entropy

Article

Optimal Nonlinear Estimation in Statistical Manifolds with Application to Sensor Network Localization

Yongqiang Cheng [1,*], Xuezhi Wang [2] and Bill Moran [2]

[1] School of Electronic Science and Engineering, National University of Defense Technology, Changsha 410073, China

[2] School of Engineering, RMIT University, Melbourne 3000, Australia; xuezhi.wang@rmit.edu.au (X.W.); bill.moran@rmit.edu.au (B.M.)

* Correspondence: nudtyqcheng@gmail.com; Tel.: +86-731-8457-4452

Received: 30 April 2017; Accepted: 24 June 2017; Published: 28 June 2017

Abstract: Information geometry enables a deeper understanding of the methods of statistical inference. In this paper, the problem of nonlinear parameter estimation is considered from a geometric viewpoint using a natural gradient descent on statistical manifolds. It is demonstrated that the nonlinear estimation for curved exponential families can be simply viewed as a deterministic optimization problem with respect to the structure of a statistical manifold. In this way, information geometry offers an elegant geometric interpretation for the solution to the estimator, as well as the convergence of the gradient-based methods. The theory is illustrated via the analysis of a distributed mote network localization problem where the Radio Interferometric Positioning System (RIPS) measurements are used for free mote location estimation. The analysis results demonstrate the advanced computational philosophy of the presented methodology.

Keywords: information geometry; statistical manifolds; nonlinear estimation; natural gradient; maximum likelihood estimation

1. Introduction

Information geometry, pioneered by Rao in the 1940s [1] and further developed by Chentsov [2], Efron [3,4] and Amari [5,6], considers the statistical relationships between families of probability densities in terms of the geometric properties of Riemann manifolds. It is the study of intrinsic properties of manifolds of probability distributions [7], where the ability of the data to discriminate those distributions is translated into a Riemannian metric. Specifically, the Fisher information gives a local measure of discrimination of the distributions which immediately provides a Riemannian metric on the parameter manifold of the distributions [1]. In particular, the collection of probability density functions called curved exponential families, which encapsulate important distributions in many real world problems, have been treated using this framework [5].

The main tenet of information geometry is that many important notions in probability theory, information theory and statistics can be treated as structures in differential geometry by regarding a space of probabilities as a differentiable manifold endowed with a Riemannian metric and a family of affine connections, including but not exclusively, the canonical Levi-Civita affine connection [6]. By providing a means to analyse the Riemannian geometric properties of various families of probability density functions, information geometry offers comprehensive results about statistical problems simply by considering them as geometrical objects. Information geometry opens new prospect to study the intrinsic geometrical nature of information theory and provides a new way to deal with statistical problems on manifolds. For example, Smith [8] studied the intrinsic Cramér-Rao bounds on estimation accuracy for the estimation problems on arbitrary manifolds where the set of intrinsic coordinates is not apparent, and derived the intrinsic bounds in the examples of covariance matrix and subspace estimation. Srivastava et al. [9,10] addressed the geometric subspace estimation and target tracking problems under a Bayesian framework. Bhattacharya and Patrangenaru [11] treated the general problem of estimation on Riemannian manifolds.

As this new general theory reveals the capability of defining a new perspective on existing questions, many researchers are extending their work on this geometric theory of information to new areas of application and interpretation. For example, a most important milestone in the area of signal processing is the work of Richardson where the geometric perspective clearly indicates the relationship between turbo-decoding and maximum-likelihood decoding [12]. The results of Amari et al. on the information geometry of turbo and low-density parity-check codes extend the geometrical framework initiated by Richardson to the information geometrical framework of dual affine connections [13]. Other investigations include the geometrical interpretation of fading in wireless networks [14]; the geometrical interpretation of the solution to the multiple hypothesis testing problem in the asymptotic limit developed by Westover [15]; and a geometric characterization of multiuser detection for synchronous DS/CDMA channels [16]. Recently, the framework of information geometry has been applied to address issues in the application of sensor networks such as target resolvability [17], radar information resolution [18] and passive sensor scheduling [19,20].

In this paper, we mainly focus on the nonlinear estimation problem and illustrate how it can benefit from the powerful methodologies of information geometry. The geometric interpretation for the solution to the maximum likelihood estimation for curved exponential families and the convergence of the gradient-based methods (such as Newton’s method and the Fisher scoring algorithm) are demonstrated via the framework developed by Efron and Amari et al. Our essential motivation of this work is to provide some insights into the nonlinear parameter estimation problems in signal processing using the theory of information geometry. By gaining a better understanding of the existing algorithms through the use of information geometric method, we are, hopefully, enabled to derive better algorithms for solving non-linear problems.

The work described in this paper consists of the following aspects. Firstly, an iterative maximum likelihood estimator (MLE) for estimating non-random parameters with measurement of the curved exponential distributions is presented. The estimator belongs to the gradient-based methods that operate on statistical manifolds and can be seen as a generalization of Newton’s method to families of probability density functions and their relevant statistics. Its interpretation in terms of differential and information geometry provides insight into its optimality and convergence. Then, by utilizing the properties of exponential families and thus identifying the parameters on statistical manifolds, the implementation of the presented MLE algorithm is simplified via reparametrization. Furthermore, it is shown that the associated stochastic estimation problem reduces to a deterministic optimization problem with respect to the measures (statistics) defined over the distributions. Finally, an example of a one-dimensional curved Gaussian is presented to demonstrate the method in the manifold. A practical example of distributed mote network localization using the Radio Interferometric Positioning System (RIPS) measurements is given to demonstrate the issues addressed in this paper. The performance of the estimator is discussed.

In the next section, classical nonlinear estimation via natural gradient MLE for curved exponential families is derived. The reparametrization from local parameters to natural parameters and expectation parameters is highlighted. In Section 3, the principles of information geometry are introduced. Further, the geometric interpretation for the presented iterative maximum likelihood estimator and the convergence of the developed algorithm is demonstrated via the properties of statistical manifolds. In Section 4, a one parameter estimation example is presented to illustrate the geometric operation of the algorithm. The performance and efficiency of the algorithm are further demonstrated via a mote localization example using RIPS measurements. Finally, conclusions are made in Section 5.

2. Nonlinear Estimation via Natural Gradient MLE

In probability and statistics, exponential families (including the normal, exponential, Gamma, Chi-squared, Beta, Poisson, Bernoulli, multinomial and many others) are an important class of distributions naturally considered. There is a framework for selecting a possible alternative parameterization of these distributions, in terms of the natural parameters, and for defining useful statistics, called the natural statistics of the families. When the natural parameters in exponential families are nonlinear functions of some "local" parameters, the distributions involved are in the curved exponential families. While curved exponential families which encapsulate important distributions are more suitable to describe many real world problems, the estimation of local parameters is often non-trivial because of the nonlinearity in parameters.

In this section, a natural gradient based maximum likelihood estimator is derived to address a nonlinear estimation problem for curved exponential families. Although the estimator has been well-known as the Fisher scoring method in the literature, the interpretation of its iterative operations via the theory of information geometry is interesting and will be presented in the following section.

The general form of a curved exponential family [5] can be expressed as

$$p(\boldsymbol{x}|\boldsymbol{u}) = \exp\Big\{C(\boldsymbol{x}) + \boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{F}(\boldsymbol{x}) - \varphi\big(\boldsymbol{\theta}(\boldsymbol{u})\big)\Big\} = p\Big(\boldsymbol{x}|\boldsymbol{\theta}(\boldsymbol{u})\Big) \tag{1}$$

where $\boldsymbol{x} \in \Omega$ is a vector valued measurement, $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ are the natural coordinates or canonical parameters, $\boldsymbol{u}$ denote local parameters and $\boldsymbol{F}(\boldsymbol{x}) = \{F_1(\boldsymbol{x}), \ldots, F_n(\boldsymbol{x})\}$ are sufficient statistics for $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$, and functions on the measurement space with elements denoted by $\boldsymbol{x}$. The function φ is called the potential function of the exponential family and it is found from the normalisation condition $\int_\Omega p_{\boldsymbol{\theta}}(\boldsymbol{x})d\boldsymbol{x} = 1$, i.e.,

$$\varphi(\boldsymbol{\theta}) = \log \int_\Omega \exp\left\{C(\boldsymbol{x}) + \sum_i^n\Big(\theta_i F_i(\boldsymbol{x})\Big)\right\} d\boldsymbol{x} \tag{2}$$

The term "curved" comes from the fact that the distribution in Equation (1) is defined by a smooth embedding $\boldsymbol{u} \longrightarrow \boldsymbol{\theta}(\boldsymbol{u})$ from the manifold parameterized by $\boldsymbol{u}$ into the canonical exponential family $p(\boldsymbol{x}|\boldsymbol{\theta})$.

