FOUNDATIONS OF STRUCTURAL STATISTICS: STATISTICAL MANIFOLDS

P. MICHL <PATRICK.MICHL@GMAIL.COM>

Abstract. Upon a consistent topological statistical theory the application of Structural Statistics required a quantification of the proximity structure of model spaces. An important tool to study these structures are (Pseudo-)Riemannian metrices, which in the category of statistical models are induced by statistical divergences. The present article is intended to extend the notation of topological statistical models by a differential structure to statistical manifolds and to introduce the differential geometric foundation to study exponential families. Exponential families are of particular importance with respect to Machine Learning and Deep Learning. In this purpose the article successively incorporates the structures of differential-, Riemannian- and symplectic geometry within an underlying topological statistical model. The last section of the chapter addresses a specific structural category, termed a dually flat statistical manifold.

1. Introduction

Since the 1960s further investigations on the invariant structures of statistical models led to various statistical theories. Notably the contributions of N. N. Chenzow [2] and Shun'ichi Amari [1], which have been concerned with the differential structure of statistical models, eventually encouraged the notation of *statistical manifolds.* Thereby Amari´s methods of *information geometry* emphasizes the *Fisher information metric*, which is obtained as a partial derivative of the Kullback-Leibler divergence. Unfortunately during the 1990s the supplementary degree of abstraction compared to its initially low applicability caused the theory to lose "momentum". Nevertheless, in the first decade of the 21st century, the increasing availability of large samples of natural observations like natural hand writing, audio samples or gene expression data, demanded a theoretic underpinning of "geometric data analysis" and "computational geometric data analysis", in particular w.r.t. the uprising Machine Learning and Deep Learning.

2. Statistical Manifolds

Topological statistical models provide the ability to characterise statistical inference without the necessity of an underlying sample space. Thereby the statistical equivalence of statistical models is provided by Kolmogorov equivalence. The topologies of those Kolmogorov quotients in turn are obtained by countable coverings of Borel sets in $\mathbb{R}$ and therefore provide the the probability to be second countable Hausdorff spaces. It is therefore straight forward to incorporate the concept of manifolds to statistical models by the Kolmogorov quotients of their induced topological statistical models. Let therefore $(S, \Sigma, \mathcal{M}, \mathcal{T})$ be topological statistical model and $n \in \mathbb{N}$. Then a *coordinate chart* (U, ϕ) within $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$ is constituted by an open set $U \in {}^{\mathcal{T}}/_{\mathrm{id}}$ and a homeomorphism $\phi : U \to \mathbb{R}^n$ into $\mathbb{R}^n$. This allows the definition of an *atlas* $\mathcal{A}$ for $\mathcal{M}$ by a family of charts $\{(U_i, \phi_i)\}_i$, that covers ${}^{\mathcal{M}}/_{\mathrm{id}}$, such that ${}^{\mathcal{M}}/_{\mathrm{id}} = \bigcup_{i\in I} U_i$. In order to extend the local Euclidean structure of the individual coordinate charts, to a global structure

over the model space, the transitions within overlapping charts are required to preserve the structure. Let therefore (U_a, ϕ_a) and (U_b, ϕ_b) be coordinate charts in $\mathcal{A}$ with a nonempty intersection $U_{a\cap b} = U_a \cap U_b$, then $\phi_a(U_{a\cap b})$ and $\phi_b(U_{a\cap b})$ generally denote different representations of $U_{a\cap b}$ in $\mathbb{R}^n$. In this case for a given $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$, the charts are regarded to be C^k-compatible, iff their *transition maps* $\phi_a \circ \phi_b^{-1}$ and $\phi_b \circ \phi_a^{-1}$ are k-times continuously differentiable. If all charts of an atlas $\mathcal{A}$ are pairwise C^k-compatible, then $\mathcal{A}$ is termed a C^k-atlas. Let then be $\mathcal{A}'$ a further C^k-atlas of $\mathcal{M}$, then $\mathcal{A}$ and $\mathcal{A}'$ are termed C^k-equivalent, if also $\mathcal{A} \cup \mathcal{A}'$ is a C^k-atlas of $\mathcal{M}$. This equivalence relationship may be used to derive a maximal atlas by completion. Let therefore $\mathcal{A}_{\max}$ be the union of all C^k-atlases of $\mathcal{M}$, that are C^k-equivalent to $\mathcal{A}$, then $\mathcal{A}_{\max}$ is unique for the C^k-equivalence class of $\mathcal{A}$ and does not depend on the choice of $\mathcal{A}$ within this class. Then any C^k-differentiable function, that is defined within the image of a chart in $\mathcal{A}_{\max}$ has a unique C^k-differentiable extension within its neighbourhood in $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$. The crux in the definition of a C^k-atlas $\mathcal{A}$ however is, that due to the Hausdorff property of $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$ and the completion of $\mathcal{A}$ by$\mathcal{A}_{\max}$ the requirement of the transition functions to be C^k-diffeomorphism in $\mathbb{R}^n$ induces a differential structure to $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$. Then not only the transition maps, but any coordinate chart by itself may be regarded as a C^k-diffeomorphism into $\mathbb{R}^n$. This defines the structure of a *statistical manifold.*

Definition (Statistical manifold)**.** *Let* $(S, \Sigma, \mathcal{M}, \mathcal{T})$ *be a topological statistical model and* $\mathcal{A}$ *an* n*-dimensional* C^k*-atlas for* $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$*. Then the tuple* $(S, \Sigma, \mathcal{M}, \mathcal{A})$ *is termed a statistical manifold.* **Remark**: *The category of* k*-differentiable statistical manifolds is denoted by* $\mathbf{StatMan}^k$*.*

Since the atlas $\mathcal{A}$ has to be defined with regard to the Kolmogorov quotient $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$ to assure the Hausdorff property, statistical manifolds have technically to be regarded as non-Hausdorff manifolds. Since the atlas $\mathcal{A}$ conversely induces a topology that equals $\mathcal{T}/_{\mathrm{id}}$ the original topological statistical model $(S, \Sigma, \mathcal{M}, \mathcal{T})$ may not be derived by $(S, \Sigma, \mathcal{M}, \mathcal{A})$. Nevertheless by the extension of the Kolmogorov quotient to the atlas $\mathrm{KQ}(\mathcal{M}, \mathcal{A})$, it follows that $\mathrm{KQ}(\mathcal{M}, \mathcal{T})$ and $\mathrm{KQ}(\mathcal{M}, \mathcal{A})$ are Kolmogorov equivalent and therefore that $(S, \Sigma, \mathcal{M}, \mathcal{T})$ and $(S, \Sigma, \mathcal{M},$ are induced by statistical equivalent models. With regard to observation based statistical inference this "irregularity" however usually has no impact, since for any *identifiable statistical manifold* $(S, \Sigma, \mathcal{M}, \mathcal{A})$, which model space $\mathcal{M}$ is identical to a parametric family $\mathcal{M}_\theta$, it holds that $\mathcal{M}/_{\mathrm{id}} = \mathcal{M}_\theta = \mathcal{M}$ and therefore that $\mathrm{KQ}(\mathcal{M}, \mathcal{A}) = (\mathcal{M}, \mathcal{A})$. Without loss of generality in the following therefore $(S, \Sigma, \mathcal{M}, \mathcal{A})$ is assumed to be an identifiable statistical manifold and therefore a manifolds in the usual context. Nevertheless, in order to provide higher structures, it is reasonable to recapitulate the usual concepts and the vocabulary of manifolds. First of all by assuming $\mathcal{A}$ to be a C^0-atlas, the transition maps, that define the structure of $(S, \Sigma, \mathcal{M}, \mathcal{A})$ are only required to be continuous and $(S, \Sigma, \mathcal{M}, \mathcal{A})$ is termed *topological.* For the case, that $\mathcal{A}$ is a C^k-atlas with $k > 0$ however, the transition functions at least have to be differentiable and therefore $(S, \Sigma, \mathcal{M}, \mathcal{A})$ is termed *differentiable.* Let now be $(\mathcal{N}, \mathcal{B})$ a further identifiable statistical manifold, where $\mathcal{B}$ is an m-dimensional C^k-atlas of $\mathcal{N}$ and $f\colon \mathcal{M} \to \mathcal{N}$ a function, that is is continuous w.r.t. the induced topologies. Then f is C^k-*differentiable* and written as $f \in C^k(\mathcal{M}, \mathcal{N})$, if for arbitrary coordinate charts $(U, \phi) \in \mathcal{A}$ and $(V, \nu) \in \mathcal{B}$ with $f(U) \subseteq V$ it holds, that $\nu \circ f \circ \phi^{-1}$ is k-times continuously differentiable. Thereby the case $(\mathcal{N}, \mathcal{B}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ occupies an exceptional position, which is abbreviated by the notation $C^k(\mathcal{M})$. At this point it is important to notice,

