Min-Max Kullback-Leibler Model Selection

Jason K. Johnson

May 25, 2002

Abstract. *This paper considers an information theoretic min-max approach to the model selection problem. The aim of this approach is to select the member of a given parameterized family of probability models so as to minimize the worst-case Kullback-Leibler divergence from an uncertain "truth" model. Uncertainty of the truth is specified by an upper-bound of the KL-divergence relative to a given reference model which provides an uncertain observation of the truth. We consider this problem in the context of regular exponential family models where the existence and uniqueness of such an optimal approximation is demonstrated. Furthermore, necessary and sufficient conditions for optimality are provided leading to the development of an iterative solution technique.*

1 Introduction

This paper considers a min-max approach for probabilistic modeling based upon information theoretic principles. Fundamental to any model selection philosophy are the notions of model uncertainty and model fidelity. Here, the Kullback-Leibler divergence is adopted to address both of these concerns. Uncertainty in the "true" probability model of the underlying process is specified by a KL-divergence relative to a given reference model (which provides an uncertain observation of the unknown truth). We then consider the problem of selecting the best approximation for this true model among some lower-order parameterized family of models. The KL-divergence is again employed as a measure of the fidelity of a candidate approximation to the truth. The proposed min-max criteria then selects the candidate model which minimizes the worst-case KL-divergence from the truth. This problem is considered in the context of regular exponential family models where the requisite information geometry is readily developed. Here, the solution of this min-max model selection problem is characterized and an iterative solution technique is proposed.

To be more precise, suppose we are performing density estimation with respect to some unknown density g considered to be a member of a known family of models $\mathcal{F}$. Suppose also that all we know about g is that it is "near" some given reference model g_0 in KL-divergence. The KL-divergence between density functions f and g is defined as $D(f||g) = E_f \log \frac{f(\mathrm{x})}{g(\mathrm{x})}$. In particular, let us assume that we are assured that the KL-divergence $D(g||g_0)$ is bounded above by ρ. The unknown density is then a member of the set of models below.

$$\mathcal{G} \equiv \{g \in \mathcal{F} | D(g||g_0) \leq \rho\} \tag{1}$$

Suppose now that we wish to perform density estimation of g within some embedded (presumably lower-order) family of models $\mathcal{H} \subset \mathcal{F}$. This will typically move us further away from the unknown true density g. However, we might expect that the divergence from the true unknown model may be bounded and that, for an appropriately chosen family of candidates, this bound may be kept small. Let us denote this bound for a given density estimate $h \in \mathcal{H}$ by $\rho(h)$ as defined below.

$$\rho(h) \equiv \sup_{g \in \mathcal{G}} D(g||h) \tag{2}$$

A natural design criteria is then to select $h \in \mathcal{H}$ so as to make this upper bound of the modeling error as small as possible. The resulting bound is then given by

$$\rho^* \equiv \inf_{h \in \mathcal{H}} \rho(h) \tag{3}$$

$$= \inf_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} D(g||h) \tag{4}$$

We are then interested both in identifying the best model $h^* \in \mathcal{H}$ which achieves the above infimum as well as (at least an upper bound of) this worst-case modeling error ρ^* so that we know how good of an approximation h^* is for the unknown truth g.

Below, this problem is considered in the context of the exponential family of models where we exploit

the information geometry of this family with respect to the KL-divergence (the properties of this geometry rely heavily upon convex analysis) and also will appeal to some results of minimax theory to characterize the solution of the above problem and suggest algorithms for solving such problems.

2 The Exponential Family

We will analyze this min-max optimization problem in the context of the exponential family of models described below. See Barndorff-Nielsen for a thorough, rigorous treatment of this family emphasizing convexity and duality principles [BN78].

An exponential family of models for a random variable x with state-space X is specified with respect to a given set of k sufficient statistics $t(x) \in R^k$. Here we take the state-space X to be R^n such that x is a continuous-valued random vector. The probability density function for x is then of the form given below.

$$f(x;\theta) = \exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{5}$$

The so-called cumulant function $\varphi(\theta)$ serves to normalize the density to contain unit probability.

$$\varphi(\theta) = \log \int \exp\{\theta \cdot t(x)\} dx \tag{6}$$

Note that the integral is positive for all finite θ such that $\varphi(\theta) > -\infty$ for all $\theta \in R^k$. The density (5) constitutes an admissible (normalizable) pdf if and only if the cumulant function is finite $\varphi(\theta) < \infty$. Hence, the admissible domain of θ is specified by this condition.

$$\mathcal{F}_\theta \equiv \{\theta \in R^k | \varphi(\theta) < \infty\}. \tag{7}$$

To construct an exponential family of models, the statistics $t(x)$ and state-space X must be chosen so that $\mathcal{F}_\theta$ is non-empty. The exponential family then consists of the set of densities $\mathcal{F} = \{f(x;\theta) \mid \theta \in \mathcal{F}_\theta\}$. This is said to constitute a *regular* exponential family if the set $\mathcal{F}_\theta$ is "full" such that it has nonempty interior in R^k. This representation of $\mathcal{F}$ is said to be *minimal* if distinct values of θ always specify distinct densities. The parameters θ are then referred to as the *exponential coordinates* of $\mathcal{F}$. For a regular exponential family, the representation is minimal if and only if the statistic functions $t(\cdot)$ are linearly independent. Below, we assume that we are working with such a minimal representation of a regular exponential family.

We now examine some useful properties of the cumulant function $\varphi(\theta)$. First, we note the moment generating property of the cumulant function. Let us denote the vector of mean sufficient statistics as $\eta \equiv E_\theta t(\mathrm{x})$ which are often referred to as the *moments*. Evaluating the vector of partial derivatives of the cumulant function with respect the exponential parameters θ yields the moments.

$$\frac{\partial \varphi(\theta)}{\partial \theta_i} = \eta_i \tag{8}$$

We adopt the notational convention that the moments η are implicitly coupled to the parameters θ so that it is understood that η varys with θ. Let $\Lambda : \mathcal{F}_\theta \to R^k$ denote this mapping from θ the η. Below, we shall see that (under the assumption of minimality) this mapping proves to be injective and hence bijective with respect to the image $\mathcal{F}_\eta \equiv \Lambda(\mathcal{F}_\theta)$.

Evaluating the matrix of second derivatives of the cumulant function reveals that this Hessian matrix is just the covariance of the sufficient statistics.

$$\frac{\partial^2 \varphi(\theta)}{\partial \theta_i \partial \theta_j} = E_\theta\{(t_i(\mathrm{x}) - \eta_i)(t_j(\mathrm{x}) - \eta_j)\} \tag{9}$$

We will denote this symmetric positive semi-definite matrix by $G(\theta) = (g_{ij}(\theta))$

$$g_{ij}(\theta) = \frac{\partial^2 \varphi(\theta)}{\partial \theta_i \partial \theta_j} \tag{10}$$

which may be identified as the Fisher information matrix with respect to the parameters θ defined as the covariance of the zero-mean "score" random vector $\mathrm{v}_\theta = \frac{\partial}{\partial \theta} \log f(\mathrm{x};\theta)$ below

$$g_{ij}(\theta) = E_\theta \left\{ \frac{\partial}{\partial \theta_i} \log f(\mathrm{x};\theta) \; \frac{\partial}{\partial \theta_j} \log f(\mathrm{x};\theta) \right\}. \tag{11}$$

Note that the positive semi-definiteness of $G(\theta)$ implies that $\varphi(\theta)$ is convex on $\mathcal{F}_\theta$. If the sufficient statistics are minimal (linearly independent) and θ indexes a regular exponential family then this covariance is positive definite and $\varphi(\theta)$ is strictly convex on $\mathcal{F}_\theta$. Also note that the Fisher information evaluated at θ indicates the first-order sensitivity of the moments with respect to small perturbations of the exponential parameters away from θ.

$$G(\theta) = \frac{\partial \eta}{\partial \theta} \tag{12}$$

So $G(\theta)$ is the Jacobian of the map $\Lambda : \theta \to \eta$. Assuming minimality, the inverse Fisher information then

specifies the local sensitivity of the exponential parameters to perturbations of the moments.

$$G^{-1}(\theta) = \frac{\partial \theta}{\partial \eta} \tag{13}$$

Exploiting the convexity of the cumulant function, we may define a "dual" function φ^* by the convex conjugate of φ as defined below.

$$\varphi^*(\beta) \equiv \sup_{\theta \in \mathcal{F}_\theta} \{\theta \cdot \beta - \varphi(\theta)\} \tag{14}$$

Because $\varphi(\theta) > -\infty$ is bounded below, the supremum is achieved. Any maximizer $\theta^* \in \mathcal{F}_\theta$ achieving the supremum must satisfy the stationarity condition $E_{\theta^*} t(\mathrm{x}) = \beta$. Since the objective function is strictly concave, this stationary condition is then sufficient and has a unique solution $\theta^*(\beta)$. Since $E_\theta t(\mathrm{x}) = \eta$ and we now see that there is only one θ^* satisfying $E_{\theta^*} t(\mathrm{x}) = \eta$ we have that $\theta^*(\eta) = \theta$. Hence, the map Λ is one-to-one with respect to the image $\mathcal{F}_\eta \equiv \Lambda(\mathcal{F}_\theta)$ and the inverse map is given by $\Lambda^{-1}(\eta) = \theta^*(\eta)$. The supremum may then be expressed in terms of a dually-coupled pair of coordinates (θ, η).

$$\varphi^*(\eta) = \theta \cdot \eta - \varphi(\theta) \tag{15}$$

This relation is sometimes referred to as the *Legendre transform*.