As an example, a curved Gaussian distribution with local parameter $\boldsymbol{u} \in \mathbb{R}^m$, mean $\boldsymbol{\mu}(\boldsymbol{u}) \in \mathbb{R}^n$ and covariance $\boldsymbol{\Sigma}(\boldsymbol{u}) \in \mathbb{R}^{n\times n}$ is expressed as

$$\boldsymbol{x} \sim \mathcal{N}\Big(\boldsymbol{\mu}(\boldsymbol{u}),\, \boldsymbol{\Sigma}(\boldsymbol{u})\Big) \tag{3}$$

By reparameterization, the standardized natural parameters $(\boldsymbol{\theta}, \boldsymbol{\Theta})$ of a curved Gaussian distribution are found to be

$$\boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{u})\,\boldsymbol{\mu}(\boldsymbol{u}) \tag{4}$$

$$\Theta = -\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{u}) \tag{5}$$

The sufficient statistics of the Gaussian distribution in Equation (3) is

$$\boldsymbol{F}(\boldsymbol{x}) = \{\boldsymbol{x},\, \boldsymbol{x}\boldsymbol{x}^T\} \tag{6}$$

and the potential function expressed in terms of local parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given by

$$\varphi(\mu, \boldsymbol{\Sigma}) = \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{n}{2}\log 2\pi. \tag{7}$$

where the superscript T signifies the transpose operation and n is the cardinality of $\boldsymbol{\mu} \in \mathbb{R}^n$.

Let $l(\boldsymbol{\theta}, \boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\theta})$ be the log likelihood of a general family of distributions $p(\boldsymbol{x}|\boldsymbol{\theta})$ as in Equation (1), and $\nabla\boldsymbol{\theta}$ the Jacobian matrix of the natural parameter $\boldsymbol{\theta}$ as a function of the local parameters $\boldsymbol{u}$. We may write the following equations by using Equation (1)

$$\begin{aligned} \nabla l(\boldsymbol{\theta}, \boldsymbol{x}) &= \nabla\left\{\boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{F}(\boldsymbol{x}) - \varphi\left(\boldsymbol{\theta}(\boldsymbol{u})\right)\right\} \\ &= \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\left[\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{\eta}(\boldsymbol{u})\right] \end{aligned} \tag{8}$$

where

$$\boldsymbol{\eta}(\boldsymbol{u}) \triangleq E\{\boldsymbol{F}(\boldsymbol{x})\} \tag{9}$$

is called the *expectation parameter* which connects to $\boldsymbol{\theta}(\boldsymbol{u})$ by the well known Legendre transformation [5], and

$$\nabla\varphi(\boldsymbol{u}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{\eta}(\boldsymbol{u}) \tag{10}$$

Both the expectation parameter $\boldsymbol{\eta}$ and Fisher information matrix $\boldsymbol{G}(\boldsymbol{\theta})$ can be obtained by differentiating the potential function $\varphi(\boldsymbol{\theta})$ with respect to natural parameters [21]

$$\eta_i = \frac{\partial}{\partial\theta_i}\varphi(\boldsymbol{\theta}) \triangleq E\left\{F_i(\boldsymbol{x})\right\} \tag{11}$$

$$g_{ij}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\varphi(\boldsymbol{\theta}) \triangleq E\left\{\left[F_i(\boldsymbol{x}) - E(F_i(\boldsymbol{x}))\right]\left[F_j(\boldsymbol{x}) - E(F_j(\boldsymbol{x}))\right]^T\right\} \tag{12}$$

The maximum likelihood estimator $\hat{\boldsymbol{u}}$ of the curved exponential family satisfies the following likelihood equation

$$\nabla l(\hat{\boldsymbol{u}}) = \nabla\log p(\boldsymbol{x}|\boldsymbol{u}) = \nabla\boldsymbol{\theta}^T(\hat{\boldsymbol{u}})\left[\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{\eta}(\hat{\boldsymbol{u}})\right] = 0 \tag{13}$$

Here $l(\hat{\boldsymbol{u}})$ is an objective function to be maximized with parameter $\boldsymbol{u}$. It was pointed out by Amari in [22] that the geometry of the Riemannian manifold must be taken into account when calculating the steepest learning directions on a manifold. He suggested the use of natural gradient (NAT) updates for the optimization on a Riemannian manifold, i.e.,

$$\boldsymbol{u}^{new} = \boldsymbol{u} + \lambda\boldsymbol{G}^{-1}(\boldsymbol{u})\nabla l(\boldsymbol{u}) \tag{14}$$

where λ is a positive learning rate that determines the step size and $\boldsymbol{G}(\boldsymbol{u})$ denotes the Riemannian metric matrix of the manifold.

For a parameterized family of probability distributions on a statistical manifold, the Riemannian metric is defined as the Fisher information matrix (FIM) [1]. For the curved exponential family in Equation (1), the FIM with respect to the local parameter $\boldsymbol{u}$ is

$$\boldsymbol{G}(\boldsymbol{u}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{G}(\boldsymbol{\theta})\nabla\boldsymbol{\theta}(\boldsymbol{u}) \tag{15}$$

where $\boldsymbol{G}(\boldsymbol{\theta})$ is the FIM with respect to the natural parameter $\boldsymbol{\theta}$. A recursive MLE of curved exponential families can then be implemented as follows

$$\begin{aligned} \boldsymbol{u}^{(k+1)} &= \boldsymbol{u}^{(k)} + \lambda\boldsymbol{G}^{-1}(\boldsymbol{u}^{(k)})\nabla l(\boldsymbol{u}^{(k)}) \\ &= \boldsymbol{u}^{(k)} + \lambda\boldsymbol{G}^{-1}(\boldsymbol{u}^{(k)})\nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(k)})\left[\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{\eta}(\boldsymbol{u}^{(k)})\right] \end{aligned} \tag{16}$$

where $\boldsymbol{\theta}(\boldsymbol{u})$ and $\boldsymbol{\eta}(\boldsymbol{u})$ denote the natural parameter and expectation parameter of the distribution, respectively. $\boldsymbol{F}(\boldsymbol{x})$ is the sufficient statistics for the measurement $\boldsymbol{x}$.

The covariance (CRLB) of the recursive MLE estimator $\boldsymbol{u}^{(k+1)}$ is the inverse of Fisher information matrix $\boldsymbol{G}^{-1}(\boldsymbol{u}^{(k+1)})$, where

$$\boldsymbol{G}(\boldsymbol{u}^{(k+1)}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(k+1)})\boldsymbol{G}\left(\boldsymbol{\theta}(\boldsymbol{u}^{(k+1)})\right)\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(k+1)}) \tag{17}$$

The proposed algorithm is summarized in Algorithm 1.

Algorithm 1: The natural gradient based iterative MLE algorithm.

```
1. Distribution reparameterization
       $p(\boldsymbol{x}|\boldsymbol{u}) = \exp\left\{C(\boldsymbol{x}) + \boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{F}(\boldsymbol{x}) - \varphi(\boldsymbol{\theta}(\boldsymbol{u}))\right\} = p\big(\boldsymbol{x}|\boldsymbol{\theta}(\boldsymbol{u})\big)$
   Identify the natural parameter $\boldsymbol{\theta}(\boldsymbol{u})$, sufficient statistics $\boldsymbol{F}(\boldsymbol{x})$ and potential function $\varphi\big(\boldsymbol{\theta}(\boldsymbol{u})\big)$.
2. Find expectation parameter $\boldsymbol{\eta}$ and Fisher information metric $\boldsymbol{G}$ to construct manifold of the
   curved exponential family $p\big(\boldsymbol{x}|\boldsymbol{\theta}(\boldsymbol{u})\big)$,
       $\boldsymbol{\eta}(\boldsymbol{u}) = E_{\boldsymbol{\theta}}\big(\boldsymbol{F}(\boldsymbol{x})\big) = \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta})$
       $\boldsymbol{G}(\boldsymbol{\theta}) = \mathrm{Cov}_{\boldsymbol{\theta}}\big(\boldsymbol{F}(\boldsymbol{x})\big) = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T\varphi(\boldsymbol{\theta}),\ \boldsymbol{G}(\boldsymbol{u}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{G}(\boldsymbol{\theta})\nabla\boldsymbol{\theta}(\boldsymbol{u})$
3. Input initial conditions
       $\boldsymbol{u}^{(0)},\ \boldsymbol{G}(\boldsymbol{u}^{(0)}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(0)})\boldsymbol{G}[\boldsymbol{\theta}(\boldsymbol{u}^{(0)})]\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(0)})$
4. Set step size $e^{(k)} > \varepsilon > 0,\ k = 0,$
   while $e^{(k)} > \varepsilon$
       Loop for the $(k+1)$th iteration
           $\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \lambda\boldsymbol{G}^{-1}(\boldsymbol{u}^{(k)})\nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(k)})\left[\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{\eta}(\boldsymbol{u}^{(k)})\right]$
           $\boldsymbol{G}(\boldsymbol{u}^{(k+1)}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(k+1)})\boldsymbol{G}\big(\boldsymbol{\theta}(\boldsymbol{u}^{(k+1)})\big)\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(k+1)})$
       Update step size
           $e^{(k+1)} = ||(\boldsymbol{u}^{(k)} - \boldsymbol{u}^{(k+1)}||$
           $k+1 \to k$
   end
```

The above natural gradient approach has a similar structure as the common gradient-based algorithms, such as the well-known steepest descent method, Newton's method and Fisher scoring algorithm. However, it does distinguish itself from the others in the following points:

