Chapter 1
Introduction

1.1 A Brief Synopsis

Parametric statistics is concerned with families

$$\mathbf{p} : M \to \mathcal{P}(\Omega) \tag{1.1}$$

of probability measures on some sample space Ω. That is, for each ξ in the parameter space M, we have a probability measure $p(\cdot;\xi)$ on Ω. And typically, one wishes to estimate the parameter ξ based on random samples drawn from some unknown probability distribution on Ω, so as to identify a particular $p(\cdot;\xi_0)$ that best fits that sampling distribution. Information geometry provides geometric tools to analyze such families. In particular, a basic question is how sensitively $p(x;\xi)$ depends on the sample x. It turns out that this sensitivity can be quantified by a Riemannian metric, the Fisher metric originally introduced by Rao. Therefore, it is natural to bring in tools from differential geometry. That metric on the parameter space M is obtained by pulling back some universal structure from $\mathcal{P}(\Omega)$ via (1.1). When Ω is infinite, which is not an untypical situation in statistics, however, $\mathcal{P}(\Omega)$ is infinite-dimensional, and therefore functional analytical problems arise. One of the main features of this book consists in a general, and as we think, most satisfactory, approach to these issues.

From a geometric perspective, it is natural to look at invariances. On the one hand, we can consider mappings

$$\kappa : \Omega \to \Omega' \tag{1.2}$$

into some other space Ω'. Such a κ is called a statistic. In some cases, Ω' might even be finite even if Ω itself is infinite. For instance, Ω' could simply be the index set of some finite partition of Ω, and $\kappa(x)$ then would simply tell us in which member of that partition the point x is found. In other cases, κ might stand for a specific observable on Ω. A natural question then concerns the possible loss of information about the parameter $\xi \in M$ from the family (1.1) when we only observe $\kappa(x)$ instead of x itself. The statistic κ is called sufficient for the family (1.1) when no information

N. Ay et al., *Information Geometry*, Ergebnisse der Mathematik und ihrer Grenzgebiete.
3. Folge / A Series of Modern Surveys in Mathematics 64,
DOI 10.1007/978-3-319-56478-4_1

is lost at all. It turns out that the information loss can be quantified by the difference of the Fisher metrics of the originally family $\mathbf{p}$ and the induced family $\kappa_* \mathbf{p}$. We shall also show that the information loss can be quantified by tensors of higher order. In fact, one of the results that we shall prove in this book is that the Fisher metric is uniquely characterized (up to a constant factor, of course) by being invariant under all sufficient statistics.

Another invariance concerns reparametrizations of the parameter space M. Of course, as a Riemannian metric, the Fisher metric transforms appropriately under such reparametrizations. However, there are particular families $\mathbf{p}$ with particular parametrizations. These naturally play an important role. In order to see how they arise, we need to look at the structure of the space $\mathcal{P}(\Omega)$ of probability measures more carefully (see Fig. 1.1). Every probability measure is a measure tout court, that is, there is an embedding

$$\imath : \mathcal{P}(\Omega) \to \mathcal{S}(\Omega) \tag{1.3}$$

into the space $\mathcal{S}(\Omega)$ of all finite signed measures on Ω. As a technical point, for a probability measure, that is, an element of $\mathcal{P}(\Omega)$, we require it to be nonnegative, but for a general (finite) measure, an element of $\mathcal{S}(\Omega)$, we do not impose this restriction. The latter space is a linear space. $p \in \mathcal{P}(\Omega)$ then is simply characterized by $\int_\Omega dp(x) = 1$, and so, $\mathcal{P}(\Omega)$ becomes a convex subset (because of the nonnegativity constraint) of an affine subspace (characterized by the condition $\int_\Omega d\mu(x) = 1$) of the linear space $\mathcal{S}(\Omega)$. On the other hand, there is also a projection

$$\pi : \mathcal{M}(\Omega) \to \mathcal{P}(\Omega) \tag{1.4}$$

of the space of nonnegative measures by assigning to each $m \in \mathcal{M}(\Omega)$ the relative measure of subsets. For any measurable subsets $A, B \subseteq \Omega$ with $m(B) > 0$, $\pi(m)$ looks at the quotients $\frac{m(A)}{m(B)}$, that is, the relative measures of those subsets. That is, a probability measure is now considered as an equivalence class of measures up to a scaling factor. Of course, as such, a probability measure and such an equivalence class is not quite the same, and therefore the target of π in (1.4) is not really $\mathcal{P}(\Omega)$, but $\mathcal{P}(\Omega)$ can be easily identified with $\pi(\mathcal{M}(\Omega))$ (modulo certain technical points that we suppress in this synopsis), by simply normalizing a measure by $m(\Omega)$ (assuming that the latter is finite). From the perspective of such relative measures, $\pi(\mathcal{M}(\Omega))$, that is by what we have just said, $\mathcal{P}(\Omega)$, can be seen as the positive part of a projective space of the linear space $\mathcal{S}(\Omega)$, that is, as the positive orthant or sector of the unit sphere in $\mathcal{S}(\Omega)$. When Ω is finite, the linear space $\mathcal{S}(\Omega)$ is finite-dimensional, and therefore, it can be naturally equipped with a Euclidean metric. This metric then also induces a metric on the unit sphere, or in the terminology developed here, the projection map π from (1.4) then induces a metric on $\mathcal{P}(\Omega)$. This is the Fisher metric, a fundamental object of our study. When Ω is infinite, then the space $\mathcal{S}(\Omega)$ is infinite-dimensional, but it does not carry the structure of a Hilbert space. Nevertheless, by considering variations of class $L^2(\Omega, \mu)$, we still obtain an L^2-scalar product, and that will again be the Fisher metric. (The space within which we vary our measure—L^1, L^∞ or L^2—will be an important technical issue

for our functional analytical considerations. In fact, L^1 will be the natural choice, as it behaves naturally under a change of base measure.)

It turns out that the structure induced on $\mathcal{P}(\Omega)$ by (1.4) in a certain sense is dual to the affine structure induced on this space by the embedding (1.3). In fact, this dual structure is affine itself, and it can be described in terms of an exponential map. In order to better understand this, let us discuss the two possible ways in which a measure can be normalized to become a probability measure. We start with a probability measure μ and want to move to another probability measure ν. We can write additively

$$\nu = \mu + (\nu - \mu) \tag{1.5}$$

and connect ν with μ by the straight line

$$\mu + t(\nu - \mu), \quad \text{with } t \in [0, 1]. \tag{1.6}$$

When we consider an arbitrary variation

$$\mu + t\xi, \tag{1.7}$$

when we want to stay within the class of probability measures, we need to subtract $\xi_0 := \xi(\Omega)$, that is, consider the measure $\xi - \xi_0$ defined by $(\xi - \xi_0)(A) := \xi(A) - \xi(\Omega)$. Thus, we get the variation

$$\mu + t(\xi - \xi_0). \tag{1.8}$$

Here, we see the problem that even if μ is nonnegative, as it should be as a probability measure, and if $\mu + t\xi$ is nonnegative as well for $t \in [0, 1]$, $\mu + t(\xi - \xi_0)$ need not always be nonnegative. Expressing this geometrically, the geodesic $\mu + t(\xi - \xi_0)$ (which is meaningful for all $t \in \mathbb{R}$) with respect to the affine structure on the simplex may leave the simplex of probability measures. Thus, this affine structure is not complete. Alternatively, we can consider a multiplicative variation and write

$$\nu = \exp\left(\frac{d\nu}{d\mu}\right)\mu \tag{1.9}$$

where $\frac{d\nu}{d\mu}$ is the Radon–Nikodym derivative of ν w.r.t. μ. A general variation would then be of the form

$$\exp(tf)\mu, \quad \text{with } t \in [0, 1], \tag{1.10}$$

where we require that the function $\exp f$ be in $L^1(\Omega, \mu)$. Here, we choose the exponential for two reasons. First, this ensures that the measure $\exp(tf)\mu$ is nonnegative if μ is. Thus, we do not run into the problem of noncompleteness as for the additive variation. Secondly, we can consider a linear space of functions f here. There is an important technical problem, though. $\exp(t_1 f) \in L^1$ does not imply that $\exp(t_2 f) \in L^1$ for $t_2 > t_1$; we shall return to this problem. For the moment, it can be circumvented by requiring that $f \in L^\infty(\Omega, \mu)$ because that implies that

