Fisher Information Neural Estimation

Tran Trong Duy*, Ly V. Nguyen*, Viet-Dung Nguyen*†, Nguyen Linh Trung*, and Karim Abed-Meraim‡§

*AVITECH Institute, VNU University of Engineering and Technology, Hanoi, Vietnam
†National Institute of Advanced Technologies of Brittany, Brest, France
‡PRISME Laboratory, University of Orléans, Orléans, France
§Academic Institute of France (IUF), 1 rue Descartes, 75005 Paris, France
Emails: {duytt, lynv, linhtrung}@vnu.edu.vn, viet.nguyen@ensta-bretagne.fr, karim.abed-meraim@univ-orleans.fr

***Abstract*—Fisher information is a fundamental quantity in information theory and signal processing. A direct analytical computation of the Fisher information is often infeasible or intractable due to the lack or sophistication of statistical models. In this paper, we propose a Fisher Information Neural Estimator (FINE) which is computationally efficient, highly accurate, and applicable for both cases of deterministic and random parameters. The proposed method solely depends on measured data and does not require knowledge or an estimate of the probability density function and is therefore universally applicable. We validate our approach using some experiments and compare with existing works. Numerical results show the high efficacy and low-computational complexity of the proposed estimation approach.**

I. Introduction

Fisher information is a well-known and well-defined concept in mathematical statistics, which is defined as a measure of the amount of information that a random variable carries about some unknown parameters. In estimation theory, the inverse of the Fisher information directly gives a well-known lower bound called Cramér-Rao bound (CRB) on the variance of any unbiased estimator of the unknown parameters. There are many other areas in which the Fisher information is applied to, e.g., Bayesian statistics, frequentist statistics, optimal experimental design, computational neuroscience, physical laws, biology, and machine learning [1]–[3].

Analytically, a closed-form expression of the Fisher information matrix (FIM) might be obtained by taking the expectation of the Hessian matrix of the log likelihood function (the score function). Unfortunately, such a straightforward computation is often impossible due to unknown statistical models. Even in circumstances where the statistical model is available, a closed-form expression of the FIM can still be intractable due to model sophistication. This difficulty raises the significance of developing FIM estimation methods.

The estimation of the FIM can be divided into two categories: plug-in and non-plug-in. In the plug-in category, the strategy is to first estimate the probability density function (pdf) based on the observed data and then use the pdf estimate for a numerical computation of the FIM. For example, in [4], Spall proposed a Monte Carlo resampling-based (MCR) method for FIM estimation. The MCR method first performs the pdf estimation for each of the perturbed experiments and then numerically computes the gradient of the log density function before sample averaging. Another existing plug-in method for FIM estimation was introduced in [5], which also estimates the pdf using the observed data and then obtains the derivatives of the pdf based on finite-difference approximation. Unlike the plug-in methods, the strategy of non-plug-in methods is to directly estimate the FIM based on the observed data. This non-plug-in strategy is particularly suitable for circumstances where the system is a black box whose operating parameters are tunable, e.g. controlled experiments [6]–[8]. One can observe data from the system for various settings of the parameters. An example of non-plug-in FIM estimation methods is in [9], which is based on a relation between the f-divergence and the FIM.

The strategy of plug-in methods is straightforward, but an accurate estimate of the pdf may not always be possible or is very difficult to obtain in scenarios where the underlying pdf is sophisticated. Non-plug-in methods do not rely on pdf estimation since the FIM is directly estimated from the observed data, and thus they are relieved of the difficulties in pdf estimation. However, existing non-plug-in methods like the one in [9] often suffer from problems of having a high computational complexity, requiring very large data sets for accurate estimation, or being specifically developed for systems where the operating parameters are deterministic.

Motivated by the above discussion and a recently developed mutual information estimation method in [10], we propose a non-plug-in FIM estimator, referred to as Fisher Information Neural Estimator (FINE), which has several advantages such as having a low computational complexity, high estimation accuracy, and applicable for both cases of deterministic and random parameters. It should also be noted that FINE employs neural networks and thus takes advantages of their nice properties such as the ability to learn non-linear and complex relationships and having low computational complexities.

