entropy

MDPI

Article

Neural Dynamics under Active Inference: Plausibility and Efficiency of Information Processing

Lancelot Da Costa [1,2,*], Thomas Parr [2], Biswa Sengupta [2,3,4] and Karl Friston [2]

[1] Department of Mathematics, Imperial College London, London SW7 2AZ, UK
[2] Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, UK; thomas.parr.12@ucl.ac.uk (T.P.); b.sengupta@imperial.ac.uk (B.S.); k.friston@ucl.ac.uk (K.F.)
[3] Core Machine Learning Group, Zebra AI, London WC2H 8TJ, UK
[4] Department of Bioengineering, Imperial College London, London SW7 2AZ, UK
* Correspondence: l.da-costa@imperial.ac.uk

Abstract: Active inference is a normative framework for explaining behaviour under the free energy principle—a theory of self-organisation originating in neuroscience. It specifies neuronal dynamics for state-estimation in terms of a descent on (variational) free energy—a measure of the fit between an internal (generative) model and sensory observations. The free energy gradient is a prediction error—plausibly encoded in the average membrane potentials of neuronal populations. Conversely, the expected probability of a state can be expressed in terms of neuronal firing rates. We show that this is consistent with current models of neuronal dynamics and establish face validity by synthesising plausible electrophysiological responses. We then show that these neuronal dynamics approximate natural gradient descent, a well-known optimisation algorithm from information geometry that follows the steepest descent of the objective in information space. We compare the information length of belief updating in both schemes, a measure of the distance travelled in information space that has a direct interpretation in terms of metabolic cost. We show that neural dynamics under active inference are metabolically efficient and suggest that neural representations in biological agents may evolve by approximating steepest descent in information space towards the point of optimal inference.

Keywords: active inference; free energy principle; process theory; natural gradient descent; information geometry; variational Bayesian inference; Bayesian brain; self-organisation; metabolic efficiency; Fisher information length

Citation: Da Costa, L.; Parr, T.; Sengupta, B.; Friston, K. Neural Dynamics under Active Inference: Plausibility and Efficiency of Information Processing. *Entropy* **2021**, *23*, 454. https://doi.org/10.3390/e23040454

Academic Editor: Hermann Haken

Received: 16 March 2021
Accepted: 6 April 2021
Published: 12 April 2021

1. Introduction

Active inference is a normative framework for explaining behaviour under the free energy principle, a theory of self-organisation originating in neuroscience [1–4] that characterises certain systems at steady-state as having the appearance of sentience [5,6]. Active inference describes agents' behaviour as following the equations of motion of the free energy principle so as to remain at steady-state, interpreted as the agent's goal [7].

Active inference describes organisms as inference engines. This assumes that organisms embody a generative model of their environment. The model encodes how the states external to the agent influence the agent's sensations. Organisms infer their surrounding environment from sensory data by inverting the generative model through minimisation of variational free energy. This corresponds to performing approximate Bayesian inference (also known as variational Bayes) [2,3,8–11], a standard method in machine learning, or minimising the discrepancy between predictions and sensations [1,12]. Active inference unifies many existing theories of brain function [13,14], such as, for example, optimal control [15–18], the Bayesian brain hypothesis [19–21] and predictive coding [19,22–24]. It has been used to simulate a wide range of behaviours in neuropsychology, machine learning and robotics. These include planning and navigation [25–31], exploration [32–36], learning

of self and others [37,38], concept learning [39–41], playing games [42–44], adapting to changing environments [45–47], active vision [48–51] and psychiatric illness [42,52–55]. Active inference agents show competitive or state-of-the-art performance in a wide variety of simulated environments [34,46,47,56].

The last decade has seen the development of a theory of how the brain might implement active inference consistently in neurobiology [1,7,49,57,58]. A number of predictions of this theory have been empirically validated, including the role of dopamine in decision-making [59,60] and free energy minimisation in exploration and choice behaviour [61,62]. This paper characterises the neural dynamics of this process theory from two complementary standpoints: (1) consistency with empirically driven models of neural population dynamics, and (2) the metabolic and computational efficiency of such dynamics.

Efficiency is an important aspect of neural processing in biological organisms [63–68] and an obvious desideratum for artificial agents. Efficiency of neural processing in biological agents is best seen in the efficient coding hypothesis by Horace Barlow [69–71], which has received much empirical support [65,66,72–77] (see, in particular, [64] for energetic efficiency) and has been a successful driver in computational neural modelling [75,78–80]. In brief, any process theory of brain function should exhibit reasonably efficient neural processing.

Active inference formalises perception as inferring the state of the world given sensory data through minimisation of variational free energy. This amounts to constantly optimising Bayesian beliefs about states of the outside world in relation to sensations. By beliefs, we mean probability distributions over states of the environment. These distributions score the extent to which an agent trusts that the environment is, or not, in this or another state. From an information theoretic viewpoint, a change of beliefs is a computation or a change in information encoded by the agent.

This belief updating has an associated energetic cost. Landauer famously observed that a change in information entails heat generation [81,82]. It follows that the energetics needed for a change in beliefs may be quantified by the change in information encoded by the agent over time (as the organism has to alter, e.g., synaptic weights, or restore transmembrane potentials) mathematically scored by the length of the path travelled by the agent's beliefs in information space. There is a direct correspondence between the (Fisher) information length of a path and the energy consumed by travelling along that path [83,84]. To ensure metabolic and computational efficiency, an efficient belief updating algorithm should reach the free energy minimum (i.e., the point of optimal inference) via the shortest possible path on average. Furthermore, since an agent does not know the free energy minimum in advance, she must find it using only local information about the free energy landscape. This is a non-trivial problem. Understanding how biological agents solve it might not only improve our understanding of the brain, but also yield useful insights into mathematical optimisation and machine learning.

