Information-driven sensor planning: Navigating a statistical manifold

(Invited Paper)

Douglas Cochran
Arizona State University
Tempe, AZ 85287-5706, USA

Alfred O. Hero III
University of Michigan
Ann Arbor, MI 48109-2122, USA

***Abstract*—Many adaptive sensing and sensor management strategies seek to determine a sequence of sensor actions that successively optimizes an objective function. Frequently the goal is to adjust a sensor to best estimate a partially observed state variable, for example, the objective function may be the final mean-squared state estimation error. Information-driven sensor planning strategies adopt an objective function that measures the accumulation of information as defined by a suitable metric, such as Fisher information, Bhattacharyya affinity, or Kullback-Leibler divergence. These information measures are defined on the space of probability distributions of data acquired by the sensor, and there is a distribution in this space corresponding to each sensor configuration. Hence, sensor planning can be posed as a problem of optimally navigating over a statistical manifold of probability distributions. This information-geometric perspective presents new insights into adaptive sensing and sensor management.**

***Index Terms*—Adaptive sensing, Sensor management, Information geometry, Hellinger distance, Multinomial class of distributions**

I. Introduction

In the 1940s, C. R. Rao [1] and H. Jeffreys [2] recognized that Fisher information induces a Riemannian metric structure on a smooth manifold of parameters that index a collection of probability distributions. This observation formed the cornerstone of the area now known as information geometry, which owes much of its development since the 1980s to S. Amari [3] and his collaborators. This information geometric perspective has been applied to problems in signal processing such as blind source separation [4], Doppler imaging [5], and dimensionality reduction [6].

Recently, the possibility of exploiting the geometric structure of Riemannian parameter manifolds in defining schemes for sensor management has been broached [7], [8]. In these papers the authors developed a correspondence between navigating over the sensor configuration space and navigating over a space of Riemannian information geometries induced by the families of measurement distributions that could be generated by the sensor. In this framework, each sensor configuration induces a different Riemannian manifold of measurement distributions and thus planning reduces to selecting among induced Riemannian metrics; which themselves form a Riemannian manifold. The present paper considers a somewhat different framework than [7], [8] in which the configuration of a sensing system may be adjusted in a manner that affects the distribution of the data it collects, but only within a fixed family of distributions. Such a framework is relevant to sensors that always output the same parametric family of distributions, e.g., a Gaussian, Wishart or multinomial distribution, albeit with different parameter values, regardless of their configuration.

More specifically, in this paper we assume that the distribution of the collected data belongs to a parametric family and the parameter value is determined by the configuration of the sensor suite. It will be assumed that the parameter space is a smooth manifold (possibly with boundary), that the mapping between the sensor configuration and the parameter is smooth, and that the sensor configuration can be controlled as a function of time in a smooth fashion. In practical situations, the assumption of smooth motion corresponds to constraining the sensor suite to be adjusted gradually (e.g., as in the position of a mobile sensor) as opposed to instantaneously (e.g., transmitted waveform in a waveform-agile radar).

The general objective is to adjust the sensor configuration according to some pre-established goal defined in terms of the distribution parameter. For example, it may be desired to "tune" the sensor system to a configuration in which the collected data follow a particular distribution. Or, as arises in some information-based sensor management strategies [9], [10], the goal might be to adjust the sensor system within a specified time to a configuration where the distribution of the collected data will be as distinct as possible, in a suitable informational sense, from some known distribution. This situation is depicted in Figure 1.

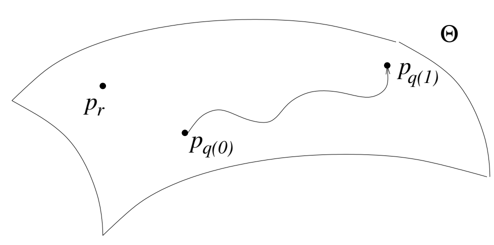

Fig. 1. An objective of navigation on the parameter manifold of interest in information-based sensor management is, starting at $q(0)$, to reach $q(1)$ such that the density $p_{q(1)}$ is as distant as possible from some reference density p_r. The distance of primary interest here is Hellinger distance d_H, which is monotonically related to information distance d_I in the multinomial family of distributions.

The mapping between the sensor configuration and the distribution parameter of the data collected is not assumed to be known *a priori*. Rather, sufficient data are collected in each configuration as the sensor is navigated to enable high-fidelity estimation of the parameter from data. This enables learning of differential relationships between sensor configuration and parameter value.