Exploiting this 1-to-1 correspondence between $\theta \in \mathcal{F}_\theta$ and $\eta \in \mathcal{F}_\eta$ we may view the exponential family of densities $\mathcal{F}$ as being indexed by the set $\mathcal{F}_\eta = \Lambda(\mathcal{F}_\theta)$. Hence, the moments η are sometimes referred to as the *moment coordinates* of a minimal exponential family. For instance, the density may be reparameterized in moment coordinates as $f^*(x;\eta) \equiv f(x;\Lambda^{-1}(\eta)$. The "dual" Fisher information matrix $G^*(\eta) = (g^*_{ij}(\eta))$ associated with this moment parameterization may then evaluated as

$$G^*(\eta) = G^{-1}(\theta) \tag{16}$$

where θ and η are dually coupled by Λ.

Employing (13), we may differentiate the Legendre transform with respect to η to reveal the following dual expressions for the gradient and Hessian of $\varphi^*(\eta)$.

$$\frac{\partial \varphi^*(\eta)}{\partial \eta_i} = \theta_i \tag{17}$$

$$\frac{\partial^2 \varphi^*(\eta)}{\partial \eta_i \partial \eta_j} = g^*_{ij}(\eta) \tag{18}$$

Note that the positive-definiteness of the Fisher information (under the assumption of minimality) insures that the dual function $\varphi^*(\eta)$ is strictly convex on $\mathcal{F}_\eta$. In this case, the convex conjugate of the dual function recovers the cumulant function completing the "duality" of this formalism.

Finally, we note that the dual function may be interpreted as the negative entropy of the model since the entropy is given by $h(\eta) \equiv -E_\theta \log f(\mathrm{x};\theta) = -\{\theta \cdot \eta - \varphi(\theta)\}$.

3 Information Geometry

We now explore the information geometry of the exponential family under the KL-divergence "distance" measure and the associated "projection" problems. This geometric viewpoint was developed by Efron, Csiszár, Amari and others [Efr78, Csi75, Ama82, Ama01]. Amari's terminology is favored here with some modification.

Kullback-Leibler divergence. The Kullback-Leibler divergence between two probability density functions $p(x)$ and $q(x)$ is defined as

$$D(p||q) \equiv E_p \log \frac{p(x)}{q(x)}. \tag{19}$$

For two exponential family models $f_1, f_2 \in \mathcal{F}$ specified by θ_1 and θ_2 the KL-divergence may be evaluating from (5) as

$$D(f_1||f_2) = \eta_1 \cdot (\theta_1 - \theta_2) - (\varphi(\theta_1) - \varphi(\theta_2)) \tag{20}$$

where $\eta_1 = \Lambda(\theta_1)$. We may exploit the Legendre transform to derive the dual form of this KL-divergence expressed in terms of the moment parameters η_1 and η_2 and the dual function as

$$D(f_1||f_2) = \theta_2 \cdot (\eta_2 - \eta_1) - (\varphi^*(\eta_2) - \varphi^*(\eta_1)) \tag{21}$$

where $\theta_2 = \Lambda^{-1}(\eta_2)$.

Bregman distances. For exponential family models, much insight into the properties of the KL-divergence may be gained from purely geometrical considerations by viewing the above expressions as *Bregman distances* respectively arising from the convexity of the functions $\varphi(\theta)$ and $\varphi^*(\eta)$. These distances are formed by taking the difference between a convex function and a tangental hyperplane underestimate of the function.

For instance, given that the cumulant function $\varphi(\theta)$ is convex, we may construct an associated Bregman distance as $B(\theta;\theta_0) \equiv \varphi(\theta) - \bar{\varphi}(\theta;\theta_0)$ where $\bar{\varphi}(\theta;\theta_0)$

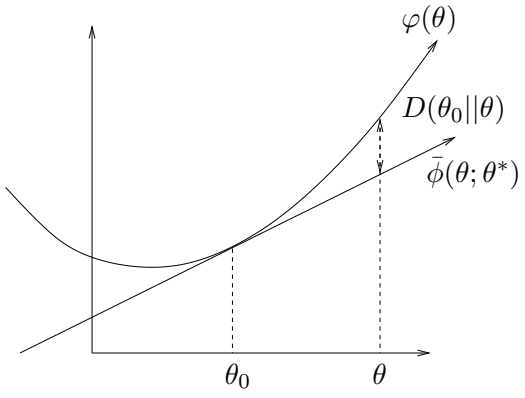

Figure 1: Illustration of Bregman distance interpretation of KL-divergence.

is an underestimate of $\varphi(\theta)$ based upon the tangental hyperplane approximation constructed at θ_0.

$$\begin{aligned}\bar{\varphi}(\theta;\theta_0) &= \varphi(\theta_0) + \nabla\varphi(\theta_0)\cdot(\theta-\theta_0) \quad (22)\\ &= \varphi(\theta_0) + \eta_0\cdot(\theta-\theta_0) \quad (23)\end{aligned}$$

In the second line above, η_0 is dually coupled to θ_0. It is then easy to verify that this Bregman distance equals the first expression for the KL-divergence defined earlier where θ_2 is viewed as the variable and θ_1 is fixed.

$$B(\theta_2;\theta_1) = D(f_1||f_2) \quad (24)$$

In a similar manner, the convex conjugate $\varphi^*(\eta)$ may be used to construct a "dual" Bregman distance in the moment parameters as $B^*(\eta) \equiv \varphi^*(\eta) - \bar{\varphi}^*(\eta;\eta_0)$ where $\bar{\varphi}^*(\eta;\eta_0) = \varphi^*(\eta_0) - \theta_0\cdot(\eta-\eta_0)$. This dual Bregman distance is then equivalent to the dual expression for the KL-divergence derived earlier by the Legendre transform where the η_1 is variable and η_2 is held fixed.

$$B^*(\eta_1;\eta_2) = D(f_1||f_2) \quad (25)$$

In either case, it is apparent from the strict convexity of φ and φ^* that the KL-divergence $D(f_1||f_2)$ is nonnegative and is zero if and only if $\theta_1 = \theta_2$ (equivalently, $\eta_1 = \eta_2$). Note that the Bregman distance inherits the convexity (in the first argument) of the convex function used to construct it. Hence, these dual interpretations of the KL-divergence reveals that the KL-divergence $D(f_1||f_2)$ is a strict convex function in either η_1 or θ_2. This suggests that the KL-divergence is most naturally expressed in the "mixed" coordinates (η_1, θ_2). The expression for the KL-divergence under this parameterization is derived from the earlier expression and the Legendre transform.

$$D(f_1||f_2) = \varphi^*(\eta_1) + \varphi(\theta_2) - \eta_1\cdot\theta_2 \quad (26)$$

This clearly represents the KL-divergence as a *convex bifunction* which is the analogue in convex analysis of a bilinear function in linear algebra. Note that this form of the KL-divergence does not require coordinate transforms.

Finally, we note that the interpretation of the KL-divergence as a Bregman distance leads to a sort of triangle law analogous to the law of cosines in Euclidean geometry. This is arrived at by decomposing the Bregman distance function $B(\theta_2;\theta_0)$ as a sum of two Bregman distances relative to a third point θ_1 plus a linear correction term

$$B(\theta_2;\theta_0) = B(\theta_1;\theta_0) + B(\theta_2;\theta_1) + \Delta\cdot(\theta_2-\theta_1) \quad (27)$$

where

$$\Delta = \nabla\varphi(\theta_1) - \nabla\varphi(\theta_0). \quad (28)$$

This then leads to the following decomposition of the KL-divergence.

$$D(f_0||f_2) = D(f_0||f_1) + D(f_1||f_2) + (\eta_0-\eta_1)\cdot(\theta_1-\theta_2) \quad (29)$$

The analogous triangle identity of Euclidean geometry may be expressed for three vectors $x, y, z \in R^n$ as

$$\frac{1}{2}||x-z||^2 = \frac{1}{2}||x-y||^2 + \frac{1}{2}||y-z||^2 + (x-y)\cdot(y-z). \quad (30)$$

Roughly speaking, the KL-divergence plays an analogous role to that of half the squared distance in Euclidean geometry. The decomposition (29) reduces to the "Pythagorean law" of information geometry which arises under the "othogonality" condition $(\eta_0-\eta_1)\cdot(\theta_1-\theta_2) = 0$. This notion of orthogonality will be explored further in a later section.

Differential Analysis. Some useful differential relations for the KL-divergence within an exponential family of models are provided below. These may be derived by differentiating the above expression for the KL-divergence and employing those differential relations developed earlier regarding the cumulant and dual functions.

Below, the first and second partial derivatives of the second argument of the $D(f^*||f)$ with respect to the exponential coordinates θ of f are given below.

$$\frac{\partial D(f^*||f)}{\partial\theta_i} = \eta_i - \eta_i^* \quad (31)$$

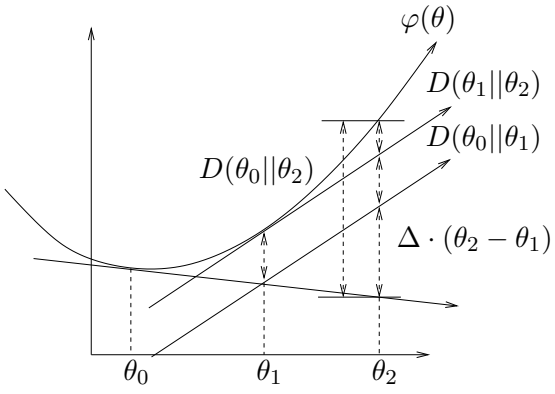

Figure 2: Illustration of "triangle law" arising from Bregman distance interpretation.

$$\frac{\partial^2 D(f^*||f)}{\partial\theta_i \partial\theta_j} = g_{ij}(\theta) \tag{32}$$

Also, the first and second order partial derivative of the first argument of the KL-divergence $D(f||f^*)$ with respect to the exponential coordinates η of f are given by

$$\frac{\partial D(f||f^*)}{\partial\eta_i} = \theta_i - \theta_i^* \tag{33}$$

$$\frac{\partial^2 D(f||f^*)}{\partial\eta_i \partial\eta_j} = g_{ij}^*(\eta) \tag{34}$$

where $G^*(\eta) \equiv G^{-1}(\theta)$ is the Fisher information matrix associated with the moment parameterization of the exponential family $f^*(x;\eta) \equiv f(x;\Lambda^{-1}(\eta))$.