The contributions of this paper are summarized as follows. First, we propose FINE – a Fisher information estimator based on neural networks for the case of deterministic parameters. The proposed FINE is based on a relation between the Fisher information and the f-divergence, which was exploited in

Correspondence: Nguyen Linh Trung (linhtrung@vnu.edu.vn). This work was funded by the National Foundation for Science and Technology Development (NAFOSTED) of Vietnam under grant number 01/2019/TN.

a previous work [9]. However, unlike [9] which computed the f-divergence by using the minimal spanning tree (MST) and the Friedman-Rafsky (FR) statistic, FINE computes the f-divergence by neural networks. Compared to [9], FINE has not only higher estimation accuracy but also a lower computational complexity. Second, we show that the proposed FINE framework can be used for the case of random parameters, i.e., FINE is applicable for the Bayesian Fisher information estimation problem. We prove that the relation between the Bayesian Fisher information matrix (B-FIM) and the f-divergence follows an expression that is similar to the case of deterministic parameters. To validate the efficacy of the proposed FINE in the Bayesian framework, we carry out some simulations about dynamical phase offset estimation in a communication system. Numerical results show that the proposed FINE gives a better estimation accuracy compared to an existing asymptotic bound.

II. Background, Problem Statement, and Related Work

A. Fisher Information and f-divergence

1) Fisher Information: Consider a random variable X whose pdf $p(x|\boldsymbol{\theta})$ is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$, a vector of d unknown parameters. When $\boldsymbol{\theta}$ is deterministic, the FIM $\mathbf{F}(\boldsymbol{\theta})$ is defined as follows [11]:

$$\begin{aligned}\mathbf{F}(\boldsymbol{\theta}) &= \mathbb{E}_{X|\boldsymbol{\theta}}\left[\big(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})\big)\big(\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})\big)^T\right] \\ &= -\mathbb{E}_{X|\boldsymbol{\theta}}\left[\mathbf{H}_{\boldsymbol{\theta}}\big(\log p(x|\boldsymbol{\theta})\big)\right], \end{aligned} \tag{1}$$

where $\nabla_{\boldsymbol{\theta}} \log p(x|\boldsymbol{\theta})$ and $\mathbf{H}_{\boldsymbol{\theta}}\big(\log p(x|\boldsymbol{\theta})\big)$ respectively denote the gradient and the Hessian matrix of the score function $\log p(x|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In case the parameter vector $\boldsymbol{\theta}$ is random, the B-FIM $\mathbf{B}$ is used instead [12], [13]

$$\begin{aligned}\mathbf{B} &= \mathbb{E}_{X,\boldsymbol{\theta}}\left[\big(\nabla_{\boldsymbol{\theta}} \log p(x,\boldsymbol{\theta})\big)\big(\nabla_{\boldsymbol{\theta}} \log p(x,\boldsymbol{\theta})\big)^T\right] \\ &= -\mathbb{E}_{X,\boldsymbol{\theta}}\left[\mathbf{H}_{\boldsymbol{\theta}}\big(\log p(x,\boldsymbol{\theta})\big)\right], \end{aligned} \tag{2}$$

where $p(x,\boldsymbol{\theta})$ is the joint pdf of X and $\boldsymbol{\theta}$.

2) f-divergence: For any convex function f such that $f(1) = 0$, the f-divergence between two probability distributions $p(x)$ and $q(x)$ is defined as a function $D_f(p\|q)$ that measures the difference between $p(x)$ and $q(x)$ [14]:

$$D_f(p\|q) = \mathbb{E}_q\left[f\left(\frac{p(x)}{q(x)}\right)\right] = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx. \tag{3}$$

The Kullback-Leibler (KL) divergence is a special case of the f-divergence where $f(t) = t\log(t)$ and is given as

$$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right] = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx. \tag{4}$$

3) Relation Between Fisher Information and f-divergence: For notational simplicity, let $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\eta}}$ denote the probability distribution of a random variable X parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\eta} = \boldsymbol{\theta} + \boldsymbol{\delta}$, respectively. Here, $\boldsymbol{\delta}$ is a small perturbation around $\boldsymbol{\theta}$. This means $p_{\boldsymbol{\theta}} = p(x|\boldsymbol{\theta})$ and $p_{\boldsymbol{\eta}} = p(x|\boldsymbol{\theta}+\boldsymbol{\delta})$. The relation between the Fisher information $\mathbf{F}(\boldsymbol{\theta})$ and the f-divergence between $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}+\boldsymbol{\delta}}$ is given in a quadratic form as [9], [15]

$$D_f(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\eta}}) \approx \frac{1}{2}\boldsymbol{\delta}^T \mathbf{F}(\boldsymbol{\theta})\boldsymbol{\delta}. \tag{5}$$

The above relation can be obtained by applying the Taylor expansion to the f-divergence. This relation indicates that if $D_f\big(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\eta}}\big)$ can be computed for at least $d(d+1)/2$ different perturbations $\boldsymbol{\delta}$, then the FIM $\mathbf{F}(\boldsymbol{\theta})$ can be obtained by solving (5). This is due to the fact that $\mathbf{F}(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ is a symmetric matrix and thus contains $d(d+1)/2$ different elements. It should be noted that the relation in (5) is for the FIM. One of our contributions is to prove that the relation between the B-FIM and the f-divergence follows an expression that is similar to (5).