A particular case of interest, which we emphasize in this paper, is when the sensing configurations generate measurements whose distributions lie in the multinomial family. This model is relevant to the many sensors that generate histograms of data, e.g., vision sensors that form histograms of visual features over a SIFT or HOG codebook, or chemical sensors that count the number of trace element particles over a number of wavelengths and/or energy levels [11]. Remarkably, in this multinomial case the statistical manifold simplifies to a simple Hellinger sphere and, given an initial distribution and a desired final distribution, the information-optimal sensor planning trajectories are geodesic great circle paths on the Hellinger sphere. Furthermore, under the multinomial model the accumulated Fisher information along the geodesic is equivalent to the Hellinger distance between the distributions. It is notable that these special properties are specific to the multinomial distribution and do not hold for other commonly encountered distributions in sensor signal processing.

The outline of the paper is as follows. The following section presents a more precise formulation of the class of problems just introduced. Subsequently, attention is directed in Section III to the multinomial family, an exponential family of probability distributions. Some elaboration upon known results that show monotonic relationships between global measures of distance, including Hellinger distance and Bhattacharyya affinity, and information distance, which is defined in terms of a local quantity (Fisher information). Geometrical interpretation of one of these monotonically equivalent distance measures offers appealing insight about the nature of optimal trajectories in this setting. Section IV describes a notional application and explains how control of the sensor system configuration can be informed by the geometry of the parameter manifold in this example. The paper concludes with some discussion of potential directions for additional research in this vein.

II. Mathematical formulation

Denote by Θ a smooth n-dimensional manifold and let $\mathcal{F}$ be a family of probability densities parameterized by Θ; i.e., $\mathcal{F} = \{p_q \,|\, q \in \Theta\}$ with each p_q a probability density function. In what follows, the Hellinger distance [12]

$$d_H(p_q, p_r) = \left[\int \left(\sqrt{p_q(x)} - \sqrt{p_r(x)}\right)^2 \, dx\right]^{1/2} \quad (1)$$
$$= \left[2 - 2\int \sqrt{p_q(x)p_r(x)} \, dx\right]^{1/2}$$

is adopted as a global metric on Θ. Because Hellinger distance is an f-divergence [13], the Riemannian metric on Θ obtained locally from the second derivative of d_H is the Fisher information. This is particularly appealing from a discrimination perspective, since the Kullback-Leibler (KL) divergence $d_{\mathrm{KL}}(p_q||p_r)$, also known as relative entropy, is a natural measure of the value of data drawn from p_q for discriminating between parameter values q and r in a hypothesis test. KL divergence is also an f-divergence, so it also leads to Fisher information as a Riemannian metric on Θ. For the purposes of what follows, it is more convenient to use Hellinger distance than KL divergence, even though it is KL divergence that provides a more intuitive interpretation of the Riemannian structure on Θ imparted by Fisher information in terms of discrimination.

The configuration of the sensor system will also be taken to be parameterized by a smooth manifold Γ of dimension $m \leq n$, and the mapping $\Phi : \Gamma \to \Theta$ taking a configuration $s \in \Gamma$ to the parameter $q = \Phi(s)$ of the distribution of data collected in that configuration will be taken to be smooth and non-degenerate. Thus a smooth trajectory $s : [0, 1] \to \Gamma$ of sensor configurations corresponds to a smooth curve $q = \Phi s$ of parameters. Further, if Ψ is a diffeomorphism between Θ and $\widetilde{\Theta}$, then $\mathcal{F}$ is equivalently parameterized by $\widetilde{\Theta}$ with a curves $q \in \Theta$ and $\tilde{q} \in \widetilde{\Theta}$ serving as interchangeable representations of the sequence of densities corresponding to the sequence $s(t)$ of sensor configurations in Γ, as depicted in Figure 2.

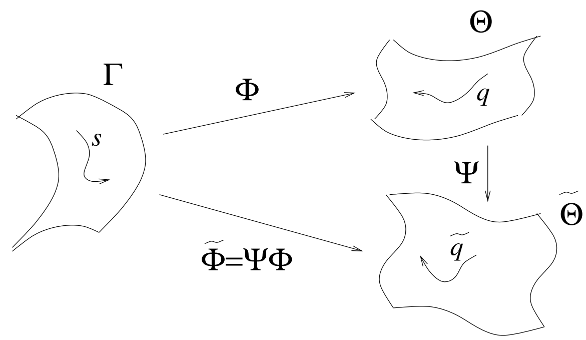

Fig. 2. Relationship between trajectories in sensor configuration space Γ and equivalent parameter spaces Θ and $\widetilde{\Theta}$ for $\mathcal{F}$.