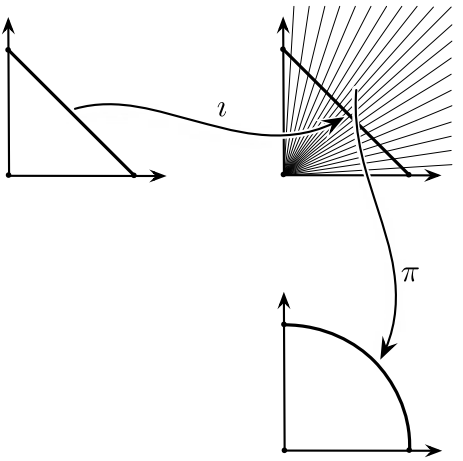

Fig. 1.1 Natural inclusion and projection

$\exp(tf) \in L^\infty$ as well for all t. Again, for a general function f, we need to impose a normalization in order to stay in the class of probability distributions. This leads to the variation

$$\frac{\exp(tf)}{Z(t)}\mu \quad \text{with } Z(t) := \int_\Omega \exp(tf)\, d\mu. \tag{1.11}$$

This multiplicative normalization is, of course, in line with our view of the probability measures as equivalence classes of measures up to a factor. Moreover, we can consider the family (1.11) as a geodesic for an affine structure, as we shall now explain. First, although the normalization factor $Z(t)$ depends on the measure μ, this does not matter as we are considering elements of a projective space which does not see a global factor. Secondly, when we have two probability measures μ, μ_1 with $\mu_1 = \phi\mu$ for some positive function ϕ with $\phi \in L^1(\Omega, \mu)$ and hence $\phi^{-1} \in L^1(\Omega, \mu_1)$, then the variations $\exp(tf)\mu$ of μ correspond to the variations $\frac{\exp(tf)}{\phi}\mu_1$ of μ_1. At the level of the linear spaces, the correspondence would be between f and $f - \log\phi$ (we might wish to require here that $\phi, \phi^{-1} \in L^\infty$ according to the previous discussion, but let us ignore this technical point for the moment). The important point here is that we can identify the variations at μ and μ_1 here in a manner that does not depend on the individual f, because the shift by $\log\phi$ is the same for all f. Moreover, when we have $\mu_2 = \psi\mu_1$, then $\mu_2 = \psi\phi\mu$, and the shift is by $\log(\psi\phi) = \log\psi + \log\phi$. But this is precisely what an affine structure amounts to. Thus, we have identified the second affine structure on the space of probability measures. It possesses a natural exponential map $f \mapsto \exp f$, is naturally adapted to our description of probability measures as equivalence classes of measures, and is complete in contrast to the first affine structure. As we shall explore in more detail in the geometric part of this book, these two structures are naturally dual to each other. They are related by a Legendre transform that generalizes the duality between entropy and free energy that is at the heart of statistical mechanics.

This pair of dual affine structures was discovered by Amari and Chentsov, and the tensor describing it is therefore called the Amari–Chentsov tensor. The Amari–Chentsov tensor encodes the difference between the two affine connections, and they can be recovered from the Fisher metric and this tensor. Like the Fisher metric, the Amari–Chentsov tensor is invariant under sufficient statistics, and uniquely characterized by this fact, as we shall also show in this book. Spaces with such a pair of dual affine structures turn out to have a richer geometry than simple affine spaces. In particular, such affine structures can be derived from potential functions. In particularly important special cases, these potential functions are the entropy and the free energy as known from statistical mechanics.

Thus, there is a natural connection between information geometry and statistical mechanics. Of course, there is also a natural connection between statistical mechanics and information theory, through the analogy between Boltzmann–Gibbs entropy and Shannon information. In many interesting cases within statistical mechanics, the interaction of physical elements can be described in terms of a graph or, more generally, in terms of a hypergraph. This leads to families of Boltzmann–Gibbs distributions that are known as hierarchical or graphical models.

In fact, information geometry also directly leads to geometric descriptions of information theoretical concepts, and this is another topic that we shall systematically explore in this book. In particular, we shall treat conditional and relative entropies from a geometric perspective, analyze exponential families, including interaction spaces and hierarchical and graphical models, and describe applications like replicator equations in mathematical biology and population game theory. Since many of those geometric properties and applications show themselves already in the case where the sample space Ω is finite and hence the spaces of measures on it are finite-dimensional, we shall start with a chapter on that case.

We consider families of measures $\mathbf{p}(\xi)$ on a sample space Ω parametrized by ξ from our parameter space M. For different ξ, the resulting measures might be quite different. In particular, they may have rather different null sets. Nevertheless, in many cases, for instance, if M is a finite-dimensional manifold, we may write such a family as

$$\mathbf{p}(\xi) = p(\cdot;\xi)\mu_0, \tag{1.12}$$

for some base measure μ_0 that does not depend on ξ. $p : \Omega \times M \to \mathbb{R}$ is the density function of $\mathbf{p}$ w.r.t. μ_0, and we then need that $p(\cdot;\xi) \in L^1(\Omega, \mu_0)$ for all ξ. This looks convenient, after all such a μ_0 is an auxiliary object, and it is a general mathematical principle that structures should not depend on such auxiliary objects. Implementing this principle systematically will, in fact, give us the crucial leverage needed to develop the general theory. Let us be more precise. As already observed above, when we have another probability measure μ_1 with $\mu_1 = \phi\mu_0$ for some positive function ϕ with $\phi \in L^1(\Omega, \mu_0)$ and hence $\phi^{-1} \in L^1(\Omega, \mu_1)$, then $\psi \in L^1(\Omega, \mu_1)$ precisely if $\psi\phi \in L^1(\Omega, \mu_0)$. Thus, the L^1-spaces naturally correspond to each other, and it does not matter which base measure we choose, as long as the different base measures are related by L^1-functions.

Second, the differential of $\mathbf{p}$ in some direction V is then given by

$$d_\xi \mathbf{p}(V) = \partial_V p(\cdot;\xi)\mu_0 \in L^1(\Omega, \mu_0), \tag{1.13}$$

assuming that this quantity exists. According to what we have just said, however, what we should consider is not $\partial_V p(\cdot;\xi)\mu_0$, which measures the change of measure w.r.t. the background measure μ_0, but rather the rate of change of $\mathbf{p}(\xi)$ relative to the measure $\mathbf{p}(\xi)$ itself, that is, the Radon–Nikodym derivative of $d_\xi \mathbf{p}(V)$ w.r.t. $\mathbf{p}(\xi)$, that is, the *logarithmic derivative*

$$\partial_V \log p(\cdot;\xi) = \frac{d\{d_\xi \mathbf{p}(V)\}}{d\mathbf{p}(\xi)}. \tag{1.14}$$

(Note that this is not a second derivative, as the outer d stands for the Radon–Nikodym derivative, that is, essentially a quotient of measures. The slightly confusing notation ultimately results from writing integration with respect to $\mathbf{p}(\xi)$ as $\int d\mathbf{p}(\xi)$.)

This then leads to the Fisher metric

$$\mathfrak{g}_\xi(V, W) = \int_\Omega \partial_V \log p(\cdot;\xi)\, \partial_W \log p(\cdot;\xi)\, d\mathbf{p}(\xi). \tag{1.15}$$

One may worry here about what happens when the density p is not positive almost everywhere. In order to see that this is not really a problem, we introduce the formal square roots

$$\sqrt{\mathbf{p}(\xi)} := \sqrt{p(\cdot;\xi)}\sqrt{\mu_0}, \tag{1.16}$$

and use the formal computation

$$d_\xi \sqrt{\mathbf{p}}(V) = \frac{1}{2}\partial_V \log p(\cdot;\xi)\sqrt{\mathbf{p}(\xi)} \tag{1.17}$$

to rewrite (1.15) as

$$\mathfrak{g}_\xi(V, W) = 4\int_\Omega d\big(d_\xi \sqrt{\mathbf{p}}(V) \cdot d_\xi \sqrt{\mathbf{p}}(W)\big). \tag{1.18}$$

Also, in a sense to be made precise, an L^1-condition on $\mathbf{p}(\xi)$ becomes an L^2-condition on $\sqrt{\mathbf{p}(\xi)}$ in (1.16), and an L^2-condition is precisely what we need in (1.18) for the derivatives. According to (1.17), this means that we should now impose an L^2-condition on $\partial_V \log p(\cdot;\xi)$. Again, all this is naturally compatible with a change of base measure.

