symmetry

MDPI

Article

Double Contingency of Communications in Bayesian Learning

Atsuhide Mori

Department of Mathematics, Osaka Dental University, Osaka 573-1121, Japan; mori-a@cc.osaka-dent.ac.jp

Abstract: In previous work, we described the geometry of Bayesian learning on a manifold. In this paper, inspired by the notion of modified double contingency of communications from sociologist Niklas Luhmann, we take two manifolds in equal parts and a potential function on their product to set up mutual Bayesian learning. Particularly, given a parametric statistical model, we consider mutual learning between two copies of the parameter space. Here, we associate the potential with the relative entropy (i.e., the Kullback–Leibler divergence). Although the mutual learning forgets all elements about the model except the relative entropy, it still substitutes for the usual Bayesian estimation of the parameter in a certain case. We propose it as a globalization of the information geometry.

Keywords: Bayesian learning; relative entropy; Kullback–Leibler divergence; information geometry; double contingency; communication; autopoiesis; cybernetics

Citation: Mori, A. Double Contingency of Communications in Bayesian Learning. *Symmetry* **2022**, *14*, 2456. https://doi.org/10.3390/sym14112456

Academic Editors: Yanlin Li, Tiehong Zhao and Sergei D. Odintsov

Received: 16 September 2022
Accepted: 15 November 2022
Published: 19 November 2022

1. Introduction

This is the sequel of the author's research [1] on the geometry of Bayesian learning. We introduce mutual Bayesian learning by taking two manifolds, each of which is the parameter space of a family of density functions on the other. This setting has the following background in sociology that seems more ideological than practical.

Talcott Parsons [2] introduced the notion of double contingency in sociology. Here, the contingency is that no event is necessary and no event is impossible. A possible understanding of this definition appeals to probability theory. Specifically, even an event with probability $P = 1$ does not always occur, and even that with $P = 0$ sometimes occurs, as a non-empty null set appears, at least conceptually. We consider the contingency as the subjective probabilistic nature of society. In fact, updating the conceptual subjective probability according to Bayes' rule should be a response to the conventional contingency that the prior probability is not a suitable predictor in reality. However, the double contingency is not straightforward, as it concerns mutually dependent social actions. In this article, we describe the double contingency by means of Bayesian learning. In our description, when one learns from another, the opposite learning also proceeds. This implies that, in contrast to sequential games such as chess, the actions in a double contingency have to be selected at once. Niklas Luhmann [3] leveraged this simultaneity to regard people not as individuals but as a single agent that he called a system. This further enabled him to apply the double contingency to any communications between systems. We introduce a function λ on the product of two manifolds to understand his systems theory.

From a practical perspective, we consider a family $\{h_x : W \to \mathbb{R}_{>0}\}_{x \in X}$ of probability density on a manifold W and regard the parameter space X as a manifold. The product $X \times X$ carries the function $\varphi : X \times X \to \mathbb{R}_{\geq 0}$ induced from the relative entropy. Recall that the information geometry [4] is a differential geometry on the diagonal set $\Delta \subset X \times X$, which deals with the 3-jet of φ at Δ. The author [5] began exploring the global geometry of $(X \times X, \varphi)$. Now we take $\exp(-\varphi)$ as the above function λ, and show that the mutual Bayesian learning between two copies of X substitutes for the original Bayesian estimation on W in a certain case. We notice that the global geometry of φ, as well as the information

geometry, forgets the original problem on W, and addresses a related problem on X. In this regard, our mutual Bayesian learning is a globalization of the information geometry.

