doi:10.20944/preprints201711.0020.v1

Article

A Measure of Information Available for Prediction

Takuya Isomura [1]*

[1] RIKEN Brain Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

* Correspondence: takuya.isomura@riken.jp; Tel.: +81-48-467-9644

Abstract: Mutual information between the brain state and the external world state represents the amount of information stored in the brain that is associated with the external world. On the other hand, surprise of sensory input indicates the unpredictability of the current input. In other words, this is a measure of prediction capability, and an upper bound of surprise is known as free energy. According to the free-energy principle (FEP), the brain continues to minimize free energy to perceive the external world. For animals to survive, prediction capability is considered more important than just memorizing information. In this study, the fact that free energy represents a gap between the amount of information stored in the brain and that available for prediction is established, where the latter will be referred to as predictive information as an analogy with Bialek's predictive information. This concept involves the FEP, the infomax principle, and the predictive information theory, and will be a useful measure to quantify the amount of information available for prediction.

Keywords: the free-energy principle; internal model hypothesis; unconscious inference; infomax principle; predictive information; independent component analysis; principal component analysis

1. Introduction

Sensory perception comprises complex responses of the brain to sensory inputs. For example, the visual cortex can distinguish objects from their background [1], while the auditory cortex can recognize a certain sound in a noisy place with high sensitivity, a phenomenon known as the cocktail party effect [2–7]. The brain has acquired these perceptual abilities without supervision, which is referred to as unsupervised learning [8–10]. Unsupervised learning, or implicit learning, is defined as the learning that happens in the absence of a teacher or supervisor; it is achieved through adaptation to environments experienced in the past, which is necessary for higher brain functions. Thus, an understanding of the physiological mechanisms that mediate unsupervised learning is fundamental to augmenting our knowledge of information processing in the brain.

One of benefits of unsupervised learning is inference, which represents the action of guessing unknown matters based on known facts or certain observations; i.e., it is the process of drawing conclusions through reasoning and estimation. While inference is thought to be an act of the conscious mind in the ordinary sense of the word, where consciousness often represents a state of self-awareness, indeed it can occur even in the unconscious mind. Hermann von Helmholtz, a 19th-century physicist/physiologist, realized that perception often requires inference by the unconscious mind and coined the word '*unconscious inference*' [11]. According to him, conscious inference and unconscious inference can be distinguished based on whether conscious knowledge is involved in the process. For example, when an astronomer computes the positions of the stars in space or their distances based on the perspective images at various times and from different parts of the orbit of the earth, he performs conscious inference. This is because the process is *"based on a conscious knowledge of the laws of optics"*; by contrast, *"in the ordinary acts of vision, this knowledge of optics is lacking"* [11]. Thus, the latter process is performed by the unconscious mind. In spite of such a difference, there is no doubt in the similarity between the results of conscious and unconscious inference. Similar to conscious inference, unconscious inference must be crucial for cognitive processes under the unconscious mind to estimate the overall picture from partial observations.

In the field of theoretical and computational neuroscience, unconscious inference has been translated as that people are constantly and unconsciously inferring (in terms of Bayesian inference) the generative process of the external world in order to achieve perception. One hypothesis, the so-called internal model hypothesis [12–18], states that people reconstruct a model of the external world in their brain through the past experiences. This internal model helps people infer hidden causes and predict future inputs automatically; in other words, this inference process happens unconsciously. This is also known as predictive coding hypothesis [19,20]. For many years, unconscious inference has been mathematically modeled under the internal model hypothesis, such as by the Helmholtz machine [12], dynamic causal modeling [14], and Markov decision process model [16]. In the 2000s, Friston proposed a mathematical foundation for unconscious inference, called the free-energy principle (FEP) [13–16], which is a candidate unified theory of higher brain functions. According to him, this principle provides a unified framework for higher brain functions including perceptual learning [14], reinforcement learning [22], motor learning [21,22], communication [23,24], emotion, mental disorders [25,26], and evolution. However, the difference between the FEP and related theories, namely the information maximization (infomax) principle [27,28] and the predictive information theory [29,30], have not been established.

In this study, the relationship between the FEP and other theories is investigated. As one of most simple and important examples, I focus on blind source separation (BSS), which is a task to separate hidden sources (or causes) from sensory inputs [31–34]. I show that BSS is a subset of the inference problem considered in the FEP, and demonstrate that free energy defined in the FEP represents the difference between the information stored in the brain (which is the measure of the infomax principle [27,28]) and the information available for predicting current and future sensory inputs (which is a measure similar to one used in the predictive information theory [29,30]).

2. Definition of a system

Let us suppose $s \equiv (s_1, \ldots, s_N)^T \sim p(s) \equiv \prod_i p(s_i)$ as hidden sources; $x \equiv (x_1, \ldots, x_M)^T \sim p(x)$ as sensory inputs; $u \equiv (u_1, \ldots, u_N)^T \sim p(u)$ as neural outputs; $z \equiv (z_1, \ldots, z_M)^T \sim p(z)$ as background noises; $\epsilon \equiv (\epsilon_1, \ldots, \epsilon_M)^T \sim p(\epsilon)$ as prediction errors; and $f \in \mathbb{R}^M$, $g \in \mathbb{R}^N$, and $h \in \mathbb{R}^M$ as nonlinear functions (see also Table 1). The generative process of the external world (or the environment) is described by a stochastic equation as

$$\text{Generative process}: \ x = f(s) + z, \tag{1}$$

and recognition and generative models of the brain are as follows:

$$\text{Recognition model}: \ u = g(x), \tag{2}$$

$$\text{Generative model}: \ x = h(u) + \epsilon. \tag{3}$$

Figure 1 illustrates the structure of the system under consideration. For the generative model, I define the prior distribution of u as $p_u(u) = \prod_i p_u(u_i)$ and the likelihood function as $p_\epsilon(\epsilon) = p^*(x|h(u)) = \mathcal{N}[\epsilon; 0, \Pi_\epsilon]$, where p^* indicates a statistical model and $\mathcal{N}$ is a Gaussian distribution. Moreover, suppose $\theta \sim p(\theta)$, $W (\in \mathbb{R}^{N \times M}) \sim p(W)$, and $V (\in \mathbb{R}^{M \times N}) \sim p(V)$ as parameter sets for f, g, and h, respectively, $\lambda \sim p(\lambda)$ as a hyper-parameter set for $p(s)$ and $p(z)$, and $\gamma \sim p(\gamma)$ as a hyper-parameter set for $p_u(u)$ and $p_\epsilon(\epsilon)$. Note that W and V are assumed as synaptic strength matrices for feedforward and backward paths, respectively, while γ is assumed as a state of neuromodulators similarly to [13–15]. Thus, Eqs. (1)-(3) are transformed into probabilistic representations

Table 1. Glossary of expressions.