1.2 An Informal Description

Let us now informally describe some of the main points of information geometry as treated in this book, and thereby perhaps already give away some of our secrets.

"Informally" here is meant seriously, indeed. That is, we shall suppress certain technical points that will, of course, be clarified in the main text.

1.2.1 The Fisher Metric and the Amari–Chentsov Structure for Finite Sample Spaces

Let first $I = \{1, \dots, n\}$, $n \in \mathbb{N} = \{1, 2, \dots\}$, be a finite sample space, and consider the set of nonnegative measures $\mathcal{M}(I) = \{(m_1, \dots, m_n) : m_i \geq 0, \sum_j m_j > 0\}$ on it. A probability measure is then either a tuple $(p_1, \dots, p_n) \in \mathcal{M}(I)$ with $\sum_j p_j = 1$, or a measure up to scaling. The latter means that we do not consider the measure m_i of an $i \in I$, or more generally, of a subset of I, but rather only quotients $\frac{m_i}{m_j}$ whenever $m_j > 0$. In other words, we look at relative instead of absolute measures. Clearly, $\mathcal{M}(I)$ can be identified with the positive sector $\mathbb{R}^n_+$ of $\mathbb{R}^n$. The first perspective would then identify the set of probability measures with the simplex $\Sigma^{n-1} = \{(y_1, \dots, y_n) : y_i \geq 0, \sum_j y_j = 1\}$, whereas the latter would rather identify it with the positive part of the projective space $\mathbb{P}^{n-1}$, that is, with the positive orthant or sector of the unit sphere S^{n-1} in $\mathbb{R}^n$, $S^{n-1}_+ = \{(q^1, \dots, q^n) : q^i \geq 0, \sum_j (q^j)^2 = 1\}$. Of course, Σ^{n-1} and S^{n-1}_+ are homeomorphic, but otherwise, their geometry is different. We shall utilize both of them. Foremost, the sphere S^{n-1} carries its natural metric induced from the Euclidean metric on $\mathbb{R}^n$. Therefore, we obtain a Riemannian metric on the set of probability measures. This is the Fisher metric. Next, let us take a measure $\mu_0 = (m_1, \dots, m_n)$ with $m_i > 0$ for all i. We call a measure $\phi\mu_0 = (\phi_1 m_1, \dots, \phi_n m_n)$ compatible with μ_0 if $\phi_i > 0$ for all i. Let us call the space of these measures $\mathcal{M}_+(I, \mu_0)$. Of course, this space does not really depend on μ_0; the only relevant aspect is that all entries of μ_0 be positive. Nevertheless, it will be instructive to look at the dependence on μ_0 more carefully. $\mathcal{M}_+(I, \mu_0)$ forms a group under pointwise multiplication. Equally importantly, we get an affine structure. Considering Σ^{n-1}, this is obvious, as the simplex is a convex subset of an $(n-1)$-dimensional affine subspace of $\mathbb{R}^n$. Perhaps somewhat more surprisingly, there is another affine structure which we shall now describe. Let $\mu_1 = \phi_1\mu_0 \in \mathcal{M}_+(I, \mu_0)$ and $\mu_2 = \phi_2\mu_1 \in \mathcal{M}_+(I, \mu_1)$. Thus, we also have $\mu_2 = \phi_2\phi_1\mu_0 \in \mathcal{M}_+(I, \mu_0)$. In particular, whenever $\mu = \phi\mu_0$ is compatible with μ_0, we have a canonical identification of $\mathcal{M}_+(I, \mu)$ with $\mathcal{M}_+(I, \mu_0)$ via multiplication by ϕ. Of course, these spaces are not linear, due to the positivity constraints. We can, however, consider the linear space $T^*_{\mu_0} \cong \mathbb{R}^n$ of $(f_1, \dots, f_n)$ with $f_i \in \mathbb{R}$. This space is bijective to $\mathcal{M}_+(I, \mu_0)$ via $\phi_i = e^{f_i}$. This is an exponential map, as familiar from Riemannian geometry or Lie group theory. (As will be explained in Chap. 2, this space can be naturally considered as the cotangent space of the space of measures at the point μ_0. For the purposes of this introduction, the difference between tangent and cotangent spaces is not so important, however, and in any case, as soon as we have a metric, there is a natural identification between tangent and cotangent spaces.) Now, there is a natural identification between T^*_μ

and $T^*_{\mu_0}$; in fact, if $\mu = \phi\mu_0$, this is achieved by the correspondence $g_i = f_i - \log\phi_i$, because then $e^g\mu = e^f\mu_0$. Since this identification is independent of f and g, and furthermore is compatible with the above product structure, we have a natural correspondence between the (co)tangent spaces at different points of $\mathcal{M}_+$, that is, an affine structure. We thus have not one, but two affine structures. These two structures are indeed different, but there is a natural duality between them. This is the Amari–Chentsov structure which we shall describe in Sect. 1.1.

1.2.2 Infinite Sample Spaces and Functional Analysis

So far, the sample space has been finite. Let us now consider a general sample space, that is, some set Ω together with a σ-algebra, so that we can consider the space of measures on Ω. We shall assume for the rest of this discussion that Ω is infinite, as we have already described the finite case, and we want to see which aspects naturally extend. Again, in this introduction, we restrict ourselves to the positive measures. Probability measures then can again either be considered as measures μ with $\mu(\Omega) = 1$, or as relative measures, that is, considering only quotients $\frac{\mu(A)}{\mu(B)}$ whenever $\mu(B) > 0$. In the first case, we would deal with an infinite dimensional simplex, in the second one with the positive orthant or sector of an infinite-dimensional sphere. Now, given again some base measure μ_0, the space of compatible measures would be $\mathcal{M}_+(\Omega, \mu_0) = \{\phi\mu_0 : \phi \in L^1(\Omega, \mu_0), \phi > 0 \text{ almost everywhere}\}$. Some part of the preceding naturally generalizes. In particular, when $\mu_1 = \phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$ and $\mu_2 = \phi_2\mu_1 \in \mathcal{M}_+(\Omega, \mu_1)$, then $\mu_2 = \phi_2\phi_1\mu_0 \in \mathcal{M}_+(\Omega, \mu_0)$. And this is precisely the property that we shall need. However, we no longer have a multiplicative structure, because if $\phi, \psi \in L^1(\Omega, \mu_0)$, then their product $\phi\psi$ need not be in $L^1(\Omega, \mu_0)$ itself. Moreover, the exponential map $f \mapsto e^f$ (defined in a pointwise manner, i.e., $e^f(x) = e^{f(x)}$) is no longer defined for all f. In fact, the natural linear space would be $L^2(\Omega, \mu_0)$, but if $f \in L^2(\Omega, \mu_0)$, then e^f need not be in $L^1(\Omega, \mu_0)$. But nevertheless, wherever defined, we have the above affine correspondence. So, at least formally, we again have two affine structures, one from the infinite-dimensional simplex, and the other as just described. Also, from the positive sector of the infinite dimensional sphere, we again get (an infinite-dimensional version of) a Riemannian metric. Now, developing the functional analysis required to make this really work is one of the major achievements of this book, see Chap. 3. Our approach is different from and more general than the earlier ones of Amari–Nagaoka [16] and Pistone–Sempi [216].

For our treatment, another simple observation will be important. There is a natural duality between functions f and measures $\phi\mu$,

$$(f, \phi\mu) = \int_\Omega f\phi \, d\mu, \tag{1.19}$$

whenever f and ϕ satisfy appropriate integrability conditions. From the perspective of the duality between functions and measures, we might require that ϕ be in L^1

and f be in L^∞. We can turn (1.19) into a symmetric pairing by rewriting it as

$$\left\langle f(\mu)^{1/2}, \phi(\mu)^{1/2} \right\rangle = \int_\Omega f (d\mu)^{1/2} \phi (d\mu)^{1/2}. \tag{1.20}$$

Since this is symmetric, we would now require that both factors be in L^2. The objects involved in (1.20), that is, those that transform like $(d\mu)^{1/2}$, that is, with the square root of the Jacobian of a coordinate transformation, are called half-densities. In particular, the group of diffeomorphisms of Ω (assuming that Ω carries a differentiable structure) operates by isometries on the space of half-densities of class L^2. Therefore, this is a good space to work with. In fact, for the Amari–Chentsov structure, we also have to consider (1/3)-densities, that is, objects that transform like a cubic root of a measure.