2. Mathematical Formulation

2.1. Geometric Bayesian Learning

We work in the C^∞-smooth category. Take a possibly non-compact and possibly disconnected manifold X equipped with a volume form $dvol_X$. Note that a discrete set is a 0-dimensional manifold on which a positive function is a volume form. Suppose that each point x of the manifold X presents a possible action of a person. A positive function $f : X \to \mathbb{R}_{>0}$ on X is called a density. If its integral $|f|_{dvol_X} := \int_X f dvol_X$ is finite, it defines the probability $f/|f|_{dvolX}$ on X. Suppose that the selection of an action x is weighted by a density f_0 on X. In our story, the person believes that a density $\rho_x : Y \to \mathbb{R}_{>0}$ on another manifold $(Y, dvol_Y)$ depends on his action x. That is why the person perceives a given point $y_0 \in Y$ by multiplying the density f_0 by the function

$$l : X \to \mathbb{R}_{>0} : X \ni x \mapsto \rho_x(y_0) > 0,$$

which is called the likelihood of the datum $y_0 \in Y$. The perception updates the prior density f_0 to the posterior density $f_1(x) := l(x)f_0(x) = \rho_x(y_0)f_0(x)$. Indeed $f_1/|f_1|_{dvol_X}$ is the Bayesian posterior probability provided that $f_0/|f_0|_{dvol_X}$ is the prior probability. The only change from the description in [1] is the aim of the learning, i.e., prediction is replaced with action. Although the word action has an active meaning, an activity consisting of countless actions would be a chain of automatic adaptations to the environment.

2.2. Mutual Learning

It is natural to symmetrize the above setting by altering the roles of X and Y. Specifically, we further suppose that a point y of the second manifold Y parameterizes a density $\rho'_y : X \to \mathbb{R}_{>0}$ of the first manifold X, and the perception of a datum $x_0 \in X$ by the second person updates a prior density $g_0 : Y \to \mathbb{R}_{>0}$ on the second manifold to the posterior density $g_1(y) = \rho'_y(x_0)g_0(y)$. This models the double contingency of Parsons [2]. We further modify it as follows. Fix volume forms $dvol_X$, $dvol_Y$, and $dvol_{X\times Y}$ on X, Y, and $X \times Y$, respectively. Take densities $f_0 : X \to \mathbb{R}_{>0}$, $g_0 : Y \to \mathbb{R}_{>0}$, and $\lambda : X \times Y \to \mathbb{R}_{>0}$. Suppose that prior densities f_0 and g_0, respectively, changes to the posterior densities

$$f_1 = \lambda(\cdot, y_0)f_0 : x \mapsto \lambda(x, y_0)f_0(x) \quad \text{and} \quad g_1 = \lambda(x_0, \cdot)g_0 : y \mapsto \lambda(x_0, y)g_0(y).$$

This models the double contingency of Luhmann [3]. We say that f_0 is *coupled* with g_0 in the mutual learning through *Luhmann's potential* λ on the product $X \times Y$. Since the potential λ is also a density, it can be coupled with a density σ_0 on another manifold Z. Specifically, if there is a datum $((x, y)_0, z_0)$ and a density $\tau_0 : (X \times Y) \times Z \to \mathbb{R}_{>0}$, the pair of two persons can change the tendency of its action selection. This mathematics enables us to consider the double contingency not only between persons but also between systems. Here we suppose that the points $(x, y)_0$ and (x_0, y_0) are given as the same point. We emphasize that what we are discussing is not how the datum appears objectively, but how we perceive it or how we learn from it subjectively. We discuss in Section 4 the discordance between $(x, y)_0$ and (x_0, y_0) to understand a proposition in Luhmann's systems theory saying that no system is a subsystem.