Note the duality between the above relations. Also, note that both Hessian matrices are positive definite verifying the earlier claim that KL-divergence is strictly convex in either η_1 or θ_2.

Finally, the above dual Hessians clarify the role the dual pair of Fisher information matrices $G(\theta)$ and $G^*(\eta) = G^{-1}(\theta)$ play as Riemannian metrics respectively in θ and η. Consider the second order Taylor expansion of $D(\theta^*||\theta)$ (this denotes $D(f^*||f)$ but where the parameterization of f^* and f is made explicit) in θ about θ^*. Because the KL-divergence and the gradient vanish when $\theta = \theta^*$ we have that for small perturbations $\delta\theta$ the KL-divergence to second order is given by

$$D(\theta^*||\theta^* + \delta\theta) \approx \frac{1}{2}\delta\theta' G(\theta^*)\delta\theta. \tag{35}$$

Likewise, in moment coordinates we have that

$$D(\eta^* + \delta\eta||\eta^*) \approx \frac{1}{2}\delta\eta' G^*(\eta^*)\delta\eta. \tag{36}$$

Hence, we again see that the KL-divergence is analogous to half the squared Euclidean distance but under the appropriate Riemannian metric. Associated to these dually-coupled Riemannian metric spaces are equivalent dual notions of "local" orthogonality with respect to the respective Riemannian metrics. These notions will be explored further below. It is also worth noting that, for a small perturbation δf away from f^* expressed in either exponential coordinates as $\delta\theta$ or, equivalently, in moment coordinates as $\delta\eta$, we have that $\delta\eta \approx G(\theta^*)\delta\theta$ so that the two KL-divergences above are equivalent to first order and may both be expressed as $\delta\eta \cdot \delta\theta$. Roughly speaking, the KL-divergence between nearby distributions is given by an inner product operation applied to their coordinate difference but where the operation of transpose is replaced by the "conjugate transpose" operation.

Flat submanifolds. There are two dual notions of "flatness" in information geometry. A subset $\mathcal{G} \subset \mathcal{F}$ is said to be *m-flat* if it's representation in moment coordinates is given by the intersection of an affine subset of R^k with $\mathcal{F}_\eta$ in which case $\mathcal{G}$ is referred to as an *m-flat submanifold*. Similarly, $\mathcal{G}$ is said to be *e-flat* if it's representation in exponential coordinates is given by the intersection of affine subset of R^k with $\mathcal{F}_\theta$. In either case, the *dimension* of the submanifold is defined as the dimension of the corresponding affine subset.

One-dimensional m-flat and e-flat submanifolds are respectively referred to as *m-geodesics* and *e-geodesics*. For instance, given two densities $f_1, f_2 \in \mathcal{F}$, we may construct a one-parameter family of densities giving the e-geodesic connecting them by mixing log-densities

$$\log f(x;\lambda) = \lambda \log f_1(x) + (1-\lambda)\log f_2(x) - \psi(\lambda) \tag{37}$$

for all values of λ such that the normalization constant

$$\psi(\lambda) = \log \int f_1^\lambda(x) f_2^{1-\lambda}(x) dx \tag{38}$$

is finite so as to insure $f(x;\lambda)$ contains unit probability. In exponential coordinates, this corresponds to the intersection of the line

$$\theta(\lambda) = \lambda\theta_1 + (1-\lambda)\theta_2 \tag{39}$$

with the set of normalizable parameter settings $\mathcal{F}_\theta$.

Similarly, the m-geodesic connecting them is given in moment coordinates by the intersection of the line

$$\eta(\lambda) = \lambda\eta_1 + (1-\lambda)\eta_2 \tag{40}$$

with the set of admissible moments $\mathcal{F}_\eta$.

Convex Sets. There are accordingly two dual notions of convex sets in information geometry. We shall say that a subset of the exponential family $\mathcal{G} \subset \mathcal{F}$ is (strictly) *e-convex* if the representation of that set in exponential coordinates is a (strictly) convex subset of R^k. This means that given any two densities $g_1, g_2 \in \mathcal{G}$ the set of log-mixtures $g(x;\lambda) = g_1^\lambda(x) g_2^{1-\lambda}(x) e^{-\psi(\lambda)}$ for $\lambda \in [0,1]$ is contained in $\mathcal{G}$. Likewise, $\mathcal{G}$ is (strictly) *m-convex* if the moment representation is a (strictly) convex set.

Strict convexity (respectively in either exponential or moment coordinates) means that no point in the surface may be expressed as a convex combination of other points in that set. Strictly convex sets have the property that every tangent intersects that set at just one point.

Othogonality in Information Geometry. Consider two connected, smooth one-parameter submanifolds of $\mathcal{F}$ given by $\mathcal{F}^{(1)} = \{f_1(x;s) | s \in S\}$ and $\mathcal{F}^{(2)} = \{f_2(x;t) | t \in T\}$ where S and T are intervals of the real numbers. Suppose also that these submanifolds intersect at $f_0(\cdot) = f_1(\cdot; s_0) = f_2(\cdot; t_0)$. Then we shall say that these submanifolds are $\mathcal{I}$-orthogonal at f_0 if the following condition holds.

$$E_{f_0}\left\{ \frac{\partial \log f_1(\mathrm{x}; s_o)}{\partial s} \frac{\partial \log f_2(\mathrm{x}; t_0)}{\partial t} \right\} = 0. \qquad (41)$$

This asserts that the zero-mean "score" statistics associated with these two families are uncorrelated at f_0. This may be interpreted geometrically, within either exponential or moment coordinates, with the appropriate Fisher information matrix acting as the Riemannian metric tensor.

Let $\theta_1(s)$ and $\theta_2(t)$ denote the curves traced by these submanifolds in exponential coordinates with tangent vectors $\dot{\theta}_1(s)$ and $\dot{\theta}_2(t)$. Then the earlier $\mathcal{I}$-orthogonality condition is equivalent to requiring that the tangent vectors at $\theta_0 = \theta(s_0) = \theta(t_0)$ are orthogonal with respect to the Fisher information metric $G(\theta)$ evaluated at θ_0.

$$\dot{\theta}_1'(s_0) G(\theta_0) \dot{\theta}_2(t_0) = 0 \qquad (42)$$

Hence, the submanifolds $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ are $\mathcal{I}$-orthogonal at f_0 if and only if their exponential coordinate representations $\mathcal{F}_\theta^{(1)}$ and $\mathcal{F}_\theta^{(2)}$ are $G(\theta_0)$-orthogonal at θ_0.

Likewise, in moment coordinates $\mathcal{I}$-orthogonality of submanifolds $\eta_1(s)$ and $\eta_2(s)$ at η_0 is expressed relative to the "dual" Fisher information $G^*(\eta) = G^{-1}(\theta)$ evaluated at η_0.

$$\dot{\eta}_1'(s_0) G^*(\eta_0) \dot{\eta}_2(t_0) = 0 \qquad (43)$$

The submanifolds $\mathcal{F}^{(1)}$ and $\mathcal{F}^{(2)}$ are $\mathcal{I}$-orthogonal at f_0 if and only if their moment coordinate representations $\mathcal{F}_\eta^{(1)}$ and $\mathcal{F}_\eta^{(2)}$ are $G^*(\eta_0)$-orthogonal at η_0.

In "mixed" coordinates, this reduces to the condition that the tangent vectors are orthogonal (in the usual Euclidean sense). For instance,

$$\dot{\eta}_1'(s_0) \dot{\theta}_2(t_0) = 0. \qquad (44)$$

This latter form of othogonality was noted by Efron in the context of the exponential family while the dual Riemannian picture is emphasized by Amari more generally.

We will say that two smooth submanifolds (not necessarily one-parameter submanifolds as above) are $\mathcal{I}$-orthogonal at an intersection point f_0 if every pair of embedded one-parameter submanifolds (taken from the respective manifolds) both containing f_0 are $\mathcal{I}$-orthogonal at f_0.

The latter "mixed" coordinate test for $\mathcal{I}$-othogonality is especially intuitive when we are working with pairs of m-flat/e-flat submanifolds (our primary concern in later sections of the paper). For instance, If $\mathcal{F}^{(1)}$ is an m-flat submanifold and $\mathcal{F}^{(2)}$ is an e-flat submanifold then the above local definition of $\mathcal{I}$-orthogonality at an intersection point f_0 is equivalent to the global characterization

$$\forall \eta_1 \in \mathcal{F}_\eta^{(1)}, \theta_2 \in \mathcal{F}_\theta^{(2)} : \ (\eta_1 - \eta_0)'(\theta_2 - \theta_0) = 0 \quad (45)$$

This is consistent with the earlier suggestion in connection to the "Pythagorean" theorem of information geometry. Note also that such an $\mathcal{I}$-orthogonal pair of m-flat/e-flat submanifolds intersect at just one point so we may omit the "at" qualification.

KL Level sets. Next, we consider the level sets of the KL-divergence. We may define two types of level sets depending upon which of the arguments of $D(f_1||f_2)$ is held fixed.