- The natural gradient estimator updates the underlying manifold metric $\boldsymbol{G}$ (i.e., FIM) at each iteration as well, which evaluates the estimate accuracy.
- Updates in the classical steepest descent types are performed via the standard gradient $\nabla l(\boldsymbol{u})$ and are well-matched to the Euclidean distance measure as well as the gradient adaptation. For the cases where the underlying parameter spaces are not Euclidean but are curved, i.e., Riemannian, $-\nabla l(\boldsymbol{u})$ does not represent the steepest descent direction in the parameter space, and thus the standard gradient adaptation is no longer appropriate. The natural gradient updates in Equation (14) improve the steepest descent update rule by taking the geometry of the Riemannian manifold into account to calculate the learning directions. In other words, it modifies the standard gradient direction according to the local curvature of the parameter space in terms of the Riemannian metric tensor $\boldsymbol{G}(\boldsymbol{u})$, thus offers faster convergence than the steepest descent method.
- The Newton method

$$\boldsymbol{u}^{new} = \boldsymbol{u} - \lambda\left[\frac{\partial^2 l(\boldsymbol{u})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right]^{-1}\nabla l(\boldsymbol{u}). \tag{18}$$

improves the steepest descent method by using the second-order derivatives of the cost function, i.e., the inverse of the Hessian of $l(\boldsymbol{u})$ to adjust the gradient search direction. When $l(\boldsymbol{u})$ is a quadratic function of $\boldsymbol{u}$, the inverse of the Hessian is equal to $\boldsymbol{G}(\boldsymbol{u})$, and thus Newton's method and the natural gradient approach are identical [22]. However, in more general contexts, the two techniques are different. Generally, the natural gradient approach increases the stability of the iteration with respect to Newton's method through replacing the Hessian by its expected value, i.e., the Riemannian metric tensor $\boldsymbol{G}(\boldsymbol{u})$.

- The natural gradient approach is identical to the Fisher scoring method in cases where the Fisher information matrix coincides with the Riemannian metric tensor of the underlying parameter space. In such cases, the natural gradient approach is a Riemannian-based version of the Fisher scoring method performed on manifolds, and it is very appropriate when the cost function is related to the Riemannian geometry of the underlying parameter space [23]. Once these methods are entered into the manifold, additional insights into their geometric meaning may be deduced in the framework of differential and information geometry.

It is worth mentioning that a strategic choice of parameterizations of the cost function may result in a faster convergence or a more meaningful implementation of an optimization algorithm, though it is quite-often non-trivial. In the proposed iterative MLE algorithm in Equation (16), an alternative parameterization of the curved exponential family, in terms of the natural and expectation parameters, are employed. Through such a reparameterization, the implementation of the natural gradient updates is facilitated by the relevant statistics of a curved exponential family.

3. Information Geometric Interpretation for Natural Gradient MLE

3.1. Principles of Information Geometry

(1) Definition of a statistical manifold: Information geometry originates from the study of manifolds of probability distributions. Consider the parameterized family of probability distributions $S = \{ p(\boldsymbol{x}|\boldsymbol{\theta}) \}$, where $\boldsymbol{x}$ is a random variable and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is a parameter vector specifying the distribution. The family S is regarded as a statistical manifold with $\boldsymbol{\theta}$ as its (possibly local) coordinate system [24].

Figure 1 illustrates the definition of a statistical manifold. For a given state of interest $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta} \in \mathbb{R}^n$, the measurement $\boldsymbol{x}$ in the sample space $\boldsymbol{X} \in \mathbb{R}^m$ is an instantiation of a probability distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$. Each probability distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ is labelled by a point $s(\boldsymbol{\theta})$ in the manifold S. The parameterized family of probability distributions $S = \{ p(\boldsymbol{x}|\boldsymbol{\theta}) \}$ forms an n-dimensional statistical manifold where $\boldsymbol{\theta}$ plays the role of a coordinate system of $\boldsymbol{S}$.

(2) Fisher information metric: The metric is the object specifying the scalar product in a particular point on a manifold in differential geometry. It encodes how to measure distances, angles and area at a particular point on the manifold by specifying the scalar product between tangent vectors at that point [25]. For a parameterized family of probability distributions on a statistical manifold, the FIM plays the role of a Riemannian metric tensor [1]. Denoted by $\boldsymbol{G}(\boldsymbol{\theta}) = [g_{ij}(\boldsymbol{\theta})]$, where

$$g_{ij} = E \left\{ \frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \log p(\boldsymbol{x}|\boldsymbol{\theta})}{\partial \theta_j} \right\}, \tag{19}$$

the FIM measures the ability of the random variable $\boldsymbol{x}$ to discriminate the values of the parameter $\boldsymbol{\theta}'$ from $\boldsymbol{\theta}$ for $\boldsymbol{\theta}'$ close to $\boldsymbol{\theta}$.

(3) Affine connection and Flatness: The affine connection ∇ (The notation ∇ is also used to denote the Jacobian in this paper. However, there should be no confusion from the context.) on a manifold S defines a linear one-to-one mapping between two neighboring tangent spaces of the manifold. When the connection coefficients of ∇ with respect to a coordinate system of S are all identically 0, then ∇ is said to be *flat*, or alternatively, *S is flat with respect to* ∇. The curvature of a flat manifold is zero everywhere. Correspondingly, flatness will result in considerable simplification of the geometry. Intuitively, a flat manifold is one that "locally looks like" a Euclidean space in terms of distances and

angles. Consequently, many operations on the flat manifold such as projection and orthogonality become more closely analogous to the case of an Euclidean space.

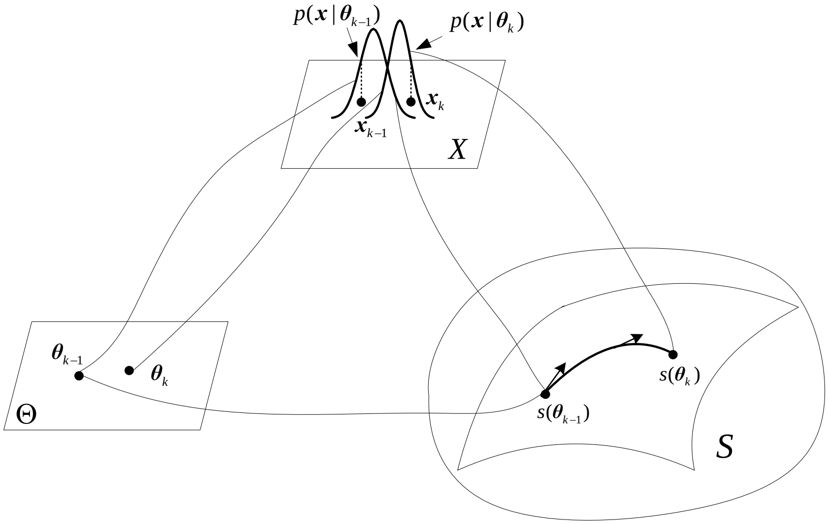

$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_{k-1})$
$p(\boldsymbol{x}\,|\,\boldsymbol{\theta}_{k})$
$\boldsymbol{x}_k$
$\boldsymbol{x}_{k-1}$
X
$\boldsymbol{\theta}_{k-1}$
$\boldsymbol{\theta}_{k}$
Θ
$s(\boldsymbol{\theta}_{k})$
$s(\boldsymbol{\theta}_{k-1})$
S

Figure 1. Definition of a statistical manifold.

It is worth mentioning that the flatness of a manifold is closely related to the definition of affine connections as well as the choice of coordinate systems of the manifold. In 1972, Chentsov [2] introduced a one-parameter family of affine connections called α-connections which were later popularized by Amari [5]:

$$\overset{\alpha}{\boldsymbol{\Gamma}}_{jim}\,(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\Big\{\partial_j\partial_i\, l(\boldsymbol{x},\,\boldsymbol{\theta})\,\partial_m l(\boldsymbol{x},\,\boldsymbol{\theta})\Big\} + \frac{1-\alpha}{2}E_{\boldsymbol{\theta}}\Big\{\partial_j l(\boldsymbol{x},\,\boldsymbol{\theta})\,\partial_i l(\boldsymbol{x},\,\boldsymbol{\theta})\,\partial_m l(\boldsymbol{x},\,\boldsymbol{\theta})\Big\} \tag{20}$$

where

$$\partial_i = \frac{\partial}{\partial\theta_i} \quad \text{and} \quad l(\boldsymbol{\theta},\,\boldsymbol{x}) = \log p(\boldsymbol{x}|\boldsymbol{\theta}). \tag{21}$$

In Equation (20), $\alpha = 0$ corresponds to the Levi-Civita connection (information connection). The case $\alpha = -1$ defines the e-connection (exponential connection) while $\alpha = -1$ defines the m-connection (mixture connection). An exponential family with natural parameter $\boldsymbol{\theta}$ as the coordinate system (parameterization) is a flat manifold under the e-connection while a mixture family with expectation parameter $\boldsymbol{\eta}$ as the coordinate system is a flat manifold under the m-connection [24]. The e-connection and m-connection play an important role in statistical inference and geometrize the estimation on the flat manifolds.