2.3. Relative Entropy

Shannon [6] introduced the notion of entropy in information theory. As for continuous distributions, Jaynes [7] pointed out that the notion of relative entropy

$$D_{dvol_X}\left(\frac{f_1}{|f_1|_{dvol_X}}\middle\|\frac{f_0}{|f_0|_{dvol_X}}\right) := \int_X \frac{f_1}{|f_1|_{dvol_X}} \log\left(\frac{f_1}{|f_1|_{dvol_X}} \cdot \frac{|f_0|_{dvol_X}}{f_0}\right) dvol_X$$

is rather foundational to the notion of entropy

$$H_{dvol_X}\left(\frac{f}{|f|_{dvol_X}}\right) := -\int_X \frac{f}{|f|_{dvol_X}} \log\left(\frac{f}{|f|_{dvol_X}}\right) dvol_X.$$

Indeed, the entropy takes all real values even for normal distributions, whereas the relative entropy is non-negative for any pair of distributions, where the non-negativity is obvious from $\log(1/t) \geq 1 - t$ and is called the Gibbs inequality. Further, if we multiply the volume form by a positive constant, the entropy changes while the relatrive entropy does not. In any case, putting $f = f_1/f_0$ and using the volume form $f_0 dvol_X$, we have

$$H_{f_0 dvol_X}\left(\frac{f_1/f_0}{|f_1/f_0|_{f_0 dvol_X}}\right) = \log|f_0|_{dvolX} - D_{dvol_X}\left(\frac{f_1}{|f_1|_{dvol_X}}\middle\|\frac{f_0}{|f_0|_{dvol_X}}\right).$$

Note that we cannot put $f_0 = 1$ unless $dvol_X$ is finite. If we multiply the volume form by a non-constant density, the relative entropy varies in general. We notice that the choice of the volume forms in the above mutual learning does not affect the result of the learning.

2.4. Mutual Learning via Relative Entropy

The information geometry [4], as well as its partial globalization by the author [1,5], starts with a family of probability distributions. Slightly more generally, we consider a manifold W equipped with a volume form $dvol_W$ and a family $\{h_x\}_{x\in X}$ of densities with finite total masses on it. We regard the parameter space X as a manifold, and define the function $\varphi : X \times X \to \mathbb{R}_{\geq 0}$ on its square by

$$\varphi(x, y) := D_{dvol_W}\left(\frac{h_x}{|h_x|_{dvol_W}}\middle\|\frac{h_y}{|h_y|_{dvol_W}}\right).$$

The information geometry focuses on the 3-jet of φ at the diagonal set $\Delta \subset X \times X$. From the Gibbs inequality, the symmetric quadratic tensor defined by the 2-jet of φ is positive semi-definite. If it is positive definite, it defines a Riemannian metric called the Fisher–Rao metric. Then the symmetric cubic tensor defined by the 3-jet of the anti-symmetrization $\varphi(x, y) - \varphi(y, x)$ directs a line of torsion-free affine connections passing through the Levi-Civita connection of the Fisher–Rao metric. This line of connections is the main subject of the information geometry. On the other hand, developing the global geometry in [1,5], we define Luhmann's potential for mutual learning between two copies of X as

$$\lambda := \exp(-\varphi) : X \times X \to \mathbb{R}_{>0}.$$

We couple a prior density $f_0 : X \to \mathbb{R}_{>0}$ on the first factor with a prior density $g_0 : X \to \mathbb{R}_{>0}$ on the second factor through the potential λ on the product $X \times X$. Here, the mutual learning updates f_0 and g_0, respectively, to the posterior densities

$$f_1 = \lambda(\cdot, y_0) f_0 : x \mapsto \lambda(x, y_0) f_0(x) \quad \text{and} \quad g_1 = \lambda(x_0, \cdot) g_0 : y \mapsto \lambda(x_0, y) g_0(y).$$

Note that the function φ changes if we multiply the volume form $dvol_W$ by a non-constant density in general. Thus, the choice of the volume form is crucial. The volume form $dvol_X$ might be related to the Fisher–Rao metric, although the choice of $dvol_X$ is indeed irrelevant to the mutual learning. We can also imagine that the other volume forms $dvol_{X\times X}$ and $dvol_W$ have been determined in earlier mutual learnings "connected" to the current one.

3. Results

We address the following problem in certain cases below.

