follows from re-ordering the terms and using the fact that $\mathscr{D}_\gamma(q||u) = \Lambda - H_\gamma(q)$ for any $q \in \mathscr{M}$. □

Corollary 1. *For any multivariate distribution p then*

$$\mathcal{N}_\gamma(p) = \mathscr{D}_\gamma(p||u) \ , \tag{48}$$

$$\mathrm{TC}_\gamma(\boldsymbol{X}^n) = \mathscr{D}_\gamma\left(p\middle|\middle|\prod_{k=1}^{n} p_{X_k}\right) \ . \tag{49}$$

Using this lemma, we can put forward our main result.

Theorem 2. *Consider $p \in \mathscr{M}$ and an orthogonal foliation $\{\tilde{\boldsymbol{M}}_k\{p\}, \tilde{\boldsymbol{E}}_k\}$ such that $\tilde{\boldsymbol{E}}_0 = \{u\}$. Then,*

$$\tilde{p}^{(k)} = \operatorname*{arg\,max}_{q \in \tilde{\boldsymbol{M}}_k\{p\}} H_\gamma(q) \ . \tag{50}$$

Additionally, the Rényi negentropy can be decomposed as

$$\mathcal{N}_\gamma(p) = \sum_{k=1}^{N} \Delta^{(k)} H_\gamma(p) \ , \tag{51}$$

with $\Delta^{(k)}H_\gamma(p) := H_\gamma\big(\tilde{p}^{(k-1)}\big) - H_\gamma\big(\tilde{p}^{(k)}\big) > 0$ quantifying the relevance of the k-th order constraints.

Proof. Because $\tilde{p}^{(k)} \in \tilde{\mathbf{M}}_k$ (see Lemma 2), then thanks to Lemma 1 any $r \in \tilde{\mathbf{M}}_k$ satisfies

$$\mathscr{D}_\gamma(r||u) = \mathscr{D}_\gamma(r||\tilde{p}^{(k)}) + \mathscr{D}_\gamma(\tilde{p}^{(k)}||u) \ . \tag{52}$$

Therefore, $\mathscr{D}_\gamma(r||u) \geq \mathscr{D}_\gamma(\tilde{p}^{(k)}||u)$ for all $r \in \tilde{\mathbf{M}}_k$, and hence it follows that

$$\tilde{p}^{(k)} = \operatorname*{arg\,min}_{q \in \tilde{\mathbf{M}}_k} \mathscr{D}_\gamma(q||u) = \operatorname*{arg\,max}_{q \in \tilde{\mathbf{M}}_k} H_\gamma(q) \ . \tag{53}$$

Above, the first equality is due to the fact that $\tilde{p}^{(k)} \in \tilde{\mathbf{M}}_k$, and the second equality uses the fact that $\mathscr{D}_\gamma(q||u) = \Lambda - H_\gamma\big(q\big)$.

To prove Eq. (51), one can use Corollary 1 and Theorem 1 to show that

$$\mathcal{N}_\gamma(p) = \mathscr{D}_\gamma(p||u) = \sum_{k=1}^{N} \mathscr{D}_\gamma(\tilde{p}^{(k)}||\tilde{p}^{(k-1)}) \ . \tag{54}$$

The desired result is then a consequence of Lemma 3. □

Above, $\Delta^{(k)}H_\gamma(p)$ accounts for the relevance of the k-th order interactions. In particular, the first order term accounts for all the non-interactive part:

$$\Delta^{(1)}H_\gamma(p) = \sum_{j=1}^{N} \mathcal{N}_\gamma(X_j) = \sum_{j=1}^{N} \Big(\log n - H_\gamma\big(X_j\big)\Big) \tag{55}$$

with $\mathcal{N}_\gamma(X_j)$ being the marginal negentropy of X_j. The remaining terms can be seen to be equal to

$$\sum_{k=2}^{N} \Delta^{(k)} H_\gamma(p) = \mathrm{TC}_\gamma(p) \tag{56}$$

showing that the TC_γ captures all the correlated part of the Rényi negentropy, following the relationship observed in Shannon's case for $\gamma = 0$ (as discussed in Ref.[75]).

C. Maximum Rényi entropy distributions over constraints on average observables

Let us now consider a collection of observables h over a system of n binary variables defined as

$$h^{i,k}(x) = \prod_{j=1}^{k} x_{I_i^k(j)} \ , \tag{57}$$

with $h^{i,k}$ being the i-th observable of k-th order, with $I_i^k(j)$ being an appropriate assignment of indices. Then, one can define the following coordinates:

$$\nu^{i,k} := \mathbb{E}\{h^{i,k}(x)\} \ . \tag{58}$$

For example, $\nu^{i,1}$ are of the form $\mathbb{E}\{x_i\}$ and $\nu^{j,2}$ of the form $\mathbb{E}\{x_r x_s\}$. Importantly, given that $x_1, \ldots, x_n$ are binary variable then one can check that, once $\nu^{i,l}$ for all i and $l \leq k$ are fixed, this in turn determines all the k-th order marginals [76]. Crucially, this implies that the parameters ν as a whole determine a unique distribution $p_\nu(x)$, and hence ν is a valid parametrisation of the corresponding statistical manifold [14, 35].