In order to make this precise, in our approach, we define the Banach spaces of *formal rth powers of (signed) measures*, denoted by $\mathcal{S}^r(\Omega)$, where $0 < r \leq 1$. For instance, $\mathcal{S}^1(\Omega) = \mathcal{S}(\Omega)$ is the Banach space of finite signed measures on Ω with the total variation as the Banach norm. The space $\mathcal{S}^{1/2}(\Omega)$ is the space of *signed half-densities* which is a Hilbert space in a natural way (the concept of a half-density will be discussed in more detail after (1.20)). Just as we may regard the set of probability measures and (positive) finite measures as subsets $\mathcal{P}(\Omega) \subseteq \mathcal{M}(\Omega) \subseteq \mathcal{S}(\Omega)$, there are analogous inclusions $\mathcal{P}^r(\Omega) \subseteq \mathcal{M}^r(\Omega) \subseteq \mathcal{S}^r(\Omega)$ of rth powers of probability measures (finite measures, respectively). In particular, we also get a rigorous definition of the (formal) tangent bundle $T\mathcal{P}^r(\Omega)$ and $T\mathcal{M}^r(\Omega)$, where $T_\mu \mathcal{M}^r(\Omega) = L^k(\Omega, \mu)$ for $k = 1/r \geq 1$, so this is precisely the tangent space which was relevant in our previous discussion.

We also define the *signed kth power* $\tilde{\pi}^k : \mathcal{S}^r(\Omega) \to \mathcal{S}(\Omega)$ for $k := 1/r \geq 1$ which is a differentiable homeomorphism between these sets and can hence be regarded as a coordinate map on $\mathcal{S}(\Omega)$ changing the differentiable structure. It maps $\mathcal{P}^r(\Omega)$ to $\mathcal{P}(\Omega)$ and $\mathcal{M}^r(\Omega)$ to $\mathcal{M}(\Omega)$, respectively. A similar approach was used by Amari [8] who introduced the concept of *α-representations*, expressing a statistical model in different coordinates by taking powers of the model. The advantage of our approach is that the definition of the parametrization $\tilde{\pi}^k$ is universally defined on $\mathcal{S}^r(\Omega)$ and does not depend on a particular parametrized measure model.

Given a statistical model $(M, \Omega, \mathbf{p})$, we interpret it as a differentiable map from M to $\mathcal{P}(\Omega) \subseteq \mathcal{S}(\Omega)$. Then the notion of k-integrability of the model from [25] can be interpreted in this setting as the condition that for $r = 1/k$, the rth power $\mathbf{p}^r$ mapping M to $\mathcal{P}^r(\Omega) \subseteq \mathcal{S}^r(\Omega)$ is continuously differentiable. Note that in the definition of the model $(M, \Omega, \mathbf{p})$, we do not assume the existence of a measure dominating all measures $\mathbf{p}(\xi)$, nor do we assume that all measures $\mathbf{p}(\xi)$ have the same null sets. With this, our approach is indeed more general than the notions of differentiable families of measures defined, e.g., in [9, 16, 25, 216, 219].

For each n, we can define a canonical n-tensor on $\mathcal{S}^r(\Omega)$ for $0 < r \leq 1/n$, which can be pulled back to M via $\mathbf{p}^r$. In the cases $n = 2$ and $n = 3$, this produces the Fisher metric and the Amari–Chentsov tensor of the model, respectively. We shall show in Chap. 5 that the canonical n-tensors are invariant under sufficient statis-

tics and, moreover, the set of tensors invariant under sufficient statistics are algebraically generated by the canonical n-tensors. This is a generalization of the results of Chentsov and Campbell to the case of tensors of arbitrary degree and arbitrary measure spaces Ω.

Let us also mention here that when Ω is a manifold and the model $\mathbf{p}(\xi)$ consists of smooth densities, the Fisher metric can already be characterized by invariance under diffeomorphisms, as has been shown by Bauer, Bruveris and Michor [44]. Thus, in the more restricted smooth setting, a weaker invariance property already suffices to determine the Fisher metric. For the purposes of this book, in particular the mathematical foundation of parametric statistics, however, the general measure theoretical setting that we have developed is essential.

There is another measure theoretical structure which was earlier introduced by Pistone–Sempi [216]; we shall discuss that structure in detail in Sect. 3.3. In fact, to appreciate the latter, the following observation is a key. Whenever $e^f \in L^1(\Omega, \mu_0)$, then for $t < 1$, $e^{tf} = (e^f)^t \in L^p(\Omega, \mu_0)$ for $p = 1/t > 1$. Thus, the set of f with $e^f \in L^1$ is not only starshaped w.r.t. the origin, but whenever we scale by a factor $t < 1$, the integrability even improves. This, however, substantially differs from our approach. In fact, in the Pistone–Sempi structure the topology used (e-convergence) is very strong and, as we shall see, it decomposes the space $\mathcal{M}_+(\Omega; \mu_0)$ of measures compatible with μ_0 into connected components, each of which is an open convex set in a Banach space. Thus, $\mathcal{M}_+(\Omega; \mu_0)$ becomes a Banach manifold with an affine structure under the e-topology. In contrast, the topology that we use on $\mathcal{P}^r(\Omega)$ is essentially the L^k-topology on $L^k(\Omega, \mu)$ for $k = 1/r$, which is much weaker. This implies that, on the one hand, $\mathcal{P}^r(\Omega)$ is *not* a Banach manifold but merely a closed subset of the Banach space $\mathcal{S}^r(\Omega)$, so it carries far less structure than $\mathcal{M}_+(\Omega; \mu_0)$ with the Pistone–Sempi topology. On the other hand, our structure is applicable to many statistical models which are not continuous in the e-topology of Pistone–Sempi.

1.2.3 Parametric Statistics

We now return to the setting of parametric statistics, because that is a key application of our theory. In parametric statistics, one considers only parametrized families of measures on the sample space Ω, rather than the space $\mathcal{P}(\Omega)$ of all probability measures on Ω. These families are typically finite-dimensional (although our approach can also naturally handle infinite-dimensional families). We consider such a family as a mapping $\mathbf{p} : M \to \mathcal{P}(\Omega)$ from the parameter space M into $\mathcal{P}(\Omega)$, and $\mathbf{p}$ needs to satisfy appropriate continuity and differentiability properties related to the L^1-topology on $\mathcal{P}(\Omega)$. In fact, some of the most difficult technical points of our book are concerned with getting these properties right, so that the abstract Fisher and Amari–Chentsov structures on $\mathcal{P}(\Omega)$ can be pulled back to such an M via $\mathbf{p}$. This makes these structures functorial in a natural way.

The task or purpose of parametric statistics then is to identify an element of such a family M that best describes the statistics obtained from sampling Ω. A map that

converts the samples from Ω into estimates for a parameter $\xi \in M$ is called an estimator. The Fisher metric quantifies the sensitivity of the dependence of the parameter ξ on the samples in Ω, and this leads to the Cramér–Rao inequality which constrains any estimator. Moreover, instead of sampling from Ω, we could consider a map $\kappa : \Omega \to \Omega'$ to some possibly much smaller space. In general, sampling from a smaller space loses some information about the parameter ξ, and consequently, the Fisher metric decreases. In fact, we shall show such a monotonicity result under very general assumptions. In informal terms, such a map κ is called a sufficient statistic for a family $\mathbf{p} : M \to \mathcal{P}(\Omega)$ if sampling from Ω' is as good for identifying the parameter ξ as sampling from Ω itself. In that case, the parameter sensitivity should be the same in either case, and according to the interpretation of the Fisher metric just given, it should be invariant under sufficient statistics. A remarkable result of Chentsov says that, conversely, the Fisher metric and the Amari–Chentsov tensor are uniquely determined by their invariance under sufficient statistics. Chentsov proved this result in the finite case only. Building upon our work in [25], we present a proof of this unique characterization of the Fisher and Amari–Chentsov structure in the general situation of an arbitrary sample space Ω. This is one of the main results derived in this book. In fact, we shall prove a very general result that classifies all tensors that are invariant under congruent Markov kernels. These statistical aspects of information geometry are taken up in Chap. 5 for the general case of an arbitrary Ω, with reference to the case of a finite Ω already treated in Chap. 2. We shall now describe this in some more detail.