Problem 1. *Does the mutual learning via the relative entropy substitute for the conventional Bayesian estimation of the parameter of the family* $\{h_x\}$*?*

Remark 1. *The mutual learning uses only the relative entropy, whereas the conventional Bayesian estimation needs all the information about the family. Thus, Problem 1 also asks if the mutual learning can "sufficiently restore" the family from the relative entropy. To clarify this point, we use the constant* 1 *as the formal prior density in the sequel even when the total volume is infinite. Then one may compare the family with the particular posterior* g_1 *to see "how much" it is restored.*

3.1. Categorical Distributions

Let W be a 0-dimensional manifold with $N+1$ unit components, i.e., $W = \{0, \dots, N\}$ with volume form $dvol_W = 1$. A point x of the open N-simplex

$$X = \{x = (x^0, \dots, x^N) \in \mathbb{R}^{N+1} \mid x^0, \dots, x^N > 0,\ x^0 + \cdots + x^N = 1\}$$

with the standard volume form $dvol_X$ presents a categorical distribution (i.e., a finite distribution) on W. We take the product manifold $X \times X$ with Luhmann's potential

$$\lambda(x, y) = \exp\Big(-x^0 \log(x^0/y^0) - \cdots - x^N \log(x^N/y^N)\Big).$$

Suppose that the prior densities are the constants $f_0(x) \equiv 1$ and $g_0(y) \equiv 1$ on the first and second factors of $X \times X$. Then, the iteration of mutual Bayesian learning yields

$$\begin{aligned} f_n(x) &= \exp\left(nx^0\overline{\log y^0} + \cdots + nx^N\overline{\log y^N} - nx^0 \log x^0 - \cdots - nx^N \log x^N\right), \\ g_n(y) &\propto \exp\left(n\overline{x^0} \log y^0 + \cdots + n\overline{x^N} \log y^N\right) \end{aligned}$$

where the overlines denote arithmetic means $\overline{x^0} = \dfrac{x_0^0 + \cdots + x_{n-1}^0}{n}$ etc.

Proposition 1. *We have the following maximum a posteriori (MAP) estimations:*

$$\begin{aligned} x^0 : \cdots : x^N = \exp(\overline{\log y^0}) : \cdots : \exp(\overline{\log y^N}) \ &\Rightarrow\ f_n(x) = \max f_n \\ y = \overline{x} = (\overline{x^0}, \cdots, \overline{x^N}) \ &\Rightarrow\ g_n(y) = \max g_n \end{aligned}$$

We notice that the probability $g_n/|g_n|_{dvol_X}$ for the posterior density g_n on the second factor of $X \times X$ is known as the Dirichlet distribution.

Definition 1. *The Dirichlet distribution* $\mathrm{Dir}(\alpha)$ *for* $\alpha = (\alpha^0, \dots, \alpha^N) \in \mathbb{R}_{>0}^{N+1}$ *is presented by the probability* $f/|f|_{dvol X}$ *on the open N-simplex* $X \subset \mathbb{R}^{N+1}$ *for the density*

$$f(x) = \exp\Big((\alpha^0 - 1) \log x^0 + \cdots + (\alpha^N - 1) \log x^N\Big).$$

In particular, the constant $\mathrm{Dir}(1, \cdots, 1)$ *is called the flat Dirichlet distribution.*

We identify the set W with the 0-skeleton of the closure $\mathrm{cl}(X)$ of the open N-simplex $X \subset \mathbb{R}^{N+1}$. If the prior is the flat Dirichlet distribution $\mathrm{Dir}(1, \dots, 1)$, the Bayesian learning from categorical data $x'_0, \dots, x'_{n-1} \in W$ yields the posterior $\mathrm{Dir}((1, \dots, 1) + x'_0 + \cdots + x'_N)$. This is the conventional Bayesian learning from categorical data. On the other hand, the above probability $g_n/|g_n|_{dvol X}$ is the Dirichlet distribution $\mathrm{Dir}((1, \dots, 1) + x_0 + \cdots + x_N)$. Here we believe that the data $x_k \in X$ obey the probability $\lambda(x, y_k)/|\lambda(x, y_k)|_{dvol_X}$, which we can consider as a continuous version of the categorical distribution. Imagine that a coarse graining of the data x_k on X yields data x'_k obeying a categorical distribution on the 0-skeleton W of the closure of X. Then the probability $g_n/|g_n|_{dvol X}$ for the new data x'_k reaches the posterior probability of the conventional Bayesian learning.