Expression	Description
Generative process	A set of stochastic equations that generate the external world dynamics
Recognition model	A model in the brain that mimics the inverse of the generative process
Generative model	A model in the brain that mimics the generative process
$s \in \mathbb{R}^N$	Hidden sources
$x \in \mathbb{R}^M$	Sensory inputs
θ	A set of parameters
λ	A set of hyper-parameters
$\vartheta \equiv \{s, \theta, \lambda\}$	A set of hidden states of the external world
$u \in \mathbb{R}^N$	Neural outputs
$W \in \mathbb{R}^{N \times M}, V \in \mathbb{R}^{M \times N}$	Synaptic strength matrices
γ	State of neuromodulators
$\varphi \equiv \{u, W, V, \gamma\}$	A set of the brain internal states
$z \in \mathbb{R}^M$	Background noises
$\epsilon \in \mathbb{R}^M$	Prediction errors
$p(x)$	The true probability density of x
$p(\varphi \mid x), p(x, \varphi), p(\varphi)$	True probability densities (posterior densities)
$p_u(u), p_\epsilon(\epsilon), p_\varphi(\varphi)$	Prior densities
$p^*(x), p^*(\varphi \mid x), p^*(x, \varphi)$	Statistical models
$dx \equiv \prod_i dx_i$	Finite spatial resolution of x
$\langle \bullet \rangle_{p(x)} \equiv \int \bullet p(x) dx$	Expectation of • over $p(x)$
$H[p(x)] \equiv \langle -\log(p(x)dx) \rangle_{p(x)}$	Shannon entropy of $p(x)dx$
$\langle -\log(p^*(x)dx) \rangle_{p(x)}$	Cross entropy of $p^*(x)dx$ over $p(x)$
$\mathcal{D}_{KL}[p(\bullet) \vert\vert p^*(\bullet)] \equiv \left\langle \log \frac{p(\bullet)}{p^*(\bullet)} \right\rangle_{p(\bullet)}$	KLD between $p(\bullet)$ and $p^*(\bullet)$
$I[x; \varphi] \equiv \mathcal{D}_{KL}[p(x, \varphi) \vert\vert p(x)p(\varphi)]$	Mutual information between x and φ
$S(x) \equiv \log \frac{p(x)}{p^*(x)}$	Surprise
$\overline{S} \equiv \langle S(x) \rangle_{p(x)}$	Surprise expectation
$F(x) \equiv S(x) + \mathcal{D}_{KL}[p(\varphi \mid x) \vert\vert p^*(\varphi \mid x)]$	Free energy
$\overline{F} \equiv \langle F(x) \rangle_{p(x)}$	Free energy expectation
$X[x; \varphi] \equiv \left\langle \log \frac{p^*(x, \varphi)}{p(x)p(\varphi)} \right\rangle_{p(x, \varphi)}$	Predictive information between x and φ

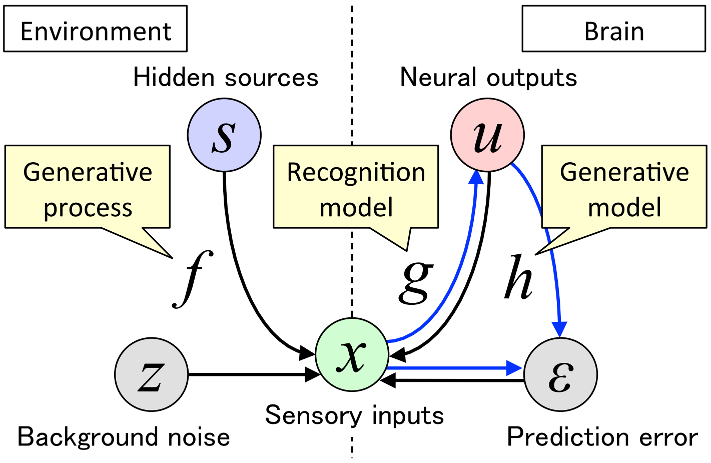

Figure 1. Schematic images of a generative process of the environment (left) and recognition and generative models of the brain (right). Note that the brain can access only the states in the right side of the dashed line, including x (see text in Section 3). Black arrows are causal relationships, while blue arrows are information flows of the neural network. See main text and Table 1 for meanings of variables and functions.

$$\begin{aligned}\text{Generative process}: \ p(s,x|\theta,\lambda) &= p(x|s,\theta,\lambda)p(s|\lambda) \\ &= \int \delta(x-f(s;\theta)-z)p(z|\lambda)p(s|\lambda)dz \\ &= p(z=x-f|\lambda)p(s|\lambda),\end{aligned} \tag{4}$$

$$\begin{aligned}\text{Recognition model}: \ p(x,u|W) &= p(x|u,W)p(u|W) \\ &= p(u|x,W)p(x) \\ &= \delta(u-g(x;W))p(x),\end{aligned} \tag{5}$$

$$\begin{aligned}\text{Generative model}: \ p^*(x,u|V,\gamma) &= p^*(x|u,V,\gamma)p_u(u|\gamma) \\ &= \int \delta(x-h(u;V)-\epsilon)p_\epsilon(\epsilon|\gamma)p_u(u|\gamma)d\epsilon \\ &= p_\epsilon(\epsilon=x-h|\gamma)p_u(u|\gamma).\end{aligned} \tag{6}$$

Note that $\delta(\bullet)$ is Dirac's delta function and $p^*(x|u,V,\gamma) \equiv p(x|u,V,\gamma,m)$ is a statistical model given a model structure m. For simplification, let us define $\vartheta \equiv \{s,\theta,\lambda\}$ as a set of hidden states of the external world and $\varphi \equiv \{u,W,V,\gamma\}$ as a set of internal states of the brain. Accordingly, by multiplying $p(\theta,\lambda)$ to Eq. (4) and $p(W,V,\gamma)$ to Eqs. (5)(6), Eqs. (4)-(6) become

$$\text{Generative process}: \ p(x,\vartheta) = p(x|\vartheta)p(\vartheta) = p(z=x-f)p(\vartheta), \tag{7}$$

$$\text{Recognition model}: \ p(x,\varphi) = p(x|\varphi)p(\varphi) = p(\epsilon=x-h)p(\varphi), \tag{8}$$

$$\text{Generative model}: \ p^*(x,\varphi) = p^*(x|\varphi)p_\varphi(\varphi) = p_\epsilon(\epsilon=x-h)p_\varphi(\varphi), \tag{9}$$

where p_φ is the prior distribution for φ and $p^*(x,\varphi) \equiv p(x,\varphi|m)$ is a statistical model given a model structure m, which is determined by the shapes of p_φ and p_ϵ. I use the expression of $p^*(x,\varphi)$ instead of $p(x,\varphi|m)$ to emphasize the difference between $p(x,\varphi)$ and $p^*(x,\varphi)$. While $p(x,\varphi)$ is the true joint probability of (x,φ) (the so-called posterior distribution), $p^*(x,\varphi)$, i.e., the product of the likelihood function and the prior distribution, represents a model that the brain hopes (x,φ) should follow. As shown later, the learning and perception in terms of the unconscious inference are achieved by minimizing the difference between $p(x,\varphi)$ and $p^*(x,\varphi)$.

3. Information stored in the brain

This section reviews the basis of information theory [35]. Information is defined as the negative log of probability. Let $\text{Prob}(x)$ be the probability of given sensory inputs x. The information in the sensory input is given by $-\log \text{Prob}(x)$ [nat], where 1 nat = 1.4427 bits. When x takes continuous values, by coarse graining, $-\log \text{Prob}(x)$ is replaced with $-\log(p(x)dx)$, where $p(x)$ is the probability density of x and $dx \equiv \prod_i dx_i$ is the product of the finite spatial resolutions of x's elements. The expectation of $-\log(p(x)dx)$ over $p(x)$ gives the Shannon entropy (or average information) [10]. Thus, in this study, Shannon entropy is defined by

$$H[p(x)] \equiv \int -\log(p(x)dx)p(x)dx \equiv \langle -\log(p(x)dx)\rangle_{p(x)} \ [\text{nat}]. \tag{10}$$

Note that $\langle\bullet\rangle_{p(x)}$ refers to the expectation of $\bullet$ over $p(x)$, $\langle\bullet\rangle_{p(x)} \equiv \int \bullet p(x)dx$. Since $d\text{Prob}(x) = p(x)dx$ takes a value between $0 \le p(x)dx \le 1$, $H[p(x)]$ takes a non-negative value, $H[p(x)] \ge 0$. Although this definition of $H[p(x)]$ is different from the original one, because a constant $-\log dx$ has been added, it is useful since $H[p(x)]$ becomes non-negative while there is no effect except sliding of the offset value. Note that $H[p(x)] = 0$ is realized if and only if $p(x)$ is Dirac's delta function. In the case of the discrete system, the change from a system where x could take two states with the same probability to a system where x could take only one state deterministically decreases 1 bit of entropy.

