CHENTSOV'S THEOREM FOR EXPONENTIAL FAMILIES

JAMES G. DOWTY

Abstract. Chentsov's theorem characterizes the Fisher information metric on statistical models as the only Riemannian metric (up to rescaling) that is invariant under sufficient statistics. This implies that each statistical model is equipped with a natural geometry, so Chentsov's theorem explains why many statistical properties can be described in geometric terms. However, despite being one of the foundational theorems of statistics, Chentsov's theorem has only been proved previously in very restricted settings or under relatively strong invariance assumptions. We therefore prove a version of this theorem for the important case of exponential families. In particular, we characterise the Fisher information metric as the only Riemannian metric (up to rescaling) on an exponential family and its derived families that is invariant under independent and identically distributed extensions and canonical sufficient statistics. We then extend this result to curved exponential families. Our approach is based on the central limit theorem, so it gives a unified proof for discrete and continuous exponential families, and it is less technical than previous approaches.

1. Introduction

Chentsov's theorem is a foundational theorem of statistics that characterizes the Fisher information metric on statistical models as the only Riemannian metric (up to rescaling) that is invariant under certain, statistically important transformations [10, 20, 9, 2, 7]. This effectively means that the Fisher information metric is the only natural metric on a statistical model, so many statistical properties of these models should be describable in terms of this metric. Known examples of this correspondence between statistical and geometric properties include: the Cramér-Rao lower bound for the variance of an unbiased estimator in terms of the inverse of the Fisher information metric [1, Thm. 2.2]; orthogonality as a criterion for first-order efficiency of estimators [1, Thm. 4.3]; the central role of statistical curvature in the information loss of an efficient estimator [13, §3.3] and in second-order efficiency [13, §3.4]; and the spontaneous emergence of the Fisher information volume [18] in the minimum description length (MDL) approach to statistical model selection [6].

The original version of Chentsov's theorem [10, 20, 9] or [1, Thm. 2.6] only applied in the restricted setting of statistical models with finite data spaces. This version of the theorem says that the Fisher information metric is the only metric (up to a multiplicative constant) that is defined on all models with finite data spaces and is invariant under all sufficient statistics. Recall that a statistical model $\mathcal{M}$ is a (sufficiently regular) set of probability measures on the same measurable space $\mathcal{X}$, which we call the data space of $\mathcal{M}$, and that a sufficient statistic for $\mathcal{M}$ is a function on $\mathcal{X}$ for which the conditional distribution of any measure P in $\mathcal{M}$, given the sufficient statistic, is the same for all P. Sufficient statistics induce corresponding maps on statistical models (the measure-theoretic push-forward maps) and the invariance assumption above is that all of these maps are isometries (i.e., distance-preserving maps).

Since the assumption of finite data spaces is very restrictive, Ay et al. [2] proved a version of Chentsov's theorem that applies to models whose data spaces are measurable subsets of a

Date: July 26, 2018.

smooth manifold $\mathcal{X}$ (though in later work [19], $\mathcal{X}$ is allowed to be an arbitrary measurable space). Their version says that the Fisher information metric is the only metric (up to rescaling) that firstly exists on all statistical models with this underlying space $\mathcal{X}$ and secondly is invariant under all sufficient statistics, including discontinuous ones. These results have subsequently been generalised to higher-order Amari-Chentsov tensors [19] and proved using an alternative approach [15]. The version of Chentsov's theorem in [2] applies to many interesting statistical models but it assumes the existence of metrics on a large class of models and very strong invariance properties for these metrics. Therefore Bauer et al. [7] proved a version of Chentsov's theorem that says the Fisher information metric is the only metric (up to rescaling) that firstly is defined on the space of all smooth, positive densities on a compact manifold $\mathcal{X}$ of dimension 2 or higher and secondly is invariant under all diffeomorphisms from $\mathcal{X}$ to itself (where diffeomorphisms are smooth maps with smooth inverses, so they are a special type of sufficient statistic). The proof of Bauer et al. [7] was based on results from the theory of generalized functions, especially the Schwartz kernel theorem [12, §6.1], and it made far weaker invariance assumptions than the proof of Ay et al. [2]. The assumption that $\mathcal{X}$ is a compact manifold without boundary excludes many cases of interest to statisticians, though Bauer et al. [7] say this assumption can be weakened.

Despite their beauty and generality, the results of Ay et al. [2] and Bauer et al. [7] leave open the possibility that there might exist a natural metric other than the Fisher information metric on an individual statistical model $\mathcal{M}$. This could occur, for example, if there is a natural metric on $\mathcal{M}$ that does not (invariantly) extend to a metric on the infinite-dimensional models of [2] and [7] that contain $\mathcal{M}$ and many unrelated models. Also, exponential families have a distinguished, finite-dimensional set of sufficient statistics, called the canonical sufficient statistics, which are related to their natural affine structures ([1, Thm. 2.4] and [3, Lemma 8.1]). Therefore, the invariance assumptions of [2] and [7] are arguably too strong for exponential families, and instead it would be more natural to consider invariance under canonical sufficient statistics rather than all sufficient statistics.

In this paper, we prove a refined version of Chentsov's theorem in the important case of (curved) exponential families. Instead of considering metrics defined on an infinite-dimensional statistical model, as in [2] and [7], we consider metrics defined only on a given exponential family $\mathcal{M}$ and some of its derived families, namely its independent and identically distributed (IID) extensions and their corresponding natural exponential families. Instead of assuming these metrics are invariant under all sufficient statistics or all diffeomorphisms, we assume invariance under canonical sufficient statistics and IID extensions. This assumption of invariance under IID extensions has no analogue in previous work, but IID extensions are natural and important transformations between statistical models (perhaps more so than sufficient statistics), so this invariance assumption is arguably more natural than invariance under sufficient statistics. Also, this extra invariance assumption is offset by the fact that we restrict our sufficient statistics to the canonical ones. Then, under a mild regularity condition, we prove that metrics with these invariance properties are multiples of the Fisher information metric (see Theorem 5.1 in Section 5). This result therefore gives a new characterisation of the Fisher information metric as the only metric on an exponential family and its derived families that is invariant under canonical sufficient statistics and IID extensions.

Our approach has a number of advantages: as discussed above, we only assume that the metric is defined on an individual model and its related models, and our invariance assumptions respect the natural affine structures of exponential families; we only consider metrics on a collection of finite-dimensional models (similar to the original version of Chentsov's theorem [10, 20, 9]), which allows us to avoid the technicalities encountered in [2] and [7] because of the

infinite-dimensionality of their statistical models; our proof is unified for discrete and continuous distributions, unlike the proofs of [10, 20, 9] and [7], so there is some hope of extending our proof to general statistical models; our proof shows that Chentsov's theorem is a corollary of the central limit theorem, which makes this result more understandable and intuitive; and our results complement those of [7], since curved exponential families are essentially the only statistical models with smooth sufficient statistics that are not diffeomorphisms, by the Pitman–Koopman–Darmois theorem [5].

The rest of this paper is set out as follows. In Section 2 we define the Fisher information metric and some relevant notions from differential geometry, as they apply in our main case of interest. In Section 3 we briefly recall the definition of an exponential family and some of its derived families. We then give precise descriptions of our assumptions in Section 4, before using these assumptions and the central limit theorem to prove our characterisation of the Fisher information metric in Section 5. Section 6 then describes extensions of our results to curved exponential families and higher-order symmetric tensors. We compare our version of Chentsov's theorem with previous versions in Section 7 before finishing with a discussion of our results and a non-technical summary of our proof in Section 8.

2. The Fisher information metric

This section briefly recalls the definitions of tangent vectors and the Fisher information metric of a statistical model. General references for the notions from Riemannian geometry described here are [13, Appendix C] and, for infinite-dimensional manifolds, [14].