We shall define an *m-ball* with center f_0 and radius ρ as the subset of exponential family models $f \in \mathcal{F}$ such that the KL-divergence $D(f||f_0)$ does not exceed ρ. Denote this set by $\mathcal{B}(\rho; f_0) \subset \mathcal{F}$.

$$\mathcal{B}(\rho; f_0) = \{ f \in \mathcal{F} \mid D(f||f_0) \leq \rho \} \qquad (46)$$

We will denote the corresponding representation of an m-ball in exponential coordinates by $\mathcal{B}_\theta(\rho; \theta_0)$ and in

moment coordinates by $\mathcal{B}_\eta(\rho;\eta_0)$ where θ_0 and η_0 are respectively the exponential and moment coordinates of the density f_0. The strict convexity of $D(f_1||f_2)$ in η_1 implies that m-balls are strictly m-convex.

The dual notion of an *e-ball* centered about the density f_0 is defined in similar fashion but where the second argument of the KL-divergence is varied while the first is held fixed.

$$\mathcal{B}^*(\rho;f_0) = \{f \in \mathcal{F} \mid D(f_0||f) \leq \rho\} \tag{47}$$

The associated coordinate representations are denoted by $\mathcal{B}^*_\theta(\rho;\theta_0)$ and $\mathcal{B}^*_\eta(\rho;\eta_0)$. The strict-convexity of $D(f_1||f_2)$ in θ_2 insures that e-balls are strictly e-convex.

KL Spheres. Also, we define related notions of *m-spheres* and *e-spheres* respectively as the surfaces of m-balls and e-balls by replacing inequalities with equalities in the above definitions. We will denote m-spheres by $\partial\mathcal{B}(\rho;f_0)$ and e-spheres by $\partial\mathcal{B}^*(\rho;f_0)$ (employing similar notation for their coordinate representations). Since $D(f_1||f_2)$ is a smooth function of η_1 or θ_2, m-spheres and e-spheres are smooth surfaces respectively in moment and exponential coordinates. Furthermore, due to the strict m-convexity of m-balls, each point in an m-sphere is associated to a unique m-flat tangent plane which intersects that sphere (and the enclosed m-ball) at just that point. Likewise, each point in an e-sphere specifies a unique e-flat tangent plane which intersects that e-sphere (e-ball) at just that point.

Compactness of KL Level-Sets. We may define two dual notions of a compact set in information geometry. We shall say that a subset $\mathcal{G} \subset \mathcal{F}$ is *e-compact* if its representation in exponential coordinates $\mathcal{G}_\theta \subset \mathcal{F}_\theta$ is compact (closed and bounded). Likewise, we shall say that $\mathcal{G}$ is m-compact when its moment representation $\mathcal{G}_\eta$ is compact.

We now argue that, assuming a minimal representation of a regular exponential family, m-balls are m-compact. (partial idea for proof). Suppose the Bregman distance function $B^*(\eta;\eta_0)$ is coercive so that $B^*(\eta;\eta_0) \to \infty$ as η either (i) diverges $||\eta|| \to \infty$ or (ii) approaches a closure point of $\mathcal{F}_\eta$. Then (i) implies that m-balls are bounded while (ii) implies that m-balls are closed (and hence compact). This is equivalent to showing that $\varphi^*(\eta)$ is "super-tangentally" coercive so that it diverges faster than any tangental supporting hyperplane as either (i) or (ii) occur. Under the assumption of minimality, $\varphi^*(\eta)$ is strictly convex such that coercivity implies super-tangental coercivity. Since $\varphi^*(\eta)$ is a convex conjugate function, its level sets are closed and convex. This implies coercivity with respect to (ii), otherwise some level-set "touches" the boundary and is not closed. We still need to show that the level-sets are bounded implying coercivity with respect to (i). (incomplete...)

By an analogous "dual" argument with respect to $B(\theta;\theta_0)$, e-balls are e-compact.

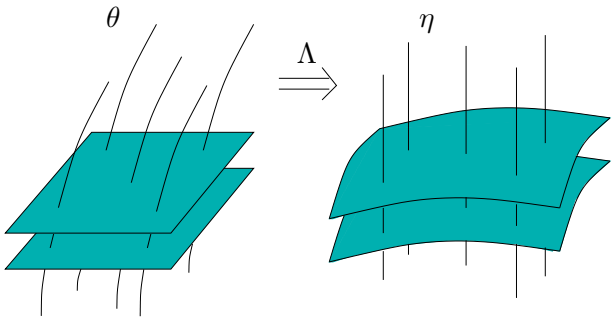

Figure 3: Illustration of e-foliation and $\mathcal{I}$-transverse m-fibration. Leaves are parallel planes in exponential coordinates while fibers are parallel lines in moment coordinates. Fibers in moment coordinates are perpendicular to leaves in exponential coordinates. Reverse picture applies for m-foliation with $\mathcal{I}$-transverse e-fibration.

Planar Foliations and Transverse Fibrations. In Euclidean geometry it is trivial to note that the space R^n may be partitioned into a set of "leaves" consisting of all $(n-1)$-dimensional hyperplanes orthogonal to a given unit vector $\hat{u}$ and that we may construct a complimentary partitioning of the space consisting of the set of lines parallel to $\hat{u}$. This set of planes may be considered as a *planar foliation* of the space with the set of lines forming a *$\mathcal{I}$-transverse fibration*. This means that each "fiber" is orthogonal to each "leaf" at their point of intersection. This picture is useful in that the fibers of the foliation then characterize the Euclidean projections to any of the leaves of the foliation. Analogous ideas arise in information geometry but where we may now define two types of flat foliations depending on whether we prefer m-flat or e-flat foliations (we shall see that each approach has an associated projection problem).

Towards this end, we define an *m-flat foliation* orthogonal to a given unit vector $\hat{u} \in R^k$ to consist of the set of all $k-1$ dimensional m-flat submanifolds

$\{M_\lambda\}$ orthogonal to $\hat{u}$ in moment-coordinates. These leaves may be defined by $M_\lambda = \{\eta \in \mathcal{F}_\eta | \hat{u} \cdot \eta = \lambda\}$ and are indexed by $\lambda \in R$. We then define the *e-flat fibration* parallel to $\hat{u}$ to consist of the set of all e-flat geodesics $\{E_\Delta\}$ parallel to $\hat{u}$ in exponential coordinates. These fibers may be defined by $E_\Delta = \{\theta \in \mathcal{F}_\theta | \exists \lambda \in R : \theta = \Delta + \lambda \hat{u}\}$ and are indexed by $\Delta \in M_0$. We now argue that the m-flat foliation is $\mathcal{I}$-*transverse* to the e-flat fibration in that each leaf is $\mathcal{I}$-orthogonal to each fiber.

Consider an arbitrary e-flat fiber E_Δ represented in exponential coordinates by $\theta_\lambda = \Delta + \lambda \hat{u}$. Let $\eta_\lambda \equiv \Lambda(\theta_\lambda)$ denote this e-geodesic in moment coordinates. For each λ (such that $\theta_\lambda \in \mathcal{F}_\theta$), the e-geodesic intersects the m-flat leaf M_λ at the single point η_λ. Since the e-geodesic is parallel to $\hat{u}$ in exponential coordinates and the m-flat submanifold is perpendicular to $\hat{u}$ in moment coordinates we have

$$\forall \theta \in E_\Delta, \eta \in M_\alpha : (\theta - \theta_\lambda) \cdot (\eta - \eta_\lambda) = 0 \qquad (48)$$

so that these are $\mathcal{I}$-orthogonal submanifolds.

In an entirely analogous manner we may define dual notions of an *e-flat foliation* and the corresponding $\mathcal{I}$-*transverse m-flat fibration*.

Spherical Foliations and Transverse Radial Fibrations. In Euclidean geometry the set of concentric spheres about a given point forms a partitioning of the space. A complimentary partitioning of the space (excluding the central point) is given by the set of "radial" lines passing through that central point. The set of spheres may be considered as a *spherical foliation* of the space with the set of radial lines providing a *transverse radial fibration* of the space. This means that the radial lines are normal to the surface of the spheres at the points of intersection. Analogous ideas arise in information geometry and prove useful for analyzing certain projection problems.

There are two dual forms of this idea depending of whether we consider spherical foliations consisting of m-spheres or e-spheres. We will define the *spherical m-foliation* of $\mathcal{F}$ with respect to $f_0 \in \mathcal{F}$ as the family of m-spheres centered at f_0. We define the *radial e-fibration* of $\mathcal{F}$ with respect f_0 as the family of e-geodesics containing f_0. We now argue that these are complimentary notions in that the spherical m-foliation is $\mathcal{I}$-transverse to the radial e-fibration.