that although the definitions of C^k-atlases and C^k-differentiable functions depend on the choice of k, this does essentially not apply for the underlying differentiable structures. The reason for this "peculiarity" may be found in the property, that for any $k > 0$ any C^k-atlas uniquely admits a "smoothing", given by a C^k-equivalent C^∞-atlas. Therefore the set of smooth functions $C^\infty(\mathcal{M})$ is well defined, independent of the underlying differentiable structure. Thereby $C^\infty(\mathcal{M})$ constitutes an associative algebra w.r.t. the pointwise product "$\cdot$", the addition "$+$" and the scalar multiplication. This allows a formal definition of *derivations* at points $P \in \mathcal{M}$ by $\mathbb{R}$-linear functions $D\colon C^\infty(\mathcal{M}) \to \mathbb{R}$, that satisfies the Leibniz rule $D(\varphi \cdot \psi) = D(\varphi)\psi(P) + \varphi(P)D(\psi)$ for all $\varphi, \psi \in C^\infty(\mathcal{M})$. Let now be $T_P\mathcal{M}$ the set of all derivations at P, then $T_P\mathcal{M}$ defines an n-dimensional $\mathbb{R}$-vector space, by the operations $(v+w)(\varphi) \coloneqq v(\varphi) + w(\varphi)$ and $(\lambda v)(\varphi) \coloneqq \lambda v(\varphi)$, for $v, w \in T_P\mathcal{M}$, $\varphi \in C^\infty(\mathcal{M})$ and $\lambda \in \mathbb{R}$. As for any given coordinate chart (U, ϕ) that contains P, any derivation $v \in T_P\mathcal{M}$ uniquely corresponds to a directional derivative in $\mathbb{R}^n$ at the point $\phi(P)$, the elements of $T_P\mathcal{M}$ are termed *tangent vectors* and $T_P\mathcal{M}$ the *tangent space* at P. Then the *partial derivatives* at P, given by $\{\partial_i\}_P$ with $\partial_i\colon P \mapsto \partial/\partial\phi^i \mid_P$ provide a basis of $T_P\mathcal{M}$, such that any $v \in T_P\mathcal{M}$ has a local *representation* by a vector $\boldsymbol{\xi} \in \mathbb{R}^n$ with $v = \xi^i\partial_i$. Let now be $f \in C^\infty(\mathcal{M}, \mathcal{N})$, then the *differential* of f at $P \in \mathcal{M}$ is a linear mapping $\mathrm{d}f_P\colon T_P\mathcal{M} \to T_{f(P)}\mathcal{N}$, which for all $v \in T_P\mathcal{M}$ and $\varphi \in C^\infty(\mathcal{N})$ is defined by $(\mathrm{d}f_P v)(\varphi) \coloneqq v(\varphi \circ f)$. Then f is an *immersion*, if for all $P \in \mathcal{M}$ the differential $\mathrm{d}f_P$ is injective. If furthermore f is injective and continuous w.r.t. to the respectively induced topologies, then f is a *smooth embedding* and the image of f a *smooth submanifold* of $\mathcal{N}$ w.r.t. the atlas, which is restricted to the image. This allows the definition of *smooth parametrisations*.

Definition (Smooth parametrisation)**.** *Let* $(\mathcal{M}, \mathcal{A})$ *be a differentiable statistical manifold,* $(V, \mathcal{B})$ *a differentiable manifold over a vector space* V *and* θ *a parametrisation for* $\mathcal{M}$ *over* V*. Then* θ *is termed a smooth parametrisation for* $(\mathcal{M}, \mathcal{A})$*, iff* $\theta^{-1}\colon \mathrm{KQ}(\mathcal{M}, \mathcal{A}) \to (V, \mathcal{B})$ *is a smooth embedding.*

Since for any smooth n-manifold the *Whitney embedding theorem* postulates the existence of a smooth embedding within $\mathbb{R}^{2n}$ any smooth statistical n-manifold $(\mathcal{M}, \mathcal{A})$ has a smooth parametrisation θ over $\mathbb{R}^{2n}$. It is therefore convenient to introduce the notation "$(\mathcal{M}_\theta, \mathcal{A})$" for a differentiable statistical manifold with a smooth parametrisation θ. Since the tuple $(\mathcal{M}, \theta^{-1})$ is an C^∞-chart that covers $\mathcal{M}$, it provides a *smooth representation* of $(\mathcal{M}, \mathcal{A})$ by the parameter space $\Theta = \theta^{-1}(\mathcal{M})$. Thereby for any coordinate chart $(U, \phi) \in \mathcal{A}$, the mapping $\phi\colon U \to \mathbb{R}^n$ provides an n-dimensional basis for the tangent space $T_P\mathcal{M}$ by the partial derivatives $\{\partial_i\}_P$. Since the smooth parametrisation θ however is an immersion, also the differentials $\mathrm{d}\theta^{-1}(\partial_i)$ also provide an n-dimensional basis of the space $T_{\theta^{-1}(P)}\Theta$. Therefore any tangent space may intuitively be identified with an n-dimensional affine subspace of Θ and any tangent vector by a directional derivative within this subspace. A smooth parametrisation then in particular allows the identification of *smooth curves* $\gamma\colon I \to \mathcal{M}$ in $\mathcal{M}$ by smooth *parametric curves* $\gamma_\theta\colon I \to \Theta$ in Θ, such that $\gamma = (\theta \circ \gamma_\theta)$. Then for any $k \in \mathbb{N}_0 \cup \{\infty, \omega\}$, it holds that $\gamma \in C^k(I, \mathcal{M})$ iff $\gamma_\theta \in C^k(I, \Theta)$. Therefore smooth parametric curves provide the foundation for a traversal of $\mathcal{M}$. Thereby the traversal along the parametric curve $\gamma_\theta(t)$ is described by the unique directional derivatives $\dot{\gamma}_\theta(t)$ in $\times$. Consequentially due to the unique correspondence between directional derivatives in $T_{\gamma_\theta(t)}\times$ and tangent vectors in $T_{\gamma(t)}\mathcal{M}$ the traversal along γ in $\mathcal{M}$ is also described by unique tangent vectors $\dot{\gamma}(t) \in T_{\gamma(t)}\mathcal{M}$. This uniqueness however only applies to

the direction in Θ but not to its representation in $T_{\gamma_\theta(t)}\times$ in terms of the chosen basis. With regard to a coordinate chart $(U,\, \phi)$ in $\mathcal{M}$ the local basis $\{\partial_i\}_P$ at $P \in U$ naturally extends over U, by regarding $\partial_i \colon P \mapsto \partial/\partial\phi^i$ as an ordered basis, termed a *local frame*, which localized as P provides $\partial_i \colon P \mapsto \partial/\partial\phi^i \mid_P$. Then the differentials $\mathrm{d}\theta^{-1}(\partial_i)$ provide a local basis of $T_{\gamma_\theta(t)}\Theta \mid_{\theta^{-1}(U)}$ and the directional derivatives $\dot\gamma_\theta(t) \in T_{\gamma_\theta(t)}\Theta \mid_{\theta^{-1}(U)}$ may uniquely be identified with tangent vectors $\dot\gamma(t) \in T_{\gamma(t)}\mathcal{M} \mid_U$ by $\dot\gamma(t) = \mathrm{d}\theta(\dot\gamma_\theta(t)) \mid_U$. In order to continue the traversal however it is required to "connect" the basis vectors of the affine spaces along the curve by an unambiguous notation, which is independent of the chosen coordinate charts. This provides the notation of an *affine connection*. At a global scale the disjoint union of all tangent spaces constitutes the *tangent bundle* $T\mathcal{M}$, which by itself is diffeomorphic to $\mathcal{M}\times\mathbb{R}^n$ and therefore in particular a differentiable manifold. This property allows to define *smooth vector fields* on $\mathcal{M}$ by smooth functions $X \in C^\infty(\mathcal{M},\, T\mathcal{M})$, or w.r.t. the sequence $\mathcal{M} \overset{X}{\hookrightarrow} T\mathcal{M} \twoheadrightarrow \mathcal{M}$ by smooth sections $X \in \Gamma(T\mathcal{M})$. Intuitively smooth vector fields assign tangent vectors to the points of the manifold, such that "small " movements on the manifold are accompanied by "small" changes within the tangent spaces. With regard to a coordinate chart $(U,\, \phi)$ a local frame $\{\partial_i\}$ may also be regarded as a localized ordered basis of the vector fields $\Gamma(T\mathcal{M} \mid_U)$. Therefore the transition of local frames may be described be derivatives of vector fields, which provides the notation of *covariant derivatives*. A covariant derivative ∇ on $\mathcal{M}$ formally defines a mapping $\nabla \colon \Gamma(T\mathcal{M})^2 \to \Gamma(T\mathcal{M})$ with $(X,\, Y) \mapsto \nabla_X Y$, which satisfies: (i) ∇ is $\mathbb{R}$-linear in both arguments, (ii) ∇ is $C^\infty(\mathcal{M})$-linear in the first argument and (iii) ∇ is a derivation in the second argument, such that $\nabla_X(\varphi\cdot Y) = X(\varphi)Y + \varphi\nabla_X Y$ for arbitrary $\varphi \in C^\infty(\mathcal{M})$ and $X,\, Y \in \Gamma(T\mathcal{M})$. An affine connection is then completely described by the specification of a covariant derivative which in turn endows a differentiable manifold with an additional structure ∇. In particular however the choice of an affine connection for any curve γ completely determines its derivative $\dot\gamma \in \Gamma(T\mathcal{M}) \mid_\gamma$ along the curve, as well as those vector fields $X \in \Gamma(T\mathcal{M})$ which are covariant constant along γ, such that $\nabla_{\dot\gamma} X = 0$. As this property however may also be applied w.r.t. the derivative along the curve itself, the choice of an affine connection ∇ in particular determines those curves γ, which derivative is covariant constant along their traversal. These curves, known as *geodesic*s, therefore generalize straight lines to differentiable manifolds.