3.2. Information Geometric Interpretation for Natural Gradient MLE

Based on the principles of information geometry introduced above, the algorithm described in Algorithm 1 can be explained via Figure 2, where the upper figure illustrates the estimation operation in Euclidean space while an alternative view of the estimation operation on statistical manifolds is analogously illustrated in the lower figure. In the upper figure, the local parameter $\boldsymbol{u}$ is to be estimated from measurements (samples) $\boldsymbol{x}$ via the likelihood function $p(\boldsymbol{x}|\boldsymbol{u})$. When the measurement model is nonlinear in the parameter, the underlying estimation of the local parameter is a nonlinear estimation or filtering problem. Usually, linear signal processing problems can be routinely solved systematically by the astute application of results from linear algebra. However, the nonlinear cases are not easy to solve. The methodology of differential and information geometry are more adaptable and capable of dealing with nonlinear problems.

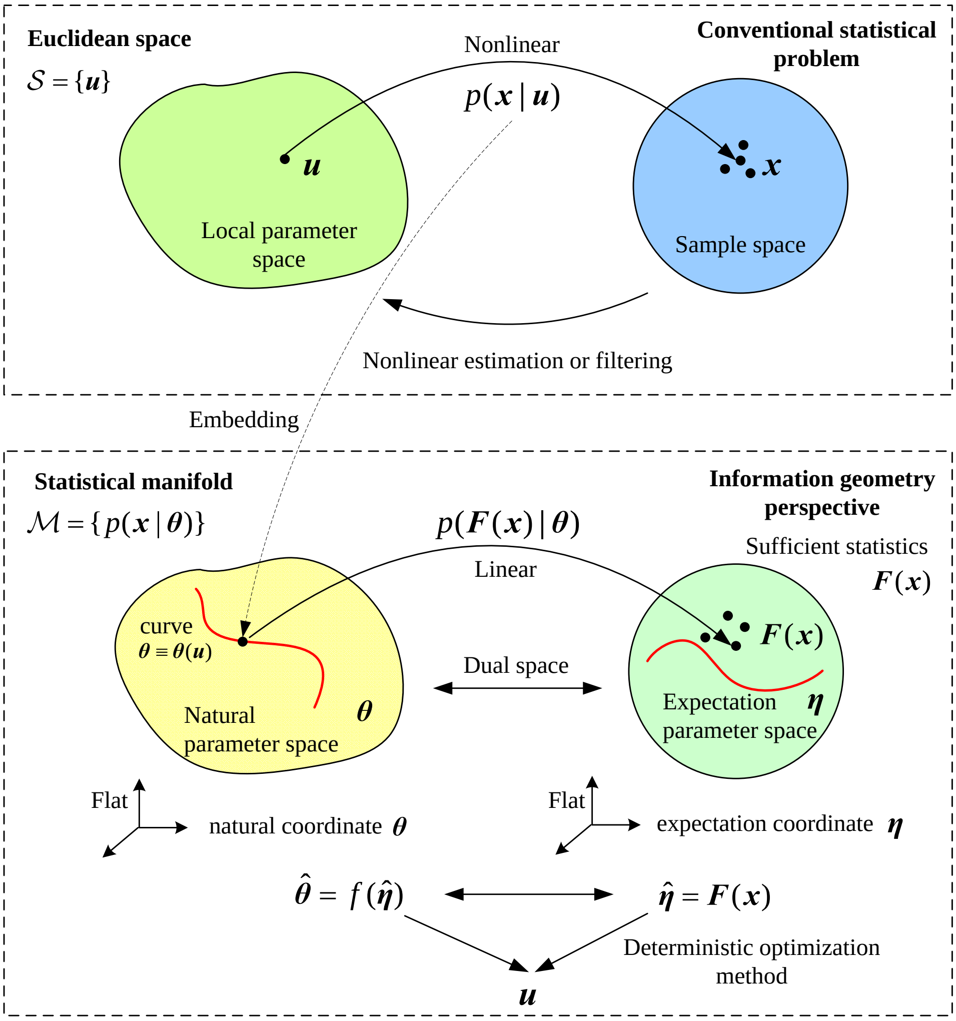

Euclidean space
$\mathcal{S} = \{\boldsymbol{u}\}$
Nonlinear
$p(\boldsymbol{x} \mid \boldsymbol{u})$
Conventional statistical problem
$\boldsymbol{u}$
$\boldsymbol{x}$
Local parameter space
Sample space
Nonlinear estimation or filtering
Embedding
Statistical manifold
$\mathcal{M} = \{p(\boldsymbol{x} \mid \boldsymbol{\theta})\}$
$p(\boldsymbol{F}(\boldsymbol{x}) \mid \boldsymbol{\theta})$
Information geometry perspective
Sufficient statistics
$\boldsymbol{F}(\boldsymbol{x})$
Linear
curve
$\boldsymbol{\theta} \equiv \boldsymbol{\theta}(\boldsymbol{u})$
Dual space
$\boldsymbol{F}(\boldsymbol{x})$
Natural parameter space
$\boldsymbol{\theta}$
Expectation parameter space
$\boldsymbol{\eta}$
Flat
natural coordinate $\boldsymbol{\theta}$
Flat
expectation coordinate $\boldsymbol{\eta}$
$\hat{\boldsymbol{\theta}} = f(\hat{\boldsymbol{\eta}})$
$\hat{\boldsymbol{\eta}} = \boldsymbol{F}(\boldsymbol{x})$
Deterministic optimization method
$\boldsymbol{u}$

Figure 2. Illustrates the nonlinear estimation concept in both Euclidean space and statistical manifolds of the dual spaces of natural parameter $\boldsymbol{\theta}$ and expectation parameter $\boldsymbol{\eta}$. **Top** figure shows relations between parameter spaces (left) and samples (right) and associated estimating mappings. **Bottom** figure shows equivalent model for manifolds of parameter (left) and sample distributions (right). In both cases, the aims are to estimate the most likely parameter values that predict the data, and vice-versa.

Given that $\mathcal{M} = \{p(\boldsymbol{x}|\boldsymbol{\theta})\}$ signifies a general set of conditional distributions, the natural parameter space $\{\boldsymbol{\theta}\} \in \mathcal{A}^n$ contains the distributions of all exponential families; to be regarded as an enveloping space. Then the curved exponential family $p(\boldsymbol{x}|\boldsymbol{u})$ in the upper figure is smoothly embedded in the enveloping space $\mathcal{A}^n$, $m \leq n$ by distribution reparameterization $\boldsymbol{u} \longrightarrow \boldsymbol{\theta}(\boldsymbol{u})$, i.e., the curved exponential family $p(\boldsymbol{x}|\boldsymbol{u})$ can be represented by a curve $\{\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{u})\}$ embedded in $\mathcal{A}^n$. Consequently the nonlinearity in the underlying estimation problem is completely characterized by the red curve inside the natural parameter space.

The circle on the lower right side of Figure 2 is the expectation parameter space $\{\boldsymbol{\eta}\} \in \mathcal{B}^n$ or sampling space, which is dual to the natural parameter space $\mathcal{A}$. The dots in the space $\mathcal{B}$ signify the "realizations" of the sufficient statistics $\boldsymbol{F}(\boldsymbol{x})$ of the distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ and they are obtained from measurements (samples) $\boldsymbol{x}$. Connected by the Legendre transformation the dual enveloping sub-manifolds $\mathcal{A}$ and $\mathcal{B}$ (i.e., natural and expectation parameter spaces) are in one-to-one correspondence [5]. By viewing the expression of the curved exponential families in Equation (1), we observe that the newly formulated likelihood $p\left(\boldsymbol{F}(\boldsymbol{x})|\boldsymbol{\theta}\right)$ is linear, which indicates the possibility for linearly estimating the natural parameter $\boldsymbol{\theta}$ by sufficient statistics $\boldsymbol{F}(\boldsymbol{x})$ firstly and then obtaining the estimation of local parameter $\boldsymbol{u}$ by a deterministic mapping from $\boldsymbol{\theta}$ to $\boldsymbol{u}$.

The nonlinear filtering is performed in the expectation parameter space $\mathcal{B}$ by projecting the samples $\boldsymbol{F}(\boldsymbol{x})$ on to the sub-manifold represented by the red curve which signifies the embedding of the curved exponential families in $\mathcal{B}$. The process is also called m-projection. As mentioned earlier, under both the e- and m-connections, the dual enveloping manifolds are flat. Therefore, the filtering (m-projection) is analogous to the projection in a deterministic Euclidean space. In consequence, the nonlinear estimation problem is finally realized as a deterministic optimization method.

The fundamental difference between the filter described here and existing nonlinear filters is that the filtering process presented is performed linearly in the dual spaces of natural parameters and expectation parameters under e- and m-connections, respectively. The filtering outcome is then mapped to local parameter space. The estimator is optimal in the MLE sense and thus has no information loss in the filtering process since it attains the CRLB [5].