In the first part of this work, we show that the dynamics prescribed by active inference for state-estimation are consistent with current models of neural population dynamics. We then show that these dynamics approximate natural gradient descent on free energy, a well-known optimisation algorithm from information geometry that follows the steepest descent of the objective in information space [85]. This leads to short paths for belief updates as the free energy encountered in discrete state-estimation is convex (see Appendix A). These results show that active inference prescribes efficient and biologically plausible neural dynamics for state-estimation and suggest that neural representations may be collectively following the steepest descent in information space towards the point of optimal inference.

2. The Softmax Activation Function in Neural Population Dynamics

This section rehearses a basic yet fundamental feature of mean-field formulations of neural dynamics; namely, the average firing rate of a neural population follows a sigmoid function of the average membrane potential. It follows that, when considering multiple competing neural populations, relative firing rates can be expressed as a softmax function

of average transmembrane potentials, as the softmax generalises the sigmoid in attributing probabilities to multivariate (resp. binary) inputs.

The sigmoid relationship between membrane potential and firing rate was originally derived by Wilson and Cowan [86], who showed that any unimodal distribution of thresholds within a neural population, whose individual neurons are modelled as a Heaviside response unit, results in a sigmoid activation function at the population level. This is because the population's activation function equals a smoothing (i.e., a convolution) of the Heaviside function with the distribution of thresholds.

The assumption that the sigmoid arises from the distribution of thresholds in a neural population remained unchallenged for many years. However, the dispersion of neuronal thresholds is, quantitatively, much less important than the variance of neuronal membrane potential within populations [87]. Marreiros and colleagues showed that the sigmoid activation function can be more plausibly motivated by considering the variance of neuronal potentials within a population [88], which is generally modelled by a Gaussian distribution under the Laplace assumption in mean-field treatments of neural population dynamics [89]. Briefly, with a low variance on neuronal states, the sigmoid function that is obtained—as a convolution of the Heaviside function—has a steep slope, which means that the neural population, as a whole, fires selectively with respect to the mean membrane potential, and vice-versa. This important fact, which was verified experimentally using dynamic causal modelling [88,90], means that the variance of membrane potentials implicitly encodes the (inverse) precision of the information encoded within the population.

Currently, the sigmoid activation function is the most commonly used function to relate average transmembrane potential to average firing rate in mean-field formulations of neural population dynamics [91,92] and deep neural networks [93,94].

It follows that when considering multiple competing neural populations, the softmax of their average membrane potentials may be used to express their relative firing rates. This can be seen from the fact that the softmax function generalizes the sigmoid in attributing probabilities in multivariate (resp. bivariate) classification tasks (p. 198) [93]. We refer to Section 5.2, Chapters 3 and 6 in [95–97] for mathematical derivations of this generalisation.

3. Neural Dynamics of Perceptual Inference

Active inference formalises perception as inferring the state of the world given sensory data through the minimisation of variational free energy [7]. For state estimation on discrete state-space generative models (e.g., partially observable Markov decision processes [98]), the free energy gradient corresponds to a generalised a prediction error [7]. This means that to infer the states of their environment, biological agents reduce the discrepancy between their predictions of the environment and their observations, or maximise the mutual information between them [63].

Variational free energy is a function of approximate posterior beliefs Q,

$$\begin{aligned} \mathrm{F}(Q) &\triangleq E_{Q(s)}\left[\ln Q(s) - \ln P(o,s)\right] \\ &= \underbrace{D_{KL}\left[Q(s) \,\|\, P(s \mid o)\right]}_{\geq 0} - \underbrace{\ln P(o)}_{\text{Log-evidence}} \\ &= \underbrace{D_{KL}\left[Q(s) \,\|\, P(s)\right]}_{\text{Complexity}} - \underbrace{E_{Q(s)}\left[\ln P(o \mid s)\right]}_{\text{Accuracy}} \end{aligned}$$

While P is the generative model: a probability distribution over hidden states (s) and observations (o) that encodes the causal relationships between them. Only past and present observations are directly accessible; hidden states can only be inferred. This means that the free energy depends on the sequence of available stimuli and is updated whenever a new observation is sampled. The symbol E_Q means the expectation (i.e., the average) of its argument under the subscripted distribution. D_{KL} is known as the Kullback–Leibler divergence or relative entropy [99–101] and is used as a non-negative measure of the discrepancy between two probability distributions. Note that this is not a measure of

distance, as it is asymmetric. The second line here shows that minimising free energy amounts to approximating the posterior over hidden states, which are generally intractable to compute directly, with the approximate posterior. Exact Bayesian inference requires the approximate and true posterior to be exactly the same, at which point free energy becomes negative log model evidence (a.k.a., marginal likelihood). This explains why the (negative) free energy is sometimes referred to as an evidence lower bound (ELBO) in machine learning [93]. The final line shows a decomposition of the free energy into accuracy and complexity, underlying the need to find the most accurate and minimally complex explanation for sensory observations (c.f., Horace Barlow's principle of minimum redundancy [69]).

When a biological organism represents some of its environment in terms of a finite number of possible states (e.g., the locations in space encoded by place cells), we can specify its belief by updating about the current state in peristimulus time. Discrete state-estimation is given as a (softmax) function of accumulated negative free energy gradients [57].

$$\begin{aligned} \dot{v} &= -\nabla_{\mathbf{s}} \mathrm{F} \\ \mathbf{s} &= \sigma(v) \end{aligned}$$

In this equation, σ is a softmax function and s represents the agent's beliefs about states. (These are the parameters of a categorical distribution Q over states). Explicitly, **s** is a vector whose i-th component is the agent's confidence (expressed as a probability) that is in the i-th state. The softmax function is the natural choice to map from free energy gradients to beliefs as the former turns out to be a logarithm [7] and the components of the latter must sum to one.