In all later sections of this paper, we will take $\mathcal{M}$ to be a regular exponential family and Θ to be its natural parameter space, but in this section we let $\mathcal{M}$ be a more general statistical model and let Θ be the parameter space for any smooth parameterisation of $\mathcal{M}$. More precisely, suppose Θ is an open subset of $\mathbb{R}^d$ and that μ is a measure on $\mathbb{R}^m$ with support $\mathcal{X}$. Then our statistical model is $\mathcal{M} = \{p_\theta \mu \mid \theta \in \Theta\}$, where each $p_\theta : \mathcal{X} \to \mathbb{R}_{>0}$ is a μ-integrable, strictly positive function that is normalized, meaning $1 = \int p_\theta d\mu$. Note that $\mathcal{M}$ is a set of probability measures on $\mathbb{R}^m$. We assume that the parameterisation of $\mathcal{M}$ by Θ is smooth, in the sense that $\theta \mapsto p_\theta(x)$ is a smooth (i.e., infinitely differentiable) function for μ-almost all x. We also assume that the parameterisation is non-singular, meaning that the parameterisation map $\Theta \to \mathcal{M}$ given by $\theta \mapsto p_\theta \mu$ is injective and that it maps non-zero tangent vectors to non-zero tangent vectors, in a sense that will become clear below.

Because Θ is an open subset of $\mathbb{R}^d$, any tangent vector u to Θ is a pair $u = (\theta, a)$ for some $\theta \in \Theta$ and some $a \in \mathbb{R}^d$, where θ is called the base-point of u. The set of all such tangent vectors, which is denoted $T\Theta$ and is called the tangent bundle of Θ, is therefore $T\Theta = \Theta \times \mathbb{R}^d$. The tangent bundle is not a vector space in general, but the set of all tangent vectors with the same base-point is. The vector space $T_\theta\Theta$ consisting of all vectors with base-point θ is called the tangent space to Θ at θ. Addition and scalar multiplication in this vector space are given by

$$su + tv = (\theta, sa + tb) \tag{2.1}$$

for any $u, v \in T_\theta\Theta$ and any $s, t \in \mathbb{R}$, where $u = (\theta, a)$ and $v = (\theta, b)$. Note that addition and scalar multiplication in $T_\theta\Theta$ effectively ignore the shared base-point θ.

Similarly, we can view each tangent vector to the statistical model $\mathcal{M}$ as a pair (P, A), where the base-point P is an element of the model $\mathcal{M}$ and A is essentially the score in a particular direction [17, §3.3]. More precisely, for each tangent vector $u = (\theta, a)$ to Θ, there

is a corresponding tangent vector $\tilde{u} = (P, A)$ to $\mathcal{M}$ given by

$$P = p_\theta \mu \text{ and } A = \sum_{i=1}^{d} a_i \frac{\partial p_\theta}{\partial \theta_i} \mu. \tag{2.2}$$

(The function taking u to $\tilde{u}$ is the differential, or tangent map, of the parameterisation $\theta \mapsto p_\theta \mu$ [14, p. 52].) Let the tangent bundle $T\mathcal{M}$ of $\mathcal{M}$ be the set of all such tangent vectors, i.e., let $T\mathcal{M} = \{\tilde{u} \mid u \in T\Theta\}$. Also, let the tangent space $T_P\mathcal{M}$ to $\mathcal{M}$ at $P \in \mathcal{M}$ be the vector space consisting of all tangent vectors $(P, A) \in T\mathcal{M}$ with base-point P. Even though we have used a particular parameterisation of $\mathcal{M}$ to define $T_P\mathcal{M}$, this tangent space is natural, in the sense that $T_P\mathcal{M}$ is the same for all smooth parameterisations [14, p. 52].

The Fisher information metric g^F on $\mathcal{M}$ is given by

$$g^F(\tilde{u}, \tilde{v}) = \int \frac{dA}{dP} \frac{dB}{dP} \, dP \tag{2.3}$$

for any tangent vectors $\tilde{u} = (P, A)$ and $\tilde{v} = (P, B)$ in the tangent space $T_P\mathcal{M}$ [7, §3], where dA/dP and dB/dP are Radon-Nikodym derivatives [8, §3.2]. It is straightforward [11, Appendix A] to show that definition (2.3) for the Fisher information metric reduces to the usual, parameterisation-dependent definition [1, eq. 2.6]. However, the formulation (2.3) will be more useful to us than the usual definition. Also, because (2.3) is phrased only in terms of natural constructions, this formula makes it clear that g^F does not depend on arbitrary choices, such as the choice of parameterisation.

A Riemannian metric on a set is just a function that puts an inner product on each of the set's tangent spaces (if the set is suitably regular and the inner products vary smoothly with the base-point). For example, a Riemannian metric on Θ can be thought of as a smooth, matrix-valued function on Θ whose value at $\theta \in \Theta$ is a $d \times d$, symmetric, positive-definite matrix $\bar{g}_\theta$, since this defines an inner product on each $T_\theta\Theta$ with the inner product of any $u, v \in T_\theta\Theta$ being $g(u, v) = a^T \bar{g}_\theta b$, where $u = (\theta, a)$ and $v = (\theta, b)$.

In our main case of interest, where $\mathcal{M}$ is an exponential family, the integral in (2.3) always converges [13, Thm. 2.2.5]. Then it is not hard to see that (2.3) defines an inner product on each tangent space to $\mathcal{M}$ (and this varies smoothly with the base-point), so the Fisher information metric g^F is a Riemannian metric on $\mathcal{M}$.

3. Exponential families and their derived families

Partly to establish our notation, this section briefly recalls the definitions of an exponential family, its IID extensions and their corresponding natural exponential families.

3.1. **Exponential families.** Let μ be a measure on $\mathbb{R}^m$ and let $T : \mathcal{X} \to \mathbb{R}^d$ be a measurable function, where $\mathcal{X} \subseteq \mathbb{R}^m$ is the support of μ. Let

$$\Theta = \left\{ \theta \in \mathbb{R}^d \;\middle|\; \int \exp(\theta \cdot T) d\mu < \infty \right\},$$

where the dot $(\cdot)$ denotes the Euclidean inner product on $\mathbb{R}^d$. For each $\theta \in \Theta$, define $p_\theta : \mathcal{X} \to \mathbb{R}_{>0}$ by

$$p_\theta(x) = \exp(\theta \cdot T(x))/Z(\theta) \tag{3.1}$$

for any $x \in \mathcal{X}$, where $Z : \Theta \to \mathbb{R}$ is the partition function $Z(\theta) = \int \exp(\theta \cdot T) d\mu$. Assume that Θ is a non-empty, open subset of $\mathbb{R}^d$ and that T is full rank, in the sense that the image of T is not contained in any $(d-1)$-dimensional hyperplane in $\mathbb{R}^d$. Then $\mathcal{M} = \{p_\theta \mu \mid \theta \in \Theta\}$

is a regular exponential family of order d with dominating measure μ and canonical sufficient statistic T, and all regular exponential families are of this form [3, §8.1]. Note that each element of $\mathcal{M}$ is a probability measure on $\mathbb{R}^m$.

3.2. **IID extensions.** The n-fold IID extension $\mathcal{M}^n$ of $\mathcal{M}$ is the set $\mathcal{M}^n = \{P^n \mid P \in \mathcal{M}\}$ of all measures of the form P^n for some $P \in \mathcal{M}$, where $P^n = P \times \cdots \times P$ (with n copies of P) is the product measure on $\mathcal{X}^n$ [8, §3.3]. In terms of the parameterisation (3.1), $\mathcal{M}^n$ is the set of all measures of the form $p_\theta^{(n)}\mu^n$ for some $\theta \in \Theta$, where $p_\theta^{(n)} : \mathcal{X}^n \to \mathbb{R}_{>0}$ is given by $p_\theta^{(n)}(x_1, \ldots, x_n) = p_\theta(x_1) \ldots p_\theta(x_n)$ and $\mu^n = \mu \times \cdots \times \mu$ is the product measure on $\mathcal{X}^n$ [3, Example 8.12(ii)]. So by (3.1),

$$p_\theta^{(n)} = \exp(n\theta \cdot T_n - n \log Z(\theta)), \tag{3.2}$$

where $T_n : \mathcal{X}^n \to \mathbb{R}^d$ is given by $T_n(x_1, \ldots, x_n) = (T(x_1) + \cdots + T(x_n))/n$ for any $x_1, \ldots, x_n \in \mathcal{X}$. Therefore $\mathcal{M}^n$ is an exponential family with dominating measure μ^n and sufficient statistic T_n (and natural parameter $n\theta$, see [13, Thm. 2.2.6]). Note that $\mathcal{M}^1 = \mathcal{M}$, $T_1 = T$ and $p_\theta^{(1)} = p_\theta$.