B. Problem Statement and Related Work

1) Problem Statement: Consider a random variable X whose pdf $p(x|\boldsymbol{\theta})$ is unknown. The parameters $\boldsymbol{\theta}$ can be either deterministic or random. We assume the parameters are tunable in the sense that they can be perturbed by a small deviation $\boldsymbol{\delta}$. The problem is to estimate the FIM $\mathbf{F}(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is deterministic or the B-FIM $\mathbf{B}$ when $\boldsymbol{\theta}$ is random using the data samples of X.

2) Related Work: In [9], Berisha and Hero exploited the relation in (5) to propose a non-plug-in Fisher information estimator based on the FR statistic [16]. Specifically, the method in [9] computes $D_f(p_{\boldsymbol{\theta}}\|p_{\boldsymbol{\eta}})$ using the MST for M different perturbations of $\boldsymbol{\delta}$ where $M \geq d(d+1)/2$. Then, the FIM $\mathbf{F}(\boldsymbol{\theta})$ is obtained by solving (5) based on the M computed divergence values.

III. Proposed FINE

A. Deterministic Parameters

Here, we consider deterministic parameters $\boldsymbol{\theta}$ and we want to estimate the FIM $\mathbf{F}(\boldsymbol{\theta})$. TFINE also exploits the relation between the Fisher information and the f-divergence in (5) but unlike the method in [9] which computes the f-divergence by using the FR statistic and the MST, the proposed FINE employs neural networks to compute the f-divergence. As will be shown later, the use of neural networks not only helps improve the estimation accuracy but also significantly reduces the computational complexity.

Our motivation for computing the f-divergence by neural networks comes from a recently developed mutual information neural estimation method in [10], which is referred to as MINE. Specifically, the mutual information between two random variables Y and Z is given as

$$\begin{aligned} I(Y,Z) &= H(Y) - H(Y \mid Z) && (6)\\ &= D_{\mathrm{KL}}(P_{YZ}\|P_Y P_Z) && (7)\\ &= \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{P_{YZ}}[T] - \log\big(\mathbb{E}_{P_Y P_Z}[e^T]\big) && (8)\\ &\geq \sup_{T\in\mathcal{F}} \mathbb{E}_{P_{YZ}}[T] - \log\big(\mathbb{E}_{P_Y P_Z}[e^T]\big) && (9)\end{aligned}$$

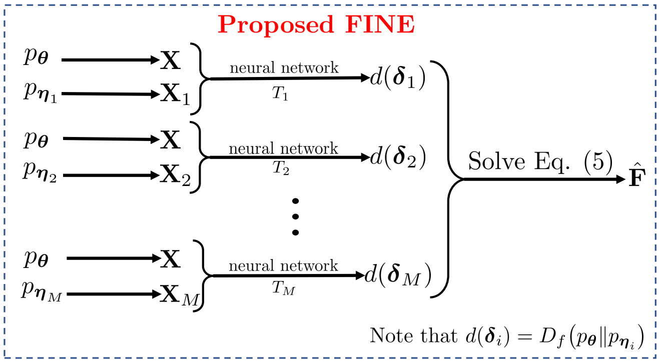

Proposed FINE
$p_{\boldsymbol{\theta}}$
$\mathbf{X}$
$p_{\boldsymbol{\eta}_1}$
$\mathbf{X}_1$
neural network
T_1
$d(\boldsymbol{\delta}_1)$
$p_{\boldsymbol{\theta}}$
$\mathbf{X}$
$p_{\boldsymbol{\eta}_2}$
$\mathbf{X}_2$
neural network
T_2
$d(\boldsymbol{\delta}_2)$
Solve Eq. (5)
$\hat{\mathbf{F}}$
$p_{\boldsymbol{\theta}}$
$\mathbf{X}$
$p_{\boldsymbol{\eta}_M}$
$\mathbf{X}_M$
neural network
T_M
$d(\boldsymbol{\delta}_M)$
Note that $d(\boldsymbol{\delta}_i) = D_f(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\eta}_i})$