Let $\kappa : \Omega \to \Omega'$ be a statistic (see (1.2)). Such a κ then induces a map κ_* on signed measures via

$$\kappa_*\mu(A) := \mu\big\{\omega \in \Omega : \kappa(\omega) \in A\big\} = \mu\big(\kappa^{-1}A\big),$$

and thus a statistical model $(M, \Omega, \mathbf{p})$ on Ω gets transformed into one on Ω', $(M, \Omega', \mathbf{p}')$. In general, it will be more difficult to recover the parameter $\xi \in M$ from $\kappa_\star p(\cdot;\xi)$ by observations on $\omega' \in \Omega'$ than from the original $p(\cdot;\xi)$ through observations of $x \in \Omega$, because κ might map several ω into the same ω'. In fact, if we put

$$\mathfrak{g}'_{kl}(\xi) = \int p'\big(\omega';\xi\big)\frac{\partial \log p'(\omega';\xi)}{\partial \xi^k}\frac{\partial \log p'(\omega';\xi)}{\partial \xi^l}\,d\mu'\big(\omega'\big) \tag{1.21}$$

then

$$(\mathfrak{g}_{kl}) \geq \big(\mathfrak{g}'_{kl}\big) \tag{1.22}$$

in the sense of tensors, that is, the difference is a nonnegative definite tensor. When no information is lost, the statistic is called sufficient, and we have equality in (1.22). (There are various characterizations of sufficient statistics, and we shall show their equivalence under our general conditions. Informally, a statistic is sufficient for the parameter ξ if having observed ω', no further information about ξ can be obtained from knowing which of the possible ω with $\kappa(\omega) = \omega'$ had occurred.)

More generally, we consider a Markov kernel, that is

$$K : \Omega \to \mathcal{P}(\Omega'). \tag{1.23}$$

For instance, we can consider conditional probability distributions $p(\omega'|\omega)$ for $\omega' \in \Omega'$. Of course, a statistic κ induces the Markov kernel K^κ where

$$K^\kappa(\omega) = \delta^{\kappa(\omega)}, \tag{1.24}$$

the Dirac measure at $\kappa(\omega)$. A Markov kernel K induces the Markov morphism

$$K_* : \mathcal{S}(\Omega) \longrightarrow \mathcal{S}(\Omega'), \qquad K_*\mu(A') := \int_\Omega K(\omega; A')\, d\mu(\omega), \tag{1.25}$$

that is, we simply integrate the kernel with respect to a measure on Ω to get a measure on Ω'. In particular, a statistic κ then induces the Markov morphism K^κ_*. It turns out that it is expedient to consider invariance properties with respect to Markov morphisms. While this will be technically important, in this Introduction, we shall simply consider statistics.

When $\kappa : \Omega \to \Omega'$ is a statistic, a Markov kernel $L : \Omega' \to \mathcal{P}(\Omega)$, i.e., going in the opposite direction now, is called κ-congruent if

$$\kappa_\star(L(\omega')) = \delta^{\omega'} \quad \text{for all } \omega' \in \Omega'. \tag{1.26}$$

In order to assess the information loss caused by going from Ω to $\mathcal{P}(\Omega')$ via a Markov kernel, there are two aspects

1. Several $\omega \in \Omega$ might get mapped to the same $\omega' \in \Omega'$. This clearly represents a loss of information, because then we can no longer recover ω from observing ω'. And if this distinction between the different ω causing the same ω' is relevant for estimating the parameter ξ, then lumping several ω into the same ω' loses information about ξ.
2. An $\omega \in \Omega$ gets diluted, that is, we have a distribution $p(\cdot|\omega)$ in place of a single value. By itself, this does not need to cause a loss of information. For instance, for different values of ω, the corresponding distributions could have disjoint supports.

In fact, any Markov kernel can be decomposed into a statistic and a congruent Markov kernel. That is, there is a Markov kernel $K^{cong} : \Omega \to \mathcal{P}(\hat{\Omega})$ which is congruent w.r.t. some statistic $\kappa_1 : \hat{\Omega} \to \Omega$, and a statistic $\kappa_2 : \hat{\Omega} \to \Omega'$ such that

$$K = \kappa_{2*} K^{cong}. \tag{1.27}$$

Moreover, we have the general monotonicity theorem

Theorem 1.1 *Let (M, Ω, p) be a statistical model on Ω as before, let $K : \Omega \to \mathcal{P}(\Omega')$ be a Markov kernel, inducing the family $p'(\cdot; \xi) = K_*(p(\cdot; \xi))$. Moreover, let $\mathfrak{g}_M$ and $\mathfrak{g}'_M$ denote the corresponding Fisher metrics. Then*

$$\mathfrak{g}_M(V, V) \geq \mathfrak{g}'_M(V, V) \quad \textit{for all } V \in T_\xi M \textit{ and } \xi \in M. \tag{1.28}$$

If $K = K^\kappa$ is the Markov kernel induced by a statistic κ as in (1.24), and if $(M, \Omega, \mathbf{p})$ has a positive regular density function, equality here holds for all ξ and all V if and only if the statistic κ is sufficient.

When $K = K^\kappa$, the difference $\mathfrak{g}_M(V, V) - \mathfrak{g}'_M(V, V)$ can then be taken as the information loss caused by the statistic κ.

Conversely, as already mentioned several times, we have

Theorem 1.2 *The Fisher metric is the unique metric, and the Amari–Chentsov tensor is the only* 3*-tensor* (*up to a constant scaling factor*) *that are invariant under sufficient statistics.*

The Fisher metric also enters into the Cramér–Rao inequality of statistics. The task of parametric statistics is to find an element in the parameter space Ξ that is most appropriate for describing the observations made in Ω. In this sense, one defines an estimator as a map

$$\hat{\xi} : \Omega \to \Xi$$

that associates to every observed datum x in Ω a probability distribution from the class Ξ. As Ξ can also be considered as a family of product measures on Ω^N ($N \in \mathbb{N}$), we can also associate to every tuple $(x_1, \ldots, x_N)$ of observations an element of Ξ. The most important example is the maximum likelihood estimator that selects that element of Ξ which assigns the highest weight to the observation x among all elements in Ξ.

Let $\vartheta : \Xi \to \mathbb{R}^d$ be coordinates on Ξ. We can then write the family Ξ as $p(\cdot; \vartheta)$ in terms of those coordinates ϑ. For simplicity, we assume that $d = 1$, that is, we have only a single scalar parameter. The general case can be easily reduced to this one.

We define the bias of an estimator $\hat{\xi}$ as

$$b_{\hat{\xi}}(\vartheta) := \mathbb{E}_\vartheta \hat{\xi} - \vartheta, \tag{1.29}$$

where E_ϑ stands for the expectation w.r.t. $p(\cdot; \vartheta)$. The Cramér–Rao inequality then says

Theorem 1.3 *Any estimator $\hat{\xi}$ satisfies*

$$\mathbb{E}_\vartheta\big((\hat{\xi} - \vartheta)^2\big) \geq \frac{(1 + b'_{\hat{\xi}}(\vartheta))^2}{\mathfrak{g}(\vartheta)} + b_{\hat{\xi}}(\vartheta)^2, \tag{1.30}$$

where $'$ stands for a derivative w.r.t. ϑ.

In particular, when the estimator is unbiased, that is, $b_{\hat{\xi}} = 0$, we have

$$\mathbb{E}_\vartheta\Big(\big(\hat{\xi} - E_\vartheta(\hat{\xi})\big)^2\Big) = \mathbb{E}_\vartheta\big((\hat{\xi} - \vartheta)^2\big) \geq \frac{1}{\mathfrak{g}(\vartheta)}, \tag{1.31}$$

that is, the variance of $\hat{\xi}$ is bounded from below by the inverse of the Fisher metric.

Here, $\mathfrak{g}(\vartheta)$ is an abbreviation for $\mathfrak{g}(\vartheta)(\frac{\partial}{\partial\vartheta}, \frac{\partial}{\partial\vartheta})$.

Thus, we see that the Fisher metric $\mathfrak{g}(\vartheta)$ measures how sensitively the probability density $p(\omega;\vartheta)$ depends on the parameter ϑ. When this is small, that is, when varying ϑ does not change $p(\omega;\vartheta)$ much, then it is difficult to estimate the parameter ϑ from the data, and the variance of an estimator consequently has to be large.