3.3. **Natural exponential families.** Recall that if $\mathcal{Y}$ and $\mathcal{Z}$ are measurable spaces, $\phi : \mathcal{Y} \to \mathcal{Z}$ is a measurable function and P is a measure on $\mathcal{Y}$ then the push-forward of P via ϕ is the measure $\phi_* P$ on $\mathcal{Z}$ given by

$$(\phi_* P)(U) = P(\phi^{-1}(U)) \tag{3.3}$$

for any measurable set U in $\mathcal{Z}$ [8, §3.6]. This immediately implies that if Y is a $\mathcal{Y}$-valued random variable with distribution P then $\phi(Y)$ is a $\mathcal{Z}$-valued random variable with distribution $\phi_* P$, which in symbols we write as

$$Y \sim P \text{ implies } \phi(Y) \sim \phi_* P. \tag{3.4}$$

Then the natural exponential family corresponding to $\mathcal{M}^n$ and T_n is the set $\mathcal{N}_n = \{T_{n*}P^n \mid P^n \in \mathcal{M}^n\}$ of measures on $\mathbb{R}^d$. By [3, Examples 8.12(ii) and 8.12(iii)], $\mathcal{N}_n = \{q_\theta^n \nu_n \mid \theta \in \Theta\}$, where ν_n is a measure on $\mathbb{R}^d$ that does not depend on θ and $q_\theta^n : \mathbb{R}^d \to \mathbb{R}_{>0}$ is given by

$$q_\theta^n(y) = \exp(n\theta \cdot y - n \log Z(\theta)) \tag{3.5}$$

for any $y \in \mathbb{R}^d$. The formula (3.5) shows that the superscript in q_θ^n is actually an exponent, so we will write q_θ for q_θ^1 (and then the notation q_θ^n is unambiguous).

Note that even though $\mathcal{M}, \mathcal{M}^2, \mathcal{M}^3, \ldots$ and $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \ldots$ are families of measures on different spaces (namely, $\mathcal{X}, \mathcal{X}^2, \mathcal{X}^3, \ldots$ and $\mathbb{R}^d, \mathbb{R}^d, \mathbb{R}^d, \ldots$, respectively), they are all parameterised by $\Theta \subseteq \mathbb{R}^d$ so they are all d-dimensional families of measures.

3.4. **The family of Bernoulli distributions.** To illustrate the general framework above, let $\mathcal{M}$ be the family of all Bernoulli distributions. This is the 1-dimensional exponential family with data space $\mathcal{X} = \{0, 1\}$, the counting measure on $\mathcal{X}$ as its dominating measure, the log-odds θ as a natural parameter, canonical sufficient statistic $T : \mathcal{X} \to \mathbb{R}$ given by $T(x) = x$ and partition function Z on the natural parameter space $\Theta = \mathbb{R}$ given by $Z(\theta) = 1 + e^\theta$ (though this description of $\mathcal{M}$ is not unique). So by (3.1), the distribution P in $\mathcal{M}$ corresponding to θ puts a mass of $p_\theta(1) = e^\theta/(1 + e^\theta)$ on the data-point $1 \in \mathcal{X}$ (so θ is the log-odds, as claimed). Then the n-fold IID extension $\mathcal{M}^n$ of $\mathcal{M}$ has data space $\mathcal{X}^n = \{0, 1\}^n$ consisting of all binary sequences of length n and the distribution P^n in $\mathcal{M}^n$ corresponding to θ puts a mass of

$$p_\theta(1)^{n\bar{x}}(1 - p_\theta(1))^{n - n\bar{x}}$$

on $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, where $\bar{x} = (\sum_{i=1}^n x_i)/n$ and $p_\theta(1) = e^\theta/(1+e^\theta)$, as above. Lastly, the sufficient statistic T_n is given by $T_n(x) = \bar{x}$ for any $x \in \mathcal{X}^n$, so the distribution Q_n in $\mathcal{N}_n$ corresponding to θ is the binomial distribution $\mathrm{Bin}(n, p_\theta(1))$ up to a linear transformation of the data space (so that Q_n has support $\{0, 1/n, 2/n, \ldots, 1\}$ instead of $\{0, 1, 2, \ldots, n\}$). Since $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$ are all parameterised by $\Theta = \mathbb{R}$, they are all 1-dimensional statistical models.

4. Invariance and regularity conditions

Let $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$ be as in Section 3 and suppose now that these spaces have been equipped with Riemannian metrics g, g^n and g_n, respectively. In this section, we will give precise conditions that formalize the notion of these metrics being invariant under IID extensions and canonical sufficient statistics, as well as giving a mild regularity condition. These conditions will then be used in Section 5 to prove our main theorem. See Section 4.4 for a number of remarks about these assumptions.

Assumptions 1. *We make the following assumptions, which are described precisely in the subsections below:*

- A1 *The metrics g and g^n are invariant under IID extensions (up to a factor of n)*
- A2 *The metrics g^n and g_n are invariant under canonical sufficient statistics*
- A3 *The norms corresponding to the metrics g_n can all be calculated by a function that satisfies a weak continuity condition*

4.1. **A1: Invariance under IID extensions.** Let $IID_n : \mathcal{M} \to \mathcal{M}^n$ be the function that maps each $P \in \mathcal{M}$ to the product measure $P^n = P \times \cdots \times P$ (see Section 3.2). Then our first assumption is that this map is an isometry (i.e., distance-preserving map) up to a factor of n.

More precisely, let $u = (\theta, a) \in T\Theta$ be any tangent vector to Θ, as in Section 2. Then similarly to (2.2), u corresponds under the smooth parameterisation (3.2) to a tangent vector $\tilde{u}^n$ to $\mathcal{M}^n$, where $\tilde{u}^n = (P^n, A^{(n)})$, $P^n = p_\theta^{(n)}\mu^n$ and $A^{(n)} = \sum_{i=1}^d a_i(\partial p_\theta^{(n)}/\partial\theta_i)\mu^n$. Let $T\mathcal{M}^n = \{\tilde{u}^n \mid u \in T\Theta\}$ be the set of all such tangent vectors to $\mathcal{M}^n$. Then our first assumption is that

$$g^n(\tilde{u}^n, \tilde{v}^n) = ng(\tilde{u}, \tilde{v}) \tag{4.1}$$

for all tangent vectors $u, v \in T\Theta$ with the same base-point. Here, $\tilde{v}$ and $\tilde{v}^n$ are the tangent vectors to $\mathcal{M}$ and $\mathcal{M}^n$ (respectively) corresponding to $v \in T\Theta$, as for u above. Note that (4.1) just says that $g^n = ng$ under the identification of $\mathcal{M}$ with $\mathcal{M}^n$ via IID_n.

The Fisher information metric is invariant under IID extensions in the sense of (4.1) by [1, eq. 4.2], so assumptions (A1)–(A3) cannot characterize the Fisher information metric unless the factor of n is included in (4.1) (though see Remark 4).