Fig. 1: Illustration of the proposed FINE method.

where $\mathcal{F}$ is any class of functions $T : \Omega \to \mathbb{R}$ satisfying the integrability constraints of the Donsker-Varadhan representation [10], and Ω is a compact domain that the two distributions P_{YZ} and $P_Y P_Z$ belong to. Here, $I(\cdot)$ and $H(\cdot)$ denote the mutual information and the entropy, respectively. Note that the equality in (8) is the Donsker-Varadhan representation. Belghazi et al. utilize a well-known property of neural networks stated as the universal approximation theorem. The idea in [10] is to treat T as a neural network and the mutual information $I(Y, Z)$ is estimated by using the observed data to train T such that $\mathbb{E}_{P_{YZ}}[T] - \log\left(\mathbb{E}_{P_Y P_Z}[e^T]\right)$ is maximized. After training, the trained objective function value reads an estimate of the mutual information. Note that the method of MINE can be applied to estimate a general f-divergence [17].

Our idea is to compute $D_f(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\eta}})$ in (5) by using the observed data to train a neural network T such that $\mathbb{E}_{p_{\boldsymbol{\theta}}}[T] - \log\left(\mathbb{E}_{p_{\boldsymbol{\eta}}}[e^T]\right)$ is maximized. Since $\mathbf{F}(\boldsymbol{\theta})$ contains $d(d+1)/2$ different elements, we need to compute $D_f(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\eta}})$ for M different perturbations of $\boldsymbol{\delta}$ where $M \geq d(d+1)/2$. Then we can obtain an estimate of $\mathbf{F}(\boldsymbol{\theta})$ by solving (5).

Let $\boldsymbol{\eta}_i = \boldsymbol{\theta} + \boldsymbol{\delta}_i$ with $i = 1, \ldots, M$ and let $\mathbf{X} \in \mathbb{R}^{N \times K}$ and $\mathbf{X}_i \in \mathbb{R}^{N \times K}$ denote the data sets observed from $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\eta}_i}$, respectively. Here, N and K are the number of data samples and the size of each sample, respectively. A neural network T_i is used to estimate the f-divergence $d(\boldsymbol{\delta}_i) = D_f\left(p_{\boldsymbol{\theta}} \| p_{\boldsymbol{\eta}_i}\right)$ based on $\mathbf{X}$ and $\mathbf{X}_i$. Specifically, T_i takes a vector of size K as the input and returns a scalar as the output. Thus, $\mathbf{X}$ and $\mathbf{X}_i$ are the input data sets of T_i. Let $\mathbf{z} = T_i(\mathbf{X}) \in \mathbb{R}^N$ and $\mathbf{z}_i = T_i(\mathbf{X}_i) \in \mathbb{R}^N$. Then T_i is trained to maximize $\mathbf{1}^T \mathbf{z}/N - \log(\mathbf{1}^T \exp\{\mathbf{z}_i\}/N)$, which is used as an estimate of $d(\boldsymbol{\delta}_i)$. For notational simplicity, we use $\exp\{\mathbf{z}_i\}$ to indicate that $\exp\{\cdot\}$ is applied to $\mathbf{z}_i$ element-wise. An illustration of the proposed FINE is given in Figure 1.

Once the f-divergences have been obtained, we need to solve (5) for $\mathbf{F}(\boldsymbol{\theta})$ under the constraint that $\mathbf{F}(\boldsymbol{\theta})$ is a symmetric positive semi-definite (PSD) matrix. We vectorize $\mathbf{F}(\boldsymbol{\theta})$ by including the distinct upper triangular values of $\mathbf{F}(\boldsymbol{\theta})$ and convert (5) to a linear function of this quantity. Let

$$\mathbf{f} = [F_{11}, \ldots, F_{dd}, F_{12}, \ldots, F_{1d}, F_{23}, \ldots, F_{2d}, \ldots, F_{(d-1)d}]^T$$

where F_{ij} is the element in the i-row and j-column of $\mathbf{F}(\boldsymbol{\theta})$, and let

$$\mathbf{u}_i = [\delta_{i1}^2, \ldots, \delta_{id}^2, 2\delta_{i1}\delta_{i2}, \ldots, 2\delta_{i1}\delta_{id}, \ldots, 2\delta_{i(d-1)}\delta_{id}]^T,$$