Definition (Geodesic)**.** *Let* $(\mathcal{M},\, \mathcal{A})$ *be a smooth statistical manifold and* ∇ *an affine connection on* $\mathrm{KQ}(\mathcal{M},\, \mathcal{A})$*. Then a smooth curve* $\gamma \colon I \to \mathcal{M}$ *is termed a geodesic w.r.t.* ∇*, iff it satisfies the geodesic equation:*

$$\nabla_{\dot\gamma}\dot\gamma = 0$$

Over and above geodesic, the choice of an affine connection ∇ admits two fundamental invariants to the differentiable structure by the *curvature* and the *torsion*. Thereby the curvature $R \colon \Gamma(T\mathcal{M})^3 \to \Gamma(T\mathcal{M})$ with $R(X,\, Y)Z \coloneqq \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z - \nabla_{[X,\, Y]}Z$ intuitively provides a description, of how tangent spaces "roll" along smooth curves under parallel transport, whereas the torsion $T \colon \Gamma(T\mathcal{M})^2 \to \Gamma(T\mathcal{M})$ with $T(X,\, Y) \coloneqq \nabla_X Y - \nabla_Y X - [X,\, Y]$ describes their "twist" about the curve. Notwithstanding these invariants however, an affine connection ∇ only adjusts local tangent spaces, but does not provide a notation of "length" or "angle". The mandatory next step, therefore regards the incorporation of local geometries within the tangent spaces, that eventually extend to a global geometry over the differentiable structure.

3. Pseudo-Riemannian Structure

Since tangent spaces are vector spaces, it is natural to obtain the local geometry by an inner product. More generally however it suffices to provide a mapping $g_P\colon T_P\mathcal{M}^2 \to \mathbb{R}$ that satisfies (i) g_P is $C^\infty(\mathcal{M})$-bilinear, (ii) g_P is symmetric and (iii) g_P is non-degenerate. In the purpose to extend the local geometries to a global geometry however, it has additionally to be claimed that the local geometries only vary smoothly w.r.t. smooth vector fields. This localization requirement yields the notation of a *pseudo-Riemannian metric* g on $(\mathcal{M}, \mathcal{A})$, which endows each point $P \in \mathcal{M}$ with a symmetric non-degenerate form g_P, such that the mapping

$$g(X, Y)\colon P \mapsto g_P(X_P, Y_P)$$

is smooth, i.e. $g(X, Y) \in C^\infty(\mathcal{M})$, for arbitrary $X, Y \in \Gamma(T\mathcal{M})$. With regard to a coordinate chart (U, ϕ) and a local frame $\{\partial_i\}$ the pseudo-Riemannian metric g has a coordinate representation by *metric coefficients* $g_{ij}\colon U \to \mathbb{R}$, with $g_{ij} = g(\partial_i, \partial_j)$ and therefore by a matrix $G_P = (g_{ij})$, termed a *fundamental matrix*. For $P \in U$ and $v, w \in T_P\mathcal{M}$, with $v = \xi^i\partial_i$ and $w = \zeta^i\partial_i$ it then follows, that $g(v, w)$ has a local representation $\langle \boldsymbol{\xi}, \boldsymbol{\zeta}\rangle_P \coloneqq \boldsymbol{\xi}^T G_P \boldsymbol{\zeta}$. In the purpose to extend the local geometry to a global geometry an affine connection ∇ has to be defined, which is compatible with g, such that $Xg(Y, Z) = g(\nabla_X Y, Z) + g(Y, \nabla_X Z)$, for all $X, Y, Z \in \Gamma(T\mathcal{M})$. In this case ∇ is termed a *metric connection* and has a coordinate representation by *connection coefficients* $\Gamma^k_{ij}\colon U \to \mathbb{R}$, with $\nabla_{\partial_i}\partial_j = \sum_{k=1}^n \Gamma^k_{ij}\partial_k$, known as the *Christoffel-symbols.* Then the geodesic equation over a parametric curve γ_θ is locally expressed by a second order ODE:

$$\nabla_{\dot\gamma_\theta}\dot\gamma_\theta = 0 \iff \ddot\gamma^k_\theta + \sum_{i,j}\Gamma^k_{ij}\dot\gamma^i_\theta\dot\gamma^j_\theta = 0,\ \forall k \tag{3.1}$$

With little effort, the *Picard-Lindelöf theorem* then assures, that for any $(P, v) \in T\mathcal{M}$ there exists a locally unique geodesic $\gamma_{P,v}\colon I \to \mathcal{M}$, that satisfies the initial conditions $\gamma_{P,v}(0) = P$ and $\dot\gamma_{P,v}(0) = v$. Thereby the locally uniqueness extends to an maximal open interval $I = (a, b)$ in $\mathbb{R}$. If ∇ is furthermore *torsion free* i.e. $T(X, Y) = 0$, for all $X, Y \in \Gamma(T\mathcal{M})$, then ∇ is termed a *Levi-Civita connection* and the Christoffel-symbols may explicitly be derived by the equation $\Gamma^k_{ij} = \frac{1}{2}\sum_l g^{kl}(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij})$, where g^{kl} denote the coefficients of the inverse fundamental matrix G_P^{-1}, which existence is assured by the properties of g_P. Therefore it follows, that any pseudo-Riemannian metric g uniquely admits a Levi-Civita connection ∇^g. For this reason the choice of a pseudo-Riemannian metric naturally induces a global geometry to a differentiable manifold and therefore w.r.t. statistical manifolds justifies the definition of *pseudo-Riemannian statistical manifolds.*

Definition ((Pseudo-)Riemannian statistical manifold)**.** *Let $(\mathcal{M}, \mathcal{A})$ be a differentiable statistical manifold and g a (Pseudo-)Riemannian metric on $\mathrm{KQ}(\mathcal{M}, \mathcal{A})$. Then the tuple $(\mathcal{M}, g)$ is termed a (Pseudo-)Riemannian statistical manifold.* **Remark**: *The category of k-differentiable (Pseudo-)Riemannia statistical manifolds is denoted by $\mathbf{StatMan}^k_R$.*

Generally (Pseudo-)Riemannian manifolds endow the notation of geodesics with an intuitive meaning as the trajectories of free particles. Thereby the equations of motion obey the principle of stationary action, whereat the *action functional* $\mathcal{S}(\gamma) \coloneqq \int_a^b \mathcal{L}\mathrm{d}t$ is defined over the Lagrangian $\mathcal{L} \coloneqq \frac{1}{2}g(\dot\gamma, \dot\gamma)$. The properties of the Levi-Civita connection then allow the transformation of the local geodesic equation 3.1 to *Euler-Lagrange equations*, over the local Lagrangian $\mathcal{L}_\theta \coloneqq \sum_{i,j}\frac{1}{2}g_{ij}\dot\gamma^i_\theta\dot\gamma^j_\theta$, such that:

$$\nabla^g_{\dot\gamma_\phi}\dot\gamma_\phi = 0 \iff \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathcal{L}_\phi}{\partial\dot\gamma^k_\phi}\right) - \frac{\partial\mathcal{L}_\phi}{\partial\gamma^k_\phi} = 0,\ \forall k \tag{3.2}$$