4.2. **A2: Invariance under canonical sufficient statistics.** Let $T_n : \mathcal{X}^n \to \mathbb{R}^d$ be the canonical sufficient statistic from Section 3.2 and let $T_{n*} : \mathcal{M}^n \to \mathcal{N}_n$ be the corresponding (measure-theoretic) push-forward map of T_n, see Section 3.3. Then our second assumption is that this map T_{n*} is an isometry (and that all other canonical sufficient statistics are isometries, in a sense that will be made precise in Section 4.3).

More precisely, let $u = (\theta, a) \in T\Theta$ be any tangent vector to Θ, as in Section 2. Then similarly to (2.2), u corresponds under the smooth parameterisation (3.5) to a tangent vector $\tilde{u}_n = (Q_n, A_n)$ to $\mathcal{N}_n$, where

$$Q_n = q_\theta^n \nu_n \text{ and } A_n = \sum_{i=1}^d a_i(\partial q_\theta^n/\partial\theta_i)\nu_n. \tag{4.2}$$

Let $T\mathcal{N}_n = \{\tilde{u}_n \mid u \in T\Theta\}$ be the set of all such tangent vectors. Then our second assumption is that

$$g_n(\tilde{u}_n, \tilde{v}_n) = g^n(\tilde{u}^n, \tilde{v}^n) \tag{4.3}$$

for all tangent vectors $u, v \in T\Theta$ with the same base-point. Here, $\tilde{v}^n$ and $\tilde{v}_n$ are the tangent vectors to $\mathcal{M}^n$ and $\mathcal{N}_n$ (respectively) corresponding to $v \in T\Theta$, as for u above. Note that (4.3) just says that $g_n = g^n$ under the identification of $\mathcal{M}^n$ with $\mathcal{N}_n$ via T_{n*}.

4.3. **A3: Calculability of norms by a function that satisfies a weak continuity condition.** Let h be the norm corresponding to g, so $h(\tilde{u}) = \sqrt{g(\tilde{u}, \tilde{u})}$ for any $\tilde{u} \in T\mathcal{M}$. Note that h determines g by the polarisation formula,

$$g(\tilde{u}, \tilde{v}) = \left[h^2(\tilde{u} + \tilde{v}) - h^2(\tilde{u} - \tilde{v})\right]/4$$

for any $\tilde{u}, \tilde{v} \in T\mathcal{M}$ with the same base-point (which follows from the bilinearity of g), so any question about g can be phrased in terms of h. However, it will be more convenient to work with h than g, because h is a function defined on $T\mathcal{M}$, whereas g is only defined on certain pairs of tangent vectors (those with the same base-point). Similarly, let h_n be the norm corresponding to g_n, so $h_n(\tilde{u}_n) = \sqrt{g_n(\tilde{u}_n, \tilde{u}_n)}$ for any $\tilde{u}_n \in T\mathcal{N}_n$.

Let $\mathcal{T}'$ be the set of all pairs (P, A), where P is a probability measure on $\mathbb{R}^d$ and A is a signed measure on $\mathbb{R}^d$, and note that $T\mathcal{N}_n \subseteq \mathcal{T}'$ for every n. Then our regularity condition (A3) is, partly, that there is subset $\mathcal{T}$ of $\mathcal{T}'$ and a function $H : \mathcal{T} \to \mathbb{R}$ so that, for each n, $T\mathcal{N}_n \subseteq \mathcal{T}$ (i.e. H is defined on each $T\mathcal{N}_n$) and

$$h_n(\tilde{u}_n) = H(\tilde{u}_n) \tag{4.4}$$

for every $\tilde{u}_n \in T\mathcal{N}_n$. In other words, we assume that there is some function H whose restriction to each $T\mathcal{N}_n$ is the norm h_n. For instance, we could take $\mathcal{T} = \cup_{n=1}^{\infty} T\mathcal{N}_n$ and then define H by the requirement that (4.4) holds, which gives a well-defined H whenever the functions h_n agree on any overlaps between the spaces $T\mathcal{N}_n$.

Further, we assume that H has the following weak continuity property. Firstly, we require that H is defined on all pairs of the form $(\Phi, f\Phi)$, where Φ is the probability measure for the standard normal distribution on $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ is a linear function (with $f(0) = 0$). Secondly, we require that

$$H(P_n, fP_n) = H(\Phi, f\Phi) \tag{4.5}$$

for any sequence P_n of probability measures on $\mathbb{R}^d$ for which $H(P_n, fP_n)$ is constant in n, $P_n \Rightarrow \Phi$ and each P_n is standardized (i.e., P_n has 0 mean and identity variance-convariance matrix), where $H(P_n, fP_n)$ is the value of the function H at $(P_n, fP_n) \in \mathcal{T}$ and $P_n \Rightarrow \Phi$ means P_n converges to Φ in the sense of the weak convergence of measures [16, Def. 1.2.1]. This condition is an extremely weak form of continuity, see Remark 1.

Lastly, as a consequence of our assumption (A2) that the metrics should be invariant under all canonical sufficient statistics, we assume that H is affine invariant (see Remark 6). Here, an invertible affine transformation of $\mathbb{R}^d$ is a map $L : \mathbb{R}^d \to \mathbb{R}^d$ of the form $L(x) = Mx + c$ for some invertible $d \times d$ matrix M and some $c \in \mathbb{R}^d$. The push-forward L_*A of any signed measure A on $\mathbb{R}^d$ is defined in a similar way to the push-forward of an (unsigned) measure, see (3.3). We define the push-forward $L_{**}(P, A)$ of any $(P, A) \in \mathcal{T}$ to be $L_{**}(P, A) = (L_*P, L_*A)$. (In this notation, L_* is the measure-theoretic push-forward, which is a map from the space of signed measures on $\mathbb{R}^d$ to itself, and L_{**} is the differential of this map if (P, A) is interpreted

as a tangent vector.) Then our condition that H is affine invariant means that $L_{**}(P, A) \in \mathcal{T}$ and

$$H(L_{**}(P, A)) = H(P, A) \tag{4.6}$$

for every $(P, A) \in \mathcal{T}$ and every invertible affine transformation L of $\mathbb{R}^d$.

For future reference, we note that if L is an invertible affine transformation, P is a probability measure and f is a P-integrable, real-valued function then

$$L_*(fP) = (f \circ L^{-1})L_*P \tag{4.7}$$

by the change of variables formula [8, Thm. 3.6.1].

4.4. **Remarks on the assumptions.**

Remark 1. *Assumptions (A1) and (A2) say that the metrics on $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$ are invariant under a countable set of transformations and, in a certain sense, under the finite-dimensional group of affine transformations of $\mathbb{R}^d$. The third assumption (A3) is an extremely weak form of continuity. Firstly, this condition says that the norms h_n agree on any overlaps between the spaces $T\mathcal{N}_n$, so that these functions can be pieced together into a single function H. Secondly, this condition says that if f is linear and $P_n \Rightarrow \Phi$ is a sequence for which (P_n, fP_n) all have the same norms then this shared norm must be $H(\Phi, f\Phi)$. By comparison, full continuity of H would require that $\lim_{n\to\infty} H(P_n, f_nP_n) = H(P, fP)$ for every sequence (P_n, f_nP_n) in $\mathcal{T}$ that converges to (P, fP) (with respect to some notion of convergence). So our third assumption is the condition for the continuity of H in the very special case where $P = \Phi$, $H(P_n, f_nP_n)$ is constant in n, $f_n = f$ for every n and f is a linear function.*

Remark 2. *Recent versions of Chentsov's theorem [2, 7] consider metrics on infinite-dimensional statistical models that are invariant under infinite-dimensional sets of transformations. This infinite dimensionality introduces technical complications and it makes strong assumptions about both the space on which the metric is defined and its symmetries. By contrast, our approach allows us to only consider metrics on a collection of finite-dimensional models, as in the original version of Chentsov's theorem [10, 20, 9]. This allows our characterisation of the Fisher information metric to be relatively free from technicalities and it allows us to make relatively weak invariance and regularity assumptions.*