where $i = 1, \ldots, M$, δ_{ij} is the the j-element of $\boldsymbol{\delta}_i$. Denote $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_M]^T$, then we have a linear system $2\mathbf{d} = \mathbf{U}\mathbf{f}$ where $\mathbf{d} = [d(\boldsymbol{\delta}_1), \ldots, d(\boldsymbol{\delta}_M)]^T$. Using the least square (LS) estimator, we can find an estimate of $\mathbf{f}$ as

$$\hat{\mathbf{f}}^{\text{LS}} = 2\left(\mathbf{U}^T\mathbf{U}\right)^{-1}\mathbf{U}^T\mathbf{d}. \tag{10}$$

This LS estimator, however, does not ensure that the resulting estimate is positive semi-definite. So we employ a semi-definite program (SDP) as follows [9]:

$$\begin{aligned} \underset{\mathbf{f}}{\text{minimize}} \quad & \|2\mathbf{d} - \mathbf{U}\mathbf{f}\|_2^2 \\ \text{subject to} \quad & f_k = \hat{f}_k^{\text{LS}},\ k = 1, \ldots, d \\ & \text{mat}(\mathbf{f}) = \mathbf{F}(\boldsymbol{\theta}) \succeq \mathbf{0}. \end{aligned} \tag{11}$$

where f_k and $\hat{f}_k^{\text{LS}}$ are the k-th element of $\mathbf{f}$ and $\hat{\mathbf{f}}^{\text{LS}}$, respectively. The $\text{mat}(\cdot)$ operator converts the vectorized FIM $\mathbf{f}$ to a full matrix representation $\mathbf{F}(\boldsymbol{\theta})$. To ensure the symmetric PSD requirement, we only need to refine the off-diagonal elements of $\mathbf{F}(\boldsymbol{\theta})$, which explains the constrains $f_k = \hat{f}_k^{\text{LS}}$, $k = 1, \ldots, d$.

B. Random Parameters

In many systems, the operating parameters $\boldsymbol{\theta}$ are not deterministic, but random. This leads to the study of the B-FIM $\mathbf{B}$. Let $\pi(\boldsymbol{\theta})$ be the distribution of $\boldsymbol{\theta}$, then

$$\begin{aligned} \mathbf{B} &= -\mathbb{E}_{X,\boldsymbol{\theta}}\left[\mathbf{H}_{\boldsymbol{\theta}}\left(\log p(x, \boldsymbol{\theta})\right)\right] \\ &= -\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{H}_{\boldsymbol{\theta}}\left(\log p(\boldsymbol{\theta})\right)\right] - \mathbb{E}_{\boldsymbol{\theta}X}\left[\mathbf{H}_{\boldsymbol{\theta}}\left(\log p(x|\boldsymbol{\theta})\right)\right] \\ &= \mathbf{F}(\pi) + \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{F}(\boldsymbol{\theta})\right]. \end{aligned} \tag{12}$$

As can be seen from (12) that the B-FIM $\mathbf{B}$ does not depend on a particular value of $\boldsymbol{\theta}$ and consists of two terms. The first term is the information of the prior distribution and the second term is the expected Fisher information. Some closely-related studies about the Bayesian Carmér-Rao bound (BCRB) were presented in [18] and [19, Chapter 7]. To the best of our knowledge, the relation between the f-divergence and the Bayesian Fisher information has not been stated formally in the literature. Here, we show that the relation between the B-FIM and the f-divergence follows an expression that is similar to the one in (5) by Theorem 1 below.

Theorem 1. *Consider a distribution $P_{X\theta}$ with the pdf $p(x, \theta + \delta)$ and another distribution $Q_{X\theta}$ with the pdf $p(x, \theta)$, where $\theta \in \Theta \subset \mathbb{R}$ is a random parameter, and $\delta \in \mathbb{R}$ is a small perturbation. For any convex function f satisfying $f(1) = 0$ and $f''(1) = 1$, the f-divergence between $P_{X\theta}$ and $Q_{X\theta}$*

$$D_f(P_{X\theta} || Q_{X\theta}) = \iint f\left(\frac{p(x, \theta + \delta)}{p(x, \theta)}\right) p(x, \theta) dx d\theta \tag{13}$$

can be approximated as follows:

$$D_f(P_{X\theta} || Q_{X\theta}) \approx \frac{1}{2}\delta^2 B \tag{14}$$

where B is the Bayesian information and defined as

$$B = \iint \left(\frac{\partial \log p(x,\theta)}{\partial \theta} \right)^2 p(x,\theta) dx d\theta. \tag{15}$$