Just as neuronal dynamics involve translation from post-synaptic potentials to firing rates, these dynamics involve translating from a vector of real numbers (v), to a vector where components are bounded between zero and one (**s**). As such, we can interpret v as the voltage potential of neuronal populations, and **s** as representing their firing rates (since these are upper bounded thanks to neuronal refractory periods). Note the softmax function here plays the same role as in mean-field formulations; it translates average potentials to firing rates. On the one hand, this view is consistent with models of neuronal population dynamics. On the other hand, it confers post-hoc face validity, as it enables the synthesis of plausible local field potentials (see Figure 1) and a wide range of other electrophysiological responses, including repetition suppression, mismatch negativity, violation responses, place-cell activity, phase precession, theta-gamma coupling, and more [57].

The idea that state-estimation is expressed in terms of firing rates is well-established when the state-space constitutes an internal representation of space. This is the raison d'être of the study of place cells [102], grid cells [103] and head-direction cells [104,105], where the states inferred are physical locations in space. Primary afferent neurons in cats have also been shown to encode kinematic states of the hind limb [106–108]. Most notably, the seminal work of Hubel and Wiesel [109] showed the existence of neurons encoding orientation of visual stimuli. In short, the very existence of receptive fields in neuroscience speaks to a carving of the world into discrete states under an implicit discrete state generative model. While many of these studies focus on single neuron recordings, the arguments presented above are equally valid and generalise the case of "populations" comprising a single neuron.

In summary, the neuronal dynamics associated with state-estimation in active inference are consistent with mean-field models of neural population dynamics. This view is strengthened a posteriori, as this allows one to generate a wide range of plausible electrophysiological responses. Yet, further validation remains to be carried out by testing these electrophysiological responses empirically. We will return to this in the discussion.