The convergence of the nonlinear iterative estimator can be geometrically explained in information geometric terms diagrammatically via Figure 3. The curved exponential family in Equation (1) is represented by the curve $\mathscr{F}_A \equiv \{\boldsymbol{\theta}(\boldsymbol{u}) : \boldsymbol{u} \in \mathbb{R}^m\}$ in $\mathcal{A}$ and also by $\mathscr{F}_B \equiv \{\boldsymbol{\eta}(\boldsymbol{u}) : \boldsymbol{u} \in \mathbb{R}^m\}$ in the dual space $\mathcal{B}$. Starting from an initial parameter $\boldsymbol{u}^{(k)}$, the algorithm constructs a vector $\mathscr{L}_{\boldsymbol{u}^{(k)}} \equiv \{\tilde{\boldsymbol{\eta}}(\boldsymbol{u}^{(k)}) = \boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{\eta}(\boldsymbol{u}^{k})\}$ from the current distribution represented by its expectation parameter $\boldsymbol{\eta}(\boldsymbol{u}^{(k)})$ to the measurement $\boldsymbol{F}(\boldsymbol{x})$. The projection of $\tilde{\boldsymbol{\eta}}(\boldsymbol{u}^{(k)})$ to the tangent vector $\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(k)})$ of the natural parameter $\boldsymbol{\theta}(\boldsymbol{u}^{(k)})$ with respect to the metric $\boldsymbol{G}(\boldsymbol{u}^{(k)})$ gives the steepest descent gradient (natural gradient) to update the current estimates (where $\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(k)})$ is represented by the dashed arrow in both $\mathcal{A}$ and $\mathcal{B}$, while the natural gradient is represented by the solid arrow in $\mathcal{B}$).

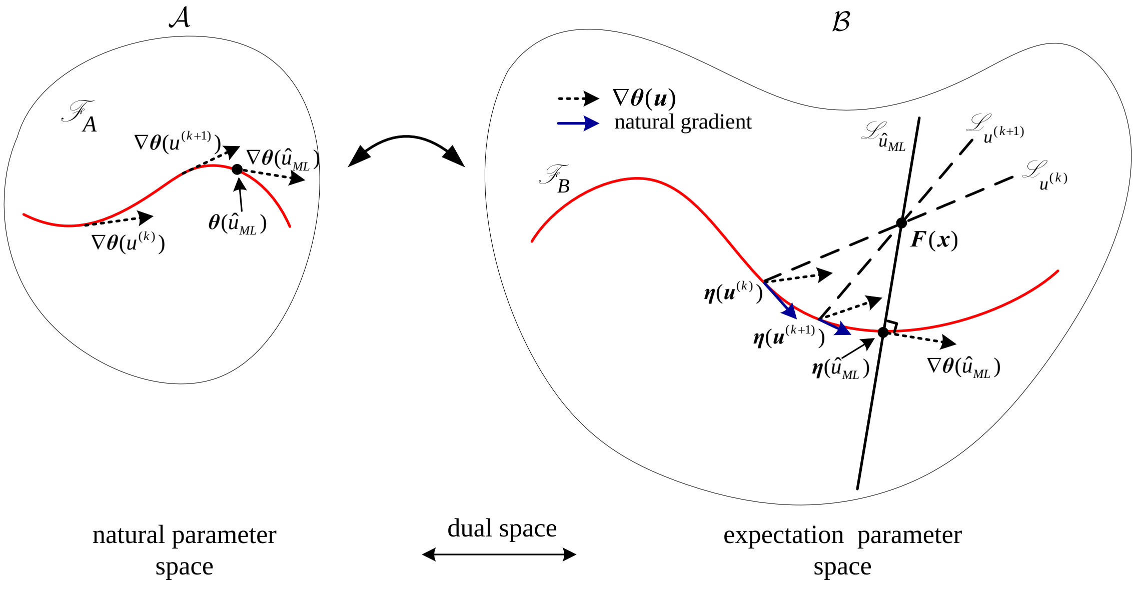

$\mathcal{A}$
$\mathscr{F}_A$
$\nabla\boldsymbol{\theta}(u^{(k+1)})$
$\nabla\boldsymbol{\theta}(\hat{u}_{ML})$
$\boldsymbol{\theta}(\hat{u}_{ML})$
$\nabla\boldsymbol{\theta}(u^{(k)})$
$\mathcal{B}$
$\nabla\boldsymbol{\theta}(\boldsymbol{u})$
natural gradient
$\mathscr{L}_{\hat{u}_{ML}}$
$\mathscr{L}_{u^{(k+1)}}$
$\mathscr{L}_{u^{(k)}}$
$\mathscr{F}_B$
$\boldsymbol{F}(\boldsymbol{x})$
$\boldsymbol{\eta}(\boldsymbol{u}^{(k)})$
$\boldsymbol{\eta}(\boldsymbol{u}^{(k+1)})$
$\boldsymbol{\eta}(\hat{u}_{ML})$
$\nabla\boldsymbol{\theta}(\hat{u}_{ML})$
natural parameter space
dual space
expectation parameter space

Figure 3. Convergence of the presented iterative maximum likelihood estimator (MLE) algorithm.

The iterations continue according to Equation (16) until the two vectors $\tilde{\boldsymbol{\eta}}(\boldsymbol{u}^{(k)})$ and $\nabla\boldsymbol{\theta}(\boldsymbol{u}^{(k)})$ are (approximately) orthogonal to each other, i.e., $\tilde{\boldsymbol{\eta}}(\hat{\boldsymbol{u}}_{ML}) \perp \nabla\boldsymbol{\theta}(\hat{\boldsymbol{u}}_{ML})$. The algorithm achieves convergence with the steepest descent gradient $\boldsymbol{G}^{-1}(\boldsymbol{u}^{(k)})\nabla\boldsymbol{\theta}^T(\boldsymbol{u}^{(k)})\tilde{\boldsymbol{\eta}}(\boldsymbol{u}^{(k)})$ vanishes and a solution to the MLE Equation (13) is obtained by projecting the data $\boldsymbol{F}(\boldsymbol{x})$ onto $\mathscr{F}_B$ orthogonally to $\nabla\boldsymbol{\theta}(\boldsymbol{u})$.

- Statistical problems can be described in manifolds in a number of ways. In the parameter estimation problems as we have discussed here the parameter belongs to a curved manifold, whereas the observations may lie on an enveloping manifold. The filtering process is thus implemented by means of projection in the manifolds.

- The iterative estimator is optimal in the MLE sense as the filtering itself involves no information loss. The stochastic filtering problem becomes an optimization problem defined over a statistical manifold.
- As seen from Algorithm 1, the algorithm implementation is relatively simple and straightforward by distribution reparameterization and operating in the dual flat manifolds. Though a Newton method-based MLE estimator can be derived directly via the likelihood. However, in most cases the operation is not trivial.
- The initial guess is important to facilitate convergence of the estimator to the true value. This can be varied and such initial value sampling may provide more certainty about reaching a global minimum. This has not been examined here.

In the next section, two examples are given to demonstrate the implementation of the developed estimator as well as its geometric interpretation.

4. Examples of Implementation of Natural Gradient MLE

4.1. An One Parameter Estimation Example of Curved Gaussian Distribution

Consider a curved Gaussian distribution

$$x \sim \mathcal{N}(u, u^2a^2) \tag{22}$$

where a is a constant and u is an unknown parameter to be estimated. The collection of distributions specified by the parameter u constitute a one-dimensional curved exponential family, which can be embedded in the natural parameter space $\mathcal{A}$ in terms of natural coordinates

$$\theta_1 = \frac{1}{a^2u}, \qquad \theta_2 = -\frac{1}{2a^2u^2} \tag{23}$$

which is a parabola (denoted by $\mathscr{F}_A$)

$$\theta_2 = -\frac{a^2}{2}\theta_1^2 \tag{24}$$

in $\mathcal{A}$. The underlying distribution in Equation (22) can be alternatively embedded in the expectation parameter space $\mathcal{B}$ in terms of expectation coordinates

$$\eta_1 = u, \qquad \eta_2 = (a^2+1)u^2 \tag{25}$$

which is also a parabola (denoted by $\mathscr{F}_B$)

$$\eta_2 = (a^2+1)\eta_1^2 \tag{26}$$

in $\mathcal{B}$.

The tangent vector $\nabla\boldsymbol{\theta}(u)$ of the curve $\mathscr{F}_A$ is

$$\nabla\boldsymbol{\theta}(u) = \frac{1}{a^2u^3}[-u, 1] \tag{27}$$

The metric $\boldsymbol{G}(u)$ has only one component g in this case, and is

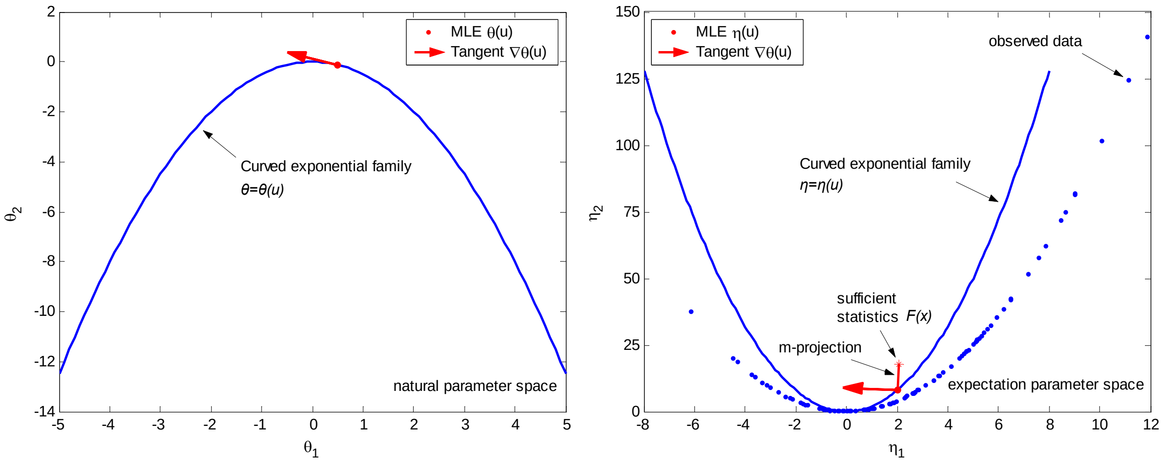
$$g = \frac{2a^2+1}{a^2u^2} \tag{28}$$

The sufficient statistics $\boldsymbol{F}(x)$ of the underlying distribution are

$$\boldsymbol{F}(x) = \{x, x^2\} \tag{29}$$

and they are obtained from measurements (samples) x.