Proof. Using the Taylor expansion about θ, we have

$$p(x, \theta + \delta) = p(x, \theta) + p'(x, \theta)\delta + o(\delta), \tag{16}$$

and thus

$$D_f(P_{X\theta} || Q_{X\theta}) = \iint f \left(1 + \frac{p'(x,\theta)\delta}{p(x,\theta)} \right) p(x,\theta) dx d\theta. \tag{17}$$

Using the Taylor expansion of f about 1, we have

$$f(1+\Delta) = f(1) + f'(1)\Delta + \frac{1}{2} f''(1)\Delta^2 + o(\Delta^2) \approx \frac{1}{2}\Delta^2. \tag{18}$$

The approximation in (18) is obtained since $f(1) = 0$, $f'(1) = 0$, and $f''(1) = 1$. We can always have $f'(1) = 0$ because $D_{f_c}(P\|Q) = D_f(P\|Q)$ where $f_c(t) = f(t) - c(t-1)$, which means if $f(t)$ does not satisfy $f'(1) = 0$, we can replace $f(t)$ by $f_c(t)$ with $c = f'(1)$. Applying (18) to (17), we obtain

$$D_f(P_{X\theta} || Q_{X\theta}) \approx \frac{\delta^2}{2} \iint \left(\frac{p'(x,\theta)}{p(x,\theta)} \right)^2 p(x,\theta) dx d\theta = \frac{\delta^2}{2} B$$

□

Although Theorem 1 is stated for one dimensional parameters, for higher dimensional parameters $\boldsymbol{\theta}$, one can use the same reasoning and obtain $D_f(P_{X\boldsymbol{\theta}} || Q_{X\boldsymbol{\theta}}) \approx \frac{1}{2}\boldsymbol{\delta^T}\mathbf{B}\boldsymbol{\delta}$. Hence, the relation between the f-divergence and the Bayesian Fisher information follows an expression similar to the case of deterministic parameters in (5). Therefore, FINE can be directly applied to this Bayesian framework. However, it should be noted that the data samples in this Bayesian framework are generated from the joint distributions of X and $\boldsymbol{\theta}$.

IV. Complexity Analysis and Test Results

A. Complexity Analysis

The complexity of FINE mostly depends on the time for training the neural networks, which is $\mathcal{O}(JWN)$ [20] where J, W, and N are the number of epochs, the number of trainable parameters, and the number of training samples, respectively. For reliable estimation, N often needs to be large, and so JW is relatively small compared to N, making the complexity of FINE roughly $\mathcal{O}(N)$. Compared to the empirical estimator proposed by Berisha and Hero [9], their method needs to construct the MSTs of dense graphs whose complexity is $\mathcal{O}(N^2)$. In the following, we will show that the run time of FINE is significantly lower than the run time of the estimator proposed in [9].

B. Test Results

Here we numerically validate and evaluate the performance of FINE for both cases of deterministic and random parameters. We used neural networks with only one hidden layer whose width is five times the width of the input layer and the ReLU activation function is used.

1) Deterministic Parameters: Consider a K-dimensional Gaussian distribution as $\mathcal{N}(\boldsymbol{\theta}, \mathbf{I}_K)$. The objective is to estimate the FIM $\mathbf{F}$ at $\boldsymbol{\theta} = \mathbf{0}$. It should be noted that $d = K$ in this scenario. The FIM has a simple closed form as $\mathbf{F}_{\text{true}} = \mathbf{I}$. Each element of $\boldsymbol{\delta}$ is drawn from $\mathcal{N}(0, 0.05)$. We use the same data size of N for all data sets $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_M$. We set $K = 4$ and $M = 5d(d+1)/2$. The proposed FINE is compared with the method proposed by Berisha and Hero in [9]. Test results are given in Figure 2. The normalized mean squared error (NMSE) is defined as $\text{NMSE} = \|\hat{\mathbf{F}} - \mathbf{F}_{\text{true}}\|_{\text{F}}^2/d^2$. It can be seen that the proposed FINE outperforms the method in [9] in terms of both accuracy and computational complexity. The run time of the method by Berisha and Hero scales quadratically with N whereas the run time of the proposed FINE decreases because it is found that the neural network converged faster with larger data sets.