Let us now consider the family of sets $\tilde{\mathbf{M}}_k$, as defined in Eq.(31) associated to this parametrisation. According to the previous discussion, $\tilde{\mathbf{M}}_k\{p\}$ is the set of all distributions for x that are compatible with the k-th order marginals. For determining the form of the corresponding k-th order γ-projection, we use the following lemma.

Lemma 4. *The solution of the optimisation problem*

$$\arg\max_{q \in \mathscr{M}} H_\gamma(q) \qquad s.t. \quad \nu^{i,l} = \mathbb{E}_q\{h^{i,l}(x)\} \tag{59}$$

for all i and $l \leq k$ gives a projection of the form

$$\tilde{p}_\theta^{(k)}(x) = e^{-z_\gamma(\theta)} \big(1 + \gamma\theta \cdot h(x)\big)^{1/\gamma} \ , \tag{60}$$

with $\theta^{i,l} = 0$ for all $l > k$, and a normalisation factor given by $z_\gamma(\theta) = \frac{1}{\gamma} \log \sum_x \big(1 + \gamma\theta \cdot h(x)\big)^{1/\gamma}$.

Proof. Using Theorem 2, it is clear that $\tilde{p}_\theta^{(k)}$ can be found by solving the extreme values of a Lagrangean of the form

$$\begin{aligned} L(q, \theta_0, \{\theta_j\}) =& H_\gamma(q) + \theta_0 \Big(\sum_i q_i - 1 \Big) \\ &+ \sum_j \theta_j \Big(\sum_k q_k F_j(x_k) - \nu_j \Big) \ , \end{aligned} \tag{61}$$

where q is a discrete distribution and θ_j are Lagrange multipliers. The desired result follows from imposing $\partial L / \partial q_i = 0$ and $\partial L / \partial \theta_j = 0$. □

Efficient numerical methods to estimate distributions of the form specified by Eq. (60) will be developed in a separate publication.

V. CONCLUSION

This paper shows how the non-Euclidean geometry of curved statistical manifolds naturally leads to a MEP that uses the Rényi entropy, generalising the traditional MEP framework based on Shannon's — which take place on flat manifolds. This generalisation of the MEP has three important consequences:

- It highlights special geometrical properties of the Rényi entropy, which make it stand apart from other generalised entropies.
- It provides a solid mathematical foundation for the numerous applications of the Rényi entropy and divergence.
- It enables a range of novel methods of analysis for the statistics of complex systems.

Rényi's entropy and divergence represent one of many routes by which the classic information-theoretic definitions can be extended. One fundamental feature of the Rényi divergence — that this work thoroughly exploits — is the correspondence that it establishes between orthogonality with respect to Fisher's metric and a Pythagorean relationship in the divergence (which does not hold in the geometry induced by e.g. the α-divergence). This correspondence is the key property that allows us to build hierarchical foliations, despite the fact that in curved manifolds the link between geometric and Fenchel-Legendre duality is generally broken. It is relevant to highlight that the correspondence between orthogonality and the Pythagorean relationship is not guaranteed by other divergences such as the α-divergence, which makes entropies such as Tsallis' [77] not well suited to extend the MEP — at least from an information geometry perspective [78]. Considering that extensions of the Renyi entropy exist (e.g. Ref. [79]), an interesting open question is to determine the range of divergences that satisfy these properties.

These findings are in agreement with recent research that is revealing special features of the Rényi entropy and divergence in the context of statistical inference and learning. In particular, Refs. [80, 81] show that the Rényi divergence can provide bounds to the generalisation error of supervised learning algorithms. Also, Ref. [82] shows that the Rényi entropy belongs to a class of functionals that are particularly well-suited for inference and estimation. Put together, these findings suggest that the Rényi entropy and divergence might be capable of playing an important role in the development of future data analysis and artificial intelligence frameworks.

This work opens the door to novel data-analyses approaches to study high-order interactions. While commonly neglected, high-order statistics have recently been proven to be instrumental in a wide range of phenomena at the heart of complex systems, including the self-organising capabilities of cellular automata [83], gene-to-

gene information flow [84], neural information processing [85], high-order brain functions [86, 87], and emergent phenomena [88, 89]. However, exhaustive modeling of high-order effects requires an exponential number of parameters; for that reason, practical investigations need to rely on heuristic modeling methods (see e.g. [90, 91]). In contrast, our framework allow us to do projections while optimising the manifold's curvature in order to best match empirical statistics. Importantly, k-th order projections on curved spaces lead to distributions that capture statistical phenomena of order higher than k without increasing the dimensionality of the parametric family. The development of this line of research is part of our future work.