Consider an arbitrary "radial" e-geodesic E containing $f_0 \in \mathcal{F}$. This e-geodesic may be parameterized as $\{f_\lambda\} \subset \mathcal{F}$ represented in moment coordinates

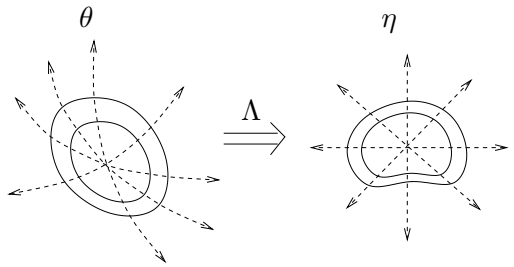

Figure 4: Illustration of spherical e-foliation and radial m-fibration. Spheres are strictly convex in exponential coordinates while radial fibers are straight lines in moment coordinates. M-fibers drawn in moment coordinates are normal to e-spheres drawn in exponential coordinates at intersection points. Reverse picture applies for spherical m-foliation and $\mathcal{I}$-transverse radial e-fibration.

by $\theta_\lambda = \theta_0 + \lambda \hat{u}$ where θ_0 is the moment coordinates of f_0 and $\hat{u}$ is some given unit vector of R^k. Denote the moment-coordinate representation of this e-geodesic by $\eta_\lambda \equiv \Lambda(\theta(\lambda))$ and let $\rho(\lambda) \equiv D(f_\lambda || f_0)$. For each value of λ (such that $\theta_\lambda \in \mathcal{F}_\theta$) the e-geodesic η_λ intersects an m-sphere $\partial\mathcal{B}_\eta(\rho(\lambda); f_0)$ at the point η_λ. Since m-spheres are the surfaces of level sets of $D(f||f_0)$ in η the gradient $\nabla_\eta D(f||f_0) = \theta - \theta_0$ is normal to the surface of the moment-coordinate representation of the m-sphere containing η. So $\theta_\lambda - \theta_0$ is perpendicular to the m-flat submanifold M_λ which is tangent to the m-sphere $\partial\mathcal{B}_\eta(\rho(\lambda); \eta_0)$ at the point η_λ. Since the points $\{\theta_\lambda\}$ form a line segment, this means that

$$\forall \theta \in E, \eta \in M_\lambda : (\theta - \theta_\lambda) \cdot (\eta - \eta_\lambda) = 0 \qquad (49)$$

so that E is $\mathcal{I}$-orthogonal to M_λ. Hence, each radial e-geodesic through f_0 is $\mathcal{I}$-othogonal to each concentric m-sphere about f_0 at the two points of intersection. In this regard, the spherical m-foliation about f_0 is $\mathcal{I}$-transverse to the radial e-fibration through f_0.

We may also define the dual notions of *spherical e-foliation* and *radial m-fibration* relative to a given "center" in the obvious manner. By similar analysis as above, these are $\mathcal{I}$-transverse in that the e-spheres of the foliation are $\mathcal{I}$-orthogonal to the radial m-geodesics of the fibration at the intersection points.

KL Optimization and Projection. In this section we consider various optimization problems involving the KL-divergence and illustrate geometric

principles characterizing the solution of these problems.

E-minimization. Consider the minimization problem

$$d^* = \inf_{g \in \mathcal{G}} D(g||f) \tag{50}$$

given the density $f \in \mathcal{F}$ and a subset of densities $\mathcal{G} \subset \mathcal{F}$. Note that the infimum is guaranteed to exist since $D(g||f)$ is bounded below by zero. We will refer to this as *e-minimization*. If there exist a unique $g^* \in \mathcal{G}$ achieving the infimum then this is referred to as the *e-projection* of f to $\mathcal{G}$. This e-projection is defined when the m-sphere $\partial\mathcal{B}(d^*; f)$ intersects the set $\mathcal{G}$ at exactly one point.

M-minimization. Consider the dual minimization problem obtained by now performing the minimization over the second argument.

$$d^* = \inf_{g \in \mathcal{G}} D(f||g) \tag{51}$$

We will refer to this as *m-minimization*. If there exists a unique $g^* \in \mathcal{G}$ achieving the infimum then this is referred to as the *m-projection* of f to $\mathcal{G}$. This m-projection is defined when the e-ball $\mathcal{B}^*(d^*; f)$ intersects $\mathcal{G}$ at exactly one point.

We may define *e-maximization* and *m-maximization* in a similar manner as above by replacing the infimum by a supremum in each of the above problems. The supremum, however, may not exist depending upon the given density and specified subset of densities over which the supremum is taken. When the supremum does exist and is achieved by a unique member of the given subset of densities, we will refer to these optimal densities as the *e-maximizer* and *m-maximizer* of the respective optimization problems.

Under certain special circumstances we may guarantee the existence and uniqueness of the projections (minimizers) and maximizers defined above. Projection to flat manifolds were considered by Čencov and Csiszár [Čen72, Csi75]. Projections to balls arise in large-deviations theory and are useful for some coding/communications design problems [CT91]. I'm not aware that maximization over balls has been considered before, but the treatment differs little from the like minimization problems. These maximization problems arise in the min-max design approach to be discussed.

E-projection to an m-flat submanifold. Consider the e-minimization problem where $\mathcal{G}$ is an m-flat submanifold. In this case, the e-projection of an arbitrary f to $\mathcal{G}$ is defined and may be characterized as follows. Let M denote the moment-coordinate representation of $\mathcal{G}$. In moment coordinates, consider "growing" the m-sphere $\partial\mathcal{B}_\eta(\rho; \eta)$ (where η is the moment coordinates of f) until it first contacts M when $\rho = d^*$. In moment coordinates, the m-sphere is strictly convex and M is flat so that the m-sphere intersects $\mathcal{G}$ at exactly one point η^*. This intersection point η^* then specifies the moment-coordinates of the e-projection of f to $\mathcal{G}$. Also, M is then tangent to the m-sphere $\partial\mathcal{B}_\eta(d^*; \eta)$ at the point η^*. Hence, the moment-coordinate representation of the radial e-geodesic connecting η to η^* is $\mathcal{I}$-orthogonal to M.

When $\mathcal{G}$ is a $(k-1)$-dimensional m-flat submanifold, this e-geodesic is an element of the e-flat fibration $\mathcal{I}$-transverse to $\mathcal{G}$. So we may think of this e-geodesic as either the element of the radial fibration through f which is $\mathcal{I}$-transverse to $\mathcal{G}$ or as the element of the m-fibration $\mathcal{I}$-transverse to $\mathcal{G}$ which contains f. This e-geodesic is then given by the intersection of the radial fibration through f with the e-flat fibration $\mathcal{I}$-transverse to $\mathcal{G}$. In this regard, we may interpret the radial e-geodesics through f as characterizing the e-projections of f to the m-flat tangent planes of the m-spheres. Likewise, an e-flat fibration characterizes the e-projections to the $\mathcal{I}$-transverse m-flat submanifolds.

M-projection to an e-flat submanifold. The dual problem of m-projection to an e-flat submanifold $\mathcal{G}$ is characterized in an entirely analogous manner by considering the e-sphere which contacts $\mathcal{G}$ at just one point. In moment coordinates, the intersection point θ^* then specifies the m-projection of f to $\mathcal{G}$. The e-flat $\mathcal{G}$ is tangent to the e-sphere at θ^* so that the m-projection is determined by the radial m-geodesic through f which is $\mathcal{I}$-transverse to $\mathcal{G}$. This m-geodesic is $\mathcal{I}$-orthogonal to the e-flat submanifold.

E-projection to an m-ball. Now consider the e-projection problem where $\mathcal{G}$ is an m-ball. Again, consider growing an m-ball in moment coordinates about f until it first contacts $\mathcal{G}$. Since both m-balls are strictly convex, these m-balls then intersect at exactly one point and have a common m-flat tangent plane at that point. This intersection point η^* then specifies the moment-coordinates of the m-projection of f to $\mathcal{G}$. Furthermore, since both radial e-geodesics (respectively through f and the center of $\mathcal{G}$) are $\mathcal{I}$-transverse to the common m-flat tangent plane through η^* (with respect to the Riemannian metric $G(\eta^*)$ in moment coordinates) these are in fact the same e-geodesic (since they must be parallel in exponential coordinates and contain a common

point θ^*). Hence, this e-geodesic may be considered as the intersection of the radial e-fibration through f with the radial e-fibration through the center of the m-ball $\mathcal{G}$. The e-projection of f to an m-ball then lies along the this e-geodesic connecting f to the center of the m-ball g_0. Moreover, this e-projection lies in the arc of this e-geodesic between f and the center of the ball and is determined by the condition that $D(g^*||g_0) = \rho$ where ρ is the radius of the m-ball $\mathcal{G}$. In exponential coordinates, this e-geodesic is parameterized as $\theta_\lambda = \theta_0 + \lambda(\theta - \theta_0)$ where θ_0 and θ are respectively the coordinates of g_0 and f. This corresponds to densities of the form

$$g_\lambda(x) = \frac{1}{Z(\lambda)} f^\lambda(x) g_0^{1-\lambda}(x) \tag{52}$$

with normalization constant

$$Z(\lambda) = \int f^\lambda(x) g_0^{1-\lambda}(x) dx \tag{53}$$

The e-projection g^* is determined by λ^* solving $D(g_\lambda||g_0) = \rho$ for $\lambda \in [0, 1]$ (there exists a unique solution within this interval). This equation may be solved numerically employing simple iterative methods provided the computations $\Lambda(\theta)$ and $\varphi(\theta)$ are available as subroutines.

M-projection to an e-ball. The dual problem of m-projection to an e-ball is analyzed in an analogous manner. The m-projection of f to the e-ball $\mathcal{G}$ lies along the m-geodesic connecting f to the center of $\mathcal{G}$ between f and the center. This m-projection corresponds to a point in moment coordinates in the line

$$\eta_\lambda = \eta_0 + \lambda(\eta - \eta_0) \tag{54}$$

where the m-projection η^* corresponds to λ^* solving $D(g_0||g_\lambda) = \rho$ for $\lambda \in [0, 1]$. This equation may be solved numerically employing simple iterative methods provided the computations $\Lambda^{-1}(\eta)$ and $\varphi^*(\eta)$ are available as subroutines.