Figure 4 shows the two dual flat spaces ($\mathcal{A}$ and $\mathcal{B}$) and illustrates the estimation process implemented in them, where the blue parabolas in two figures denote the embeddings of the curved Gaussian distribution specified by parameter u. Without loss of generality, $a = 1, u = 2$ are assumed. The red arrows in two figures show the tangent vector $\nabla\boldsymbol{\theta}(u)$ of the curve $\mathscr{F}_A$ while the two red dots on the parabolas denote the "realizations" of the distribution in Equation (22) specified by the given parameter $u = 2$. One hundred observed data (measurements) are shown by the blue dots in the expectation parameter space specified by coordinates (x, x^2). The red asterisk denotes the sufficient statistics $\boldsymbol{F}(x)$ obtained from the statistical mean of the measurements (samples). By projecting the data $\boldsymbol{F}(x)$ on to the sub-manifold represented by $\mathscr{F}_B$ orthogonally to $\nabla\boldsymbol{\theta}(u)$, i.e., $(\boldsymbol{F}(x) - \boldsymbol{\eta}(\hat{u}_{ML})) \perp \nabla\boldsymbol{\theta}(\hat{u}_{ML})$, the MLE estimation of parameter u is obtained. By viewing Equation (25) and Figure 4b, the estimation $\hat{u}_{ML} = \hat{\eta}_1$ is with high accuracy to the true value of u in this example.



(a) (b)

Figure 4. Estimation example of one-dimensional curved Gaussian distribution. (**a**) Shows the natural parameter space and embedding of the curved Gaussian distribution in it; (**b**) Shows the dual expectation parameter space and the estimation process in it.

4.2. A Mote Localization Example via RIPS Measurements

The Radio Interferometric Positioning System (RIPS) is an efficient sensing technique for sensor networks of small, low cost sensors with various applications. It utilizes radio frequency interference to obtain a sum of distance differences between the locations of a quartet of the motes which is initially reported in [26] and further discussed in [27]. In this paper, we take this mote localization problem via RIPS measurements as an application of the presented natural gradient MLE.

(1) Problem description: The Radio Interferometric Positioning System (RIPS) measurement model is described in [28]. A single RIPS measurement, as described in [26], involves four motes. A mote is a node in a sensor network that is capable of performing some processing, gathering sensory information and communicating with other connected nodes in the network. The main components of a sensor mote are a microcontroller, transceiver, external memory, power source and sensing hardware device. Figure 5 illustrates a collection of four motes A, B, C, D, where A, B and C are *anchor* motes (i.e., their location states are already known) and D is a free mote with unknown location. Two motes act as transmitters, sending a pure sine wave at slightly different frequencies. This results in an interference signal at a low beat frequency that is received by the other two motes (acting as receivers). A sum of range differences between the four motes can be obtained from the phase difference of the received interference signals at the two receiver locations. If motes A and B serve as transmitters and motes C and D form the receiver pair, then the corresponding RIPS measurement, denoted $k_{A,B,C,D}$, measures the distance differences

$$\begin{aligned} k_{A,B,C,D} &= ||\boldsymbol{X}_D - \boldsymbol{X}_A|| - ||\boldsymbol{X}_D - \boldsymbol{X}_B|| \\ &+ ||\boldsymbol{X}_B - \boldsymbol{X}_C|| - ||\boldsymbol{X}_A - \boldsymbol{X}_C|| \end{aligned} \tag{30}$$

which may be simply written as

$$k_{A,B,C,D} = d_{AD} - d_{BD} + d_{BC} - d_{AC}. \tag{31}$$

In the absence of noise, two independent RIPS measurements can be found. Therefore, the other independent measurement is given by

$$k_{A,C,B,D} = d_{AD} - d_{CD} + d_{BC} - d_{AB} \tag{32}$$

which uses motes A and C as a transmitter pair and B and D as a receiver pair.

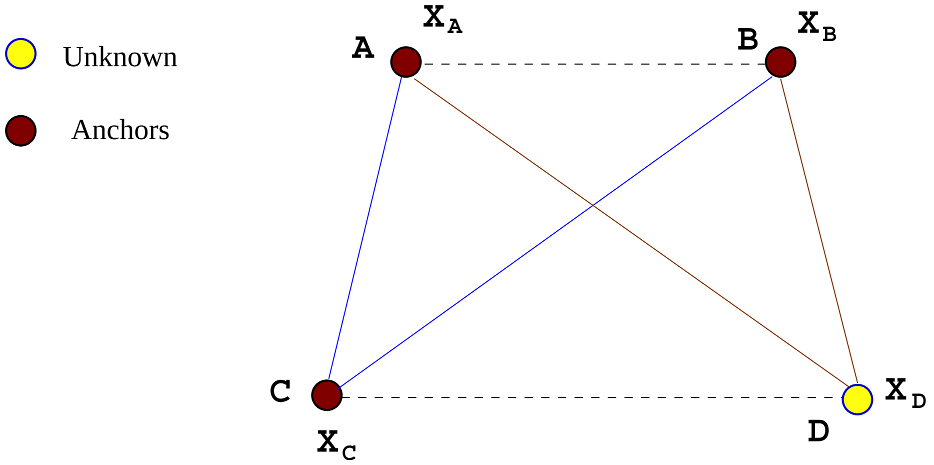


Figure 5. Radio Interferometric Positioning System (RIPS) measurement involving four sensors with three known anchors and one unknown sensor.

The two independent RIPS measurements can be written as

$$\boldsymbol{k}(\boldsymbol{u}) = \begin{bmatrix} k_{A,B,C,D} \\ k_{A,C,B,D} \end{bmatrix} = \begin{bmatrix} \delta_{ab} + \sqrt{(x_a - u_x)^2 + (y_a - u_y)^2} - \sqrt{(x_b - u_x)^2 + (y_b - u_y)^2} \\ \delta_{ac} + \sqrt{(x_a - u_x)^2 + (y_a - u_y)^2} - \sqrt{(x_c - u_x)^2 + (y_c - u_y)^2} \end{bmatrix} \tag{33}$$

where $\boldsymbol{u} = [u_x, u_y]'$ is the unknown location of the free node D, and

$$\begin{aligned} \delta_{ab} &= d_{BC} - d_{AC} \\ &= \sqrt{(x_b - x_c)^2 + (y_b - y_c)^2} - \sqrt{(x_a - x_c)^2 + (y_a - y_c)^2} \\ \delta_{ac} &= d_{BC} - d_{AB} \\ &= \sqrt{(x_b - x_c)^2 + (y_b - y_c)^2} - \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2} \end{aligned}$$

are both known constants in which $x_i, y_i, i = a, b, c$, are the location coordinates of the three anchor motes.

Accordingly, a generic RIPS measurement model can be written as

$$\boldsymbol{x} = \boldsymbol{k}(\boldsymbol{u}) + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \sigma^2 I_{2\times 2} \tag{34}$$

The underlying localization problem is to estimate the location of the free mote D based on RIPS measurement in Equation (34), where we assume that the knowledge of anchor node locations are known. The problem of locating the node D from RIPS measurements corrupted with Gaussian noise in Equation (34) reduces to a nonlinear parameter estimation problem. We adopt the natural gradient based MLE estimator described above to address it.