Another set of promising applications is found in condensed matter systems, where the Rényi entropy is often introduced as a measure of the degree of quantum entanglement. In particular, the Rényi entropy results from an heuristic generalisation of the Von Neumann entropy, which has important benefits in being (i) more suitable to numerical simulations [92] and (ii) being easier to measure by experiments [93]. In particular, the Rényi entropy has been shown to be sensible to features of quantum systems such as central charge [94], and knowledge of it at all orders encodes the whole entanglement spectrum of a quantum state [95]. Moreover, in strongly coupled systems, Rényi entropies have been essential for establishing a connection between quantum entanglement and gravity [96, 97]. More recently, the Rényi mutual information has been taking a central role in the identification of phase transitions [43, 44, 98]. The mathematical framework established in this work serves as a solid basis for these investigations, and further allows the exploration of novel application of information geometry tools in these scenarios.

It is our hope that this contribution may serve to widen the range of applicability of the MEP, while fostering theoretical and practical investigations related to the properties of curved statistical manifolds.

ACKNOWLEDGMENTS

The authors thank Shunichi Amari for careful reading of the manuscript and a number of insightful suggestions, and Ryota Kanai and Yike Guo for supporting this research. F.E.R. is supported by the Ad Astra Chandaria foundation.

Appendix A: Deformed exponential family distributions

For completeness, this appendix presents a derivation of the functional form of $\tilde{p}_\xi$ as presented by Eq. (14) that follows Ref. [39, Sec. 4.1]. For this, let us consider an "exponentially-flat" manifold [99], i.e. a manifold $\mathscr{M}$ with a parametrisation ξ such that all $p \in \mathscr{M}$ can be expressed as

$$p_\xi(x) = e^{-\xi\cdot h(x)+\phi(\xi)}, \tag{A1}$$

where $h(x)$ is a vector of sufficient statistics of x, and $-\phi(\xi)$ is the cumulant generating function. Note that this "natural parametrisation" of $\mathscr{M}$ allows to express the corresponding contrast function of the KL, $\mathcal{D}_{\mathrm{KL}}[\xi;\xi'] := \mathcal{D}_{\mathrm{KL}}(p_\xi||p_{\xi'})$, as a Bregman divergence:

$$\mathcal{D}_{\mathrm{KL}}[\xi;\xi'] = (\xi-\xi')\eta - \phi(\xi) + \phi(\xi') \ . \tag{A2}$$

To find a "deformed" exponential distribution $\tilde{p} \in \mathscr{M}$, one needs to find the natural parametrisation of $\mathscr{M}$ that allows to express the Rényi entropy as a Bregman-like divergence. For this purpose, one can rewrite Eq. (A1) in its self-dual form to find

$$\log p_\xi(x) = -\mathcal{D}_{\mathrm{KL}}[\xi:\xi'] - \psi\big(h(x)\big), \tag{A3}$$

with ψ the conjugate of ϕ, and $h(x)$ plays the role of the dual variable η'. Then, one can re-write Eq. (A3) replacing $\mathcal{D}_{\mathrm{KL}}$ with $\mathscr{D}_\gamma$, and use Eq. (16) to obtain

$$\log\tilde{p}_\xi(x) = -\mathscr{D}_\gamma[\xi:\xi'] - \psi_\gamma(h(x)) \tag{A4}$$

$$= -\frac{1}{\gamma}\log\big(1+\gamma\xi\cdot h(x)\big) + \varphi_\gamma(\xi) \ , \tag{A5}$$

which leads to

$$\tilde{p}_\xi(x) = \big(1+\gamma\xi\cdot h(x)\big)^{-\frac{1}{\gamma}} e^{-\varphi_\gamma(\xi)} \tag{A6}$$

with a normalising potential given by Eq. (15). Importantly, one can show that [39, Th.13]

$$\mathscr{D}_\gamma(\tilde{p}_\xi||\tilde{p}_{\xi'}) = \mathscr{D}_\gamma[\xi;\xi'] \ , \tag{A7}$$

which confirms that the parametrisation of $\mathscr{M}$ determined by Eq. (14) is the natural (in the Bregman-like sense) parametrisation of the deformed geometry induced by $\mathscr{D}_\gamma$.

Appendix B: Analysis of deformed expectation values

The deformed expectation values given by Eq. (20) are non-trivial to interpret, and their explicit dependence on ξ makes numerical simulation challenging. However, exploring some ranges of values of γ can help us to flesh out an interpretation for η.