This means that the brain memorizes the 1-bit information; i.e., the brain state corresponds to 1 bit of the external world state. Whereas, in the case of the continuous system, a constraint should be added to avoid divergence; this will be referred to as internal energy [14]. Internal energy has the same unit as Shannon entropy. The information loss increases if a state goes away from the energy landscape.

Let us consider the case where the sensory inputs are determined by the hidden states. Again, suppose x as sensory inputs; $\vartheta = \{s, \theta, \lambda\}$ as a set of the external world hidden states, i.e., a set of hidden sources s, parameters θ, and hyper-parameters λ; and $\varphi = \{u, W, V, \gamma\}$ as a set of the brain internal states, i.e., a set of neural outputs u, synaptic strength matrices W and V, and neuromodulators γ. The external world states are determined by a set of x and ϑ, (x, ϑ). Mathematically, the information shared between the external world states (x, ϑ) and the brain internal states φ is defined by mutual information $I[(x, \vartheta); \varphi]$, which is defined in terms of the Kullback-Leibler divergence (KLD) [10] as

$$I[(x,\vartheta);\varphi] \equiv \mathcal{D}_{KL}\Big[p(x,\vartheta,\varphi)||p(x,\vartheta)p(\varphi)\Big] \equiv \left\langle \log \frac{p(x,\vartheta,\varphi)}{p(x,\vartheta)p(\varphi)} \right\rangle_{p(x,\vartheta,\varphi)} \quad [\mathrm{nat}]. \tag{11}$$

Note that $p(x, \vartheta, \varphi)$ is the joint probability of (x, ϑ) and φ, and $p(x, \vartheta)$ and $p(\varphi)$ are their marginal distributions, respectively. KLD indicates the distance between two distributions; thus, $I[(x, \vartheta); \varphi]$ represents how different $p(x, \vartheta, \varphi)$ is from $p(x, \vartheta)p(\varphi)$. If (x, ϑ) and φ are independent of each other, $I[(x, \vartheta); \varphi]$ becomes zero as $p(x, \vartheta, \varphi) = p(x, \vartheta)p(\varphi)$ holds. Otherwise, $I[(x, \vartheta); \varphi]$ takes a positive value because of the non-negativity of KLD [10].

However, there is a clear requirement in practice that "information that the brain can access consists only of the sensory input"; i.e., the brain can access only the sensory input x. Thus, the brain needs to increase $I[(x, \vartheta); \varphi]$ without accessing ϑ directly, so that ϑ are referred to as hidden states. Accordingly, because ϑ given x is independent of φ given x, $p(\vartheta, \varphi|x) = p(\vartheta|x)p(\varphi|x)$, I have

$$I[(x,\vartheta);\varphi] = \left\langle \log \frac{p(\vartheta|x)p(\varphi|x)p(x)}{p(\vartheta|x)p(x)p(\varphi)} \right\rangle_{p(\vartheta|x)p(\varphi|x)p(x)} = \left\langle \log \frac{p(\varphi|x)}{p(\varphi)} \right\rangle_{p(\varphi,x)} = I[x;\varphi]. \tag{12}$$

Using Shannon entropy, $I[x; \varphi]$ becomes

$$I[x;\varphi] = H[p(x)] - H[x|\varphi] \quad [\mathrm{nat}], \tag{13}$$

where

$$H[x|\varphi] \equiv \Big\langle -\log\Big(p(x|\varphi)dx\Big)\Big\rangle_{p(x,\varphi)} \equiv \langle H[p(\epsilon)]\rangle_{p(\varphi)} \equiv \Big\langle -\log\Big(p(\epsilon)dx\Big)\Big\rangle_{p(\epsilon)p(\varphi)} \tag{14}$$

is the conditional entropy of x given φ. Thus, maximization of $I[(x, \vartheta); \varphi]$ is the same meaning as maximization of $I[x; \varphi]$ for the brain. As $I[x; \varphi]$, $H[p(x)]$, and $H[x|\varphi]$ are non-negative, $I[x; \varphi]$ has the range of $0 \leq I[x; \varphi] \leq H[p(x)]$. Note that $I[x; \varphi] = 0$ occurs if and only if x and φ are independent of each other, while $I[x; \varphi] = H[p(x)]$ occurs if and only if x is fully explained by φ. In this manner, $I[x; \varphi]$ describes the information on the external world stored in the brain. According to the infomax principle, the brain maximizes $I[x; \varphi]$ to perceive the external world [27,28]. However, $I[x; \varphi]$ does not fully explain the prediction performance of the brain. For example, if neural outputs just express the sensory input itself ($u = x$), $I[x; \varphi] = H[p(x)]$ is easily achieved, but it does not mean that the brain can predict input statistics. This will be considered in the next section.

4. The free-energy principle

If one has a statistical model determined by model structure m, the information calculated based on m is given by the negative log likelihood $-\log p(x|m)$, which is termed as the surprise of the sensory input. The surprise represents the unpredictability of the sensory input for the individual. For example, a visual input such as that of a chicken flying across the sky has a high surprise value because this scene has never been seen, but the surprise will decrease after one learns that this can happen. The brain is considered to minimize the surprise in the sensory input based on the prior knowledge of the external world, in order to perform unconscious inference and optimize their perception [13]. To infer if an event is likely to happen based on the past observation, a statistical model is necessary; otherwise it is difficult for the brain to generalize sensory inputs [36]. As in Section 2, I express a statistical model as $p^*(x) \equiv p(x|m)$ to clarify the difference from true probability density $p(x)$. Notably, the cross entropy $\langle -\log(p^*(x)dx)\rangle_{p(x)}$ is always larger than or equal to Shannon entropy $H[p(x)]$ because of the non-negativity of KLD. Hence, in this study, I define the input surprise by

$$S(x) \equiv -\log p^*(x) + \log p(x) \quad [\text{nat}] \tag{15}$$

and its expectation over $p(x)$ by

$$\begin{aligned} \overline{S} &\equiv \langle S(x)\rangle_{p(x)} = \mathcal{D}_{KL}[p(x)||p^*(x)] \\ &= \langle -\log(p^*(x)dx)\rangle_{p(x)} - H[p(x)] \quad [\text{nat}]. \end{aligned} \tag{16}$$

This definition of $S(x)$ is different from the original one [13] as $\log p(x)$ has been added, but it is useful since $\overline{S} \geq 0$ and $\overline{S} = 0$ holds if and only if $p^*(x) = p(x)$ while there is no effect except sliding of the offset value.