III. Analysis in the multinomial family

As noted in [14] and developed further here, the information geometry of the multinomial family $\mathcal{F}_M$ engenders properties that are particularly desirable for development of the kinds of navigation strategies sought here. First, Hellinger distance in this family is monotonically related to information distance, which is defined as an integral of a local quantity (Fisher information). Moreover, the geodesic curves defining information distance in the standard parameter space Θ correspond to great circles in an alternative spherical parameter space $\widetilde{\Theta}$, connecting navigation problems in $\mathcal{F}_M$ to classical navigation on the Earth's surface.

For the multinomial family K outcomes and $N = 1$ trial, densities are parameterized by a vector parameter $q = (q_1, \ldots, q_K)$ where q_k is the probability of outcome k. Each q_k is non-negative and $q_1 + \cdots + q_K = 1$, so the parameter space Θ is the face of the unit simplex in the non-negative orthant in $\mathbb{R}^K$. An equivalent parameterization is by vectors

$$\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_K) = (\sqrt{q_1}, \ldots, \sqrt{q_K}),$$

which are unit vectors in $\mathbb{R}^K$ with non-negative elements; i.e., $\tilde{q} \in \widetilde{\Theta}$, the portion of the unit sphere intersecting the non-negative orthant in $\mathbb{R}^K$. Both Θ and $\widetilde{\Theta}$ are smooth $(K-1)$-dimensional manifolds with boundary, and they are diffeomorphic under the non-degenerate map Ψ taking q to $\tilde{q}$. If $N > 1$ (and is known), the multinomial family is still parameterized by $q \in \Theta$ and equivalently by $\tilde{q} \in \widetilde{\Theta}$, so much of the following discussion applies without significant change.

With the assumptions noted in Section II, each sensor configuration $s(t)$ on a smooth curve $s : [0,1] \to \Gamma$ corresponds to a parameter $q(t)$ on a smooth curve $q = \Phi s$ in Θ or equivalently to a parameter $\tilde{q}(t) = \Psi q(t)$ on a smooth curve $\tilde{q}$ in $\widetilde{\Theta}$. The probability density function corresponding to the parameter $q \in \Theta$ will be denoted by p_q. With this notation, the Hellinger distance between $p_{q(0)}$ and $p_{q(t)}$ is given by (1) as

$$\begin{aligned} d_H^2(p_{q(0)}, p_{q(t)}) &= 2 - 2\int \sqrt{p_{q(0)}(x) p_{q(t)}(x)}\, dx \qquad (2) \\ &= 2 - 2\,\langle \tilde{q}(0), \tilde{q}(t) \rangle . \end{aligned}$$

In this expression, the integral is simply the sum over the discrete outcome set $\{1, \ldots, K\}$ and $\langle \cdot, \cdot \rangle$ is thus the standard inner product in $\mathbb{R}^K$. Thus, in this family of distributions, $d_H(p_q, p_r)$ depends only on the angle between the unit vectors $\tilde{q}$ and $\tilde{r}$ as defined by $\langle \tilde{q}, \tilde{r} \rangle$. This inner product is known as the Bhattacharyya affinity between p_q and p_r.

The information distance $d_I(p_{q(0)}, p_{q(t)})$ between $p_{q(0)}$ and $p_{q(t)}$ is defined as the minimum value, over all curves q sharing the endpoint values $q(0)$ and $q(t)$, of

$$\int_0^t \sqrt{I(u)}\, du.$$

Here, $I(u)$ is the Fisher information at $q(u) \in \Theta$. The log likelihood function is given by $\log q_k(u)$, $k \in \{1, \ldots, K\}$. So the score function is

$$\frac{1}{q_k(u)} \frac{d}{du} q_k(u), \quad k \in \{1, \ldots, K\}$$

and the Fisher information is

$$\begin{aligned} I(u) &= \sum_{k=1}^K \left(\frac{1}{q_k(u)} \frac{d}{du}\, q_k(u) \right)^2 q_k(u) \qquad (3) \\ &= \sum_{k=1}^K \left(\frac{d}{du}\, q_k(u) \right)^2 \left(\frac{1}{q_k(u)} \right). \end{aligned}$$

Note that

$$\begin{aligned} \left| \frac{d}{du} \tilde{q}(u) \right|^2 &= \sum_{k=1}^K \left(\frac{d}{du} \sqrt{q_k(u)} \right)^2 \\ &= \frac{1}{4} \sum_{k=1}^K \frac{1}{q_k(u)} \left(\frac{d}{du}\, q_k(u) \right)^2 = \frac{1}{4}\, I(u). \end{aligned}$$

So the integral of $\sqrt{I(u)}$ along any smooth curve q in Θ is twice the length of the corresponding curve $\tilde{q}$ in $\widetilde{\Theta}$. In particular, the information distance between $p_{q(0)}$ and $p_{q(t)}$ is twice the great circle (geodesic) distance between $\tilde{q}(\tau)$ and $\tilde{q}(t)$ in the spherical segment $\widetilde{\Theta}$.