To this end, let us start by considering the Taylor series expansion of the Z_ξ field given by

$$Z_\xi(h) = h(X)\sum_{n=0}^{\infty}(-1)^n(\gamma\xi\cdot h(X))^n \ . \tag{B1}$$

Small values of γ ensure convergence of the series. Now, one may write its expectation value as

$$\mathbb{E}_\xi\{Z^i_\xi(h)\} \simeq \mathbb{E}_\xi\{h^i\} - \gamma\xi^j\mathbb{E}_\xi\{h^ih_j\}$$

$$+\gamma^2\xi^j\xi^k\mathbb{E}_\xi\{h^i h_j h_k\}\ , \tag{B2}$$

where we have retained up to second order corrections. Similarly for η, one can find that

$$\begin{aligned}\eta_i \simeq\ & \mathbb{E}_\xi\{h_i\} - \xi^j(\mathbb{E}_\xi\{h_i h_j\} + \mathbb{E}_\xi\{h_j\}\mathbb{E}_\xi\{h_i\})\gamma \\ & + \xi^j\xi^k(\mathbb{E}_\xi\{h_i h_j h_k\} + \mathbb{E}_\xi\{h_j h_k\}\mathbb{E}_\xi\{h_i\} \\ & \quad + \mathbb{E}_\xi\{h_j\}\mathbb{E}_\xi\{h_k h_i\} \\ & \quad + \mathbb{E}_\xi\{h_j\}\mathbb{E}_\xi\{h_k\}\mathbb{E}_\xi\{h_i\})\gamma^2\ .\end{aligned} \tag{B3}$$

This implies that these Bregman-like dual coordinate generally deviates from the one obtained for $\gamma = 0$ through higher orders moments, which becomes more prominent as one increases the order of its γ-expansion.

Appendix C: Pythagorean relation

This appendix provides a proof for Lemma 1, which follows results presented in Ref. [39].

Proof. Let's consider a primal geodesic connecting p and q with coordinates ξ and a dual geodesic connecting r and q with coordinates η. The geodesics are then proportional to $\xi^i_r - \xi^i_q$ and $\eta_{p,j} - \eta_{q,j}$ respectively. Then, let's define

$$A = \sum_i (\xi^i_r - \xi^i_q)\partial_{\xi^i}\ , \tag{C1}$$

$$B = \sum_j (\eta_{p,j} - \eta_{q,j})\partial_{\eta_j}\ , \tag{C2}$$

and take a look of their inner product

$$\langle A, B\rangle = \Big\langle \sum_i (\xi^i_r - \xi^i_q)\partial_{\xi^i}, \sum_j (\eta^j_q - \eta^j_r)\partial_{\eta_j}\Big\rangle \tag{C3}$$

$$= \sum_{i,j}(\xi^i_r - \xi^i_q)(\eta_{p,j} - \eta_{q,j})\langle\partial_{\xi^i}, \partial_{\eta_j}\rangle\ . \tag{C4}$$

In other words, we rely on the evaluation of (C4), which requires that the inner product of the primal and dual bases induced by the divergence (13), vanish. That is,

$$\langle\partial_{\xi^i}, \partial_{\eta_j}\rangle = \Big\langle \partial_{\xi^i}, \sum_m \partial_{\eta_j}\xi^m\partial_{\xi^m}\Big\rangle \tag{C5}$$

$$= \sum_m \partial_{\eta_j}\xi^m\langle\partial_{\xi^i}, \partial_{\xi^m}\rangle\ , \tag{C6}$$

whose intern product can be directly obtained from the divergence as

$$\tilde{g}_{im}(\xi) = -\partial_i\partial_{m'}\mathscr{D}_\gamma[\xi, \xi']|_{\xi'=\xi} \tag{C7}$$

$$= \left\{\frac{-\partial_{\xi'^m}\eta'_i}{\Pi(\xi,\eta')} + \sum_l \frac{\gamma\,\eta'_i\xi^l}{\Pi(\xi,\eta')^2}\partial_{\xi'^m}\eta'_l\right\}\Bigg|_{\xi'=\xi}\ , \tag{C8}$$

where we use the shorthand notation $\Pi(\xi,\eta') := (1+\gamma\xi\cdot\eta')$. Replacing this expression into (C6) yields

$$\langle\partial_{\xi^i}, \partial_{\eta_j}\rangle = \frac{-1}{\Pi(\xi,\eta)}\delta^j_i + \frac{\alpha}{\Pi(\xi,\eta)^2}\eta_i\xi^j\ . \tag{C9}$$

Using this in Eq. (C4), and adopting $\Pi_q := \Pi(\xi_q, \eta_q)$ for brevity, one finds that

$$\langle A, B\rangle = \sum_{i,j}(\xi^i_r - \xi^i_q)(\eta_{p,j} - \eta_{q,j})\left(\frac{-1}{\Pi_q}\delta^j_i - \frac{\alpha}{\Pi^2_q}\eta_{q,i}\xi^j_q\right)\ . \tag{C10}$$