The following is the summary of the above.

Theorem 1. *Instead of the conventional Bayesian learning from categorical data, we consider the mutual learning on the product of two copies of the space of categorical distributions via the relative entropy. Then a coarse graining of the data of the first factor into the* 0*-skeleton of the closure of the domain deforms the second factor of the mutual learning into the conventional Bayesian learning.*

Thus, the answer to Problem 1 is affirmative in this case.

3.2. Normal Distributions

In the case where X is the space of normal distributions, we would like to change the coordinates of the second factor of the product $X \times X$ to make the expression simpler, although one can reach the same result through a straightforward calculation.

3.2.1. The Coordinate System

Let X be the upper-half plane $\{(m,s) \mid m \in \mathbb{R},\ s \in \mathbb{R}_{>0}\}$ and W the line $\{w \mid w \in \mathbb{R}\}$. Suppose that any point (m,s) of X presents the normal distribution $\mathrm{N}(m,s^2)$ on W with mean m and standard deviation s. The relative entropy is expressed as

$$\begin{aligned} D_{dw}\Big(\mathrm{N}(m,s^2)\Big\|\mathrm{N}(m',s'^2)\Big) &= \frac{(m-m')^2+s^2-s'^2}{2s'^2} - \log\frac{s}{s'} \\ &= \frac{(m-m')^2+(s-s')^2}{2s'^2} + \frac{s-s'}{s'} - \log\left(1+\frac{s-s'}{s'}\right). \end{aligned}$$

This implies that the Fisher–Rao metric is the half of the Poincaré metric. We put

$$dvol_X := \frac{1}{s^2} dm \wedge ds = d\left(\frac{1}{s} dm\right),$$

and consider the symplectic product $(X, dvol_X) \times (X, dvol_X) = (X \times X', dvol_X - dvol_{X'})$. In [5], the author fixed the Lagrangian correspondence

$$N = \left\{ ((m,s),(M,S)) \in X \times X \;\middle|\; \frac{m}{s} + \frac{M}{S} = 0,\ \ sS = 1 \right\},$$

which is the graph of the symplectic involution

$$F : X \to X : (m,s) \mapsto (M,S) = \left(-\frac{m}{s^2}, \frac{1}{s}\right).$$

Using it, the author took the "stereograph" $\widetilde{D} : X \times X \to \mathbb{R}_{\geq 0}$ of the relative entropy as follows. Regard a value D of the relative entropy $D_{dw}\big(\mathrm{N}(m,s^2)\big\|\mathrm{N}(m',s'^2)\big)$ as a function of the pair of two points (m,s) and (m',s') on the first factor of the product $X \times X$; take the point $(M,S) = \left(-\frac{m'}{s'^2}, \frac{1}{s'}\right)$ on the second factor, which N corresponds to the point (m',s') on the first factor; and regard the value D as the value of a function $\widetilde{D}$ of the point $(m,s,M,S) \in X \times X$. That is, the function $\widetilde{D}$ is defined by

$$\begin{aligned} \widetilde{D}(m,s,M,S) :=& D_{dw}\Big(\mathrm{N}(m,s^2)\Big\|\mathrm{N}(-M/S^2, 1/S^2)\Big) \\ =& \frac{1}{2}\left(\frac{M}{S} + sS\frac{m}{s}\right)^2 + \frac{s^2S^2 - 1 - \log(s^2S^2)}{2}. \end{aligned}$$

The function $\widetilde{D}$ enjoys symplectic/contact geometric symmetry as well as the submanifold N. See [1] for the multivariate versions of $\widetilde{D}$ and N with Poisson geometric symmetry.