Consequentially the geodesics of (Pseudo-)Riemannian manifolds are stationary solutions of the action functional $\mathcal{S}(\gamma)$, i.e. $\delta\mathcal{S} = 0$. By regarding the tangent bundle $T\mathcal{M}$ as the *configuration space* of a moving particle and its elements $(q,\, \dot{q}) \in T\mathcal{M}$ as the *generalized coordinates*, the Lagrangian equals its *kinetic term*. With regard to equation 3.2 the geodesics of $(\mathcal{M},\, g)$ then coincide with the trajectories of free particles. This encourages the interpretation of $(\mathcal{M},\, g)$ as a dynamical system, where the temporal evolution is determined by the *geodesic flow* $\Phi^t : T\mathcal{M} \to T\mathcal{M}$ with $\Phi^t(q,\, \dot{q}) = (\gamma_{q,\dot{q}}(t),\, \dot{\gamma}_{q,\dot{q}}(t))$, where $\gamma_{q,\dot{q}}(t)$ is the locally unique geodesic, that satisfies the initial conditions $\gamma_{q,\dot{q}}(0) = q$ and $\dot{\gamma}_{q,\dot{q}}(0) = \dot{q}$. Then due to $\frac{\mathrm{d}}{\mathrm{d}t} g(\dot{q},\, \dot{q}) = g(\nabla^g_{\dot{q}}\dot{q},\, \dot{q}) = 0$ the geodesic flow preserves the kinetic term along its trajectories, and therefore generalizes Newton's first law of motion to curvilinear and pseudo-Euclidean spaces. In appreciation of its origins in the conceptualization of spacetime, a geodesic γ is therefore termed *spacelike* if $g(\dot{q},\, \dot{q}) > 0$, *lightlike* if $g(\dot{q},\, \dot{q}) = 0$ and *timelike* if $g(\dot{q},\, \dot{q}) < 0$. Moreover the (Pseudo-)Riemannian metric g induces a canonical isomorphism between the tangent spaces $T_q\mathcal{M}$ and their respective dual spaces $T^*_q\mathcal{M}$, the *cotangent spaces*, which assigns a *cotangent vector* $p \in T^*_q\mathcal{M}$ to each tangent vector $\dot{q} \in T_q\mathcal{M}$ by $p(v) \coloneqq g_q(\dot{q},\, v)$. Then also the choice of a local frame $\{\partial_i\}$ uniquely induces a *local coframe* $\{\mathrm{d}q^i\}$ by $\mathrm{d}q^i \coloneqq \partial_i^T G_q$, such that any $p \in T^*_q\mathcal{M}$ has a local representation $p = p_i \mathrm{d}q^i$. As the geodesic flow however preserves the kinetic term it holds, that $\frac{\mathrm{d}}{\mathrm{d}t} p(\dot{q}) = \frac{\mathrm{d}}{\mathrm{d}t} g(\dot{q},\, \dot{q}) = 0$, such that p equals the *conjugate momentum* of $\dot{q}$. Consequentially the disjoint union of all cotangent spaces, given by the *cotangent bundle* $T^*\mathcal{M}$ equals the *phase space* of the dynamical system. Finally by the definition of the *Hamiltonian* $\mathcal{H}(q,\, p) \coloneqq \frac{1}{2} g^{ij} p_i p_j$ with $(g^{ij}) = G_q^{-1}$ it follows, that $(\mathcal{M},\, g)$ uniquely corresponds to a *Hamiltonian system*, since (i) $\dot{q}^i = g^{ij} p_j = \frac{\partial \mathcal{H}}{\partial p_i}$ and (ii) $\dot{p}^i = -\frac{\partial}{\partial \gamma_i} \frac{1}{2} g^{ij} p_i p_j = -\frac{\partial \mathcal{H}}{\partial q_i}$. This representation allows to reformulate the principle of stationary action in *canonical coordinates* $(q,\, p) \in T^*\mathcal{M}$ by the *curve integral* $\mathcal{S}(q) = \int_a^b \mathcal{L}\mathrm{d}t = \frac{1}{2} \int_q p$. In particular this formulation, known as *Maupertuis' principle* then describes the trajectory of a free particle by its geometric shape instead of its temporal evolution. Due to this geometric interpretation the action functional of a curve may also be defined with regard to a given vector field of conjugate momenta $p \in \Gamma(T^*\mathcal{M})$, by $\mathcal{S}_p(q) \coloneqq \frac{1}{2} \int_q p$. Then for arbitrary smooth curves $q\colon [a,\, b] \to \mathcal{M}$ the action $\mathcal{S}_p(q)$ is completely determined by its boundary values localized at q_a and q_b, such that:

$$\mathcal{S}_p(q) = \frac{1}{2} \int_q p = \mathcal{S}_p(q_b) - \mathcal{S}_p(q_a)$$

Although w.r.t. the fundamental theorem of calculus this insight seems rather trite, it provides a far-reaching generalisation. Thereby in the very same manner as the smooth sections of $T^*\mathcal{M}$ constitute the smooth linear forms over $T\mathcal{M}$, the smooth alternating multilinear forms over $T\mathcal{M}^k$, termed *differential k-forms* are given by smooth sections of the *outer product* $\Lambda^k(T^*\mathcal{M})$. These k-forms then provide the natural integrands over curves, surfaces, volumes or higher-dimensional k-manifolds and therefore may be thought as measures of the flux through infinitesimal k-parallelotopes. In this sense the smooth functions over $\mathcal{M}$ are 0-*forms* and the smooth vector fields over $\mathcal{M}$ are 1-*forms*. In particular however since for any smooth function $f \in C^\infty(\mathcal{M})$ the differential $\mathrm{d}f$ is a smooth vector field, it appears that d by itself defines an $\mathbb{R}$-linear mapping $\mathrm{d}\colon \Omega^0(\mathcal{M}) \to \Omega^1(\mathcal{M})$, where $\Omega^k(\mathcal{M}) \coloneqq \Gamma(\Lambda^k(T^*\mathcal{M}))$. This encourages to extend the notation of a differential to arbitrary k-forms by the *exterior derivative*, given by an $\mathbb{R}$-linear mapping $\mathrm{d}_k\colon \Omega^k(\mathcal{M}) \to \Omega^{k+1}(\mathcal{M})$, that satisfies (i) d_k is an antiderivation for any $k \in \mathbb{N}_0$, (ii) $\mathrm{d}_{k+1} \circ \mathrm{d}_k = 0$ for any $k \in \mathbb{N}_0$ and (iii) d_0 is the differential. In

more detail (i) claims, that for any $\alpha \in \Omega^k(\mathcal{M})$ and $\beta \in \Omega^l(\mathcal{M})$ it follows, that $\mathrm{d}_{k+l}(\alpha \wedge \beta) = \mathrm{d}_k\alpha \wedge \beta + (-1)^l(\alpha \wedge \mathrm{d}_l\beta)$. This provides the property, that infinitesimal changes of the volume $\alpha \wedge \beta$ are expressible as the sum of infinitesimal changes in their orthocomplemented constituent volumes. Then the additional claim (ii) assures the symmetry of second derivatives and (iii) the compatibility with the differential. In order to provide a measure of length however the (Pseudo-)Riemannian metric g is additionally required to be positive definite, i.e. such that $\forall P \in \mathcal{M}$ the g_p are positive definite. Then g is termed a *Riemannian metric* and a statistical manifold $(\mathcal{M}, g)$ a *Riemannian statistical manifold*. In this case the Riemannian metric defines an inner product $\langle \cdot, \cdot \rangle_g \colon T_P\mathcal{M}^2 \to \mathbb{R}$ by $(v, w) \mapsto g_P(v, w)$ and therefore induces a norm $\|\cdot\|_g \colon T_P\mathcal{M} \to \mathbb{R}$ by $\|v\|_g \coloneqq \sqrt{\langle v, v \rangle_g}$. This allows the definition of the *length* functional of a piecewise smooth curve.

Definition (Arc length)**.** *Let $(\mathcal{M}, g)$ be a Riemannian statistical manifold and $\gamma \colon [a, b] \to \mathcal{M}$ a piecewise smooth curve in* $\mathrm{KQ}(\mathcal{M}, g)$*. Then the arc length of γ is given by:*

$$L(\gamma) \coloneqq \int_a^b \|\dot{\gamma}(t)\|_g \mathrm{d}t \tag{3.3}$$

Analogues to the action functional, the length functional may be written by a Lagrangian, which is given by $\mathcal{L}_L(\gamma, \dot{\gamma}, t) \coloneqq \sqrt{g(\dot{\gamma}, \dot{\gamma})}$, such that $\mathcal{L}_L = \sqrt{2\mathcal{L}}$. Then the Euler-Lagrange equations for the length functional are equivalent to the Euler-Lagrange equations for action functional, such that the stationary solutions of the length and action functional coincide. This property allows to equip Riemannian statistical manifolds with a *distance*.

Definition (Distance)**.** *Let $(\mathcal{M}, g)$ be a Riemannian statistical manifold, then the distance $d \colon \mathcal{M}^2 \to \mathbb{R}$ of $P, Q \in \mathcal{M}$ is defined by:*

$$d(P, Q) \coloneqq \begin{cases} \infty & \text{, if } P \text{ and } Q \text{ are not} \\ & \text{path connected in } \mathcal{M} \\ \inf L(\gamma) & \text{, where } \gamma \colon [a, b] \to \mathcal{M} \\ & \text{with } \gamma(a) = P, \gamma(b) = Q \end{cases} \tag{3.4}$$

Due to its definition the distance d of a Riemannian statistical manifold for arbitrary $P, Q, R \in \mathcal{M}$ satisfies, that: (i) $d(P, P) = 0$, (ii) $d(P, Q) = d(Q, R)$ and (iii) $d(P, Q) + d(Q, R) \leq d(P, R)$. In order to show, that $(\mathcal{M}, d)$ is a metric space it therefore suffices to prove, that $d(P, Q) > 0$ for $P \neq Q$. Let $(\mathcal{M}, g)$ and $(\mathcal{N}, g')$ be Pseudo-Riemannian manifolds, and $f \in C^\infty(\mathcal{M}, \mathcal{N})$, then f is an isometry, iff $g_P(v, w) = g'_{f(P)}(\mathrm{d}f_P(v), \mathrm{d}f_w(P))$ for all $P \in \mathcal{M}$ and $v, w \in T_P\mathcal{M}$.

Let $(\mathcal{P}, \mathcal{A})$ be a differentiable statistical manifold. *Then a mapping $D(\cdot \parallel \cdot) \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$ is termed a divergence over $\mathcal{P}$, iff for all $P, Q \in \mathcal{P}$ it holds, that $D(P \parallel Q) \geq 0$, with $D(P \parallel Q) = 0 \Leftrightarrow P = Q$.* If

these scalar products are usually induced by derivatives of locally linear divergences.