E-maximizer over an m-ball. Now consider the problem of determining the point in an m-ball "furthest" from a specified model f in the sense of e-maximization. This problem of determining the e-maximizer over an m-ball may be analyzed as follows. In the case that $f \notin \mathcal{G}$, consider growing the radius ρ of another m-ball about f until it first contains the given m-ball $\mathcal{G}$ when $\rho = d^*$. The m-sphere surfaces of the two m-balls then intersect at a single point η^* and have a common m-flat tangent plane at this point. Since the radial e-geodesics from η to η^* and from the center of the m-ball η_0 to η^* both contain η^* and are both $\mathcal{I}$-orthogonal to that common m-flat tangent plane, these are the same e-geodesic. So the e-maximizer over an m-ball is obtained by extrapolating the e-geodesic from f to the center of the m-ball g_0 beyond the center until it intersects the far surface of $\mathcal{G}$. A similar argument applies when $f \in \mathcal{G}$ provided f is not the center of $\mathcal{G}$ (in which case the e-maximizer is not defined). In this case, we grow the m-ball about f until it first contacts the surface of $\mathcal{G}$. The argument above then applies. The e-maximizer is again of the form (52) but now corresponds to λ^* solving $D(g_\lambda||g_0) = \rho$ for $\lambda < 0$.

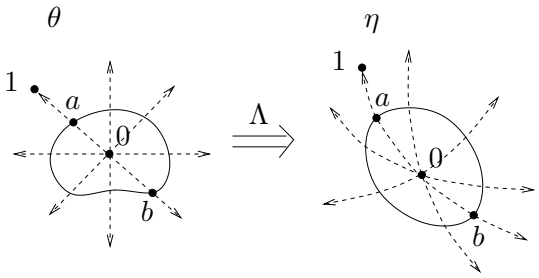

Figure 5: Illustration of e-minimization and e-maximization over an m-ball. The points labeled a and b are respectively the e-minimizer and e-maximizer of the point labeled 1 over the depicted m-ball. Note that all 3 points lie in the same radial e-geodesic which also contains the center of the ball labeled 0. The reverse picture applies for m-minimization/maximization over an e-ball.

M-maximizer over an e-ball. The dual problem of determining the m-maximizer over an e-ball is analyzed in an analogous manner as above. The m-maximizer is determined by extrapolating the m-geodesic from f to the center of the e-ball g_0 beyond the center until the far surface of the e-ball is reached. The m-maximizer is again of the form (54) but now corresponds to λ^* solving $D(g_0||g_\lambda) = \rho$ for $\lambda < 0$.

We have now developed the necessary background material to treat the min-max problem posed in the introduction.

4 KL Min-Max Problems

Now we return to the min-max modeling problem posed in the introduction. We will restate the min-max problem as follows. Given an m-ball $\mathcal{G} \subset \mathcal{F}$ of radius ρ about a given density model g_0, determine the

model in a specified e-flat submanifold $\mathcal{H} \subset \mathcal{F}$ which minimizes the worst-case KL-divergence $D(g||h)$ for $g \in \mathcal{G}$ and $h \in \mathcal{H}$. This corresponds to the min-max problem below.

$$\inf_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} D(g||h) \tag{55}$$

Conceptually, we are trying to determine the "best" model $h^* \in \mathcal{H}$ for an unknown model $g \in \mathcal{G}$ where all we know about g is that it is near a given reference model g_0 in the sense that $D(g||g_0) \leq \rho$. Hence, ρ characterizes the modeling uncertainty of the density estimate g_0. This may be considered as a generalization of the problem of m-projection to an e-flat submanifold so as to allow for uncertainty in the given model g_0. In fact, the above min-max design problem reduces to m-projection as ρ goes to zero.

Consider the inner supremum of the above max-min problem. This corresponds to performing e-maximization over the m-ball $\mathcal{G}$ from a given density h. As discussed previously, the optimal e-maximizer is determined by tracing the radial e-geodesic containing h through g_0 to the far side of the m-sphere surface of $\mathcal{G}$. The point of intersection of this e-geodesic with the far side of the m-sphere then gives the e-maximizer g^*. Denote this mapping by $\mu : \mathcal{H} \to \mathcal{G}$ so that $g^* = \mu(h)$. The outer infimum then chooses $h \in \mathcal{H}$ so as to minimize the KL-divergence $D(\mu(h)||h)$ so that h is as "near" as possible to its e-maximizer over $\mathcal{G}$.

Now let us consider the outer infimum of the min-max problem. The objective here is to minimize $D(\mu(h)||h)$ subject to $h \in \mathcal{H}$. Let θ and η denote the dually coupled exponential and moment coordinates of h. Also, let $\hat{\theta}$ and $\hat{\eta}$ denote the dually coupled coordinates of $\hat{h} \equiv \mu(h)$. Then consider the parameterization of $D(\hat{h}||h)$ in $(\hat{\eta}, \theta)$ below

$$D(\hat{\eta}||\theta) = \varphi^*(\hat{\eta}) + \varphi(\theta) - \hat{\eta} \cdot \theta. \tag{56}$$

Now take the gradient of this KL-divergence viewing $\hat{\eta}$ as a function of θ.

$$\begin{aligned} \frac{\partial D(\hat{\eta}(\theta)||\theta)}{\partial \theta} &= \frac{\partial \hat{\eta}}{\partial \theta} \cdot \frac{\varphi^*(\hat{\eta})}{\partial \eta} + \frac{\partial \varphi(\theta)}{\partial \theta} - \left(\frac{\partial \hat{\eta}}{\partial \theta} \cdot \theta + \hat{\eta} \right) \\ &= \frac{\partial \hat{\eta}}{\partial \theta} \cdot \hat{\theta} + \eta - \left(\frac{\partial \hat{\eta}}{\partial \theta} \cdot \theta + \hat{\eta} \right) \\ &= (\eta - \hat{\eta}) - \frac{\partial \hat{\eta}}{\partial \theta} \left(\theta - \hat{\theta} \right) \end{aligned} \tag{57}$$

In order to be minimize of $D(\hat{\eta}(\theta)||\theta)$ it is necessary that the above gradient is orthogonal to the exponential representation $\mathcal{H}_\theta$ of the e-flat submanifold. We shall consider this stationary condition further momentarily. But first, a digression into the topic of minimax theory lends insight as to how to proceed.

Minimax Theory. Consider the associated max-min problem obtained by switching the order of the optimizations.

$$\sup_{g \in \mathcal{G}} \inf_{h \in \mathcal{H}} D(g||h) \tag{58}$$

The inner infimum corresponds to the problem of m-projection of a given density $g \in \mathcal{G}$ to an e-flat submanifold $\mathcal{H}$. As discussed previously, a unique solution exists and is obtained by tracing the m-geodesic containing g which is $\mathcal{I}$-orthogonal to $\mathcal{H}$. This m-geodesic is given by the intersection of the m-fibration $\mathcal{I}$-transverse to $\mathcal{H}$ with the radial m-fibration through g. The point h^* at which this m-geodesic intersects $\mathcal{H}$ is then the m-projection which uniquely achieves the infimum. Let us denote the projection operation by the function $\pi : \mathcal{G} \to \mathcal{H}$ so that $h^* = \pi(g)$. The outer supremum then chooses $g \in \mathcal{G}$ so as to maximize $D(g||\pi(g))$ such that g is furthest from its m-projection as possible.

We now summarize some important results of minimax theory [Roc69] with respect to a function $K(u, v)$ over domain $U \times V$. First, the *minimax inequality* relates the min-max and max-min problems showing that the max-min value under-estimates the min-max value.

$$\sup_{v \in V} \inf_{u \in U} K(u, v) \leq \inf_{u \in U} \sup_{v \in V} K(u, v) \tag{59}$$

However, under many general circumstances it is known that in fact the *minimax equality* holds so that the min-max and max-min values are equal and are then referred to as the minimax or saddle value of $K(u, v)$.

$$\sup_{v \in V} \inf_{u \in U} K(u, v) = \inf_{u \in U} \sup_{v \in V} K(u, v) \tag{60}$$

The following notion is useful for characterizing when the minimax equality holds.

The pair $(\bar{u}, \bar{v}) \in U \times V$ is said to be a *saddle point* of the function $K(u, v)$ over $U \times V$ when both

1. $\bar{u}$ is a minimizer of $K(\cdot, \bar{v})$

$$\bar{u} \in \arg\min_{u \in U} K(u, \bar{v}) \tag{61}$$

2. $\bar{v}$ is a maximizer of $K(\bar{u}, \cdot)$

$$\bar{v} \in \arg\max_{v \in V} K(\bar{u}, v) \tag{62}$$

or, equivalently,

$$K(u,\bar{v}) \leq K(\bar{u},\bar{v}) \leq K(\bar{u},v) \;\; \forall u \in U, v \in V. \quad (63)$$

A well known minimax theorem then states that a point $(\bar{u},\bar{v}) \in U \times V$ is a saddle point if and only the following three conditions hold.

1. the minimax equality holds,
2. $\bar{u}$ solves the min-max problem

$$\bar{u} \in \arg\min_{u \in U} \sup_{v \in V} K(u,v) \quad (64)$$

3. $\bar{v}$ solves the max-min problem

$$\bar{v} \in \arg\max_{v \in V} \inf_{u \in U} K(u,v) \quad (65)$$

So, if we can establish the existence of a saddle point then we know that the minimax equality holds and that determining such a saddle point simultaneously solves both the min-max and max-min problems.