Because x is generated by the external world generative process, consideration of the structure and dynamics behind the sensory input can provide accurate inference. According to the internal model hypothesis, animals develop the internal model in their brain to increase the accuracy and efficiency of inference [12–18]; thus, the brain internal states φ are hypothesized to mimic the hidden states of the external world ϑ. A problem is that $-\log p^*(x) = -\log(\int p^*(x,\varphi)d\varphi)$ is intractable for animals, because they have to deal with the integral of $p^*(x,\varphi)$ placed in the logarithm function. The FEP hypothesizes that animals calculate an upper bound of $-\log p^*(x)$ instead that is tractable for them and terms this bound as free energy $F(x)$ [13].

$$\begin{aligned} F(x) &\equiv S(x) + \mathcal{D}_{KL}[p(\varphi|x)||p^*(\varphi|x)] \\ &= \langle -\log p^*(x,\varphi) + \log p(x,\varphi)\rangle_{p(\varphi|x)} \, [\text{nat}]. \end{aligned} \tag{17}$$

Again, this definition of $F(x)$ is different from the original one [13] as $\log p(x)$ has been added. Note that $p(\varphi|x)$ is the conditional probability of the internal model in the brain, termed as the recognition density. Due to the non-negativity of KLD, $F(x)$ provides an upper bound of $S(x)$ and $F(x) = S(x)$ holds if and only if $p^*(\varphi|x) = p(\varphi|x)$. Furthermore, the expectation of $F(x)$ over $p(x)$ is defined by

$$\begin{aligned} \overline{F} &\equiv \langle F(x)\rangle_{p(x)} = \mathcal{D}_{KL}[p(x,\varphi)||p^*(x,\varphi)] \\ &= \langle U(x,\varphi)\rangle_{p(x,\varphi)} - H[p(x,\varphi)] \quad [\text{nat}], \end{aligned} \tag{18}$$

where $U(x,\varphi) \equiv -\log(p^*(x,\varphi)dxd\varphi)$ is termed as the internal energy and $H[p(x,\varphi)] \equiv \langle -\log(p(x,\varphi)dxd\varphi)\rangle_{p(x,\varphi)}$ is the joint entropy of x and φ. $\overline{F}$ indicates the difference between the actual probability $p(x,\varphi)$ and its statistical model $p^*(x,\varphi)$. Because of the non-negativity of KLD, $\overline{F}$ is always larger than or equal to $\overline{S}(\geq 0)$ and $\overline{F} = \overline{S} = 0$ holds if and only if $p^*(x,\varphi) = p(x,\varphi)$. Internal energy $U(x,\varphi)$ quantifies the amplitude of the prediction error at a given moment [13]. Minimization of $\langle U(x,\varphi)\rangle_{p(x,\varphi)}$ is the so-called maximum a posteriori (MAP) estimation (or the maximum likelihood estimation if the priors are uniform distributions) [10] and provides a solution that (at least locally) minimizes the prediction error. Whereas, maximization of $H[p(x,\varphi)]$ increases the independency between internal states, which helps neurons to establish an efficient representation as pointed out by Jaynes' max entropy principle [37,38]. This is essential for BSS [31–34] because the optimal parameters that minimize $\langle U(x,\varphi)\rangle_{p(x,\varphi)}$ are not always determined identically. Due to this, the MAP estimation alone does not always identify the generative process behind the sensory inputs. As $\overline{F}$ is the sum of costs for the MAP estimation and BSS, free-energy minimization is the rule to simultaneously minimize the prediction error and maximize the independency of the internal states.

5. Information available for prediction

Then, let us consider how free energy expectation $\overline{F}$ relates to mutual information $I[x;\varphi]$. According to Helmholtz's unconscious inference and the internal model hypothesis, the aim of the brain is to predict x, and for this purpose, the brain shifts the actual probability $p(x,\varphi) = p(\epsilon)p(\varphi)$ closer to the statistical model $p^*(x,\varphi) = p_\epsilon(\epsilon)p_\varphi(\varphi)$ that the brain hopes (x,φ) should follow. Thus, the difference between these two distributions is associated with the loss of information. The amount of information available for the prediction can be calculated in the following manner: as $H[p(x)]$ is information of the sensory input and $I[x;\varphi]$ is information stored in the brain, $H[p(x)] - I[x;\varphi] = \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ indicates the information loss in the recognition model (Fig. 2). By contrast, the distance between actual and desired (prior) distributions of internal states $\mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)]$ quantifies the information loss for inferring internal states (i.e., blind state separation). Moreover, the distance between distributions of the actual reconstruction error and the prediction error under the given model $\langle \mathcal{D}_{KL}[p(x|\varphi)||p^*(x|\varphi)]\rangle_{p(\varphi)} = \langle \mathcal{D}_{KL}[p(\epsilon)||p_\epsilon(\epsilon)]\rangle_{p(\varphi)}$ quantifies the information loss for predicting inputs using internal states. Therefore, by subtracting these three values from $H[p(x)]$, I obtain a mutual-information-like measure representing the prediction capability,

$$\begin{aligned} X[x;\varphi] &\equiv H[p(x)] - \langle H[p(\epsilon)]\rangle_{p(\varphi)} - \mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)] - \langle \mathcal{D}_{KL}[p(\epsilon)||p_\epsilon(\epsilon)]\rangle_{p(\varphi)} \\ &= \left\langle \log \frac{p^*(x,\varphi)}{p(x)p(\varphi)} \right\rangle_{p(x,\varphi)} \quad [\text{nat}], \end{aligned} \tag{19}$$

which I will refer to as predictive information as an analogy with Bialek's predictive information [29,30]. Their relationship is discussed in the next section. This predictive information $X[x;\varphi]$ is defined by replacing $p(x,\varphi)$ in $I[x;\varphi]$ with $p^*(x,\varphi)$. Thus, immediately, I obtain

$$\overline{F} = I[x;\varphi] - X[x;\varphi] \quad [\text{nat}]. \tag{20}$$

Hence, $\overline{F}$ represents a gap between the amount of information stored in the brain and that available for prediction, which is equivalent to the information loss in the generative model. It is interesting to note that the sum of losses in recognition and generative models $H[p(x)] - X[x;\varphi] = \overline{F} + \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ is an upper bounds of $\overline{F}$ because of the non-negativity of $\langle H[p(\epsilon)]\rangle_{p(\varphi)}$ (Fig. 2). However, since $\langle H[p(\epsilon)]\rangle_{p(\varphi)}$ is generally nonzero, $F(x) + \langle H[p(\epsilon)]\rangle_{p(\varphi)}$ may not reach zero even when $p(x,\varphi) = p^*(x,\varphi)$.

Furthermore, $X[x;\varphi]$ is transformed as

$$X[x;\varphi] = H[p(x)] - L_X - L_A, \tag{21}$$

where

$$L_X \equiv \langle -\log(p_\epsilon(\epsilon)dx)\rangle_{p(\epsilon)p(\varphi)} \tag{22}$$

is the so-called reconstruction error similar to that for principal component analysis (PCA) [39], while

$$L_A \equiv \mathcal{D}_{KL}[p(\varphi)||p_\varphi(\varphi)] \tag{23}$$

is an enhancement of Amari's cost function for independent component analysis (ICA) [40]. PCA is one of the most popular dimensionality reduction methods to remove background noise and extract important features from sensory inputs [10,39,41], while ICA is one of BSS methods to decompose a mixture set of sensory inputs into independent hidden sources [32,34,40,42,43]. Theoreticians hypothesize that the PCA- and ICA-like learning underlies BSS in the brain [3]. Equation (21) indicates that $X[x;\varphi]$ consists of the PCA- and ICA-like parts, i.e., maximization of $X[x;\varphi]$ can perform both dimensionality reduction and BSS (Fig. 2). Their relationships are discussed in the next section.