Definition (Statistical divergence)**.** *Let*

$$(X, \mathcal{P}) \in \mathrm{ob}(\mathbf{Stat})$$

Then a mapping $D(\cdot \parallel \cdot) \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$ is termed a (statistical) divergence over $\mathcal{P}$, iff for all $P, Q \in \mathcal{P}$ it holds, that

$$D(P \parallel Q) \geq 0, \forall P, Q \in \mathcal{P} \tag{3.5}$$

$$D(P \parallel Q) = 0 \Leftrightarrow P = Q, \forall P, Q \in \mathcal{P} \tag{3.6}$$

Definition (Locally linear divergence)**.** *Let*

$$(X, \mathcal{P}_\xi) \in \mathrm{ob}(\mathbf{StatMan}^k)$$

and let D be a statistical divergence over $(X, \mathcal{P})$. Then D is termed locally linear, iff for all $P \in \mathcal{P}$ the linearisation of D at P is given by a positive

definite matrix $G_\xi(P)$, such that:

$$D[P_\xi \parallel P_\xi + \mathrm{d}P] = \frac{1}{2}\mathrm{d}\boldsymbol{\xi}^T G_\xi(P_\xi)\mathrm{d}\boldsymbol{\xi} + O(n^3) \tag{3.7}$$

Example (Kullback-Leibler divergence)**.** *Let*

$$(X, \mathcal{P}) \in \mathrm{ob}(\mathbf{Stat})$$

Then for $P, Q \in \mathcal{P}$ the Kullback-Leibler divergence $D_{KL}[\cdot \parallel \cdot] : \mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$ is defined by:

$$D_{\mathrm{KL}}[P \parallel Q] \coloneqq \int_X \mathrm{d}_\mu P(x) \log \frac{\mathrm{d}_\mu P(x)}{\mathrm{d}_\mu Q(x)} \mathrm{d}\mu(x) \tag{3.8}$$

Remark: *The Kullback-Leibler divergence measures the amount of information, which is gained when one revises ones beliefs from the prior probability distribution P to the posterior probability distribution Q.*

Riemannian statistical manifolds $(X, \mathcal{P}, D)$ are identified with Riemannian manifolds (M, g), where M is given by $(X, \mathcal{P})$ and the Riemannian metric g by the linearisation of D. This identification is well-defined, since $(X, \mathcal{P})$ is by definition a smooth manifold and the linearisation of a locally linear divergence for any $P \in \mathcal{P}$ yields a positive definite matrix $G(P)$. Let ξ be a differentiable parametrisation of $(X, \mathcal{P})$, then the line element $\mathrm{d}s^2$ has a local representation:

$$\mathrm{d}s_P^2 = 2D[P \parallel P + \mathrm{d}P] = \mathrm{d}\boldsymbol{\xi}^T G_\xi(P_\xi)\mathrm{d}\boldsymbol{\xi} \tag{3.9}$$

This allows the determination of distances in $(X, \mathcal{P})$ by the length of continuously differentiable curves.

Lemma 1. *Let $(X, \mathcal{P}, D)$ be a Riemannian statistical manifold and $P, Q \in \mathcal{P}$. Then the length of continuously differentiable curves $\gamma_{P,Q} : [a, b] \to (X, \mathcal{P})$ with $\gamma(a) = P$ and $\gamma(b) = Q$ has a unique infimum $d_{P,Q}$, such that $d_{P,Q} \geq 0$ and $d_{P,Q} = 0 \Leftrightarrow P = Q$.*

Proof. Since D is locally linear, it follows that for any curve γ from P to Q it holds, that:

$$L(\gamma_{P,Q}) \geq D[\gamma(a) \parallel \gamma(b)] \geq 0 \tag{3.10}$$

Therefore the length of all continuously differentiable curves from P to Q has a unique infimum $d_{P,Q} \geq 0$. Let $P = Q$, then the continuously differentiable curves connecting P and Q may be contracted at P such that $d_{P,Q} = \inf L(\gamma_{P,Q}) = \liminf_{Q \to P} D[P \parallel Q] = D[P \parallel P] = 0$. Conversely let $P \neq Q$, then $L(\gamma_{P,Q}) \geq D[P \parallel Q] > 0$. □

Then for $P, Q \in \mathcal{P}$ a geodesic from P to Q is given by a continuous differentiable curve $\gamma_{P,Q} : [a, b] \to (X, \mathcal{P})$ with $\gamma_{P,Q}(a) = P$ and $\gamma_{P,Q}(b) = Q$, such that $\gamma_{P,Q}$ minimizes the length among all continuously differentiable curves from P to Q. In the purpose to preserve the distance of the Riemannian structure, the parametrisation of $(X, \mathcal{P}, D)$ requires to preserve the property of a curve to be a geodesic within the parameter space. These preserved geodesics are termed affine geodesics. If a differentiable parametrisation globally preserves the distances, then it is given by an isometric embedding of $(X, \mathcal{P}, D)$ and termed an affine parametrisation.

Definition (Affine parametrisation)**.** *Let $(X, \mathcal{P}, D)$ be a Riemannian statistical manifold. Then a parametrisation ξ is termed an affine parametrisation for $(X, \mathcal{P}, D)$, iff the geodesics in $(X, \mathcal{P}, D)$ are ξ-affine geodesics.* **Remark**: *A Riemannian statistical manifolds, given by the notation $(X, \mathcal{P}_\xi, D)$ implicates an affine parametrisation ξ.*

The embedding of a smooth statistical manifold $(X, \mathcal{Q})$ within a Riemannian statistical manifold $(X, \mathcal{P}, D)$ naturally induces the Riemannian metric to the submanifold $(X, \mathcal{Q})$, such that also $(X, \mathcal{Q}, D)$ is a Riemannian statistical manifold. This is of particularly importance for the approximation of high dimensional statistical models by lower dimensional submanifolds. For this purpose the Riemannian metric is fundamental to obtain a projection from

probability distributions in $\mathcal{P}$ to their closest approximation in $\mathcal{Q}$. This projection is a geodesic projection:

Definition (Geodesic projection)**.** *Let* $(X, \mathcal{P}, D)$ *be a Riemannian statistical manifold and* $(X, \mathcal{Q})$ *a smooth submanifold. Then a mapping* $\pi : \mathcal{P} \longrightarrow \mathcal{Q}$ *is termed a geodesic projection, iff any point* $P \in \mathcal{P}$ *is mapped to a point* $\pi(P) \in \mathcal{Q}$*, that minimizes the distance* $d(P, \pi(P))$*.* **Remark:** *By it's definition* $d(P, \pi(P)) < \infty$ *iff* P *and* $\pi(P)$ *are path-connected. Therefore geodesic projections are by convention restricted to the common topological components of* $\mathcal{P}$ *and* $\mathcal{Q}$*.*

4. Dually flat Structure

In Riemannian manifolds the problem to determine geodesic projections to submanifolds is generally hard to solve, since the distance has to be minimized over all continuously differentiable curves that connect points to the submanifold. A particular convenient geometry however arises by a flat Riemannian metric, whereas the flatness of the metric is related to the direction of curves. The claim for a further flat structure, which is given by the dual Riemannian metric allows a generalization of the Pythagorean theorem and therefore an explicit calculation rule for geodesic projections by dual affine linear projections.

Definition (Dual Riemannian metric)**.** *Let* (M, g) *be a Riemannian manifold, then the Riemannian metric tensor* g *is given by a family of positive definite matrices* $\{g_P\}_{P \in M}$*. Then metric* g^**, with is dual to* g *is given by the family of the inverse Riemannian metric tensors, such that* $g^*_P = g_P^{-1}$, $\forall P \in M$*.*

The dual Riemannian metric g^* may be regarded as the Riemannian metric with a locally inverse direction. In Riemannian statistical manifolds this definition corresponds to the dual divergence.

Definition (Dual divergence)**.** *Let*

$$(X, \mathcal{P}) \in \mathrm{ob}(\mathbf{Stat})$$

and D *be a locally linear divergence over* $(X, \mathcal{P})$*. Then the dual divergence* D^* *w.r.t.* D *is given by:*

$$D^*[P \parallel Q] = D[Q \parallel P], \forall P, Q \in \mathcal{P} \tag{4.1}$$

In its most simple case the Riemannian metric g, induced by the divergence D equals the dual Riemannian metric g^*, induced by the dual divergence D^*. In this case the Riemannian metric and the divergence are termed self-dual.

Definition (Self-dual Riemannian metric)**.** *Let* (M, g) *be a Riemannian manifold, then the Riemannian metric tensor* g *is termed self-dual, iff* $g^* = g$*.*

4.1. Dual parametrisation and Legendre transformation.

By considering a differentiable statistical manifold $(X, \mathcal{P}_\xi)$ and a real valued differentiable convex function $\psi : \mathrm{img}\xi \to \mathbb{R}$, the differentiability of ψ may be used to introduce a further differentiable parametrisation of $(X, \mathcal{P})$ by $\boldsymbol{\xi}^*_P := \nabla_\xi \psi(\boldsymbol{\xi}_P)$. Furthermore since ψ is convex, the Jacobian determinant is positive for any $\boldsymbol{\xi}_P \in \mathrm{dom}\xi$ and therefore the transformation $\xi \to \xi^*$ is globally invertible. This defines a bijective relationship between parameter vectors $\boldsymbol{\xi}_P \in \mathrm{dom}\xi$ and their respectively normal vectors in the tangent space, given by $\boldsymbol{\xi}^*_P \in \mathrm{dom}\xi^*$. Since ξ is an identifiable parametrisation and the transformation $\xi \to \xi^*$ is globally invertible, also ξ^* is an identifiable parametrisation. This justifies the following definition:

Definition (Dual parametrisation)**.** *Let* $(X, \mathcal{P}_\xi)$ *be a differentiable statistical manifold and* $\psi : \mathrm{dom}\xi \to \mathbb{R}$ *a sufficiently differentiable convex function. Then the dual parametrisation for* $(X, \mathcal{P})$ *w.r.t.* ψ *is given by:*

$$\boldsymbol{\xi}^*_P = \nabla_\xi \psi(\boldsymbol{\xi}_P), \forall P \in \mathcal{P} \tag{4.2}$$

Due to the convexity of ψ, also the inverse transformation $\xi^* \to \xi$ may be represented by the partial

derivation of dual function $\psi^* : \mathrm{dom}\xi^* \to \mathbb{R}$. This yields a transformation $(\xi, \psi) \to (\xi^*, \psi^*)$, which is defined by a dualistic relationship between (ξ, ψ) and (ξ^*, ψ^*), such that additional to equation 4.2 also:

$$\boldsymbol{\xi}_P = \nabla_{\xi^*}\psi^*(\boldsymbol{\xi}^*_P), \forall P \in \mathcal{P} \tag{4.3}$$

This transformation is known as the *Legendre transformation* and the function ψ^* as the *Legendre dual function* of ψ.