Remark 3. *It is not hard to see that the Fisher information metric satisfies assumptions (A1)–(A3). For it is well known that the Fisher information metric is invariant under both IID extensions (in the sense of (4.1)) and sufficient statistics [1, eq. 4.2 and Thm. 2.1]. Also, given any probability measure P on $\mathbb{R}^d$, let $\mathcal{T}_P = \{(P, fP) \mid f \in L^2(\mathbb{R}^d, P)\}$, and let $\mathcal{T}$ be the union of these spaces $\mathcal{T}_P$ as P ranges over the set of all probability measures on $\mathbb{R}^d$. Then by (2.3), the Fisher information norm $H^F(P, fP)$ of any $(P, fP) \in \mathcal{T}$ is just the $L^2(\mathbb{R}^d, P)$-norm of f. So if f is a linear function on $\mathbb{R}^d$, say $f(y) = c \cdot y$ for some $c \in \mathbb{R}^d$, and Q is any standardized probability measure on $\mathbb{R}^d$ then*

$$H^F(Q, fQ) = \sqrt{\int (c \cdot y)^2 dQ(y)} = \sqrt{c^T \left(\int yy^T dQ(y) \right) c} = \sqrt{c^T I c} = \|c\|,$$

where $\|c\|$ is the Euclidean norm of $c \in \mathbb{R}^d$. So for any sequence P_n of standardized probability measures (whether weakly convergent to Φ or not), $H^F(P_n, fP_n) = \|c\| = H^F(\Phi, f\Phi)$, so H^F satisfies the weak continuity condition (4.5). Also, this function H^F is affine invariant (4.6) by the change of variables formula (4.7).

Remark 4. *In some ways the factor of n in (4.1) is not essential, since we could instead formulate our assumptions and theorems in terms of the metrics $\dot{g}^n = g^n/n$ and $\dot{g}_n = g_n/n$, in which case (4.1) would be equivalent to the equation that describes exact invariance under the map IID_n, rather than invariance up to a factor of n (though H as in (4.4) might not exist without the factor of n). However, it is natural to include the factor of n in our formulation of IID invariance, firstly because the Fisher information metric is IID invariant in the sense of (4.1) [1, eq. 4.2], so assumptions (A1)–(A3) would not characterise the Fisher information metric without this factor, and secondly because the factor of n arises from a natural construction from differential geometry (see Remark 5).*

Remark 5. *Given an arbitrary Riemannian metric g on $\mathcal{M}$, a natural construction from differential geometry gives a metric on the n-fold IID extension $\mathcal{M}^n$ of $\mathcal{M}$ equal to the metric g^n satisfying (4.1), as follows. The Cartesian product $\prod^n \mathcal{M}$ of $\mathcal{M}$ with itself n times is the space whose points are n-tuples $(P_1, \dots, P_n)$ of measures $P_1, \dots, P_n \in \mathcal{M}$ on $\mathcal{X}$. Given such an n-tuple, there is a corresponding product measure $P_1 \times \cdots \times P_n$ on $\mathcal{X}^n$, and conversely we can recover each P_i from $P_1 \times \cdots \times P_n$ by marginalizing, so we can identify $(P_1, \dots, P_n)$ with the product measure $P_1 \times \cdots \times P_n$ on $\mathcal{X}^n$. This product measure is the joint distribution of independent random variables $X_1, \dots, X_n$ whose marginal distributions are $P_1, \dots, P_n$, respectively. So if $(P_1, \dots, P_n) \in \prod^n \mathcal{M}$ satisfies $P_1 = \cdots = P_n$ then $P_1 \times \cdots \times P_n$ is the joint distribution of IID random variables $X_1, \dots, X_n$. Therefore we can identify the diagonal*

$$\Delta = \left\{ (P_1, \dots, P_n) \in \prod^n \mathcal{M} \,\middle|\, P_1 = \cdots = P_n \right\}$$

of $\prod^n \mathcal{M}$ with the n-fold IID extension $\mathcal{M}^n$ of $\mathcal{M}$. But a Riemannian metric on $\mathcal{M}$ induces a Riemannian metric on the Cartesian product $\prod^n \mathcal{M}$, and then Δ inherits a metric from $\prod^n \mathcal{M}$. Under the above identification between Δ and $\mathcal{M}^n$, this metric is the metric g^n on $\mathcal{M}^n$ that satisfies (4.1).

Remark 6. *The canonical sufficient statistics for an exponential family are only unique up to affine transformations [3, Lemma 8.1], meaning that if L is an invertible affine transformation of $\mathbb{R}^d$ and $T_n : \mathcal{X}^n \to \mathbb{R}^d$ is a canonical sufficient statistic then $L \circ T_n$ is also a canonical sufficient statistic (and every canonical sufficient statistic is of this form). Replacing T_n by $L \circ T_n$ effectively replaces each tangent vector $\tilde{u}_n \in T\mathcal{N}_n$ by $L_{**}\tilde{u}_n$, so (4.3), (4.4) and the analogous equations for $L \circ T_n$ imply $H(L_{**}\tilde{u}_n) = H(\tilde{u}_n)$ for every $\tilde{u}_n \in T\mathcal{N}_n$. So since L is arbitrary, H is affine invariant.*

5. The main theorem

We can now prove our version of Chentsov's theorem. This theorem characterises the Fisher information metric as the only metric (up to a multiplicative constant) on an exponential family that is invariant under IID extensions and canonical sufficient statistics.

Let g^F, g^{nF} and g^F_n be the Fisher information metrics on $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$, respectively.

Theorem 5.1. *Suppose that assumptions (A1)–(A3) of Section 4 hold. Then there is some $c > 0$ so that $g = cg^F$, $g^n = cg^{nF}$ and $g_n = cg^F_n$ for every integer $n \geq 1$.*

Proof. Let any integer $n \geq 1$ and any $\theta \in \Theta$ be given, and let $Q_1 = q_\theta \nu_1 \in \mathcal{N}_1$ and $Q_n = q^n_\theta \nu_n \in \mathcal{N}_n$ be the corresponding distributions in $\mathcal{N}_1$ and $\mathcal{N}_n$. By Theorem 2.2.6 of [13] and the comments preceding it, if $Y_1, \dots, Y_n$ are independent random variables all distributed

according to Q_1 then their mean is distributed as Q_n, which we write as

$$(Y_1 + \cdots + Y_n)/n \sim Q_n. \tag{5.1}$$

Alternatively, it is not hard to prove (5.1). For if $X_1, \ldots, X_n \sim P$ are IID, where $P = p_\theta \mu$, and if $Y_i' = T(X_i)$ then $Y_1', \ldots, Y_n' \sim Q_1$ are IID, since $Q_1 = T_* P$ by definition. Therefore $Y_1, \ldots, Y_n$ and $Y_1', \ldots, Y_n'$ both have the same joint distribution, so their means have the same distribution, by an application of (3.4). But $(Y_1' + \cdots + Y_n')/n = T_n(X_1, \ldots, X_n) \sim Q_n$, by (3.4) and since $Q_n = T_{n*} P^n$ by definition, so (5.1) follows.

By (5.1), the mean τ_θ for Q_1 is the same as that for Q_n, i.e.

$$\tau_\theta = \int y dQ_1(y) = \int y dQ_n(y), \tag{5.2}$$

and the variance-covariance matrix Σ_θ for Q_1 is n times that for Q_n, i.e.

$$\Sigma_\theta = \int (y - \tau_\theta)(y - \tau_\theta)^T dQ_1(y) = n \int (y - \tau_\theta)(y - \tau_\theta)^T dQ_n(y). \tag{5.3}$$

Now, let $u = (\theta, a) \in T_\theta \Theta$ be any tangent vector to Θ at θ, and define $f : \mathbb{R}^d \to \mathbb{R}$ by $f(y) = (\Sigma_\theta^{1/2} a) \cdot y$ for any $y \in \mathbb{R}^d$. Here, the square root $\Sigma_\theta^{1/2}$ is defined in the standard way via a diagonalisation of the symmetric, positive-definite matrix Σ_θ. As before, let $\tilde{u}$ and $\tilde{u}_n$, respectively, be the tangents to $\mathcal{M}$ and $\mathcal{N}_n$ that correspond to u under the parameterisations (3.2) and (3.5).