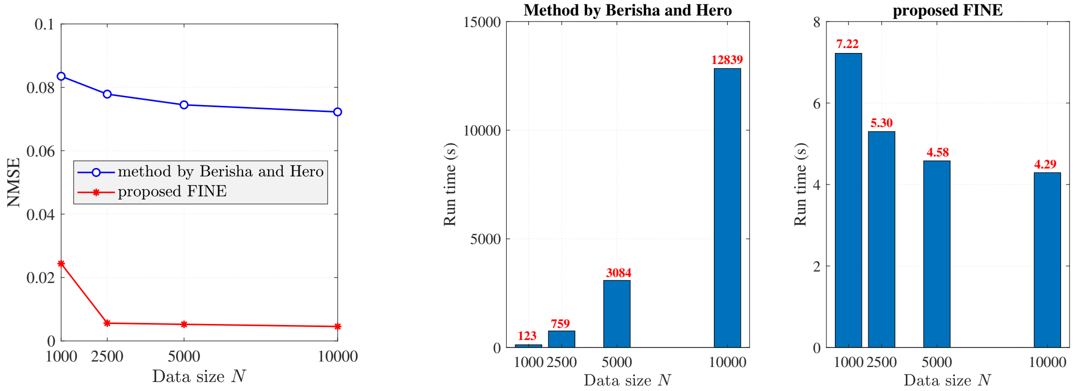
2) Random Parameters: Consider the transmission of L BPSK symbols $\boldsymbol{a} = [a_1, \ldots, a_L]^T$ over an additive white Gaussian noise (AWGN) channel affected by carrier phase offsets $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_L]^T$, the received signal is given as

$$y_l = a_l e^{j\theta_l} + n_l,$$

where n_l is the additive white Gaussian noise and distributed as $\mathcal{N}(0, \sigma_n^2)$ and $\boldsymbol{\theta}$ follows the Wiener phase-offset evolution, i.e., $\theta_l = \theta_{l-1} + w_l$, where $w_l \sim \mathcal{N}(0, \sigma_w^2)$. The receiver needs to estimate the carrier-phase offsets $\boldsymbol{\theta}$. In [21], a closed-form BCRB was derived for this scenario. In addition, the authors also proposed an asymptotic bound, which is referred to as ABCRB. We adopt the observation block length $L = 4$ and various values of σ_w^2. Comparison results are shown in Figure 3. It is observed that with a low value of σ_w^2, both the estimated BRCB (by FINE) and ABCRB are close to the true BCRB. Figures 3b and 3c show that our proposed FINE produces remarkable improvements over the ABCRB when σ_w^2 is higher. Thus, the results in Fig. 3 verify the efficacy of the proposed FINE in case of random parameters.

V. Conclusion

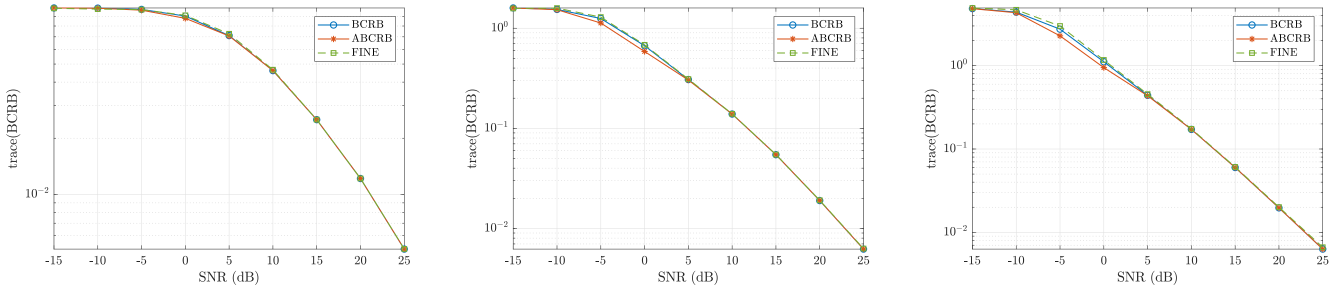
This paper proposes a new method to estimate the Fisher information based on neural networks. The proposed FINE approach has a low-computational complexity, high estimation accuracy, and is applicable for both cases of deterministic and random parameters. Numerical test results show that the proposed FINE not only performs the estimation task faster but also yields more accurate estimates compared to existing works. For example, the computational complexity of the proposed FINE is roughly linear with the data size, whereas the computational complexity of an existing method in [9] scales quadratically with the data size. Another important advantage of the proposed approach is that it can be applied to systems where the underlying statistical model is unknown or at a high level of sophistication since only observed data samples are needed in the estimation process.