Lemma 2. *Let $(X, \mathcal{P}_\xi)$ be a differentiable statistical manifold, $\psi : \mathrm{dom}\xi \to \mathbb{R}$ a differentiable convex function and $(\xi, \psi) \to (\xi^*, \psi^*)$ a Legendre transformation of (ξ, ψ), then the Legendre dual function $\psi^* : \mathrm{dom}\xi^* \to \mathbb{R}$ is given by:*

$$\psi^*(\boldsymbol{\xi}^*_P) = \arg\max_P \left(\boldsymbol{\xi}_P \cdot \boldsymbol{\xi}^*_P - \psi(\boldsymbol{\xi}_P)\right), \forall P \in \mathcal{P} \tag{4.4}$$

Proof. By applying the definition of the dual parametrisation ξ^* it has only to be proofed, that the function ψ^*, given by 4.4 satisfies the conditions of the Legendre dual function, given by equation 4.3. Let $P \in \mathcal{P}$, then:

$$\begin{aligned}
&\nabla_{\xi^*}\psi^*(\boldsymbol{\xi}^*_P) \\
&\overset{4.4}{=} \boldsymbol{\xi}_P + (\partial_{\xi^*}\boldsymbol{\xi}_P) \cdot \boldsymbol{\xi}^*_P - \nabla_\xi\psi(\boldsymbol{\xi}_P) \cdot (\partial_{\xi^*}\boldsymbol{\xi}_P) \\
&= \boldsymbol{\xi}_P + (\partial_{\xi^*}\boldsymbol{\xi}_P) \cdot \nabla_\xi\psi(\boldsymbol{\xi}_P) - \nabla_\xi\psi(\boldsymbol{\xi}_P) \cdot (\partial_{\xi^*}\boldsymbol{\xi}_P) \\
&= \boldsymbol{\xi}_P + (\nabla_\xi\psi(\boldsymbol{\xi}_P) - \nabla_\xi\psi(\boldsymbol{\xi}_P)) \cdot (\partial_{\xi^*}\boldsymbol{\xi}_P) = \boldsymbol{\xi}_P
\end{aligned}$$

□

4.2. **Bregman divergence.** The dualistic relationship between dual parametrisations, given by the Legendre transformation shall be extended to Riemannian metrices. To this end a family of locally linear divergences is introduced, that generates a dualistic relationship structure:

Definition (Bregman divergence)**.** *Let $(X, \mathcal{P}_\xi)$ be a differentiable statistical manifold and $\psi : \mathrm{dom}\xi \to \mathbb{R}$ a differentiable convex function. Then for $P, Q \in \mathcal{P}$ the **Bregman divergence** D_ψ w.r.t. the differentiable parametrisation ξ is given by:*

$$D_\psi[P \parallel Q] = \psi(\boldsymbol{\xi}_P) - \psi(\boldsymbol{\xi}_Q) - \nabla_\xi\psi(\boldsymbol{\xi}_P) \cdot (\boldsymbol{\xi}_Q - \boldsymbol{\xi}_P) \tag{4.5}$$

Lemma 3. *Let D_ψ be a Bregman divergence with regard to a sufficiently differentiable parametrisation ξ, then D_ψ is locally linear and the Riemannian metric, induced by D_ψ, is given by:*

$$g_P = \nabla^2_\xi\psi(\boldsymbol{\xi}_P) \tag{4.6}$$

Proof. By applying the definition of a differentiable convex function it follows, that D_ψ is locally linear and the linearisation term of the Taylor expansion yields $G_\xi(P_\xi) = \nabla^2_\xi\psi(\boldsymbol{\xi}_P)$. Since by the definition of a Riemannian statistical manifold $g_P = G_\xi(P_\xi)$, it follows that $\mathrm{g_P} = \nabla^2_\xi\psi(\boldsymbol{\xi}_P)$ □

Lemma 4. *Let D_ψ be a Bregman divergence, then the dual divergence D^*_ψ is given by the Bregman divergence of the Legendre dual function ψ^*, such that:*

$$D^*_\psi[P \parallel Q] = D_{\psi^*}[P \parallel Q] \tag{4.7}$$

Proof. Let $G_\xi(P_\xi)$ be the linearisation of D_ψ at $P \in \mathcal{P}$, then:

$$\begin{aligned}
&G_\xi(P_\xi) \\
&= \nabla^2_\xi\psi(\boldsymbol{\xi}_P) = \nabla_\xi\boldsymbol{\xi}^*_P \\
&= (\nabla_{\xi^*})^{-1}\boldsymbol{\xi}_P = (\nabla^2_{\xi^*}\psi^*)^{-1}(\boldsymbol{\xi}^*_P) \\
&= G^{-1}_{\xi^*}(P_{\xi^*})
\end{aligned}$$

And therefore:

$$G_\xi(P_\xi) = G^{-1}_{\xi^*}(P_{\xi^*}), \forall P \in \mathcal{P} \tag{4.8}$$

Since D_ψ is locally linear $G_\xi(P_\xi)$ is positive definite $\forall P \in \mathcal{P}$. From equation 4.8 it therefore follows, that also $G^{-1}_{\xi^*}(P_{\xi^*})$ is positive definite $\forall P \in \mathcal{P}$ and since the inverse matrix of a positive definite matrix is also positive definite it follows, that $G_{\xi^*}(P_{\xi^*})$ is a positive definite $\forall P \in \mathcal{P}$. Furthermore by the definition of the Legendre transformation $G_{\xi^*}(P_{\xi^*})$ is the Hessian matrix of $\psi^*(\boldsymbol{\xi}^*_P)$ and therefore ψ^*

is a convex function of $\boldsymbol{\xi}_P^* \in \mathrm{dom}\xi^*$. Therefore ψ^*satisfies the requirement for the definition of a Bregman divergence. Let P, $Q \in \mathcal{P}$, then:

$$\begin{aligned}
&D_{\psi^*}[P \parallel Q] \\
&= \psi^*(\boldsymbol{\xi}_P^*) - \psi^*(\boldsymbol{\xi}_Q^*) - \nabla_{\xi^*}\psi^*(\boldsymbol{\xi}_Q^*)(\boldsymbol{\xi}_Q^* - \boldsymbol{\xi}_P^*) \\
&= \psi(\boldsymbol{\xi}_Q) - \psi(\boldsymbol{\xi}_P) - \nabla_{\xi}\psi(\boldsymbol{\xi}_P)(\boldsymbol{\xi}_P - \boldsymbol{\xi}_Q) \\
&\overset{\text{def}}{=} D_{\psi}[Q \parallel P] \\
&\overset{\text{def}}{=} D_{\psi}^*[P \parallel Q]
\end{aligned}$$

□

Proposition 5. *Let $(X, \mathcal{P}_\xi, D_\psi)$ be a Riemannian statistical manifold with a Bregman divergence D_ψ. Then the dual Riemannian metric g^* is induced by the Bregman divergence D_{ψ^*} of the Legendre dual function ψ^*.*

Proof. By applying the definition for the dual Riemannian metric for $P \in \mathcal{P}$ it follows, that::

$$g_P^* \overset{\text{def}}{=} g_P^{-1} \overset{\text{def}}{=} G_\xi^{-1}(P_\xi) \overset{4.8}{=} G_\xi^*(P_\xi) \overset{4.6}{=} \nabla_{\xi^*}^2\psi^*(\boldsymbol{\xi}_P^*)$$

This is the linearisation of the Bregman divergence D_{ψ^*}. □

4.3. **Dually flat statistical manifolds.**

Definition (Dually flat manifold)**.** *Let (M, g) be a Riemannian manifold. Then (M, g) is termed a dually flat (Riemannian) manifold, iff:*

(1) *g is a flat Riemannian metric of M*
(2) *g^* is a flat Riemannian metric of M*

Example (Euclidean space)**.** An example for a dually flat manifold is given by Euclidean spaces. Let ξ be the Cartesian coordinates of an Euclidean space E, then ξ is an affine parametrisation of E and for P, $Q \in E$ the Euclidean metric of E is induced by the Euclidean divergence $D[P \parallel Q] = \frac{1}{2}|\boldsymbol{\xi}_P - \boldsymbol{\xi}_Q|^2$. In this case the divergence is symmetric and therefore the Euclidean metric is self-dual Since E is flat with regard to the Euclidean metric E is also flat with regard to the dual Euclidean metric and therefore E is a dually flat manifold.