Claim 1: $h(\tilde{u}) = H(\Phi, f\Phi)$. By (3.5), (4.2) and the fact that τ_θ is the gradient of $\log Z$ at θ [13, Thm. 2.2.1], $\tilde{u}_n = (Q_n, A_n)$ with $Q_n = q_\theta^n \nu_n$ and

$$A_n = \sum_{i=1}^d a_i \frac{\partial q_\theta^n}{\partial \theta_i} \nu_n = \sum_{i=1}^d a_i n \left(\iota_i - \frac{\partial \log Z}{\partial \theta_i} \right) q_\theta^n \nu_n = na \cdot (\iota - \tau_\theta) Q_n, \tag{5.4}$$

where $\iota_i(y) = y_i$ and $\iota(y) = y$ for any $y \in \mathbb{R}^d$.

Let L be the affine transformation on $\mathbb{R}^d$ given by $L(y) = \sqrt{n} \Sigma_\theta^{-1/2} (y - \tau_\theta)$, and note that the inverse $\Sigma_\theta^{-1/2}$ exists because Σ_θ is positive-definite. By (5.2) and (5.3), this choice of L ensures that $L_* Q_n$ is standardised, i.e., that $L_* Q_n$ has mean 0 and variance-covariance matrix equal to the $d \times d$ identity matrix. Note that L depends on n, so we could instead write this as L_n, but for notational simplicity we will drop the subscript. Then by (4.7) and (5.4),

$$L_* A_n = na \cdot (\iota \circ L^{-1} - \tau_\theta) L_* Q_n = \sqrt{n} f L_* Q_n, \tag{5.5}$$

where f is as in the statement of the claim.

So recalling the notation $L_{**} \tilde{u}_n = L_{**}(Q_n, A_n) = (L_* Q_n, L_* A_n)$, we have

$$\begin{aligned} h(\tilde{u}) &= n^{-1/2} h_n(\tilde{u}_n) \text{ by (4.1) and (4.3)} \\ &= h_n(n^{-1/2} \tilde{u}_n) \text{ by the bilinearity of } g_n \\ &= H(n^{-1/2} \tilde{u}_n) \text{ by (4.4)} \\ &= H(n^{-1/2} L_{**} \tilde{u}_n) \text{ by (4.6)} \\ &= H(L_* Q_n, f L_* Q_n) \text{ by (2.1) and (5.5).} \end{aligned} \tag{5.6}$$

By (5.1), the central limit theorem (e.g. see [16, Cor. 8.1.10]) and the fact that L_*Q_n is standardised, $L_*Q_n \Rightarrow \Phi$. Therefore,

$$\begin{aligned} h(\tilde{u}) &= H(L_*Q_n, fL_*Q_n) \text{ for all } n, \text{ by (5.6)} \\ &= H(\Phi, f\Phi) \text{ by (4.5)}, \end{aligned} \tag{5.7}$$

so the claim is proved.

Now, let $v = (\phi, b) \in T\Theta$ be any tangent vector to Θ, not necessarily with the same base-point as u, and let $\tilde{v} \in T\mathcal{M}$ be the corresponding tangent vector to $\mathcal{M}$.

Claim 2: $a^T\Sigma_\theta a = b^T\Sigma_\phi b$ implies $h(\tilde{u}) = h(\tilde{v})$. To prove this, assume that $a^T\Sigma_\theta a = b^T\Sigma_\phi b$, i.e. that $\Sigma_\theta^{1/2}a$ and $\Sigma_\phi^{1/2}b$ have the same Euclidean norm. Then there exists a $d \times d$ orthogonal matrix M so that

$$M\Sigma_\theta^{1/2}a = \Sigma_\phi^{1/2}b. \tag{5.8}$$

Also, $M_*\Phi = \Phi$ because M is orthogonal, so

$$M_{**}(\Phi, f\Phi) = (M_*\Phi, M_*(f\Phi)) = (M_*\Phi, (f \circ M^{-1})M_*\Phi) = (\Phi, e\Phi) \tag{5.9}$$

by (4.7), where $e : \mathbb{R}^d \to \mathbb{R}$ is given by

$$e(y) = f(M^{-1}(y)) = (\Sigma_\theta^{1/2}a) \cdot M^{-1}y = (\Sigma_\theta^{1/2}a)^T M^{-1}y = (\Sigma_\phi^{1/2}b) \cdot y \tag{5.10}$$

for any $y \in \mathbb{R}^d$, by (5.8) and $M^{-1} = M^T$ (since M is orthogonal). So

$$\begin{aligned} h(\tilde{v}) &= H(\Phi, e\Phi) \text{ by Claim 1 applied to } v \text{ and by (5.10)} \\ &= H(M_{**}(\Phi, f\Phi)) \text{ by (5.9)} \\ &= H(\Phi, f\Phi) \text{ by (4.6)} \\ &= h(\tilde{u}) \text{ by Claim 1}, \end{aligned}$$

which proves Claim 2.

Claim 3: There is some $c > 0$ so that $h(\tilde{v}) = c\, h^F(\tilde{v})$ for all tangent vectors $\tilde{v} \in T\mathcal{M}$. It is well known [13, Thms. 2.2.1 and 2.2.5] that the Fisher information metric on the natural parameter space is the variance-covariance matrix of the corresponding sufficient statistic, so $g^F(\tilde{u}, \tilde{u}) = a^T\Sigma_\theta a$. Alternatively, this follows easily from setting $n = 1$ in (5.4) and combining this with (2.3) and the invariance of g^F under sufficient statistics [1, Thm. 2.1], since these give

$$g^F(\tilde{u}, \tilde{u}) = g_1^F(\tilde{u}_1, \tilde{u}_1) = a^T \left(\int (y - \tau_\theta)(y - \tau_\theta)^T dQ_1(y) \right) a = a^T\Sigma_\theta a, \tag{5.11}$$

where $\tilde{u}_1 \in T\mathcal{N}_1$ is the tangent vector to $\mathcal{N}_1$ corresponding to $u \in T\Theta$. So Claim 2 is equivalent to

$$h^F(\tilde{u}) = h^F(\tilde{v}) \text{ implies } h(\tilde{u}) = h(\tilde{v}), \tag{5.12}$$

for all tangent vectors $\tilde{u}, \tilde{v} \in T\mathcal{M}$, even if they have different base-points.

Now, fix $\tilde{u}$ to be some non-zero vector with $h^F(\tilde{u}) = 1$, and let $c = h(\tilde{u})$. Note that $c > 0$ because g is an inner product on each tangent space so the norm of any non-zero tangent vector is strictly positive. Then for any non-zero $\tilde{v}$, $h^F(\tilde{v}/h^F(\tilde{v})) = h^F(\tilde{v})/h^F(\tilde{v}) = 1$ by the bilinearity of g^F. So $h^F(\tilde{u}) = h^F(\tilde{v}/h^F(\tilde{v}))$ and hence, by (5.12), $h(\tilde{u}) = h(\tilde{v}/h^F(\tilde{v}))$. Therefore $c = h(\tilde{u}) = h(\tilde{v}/h^F(\tilde{v})) = h(\tilde{v})/h^F(\tilde{v})$ by the bilinearity of g, so rearranging this equation proves the claim for all non-zero tangent vectors $\tilde{v} \in T\mathcal{M}$. But the claim holds trivially for any zero tangent vector $\tilde{v}$, since $0 = h(\tilde{v}) = h^F(\tilde{v})$ by the bilinearity of g and g^F, so the claim is proved.