All these statistical results will be shown in the general framework developed in Chap. 3, that is, in much greater generality than previously known.

1.2.4 Exponential and Mixture Families from the Perspective of Differential Geometry

Before those statistical applications, however, in Chap. 4, we return to the differential geometric aspects already introduced in Chap. 2. Recall that there are two different affine structures on our spaces of probability measures, one coming from the simplex, the other from the exponential maps. Consequently, for each of these structures, we have a notion of a linear family. For the first structure, these are the so-called mixture families

$$p(x;\eta) = c(x) + \sum_{i=1}^{d} g^i(x)\eta_i,$$

depending on functions g^i and c (which has to be adjusted to make $p(\cdot;\eta)$ into a probability measure), where $\eta_1,\dots,\eta_n$ are the parameters. For the second structure, we have the exponential families

$$p(x;\vartheta) = \exp\bigl(\gamma(x) + f_i(x)\vartheta^i - \psi(\vartheta)\bigr), \tag{1.32}$$

depending on functions f_i (observables in a statistical mechanics interpretation) with parameters ϑ^i and

$$\psi(\vartheta) = \log \int \exp\bigl(\gamma(x) + f_i(x)\vartheta^i\bigr)\,dx \tag{1.33}$$

being the normalization required to make $p(\cdot;\vartheta)$ a probability distribution. In fact, in statistical mechanics, ψ is known as the *free energy*. Of course, we can try to write one and the same family in either the η or the ϑ parameters. Of course, the relationship between them will be nonlinear. Remarkably, when working with the ϑ parameters, we can obtain the Fisher metric from

$$g_{ij} = \partial_i\partial_j\psi(\vartheta),$$

and the Amari–Chentsov tensor from

$$T_{ijk} = \partial_i\partial_j\partial_k\psi(\vartheta),$$

where ∂_i is the derivative w.r.t. ϑ^i. In particular, $\psi(\vartheta)$ is a strictly convex function because its second derivatives are given by the Fisher metric, hence are positive definite. It is important to point out at this stage that convexity here is meant in the sense of affine geometry, and not in the sense of Riemannian geometry. Convexity here simply means that the matrix of ordinary second derivatives is positive semidefinite, and this property is invariant under affine coordinate transformations only. In Riemannian geometry, one would rather require that the matrix of second covariant derivatives be positive semidefinite, and this property is invariant under arbitrary coordinate transformations because the transformation rules for covariant derivatives involve a metric dependent term that compensates the possible nonlinearities of coordinate transformations. Thus, even though the second derivatives of our function yield the metric tensor, the convexity involved here is an affine notion. (This is somewhat similar to Kähler geometry where a Kähler metric is a Hermitian metric that is locally given by the complex Hessian of some potential function. Here, the allowed transformations are the holomorphic ones, as opposed again to general coordinate transformations. In fact, it turns out that those affine structures that we are considering here, that is, those that are locally derived from some strictly convex potential function, can be seen as real analogues of Kähler structures.) In any case, since we are dealing with a convex function, we can therefore pass to its Legendre transform φ. This also induces a change of parameters, and remarkably, this is precisely the transition from the ϑ to the η parameters. With respect to the latter, φ is nothing but the negative of the *entropy* of the probability distribution p, that is,

$$\varphi = \int p(x;\vartheta)\log p(x;\vartheta)\,dx. \tag{1.34}$$

This naturally yields the relations for the inverse metric tensor

$$g^{ij} = \partial^i\partial^j\varphi(\eta), \tag{1.35}$$

where now ∂^i is a derivative w.r.t. η_i. Moreover, we have the duality relations

$$\eta_i = \partial_i\psi(\vartheta), \quad \vartheta^i = \partial^i\varphi(\eta).$$

These things will be explained and explored within the formalism of differential geometry.

1.2.5 Information Geometry and Information Theory

The entropy occurring in (1.34) is also known as the *Shannon information*, and it is the basic quantity of information theory. This naturally leads to the question of the relations between the Fisher information, the basic quantity of information geometry, and the Shannon information. The Fisher information is an infinitesimal quantity, whereas the Shannon information is a global quantity. One such relation

is given in (1.35): The inverse of the Fisher metric is obtained from the second derivatives of the Shannon information. But there is more. Given two probability distributions, we have their Kullback–Leibler divergence

$$D_{KL}(q\|p) := \int \big(\log q(x) - \log p(x)\big) q(x)\, dx. \tag{1.36}$$

Here, we shall take as our base measure simply dx.

This quantity is nonnegative ($D_{KL}(q\|p) \geq 0$, with equality only for $p = q$), but not symmetric in q and p ($D_{KL}(q\|p) \neq D_{KL}(p\|q)$ in general), and so, we cannot take it as the square of a distance function. It turns out that the Fisher metric can be obtained by taking second derivatives of $D_{KL}(q\|p)$ w.r.t. p at $q = p$, whereas taking second derivatives there in the dual coordinates w.r.t. q yields the inverse of the Fisher metric. In fact, this non-symmetry makes the relation between Shannon entropy and information geometry more subtle and more interesting, as we shall now briefly explain.

Henceforth, for simplicity, we shall put $\gamma(x) = 0$ in (1.32), as γ will play no essential role for the moment. As in (1.32), we assume that some functions (observables) f_i, $i = 1, \ldots, n$, are given, and that w.r.t. q, they have certain expectation values,

$$\mathbb{E}_q(f_i) = \bar{f}_i, \quad i = 1, \ldots, n. \tag{1.37}$$

For any $0 \leq m \leq n$, we then look for the probability distribution $p^{(m)}$ that has the same expectation values for the functions f_j, $j = 1, \ldots, m$,

$$\mathbb{E}_{p^{(m)}}(f_j) = \bar{f}_j, \quad j = 1, \ldots, m, \tag{1.38}$$

and that maximizes the entropy

$$H\big(p^{(m)}\big) = -\int \log p^{(m)}(x)\, p^{(m)}(x)\, dx \tag{1.39}$$

among all distributions satisfying (1.38). An easy calculation shows that such a $p^{(m)}$ is necessarily of the form (1.32) on the support of $p^{(m)}$, that is,

$$p^{(m)}(x) = \exp\left(\sum_{j=1}^{m} f_j(x)\vartheta^j - \psi(\vartheta)\right) =: p^{(m)}(x;\vartheta) \tag{1.40}$$

for suitable coefficients ϑ^j which are determined by the requirement (1.38). This means that among all distributions with the same expectation values (1.38), the exponential distribution (1.40) has the largest entropy. For this $p^{(m)}(x;\vartheta)$, the Kullback–Leibler distance from q becomes

$$D_{KL}\big(q\,\|\,p^{(m)}\big) = -H(q) + H\big(p^{(m)}\big). \tag{1.41}$$

Since, as noted, $p^{(m)}$ maximizes the entropy among all distributions with the same expectation values for the f_j, $j = 1, \ldots, m$, as q, the Kullback–Leibler divergence in (1.41) is nonnegative, as it should be. Moreover, among all exponential distributions $p(x; \theta)$, we have

$$D_{KL}\big(q \,\|\, p^{(m)}(\cdot; \vartheta)\big) = \inf_{\theta} D_{KL}\big(q \,\|\, p^{(m)}(\cdot; \theta)\big) \tag{1.42}$$

when the coefficients ϑ^j are chosen to satisfy (1.38), as we are assuming. That is, among all such exponential distributions that with the same expectation values as q for the functions f_j, $j = 1, \ldots, m$, minimizes the Kullback–Leibler divergence. We may consider this as the projection of the distribution q onto the family of exponential distributions $p^{(m)}(x; \theta) = \exp(\sum_{j=1}^{m} f_j(x)\theta^j - \psi(\theta))$. Since, however, the Kullback–Leibler divergence is not symmetric, this is not obtained by the geodesic projection w.r.t. the Fisher metric; rather, two affine flat connections enter which are dual w.r.t. the Fisher metric. These affine flat connections come from the two affine flat structures described above.