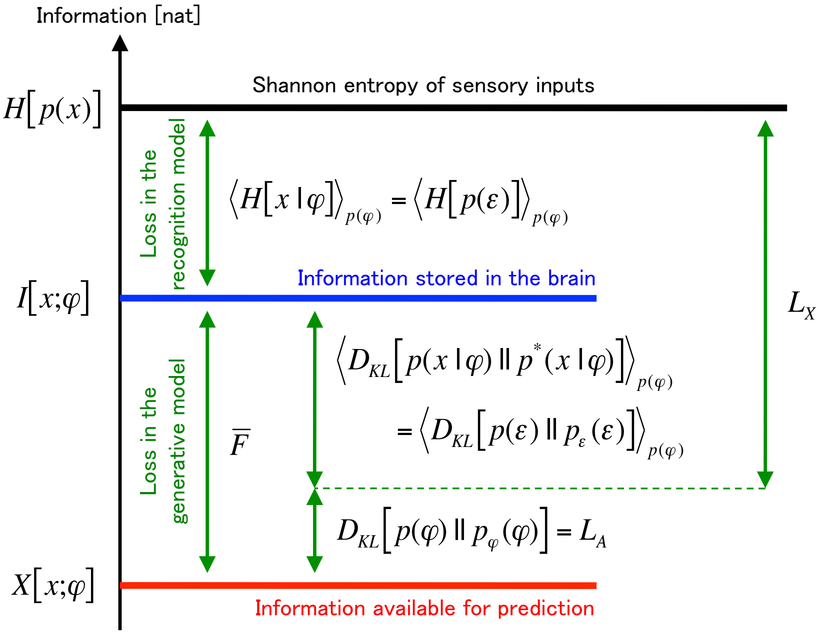
Information [nat]
$H[p(x)]$
Shannon entropy of sensory inputs
Loss in the recognition model
$\langle H[x|\varphi]\rangle_{p(\varphi)} = \langle H[p(\varepsilon)]\rangle_{p(\varphi)}$
Information stored in the brain
$I[x;\varphi]$
L_X
Loss in the generative model
$\overline{F}$
$\langle D_{KL}[p(x|\varphi) \| p^*(x|\varphi)]\rangle_{p(\varphi)}$
$= \langle D_{KL}[p(\varepsilon) \| p_\varepsilon(\varepsilon)]\rangle_{p(\varphi)}$
$D_{KL}[p(\varphi) \| p_\varphi(\varphi)] = L_A$
$X[x;\varphi]$
Information available for prediction

Figure 2. Schematic of information level. Relationship between free energy, mutual information, and predictive information is illustrated. Owing to the non-negativity of KLD, $\langle -\log p^*(x)\rangle_{p(x)}$ is always larger than or equal to $\langle -\log p(x)\rangle_{p(x)}$ and $F[p(\vartheta), x]$ provides an upper bound of $\langle -\log p^*(x)\rangle_{p(x)}$.

6. Comparison between the free-energy principle and related theories

In this section, I compare the FEP with other theories and methods. As describe in the above sections, the aim of the infomax principle is to maximize mutual information $I[x;\varphi]$ (Eq. (13)), while the aim of the FEP is to minimize free energy expectation $\overline{F}$ (Eq. (18)), and maximization of predictive information $X[x;\varphi]$ (Eq. (19)) means to do both of them simultaneously. Let us see how they are different from each other using a simple example.

6.1. Infomax principle

The generative process and recognition- and generative models defined in Section 2 are assumed. For simplification, suppose W, V and γ follow Dirac's delta functions; then, the goal of the infomax principle is simplified as maximization of mutual information between x and u,

$$I[x;u] = \left\langle \log \frac{p(x,u)}{p(x)p(u)} \right\rangle_{p(x,u)} = H[p(x)] - H[x|u] = H[p(u)] - H[u|x], \tag{24}$$

where $H[p(u)] = \langle -\log(p(u)du)\rangle_{p(u)}$ and $H[u|x] = \langle -\log(p(u|x)du)\rangle_{p(u,x)}$. If $\dim(x) \geq \dim(u)$ and a linear recognition model $u = g(x) = Wx$ with full-rank matrix W is considered, since $H[u|x] = 0$ and u has an infinite range, $I[x;u] = H[p(u)]$ monotonically increases as the variance of u increases. Thus, maximization of $I[x;u]$ cannot perform either PCA or ICA. To perform PCA and ICA based on the infomax principle, one needs to consider mutual information between sensory inputs and nonlinearly transformed neural outputs. When nonlinear neural outputs have a finite range, the variance of them should be maintained in the appropriate range. The infomax based PCA and ICA [42,43] are formulated based on this requirement. Mutual information between x and neural outputs transformed by an injective nonlinear function $\psi(\bullet)$, $\psi(u) = (\psi(u_1), \dots, \psi(u_N))^T$, is given by

$$I[x;\psi(u)] = \left\langle \log \frac{p(x,\psi(u))}{p(x)p(\psi(u))} \right\rangle_{p(x,\psi(u))} = H[p(\psi(u))] - H[\psi(u)|x], \tag{25}$$

where $H[p(\psi(u))] = \langle -\log(p(\psi(u))du) \rangle_{p(\psi(u))}$ and $H[\psi(u)|x] = \langle -\log(p(\psi(u)|x)du) \rangle_{p(\psi(u),x)}$. By the relationship of $p(\psi(u)) = |\partial u/\partial\psi(u)|p(u) = (\prod_i \psi'(u_i))^{-1}p(u)$, I have $H[p(\psi(u))] = \langle -\log\{(\prod_i \psi'(u_i))^{-1}p(u)du\} \rangle_{p(u)} = H[p(u)] + \langle \sum_i \log\psi'(u_i) \rangle_{p(u)}$. Since $H[\psi(u)|x] = 0$ hold, Eq. (25) becomes

$$I[x;\psi(u)] = H[p(u)] + \left\langle \sum_i \log\psi'(u_i) \right\rangle_{p(u)}. \tag{26}$$

As I will describe in the following, maximization of Eq. (26) performs PCA and ICA.

6.2. Principal component analysis

Both the infomax principle and the FEP give a cost function of PCA. Suppose $\dim(x) > \dim(u)$, $V = W^T$, and $-\log\psi'(u_i) = u_i^2$. From Eq. (24), $H[p(u)] = H[p(x)] - \langle H[p(\epsilon)] \rangle_{p(\varphi)}$ holds. Since the prediction error is given by $\epsilon = x - W^T u = (I - W^T W)x$, I have $\langle H[p(\epsilon)] \rangle_{p(\varphi)} = \langle -\log\{p(x)|\partial x/\partial\epsilon|dx\} \rangle_{p(x,\varphi)} = H[p(x)] + \langle \log|I - W^T W| \rangle_{p(\varphi)}$. Thus, Eq. (26) becomes

$$I[x;\psi(u)] = -\left\langle \log|I - W^T W| \right\rangle_{p(\varphi)} - \left\langle |u|^2 \right\rangle_{p(u)}. \tag{27}$$

The first term of Eq. (27) becomes the maximum if W holds $WW^T = I$ (i.e., an orthogonal matrix). To maximize the second term, outputs u need to be involved in a subspace spanned by the first to the Nth major principal components of x. Therefore, maximization of Eq. (27) performs PCA.