(2) Mote localization via RIPS measurements: Based on the RIPS measurement model in Equation (34), we can write the likelihood function in the form

$$p(\boldsymbol{x}|\boldsymbol{u}) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[\boldsymbol{x} - \boldsymbol{k}(\boldsymbol{u})]^T \Sigma^{-1} [\boldsymbol{x} - \boldsymbol{k}(\boldsymbol{u})]\right\} \tag{35}$$

Rearranging Equation (35) to describe in terms of the canonical curved exponential family, we obtain

$$p(\boldsymbol{x}|\boldsymbol{u}) = \exp\Big\{C(\boldsymbol{x}) + \boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{F}(\boldsymbol{x}) - \varphi[\boldsymbol{\theta}(\boldsymbol{u})]\Big\}, \tag{36}$$

which yields (The required quantities can be obtained directly from the standard relations between Equations (4)–(7) for curved Gaussian distributions.)

$$\boldsymbol{\theta}(\boldsymbol{u}) = \Sigma^{-1}\boldsymbol{k}(\boldsymbol{u}) \tag{37}$$

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{x} \tag{38}$$

$$\varphi[\boldsymbol{\theta}(\boldsymbol{u})] = -\frac{1}{4}\mathrm{tr}(\Theta^{-1}\boldsymbol{\theta}\boldsymbol{\theta}^T) - \frac{1}{2}\log|-\Theta| + \log\pi, \quad \Theta = -\frac{1}{2}\Sigma^{-1} \tag{39}$$

The expectation parameter and FIM on natural parameter are given by

$$\boldsymbol{\eta}(\boldsymbol{u}) = \nabla_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta}) = -\frac{1}{2}\Theta^{-1}\boldsymbol{\theta} = \boldsymbol{k}(\boldsymbol{u}) \tag{40}$$

$$\boldsymbol{G}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^T\varphi(\boldsymbol{\theta}) = -\frac{1}{2}\Theta^{-1} = \Sigma \tag{41}$$

The Jacobian matrix of natural parameter $\boldsymbol{\theta}$ with respect to local parameter $\boldsymbol{u}$ is given by

$$\nabla\boldsymbol{\theta}(\boldsymbol{u}) = \Sigma^{-1}\nabla_{\boldsymbol{u}}\boldsymbol{k}(\boldsymbol{u}) = \Sigma^{-1}\begin{bmatrix} \frac{\partial k_{A,B,C,D}}{\partial u_x} & \frac{\partial k_{A,B,C,D}}{\partial u_y} \\ \frac{\partial k_{A,C,B,D}}{\partial u_x} & \frac{\partial k_{A,C,B,D}}{\partial u_y} \end{bmatrix} \tag{42}$$

where

$$\frac{\partial k_{A,B,C,D}}{\partial u_x} = \frac{x_b - u_x}{\sqrt{(x_b-u_x)^2 + (y_b-u_y)^2}} - \frac{x_a - u_x}{\sqrt{(x_a-u_x)^2 + (y_a-u_y)^2}} \tag{43}$$

$$\frac{\partial k_{A,B,C,D}}{\partial u_y} = \frac{y_b - u_y}{\sqrt{(x_b-u_x)^2 + (y_b-u_y)^2}} - \frac{y_a - u_y}{\sqrt{(x_a-u_x)^2 + (y_a-u_y)^2}} \tag{44}$$

$$\frac{\partial k_{A,C,B,D}}{\partial u_x} = \frac{x_c - u_x}{\sqrt{(x_c-u_x)^2 + (y_c-u_y)^2}} - \frac{x_a - u_x}{\sqrt{(x_a-u_x)^2 + (y_a-u_y)^2}} \tag{45}$$

$$\frac{\partial k_{A,C,B,D}}{\partial u_y} = \frac{y_c - u_y}{\sqrt{(x_c-u_x)^2 + (y_c-u_y)^2}} - \frac{y_a - u_y}{\sqrt{(x_a-u_x)^2 + (y_a-u_y)^2}} \tag{46}$$

The FIM with respect to the local parameter $\boldsymbol{u}$, i.e., Equation (15) becomes

$$\boldsymbol{G}(\boldsymbol{u}) = \nabla\boldsymbol{\theta}^T(\boldsymbol{u})\boldsymbol{G}(\boldsymbol{\theta})\nabla\boldsymbol{\theta}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}^T\boldsymbol{k}(\boldsymbol{u})\Sigma^{-1}\nabla_{\boldsymbol{u}}\boldsymbol{k}(\boldsymbol{u}) \tag{47}$$

Therefore, the iterative MLE estimator for estimating the location of a free mote using RIPS measurements is implemented as

$$\boldsymbol{u}^{(k+1)} = \boldsymbol{u}^{(k)} + \lambda\left[\nabla_{\boldsymbol{u}}\boldsymbol{k}(\boldsymbol{u}^{(k)})\right]^{-1}[\boldsymbol{x} - \boldsymbol{k}(\boldsymbol{u}^{(k)})] \tag{48}$$

$$\boldsymbol{G}(\boldsymbol{u}^{(k+1)}) = \left[\nabla_{\boldsymbol{u}}\boldsymbol{k}(\boldsymbol{u}^{(k+1)})\right]^T\Sigma^{-1}\nabla_{\boldsymbol{u}}\boldsymbol{k}(\boldsymbol{u}^{(k+1)}) \tag{49}$$

The covariance of the estimator $\boldsymbol{u}^{(k+1)}$ is the inverse of the Fisher information matrix given in Equation (49).

As with other gradient optimization algorithms, a reasonable guess of the initial state value $\boldsymbol{u}^{(0)}$ is required to facilitate the optimization converges to the correct local minimum. In this application, the initial state $\boldsymbol{u}^{(0)}$ may be obtained from the RIPS measurements via the RIPS trilateration algorithm described in [29], which is then used to calculate $\boldsymbol{G}(\boldsymbol{u}^{(0)})$.

The performance of the proposed localization algorithm is illustrated by analyzing a scenario illustrated in Figure 6. In this example, three RIPS nodes are located at $(40, 50)$, $(70, 50)$, and $(60, 70)$ m. The noise of a RIPS measurement is assumed to be zero-mean Gaussian with a standard deviation $\sigma = 1$ m.

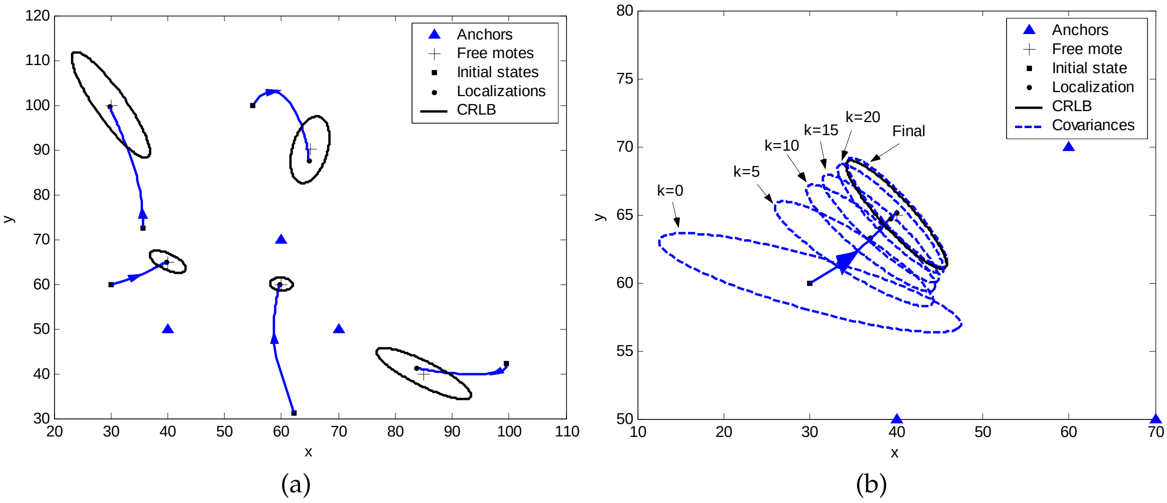


Figure 6. (**a**) Shows how the algorithm correctly localizes 5 unknown sensors ("+") from 3 motes ("Δ") with the iteration beginning at initial (guessed) locations; (**b**) Shows an example of the convergence of the iterative MLE estimator covariance to CRLB in the localization process.

Figure 6a shows 5 cases of localization results of the algorithm. The initial values are randomly generated in the simulations. We use the label CRLB to signify the ellipse which corresponds to the inverse of the Fisher information matrix of the network $\boldsymbol{G}(\boldsymbol{u})$, centered at the true location $\boldsymbol{u}$. Figure 6b shows the covariances of the iterative MLE estimator at different iterations, where the kth error ellipse is calculated using the inverse of $\boldsymbol{G}(\boldsymbol{u}^{(k)})$ in Equation (49) and is centered at $\boldsymbol{u}^{(k)}$. Figure 7a,b demonstrate the localization results of different initial state values with the same measurements and results based on a set of measurements with the same initial state values, respectively. The estimator performance under this scenario is summarized in Table 1.

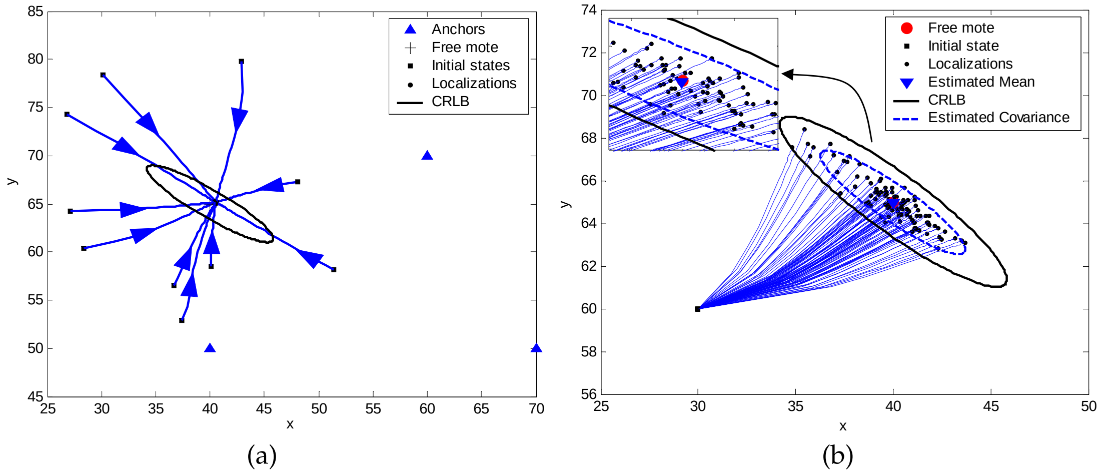


Figure 7. (**a**) An example of localization with the same three anchor motes as in Figure 6 for iteration beginning at 10 different initial locations; (**b**) Localization results for a set of measurements with the same initial state values.