3.2.2. The Mutual Learning

In the above setting, we define Luhmann's potential by

$$\begin{aligned}\lambda(m,s,M,S) :=& \exp(-\widetilde{D}(m,s,M,S)) \\ =& sS\exp\left(-\frac{1}{2}\left(\frac{M}{S}+sS\frac{m}{s}\right)^2-\frac{s^2S^2-1}{2}\right)\end{aligned}$$

Put $f_0(m,s)\equiv 1$ and $g_0(M,S)\equiv 1$. Then, the iteration of the mutual learning yields

$$\begin{aligned}f_n(m,s) &\propto s^n\exp\left(-\frac{n\overline{S^2}}{2}\left\{\left(m-\frac{-\overline{M}}{\overline{S^2}}\right)^2+s^2\right\}\right)\\ &\le s^n\exp\left(-\frac{n\overline{S^2}}{2}s^2\right),\\ g_n(M,S) &\propto S^n\exp\left(-\frac{nS^2}{2}\left\{\left(\frac{-M}{S^2}-\overline{m}\right)^2+\overline{m^2}-\overline{m}^2+\overline{s^2}\right\}\right)\\ &= (s')^{-n}\exp\left(-\frac{n}{2(s')^2}\left\{(m'-\overline{m})^2+\overline{m^2}-\overline{m}^2+\overline{s^2}\right\}\right).\end{aligned}$$

Since $\frac{d}{ds}s^n\exp\left(-\frac{n\overline{S^2}}{2}s^2\right)=\left(ns^{n-1}-n\overline{S^2}s^{n+1}\right)\exp\left(-\frac{n\overline{S^2}}{2}s^2\right)$, we see that the density f_n reaches the maximum at $(m,s)=\left(\frac{-\overline{M}}{\overline{S^2}},\sqrt{\frac{1}{\overline{S^2}}}\right)$. Similarly, we can see that the density g_n reaches the maximum when $m'=\frac{-M}{S^2}=\overline{m}$ and $s'^2=\frac{1}{S^2}=\overline{m^2}-\overline{m}^2+\overline{s^2}$ hold.

Definition 2. *The normal-inverse-Gamma distribution* $\mathrm{NIG}(\mu,\nu,\alpha,\beta)$ *on the upper-half plane* $\hat{X}=\{(m,v)\mid(m,s)\in X, v=s^2\}$ *equipped with the volume form* $dvol_{\hat{X}}=dm\wedge dv$ *is the probability density proportional to*

$$v^{-\alpha-1}\exp\left(-\frac{\nu(m-\mu)^2}{2v}-\frac{\beta}{v}\right).$$

Its density form is the volume form with unit total mass, which is proportional to

$$v^{-\alpha-1}\exp\left(-\frac{\nu(m-\mu)^2}{2v}-\frac{\beta}{v}\right)dvol_{\hat{X}}.$$

Using our volume form $dvol_X$, we can write the density form of $\mathrm{NIG}(\mu,\nu,\alpha,\beta)$ as

$$\mathrm{const}\cdot s^{-2\alpha+1}\exp\left(-\frac{\nu}{2s^2}\left\{(m-\mu)^2+\frac{2\beta}{\nu}\right\}\right)dvol_X.$$

This is proportional to $g_n dvol_X$ on the second factor of $X\times X$ when

$$(\mu,\nu,\alpha,\beta)=\left(\overline{m},n,\frac{n+1}{2},\frac{n(\overline{m^2}-\overline{m}^2+\overline{s^2})}{2}\right).$$

We identify the line W with the boundary of X. The conventional Bayesian learning of the normal data $m_0,\dots,m_{n-1}$ yields the posterior $\mathrm{NIG}\left(\overline{m},n,\frac{n+1}{2},\frac{n(\overline{m^2}-\overline{m}^2)}{2}\right)$ provided that the prior is formally 1. Thus, we have the following result similar to Theorem 1.