Lemma 6. *Let $(X, \mathcal{P}_\xi, D_\psi)$ be a Riemannian statistical manifold with a Bregman divergence D_ψ. Then $(X, \mathcal{P}_\xi, D_\psi)$ is a dually flat statistical manifold, iff:*

(1) *The ξ-affine geodesics are flat with regard to the Riemannian metric, induced by D_ψ*
(2) *The ξ^*-affine geodesics are flat with regard to the Riemannian metric, induced by D_{ψ^*}*

Proof. Since the Legendre transformation generally does not preserve the Riemannian metric, the flatness of $(X, \mathcal{P}, D_\psi)$ and $(X, \mathcal{P}, D_{\psi^*})$ are indeed independent properties. Let the ξ-affine geodesics be flat with regard to the Riemannian metric g, induced by D_ψ, then also $(X, \mathcal{P})$ is flat w.r.t. g. Let further the ξ^*-affine geodesics be flat with regard to the Riemannian metric $\tilde{g}$, induced by D_{ψ^*}, then by 5 it follows, that $\tilde{g} = g^*$ and therefore g^* is a flat Riemannian metric of $(X, \mathcal{P})$. Conversely let $(X, \mathcal{P}_\xi, D_\psi)$ be a dually flat statistical manifold with a Bregman divergence D_ψ. Then by convention ξ is an affine parametrisation of $(X, \mathcal{P}, D_\psi)$ and the geodesics in $(X, \mathcal{P}, D_\psi)$ are ξ-affine geodesics and flat with regard to the Riemannian metric, induced by D_ψ. Furthermore the dual Riemannian metric g^* induced by D_ψ^* is a flat Riemannian metric of $(X, \mathcal{P})$ and since D_ψ is a Bregman divergence it follows that $D_\psi^* = D_{\psi^*}$. Then the dual parametrisation $\xi^*$is an affine parametrisation of $(X, \mathcal{P}, D_\psi^*)$ and the geodesics in $(X, \mathcal{P}, D_\psi)$ are ξ^*-affine geodesics and flat with regard to the Riemannian metric, induced by D_{ψ^*} . □

Definition (Dual geodesic projection)**.** *Let $(X, \mathcal{P}, D)$ be a Riemannian statistical manifold with a smooth submanifold $(X, \mathcal{Q})$. Then a mapping $\pi^* : \mathcal{P} \longrightarrow \mathcal{Q}$ is termed a **dual geodesic projection**, iff any point $P \in \mathcal{P}$ is mapped to a point $\pi^*(P) \in \mathcal{Q}$, that minimizes the distance $d(P, \pi^*(P))$ w.r.t. the dual Riemannian metric, which is induced by D^*.*

In the case of a dually flat statistical manifold, the dual affine structure induces a correspondence

relationship between the Riemannian metrices, induced by D and D^*.

Lemma 7. *Let $(X, \mathcal{P}_\xi, D_\psi)$ be a Riemannian statistical manifold with a Bregman divergence D_ψ. Then D_ψ has a mixed representation in the parametrisations ξ and ξ^*, which is given by:*

$$D_\psi[P \parallel Q] = \psi(\boldsymbol{\xi}_P) + \psi^*(\boldsymbol{\xi}^*_Q) - \boldsymbol{\xi}_P \cdot \boldsymbol{\xi}^*_Q \tag{4.9}$$

Proof. By applying the definition of the dual divergence and Lemma 4 it follows, that:

$$D_\psi[P \parallel Q] \overset{4.7}{=} D_{\psi^*}[Q \parallel P]$$

The right side of the equation is calculated by the definition of the Bregman divergence and the Legendre dual function, such that:

$$\begin{aligned} & D_{\psi^*}[Q \parallel P] \\ & \overset{4.5}{=} \psi^*(\boldsymbol{\xi}^*_Q) - \psi^*(\boldsymbol{\xi}^*_P) - \nabla\psi^*(\boldsymbol{\xi}^*_P)(\boldsymbol{\xi}^*_P - \boldsymbol{\xi}^*_Q) \\ & = \psi(\boldsymbol{\xi}_P) + \psi^*(\boldsymbol{\xi}^*_Q) - \boldsymbol{\xi}_P \cdot \boldsymbol{\xi}^*_Q \end{aligned}$$

□

Theorem 8 (*Amari Pythagorean Theorem*)**.** *Let $(X, \mathcal{P}_\xi, D_\psi)$ be a dually flat statistical manifold, which is given by a Bregman divergence D_ψ and let $P, Q, R \in \mathcal{P}$ be an orthogonal triangle in the sense, that the ξ^*-affine geodesic $\gamma^*_{P,Q}$ from P to Q is orthogonal to the ξ-affine geodesic $\gamma_{Q,R}$ from Q to R, then:*

$$D_\psi[P \parallel R] = D_\psi[P \parallel Q] + D_\psi[Q \parallel R] \tag{4.10}$$

Proof. The ξ^*-affine geodesic $\gamma^*_{P,Q} : [0, 1] \to (X, \mathcal{P})$, with $\gamma^*_{P,Q}(0) = P$ and $\gamma^*_{P,Q}(1) = Q$ is parametrized by:

$$\boldsymbol{\xi}^*_{P,Q}(t) = t\boldsymbol{\xi}^*_Q + (1-t)\boldsymbol{\xi}^*_P,\ t \in [0, 1]$$

and the ξ-affine geodesic $\gamma_{Q,R} : [0, 1] \to (X, \mathcal{P})$, with $\gamma_{Q,R}(0) = Q$ and $\gamma_{Q,R}(1) = R$ by:

$$\boldsymbol{\xi}_{Q,R}(t) = t\boldsymbol{\xi}_R + (1-t)\boldsymbol{\xi}_Q,\ t \in [0, 1]$$

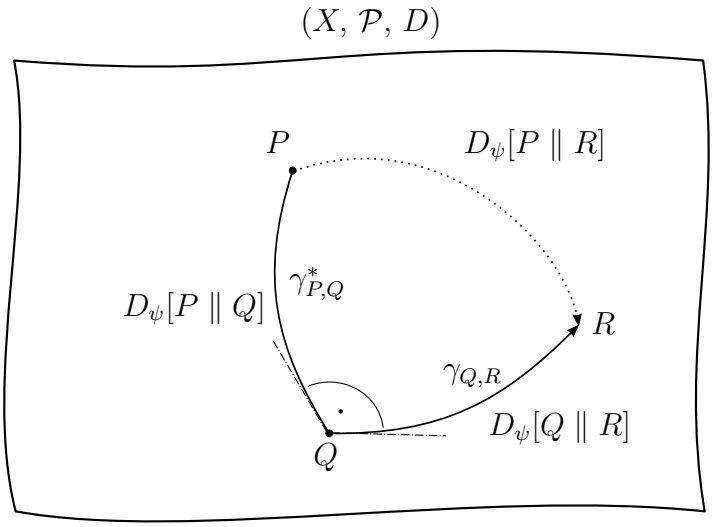

FIGURE 4.1. Pythagorean theorem for dually flat manifolds

Let $\langle \cdot, \cdot \rangle_g$ denote the local scalar product, which is induced by the Bregman divergence D_ψ. By applying the definition of the Bregman divergence, the local scalar product at the point Q is given by:

$$\begin{aligned} & \langle \frac{\mathrm{d}}{\mathrm{d}t}\gamma^*_{P,Q}(t)|_{t=1},\ \frac{\mathrm{d}}{\mathrm{d}t}\gamma_{Q,R}(t)|_{t=0} \rangle_g \\ & \overset{4.5}{=} (\boldsymbol{\xi}^*_Q - \boldsymbol{\xi}^*_P) \cdot (\boldsymbol{\xi}_R - \boldsymbol{\xi}_Q) \\ & \overset{4.9}{=} \boldsymbol{\xi}^*_Q \cdot \boldsymbol{\xi}_R - \boldsymbol{\xi}^*_P \cdot \boldsymbol{\xi}_R + \boldsymbol{\xi}^*_P \cdot \boldsymbol{\xi}_Q - \psi(\boldsymbol{\xi}_Q) - \psi^*(\boldsymbol{\xi}^*_Q) \\ & \overset{4.9}{=} D_\psi[P \parallel Q] + D_\psi[Q \parallel R] - D_\psi[P \parallel R] \end{aligned}$$

Since $\gamma^*_{P,Q}$ and $\gamma_{Q,R}$ are required to be orthogonal in the point Q, the left side of the equation equals zero and therefore it follows, that:

$$D_\psi[P \parallel Q] + D_\psi[Q \parallel R] - D_\psi[P \parallel R] = 0$$

□

Due to the generic asymmetry of Bregman divergences the generalized Pythagorean theorem has a corresponding dual theorem, which mutatis mutandis is given by:

$$D^*_\psi[P \parallel R] = D^*_\psi[P \parallel Q] + D^*_\psi[Q \parallel R] \tag{4.11}$$

If ψ is chosen, such that D_ψ is symmetric, the induced Riemannian metric of D_{ψ^*} is identical to that of D_ψ, since:

$$D_\psi[P \parallel Q] = D_\psi[Q \parallel P] = D_{\psi^*}[P \parallel Q]$$

In this case the generalized Pythagorean theorem and its dual corresponding are equivalent and the induced Riemannian metric is self-dual.