Now, since $\widetilde{\Theta}$ contains only points on the unit sphere that lie in the non-negative orthant, the great circle distance between two points $\tilde{q}(0)$ and $\tilde{q}(t)$ in $\widetilde{\Theta}$ is the angle subtended by the unit vectors $\tilde{q}(\tau)$ and $\tilde{q}(t)$; i.e., $d_I(p_{q(0)}, p_{q(t)}) = 2 \arccos \langle \tilde{q}(0), \tilde{q}(t) \rangle$. But, from (2),

$$\langle \tilde{q}(0), \tilde{q}(t) \rangle = 1 - \frac{1}{2}\, d_H^2(p_{q(0)}, p_{q(t)}).$$

Thus,

$$d_I(p_{q(0)}, p_{q(t)}) = 2 \arccos \left(1 - \frac{1}{2}\, d_H^2(p_{q(0)}, p_{q(t)}) \right)$$

and

$$d_H(p_{q(0)}, p_{q(t)}) = 2 \sin \left(\frac{d_I(p_{q(0)}, p_{q(t)})}{4} \right). \qquad (4)$$

The preceding development shows that the multinomial family offers attractive properties for the purposes of information-geometric sensor management. Not only do optimal trajectories for Hellinger and information distance metrics coincide, but they have particularly appealing interpretations as great circle segments in $\widetilde{\Theta}$. Methods for following great circles have been well studied in connection with classical navigation, which may provide a source of insight about sensor management schemes in this setting.

IV. Notional Application

The distribution of particle types 1, 2, and 3 impinging on a detector depends on the fraction of the aperture covered by filtering materials A and B. The objective is to tune the instrument to admit a desired mix of particle types; i.e., the desired final density is $p_{q(1)}$ with $q(1) = (q_1(1), q_2(1), q_3(1))$ pre-established. The distribution of particle types is controlled via a two-dimensional control vector $\varphi = (\varphi_1, \varphi_2)$, which determines the fraction of the aperture covered with materials A and B. The initial value of φ is $\varphi(0) = (\varphi_1(0), \varphi_2(0))$ and the values of the components of φ can be adjusted independently in a smooth fashion. The relationship between the sensor configuration $\varphi(t)$ and the parameter of the corresponding particle-type distribution is fixed but unknown.

By collecting data in the initial configuration, $q(0)$ is estimated. From the analysis in Section III, the trajectory along which the Hellinger distance between the initial and final distributions will be reduced most quickly by a small change in configuration space follows the great circle in the two-sphere $\widetilde{\Theta}$ joining $\tilde{q}(0)$ and $\tilde{q}(1)$.

At any time t, the local relationship between the control parameters $(\varphi_1(t), \varphi_2(t))$ and the parameter $q(t)$ of the corresponding distribution can be estimated as follows:

1) With ϵ small, estimate $\frac{\partial q_i}{\partial \varphi_1}(t)$ for $i = 1, 2, 3$ by setting the sensor configuration to $(\varphi_1(t) + \epsilon, \varphi_2(t))$ and

observing the change in the value of q_i estimated from data collected in the new configuration.

2) Similarly, estimate $\frac{\partial q_i}{\partial \varphi_2}(t)$ by setting the sensor configuration to $(\varphi_1(t), \varphi_2(t) + \epsilon)$.
3) Using the local relationship between $\varphi(t)$ and $\tilde{q}(t)$ estimated in this way, set $\varphi(t+\Delta t)$ so that $\tilde{q}(t+\Delta t)$ falls (approximately) on the desired great circle trajectory in $\widetilde{\Theta}$.
4) Collect data to estimate $q(t + \Delta t)$ and repeat steps 1-3 seeking to follow the great circle trajectory from $\tilde{q}(t + \Delta t)$ to $\tilde{q}(1)$.

This crude algorithm illustrates in principle how a practical system might learn enough about the local relationship between motion on the sensor configuration manifold Γ and the motion it induces in the parameter manifold Θ to approximately navigate a desired trajectory in Θ. More sophisticated learning schemes, possibly informed by some *a priori* knowledge about the map Φ beyond its differentiability, would likely be possible in an actual application of this kind.