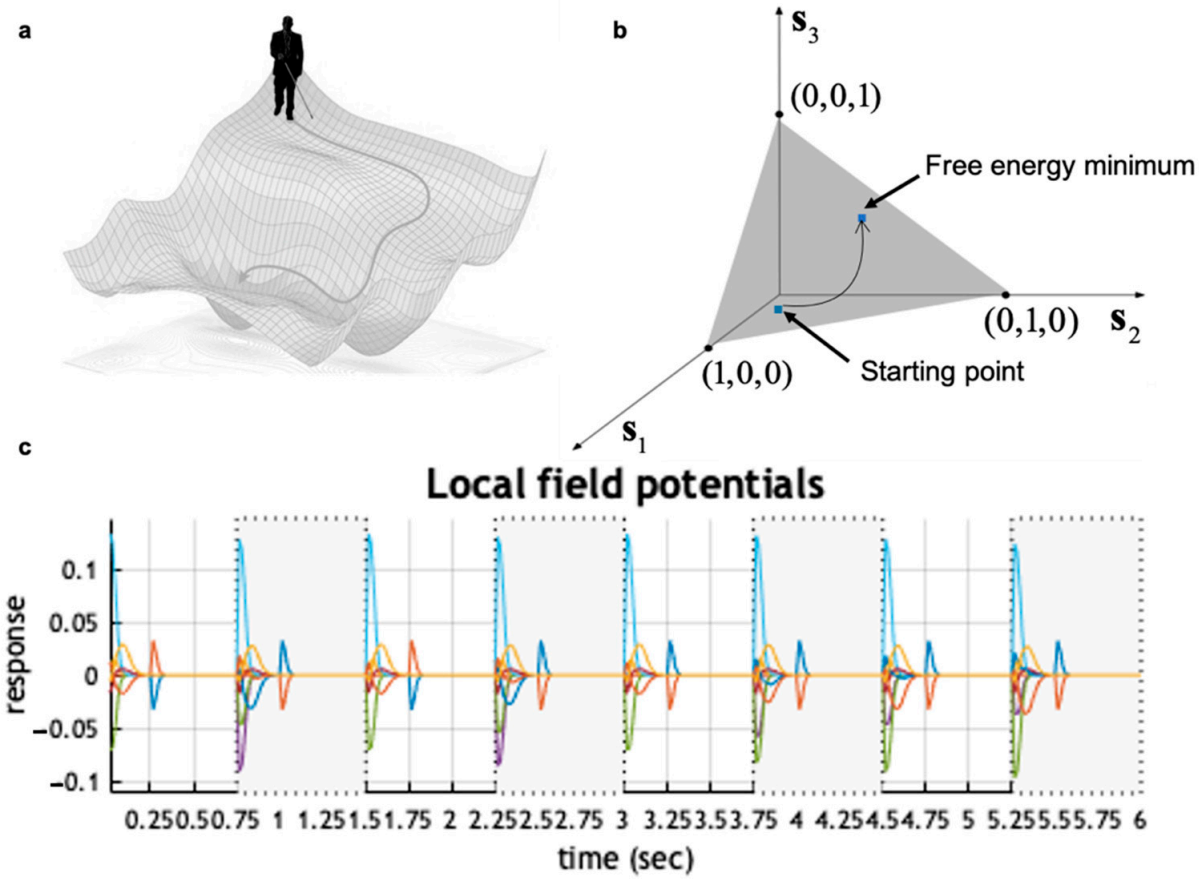

Figure 1. Active inference for state estimation (discrete state-space). Panel (**a**) summarises the problem of finding the minimum of a function (e.g., the free energy). One possibility would be taking the shortest path, which involves climbing up a hill; however, in the nescience of the minimum, a viable strategy consists of myopically taking the direction of steepest descent. In panel (**b**), we depict an example of a trajectory of an agent's beliefs during the process of perception, which consists of updating beliefs about the states of the external world to reach the point of optimal inference (i.e., free energy minimum). In this example, the state-space comprises only three states (e.g., three different locations in a room). As they are probabilities over states, the components of s are non-negative and sum to one; hence, the agent's beliefs naturally live on a triangle in three-dimensional space. Mathematically, this object is called a (two-dimensional) simplex. This constitutes the belief space, or information space, on which the free energy is defined. Technically, this object is a smooth statistical manifold, which corresponds to the set of parameters of a categorical distribution. To optimise metabolic and computational efficiency, agents must update their beliefs to reach the free energy minimum via the shortest possible path on this manifold. In panel (**c**), we exhibit simulated local field potentials that arise by interpreting the rate of change of v in terms of depolarisations, over a sequence of eight observations (e.g., saccadic eye-movements). As the rate of change is given by the free energy gradients, the decay of these local field potentials to zero coincides with the reaching of the free energy minimum (at which the gradient is zero by definition). Each colour expresses the mean voltage potential of a neural population, which encodes beliefs about one hidden state. These were obtained during the first eight trials of the T-Maze numerical simulation described in Section 6.1.1. Local field potentials accompanying the rule learning simulation—and for subsequent trials of the T-Maze task—can be found in [39,57], respectively. For more details on the generation of simulated electrophysiological responses, see [57].

4. A Primer on Information Geometry and Natural Gradient Descent

To assess the computational and metabolic efficiency of a belief trajectory, it becomes necessary to formalise the idea of "belief space". These are well-studied structures in the field of information geometry [110–112], called statistical manifolds. In our case, these are (smooth) manifolds, where each point corresponds to a parameterisation of the probability distribution in consideration (see Figure 2). One is then licensed to talk about a change in beliefs as a trajectory on a statistical manifold.

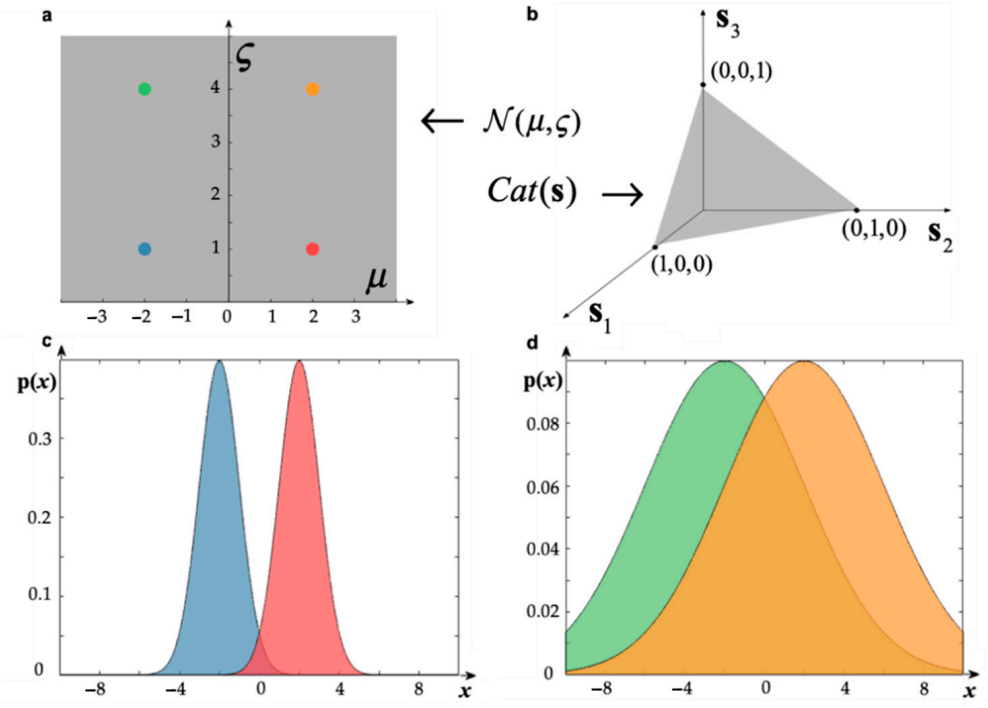

Figure 2. Statistical manifolds and information length. Panels (**a**,**b**) illustrate the statistical manifolds associated with two well-known probability distributions: the normal distribution and the categorical distribution, respectively. The statistical manifold associated with a probability distribution is the set of all possible parameters that it can take. For the univariate normal distribution, parameterised with mean μ and positive standard deviation ς, the associated statistical manifold is the upper half plane (panel **a**). For the categorical distribution, in the case of three possible states, the statistical manifold is the 2-dimensional simplex (panel **b**). More generally, in the case of n possible states, the statistical manifold of the categorical distribution is the set of all vectors with positive components that sum to one, i.e., the $(n-1)$-dimensional simplex. This is a higher-dimensional version of the triangle or the tetrahedron. In panels (**c**,**d**) we illustrate why the Euclidean distance is ill-suited to measure the information distance between probability distributions. To show this, we selected four distributions that correspond to points on the statistical manifold of the normal distribution. One can see that the Euclidean distance between the modes of the red and the blue distributions is the same as that from the orange and the green; however, the difference in information of each respective pair is quite different. In panel (**c**), the two distributions correspond to two drastically different beliefs, since there is such little overlap; on the contrary, the beliefs in panel (**d**) are much more similar. This calls for a different notion of distance that measures the difference in (Fisher) information between distributions, namely, the information length.