Now, let us consider these ideas in the context of the KL min-max design problem. First, consider the geometric interpretation of a saddle point $(\bar{g},\bar{h})$ of the KL-divergence $D(g||h)$ over the sets $g \in \mathcal{G}$ and $h \in \mathcal{H}$. From the definition of a saddle point we have the necessary and sufficient conditions that (i) $\bar{g}$ is the e-maximizer of $\bar{h}$ over $\mathcal{G}$

$$\bar{g} = \mu(\bar{h}) \quad (66)$$

S and that (ii) $\bar{h}$ is the m-projection of $\bar{g}$ to $\mathcal{H}$

$$\bar{h} = \pi(\bar{g}). \quad (67)$$

This is equivalent to the condition that the model $\bar{h} \in \mathcal{H}$ is a fixed-point of the composition of these two operations which we denote by the mapping $\psi : \mathcal{H} \to \mathcal{H}$ defined as $\psi(h) = \pi(\mu(h))$.

$$\bar{h} = \psi(\bar{h}) \quad (68)$$

So "projecting" $\bar{h}$ to $\partial\mathcal{G}^+$ (the "upper" hemisphere of $\mathcal{G}$ defined by $\mu(\mathcal{H})$) along the e-geodesics through g_0 and then m-projecting back to $\mathcal{H}$ recovers $\bar{h}$. Geometrically, $\bar{h}$ is a fixed point of ψ if and only if the radial e-geodesic of g_0 containing $\bar{h}$ and the m-geodesic through $\bar{h}$ $\mathcal{I}$-transverse to $\mathcal{H}$ intersect at a point $\bar{g} \in \partial\mathcal{G}^+$. We would like to demonstrate both the existence and the uniqueness of such a fixed point. Determining this fixed point then solves the min-max model selection problem.

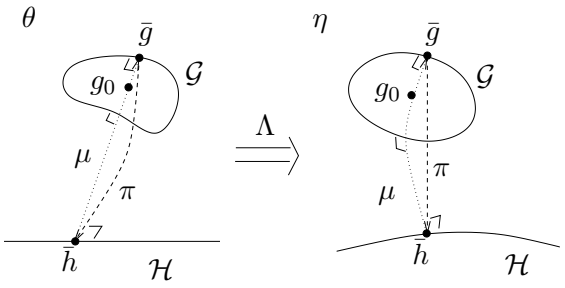

Figure 6: Illustration of saddle-point solution $(\bar{h},\bar{g})$ of minimax problem where $\mathcal{G}$ is an m-ball and $\mathcal{H}$ is an e-flat submanifold. The m-geodesic (dashed line) in moment coordinates is perpendicular to the e-flat $\mathcal{H}$ in exponential coordinates. The e-geodesic (dotted line) in exponential coordinates in perpendicular to the m-sphere surface of $\mathcal{G}$ at both points of intersection. Note that the e-geodesic contains the center of the m-ball g_0.

With these ideas in mind, let us now return to the necessary condition for h being a solution of the min-max problem. If $\theta \in \mathcal{H}_\theta$ is a minimum of $\rho(\theta) = D(\hat{\eta}(\theta)||\theta)$ then the gradient is perpendicular to $\mathcal{H}_\theta$ at θ. This condition is expressed as

$$\forall \theta' \in \mathcal{H}_\theta : \left\{ (\hat{\eta} - \eta) - \frac{\partial \hat{\eta}}{\partial \theta} \cdot (\hat{\theta} - \theta) \right\} \cdot (\theta' - \theta) = 0 \quad (69)$$

Based upon minimax theory, we then entertain the conjecture that the solution of the min-max problem corresponds to a saddle point. This is the case if and only if θ is the m-projection of $\hat{\theta}$ to $\mathcal{H}$. A necessary and sufficient condition for θ being the m-projection of $\hat{\theta}$ is that the straight line through η and $\hat{\eta}$ in moment coordinates is perpendicular to $\mathcal{H}_\theta$. This m-projection condition is expressed as

$$\forall \theta' \in \mathcal{H}_\theta : \; (\hat{\eta} - \eta) \cdot (\theta' - \theta) = 0. \quad (70)$$

Comparing this to the min-max condition shows that the earlier necessary min-max condition implies the m-projection condition provided the extra term vanishes.

$$\frac{\partial \hat{\eta}}{\partial \theta} \cdot (\hat{\theta} - \theta) = 0 \quad (71)$$

In fact, we now provide a geometric argument showing that this term is actually zero for all θ (not just at the minimum). Consider how the point $\hat{\eta}$ in moment coordinates may be determined geometrically given the point θ is exponential coordinates (refer to Figure 7). Trace a straight line from θ through the center

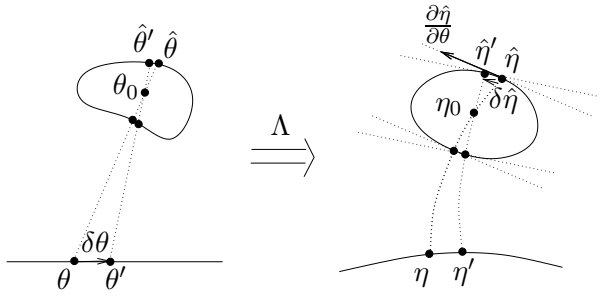

Figure 7: Illustration (in one parameter) that displacement $\partial\hat{\eta}$ due to vanishingly small perturbation $\partial\theta$ is tangent to surface of m-ball $\mathcal{G}$. Note that $\hat{\theta}$ is determined by tracing straight line from θ through θ_0. This radial e-geodesic is $\mathcal{I}$-orthogonal to the m-sphere so that $\hat{\eta}$ is the intersection point of "upper" supporting hyperplane with normal vector $(\theta - \theta_0)$. Hence, $\delta\hat{\eta}/\delta\theta \to \partial\hat{\eta}/\partial\theta$ as $\delta\theta \to 0$ which lies in the m-flat tangent plane intersecting $\hat{\eta}$.

of the m-ball θ_0 in exponential coordinates. This is the exponential representation of the e-geodesic connecting h and g_0. This geodesic intersects the surface of $\mathcal{G}$ at exactly two points. The point nearest to θ (between θ and θ_0) gives the e-projection of h to $\mathcal{G}$ while the further point gives the e-maximizer of h over $\mathcal{G}$.

Now consider this picture in moment coordinates. The intersection points in moment coordinates may be determined geometrically as follows. The e-geodesic is a radial e-geodesic of g_0 and hence $\mathcal{I}$-orthogonal to the m-sphere surface of $\mathcal{G}$. This means that the m-flat tangent planes at these two intersection points are normal to the vector $(\theta - \theta_0)$. Hence, these intersection points may be determined by "sandwiching" the moment representation of $\mathcal{G}$ between "upper" and "lower" supporting hyperplanes normal to $(\theta - \theta_0)$ where the intersection point of the "upper" hyperplane gives $\hat{\eta}$.

Finally, consider how a small perturbation $\delta\theta$ affects this construction of $\hat{\eta}$. This produces a small rotation of the normal vector of the upper supporting hyperplane which then contacts the m-ball at a nearby point $\hat{\eta} + \delta\hat{\eta}$ in the surface of the m-ball. Because m-spheres are smooth surfaces, this perturbed point $\hat{\eta} + \delta\hat{\eta}$ approaches $\hat{\eta}$ as $\delta\theta$ vanishes such that, for vanishingly small perturbations $\delta\theta$, the displacement $\delta\hat{\eta}$ is tangent to the surface of the m-ball. Then, by considering positive perturbations of single coordinates θ_i we conclude that each row of $\partial\hat{\eta}/\partial\theta$ lies in the subspace parallel to this tangental hyperplane having normal vector $(\theta - \theta_0)$. The parallel vector $(\theta - \hat{\theta})$ then lies in the null space of this matrix so that (77) holds.

Hence, the gradient derived earlier actually simplifies to

$$\frac{\partial D(\hat{\eta}(\theta)||\theta)}{\partial\theta} = \eta - \hat{\eta}. \tag{72}$$

which is precisely the result when $\hat{\eta}$ is viewed as fixed (independent of θ). Consequently, if $\bar{h}$ is a minimizer of the min-max problem then it is the m-projection $\pi(\bar{g})$ where $\bar{g} = \mu(\bar{h})$ so that $(\bar{h}, \bar{g})$ is a saddle point of $D(g||h)$ over $g \in \mathcal{G}$ and $h \in \mathcal{H}$.

Employing a similar strategy, we may also show that any solution of the max-min problem is a saddle point solution. Consider the necessary optimality conditions for maximizing $D(g||\pi(g))$ over $g \in \mathcal{G}$. We now let θ and η denote the dually coupled exponential and moment coordinates of g and abbreviate it's m-projection to $\mathcal{H}$ as $\hat{g} \equiv \pi(g)$ with coordinates denoted by $\hat{\theta}$ and $\hat{\eta}$. Consider $D(g||\pi(g))$ as parameterized by $(\eta, \hat{\theta})$ so that we are interested in minimizing the function $\rho^*(\eta) \equiv D(\eta||\hat{\theta}(\eta)) = \varphi^*(\eta) + \varphi(\hat{\theta}) - \eta \cdot \hat{\theta}$ over the m-ball $\mathcal{B}_\eta(\rho; \eta_0)$. The gradient is evaluated below where $\hat{\theta}$ is viewed as a function of η.

$$\frac{\partial D(\eta||\hat{\theta}(\eta))}{\partial\eta} = (\theta - \hat{\theta}) - \frac{\partial\hat{\theta}}{\partial\eta} \cdot (\eta - \hat{\eta}) \tag{73}$$

However, the operation of m-projection traces the m-geodesic containing η $\mathcal{I}$-orthogonal to $\mathcal{H}$. This m-geodesic is a straight line in moment coordinates which is parallel to the vector $(\eta - \hat{\eta})$ and hence normal to $\mathcal{H}_\theta$. Hence, the rows of the Jacobian $\partial\hat{\theta}/\partial\eta$ lie in the orthogonal subspace of $(\eta - \eta_0)$ so that the second term above is zero and the above gradient formula reduces to

$$\frac{\partial D(\eta||\hat{\theta}(\eta))}{\partial\eta} = \theta - \hat{\theta} \tag{74}$$

which is the same as if $\hat{\theta}$ were viewed as fixed (independent of η). Consequently, the necessary optimality conditions for $\bar{g}$ being a solution of the max-min problem imply that $\bar{g}$ is the e-maximizer $\mu(\bar{h})$ over $\mathcal{G}$ of it's m-projection $\bar{h} \equiv \pi(\bar{g})$ to $\mathcal{H}$ so that $(\bar{h}, \bar{g})$ is again a saddle point.