PCA is also derived by minimization of L_X (Eq. (22)) under the assumption that $p_\epsilon(\epsilon)$ is a Gaussian distribution $p_\epsilon(\epsilon) = \mathcal{N}[\epsilon; 0, \Pi_\epsilon]$ with precision matrix Π_ϵ (the inverse of covariance matrix). If I suppose $\Pi_\epsilon = \gamma_1 I + \gamma_2(\langle \epsilon\epsilon^T \rangle_{p(\epsilon)})^{-1}$ with positive hyper-parameters γ_1, γ_2, L_X becomes

$$L_X = \left\langle \frac{\gamma_1}{2}\langle \epsilon^T\epsilon \rangle_{p(\epsilon)} + \frac{\gamma_2}{2} - \frac{1}{2}\log\left|\gamma_1 I + \gamma_2(\langle \epsilon\epsilon^T \rangle_{p(\epsilon)})^{-1}\right| \right\rangle_{p(\varphi)} + \text{const.} \tag{28}$$

In the special case of $\gamma_2 = 0$, L_X becomes a common cost function for the least square error PCA [39] and auto-encoder [44], and its derivative $\partial L_X/\partial W$ is similar to the well-known Oja's subspace rule for PCA [41]. Moreover, since $\langle H[p(\epsilon)] \rangle_{p(\varphi)} = \langle \log|I - W^T W| \rangle_{p(\varphi)} + \text{const.} = \langle 1/2 \cdot \log|\langle \epsilon\epsilon^T \rangle_{p(\epsilon)}| \rangle_{p(\varphi)} +$ const., when the priors of W, V, and γ are flat and $1 \ll \gamma_1 \ll \gamma_2$, free energy expectation (Eq. (18)) approximately becomes

$$\begin{aligned} \overline{F} &= L_X - \langle H[p(\epsilon)] \rangle_{p(\varphi)} + \mathcal{D}_{KL}[p(u)||p_0(u)] \\ &= \left\langle \frac{\gamma_1}{2}\langle \epsilon^T\epsilon \rangle_{p(\epsilon)} + \frac{\gamma_2}{2} - \frac{1}{2}\log\left|\gamma_1\langle \epsilon\epsilon^T \rangle_{p(\epsilon)} + \gamma_2 I\right| \right\rangle_{p(\varphi)} + \mathcal{D}_{KL}[p(u)||p_0(u)] + \text{const.} \\ &\approx \left\langle \frac{\gamma_1}{2}\langle \epsilon^T\epsilon \rangle_{p(\epsilon)} + \frac{\gamma_2}{2} \right\rangle_{p(\varphi)} + \text{const.} \end{aligned} \tag{29}$$

Therefore, $\overline{F}$ is approximately transformed as $\overline{F} \approx L_X + \text{const.}$

6.3. Independent component analysis

Both the infomax principle and the FEP give a cost function of ICA. Suppose that sources $s_1, \dots, s_N$ independently follow an identical distribution $p_0(s_i)$. The infomax based ICA is derived from Eqs. (25)-(26) [42,43]. If $\psi(u_i)$ is defined to satisfy $\psi'(u_i) = p_0(u_i)$, negative mutual information $-I[x;\psi(u)]$ becomes KLD between actual and prior distributions up to constant term,

$$-I[x;\psi(u)] - \log du = \Big\langle \log p(u) - \log p_0(u) \Big\rangle_{p(u)} = \mathcal{D}_{KL}[p(u)||p_0(u)] \equiv L_A. \tag{30}$$

L_A is known as Amari's ICA cost function [40]. While both $-I[x;\psi(u)]$ and L_A provide the same gradient descent rule, the nonlinearly transformed neural outputs $\psi(u)$ are required to formulate $I[x;\psi(u)]$. By contrast, L_A straightforwardly represents that minimization of KLD between $p(u)$ and $p_0(u)$ performs ICA similarly to the FEP. Indeed, if $\dim(u) = \dim(x) = N$, $u = g(x)$ is an injective function, and the priors of W, V, and γ are flat, I obtain $\overline{F} = \mathcal{D}_{KL}[p(u)||p_0(u)] = L_A$. Therefore, ICA is a subset of the inference problem considered in the FEP, and the derivation from the FEP is simpler while both the infomax principle and the FEP can perform ICA.

Furthermore, when $\dim(x) > \dim(u)$, minimization of $\overline{F}$ can perform both dimensionality reduction and BSS. When the priors of W, V, and γ are flat and $\gamma_1 \ll \gamma_2$, free energy expectation (Eq. (18)) approximately becomes

$$\overline{F} \approx \Big\langle \frac{\gamma_1}{2} \langle \epsilon^T \epsilon \rangle_{p(\epsilon)} + \frac{\gamma_2}{2} \Big\rangle_{p(\varphi)} + L_A + \text{const.} \tag{31}$$

Therefore, $\overline{F}$ is approximately transformed as $\overline{F} \approx L_X + L_A + \text{const.}$ and can switch the weights of PCA- and ICA parts by controlling γ_1. Whereas, if γ has a sufficient dimension and $\Pi_\epsilon(\gamma)$ is fine tuned to minimize $\overline{F}$, I get $\Pi_\epsilon = (\langle \epsilon\epsilon^T \rangle_{p(\epsilon)})^{-1}$ by solving $\partial \overline{F} / \partial \Pi_\epsilon = 0$. Under this condition, since L_A is equal to $H[x|u]$ up to constant term, I find

$$\overline{F} = L_A + \text{const.} \tag{32}$$

Thus, $\overline{F}$ consists only of the ICA part when $\Pi_\epsilon(\gamma)$ is fine tuned.

6.4. Predictive information

Predictive information is a measure proposed by Bialek to quantify the average generalization power of sensory inputs [29,30], which is defined by

$$I_p[x_{future}; x_{past}] \equiv \left\langle \log \frac{p^*(x_{future}, x_{past})}{p^*(x_{future}) p(x_{past})} \right\rangle_{p(x_{future}, x_{past})}, \tag{33}$$

where x_{future} and x_{past} indicate future and past sensory inputs, respectively. Note that $p^*(x_{future}, x_{past})$ and $p^*(x_{future})$ are the likelihood function (a statistical model) and the prior distribution, respectively, while $p(x_{past})$ and $p(x_{future}, x_{past})$ are true probability distributions. If I suppose that the internal state φ represents information based on the past observation while x represents the current sensory inputs, Bialek's predictive information $I_p[x;\varphi]$ becomes

$$I_p[x;\varphi] = \left\langle \log \frac{p^*(x,\varphi)}{p^*(x)p(\varphi)} \right\rangle_{p(x,\varphi)} = \left\langle \log \frac{p_\epsilon(\epsilon) p_\varphi(\varphi)}{p^*(x)p(\varphi)} \right\rangle_{p(x,\varphi)}, \tag{34}$$

While this definition of $I_p[x;\varphi]$ supposes that x exactly follows $p(x) = p^*(x)$, it is difficult to directly know and mimic the exact shape of $p(x)$ in practice. If I suppose $p^*(x)$ can be different from $p(x)$, I obtain $X[x;\varphi]$ as a lower bound of $I_p[x;\varphi]$,

$$I_p[x;\varphi] \geq X[x;\varphi]. \tag{35}$$

If and only if I can design $p^*(x)$ as the exactly same shape as $p(x)$, $I_p[x;\varphi] = X[x;\varphi]$ holds, while $I_p[x;\varphi] > X[x;\varphi]$ when $p^*(x)$ is different from $p(x)$ because of the non-negativity of KLD. Therefore, $X[x;\varphi]$ is a generalized measure of $I_p[x;\varphi]$.