Smooth statistical manifolds are naturally equipped with a different notion of distance, even though they may be subsets of Euclidean space. This is because the Euclidean distance measures the physical distance between points, while the information length measures distance in terms of the (accumulated) change in (Fisher) information (see Figure 2) along a path. The canonical choice of information length on a statistical manifold is associated with the Fisher information metric tensor $\mathbf{g}$ [113–115]. Technically, a metric tensor is a smoothly varying choice of symmetric, positive definite matrices at each point of the manifold. This enables computation of the length of paths as well as the distance between points, by measuring the length of the shortest path (see Appendix B). Mathematically, the Fisher information metric can be defined the Hessian of the KL divergence between two infinitesimally close distributions (see Appendix B). This means that the information length of a trajectory on a statistical manifold is given by accruing infinitesimally small changes in the KL divergence along it.

Amari's natural gradient descent [85,116] is a well-known optimisation algorithm for finding the minimum of functions defined on statistical manifolds such as the variational free energy. It consists of preconditioning the standard gradient descent update rule with the inverse of the Fisher information metric tensor:

$$\dot{\mathbf{s}} = -\nabla_{\mathbf{s}} \mathrm{F} \;\rightarrow\; \dot{\mathbf{s}} = -\mathbf{g}^{-1}(\mathbf{s})\nabla_{\mathbf{s}} \mathrm{F}$$

Preconditioning by the inverse of $\mathbf{g}$ means that the natural gradient privileges directions of low information length. One can see this, since the directions of greatest (resp. smallest) information length are the eigenvectors of the highest (resp. lowest) eigenvalues of $\mathbf{g}$.

In fact, Amari proved that the natural gradient follows the direction of steepest descent of the objective function in information space [85]. Technically, the natural gradient generalises gradient descent to functions defined on statistical manifolds. As the free energy for discrete state-estimation is convex (see Appendix A), this means that natural gradient descent will always converge to the free energy minimum via a short path.

To summarise, agents' beliefs naturally evolve on a statistical manifold towards the point of optimal inference. These manifolds are equipped with a different notion of distance, namely, the information length. In the case of discrete state-estimation, beliefs evolve on the simplex towards the free energy minimum. Reaching the minimum with a short path translates into higher computational and metabolic efficiency. One scheme that achieves short paths for finding the minimum of the free energy is the natural gradient. In the next section, we will show that the neuronal dynamics entailed by active inference approximate natural gradient descent.

5. Active Inference Approximates Natural Gradient Descent

Discretising the (neuronal) dynamics prescribed by active inference and natural gradient descent give us the following state-estimation belief updates, respectively:

$$\mathbf{s}^{(t+1)} \leftarrow \sigma\Big(\ln \mathbf{s}^{(t)} - \epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\Big) \quad \mathbf{s}^{(t+1)} \leftarrow \frac{\mathbf{s}^{(t)} - \epsilon \, \mathbf{g}^{-1}(\mathbf{s}^{(t)})\nabla_{\mathbf{s}^{(t)}} \mathrm{F}}{\sim}$$

In these equations, the logarithm is taken component-wise; ε is the step-size used in the discretisation and $\sim$ denotes normalisation by the sum of the components to ensure that $\mathbf{s}^{(t+1)}$ lies on the simplex. (Natural gradient descent dynamics do not necessarily remain on the statistical manifold. This problem has been the object of numerous works that supplemented the natural gradient update with a projection step [117–121]. Here, we choose the simplest projection step to ensure that the result remains on the simplex: normalising with the sum of the components.)

These dynamics are approximately the same. We can equate them under a first order Taylor approximation of the exponential inside the softmax function:

$$\begin{aligned}
\mathbf{s}^{(t+1)} \quad &\leftarrow \sigma\Big(\ln \mathbf{s}^{(t)} - \epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\Big) \\
&= \frac{\exp\Big[\ln \mathbf{s}^{(t)} - \epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\Big]}{\sim} \\
&= \frac{\mathbf{s}^{(t)} \odot \exp\Big[-\epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\Big]}{\sim} \\
&\simeq \frac{\mathbf{s}^{(t)} \odot \Big[\vec{1} - \epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\Big]}{\sim} \\
&= \frac{\mathbf{s}^{(t)} - \epsilon \, \mathbf{s}^{(t)} \odot \nabla_{\mathbf{s}^{(t)}} \mathrm{F}}{\sim} \\
&= \frac{\mathbf{s}^{(t)} - \epsilon \, \mathbf{g}^{-1}(\mathbf{s}^{(t)})\nabla_{\mathbf{s}^{(t)}} \mathrm{F}}{\sim}
\end{aligned}$$

The symbol $\odot$ denotes the Hadamard product (elementwise multiplication). The last line follows since, on the simplex, the inverse of the Fisher information metric tensor is simply a diagonal matrix whose diagonal is $\mathbf{s}$ (see Appendix C).

Although these dynamics are approximately the same, this does not guarantee that the paths taken in the limit of infinitesimally small time steps (which correspond to continuous-time dynamics in physical and biological systems) will be the same. One can see this algebraically, since the number of time steps needed to reach the free energy minimum increases as the step-size decreases; thus, the difference between paths, which equals the sum of the differences at each timestep, is not guaranteed to converge to zero. Hence, it is necessary to verify that this approximation holds well in practice by analysing the discrepancy between paths using numerical simulations.