Combining the earlier minimax theorem with these information geometric results, it is apparent that there can exist at most one such saddle point. To prove this, let $(\bar{h}_1, \bar{g}_2)$ and $(\bar{h}_2, \bar{g}_2)$ both be saddle points. Then, appealing the the minimax theorem,

the point $(\bar{h}_1, \bar{g}_2)$ must also be a saddle point. Appealing to information geometry, we then have that both $\bar{h}_1 = \pi(\bar{g}_2) = \bar{h}_2$ and likewise that $\bar{g}_2 = \mu(\bar{h}_1) = \bar{g}_1$ so that these cannot be distinct saddle points. Hence, if $\bar{h}$ is a feasible solution of the min-max problem then it is the unique solution and $\mu(\bar{h})$ is the unique solution of the max-min problem. Likewise, if $\bar{g}$ is a feasible solution of max-min problem, then it is unique and $\pi(\bar{g})$ is the unique solution of the min-max problem. In either case, there exists a unique saddle point.

Yet, we still have not demonstrated that either the min-max or max-min problems has a feasible solution. To summarize, here are the two possible scenarios consistent with earlier claims.

(1) There exists a unique saddle-point $(\bar{h}, \bar{g})$ in which case $\bar{h}$ is the unique solution of the min-max problem and $\bar{g}$ is the unique solution of the max-min problem.

(2) There does not exist a saddle-point in which case neither the min-max nor the max-min problems have feasible solutions (neither the infimum of the min-max problem nor the supremum of the max-min problem are achieved).

However, given that the m-ball $\mathcal{G}$ is m-compact and that the function $f(\eta) = D(g||\pi(g))$ (where η are the moment coordinates of g) is a continuous function due to the continuity of the KL-divergence, then by the Weierstrass theorem [Rud76] the supremum of the max-min problem is achieved by some $\eta \in \mathcal{G}_\eta$. Hence, the max-min problem must have a solution ruling out the scenario (2) above such that (1) holds. This completes the analysis.

5 Conclusion

The main results of this the paper may be summarized as follows. Given an m-ball $\mathcal{G}$ and an e-flat submanifold $\mathcal{H}$ of a regular exponential family of models $\mathcal{F}$, there exists exactly one saddle point $(\bar{g}, \bar{h})$ of the KL-divergence $D(g||h)$ over the domain $g \in \mathcal{G}$ and $h \in \mathcal{H}$. This asserts that $\bar{g}$ is the e-maximizer of $\bar{h}$ over $\mathcal{G}$

$$\bar{g} = \arg\max_{g \in \mathcal{G}} D(g||\bar{h}) \tag{75}$$

while $\bar{g}$ is the m-projection of $\bar{h}$ to $\mathcal{H}$.

$$\bar{h} = \arg\min_{h \in \mathcal{H}} D(\bar{g}||h) \tag{76}$$

Furthermore, $\bar{h}$ is then the unique solution of the min-max problem

$$\bar{h} = \arg\min_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} D(g||h) \tag{77}$$

while $\bar{g}$ is the unique solution of the max-min problem

$$\bar{g} = \arg\max_{g \in \mathcal{G}} \inf_{h \in \mathcal{H}} D(g||h) \tag{78}$$

and the two problem have the same value given by $D(\bar{g}||\bar{h})$.

In regards to the min-max problem, there exists a unique solution h^* and the necessary and sufficient optimality condition is the fixed-point condition $h^* = \pi(\mu(h^*))$ so that the m-projection of the e-maximizer of h^* recovers h^*.

The above analysis may also be duplicated with respect to the "dual" min-max problem obtained by reversing the sense of the KL-divergence

$$\inf_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} D(h||g) \tag{79}$$

but where $\mathcal{G}$ is now an e-ball (instead of an m-ball) and $\mathcal{H}$ is now m-flat (instead of e-flat). The conclusion is again that there is no duality gap and that there exists a unique saddle point $(\bar{h}, \bar{g})$ so that $\bar{h}$ uniquely solves the above min-max problem while $\bar{g}$ uniquely solves the corresponding max-min problem. But now the geometrical interpretation is that $\bar{g}$ is the m-maximizer (instead of e-maximizer) of $\bar{h}$ over $\mathcal{G}$ while $\bar{h}$ is the e-projection (instead of m-projection) of $\bar{g}$ to $\mathcal{H}$. We believe the earlier formulation is more appropriate for performing model selection/reduction in the context of exponential family models. However, this latter dual version of the min-max problem may prove more appropriate and/or tractable in the case of mixture families.

An iterative algorithm for solving the min-max problem has been developed and implemented in the specific context of Gauss-Markov processes. This algorithm is briefly discussed in the appendix. Time does not permit further exploration of the utility of this method. So, in closing, I suggest that these methods may prove to have useful applications in the areas of robust parameter estimation, order estimation, model reduction and recursive methods of approximate inference.

A An Iterative Solution Technique.

A simple iterative algorithm has been developed for determining the min-max model in the context

of Gauss-Markov processes. The restriction to e-flat manifolds very naturally corresponds to imposing Markov structure upon an exponential family. For instance, a natural problem to consider is to find the best "fit" to data among the family of (i) fully-factorized models (assuming independence), (ii) singly-connected models (assuming simple Markov structure such as a Markov chain or tree), or (iii) more general sparse albeit "loopy" Markov structures such as 2d grids for image processing. For exponential family models, this corresponds to forcing certain exponential parameters to zero so that no sufficient statistics couple conditionally independent random variables. The data set is specified by its maximum-likelihood moment parameters given by the "empirical" moments. For the "full" Gaussian density (not presuming any Markov structure) this is the sample mean and covariance. The problem of selecting the best model among a family of models presuming some sparse Markov structure then reduces to m-projection if the maximum-likelihood criterion is employed. As an alternative approach, we might specify an m-ball about these empirical moment, say with radius $\rho = 1/N$ in keeping with the AIC principle. The min-max criterion may then be employed to select the "best" Markov model among that subfamily.

Without getting into specifics, the structure of the general algorithm is outlined below. For clarity, we denote densities below but the actual algorithm stores either exponential or moment coordinates (or both) as convenient.

1. Initialization: Set $\bar{h} = \pi(g_0)$, the m-projection of the center of the m-ball g_0 to the e-flat manifold $\mathcal{H}$.

2. Evaluate $\hat{h} = \pi(\mu(\bar{h}))$ the m-projection of the e-maximizer of $\bar{h}$.

3. Line minimization. Minimize $D(\mu(h)||h)$ over the e-geodesic connecting $\bar{h}$ to $\hat{h}$. Update $\bar{h}$ to the minimizer over this embedded e-geodesic of the e-flat manifold.

4. Iterate 2-3 until $\bar{h} = \pi(\mu(\bar{h}))$ which is then the optimal solution.

I've found that, for the Gauss-Markov problems I've considered (involving on the order of a dozen random variables), this algorithm converges to machine precision after 6-12 line minimizations for $\rho < 1$ but may require several dozen line minimizations for $1 < \rho < 10$. It should also be remarked that the line-minimization is itself a one-parameter min-max problem, but has simple convex structure such that it may be solved by determining the unique zero of $d \cdot (\pi(\mu(\theta)) - \theta)$ where $d = \hat{\theta} - \bar{\theta}$ is the search direction in exponential coordinates and where θ, $\hat{\theta}$, and $\bar{\theta}$ are respectively the exponential coordinates of h, $\hat{h}$ and $\bar{h}$. Once the minimum is bounded (with the endpoints having opposite signs of the above inner product), simple bisection/interpolation methods converge quickly typically requiring evaluation of a dozen or so trial points.

The required m-projections may be evaluated directly when the presumed Markov structure is singly-connected, but more generally requires iterative methods such as the iterative proportional fitting procedure (IPF) discussed in Csiszár. The e-maximization requires solution of the one-parameter equation $D(g_\lambda||g_0) = \rho$ for $\lambda < 0$ as discussed in the main text. This is also solved by bounding the solution and employing bisection/interpolation methods.

It is also interesting to note that comparison of the min-max and maximum-likelihood models shows that the min-max models always have higher entropy then the corresponding ML models where the entropy-gain (information-loss) increases monotonically with ρ. This suggests that the min-max approach is similar to ML but with a bias towards high-entropy models. In scenarios where the ML estimator itself is biased, this may lead to some improvement under the min-max approach.

The reader is invited to examine/execute the matlab code in my directory

`/Public/6.291-proj/matlab`

located under

`/afs/athena.mit.edu/user/j/a/jasonj`

A sample experiment is invoked by issuing the command `run_minimax` to the matlab command-line interpreter.

References

[Ama82] S. Amari. Differential geometry of curved exponential families – curvature and information loss. *Annals of Statistics*, 10(2):357–385, June 1982.

[Ama01] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, July 2001.

[BN78] O. Barndorff-Nielsen. *Information and Exponential Families*. Wiley series in probability and mathematical statistics. John Wiley, 1978.

[Čen72] N.N. Čencov. *Statistical Decision Rules and Optimal Decisions*. Nauka, Moscow, 1972.

[Csi75] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, February 1975.

[CT91] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[Efr78] B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

[Roc69] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1969.

[Rud76] W. Rudin. *Principles of Mathematical Analysis, 3rd ed.* McGraw Hill, 1976.