The theorem now follows from Claim 3 and by (4.1), (4.3) and the analogous equations for the Fisher information metrics g^F, g^{nF} and g^F_n, which hold by [1, eq. 4.2 and Thm. 2.1]. □

6. Extensions

6.1. **Extensions to curved exponential families.** The proof of Theorem 5.1, without any essential changes, also characterises the Fisher information metric on curved exponential families.

For suppose that $\widetilde{\mathcal{M}}$ is a submanifold of the exponential family $\mathcal{M}$ of Section 3, meaning that $\widetilde{\mathcal{M}} = \{p_\theta \mu \mid \theta \in \widetilde{\Theta}\}$, where $\widetilde{\Theta}$ is a submanifold of the natural parameter space $\Theta \subseteq \mathbb{R}^d$ of $\mathcal{M}$ (so $\widetilde{\Theta}$ is either an open subset of Θ or a submanifold of lower dimension). The tangent bundle $T\widetilde{\Theta}$ of $\widetilde{\Theta}$ is usually more complicated than $T\Theta$ (see standard textbooks such as [13, Appendix C] or [14]) but the tangent vectors to $\widetilde{\Theta}$ are a subset of the tangent vectors to Θ. So given any $\theta \in \widetilde{\Theta}$ and a tangent vector $u \in T_\theta\widetilde{\Theta}$, there is a corresponding tangent vector $\tilde{u}$ to $\widetilde{\mathcal{M}}$ given by (2.2). The n-fold IID extension $\widetilde{\mathcal{M}}^n$ of $\widetilde{\mathcal{M}}$ is defined as in Section 3.2, and $\widetilde{\mathcal{M}}^n$ is clearly a submanifold of $\mathcal{M}^n$ (since it can be parameterised by $\widetilde{\Theta}$). The natural exponential family $\widetilde{\mathcal{N}}_n$ is defined as in Section 3.3, and it is again clear that $\widetilde{\mathcal{N}}_n$ is a submanifold of $\mathcal{N}_n$. Assumptions (A1) and (A2) then just become equations (4.1) and (4.3) for all tangent vectors $u, v \in T\widetilde{\Theta}$ with the same base-point. Assumption (A3) is also the same except that $T\widetilde{\mathcal{N}}_n$ replaces $T\mathcal{N}_n$, so the weak continuity property (4.5) is essentially unchanged. Also, given any invertible affine transformation L of $\mathbb{R}^d$, if $T_n : \mathcal{X}^n \to \mathbb{R}^d$ is a canonical sufficient statistic then so is $L \circ T_n$, hence H is affine invariant (as in Remark 6).

The statement of Theorem 5.1 is exactly the same, though all metrics are now understood to be on $\widetilde{\mathcal{M}}$, $\widetilde{\mathcal{M}}^n$ and $\widetilde{\mathcal{N}}_n$ rather than $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$. For any $Q_1 \in \widetilde{\mathcal{M}}_1$ and any IID $Y_1, \ldots, Y_n \sim Q_1$, we have

$$(Y_1 + \cdots + Y_n)/n \sim Q_n,$$

as in (5.1). Therefore the mean and variance formulas ((5.2) and (5.3)) hold, and we can also apply the central limit theorem to prove Claim 1 for any $u \in T\widetilde{\Theta}$. Claim 2 then follows for any tangent vector $v \in T\widetilde{\Theta}$. Lastly, (5.11) holds for any $\tilde{u}$ in $T\mathcal{M}$ and hence for any $\tilde{u}$ in the subset $T\widetilde{\mathcal{M}}$ of $T\mathcal{M}$, so Claim 3 follows and hence so does the theorem.

6.2. **Extensions to higher-order symmetric tensors.** The proof of Theorem 5.1 also extends with almost no changes to characterise symmetric, order-k tensors $\hat{g}$ and $\hat{g}_n$ on $\mathcal{M}$ and $\mathcal{N}_n$, respectively, that satisfy conditions closely analogous to assumptions (A1)–(A3) of Section 4. Given such tensors $\hat{g}_n$, define $\hat{h}_n(\tilde{u}_n) = \sqrt[k]{\hat{g}_n(\tilde{u}_n, \ldots, \tilde{u}_n)}$, where there are k copies of $\tilde{u}_n$ in the right-hand side of this equation. Assume that

$$\hat{g}_n(\tilde{u}_n, \ldots, \tilde{u}_n) = n^{k/2} \hat{g}_1(\tilde{u}_1, \ldots, \tilde{u}_1), \tag{6.1}$$

which is a generalisation of (4.1) from $k = 2$ to general k. Then as in the proof of Theorem 5.1, $\hat{h}_n(\tilde{u}_n) = \sqrt{n}\hat{h}_1(\tilde{u}_1)$ and $\hat{h}_n(\alpha\tilde{u}_n) = \alpha\hat{h}_n(\tilde{u}_n)$ for any $\alpha \geq 0$ (by (6.1) and the multi-linearity of $\hat{g}_n$). So with $\hat{h}$ in place of h, the proof of Theorem 5.1 implies that $\hat{h}(\tilde{u}) = c\, h^F(\tilde{u})$ for some $c \in \mathbb{R}$, where h^F is the norm of the Fisher information metric. Raising this equation to the power of k gives

$$\hat{g}(\tilde{u}, \ldots, \tilde{u}) = c^k \left[g^F(\tilde{u}, \tilde{u})\right]^{k/2}. \tag{6.2}$$

If k is odd then the left-hand side is an odd function of $\tilde{u}$ (i.e. it changes sign when $\tilde{u}$ is replaced by $-\tilde{u}$) while the right-hand side is an even function, which is a contradiction unless

both sides vanish, so $c = 0$. If k is even, then since $\hat{g}$ is determined by (6.2) (by the polarisation formula for symmetric tensors), $\hat{g}$ must be a constant times the symmetric part of $(g^F)^{k/2}$. For example, when $k = 4$ then there is some $c' \in \mathbb{R}$ so that

$$\hat{g}(\tilde{u}, \tilde{v}, \tilde{w}, \tilde{m}) = c' \left[g^F(\tilde{u}, \tilde{v}) g^F(\tilde{w}, \tilde{m}) + g^F(\tilde{u}, \tilde{w}) g^F(\tilde{v}, \tilde{m}) + g^F(\tilde{u}, \tilde{m}) g^F(\tilde{v}, \tilde{w}) \right]$$

for any $\tilde{u}, \tilde{v}, \tilde{w}, \tilde{m} \in T\mathcal{M}$.

Remark 7. *It might also be possible to adapt the proof of Theorem 5.1 to characterise the higher-order Amari-Chentsov tensors, which are symmetric, order-k tensors that coincide with the Fisher information metric when $k = 2$ and in general are given by an equation similar to (2.3), e.g. see [2, eq. 2.4] for the $k = 3$ case. Claim 1 in the proof of Theorem 5.1 does not seem to hold for these tensors in general. However, if we replace the $k/2$ in (6.1) by other powers and strengthen the weak continuity condition on H then it might be possible to replace Claim 1 by $\hat{h}(\tilde{u}) = H(K\Phi, fK\Phi)$, where K is an Edgeworth polynomial (see [4] or [13, §4.5]). Then a symmetry argument, similar to the one in the proof of Theorem 5.1, should give the desired characterisation.*

7. Comparison to previous versions of Chentsov's theorem

The original version of Chentsov's theorem applied to statistical models with finite data spaces, and these models are all (curved) exponential families. So in this section we compare our version of Chentsov's theorem to the original version.

The original version of Chentsov's theorem [10, 20, 9] characterises the Fisher information metric as the only metric (up to rescaling) that firstly is defined on all models with finite data spaces and secondly is invariant under (the measure-theoretic push-forwards of) all sufficient statistics. By comparison, our version characterises the Fisher information metric as the only metric (up to rescaling) that firstly is defined on an individual exponential family and its derived families and secondly is invariant under IID extensions and (the measure-theoretic push-forwards of) canonical sufficient statistics. We now show that the two versions of Chentsov's theorem differ in both respects, i.e. they differ in both the set of models on which the invariant metric is defined and on the assumed invariance properties of the metric.