6. Numerical Simulations

In this section, we use numerical simulations of decision-making tasks under uncertainty to assess to what extent the inferential dynamics of active inference match the natural gradient descent on free energy. Our simulations compared the information length of the belief trajectories taken by both schemes, which reflects their computational and metabolic efficiency.

6.1. Methodology

We simulate two sequential decision-making tasks under uncertainty—a simple two-step maze task and a more complex abstract rule learning task. These tasks involve an agent equipped with a generative model (in our case, a partially observed Markov decision process [98]) and a repeated sequence of the following steps [7]:

1. **Observation**: The agent receives an observation;
2. **Perceptual inference:** The agent infers hidden states from observations by minimizing free energy;
3. **Planning as inference:** The agent evaluates courses of action by evaluating their expected free energy, an objective which favours exploitative (goal-seeking) and explorative (ambiguity resolving) behaviour [7,33];
4. **Decision-making:** The agent executes the action with the lowest expected free energy.
5. **Learning:** The agent learns the generative model by accumulating data on hidden state transitions and on the likelihood mapping between states and outcomes.

6.1.1. Overview of Tasks

The first paradigm simulates a rat in a T-Maze (see Figure 3). The T-Maze has a reward which is placed either in the right or left upper arm (in red). The bottom arm contains a cue that specifies the location of the reward. The rat's initial location is the middle of the T-Maze. The initial conditions are specified such that the rat believes that it will claim the reward and avoid the unbaited upper arm. In the first three trials, the reward's location is random. The optimal strategy then consists of collecting the cue (exploration) and using this information to collect the reward (exploitation). To achieve this, the rat must infer its location and the configuration of the maze, in addition to the route it will take. In addition, the agent has to learn that the cue provides reliable information that enables it to infer the reward's location. After the first three trials, the reward's location remains constant across trials. As there is no residual uncertainty to resolve by checking the cue, the optimal strategy then consists of directly collecting the reward. To do this, the agent has to learn the reward's location.

The second paradigm is a simulation of abstract rule learning. In each trial, the agent is presented with a sequence of stimuli and has to guess the next in the sequence, upon which it receives positive or negative feedback. The stimuli are generated according to a rule unknown to the agent, so that there is always one unique correct answer. The goal of the agent is to provide correct answers while avoiding incorrect ones. To achieve its goal, the agent has to learn the rule generating the stimuli. In the first trials—and

in the absence of knowledge—the agent guesses the missing stimulus. After each trial, the agent updates its prior beliefs based on past observations. After several trials, the agent has learned the contingencies of the task, whence it is able to predict the missing stimulus correctly and reliably. A complete specification of the generative model for this rule learning task may be found in [39]. The important consideration for our purposes is that this simulation is a more complex version of the T-Maze task, in the sense that it uses the same belief updating scheme and generative model. The key difference is that the generative model contains many more states: sixteen control states, 144 hidden states and 48 possible observations [39], Figure 3. This means that the simplex on which the free energy is defined has 143 dimensions.