6.5. Simulation

The difference between the infomax principle and the FEP is illustrated by a simple simulation using a linear generative model and a linear neural network (Fig. 3). For simplification, I assume that dynamics of u quickly converge to the optimum that minimizes $F(x)$ compared to the change of s (adiabatic approximation). First, when W is randomly chosen and V is defined by $V = W^T$, both $H[x|\varphi]$ and $\overline{F}$ are scattered (black circles in Fig. 3) since neural outputs represent random mixtures of sources and noises. Next, when W is optimized according to either Eq. (27) or (28) under the condition where $V = W^T$, neural outputs express major principal components of inputs (i.e., PCA; blue circles in Fig. 3). This is the case where $H[x|\varphi]$ is minimized; thus, PCA performs the infomax optimization. Whereas, when W, V and $\Pi_\epsilon(\gamma)$ are optimized according to the FEP (see Eq. (32)), neural outputs represent independent components that match to the prior source distribution, i.e., performing BSS (i.e., ICA), while minimizing the prediction error (red circles in Fig. 3). For the linear generative process as shown in Fig. 3, minimization of $\overline{F}$ can reliably and accurately perform both dimensionality reduction and BSS, because outputs become independent of each other and match the prior belief if and only if outputs represent true sources up to permutation and sign-flip. Since $X[x;\varphi]$ consists of PCA- and ICA- cost functions (see Eq. (21)), maximization of $X[x;\varphi]$ finds a solution that intermediates between solutions of the infomax principle and the FEP. Interestingly, the infomax optimization (i.e., PCA) provided W that makes $\overline{F}$ closer to zero than random states; i.e., the infomax optimization can contribute free energy minimization. Note that, in the case of the nonlinear system, there are many different transformations that make outputs independent of each other [45]. Hence, there is no guarantee that minimization of $\overline{F}$ can identify true sources of nonlinear generative models.

In sum, the aims of the FEP, the infomax principle, and the predictive information theory are similar to each other; especially, when both of sources and noises follow Gaussian distributions, their aims become the same meaning. By contrast, the optimal synaptic weights for the FEP can be different from that for the infomax principle when sources follow non-Gaussian distributions. Under this condition, the predictive information theory finds an intermediate solution between those for the FEP and the infomax principle.

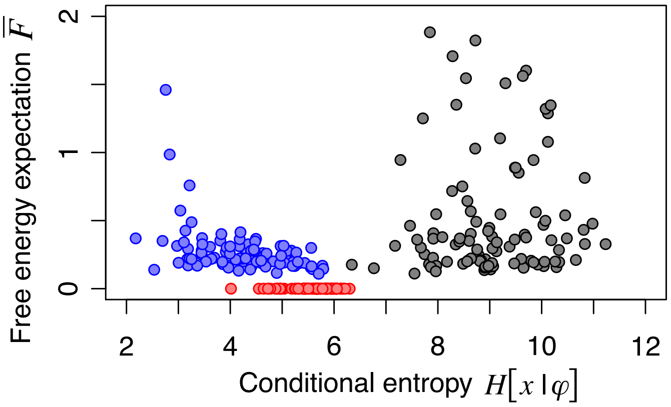

Figure 3. The difference between the infomax principle and the FEP when sources follow a non-Gaussian distribution. Suppose s as two-dimensional hidden sources following an identical Laplace distribution with zero mean and unit variance; x as four-dimensional sensory inputs; u as two-dimensional neural outputs; z as four-dimensional background Gaussian noises following $\mathcal{N}[z; 0, \Pi_z]$; θ as a 4×2-dimensional mixing matrix; W as a 2×4-dimensional synaptic strength matrix for the bottom-up path; V as a 4×2-dimensional synaptic strength matrix for the top-down path; and the priors of W, V, and γ as flat priors. Sensory inputs were determined by $x = \theta s + z$, while neural outputs were determined by $u = Wx$. The prediction error was given by $\epsilon = x - Vu$ and used to calculate $H[p(\epsilon)]$ and L_A. Horizontal and vertical axes are conditional entropy $H[x|\varphi]$ (Eq. (14)) and free energy expectation $\overline{F}$ (Eq. (18)), respectively. Black, blue, and red circles indicate the results when W is a random matrix, optimized for the infomax principle (i.e., PCA), and optimized for the FEP, respectively. Simulations were conducted 100 times with randomly selected θ and Π_z for each condition. For each simulation, 10^8 random sample points were generated and probability distributions were calculated by the histogram method.

7. Discussion

In this study, the FEP is linked with the infomax principle and the predictive coding theory. It is more likely that the purpose of the brain is to minimize the surprise of sensory inputs to realize better perception rather than maximize the amount of stored information. For example, the visual input captured by a video camera contributes to the stored information, but it cannot be used for prediction directly. Whereas, the brain is capable of inference and prediction using stored information. Surprise expectation $\overline{S}(\geq 0)$ represents the difference between actual observation and prediction under the statistical model, and free energy expectation $\overline{F}$ provides its upper bound. Predictive information $X[x; \varphi]$ is introduced to quantify the prediction and generalization capability of sensory inputs, which is defined by slightly modifying the definition in the previous studies [29,30]. Using this, $\overline{F}$ is explained as a gap between information stored in the brain $I[x, \varphi]$ and that available for prediction $X[x; \varphi]$ (Eq. (20)).

Moreover, the derivation of ICA is simplified by the FEP. To perform ICA based on the infomax principle, one needs to tune the nonlinearity of neural outputs such that its derivative matches the prior distribution. By contrast, under the FEP, ICA is straightforwardly derived from KLD between the true probability distribution and the prior distribution of u. Especially, in the absence of background noise and prior knowledge on parameters and hyper-parameters, free energy expectation $\overline{F}$ (Eq. (18)) is equivalent to surprise expectation $\overline{S}$ (Eq. (16)) and Amari's ICA cost function L_A (Eq. (30)). Thus, ICA is a subproblem of the FEP.

The FEP is a useful theory from theoretical and engineering view points, since various learning rules can be derived from common cost function $F(x)$ [14,15]. However, to be a physiologically plausible theory of the brain, the FEP needs to satisfy certain physiological requirements. There are two major requirements: first, physiological evidence that shows the existence of learning or self-organizing processes under the FEP is required. The model structure under the FEP is consistent with previous biological knowledge and proposes the possible function of the cortical microcircuits [18]. Moreover, BSS performed by *in vitro* neural networks reduce free energy in the network [46], and the spontaneous prior activity of a visual area is known to learn the properties of natural pictures [47]. These results suggest the physiological plausibility of the FEP. Nevertheless, further experiments and consideration of information theoretical optimization under physiological constraints [48] are required to prove the existence of the FEP in the biological brain. Second, the update rule must be a biologically plausible local learning rule; i.e., synaptic strengths much be changed by signals from connected inputs. While the synaptic update rule for the discrete system is local [16], the current rule for the continuous system [14] is a non-local rule. Recently developed biologically-plausible three-factor learning models in which Hebbian learning is mediated by the third modulatory factor [49–51] may help to understand the neuronal mechanism underling unconscious inference and the FEP. Therefore, it is necessary to investigate how actual neural networks infer the dynamical system behind the sensory input. This will help develop a biologically plausible learning algorithm through which the actual neural network might develop the internal model in a manner consistent with the physiological experimental observations.