Evaluating the sum, one finds that this expression is proportional to

$$\Pi_q(\xi_r - \xi_q)\cdot(\eta_p - \eta_q) + \alpha\xi_q\cdot(\eta_p - \eta_q)\eta_q\cdot(\xi_r - \xi_q) \tag{C11}$$

Finally, the Pythagorean relationship in Eq. (36) holds

$$\begin{aligned}\iff & (1+\gamma\xi_q\cdot\eta_p)(1+\gamma\xi_r\cdot\eta_q) = \\ & \qquad (1+\gamma\xi_r\cdot\eta_p)(1+\gamma\xi_q\cdot\eta_q) \qquad (C12)\\ \iff & (\xi_r - \xi_q)\cdot(\eta_p - \eta_q) = \\ & \qquad \gamma(\xi_q\cdot\eta_p)(\xi_r\cdot\eta_q) - \gamma(\xi_r\cdot\eta_p)(\xi_q\cdot\eta_q) \qquad (C13)\end{aligned}$$

as it can be seen directly from its logarithmic dependence and the Fenchel-Lengendre relation for the scalar potentials on point q. Since the primal geodesic and its dual are orthogonal at q, this (C11) must vanish resulting in (C13), hence the Pythagorean relation holds. □

[1] Stefan Thurner, Rudolf Hanel, and Peter Klimek, *Introduction to the theory of complex systems* (Oxford University Press, 2018).

[2] Edwin T Jaynes, "Information theory and statistical mechanics," Physical review **106**, 620 (1957).

[3] Edwin T Jaynes, *Probability theory: The logic of science* (Cambridge university press, 2003).

[4] Rodrigo Cofré, Rubén Herzog, Derek Corcoran, and Fernando E Rosas, "A comparison of the maximum entropy principle across biological spatial scales," Entropy **21**, 1009 (2019).

[5] Marc Santolini, Thierry Mora, and Vincent Hakim, "A general pairwise interaction model provides an accurate description of in vivo transcription factor binding sites," PloS one **9**, e99015 (2014).

[6] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," Proceedings of the National Academy of Sciences **106**, 67–72 (2009).

[7] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo

Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," Proceedings of the National Academy of Sciences **108**, E1293–E1301 (2011).

[8] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan, "Maximum entropy models for antibody diversity," Proceedings of the National Academy of Sciences **107**, 5405–5410 (2010).

[9] Yuval Elhanati, Anand Murugan, Curtis G Callan, Thierry Mora, and Aleksandra M Walczak, "Quantifying selection in immune receptor repertoires," Proceedings of the National Academy of Sciences **111**, 9875–9880 (2014).

[10] Elad Schneidman, Michael J Berry, Ronen Segev, and William Bialek, "Weak pairwise correlations imply strongly correlated network states in a neural population," Nature **440**, 1007–1012 (2006).

[11] Aonan Tang, David Jackson, Jon Hobbs, Wei Chen, Jodi L Smith, Hema Patel, Anita Prieto, Dumitru Petrusca, Matthew I Grivich, Alexander Sher, *et al.*, "A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro," Journal of Neuroscience **28**, 505–518 (2008).

[12] Olivier Marre, Sami El Boustani, Yves Frégnac, and Alain Destexhe, "Prediction of spatiotemporal patterns of neural activity from pairwise correlations," Physical review letters **102**, 138101 (2009).

[13] Rodrigo Cofre and Bruno Cessac, "Exact computation of the maximum-entropy potential of spiking neural-network models," Physical Review E **89**, 052117 (2014).

[14] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M Walczak, "Statistical mechanics for natural flocks of birds," Proceedings of the National Academy of Sciences **109**, 4786–4791 (2012).

[15] Andrea Cavagna, Irene Giardina, Francesco Ginelli, Thierry Mora, Duccio Piovani, Raffaele Tavarone, and Aleksandra M Walczak, "Dynamical maximum entropy approach to flocking," Physical Review E **89**, 042707 (2014).

[16] Yair Shemesh, Yehezkel Sztainberg, Oren Forkosh, Tamar Shlapobersky, Alon Chen, and Elad Schneidman, "High-order social interactions in groups of mice," Elife **2**, e00759 (2013).

[17] John Harte, *Maximum entropy and ecology: a theory of abundance, distribution, and energetics* (OUP Oxford, 2011).

[18] John Harte and Erica A Newman, "Maximum information entropy: a foundation for ecological theory," Trends in ecology & evolution **29**, 384–389 (2014).

[19] Edward D Lee, "Partisan intuition belies strong, institutional consensus and wide Zipf's law for voting blocs in US Supreme Court," Journal of Statistical Physics **173**, 1722–1733 (2018).

[20] Christopher W Lynn, Lia Papadopoulos, Daniel D Lee, and Danielle S Bassett, "Surges of collective human activity emerge from simple pairwise correlations," Physical Review X **9**, 011022 (2019).