Theorem 2. *Instead of the conventional Bayesian learning from normal data on $\mathbb{R}$, we consider the mutual learning on the product of two copies of the space X of normal distributions via the relative entropy. Then a coarse graining of the data of the first factor into the boundary $\partial X = \mathbb{R}$ by taking $s \to 0$ deforms the second factor of the mutual learning into the conventional Bayesian learning.*

Thus, the answer to Problem 1 is also affirmative in this case.

3.3. Von Mises Distributions with Fixed Concentration in Circular Case

A von Mises distribution $M_k(m)$ with a fixed large concentration $k(\gg 1)$ is a circular analogue of a normal distribution with a fixed small variance that is parametrized by a point m of $X = \mathbb{R}/2\pi\mathbb{Z}$. Its density is proportional to the restriction of the function $\exp(k\cos(m)x + k\sin(m)y)$ to the circle $W = \{(x,y) \mid x = \cos w, y = \sin w, w \in \mathbb{R}/2\pi\mathbb{Z}\}$ with $dvol_W = dw$. Then, using the easy formula $\int_0^{2\pi} \exp(k\cos x)\sin x\, dx = 0$, we obtain the following expression of the relative entropy:

$$\begin{aligned} D(M_k(m)\|M_k(m')) &= \int_0^{2\pi} \exp(k\cos(w-m'))\big(k\cos(w-m') - k\cos(w-m)\big)dw \\ &= c(1-\cos(m-m')) \end{aligned}$$

where c is a positive constant. (When $k \in \mathbb{Z}$, using modified Bessel, we have $c = \dfrac{kI_1(k)}{I_0(k)}$.) Thus, Luhmann's potential is $\lambda(m,m') = \exp(-c(1-\cos(m-m'))$. We put $f_0(m) \equiv 1$ and $g_0(m') \equiv 1$. Then, the iteration of mutual Bayesian learning on the torus $X \times X$ yields

$$\begin{aligned} f_n(m) &= \exp\Big(-nc(1 - \overline{\cos m'}\cos m - \overline{\sin m'}\sin m)\Big), \\ g_n(m') &= \exp\big(-nc(1 - \overline{\cos m}\cos m' - \overline{\sin m}\sin m')\big). \end{aligned}$$

On the other hand, the conventional Bayesian learning on W yields the posterior probability density proportional to $\exp\big(nk\overline{\cos}(m)\cos w + nk\overline{\sin}(m)\sin w\big)$, which looks like $g_n(w)$. This suggests the affirmative answer to Problem 1.

3.4. Conclusions

We have observed that the answer to Problem 1 is affirmative in some cases. Specifically, the mutual Bayesian learning covers at least a non-empty area of parametric statistics. The author expects that it could cover the whole from some consistent perspective.

4. Discussion

4.1. On Socio-Cybernetics

In our setup of mutual learning, a system must be organized as the product of two manifolds with Luhmann's potential before each member learns. Further, the potential is the result of an earlier mutual learning in which the system was a member. In Luhmann's description [3], the unit of society is not the agent of an action but a communication or rather a chain of communications. In mathematics, a manifold is locally a product of manifolds and is characterized as the algebraic system of functions on it. By analogy, Luhmann's society seems to be a system of relations between certain systems of functions. Some authors criticize his theory for failing to acquire individual identity, but an individual is a relation between identities that are already represented by manifolds.

As a matter of course, reality cannot be explained by theories. Instead, a theory which can better explain something on reality is chosen. In Section 2.2, we have assumed that $(x,y)_0$ and (x_0,y_0) are given as the same point in reality. Then there are two possibilities: (1) The potential λ is updated by using $(x,y)_0$ as a component of a datum, or (2) the mutual learning of (x_0,y_0) is performed under the undated potential λ. The discordance between (1) and (2) does not affect the reality. Further, there is no consistent hierarchy among

Luhmann's systems that choose either (1) or (2), and therefore there is no system that is a proper subsystem of another system. Perhaps, the social system chooses either (1) or (2), which can better explain the "fact" in relation to other "facts" in a story on reality. Undertaking all of the above, the notion of autopoiesis that Maturana and Varela [8] found in living organisms can be the foundation of Luhmann's socio-cybernetics.