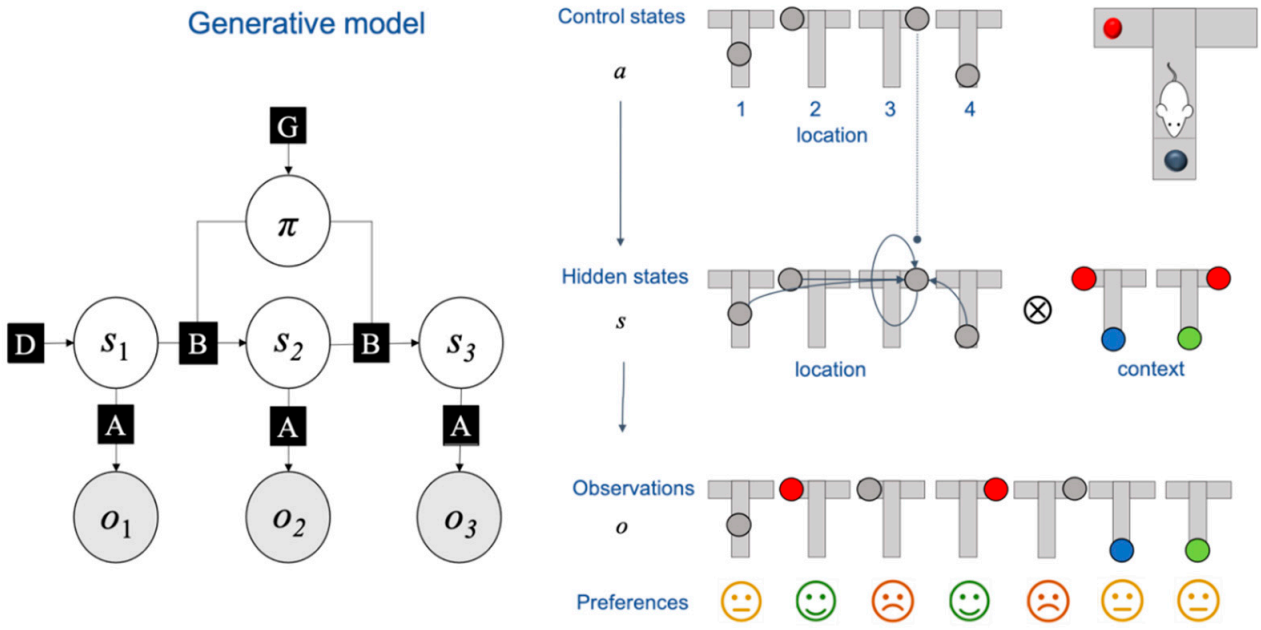

Figure 3. Generative model and T-Maze task. This figure describes the generative model used in the simulations and the specificities of the T-Maze task. The generative model is a partially observed Markov decision process (POMDP), a canonical model for sequential decision-making tasks under uncertainty. On the left-hand-side, the generative model is expressed as a factor graph [122]: the circles represent random variables while the squares represent factors. The shaded variables are observations, which the agent samples in discrete time. The unshaded variables are random variables that are inferred by the agent. These are either hidden states (s) or policies (π, i.e., action sequences), which need to be inferred from observations. The factors encode the relationships between the variables that they link. For example, D is a vector of non-negative components that sum to one that expresses the prior belief of being in this or another hidden state; B is a stochastic matrix that encodes the transition probabilities between hidden states given actions; similarly, A is a matrix that gives the probability of observations given states. See [7] for more details. Importantly, the factors can themselves be learned over time. On the right-hand side, we further unpack the generative model of the T-Maze task. The task is as pictured in the upper-right corner. The rat's initial location is the centre of the T-Maze. At each time-step, the rat can choose between four control states (i.e., actions) that lead the rat to move to one of the four maze locations (middle, top-left, top-right or bottom). There are four hidden states that correspond to the location of the rat in the T-Maze and another two that correspond to the location of the reward (right arm or left arm), which leads to eight hidden states in total, meaning that the associated simplex on which the free energy is defined is seven-dimensional. There are seven possible observations that correspond to being in the middle of the T-Maze, in each one of the upper arms with (in red) or without the reward (in grey), and to collecting the cue, which reveals the location of the reward. The agent has a preference for each of these observations: the agent wants to claim the reward and avoid the unbaited arm. The agent is neutral w.r.t. all other observations. See [57] for more details of this paradigm.

6.1.2. Two Schemes

We simulate each task with two different schemes: a standard active inference scheme, and a modified active inference scheme that differs only in that it performs perceptual inference using natural gradient descent. More explicitly, these agents, respectively, infer hidden states from observations in peristimulus time using the learning rule

$$\mathbf{s}^{(t+1)} \leftarrow \sigma\left(\ln \mathbf{s}^{(t)} - \epsilon \, \nabla_{\mathbf{s}^{(t)}} \mathrm{F}\right) \quad \mathbf{s}^{(t+1)} \leftarrow \frac{\mathbf{s}^{(t)} - \epsilon \, \mathbf{g}^{-1}(\mathbf{s}^{(t)}) \nabla_{\mathbf{s}^{(t)}} \mathrm{F}}{\sim},$$

with a common step size of $\varepsilon = 0.25$. Note that this optimization procedure is repeated at each step of the decision-making task, as a new observation updates the free energy landscape. Note also that the free energy changes in between trials due to the fact that the generative model is learned over time. The explicit form of the free energy for a partially observed Markov decision process is given in Appendix A.

All other aspects of belief updating (planning as inference, decision-making and learning) were conserved across schemes, following [7,39,57] and their publicly available software implementation (see Software note).

6.1.3. Measuring Information Length

We simulate both tasks with 128 agents of each scheme and for 24 trials. For each agent, we measure the information length of perceptual inference by accumulating the information distance (Appendix E) between consecutive updates of the learning rule $\mathbf{s}^{(t+1)}, \mathbf{s}^{(t)}$.

$$d_{\mathbf{g}}(\mathbf{s}^{(t+1)}, \mathbf{s}^{(t)}) = 2\| \sqrt{\mathbf{s}^{(t+1)}} - \sqrt{\mathbf{s}^{(t)}} \|$$

6.2. Results

We plot the results of our numerical experiments in Figure 4.

We observe that, in the T-Maze task, standard active inference performs better than the active inference scheme modified with natural gradient descent. However, the opposite is true in the abstract rule learning task. Overall, the differences in information length between the two schemes are small compared to the overall information length of belief trajectories (less than 4% deviation on average in both tasks). This suggests that the difference between the two schemes is small.

The paradigm-specific difference in performance is due to differences in the generative model and the corresponding free energy landscapes. This means that it should be possible to find other paradigms where either active inference or the natural gradient performs better. Yet, any discrete state-estimation task will exhibit a quantitatively similar, convex-free energy landscape (Appendix A), which implies that the inference problem will be mathematically analogous. Therefore, we expect small performance differences between the two schemes for related sequential decision-making tasks (with an associated partially observed Markov decision process).

To summarise, our results suggest that state-estimation in active inference is a good approximation to natural gradient descent on free energy. Natural gradient descent follows the steepest descent (as defined in information space) on a convex free energy landscape (see Appendix A). This means that neural dynamics under active inference take short paths to the free energy minimum, which are computationally and metabolically efficient.