[21] Constantino Tsallis, Murray Gell-Mann, and Yuzuru Sato, "Asymptotically scale-invariant occupancy of phase space makes the entropy Sq extensive," Proceedings of the National Academy of Sciences **102**, 15377–15382 (2005).

[22] Petr Jizba and Toshihico Arimitsu, "The world according to Rényi: thermodynamics of multifractal systems," Annals of Physics **312**, 17–59 (2004).

[23] Stefan Thurner and Rudolf Hanel, "Entropies for complex systems: generalized-generalized entropies," in *AIP Conference Proceedings*, Vol. 965 (American Institute of Physics, 2007) pp. 68–75.

[24] Constantino Tsallis, *Introduction to nonextensive statistical mechanics: approaching a complex world* (Springer Science & Business Media, 2009).

[25] Constantino Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," Journal of statistical physics **52**, 479–487 (1988).

[26] Christian Beck and Ezechiel GD Cohen, "Superstatistics," Physica A: Statistical mechanics and its applications **322**, 267–275 (2003).

[27] Petr Jizba and Toshihico Arimitsu, "Observability of Rényi's entropy," Physical Review E **69**, 026128 (2004).

[28] Rudolf Hanel, Stefan Thurner, and Murray Gell-Mann, "Generalized entropies and logarithms and their duality relations," Proceedings of the National Academy of Sciences **109**, 19151–19154 (2012).

[29] Henrik Jeldtoft Jensen and Piergiulio Tempesta, "Group entropies: From phase space geometry to entropy functionals via group theory," Entropy **20**, 804 (2018).

[30] S-I Amari, "Information geometry on hierarchy of probability distributions," IEEE transactions on information theory **47**, 1701–1711 (2001).

[31] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of mathematical monographs (American Mathematical Society, 2000).

[32] Shun-ichi Amari, *Information geometry and its applications*, Vol. 194 (Springer, 2016).

[33] Elad Schneidman, Susanne Still, Michael J Berry, William Bialek, *et al.*, "Network information and connected correlations," Physical review letters **91**, 238701 (2003).

[34] Eckehard Olbrich, Nils Bertschinger, and Johannes Rauh, "Information decomposition and synergy," Entropy **17**, 3501–3517 (2015).

[35] Fernando E. Rosas, Vasilis Ntranos, Christopher J Ellison, Sofie Pollin, and Marian Verhelst, "Understanding interdependency through complex information sharing," Entropy **18**, 38 (2016).

[36] Shun-ichi Amari, Shiro Ikeda, and Hidetoshi Shimokawa, "Information geometry of α-projection in mean-field approximation," Recent Developments of Mean Field Approximation, M. Opper, D. Saad, Eds., MIT Press, Cambridge (2000).

[37] Takashi Kurose, "Conformal-Projective Geometry of Statistical Manifolds," Interdisciplinary Information Sciences **8**, 89–100 (2002).

[38] Hiroshi Matsuzoe *et al.*, "Statistical manifolds and affine differential geometry," in *Probabilistic Approach to Geometry* (Mathematical Society of Japan, 2010) pp. 303–321.

[39] Ting-Kam Leonard Wong, "Logarithmic divergences from optimal transport and Rényi geometry," Information Geometry **1**, 39–78 (2018).

[40] Antonio M Scarfone, Hiroshi Matsuzoe, and Tatsuaki Wada, "A study of Rényi entropy based on the information geometry formalism," Journal of Physics A: Mathematical and Theoretical **53**, 145003 (2020).

[41] Dmitry S Shalymov and Alexander L Fradkov, "Dynamics of non-stationary processes that follow the maximum of the Rényi entropy principle," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **472**, 20150324 (2016).

[42] AG Bashkirov, "On maximum entropy principle, superstatistics, power-law distribution and Renyi parameter," Physica A: Statistical Mechanics and its Applications **340**, 153–162 (2004).

[43] Jean-Marie Stéphan, Stephen Inglis, Paul Fendley, and Roger G. Melko, "Geometric Mutual Information at Classical Critical Points," Phys. Rev. Lett. **112**, 127204 (2014).

[44] Jason Iaconis, Stephen Inglis, Ann B. Kallin, and Roger G. Melko, "Detecting classical phase transitions with renyi mutual information," Phys. Rev. B **87**, 195134 (2013).

[45] Shun-ichi Amari, "Information geometry," Japan. J. Math **16**, 1–48 (2021).

[46] Vincenzo Vitagliano, Thomas P. Sotiriou, and Stefano Liberati, "The dynamics of metric-affine gravity," Annals Phys. **326**, 1259–1273 (2011), [Erratum: Annals Phys. 329, 186–187 (2013)], arXiv:1008.0171 [gr-qc].

[47] Vincenzo Vitagliano, "The role of nonmetricity in metric-affine theories of gravity," Class. Quant. Grav. **31**, 045006 (2014), arXiv:1308.1642 [gr-qc].