4.2. On the Total Entropy

In objective probability theory, one considers a continuous probability distribution as the limit of a family of finite distributions presented by relative frequency histograms and the entropy of the limit as the limit of the entropies. Since the entropy of a finite distribution whose support is not a singleton is positive, a distribution with negative entropy, e.g., a normal distribution with small variance, does not appear. On the other hand, we take the position of subjective probability theory, and regard a positive function on a manifold that has the unit mass with respect to a fixed volume form as a probability density. From our point of view, the relative entropy between two probability densities is essential as it is non-negative; it presents the information gain; and it does not change (while even the sign of entropy does change) by multiplying the volume form by any positive constant. We notice that an objective probability is a subjective probability, and not vice-versa.

We know that the lowest entropy at the beginning of the universe must be relative to higher entropy in the future. In this regard, the total amount of information decreases as the order of time. However, it is still possible that the amount of consumable information increases, and perhaps that is how this world works. Here we would like to distinguish the world from the universe, even though they concern the same reality and therefore communicate with each other. The world consists of human affairs, including the possible variations of knowledge on facts in the universe—there is no love in the universe, but love is the most important consumable thing in the world. We consider that the notion of complexity in Luhmann's systems theory concerns such consumability as it relates to coupling of systems. Now the problem is not the total reserve of information, but how to strike it and refine it like oil. At present, autopoiesis is gaining ground against mechanistic cybernetics. Our research goes against this stream: Its goal is to invent a learning machine to exploit information resources to be consumed by humans and machines.

4.3. On Geometry

In this paper, we have quickly gone from the general definition of mutual learning to a discussion of the special mutual learning via relative entropy. However, it may be worthwhile to stop and study various types of learning according to purely geometric interests. For example, the result of previous work [1] is apparently related to the geometry of dual numbers, and fortunately this special issue includes a study [9] on a certain pair of dual number manifolds. Considering mutual learning for pairs of related manifolds such as this is something to be investigated in the future.

In addition, in proceeding to the case of the mutual learning via relative entropy, one basic problem was left unaddressed: Given a non-negative function φ on a squared manifold $M \times M$ that takes zero on the diagonal set, can we take a family of probability densities with parameter space M so that the relative entropy induces the function φ?

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Mori, A. Global Geometry of Bayesian Statistics. *Entropy* **2020**, *22*, 240. [CrossRef] [PubMed]
2. Parsons, T. *The Social System*; Free Press: Glencoe, IL, USA, 1951.
3. Luhmann, N. The autopoiesis of the social system. In *Sociocybernetic Paradoxes: Observation, Control and Evolution of Self-Steering Systems*; Geyer, R.F., van der Zouwen, J., Eds.; Sage: London, UK, 1986; pp. 172–192.
4. Amari, S. *Information Geometry and Its Applications*; Springer: Tokyo, Japan, 2016.
5. Mori, A. Information geometry in a global setting. *Hiroshima Math. J.* **2018**, *48*, 291–305. [CrossRef]
6. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
7. Jaynes, E. Prior Probabilities. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 227–241. [CrossRef]
8. Maturana, H.; Varela, F. *Autopoiesis and Cognition: The Realization of the Living, Boston Studies in the Philosophy and History of Science 42*; Reidel: Dordrecht, The Netherlands, 1972.
9. Li, Y.; Alluhaibi, N.; Abdel-Baky, R.A. One-parameter Lorentzian dual spherical movements and invariants of the axodes. *Symmetry* **2022**, *14*, 1930. [CrossRef]