(a) NMSE.

(b) Run time.

Fig. 2: Estimation accuracy and computational complexity comparison for the case of deterministic $\boldsymbol{\theta}$.



(a) $\sigma_w^2 = 0.1^2$

(b) $\sigma_w^2 = 0.4^2$

(c) $\sigma_w^2 = 0.7^2$

Fig. 3: Validation of FINE for the case of random $\boldsymbol{\theta}$.

References

[1] A. Hannun, C. Guo, and L. van der Maaten, "Measuring data leakage in machine-learning models with Fisher information," in *Proc. Uncertainty in Artificial Intelligence*, vol. 161, 2021, pp. 760–770.

[2] A. Ly, M. Marsman, J. Verhagen, R. P. Grasman, and E.-J. Wagenmakers, "A tutorial on Fisher information," *Journal of Mathematical Psychology*, vol. 80, pp. 40–55, Oct. 2017.

[3] S. A. Frank, "Natural selection maximizes Fisher information," *Journal of Evolutionary Biology*, vol. 22, no. 2, pp. 231–244, Jan. 2009.

[4] J. C. Spall, "Monte Carlo computation of the Fisher information matrix in nonstandard settings," *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 889–909, Jan. 2005.

[5] O. Har-Shemesh, R. Quax, B. Minano, A. G. Hoekstra, and P. M. Sloot, "Nonparametric estimation of Fisher information from real data," *Physical Review E*, vol. 93, no. 2, p. 023301, Feb. 2016.

[6] L. Gilbertson and R. A. Lutfi, "Correlations of decision weights and cognitive function for the masked discrimination of vowels by young and old adults," *Hearing Research*, vol. 317, pp. 9–14, 2014.

[7] S. Nelander, W. Wang, B. Nilsson, Q.-B. She, C. Pratilas, N. Rosen, P. Gennemark, and C. Sander, "Models from experiments: Combinatorial drug perturbations of cancer cells," *Molecular Systems Biology*, vol. 4, no. 1, p. 216, 2008.

[8] R. C. Jansen, "Studying complex biological systems using multifactorial perturbation," *Nature Reviews Genetics*, vol. 4, no. 2, pp. 145–151, 2003.

[9] V. Berisha and A. O. Hero, "Empirical non-parametric estimation of the Fisher information," *IEEE Signal Process. Letters*, vol. 22, no. 7, pp. 988–992, July 2015.

[10] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proc. of the 35th International Conference on Machine Learning*, vol. 80, Jul. 2018, pp. 531–540.

[11] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical transactions of the Royal Society of London. Series A*, vol. 222, no. 594-604, pp. 309–368, 1922.

[12] H. L. van Trees, *Detection, Estimation and Modulation Theory Part I.* New York, NY, USA: John Wiley & Sons, 1968.

[13] R. D. Gill and B. Y. Levit, "Applications of the van Trees Inequality: A Bayesian Cramér-Rao Bound," *Bernoulli*, vol. 1, no. 1/2, pp. 59–79, 1995. [Online]. Available: http://www.jstor.org/stable/3318681

[14] A. Rényi, "On measures of entropy and information," in *Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 4, 1961, pp. 547–562.

[15] S.-i. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, Mar. 2010.

[16] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *The Annals of Statistics*, pp. 697–717, 1979.

[17] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Information Theory*, vol. 56, no. 11, pp. 5847–5861, Nov. 2010.

[18] M. A. Kumar and K. V. Mishra, "Information geometric approach to bayesian lower error bounds," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, June 2018, pp. 746–750.

[19] Y. Wu, "Lecture notes on information-theoretic methods for high-dimensional statistics," Yale University, USA, Jan. 2020. [Online]. Available: http://www.stat.yale.edu/∼yw562/teaching/it-stats.pdf

[20] E. Mizutani and S. E. Dreyfus, "On complexity analysis of supervised MLP-learning for algorithmic comparisons," in *Proc. International Joint Conference on Neural Networks*, vol. 1, 2001, pp. 347–352.

[21] S. Bay, C. Herzet, J.-M. Brossier, J.-P. Barbot, and B. Geller, "Analytic and asymptotic analysis of Bayesian Cramér–Rao bound for dynamical phase offset estimation," *IEEE Trans. Signal Process.,*, vol. 56, pp. 61–70, Feb. 2008.