The procedure can be iterated w.r.t. m, by projecting $p^{(m)}(\cdot; \vartheta)$ onto the exponential family of distributions $p^{(m-1)}(\cdot; \theta)$. As defined above in (1.38), these families are obtained by fixing the expectation values of more and more observables. For $m = 0$, we simply obtain the uniform distribution p_0, and we have

$$D_{KL}\big(q \,\|\, p^{(k)}\big) = D_{KL}\big(q \,\|\, p^{(n)}\big) + D_{KL}\big(p^{(n)} \,\|\, p^{(n-1)}\big) + \cdots + D_{KL}\big(p^{(k+1)} \,\|\, p^{(k)}\big) \tag{1.43}$$

for $k = 0, \ldots, n$, as will be shown in Sect. 4.3. This decomposition will be systematically explored in Sect. 6.1.

Other applications of information geometry presented in Chap. 6 will include Monte Carlo methods, infinite-dimensional Gibbs families, and evolutionary dynamics. The latter concerns the dynamics of biological populations subjected to the effects of selection, mutation, and random sampling. Those structures can be naturally interpreted in terms of information geometric concepts.

1.3 Historical Remarks

In 1945, in his fundamental paper [219] (see also [220]), Rao used Fisher information to define a Riemannian metric on a space of probability distributions and, equipped with this tool, to derive the Cramér–Rao inequality. Differential geometry was not well-known at that time, and only 30 years later, in [89], Efron extended Rao's ideas to the higher-order asymptotic theory of statistical inference.[1] He defined smooth subfamilies of larger exponential families and their statistical

[1] In [11, p. 67] Amari uncovered a less known work of Harold Hotelling on the Fisher information metric submitted to the American Mathematical Society Meeting in 1929. We refer the reader to [11] for details.

curvature, which, in the language of Riemannian geometry, is the second fundamental form of the subfamilies regarded as Riemannian submanifolds in the Riemannian manifold of the underlying exponential family provided with the Fisher metric. (In [16, p. 23] statistical curvature is also called *embedding curvature* or *e-curvature* and totally geodesic submanifolds are called *autoparallel submanifolds*.) Efron named those smooth subfamilies "curved exponential families." In 1946–1948, the geophysicist and Bayesian statistician Jeffreys introduced what we today call the Kullback–Leibler divergence, and discovered that for two distributions that are infinitely close we can write their Kullback–Leibler divergence as a quadratic form whose coefficients are given by the elements of the Fisher information matrix [131, 132]. He interpreted this quadratic form as the length element of a Riemannian manifold, with the Fisher information playing the role of the Riemannian metric. From this geometrization of the statistical model, he derived his prior distributions as the measures naturally induced by the Riemannian metric.

In 1955, in his lectures at the H. Poincaré Institute, Kolmogorov discussed the problem of the existence of *natural* differentiable structures on ensembles of probability distribution. Following a suggestion by Morozova, see [188], Chentsov defined an affine flat connection (the e-connection) on the set $\mathcal{P}_+(\Omega, \mu)$ [60]. Further, analyzing the "naturality" condition for differentiable structures, Chentsov invented the category of mathematical statistics [61]. This category was introduced independently and almost at the same time by Morse and Sacksteder [189], using foundational ideas of Wald [252] and Blackwell [49] in the statistical decision theory and under the influence of the categorical approach in algebraic topology that was very fashionable at that time. The morphisms in the Chentsov category of mathematical statistics are Markov morphisms and geometric notions on probability distribution ensembles are required to be invariant under Markov morphisms [65]. In his influential book [65] Chentsov considered only geometry on probability distribution spaces $\mathcal{P}_+(\Omega, \mu)$ for finite sample spaces Ω, referring to technical difficulties of treating infinite-dimensional differentiable manifolds. The only exceptions are curved exponential families—subfamilies of the canonical exponential families defined by the e-connection in [60]. Using the categorical approach, in particular Markov morphisms and a related notion of *congruent embedding*, see Definition 5.1, Chentsov discovered the Amari–Chentsov connections and proved the uniqueness of the Amari–Chentsov structure by their invariance under sufficient statistics [65].

Independently, inspired by the Efron paper and Dawid's discussion on it [89], Amari defined in [6, 7] the notion of α-connections and showed its usefulness in the asymptotic theory of statistical estimation. In particular, using geometric methods, Amari achieved Fisher's life-long dream of showing that the maximal likelihood estimator is optimal [7, 8, 16], see also Sect. 5.2 below. Shortly after this, Amari and Nagaoka introduced the notion of dual connections, developed the general theory of dually flat spaces, and applied it to the geometry of α-connections [194].

These events prepared the birth of information geometry, whose name appeared for the first time (in English) in [15], which was known before also as the differential geometrical theory of statistics. (Certain aspects of information geometry, e.g.,

results due to Morozova and Chentsov concerning invariants of pairs of probability distributions, belong to Markovian categorial geometry [63, 188], which is not necessarily a part of differential geometry. Thus the name "information geometry" is more suitable, and it also sounds more attractive.) The field of information geometry developed in particular thanks to the work of Amari and his school. Further developments of information geometry were mainly focused on divergence functions and their generalizations, and devoted to applications in statistical inference and information theory, especially in higher-order asymptotic theory of statistical inference. Here we would like to mention the papers by Csiszár [72–74] on divergence functions, especially f-divergences and their invariant properties, and by Eguchi on the dualistic geometry of general divergences (contrast functions) [91–93], and the papers on the geometry of manifolds with dually flat connections [13, 34, 59, 236]. We recommend [11, 56] for a survey and bibliography on divergence geometry and [9, 11, 16, 148, 188] for a survey of applications of information geometry in the early period. Later applications of information geometry include neural networks, machine learning, evolutionary biology, etc. [11, 16]; see also Chap. 6 in our book. Regarding Markovian categorial geometry, in addition to the aforementioned papers by Chentsov and Morozova–Chentsov, we also would like to mention the paper by Campbell [57] on an extension of the Chentsov theorem for finite sample spaces, which will be significantly generalized in Chaps. 2 and 5. The papers [186, 187] (see also [188, §6]) by Morozova and Chentsov on the geometry of α-connections, in particular, giving an explicit description of totally geodesic submanifolds of the manifold $\mathcal{M}_+(I)$ provided with an α-connection, belong to the intersection of Markovian categorial geometry and divergence geometry. We note that the dualistic geometries considered in the early period of information geometry, in particular the geometry of curved exponential families, are not necessarily related to finite sample spaces but they are supposed to be finite-dimensional [39, 41, 62, 90]. Amari [9], Lauritzen [160], Murray–Rice [192] have proposed general concepts of finite-dimensional statistical models. Among further advancements in information geometry are Lauritzen's introduction of the notion of statistical manifolds and Lê's immersion theorem, which we shall discuss in Chap. 4. The infinite-dimensional information geometry and, in particular, infinite-dimensional families of probability distributions were first considered in Pistone–Sempi's work [216] in 1995, see also our discussion in Chap. 3, and later in subsequent papers by Pistone and coauthors; see, e.g., [106], and recently in [25], which combines the approach of Markovian categorial geometry with functional analytical techniques. As an application we have proved a version of the Chentsov theorem which will be generalized further in this book. We would also like to mention [164] for another view on Chentsov's theorem and its generalizations.

Finally, we note that information geometry has been generalized for quantum systems, but the related circle of questions lies outside the scope of our book, and we refer the interested reader to excellent reviews in [16, 188, 214].