Definition (Affine projection)**.** *Let* $(X, \mathcal{P}_\xi, D)$ *be a Riemannian statistical manifold and* $(X, \mathcal{Q})$ *a smooth submanifold. Then a projection* $\pi_\xi^\perp : \mathcal{P} \to \mathcal{Q}$ *is termed an* ξ***-affine projection*** *from* $(X, \mathcal{P}, D)$ *to* $(X, \mathcal{Q}, D)$*, iff for any* $P \in \mathcal{P}$ *the* ξ*-affine geodesics from* P *to* $\pi_\xi^\perp(P)$ *are orthogonal to* $\mathcal{Q}$*.*

Lemma 9. *Let* $(X, \mathcal{P}_\xi, D_\psi)$ *be a dually flat statistical manifold with a smooth submanifold* $(X, \mathcal{Q})$*. Then there exists an* ξ*-affine projection as well as an* ξ^**-affine projection from* $(X, \mathcal{P}, D)$ to $(X, \mathcal{Q})$*.*

Proof. Since $(X, \mathcal{P}_\xi, D_\psi)$ is Riemannian statistical manifold by convention ξ is an affine parametrisation and therefore by definition any $P, Q \in \mathcal{P}$ are connected by a ξ-affine geodesic $\gamma_{P,Q} : [0, 1] \to (X, \mathcal{P})$ with $\gamma_{P,Q}(0) = P$ and $\gamma_{P,Q}(1) = Q$. Let's assume, that for a given $P \in \mathcal{P}$ there is no $Q \in \mathcal{Q}$, such that $\gamma_{P,Q} \perp \mathcal{Q}$, then due to the *mean value theorem* $\mathcal{Q}$ is not differentiable with regard to the affine parametrisation ξ and since ξ is a homeomorphism $\mathcal{Q}$ is also not differentiable in $\mathcal{P}$. However since $\mathcal{Q}$ is a smooth submanifold this does not hold, such that there exists a $Q \in \mathcal{Q}$ with $\gamma_{P,Q} \perp \mathcal{Q}$. The argument is true for any $P \in \mathcal{P}$ and therefore proves the existence of an ξ-affine projection. Since $(X, \mathcal{P}_\xi, D_\psi)$ is dually flat also ξ^* is an affine parametrisation. Then the argument, given for the ξ-affine projection mutatis mutandis proves the existence of an ξ^*-affine projection is argument may analogous be applied to the dual space, and the also proves the existence of a dual affine projection. □

Theorem 10 (*Amari Projection theorem*)**.** *Let* $(X, \mathcal{P}_\xi, D_\psi)$ *be a dually flat statistical manifold and* $(X, \mathcal{Q})$ *a smooth submanifold. Then the geodesic projection* $\pi : \mathcal{P} \to \mathcal{Q}$ *is an* ξ^**-affine projection and the dual geodesic projection* $\pi^* : \mathcal{P} \to \mathcal{Q}$ *is an* ξ*-affine projection.*

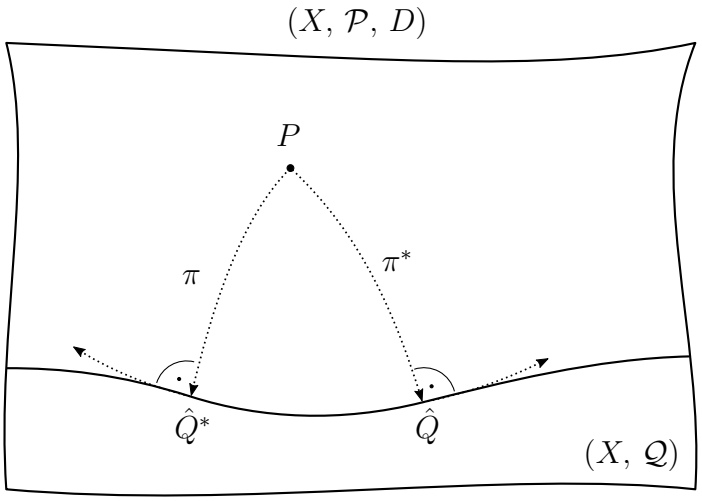

FIGURE 4.2. Projection theorem for dually flat manifolds

Proof. Let $P \in \mathcal{P}$, then due to Lemma 9 a dual affine projection $\pi_{\xi^*}^\perp$ may be chosen, such that the dual affine curve from P_{ξ^*} to $\pi_{\xi^*}^\perp(P)$ is orthogonal to $\mathcal{Q}_{\xi^*}$. Let $Q = \pi_{\xi^*}^\perp(P)$. Then for any sufficiently close $R = Q + \mathrm{d}Q \in \mathcal{Q}$ with $\boldsymbol{\xi}_R = \boldsymbol{\xi}_{\hat{Q}} + \mathrm{d}\boldsymbol{\xi}$ and $\mathrm{d}\boldsymbol{\xi} \neq 0$ the triangle $P, Q, R \in \mathcal{P}$ is orthogonal in $\mathcal{Q}$ and Theorem 8 gives the relation $D[P \parallel R] > D[P \parallel Q]$. This shows, that Q is a critical point w.r.t. the divergence $D[P \parallel Q]$. Conversely since $\mathcal{Q}$ is a smooth submanifold the mean value theorem shows that for any critical point $Q \in \mathcal{Q}$, w.r.t. the divergence $D[P \parallel Q]$ a dual affine projection from P to Q exists and therefore in particular for the points $\hat{Q} \in \mathcal{Q}$ that minimizes the divergence. From equation 3.10 we obtain for the distance that $D[P \parallel \hat{Q}] \leq d(P, \hat{Q})$. Furthermore by definition $d(P, \hat{Q})$ is the minimal length of a curve from P to $\hat{Q}$, but since there exists a dual affine projection from P to $\hat{Q}$, which has the length $D[P \parallel \hat{Q}]$ it follows that $d(P, \hat{Q}) = D[P \parallel \hat{Q}]$ and therefore the geodesic projection is a dual affine projection. By applying the dual version of Theorem 8 this argument mutatis mutandis also holds for the dual geodesic projection w.r.t. the affine projection. □

Corollary 11. *Let* $(X, \mathcal{P}_\xi, D_\psi)$ *be a dually flat statistical manifold and* $(X, \mathcal{Q})$ *and* $(X, \mathcal{S})$ *smooth submanifolds. Let further be* $(X, \mathcal{Q})$ *flat w.r.t.* D_{ψ^*}

and $(X, \mathcal{S})$ *flat w.r.t.* D_ψ*. Then the geodesic projection* $\pi : \mathcal{P} \to \mathcal{Q}$ *is uniquely given by an* ξ^**-affine projection and the dual geodesic projection* $\pi^* : \mathcal{P} \to \mathcal{S}$ *is uniquely given by an* ξ*-affine projection.*

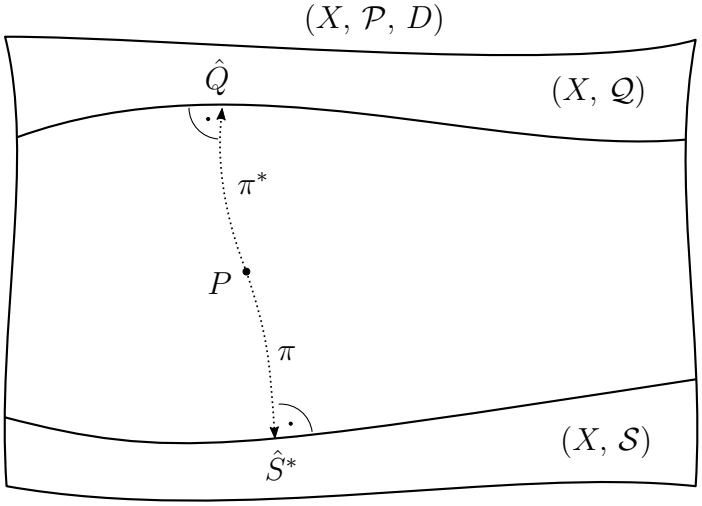

FIGURE 4.3. Unique projections in dually flat manifolds

Proof. By virtue of Theorem 10 it suffices to proof the uniqueness of the affine projection and the dual affine projection. Let $p \in \mathcal{P}$, then Lemma 9 asserts the existence of a dual affine projection of P to a point $\pi(P) = \hat{Q} \in \mathcal{Q}$ and since $(X, \mathcal{Q})$ is flat it follows, that $\mathcal{Q} \subseteq T_Q\mathcal{Q}$ such that for any $R \in \mathcal{Q}$ Theorem 8 shows that:

$$D[P \parallel R] = D[P \parallel Q] + D[Q \parallel R] \geq D[P \parallel Q]$$

Therefore Q is the global minimum and $\pi(P)$ is unique. By the application of the dual version of Theorem 8 to the submanifold $(X, \mathcal{S})$ the argument mutatis mutandis also proves, that $\pi^*(S) = \hat{S} \in \mathcal{S}$ is unique in $(X, \mathcal{S})$. □

REFERENCES

[1] Shun-ichi Amari. Differential geometrical theory of statistics. *Differential geometry in statistical inference*, pages 19–94, 1987.

[2] Nikolai Nikolajewitsch Chenzow. Categories of mathematical statistics. *Uspekhi Mat. Nauk*, 14(2 (86)):87–158, 1965.