[48] The dual transport operator acts on cotangent vectors, and is defined by the condition of guaranteeing $g_q(\Pi V, \Pi^* W) = g_p(V, W)$ for all $W \in T_p\mathscr{M}$ and $V \in T_p^*\mathscr{M}$).

[49] Shun-ichi Amari and Andrzej Cichocki, "Information geometry of divergence functions," Bulletin of the polish academy of sciences. Technical sciences **58**, 183–195 (2010).

[50] Divergences are in general weaker than distances, as they don't need to be symmetric in their arguments and don't need to respect the triangle inequality.

[51] Ovidiu Calin and Constantin Udrişte, *Geometric modeling in probability and statistics* (Springer, 2014).

[52] NN Chentsov, "Statistical decision rules and optimal inference. Transl. Math," Monographs, American Mathematical Society, Providence, RI (1982).

[53] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer, "Information geometry and sufficient statistics," Probability Theory and Related Fields **162**, 327–364 (2015).

[54] Hông Vân Lê, "The uniqueness of the Fisher metric as information metric," Annals of the Institute of Statistical Mathematics **69**, 879–896 (2017).

[55] James G Dowty, "Chentsov's theorem for exponential families," Information Geometry **1**, 117–135 (2018).

[56] Jaehyung Choi and Andrew Mullhaupt, "Kählerian Information Geometry for Signal Processing," Entropy **17**, 1581–1605 (2015).

[57] Jun Zhang and Fubo Li, "Symplectic and Kähler structures on statistical manifolds induced from divergence functions," in *International Conference on Geometric Science of Information* (Springer, 2013) pp. 595–603.

[58] Takao Matumoto *et al.*, "Any statistical manifold has a contrast function—On the C3-functions taking the minimum at the diagonal of the product manifold," Hiroshima Math. J **23**, 327–332 (1993).

[59] Nihat Ay and Shun-ichi Amari, "A novel approach to canonical divergences within information geometry," Entropy **17**, 8111–8129 (2015).

[60] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory," IEEE Transactions on Information Theory **52**, 4394–4412 (2006).

[61] Following Ref. [39], we use a non-standard definition of Bregman divergences based on concave (instead of convex) functions.

[62] Shun-Ichi Amari, "α-divergence is unique, belonging to both f-divergence and bregman divergence classes," .

[63] Nikolai Nikolaevich Cencov, *Statistical decision rules and optimal inference*, 53 (American Mathematical Soc., 2000).

[64] Shun-Ichi Amari, "Differential geometry of curved exponential families-curvatures and information loss," The Annals of Statistics , 357–385 (1982).

[65] Given two divergences $\mathcal{D}$ and $\tilde{\mathcal{D}}$, there is always a function $F : \mathbb{R}^3 \to \mathbb{R}$ such that $\tilde{\mathcal{D}}[\xi;\xi'] = F(\mathcal{D}[\xi;\xi'], \xi, \xi)$. Building on this fact, one can consider three levels of similarity: (i) when F depends only on the first argument — which then implies the corresponding geometries are essentially the same, (ii) when F can be expressed as $F(x, y, z) = f(x)g(y, z)$ — which implies conformal-projective equivalence (see Sec. II D, and (iii) the more general case.

[66] Geodesics are the straight curves established by the connection, which in non-Riemannian geometries are not the same as the shortest curves between two points.

[67] Francisco Valverde-Albacete and Carmen Peláez-Moreno, "The Case for Shifting the Renyi Entropy," Entropy **21**, 46 (2019).

[68] Equivalently, projective equivalence can be defined by $\nabla_X Y = \tilde{\nabla}_X Y + \nu(X)Y + \nu(Y)X$ for any smooth pair of vector fields X and Y.

[69] In general λ could depend on both ξ and ξ' [37, 45]; however, for the purposes of this paper we restrict ourselves to consider only "left conformal-projective factors" (i.e. $\lambda(\xi)$).

[70] Ting-Kam Leonard Wong and Jiaowen Yang, "Logarithmic divergences: geometry and interpretation of curvature," in *International Conference on Geometric Science of Information* (Springer, 2019) pp. 413–422.

[71] Note that the metrics coming from the α and Rényi divergences are conformally related $\tilde{g} = (\gamma+1)g$ as seen by (12).

[72] Shunichi Amari, S Ikeda, and H Shimokawa, "Information geometry of-projection in mean field approximation," Advanced Mean Field Methods , 241–258 (2001).

[73] One can interpret a continuous decrease in the constant sectional α-curvature of the manifold manifesting as a decrease in the order of Rényi's entropy in statistics, eventually converging to Shannon's for $\gamma \to 0$ limit.

[74] Milan Studenỳ and Jirina Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in graphical models* (Springer, 1998) pp. 261–297.

[75] Fernando E. Rosas, Pedro A. M. Mediano, Michael Gastpar, and Henrik J Jensen, "Quantifying high-order interdependencies via multivariate extensions of the mutual information," Physical Review E **100**, 032305 (2019).