Because IID extensions strictly increase the Fisher information metric (by a factor of n) [1, eq. 4.2], these transformations of statistical models cannot be induced by sufficient statistics (or any other Markov morphism) due to the monotonicity property of the Fisher information metric, e.g. see [1, p. 30–31]. So the invariance assumption of our version of Chentsov's theorem and the original version are different. Also, the original version applies to metrics defined on a countable collection of models of dimension $1, 2, 3, \ldots$ whereas our version applies to a countable collection of models that all have the same dimension (the dimension of Θ). So our assumption about the set of models on which the invariant metric is defined is also different from the original version of the theorem.

More concretely, the original version of Chentsov's theorem concerns a metric that is defined on each probability simplex

$$\triangle_{n-1} = \left\{ (p_1, \ldots, p_n) \in \mathbb{R}^n \,\middle|\, p_i > 0 \text{ and } 1 = \sum_{i=1}^{n} p_i \right\}$$

(with $(p_1, \ldots, p_n) \in \triangle_{n-1}$ putting mass p_i on the i^{th} point of the data space) and is invariant under all sufficient statistics of all submodels of these probability simplices, e.g. see [9] or [1, Thm. 2.6]. By contrast, our result applies to a metric that is only defined on an individual exponential family $\mathcal{M}$ and its derived families. For example, let $\mathcal{M}$ be the family of all

Bernoulli distributions and let $\mathcal{M}^n$ and $\mathcal{N}_n$ be the corresponding the derived families, as in Section 3.4. Then $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$ are all parameterised by $\Theta = \mathbb{R}$ so they are all 1-dimensional statistical models. In fact, $\mathcal{M}$, $\mathcal{M}^n$ and $\mathcal{N}_n$ are 1-dimensional submodels of $\triangle_1$, $\triangle_{2^n-1}$ and $\triangle_n$, respectively. Our invariance assumptions (A1) and (A2) say that the maps between these 1-dimensional submodels given by $P \mapsto P^n$ and $P^n \mapsto Q_n$ are isometries (up to a factor of n, for the first map), where P, P^n and Q_n are as given in Section 3.4 and all correspond to the same value of θ. The map $P^n \mapsto Q_n$ is induced by a sufficient statistic but, as argued above, the map $P \mapsto P^n$ is not, by the monotonicity of the Fisher information metric.

This shows that our version of Chentsov's theorem and the original version differ in both the set of models on which the invariant metric is defined and on the assumed invariance properties of the metric. Similar considerations also apply to the more recent versions of Chentsov's theorem [2, 7].

8. Discussion

Our version of Chentsov's theorem characterises the Fisher information metric as the unique Riemannian metric (up to rescaling) on a curved exponential family $\mathcal{M}$ that is invariant under IID extensions and canonical sufficient statistics. We proved this by considering metrics g on $\mathcal{M}$, g^n on the n-fold IID extension $\mathcal{M}^n$ of $\mathcal{M}$, and g_n on the natural exponential family $\mathcal{N}_n$ corresponding to $\mathcal{M}^n$. Then, under the above invariance conditions, g can be calculated in terms of g_n, for any n. But for large n, the central limit theorem and a property (5.1) of exponential families imply that $\mathcal{N}_n$ consists of distributions that are all approximately normally distributed, so each distribution in $\mathcal{N}_n$ is determined to a good approximation by its mean and variance-covariance matrix. Further, each tangent vector to $\mathcal{N}_n$ is essentially a linear function f times a distribution in $\mathcal{N}_n$. Combining these facts shows that (the norm corresponding to) g is approximately equal to a simple function of f and the mean and variance-covariance matrix of the relevant distribution in $\mathcal{N}_n$. Our regularity condition implies that this approximation becomes exact in the limit as $n \to \infty$. Then our main result follows from an identity (5.11) relating the variance-covariance matrix to the Fisher information metric on an exponential family.

In general, Chentsov's theorem characterizes the Fisher information metrics on statistical models as the only Riemannian metric (up to rescaling) that is invariant under certain, statistically important transformations. Previous studies have taken these transformations to be either all sufficient statistics or a large, regular subset of these. By contrast, we take these statistically important transformations to be the IID extensions and canonical sufficient statistics. This class of transformations is arguably more natural than the class of all sufficient statistics, it is more appropriate for exponential families and it is a relatively small class so our invariance assumptions are weaker than those of previous studies. Our regularity assumptions also appear to be weaker than previous studies, ultimately due to the fact that our approach only requires us to study a collection of finite-dimensional models, rather than an infinite-dimensional model.

We have given a new characterisation of the Fisher information metric on an (curved) exponential family and we have shown that this result is an intuitive consequence of the central limit theorem. The main limitation of this paper is that our main result is only proved for exponential families. However, these families are an important class of statistical models, being well studied and widely used in applications. Also, our proof treats discrete and continuous models in a uniform way, so there is some hope that our approach can be adapted to give a proof of Chentsov's theorem for general statistical models. Lastly, our focus

on exponential families complements the focus of Bauer et al. [7] on diffeomorphism-invariant metrics, since curved exponential families are essentially the only statistical models that have smooth sufficient statistics that are not diffeomorphisms, by the Pitman–Koopman–Darmois theorem [5].

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of mathematical monographs*. American Mathematical Society, Providence, 2000.

[2] N. Ay, J. Jost, H. V. Lê, and L. Schwachhöfer. Information geometry and sufficient statistics. *Probab. Theory Relat. Fields*, 162:327–364, 2015.

[3] O. Barndorff-Nielsen. *Information and exponential families*. John Wiley & Sons, New York, 1978.

[4] O. Barndorff-Nielsen and D. R. Cox. Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society Series B (Methodological)*, 41(3):279–312, 1979.

[5] O. Barndorff-Nielsen and K. Pedersen. Sufficient data reduction and exponential families. *Math. Scand.*, 22:197–202, 1968.

[6] A. R. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, October 1998.

[7] M. Bauer, M. Bruveris, and P. W. Michor. Uniqueness of the Fisher-Rao metric on the space of smooth densities. *Bull. London Math. Soc.*, 48:499–506, 2016.

[8] V. I. Bogachev. *Measure Theory, Volume I*. Springer, Berlin, 2007.

[9] L. L. Campbell. An extended Cencov characterization of the information metric. *Proc. Am. Math. Soc.*, 98:135–141, 1986.

[10] N. N. Chentsov. Algebraic foundation of mathematical statistics. *Math. Operationsforsch. statist.*, 9:267–276, 1978.

[11] J. G. Dowty. Chentsov's theorem for exponential families. *arXiv:1701.08895*, 2017.

[12] F. G. Friedlander and M. Joshi. *Introduction to the Theory of Distributions (second edition)*. Cambridge University Press, Cambridge UK, 1998.

[13] R. E. Kass and P. W. Vos. *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons, New York, 1997.

[14] S. Lang. *Fundamentals of Differential Geometry*, volume 191 of *Graduate Texts in Mathematics*. Springer, New York, 1999.

[15] H. V. Lê. The uniqueness of the fisher metric as information metric. *arXiv:1306.1465*, 2013.

[16] M. M. Meerschaert and H.-P. Scheffler. *Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice*. Wiley series in probability and statistics. John Wiley & Sons, New York, 2001.

[17] G. Pistone and C. Sempri. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Ann. Stat.*, 23(5):1543–1561, 1995.

[18] J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.

[19] L. Schwachhöfer, N. Ay, J. Jost, and H. V. Lê. Congruent families and invariant tensors. *arXiv: 1705.11014*, 2017.

[20] N. N. Čencov. *Statistical Decision Rules and Optimal Inference*, volume 53 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, 1982.

Centre for Epidemiology and Biostatistics, University of Melbourne, Parkville, Victoria, 3010, Australia
E-mail address: jdowty at unimelb.edu.au