V. Discussion

The preceding sections have described how certain sensor management scenarios can manifest in terms of the metric structure and geometry of a related statistical manifold. In the multinomial family, the particular form (3) of the Fisher information at each point along a curve in the parameter manifold Θ motivates its interpretation as differential arc length in the spherical segment $\widetilde{\Theta}$, immediately implying that the geodesics are great circle arcs in this parameter manifold. This observation makes clear the monotonic relationship between Hellinger distance and information distance in the multinomial family. To the authors knowledge, such simple relationships do not hold for any other important families of distributions. For example, for statistical manifolds induced by the multivariate Gaussian, student-t, Dirichlet, and Wishart distributions, geodesics are not great circle arcs. Further investigation into such issues, including characterization of situations in which d_H and d_I are not consistent, appears warranted.

Information distance is explicitly defined in terms of accumulation of Fisher information along a path, while Hellinger distance is defined strictly in terms of the end points of the path. But Hellinger distance does have a derivative that evolves along a given path. So it may be enlightening to investigate differential relationships between d_I and d_H along paths, in particular for assessing the equivalence of "energy integrals" and the geodesics that arise as solutions of variational problems on these integrals.

It will be valuable to explore the relationships between the parametric framework presented in this paper and the non-parametric framework presented in [7] and [8]. It is certainly plausible, but remains to be proved, that there are asymptotic regimes where the non-parametric framework reduces to the parametric framework. Specifically, when multiple (n) i.i.d. measurements are collected from each sensor configuration along the planning path, and these measurements are quantized to a histogram, large n asymptotic theory asserts that the histograms converge to a multinomial distribution. Thus in this asymptotic case the sensor output distributions can be expected to be well approximated by the multinomial family investigated here. Under different assumptions the statistical distributions may converge to other parametric families, for example, to the multivariate Gaussian distribution under the CLT.

Finally, another important area of investigation will be the impact of navigation constraints that may prevent navigation along a desired trajectory in the parameter manifold. For example, there may be no sensor configuration in Γ corresponding to certain points or regions in Θ in the vicinity of the geodesic path. Approaches for accommodating such constraints will be essential if ideas like those outlined in this paper are to become useful tools in real-world sensor management applications.

Acknowledgment

This work was supported in part by the U. S. Army Research Office under MURI grant No. W911NF-11-1-0391 and by the U.S. Air Force Office of Scientific Research under grant No. FA9550-09-1-0561. The authors are grateful to Brian Sadler for his ongoing collaboration in the area of this paper.

References

[1] C. R. Rao, "Information and accuracy attainable in estimation of statistical parameters" *Bulletin of the Calcutta Mathematical Society*, vol. 37, pp. 81–91, 1945.

[2] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London*. Series A, vol. 186, issue 1007, pp. 453–461, 1946.

[3] S. Amari and H. Nagaoka, *Methods of Information Geometry*, AMS Translations of Mathematical Monographs, vol. 191, 2000.

[4] S.-I. Amari and J.-F. Cardoso, "Blind source separation-semiparametric statistical approach," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2692–2700, 1997.

[5] F. Barbaresco, "Interactions between symmetric cone and information geometries: Bruhat-tits and Siegel spaces models for high resolution autoregressive Doppler imagery," in *Emerging Trends in Visual Computing*, pp. 124–163, Springer, 2009.

[6] K. M. Carter, R. Raich, W. G. Finn, and A. O. Hero, "Information-geometric dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 89–99, 2011.

[7] W. Moran, S. D. Howard, D. Cochran, and S. Suvorova "Sensor management via Riemannian geometry," *Proceedings of the Allerton Conference on Communications, Control, and Computing*, pp. 358–362, October 2012.

[8] B. Moran, S. D. Howard, and D. Cochran, "An information-geometric approach to sensor scheduling," *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 52615264, April 2012.

[9] A. O. Hero, D. A. Castañón, D. Cochran, and K. Kastella, eds., *Foundations and Applications of Sensor Management*, New York, Springer, 2008.

[10] A. O. Hero III and D. Cochran, "Sensor management: Past, present, and future," *IEEE Sensors Journal*, vol. 11, no. 12, pp. 3064–3075, December 2011.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection,"*Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893. 2005.

[12] L. L. Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer, 2000.

[13] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[14] R. E. Kass and P. W. Voss, *Geometrical Foundations of Asymptotic Inference*, Wiley Interscience, 1997.