[76] The k-th order marginals of p are the distributions considering k of the n variables that compose x, which are obtained by marginalising the other $n - k$ variables.

[77] For an explanation of the close relationship between the Tsallis entropy and the α-divergence, please see [100].

[78] For an interesting related discussion, including thermodynamic aspects, see Ref. [40].

[79] David C De Souza, Rui F Vigelis, and Charles C Cavalcante, "Geometry induced by a generalization of Rényi divergence," Entropy **18**, 407 (2016).

[80] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa, "Generalization Error Bounds Via Rényi-, f-Divergences and Maximal Leakage," arXiv preprint arXiv:1912.01439 (2019).

[81] Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa, "Robust Generalization via f- Mutual Information," in *2020 IEEE International Symposium on Information Theory (ISIT)* (IEEE, 2020) pp. 2723–2728.

[82] Petr Jizba and Jan Korbel, "Maximum entropy principle in statistical inference: Case for non-Shannonian entropies," Physical review letters **122**, 120601 (2019).

[83] Fernando E. Rosas, Pedro AM Mediano, Martín Ugarte, and Henrik J Jensen, "An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems," Entropy **20**, 793 (2018).

[84] Zixuan Cang and Qing Nie, "Inferring spatial and signaling relationships between cells from single cell transcriptomic data," Nature communications **11**, 1–13 (2020).

[85] Michael Wibral, Viola Priesemann, Jim W Kay, Joseph T Lizier, and William A Phillips, "Partial information decomposition as a unified approach to the specification of neural goal functions," Brain and cognition **112**, 25–38 (2017).

[86] Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Negin Holland, Tim D Fryer, John T O'Brien, James B Rowe, David K Menon, Daniel Bor, and Emmanuel A Stamatakis, "A synergistic core for human brain evolution and cognition," bioRxiv (2020).

[87] Andrea I Luppi, Pedro AM Mediano, Fernando E Rosas, Judith Allanson, John D Pickard, Robin L Carhart-Harris, Guy B Williams, Michael M Craig, Paola Finoia, Adrian M Owen, *et al.*, "A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition," bioRxiv (2020).

[88] Fernando E Rosas, Pedro AM Mediano, Henrik J Jensen, Anil K Seth, Adam B Barrett, Robin L Carhart-Harris, and Daniel Bor, "Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data," arXiv preprint arXiv:2004.08220 (2020).

[89] Thomas Varley and Erik Hoel, "Emergence as the conversion of information: A unifying theory," arXiv preprint arXiv:2104.13368 (2021).

[90] Elad Ganmor, Ronen Segev, and Elad Schneidman, "Sparse low-order interaction network underlies a highly correlated and learnable neural population code," Proceedings of the National Academy of sciences **108**, 9679–9684 (2011).

[91] Hideaki Shimazaki, Kolia Sadeghi, Tomoe Ishikawa, Yuji Ikegaya, and Taro Toyoizumi, "Simultaneous silence organizes structured higher-order interactions in neural populations," Scientific reports **5**, 1–13 (2015).

[92] Matthew B Hastings, Iván González, Ann B Kallin, and Roger G Melko, "Measuring Renyi entanglement entropy in quantum Monte Carlo simulations," Physical review letters **104**, 157201 (2010).

[93] Rajibul Islam, Ruichao Ma, Philipp M Preiss, M Eric Tai, Alexander Lukin, Matthew Rispoli, and Markus Greiner, "Measuring entanglement entropy in a quantum many-body system," Nature **528**, 77–83 (2015).

[94] Jean-Marie Stéphan, Stephen Inglis, Paul Fendley, and Roger G Melko, "Geometric mutual information at classical critical points," Physical review letters **112**, 127204 (2014).

[95] Jean-Marie Stéphan, "Shannon and Rényi mutual information in quantum critical spin chains," Phys. Rev. B **90**, 045424 (2014).

[96] Xi Dong, "The Gravity Dual of Renyi Entropy," Nature Commun. **7**, 12472 (2016), arXiv:1601.06788 [hep-th].

[97] Taylor Barrella, Xi Dong, Sean A. Hartnoll, and Victoria L. Martin, "Holographic entanglement beyond classical gravity," JHEP **09**, 109 (2013), arXiv:1306.4682 [hep-th].

[98] Michael P. Zaletel, Jens H. Bardarson, and Joel E. Moore, "Logarithmic Terms in Entanglement Entropies of 2D Quantum Critical Points and Shannon Entropies of Spin Chains," Phys. Rev. Lett. **107**, 020402 (2011).

[99] A direct calculation shows that parametrisations based on exponential family distributions generate a flat connection.

[100] Andrzej Cichocki and Shun-ichi Amari, "Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities," Entropy **12**, 1532–1568 (2010).