1.4 Organization of this Book

Let us summarize the organization of our book. After this introduction, in Chap. 2, we shall explain the basic constructions in the finite-dimensional case, that is, when the underlying space Ω on which we study probability distributions has only finitely many elements. This chapter also provides the natural context for a formal treatment of interaction spaces and of hierarchical models, emphasizing the important special class of graphical models. The space of probability distributions on such a finite space as treated in this chapter is finite-dimensional. In the next Chap. 3, we consider a general space Ω. Consequently, we shall have to deal with infinite-dimensional spaces of probability measures, and technical complications emerge. We are able, however, to develop a functional analytic framework within which these complications can be overcome. We shall introduce and develop the important notion of parametrized measure models and define suitable integrability properties, in order to obtain the analogue of the structures considered in Chap. 2. These structures will not depend on the choice of a base measure because we shall set up the framework in such a way that all objects transform appropriately under a change of base measure. We shall also discuss the structure of Pistone and Sempi. The following Chap. 4 will develop the differential geometry of statistical models. This includes dualistic structures, consisting of a Riemannian metric and two connections that are dual to each other with respect to that metric. When these connections are torsion-free, such a structure can more compactly be described as a statistical model, given by a metric, that is, a symmetric positive definite 2-tensor, and a symmetric 3-tensor. These are abstract versions of the Fisher metric and the Amari–Chentsov tensor. Alternatively, it can be described through a divergence. Any statistical model can be isostatistically immersed into a standard model defined by an Amari–Chentsov structure, and this then provides the link between the abstract differential geometry of Chap. 4 and the general functional analysis of Chap. 3. When these connections are even flat, the structure can be locally obtained from potential functions, that is, convex functions whose second derivatives yield the metric and whose third derivatives yield the 3-tensor. Here, convexity is considered as an affinely invariant notion, and consequently, we need to discuss the underlying affine structure. This also gives rise to a dual structure via the Legendre transform of the convex function. That is, we shall find a pair of dual affine structures, and this is the geometry discovered by Amari and Chentsov. Chapter 5 will turn to the statistical aspects. We shall present one of the main results of information geometry, that the Fisher metric and the Amari–Chentsov tensor are characterized by their invariance under sufficient statistics. Our treatment of sufficient statistics here is more general than what can be found in statistics texts. Also, we shall discuss estimators and derive a general version of the Cramér–Rao inequality within our framework. In the last chapter, we shall connect our treatment of information geometry with various applications and other fields. Building upon the treatment of interaction spaces and of hierarchical models in Chap. 2, we shall describe information theoretical complexity measures and applications of information geometry to Markov chains. This yields an approach to the analysis of systems of interacting units. Moreover, we shall discuss how information geometry provides a natural setting for the basic structures of mathematical

biology, like mathematical population genetics, referring for a more detailed presentation to [124], however. We shall also briefly sketch a formal relationship of information geometry with functional integrals, the Gibbs families of statistical mechanics. In fact, these connections with statistical physics already shine through in Sect. 4.3. Here, however, we only offer an informal way of thinking without technical rigor. We hope that this will be helpful for a better appreciation of the meaning of various concepts that we have treated elsewhere in this book. In an appendix, we provide a systematic, but brief, overview of the basic concepts and results from measure theory, Riemannian geometry and Banach manifolds on which we shall freely draw in the main text.

The standard reference for the development based on differential geometry is Amari's book [8] (see also the more recent treatment [16], and also [192]). In fact, from statistical mechanics, Balian et al. [38] arrived at similar constructions. The development of information geometry without the requirement of a differentiable structure is based on Csiszár's work [75] and has been extended in [76] and [188]. In identifying unique mathematical structures of information geometry, invariance assumptions due to Chentsov [65] turn out to be fundamental. A systematic approach to the theory has been developed in [25]. The geometric background material can be found in [137].

When we speak about geometry in this book, we mean differential geometry. In fact, however, differential geometry is not the only branch of geometry that is useful for statistics. Algebraic geometry is also important, and the corresponding approach is called algebraic statistics. Algebraic statistics treats statistical models for discrete data whose probabilities are solution sets of polynomial equations of the parameters. It uses tools of computational commutative algebra to determine maximum likelihood estimators. Algebraic statistics also works with mixture and exponential families; they are called linear and toric models, respectively. While information geometry is concerned with the explicit representation of models through parametrizations, algebraic statistics highlights the fact that implicit representations (through polynomial equations) provide additional important information about models. It utilizes tools from algebraic geometry in order study the interplay between explicit and implicit representations of models. It turns out that this study is particularly important for understanding closures of models. In this regard, we will present implicit descriptions of exponential families and their closures in Sect. 2.8.2. In the context of graphical models and their closures [103], this leads to a generalization of the Hammersley–Clifford theorem. We shall prove the original version of this theorem in Sect. 2.9.3, following Lauritzen's presentation [161]. In this book, we do not address further aspects of algebraic statistics and refer to the monographs [84, 107, 209, 215] and the seminal work of Diaconis and Sturmfels [82].

When we speak about information, we mean classical information theory à la Shannon. Nowadays, there also exists the active field of quantum information theory. The geometric aspects of quantum information theory are explained in the monograph [45].

We do not assume that the reader possesses a background in statistics. We rather want to show how statistical concepts can be developed in a manner that is both

rigorous and general with tools from information theory, differential geometry, and functional analysis.

On the Notation and Some Conventions

We have two kinds of spaces, a measure space Ω on which our measures live, and a parameter space M that parametrizes a family of measures. The elements of Ω will be denoted by either x or ω. Depending on the circumstances, Ω may carry some further structure, like a topology or a differentiable structure. When the measure space is finite, we denote it by I instead, and its elements by i, to conform to standard conventions. In the finite case, these elements i will usually be written as indices.

The parameters in M will usually be denoted by ξ. When we have particular parametrized families of measures, we use other Greek minuscules, more precisely ϑ for the parameters of exponential families, and η for those of mixture families. An estimator for the parameter ξ is written as $\hat{\xi}$, as usual in the statistics literature.

We use the letter μ, ν, or m to indicate a general finite measure, and p to stand for a probability measure. Calligraphic letters will stand for spaces of measures, and so, $\mathcal{M}(\Omega)$ and $\mathcal{P}(\Omega)$ will denote spaces of general or probability measures on Ω. Since our (probability) measures will live on Ω, but depend on a parameter $\xi \in M$, we write $p(x;\xi)$ to indicate these dependencies. When the element of Ω plays no role, we may also simply write $\mathbf{p}(\xi)$. Thus, in the context of ξ-dependence, $\mathbf{p}(\xi)$ is a general finite measure, not necessarily a probability measure.

We shall often need to use some base measure on Ω, which will be denoted by μ or μ_0. The integration of an integrable (w.r.t. μ) function f, that is, $f \in L^1(\Omega,\mu)$, will be written as $\int f(x)d\mu(x)$; thus, when we carry out an integration, we shall write $d\mu$ in place of μ. Also, the pairing between a function $f \in L^\infty(\Omega,\mu)$ and a measure $\phi\mu$ with $\phi \in L^1(\Omega,\mu)$ will be written as $(f,\phi) = \int f(x)\phi(x)d\mu(x)$ whereas an L^2-product will be denoted by $\langle h,k\rangle = \int h(x)k(x)d\mu(x)$.

In the finite case, we shall use somewhat different conventions. Of course, we shall then use sums in place of integrals. Here, $\Sigma^{n-1} = \{(p_1,\dots,p_n) : p_i \geq 0, \sum_j p_j = 1\}$ is the unit simplex of probability measures on a space I of n elements. As will become clear in our main text, instead of Σ^{n-1}, it is often more natural to consider the positive sector of the sphere; for a somewhat annoying technical reason, it is better to take the sphere of radius 2 instead of the unit sphere. Therefore, we shall work with $S^{n-1}_{2,+} = \{(q^1,\dots,q^n) : q^i \geq 0, \sum_j (q^j)^2 = 4\}$. Σ^{n-1} and $S^{n-1}_{2,+}$ are homeomorphic, of course, but often, the latter is better suited for our purposes than the former.

$\mathbb{E}_p$ will mean the expectation value w.r.t. the (probability) measure p.

We shall often use the normal or Gaussian distribution on $\mathbb{R}^d$ with center or mean x and covariance matrix $\Lambda = (\lambda_{ij})$,

$$\mathcal{N}(y;x,\Lambda) = \frac{1}{\sqrt{2\pi^d|\Lambda|}} \exp\left(-\frac{\lambda^{ij}(x^i-y^i)(x^j-y^j)}{2}\right), \tag{1.44}$$

putting $|\Lambda| := \det(\lambda_{ij})$ and denoting the inverse of Λ by $\Lambda^{-1} = (\lambda^{ij})$, and where the standard summation convention is used (see Appendix B). We shall also simply write $\mathcal{N}(x, \Lambda)$ for $\mathcal{N}(\cdot; x, \Lambda)$.

Our key geometric objects are the Fisher metric and the Amari–Chentsov tensor. Therefore, they are denoted by single letters, $\mathfrak{g}$ for the Fisher metric, and $\mathbf{T}$ for the Amari–Chentsov tensor. Consequently, for instance, other metrics will be denoted differently, by g or by other letters, like h.

We shall always write log for the logarithm, and in fact, this will always mean the natural logarithm, that is, with basis e.

http://www.springer.com/978-3-319-56477-7

Information Geometry

Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L.

2017, XI, 407 p. 15 illus., Hardcover
ISBN: 978-3-319-56477-7