In summary, I investigated the differences between two types of information—information stored in the brain and that available for prediction. It was demonstrated that free energy represents the gap between these two information. This result clarified the difference between the FEP and related theories and will utilize for understanding unconscious inference from theoretical view points.

Acknowledgments: This work was supported by RIKEN Brain Science Institute.

Conflicts of Interest: The author declares no competing financial interests. The founding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. DiCarlo, J.J.; Zoccolan, D.; Rust, N.C. How does the brain solve visual object recognition? *Neuron* **2012**, *73*, 415-434.
2. Bronkhorst, A.W. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acustica united with Acustica* **2000**, *86*, 117-128.
3. Brown, G.D.; Yamada, S.; Sejnowski, T.J. Independent component analysis at the neural cocktail party. *Trends in neurosciences* **2001**, *24*, 54-63.
4. Haykin, S.; Chen, Z. The cocktail party problem. *Neural Comput* **2005**, *17*, 1875-1902.
5. Narayan, R.; Best, V.; Ozmeral, E.; McClaine, E.; Dent, M.; Shinn-Cunningham, B.; Sen, K. Cortical interference effects in the cocktail party problem. *Nat Neurosci* **2007**, *10*, 1601-1607.
6. Mesgarani, N.; Chang, E.F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **2012**, *485*, 233-236.
7. Golumbic, E.M.Z.; Ding, N.; Bickel, S.; Lakatos, P.; Schevon, C.A.; McKhann, G.M.; Schroeder, C.E. Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron* **2013**, *77*, 980-991.
8. Dayan, P.; Abbott, L.F. *Theoretical neuroscience: computational and mathematical modeling of neural systems*; MIT Press, London, 2001.
9. Gerstner, W.; Kistler, W. *Spiking Neuron Models. Single Neurons, Populations, Plasticity*; Cambridge University Press, Cambridge, 2002.
10. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer Verlag, 2006.
11. von Helmholtz, H.; Southall, J.P.C. *Treatise on physiological optics (Vol. 3)*; Courier Corporation, 2005.
12. Dayan, P.; Hinton, G.E.; Neal, R.M.; Zemel, R.S. The helmholtz machine. *Neural Comput* **1995**, *7*, 889-904.
13. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *Journal of Physiology-Paris* **2006**, *100*, 70-87.
14. Friston, K.J. Hierarchical model in the brain. *PLoS Comput Biol* **2008**, *4*, e1000211.
15. Friston, K. The free-energy principle: a unified brain theory? *Nat Rev Neurosci* **2010**, *11*, 127-138.

16. Friston, K.; FitzGerald, T.; Rigoli, F.; Schwartenbeck, P.; Pezzulo, G. Active inference: A process theory. *Neural Comput* **2017**, *29(1)*, 1-49.
17. George, D.; Hawkins, J. Towards a mathematical theory of cortical micro-circuits. *PLoS Comput Biol*, **2009**, *5*, e1000532.
18. Bastos, A.M.; Usrey, W.M.; Adams, R.A.; Mangun, G.R.; Fries, P.; Friston, K.J. Canonical microcircuits for predictive coding. *Neuron* **2012**, *76*, 695-711.
19. Rao, R.P.; Ballard, D.H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* **1999**, *2*, 79-87.
20. Friston, K. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* **2005**, *360*, 815-836.
21. Kilner, J.M.; Friston, K.J.; Frith, C.D. Predictive coding: an account of the mirror neuron system. *Cognitive Processing* **2007**, *8*, 159-166.
22. Friston, K.; Mattout, J.; Kilner, J. Action understanding and active inference. *Biological Cybernetics*, **2011**, *104*, 137-160.
23. Friston, K.J.; Frith, C.D. Active inference, communication and hermeneutics. *Cortex* **2015**, *68*, 129-143.
24. Friston, K.; Frith, C. A duet for one. *Consciousness and Cognition* **2015**, *36*, 390-405.
25. Fletcher, P.C.; Frith, C.D. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nat Rev Neurosci*, **2009**, *10*, 48-58.
26. Friston, K.J.; Stephan, K.E.; Montague, R.; Dolan, R.J. Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, **2014**, *1*, 148-158.
27. Linsker, 1992. The title of the cited article. *Journal Abbreviation* **2008**, *10*, 142-149.
28. Lee, T.W.; Girolami, M.; Bell, A.J.; Sejnowski, T.J. A unifying information-theoretic framework for independent component analysis. *Comput Math Appl* **2000**, *39*, 1-21.
29. Bialek, W.; Tishby, N. Predictive information. *arXiv* **1999**, arXiv:cond-mat/9902341.
30. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, complexity, and learning. *Neural Comput* **2001**, *13(11)*, 2409-2463.
31. Belouchrani, A.; Abed-Meraim, K.; Cardoso, J.F.; Moulines, E. A blind source separation technique using second-order statistics. *Signal Processing IEEE Trans on* **1997**, *45*, 434-444.
32. Choi, S.; Cichocki, A.; Park, H.M.; Lee, S.Y. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews* **2005**, *6*, 1-57.
33. Cichocki, A.; Zdunek, R.; Phan, A.H.; Amari, S.I. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*; John Wiley & Sons, 2009.
34. Comon, P.; Jutten, C. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*; Academic Press, 2010.
35. Shannon, C.E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press, 1998.
36. Arora, S.; Risteski, A. Provable benefits of representation learning. *arXiv*, arXiv:1706.04601.
37. Jaynes, E.T. Information theory and statistical mechanics. *Physical Review*, **1957**, *106*, 620-630.
38. Jaynes, E.T. Information theory and statistical mechanics. II. *Physical Review*, **1957**, *108*, 171-190.
39. Xu, L. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural Netw* **1993**, *6*, 627-648.
40. Amari, S.I.; Cichocki, A.; Yang, H.H. A new learning algorithm for blind signal separation. *Adv Neural Inf Proc Sys* **1996**, *8*, 757-763.
41. Oja, E. Neural networks, principal components, and subspaces. *Int J Neural Syst* **1989**, *1*, 61-68.
42. Bell, A.J.; Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* **1995**, *7*, 1129-1159.
43. Bell, A.J.; Sejnowski, T.J. The “independent components” of natural scenes are edge filters. *Vision Res* **1997**, *37*, 3327-3338.
44. Hinton, G.E.; Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504-507.
45. Hyvärinen, A.; Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw* **1999**, *12*, 429-439.
46. Isomura, T.; Kotani, K.; Jimbo, Y. Cultured Cortical Neurons Can Perform Blind Source Separation According to the Free-Energy Principle. *PLoS Comput Biol* **2015**, *11*, e1004643.

47. Berkes, P.; Orbán, G.; Lengyel, M.; Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **2011**, *331*, 83-87.
48. Sengupta, B.; Stemmler, M.B.; Friston, K.J. Information and efficiency in the nervous system–a synthesis. *PLoS Comput Biol* **2013**, *9*, e1003157.
49. Frémaux, N.; Gerstner, W. Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Front Neural Circuits* **2016**, *9*, doi:10.3389/fncir.2015.00085.
50. Isomura, T.; Toyoizumi, T. A Local Learning Rule for Independent Component Analysis. *Sci Rep* **2016**, *6*, 28073.
51. Kuśmierz, L.; Isomura, T.; Toyoizumi T. Learning with three factors: modulating Hebbian plasticity with errors. *Current Opinion in Neurobiology* **2017**, *46*, 170-177.