Table 1. Estimator performance summary for the mote localization scenario shown in Figure 7b.

Measure	**Results**
Number of Monte Carlo runs	$N = 100$
Standard deviation of sensor noise	$\sigma = 2$ m
Location of the free mote	$\boldsymbol{u} = [40, 65]^T$
CRLB for estimating the state $\boldsymbol{u}$	$\boldsymbol{G}^{-1}(\boldsymbol{u}) = \begin{bmatrix} 24.4279 & -15.2115 \\ -15.2115 & 11.4530 \end{bmatrix}$
Sample mean of the estimator	$E\{\hat{\boldsymbol{u}}\} = \frac{1}{N}\sum_{n=1}^{N}\hat{\boldsymbol{u}}_n = [39.9776, 64.9758]^T$
Sample covariance of the estimator	$\mathrm{Cov}(\hat{\boldsymbol{u}}) = \frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{\boldsymbol{u}}_n - E\{\hat{\boldsymbol{u}}\}\right)\left(\hat{\boldsymbol{u}}_n - E\{\hat{\boldsymbol{u}}\}\right)^T = \begin{bmatrix} 2.9817 & -1.7683 \\ -1.7683 & 1.2706 \end{bmatrix}$
Average RMS location error	$\frac{1}{N}\sum_{n=1}^{N}\lvert\lvert\hat{\boldsymbol{u}}_n - \boldsymbol{u}\rvert\rvert = 1.5928$ m
Number of iterations M (ε—iteration stopping threshold) (λ—learning rate)	$M = 44$, when $\lambda = 0.1$, $\varepsilon = 0.01$ m $M = 25$, when $\lambda = 0.2$, $\varepsilon = 0.01$ m $M = 35$, when $\lambda = 0.2$, $\varepsilon = 0.001$ m

5. Conclusions

In this paper, an iterative maximum likelihood estimator based on the natural gradient method is described to address a class of nonlinear estimation problems for distributions of curved exponential families. We show that the underlying nonlinear stochastic filtering problem is solved by a natural gradient optimization technique which operates over statistical manifolds under dual affine connections. In this way, information geometry offers an interesting insight into the natural gradient algorithm and connects the stochastic estimation problem to a deterministic optimization problem. In this respect, the underlying philosophy is far more significant than the algorithm itself. Furthermore, based on an information geometric analysis it is promising that better algorithms for solving non-linear estimation problems can be derived. For instance, a "whitened gradient" which whitens the tangent space of a manifold has been presented in [30]. The whitened gradient replaces the Riemannian metric $\boldsymbol{G}(\boldsymbol{u})$ in the natural gradient updates by its square root $\boldsymbol{G}^{-\frac{1}{2}}(\boldsymbol{u})$ and results in a faster and more robust convergence.

The work in this paper indicates that the methods of differential/information geometry provide useful tools for systematically solving certain non-linear problems commonly encountered in signal processing. Future work involves extrapolation of these techniques to handle the filtering problem for nonlinear stochastic dynamics.

Acknowledgments: This work was supported in part by the U.S. Air Force Office of Scientific Research under grant No. FA9550-12-1-0418.

Author Contributions: Yongqiang Cheng put forward the original ideas and performed the research. Xuezhi Wang conceived and designed the application of the estimator to RIPS mote network localization. Bill Moran reviewed the paper and provided useful comments. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
2. Chentsov, N.N. *Statistical Decision Rules and Optimal Inference*; Leifman, L.J., Ed.; Translations of Mathematical Monographs; American Mathematical Society: Providence, RI, USA, 1982; Volume 53.
3. Efron, B. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Stat.* **1975**, *3*, 1189–1242.
4. Efron, B. The geometry of exponential families. *Ann. Stat.* **1978**, *6*, 362–376.

5. Amari, S. Differential geometry of curved exponential families-curvatures and information loss. *Ann. Stat.* **1982**, *10*, 357–385.
6. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Kobayashi, S., Takesaki, M., Eds.; Translations of Mathematical Monographs; American Mathematical Society: Providence, RI, USA, 2000; Volume 191.
7. Amari, S. Information geometry of statistical inference—An overview. In Proceedings of the IEEE Information Theory Workshop, Bangalore, India, 20–25 October 2002.
8. Smith, S.T. Covariance, subspace, and intrinsic Cramér–Rao bounds. *IEEE Trans. Signal Process.* **2005**, *53*, 1610–1630.
9. Srivastava, A. A Bayesian approach to geometric subspace estimation. *IEEE Trans. Signal Process.* **2000**, *48*, 1390–1400.
10. Srivastava, A.; Klassen, E. Bayesian and geometric subspace tracking. *Adv. Appl. Probab.* **2004**, *36*, 43–56.
11. Bhattacharya, R.; Patrangenaru, V. Nonparametric estimation of location and dispersion on Riemannian manifolds. *J. Stat. Plan. Inference* **2002**, *108*, 23–35.
12. Richardson, T. The geometry of turbo-decoding dynamics. *IEEE Trans. Inf. Theory* **2000**, *46*, 9–23.
13. Ikeda, S.; Tanaka, T.; Amari, S. Information geometry of turbo and low-density parity-check codes. *IEEE Trans. Inf. Theory* **2004**, *50*, 1097–1114.
14. Haenggi, M. A geometric interpretation of fading in wireless networks: Theory and application. *IEEE Trans. Inf. Theory* **2008**, *54*, 5500–5510.
15. Westover, M.B. Asymptotic geometry of multiple hypothesis testing. *IEEE Trans. Inf. Theory* **2008**, *54*, 3327–3329.
16. Li, Q.; Georghiades, C.N. On a geometric view of multiuser detection for synchronous DS/CDMA channels. *IEEE Trans. Inf. Theory* **2000**, *46*, 2723–2731.
17. Cheng, Y.; Wang, X.; Moran, B. Sensor network performance evaluation in statistical manifolds. In Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, 26–29 July 2010.
18. Cheng, Y.; Wang, X.; Caelli, T.; Li, X.; Moran, B. On information resolution of radar systems. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 3084–3102.
19. Wang, X.; Cheng, Y.; Moran, B. Bearings-only tracking analysis via information geometry. In Proceedings of the 13th International Conference on Information Fusion, Edinburgh, UK, 26–29 July 2010.
20. Wang, X.; Cheng, Y.; Morelande, M.; Moran, B. Bearings-only sensor trajectory scheduling using accumulative information. In Proceedings of the International Radar Symposium, Leipzig, Germany, 7–9 September 2011.
21. Altun, Y.; Hofmann, T.; Smola, A.J.; Hofmann, T. Exponential families for conditional random fields. In Proceedings of the 20th Annual Conference on Uncertainty in Artificial Intelligence, Banff, AB, Canada, 7–11 July 2004.
22. Amari, S.; Douglas, S.C. Why natural gradient? In Proceedings of the IEEE International Conference on Acoustics Speech Signal Process, Seattle, WA, USA, 12–15 May 1998.
23. Manton, J.H. On the role of differential geometry in signal processing. In Proceedings of the IEEE International Conference on Acoustics Speech Signal Process, Philadelphia, PA, USA, 18–23 March 2005.
24. Amari, S. Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **2001**, *47*, 1701–1711.
25. Brun, A.; Knutsson, H. Tensor glyph warping-visualizing metric tensor fields using Riemannian exponential maps. In *Visualization and Processing of Tensor Fields: Advances and Perspectives, Mathematics and Visualization*; Laidlaw, D.H., Weickert, J., Eds.; Springer: Berlin, Germany, 2009.
26. Maroti, M.; Kusy, B.; Balogh, G.; Volgyesi, P.; Molnar, K.; Nadas, A.; Dora, S.; Ledeczi, A. *Radio Interferometric Positioning*; Tech. Rep. ISIS-05-602; Institute for Software Integrated Systems, Vanderbilt University: Nashville, TN, USA, 2005.
27. Kusy, B.; Ledeczi, A.; Maroti, M.; Meertens, L. Node-density independent localization. In Proceedings of the 5th International Conference on Information Processing in Sensor Networks, Nashville, TN, USA, 19–21 April 2006.
28. Wang, X.; Moran, B.; Brazil, M. Hyperbolic positioning using RIPS measurements for wireless sensor networks. In Proceedings of the 15th IEEE International Conference on Networks (ICON2007), Adelaide, Australia, 19–21 November 2007.

29. Scala, B.F.; Wang, X.; Moran, B. Node self-localisation in large scale sensor networks. In Proceedings of the International Conference on Information, Decision and Control (IDC 2007), Adelaide, Australia, 11–14 February 2007.
30. Yang, Z.; Laaksonen, J. Principal whitened gradient for information geometry. *Neural Netw.* **2008**, *21*, 232–240.