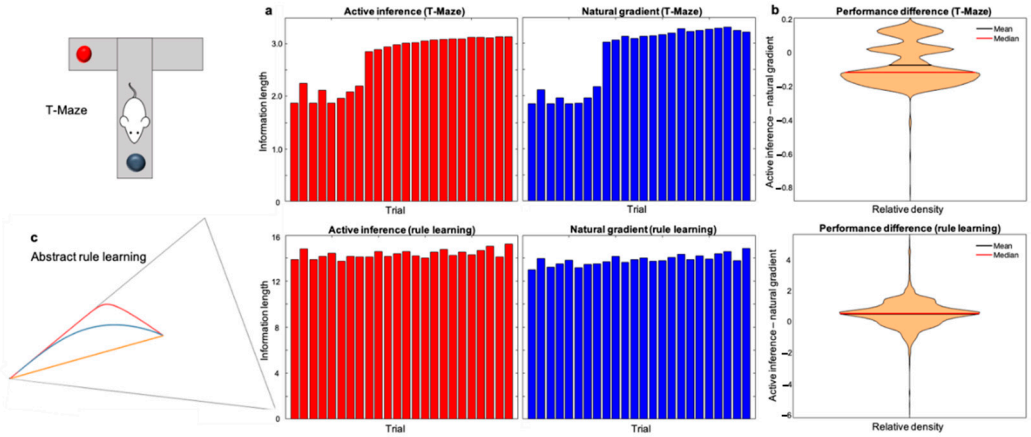

Figure 4. Information length of perceptual inference in active inference and natural gradient. The upper-half of the figure plots the results of the numerical experiments with T-Maze task, while the lower half the abstract rule learning task. In Panel (**a**), the histograms show the information length of perceptual inference in active inference (in red) and natural gradient (in blue). Specifically, this is (mean = 7075, median = 2.9776) for active inference, and (mean = 2.7793, median = 3.1163) for the natural gradient; in the rule learning task, the information lengths are (mean = 14.3573, median = 16.2099) for active inference and (mean = 13.8588, median = 16.1058) for the natural gradient. Information trajectories are longer in the abstract rule learning task, due to its increased complexity. In other words, the greater number of hidden states—in the generative model—mandate more belief updating. This means the model used for this high-dimensional task entails greater complexity in the technical sense: as belief updating moves posterior beliefs further from their priors. (Parenthetically, this means that the agents took much longer to update their beliefs). The reasons for systematic variation in information length across trials within tasks is that (1) the task configuration varied from trial to trial and (2) the generative models were themselves learned over trials. The sudden increase in information length in the T-Maze task at the ninth trial occurs when agents learn that the reward's location is constant. As the reward's location is unambiguous, belief trajectories are longer as beliefs about context move from the middle of the simplex to the vertex representing the reward's location at *each* step of the paradigm. This means that we can improve the biological plausibility of the implementation by starting each belief update at the previous free energy minimum instead of an agnostic location in the middle of the simplex. This would avoid beliefs travelling to the same vertex multiple times in the same trial and hence would make belief trajectories *shorter* once the reward location is learned. This consideration does not affect our characterisation, which simply compares the information length given consistent starting and end points. As the context could be deduced only at the last step of the task, a similar sudden change in information length did not occur in the abstract rule learning task. However, the slight increase in information length across trials occurs for the same reason: learning the generative model made inferences increasingly precise. In Panel (**b**), the violin plots illustrate the difference in information length by subtracting the information length of the natural gradient from the information length of active inference. In the T-Maze paradigm, active inference mostly takes shorter paths, which is why the violin plot's values are mostly negative (mean= -0.0718, median= -0.1146). However, we obtain the opposite pattern in the abstract rule learning task (mean = 0.4985, median= 0.5401). Panel (**c**) shows an example of the belief trajectory taken during state estimation in the abstract rule learning task. This was possible to include as the generative model includes a hidden state dimension with three possible alternatives [39], enabling a two-dimensional representation of the associated simplex. The red trajectory corresponds to active inference, the blue is natural gradient descent, and the orange is the shortest path to the free energy minimum (i.e., the geodesic, see Appendix D). This example is not representative of the average and was chosen for purely illustrative purposes as the trajectories are very distinct, lengthy and do not coincide with the geodesic. The fact that both schemes were significantly suboptimal in this example was unsurprising as beliefs evolve to the free energy minimum using only local information about the free energy landscape. Note that this suboptimality occurred only in a small minority of the trials considered here.

7. Discussion

In the first part of this paper, we showed that neural dynamics for discrete state-estimation under active inference are consistent with mean-field models' neural population dynamics. This construct validity is further supported by the wide range of plausible electrophysiological responses that can be synthesised with active inference [57]. Yet, to fully endorse this view, the electrophysiological responses simulated during state-

estimation need empirical validation. To do this, one would have to specify the generative model that a biological agent employs to represent a particular environment. This may be identified by comparing alternative hypothetical generative models with empirical choice behaviour and computing the relative evidence for each model (e.g., [123]). Once the appropriate generative model is found, one would need to compare the evidence for a few possible practical implementations of active inference, which come from various possible approximations to the free energy [27,124,125], each of which yields different belief updates and simulated electrophysiological responses. Note that, of possible approximations of the free energy, the marginal approximation which was used in our simulations currently stands as the most biologically plausible [124]. Finally, one would be able to assess the explanatory power of active inference in relation to empirical measurements and compare it with other process theories.

In the second part of this paper, we showed that the neuronal process theory associated with active inference approximates natural gradient descent for discrete state-estimation. (Note that this work does not treat continuous state-estimation, e.g., inferring the temperature in a room, which needs to be investigated separately). Given that the natural gradient follows the direction of steepest descent of the free energy in information space [85] and that the free energy landscape at hand is convex (see Appendix A), this ensures that agent's beliefs reach the point of optimal inference via short trajectories in information space. This means that active inference entails neuronal dynamics that are both computationally and energetically efficient, an important feature of any reasonable process theory of the brain.

In the case of simulated (i.e., discretised) belief dynamics, active inference and natural gradient perform similarly on average. Performance is scored by the information length accrued during belief updating, which measures efficiency. In some cases, however, the belief trajectories taken by both schemes were significantly longer than the shortest path to the point of optimal inference. This is unsurprising since agents' beliefs move myopically to the free energy minimum. In short, our analysis suggests that biological agents can perform natural gradient descent in a biologically plausible manner. From an engineering perspective, this means that we can relate variational message passing and belief propagation, two inferential algorithms based on free energy minimisation [124,126,127], with the natural gradient.

A general point is that the tools furnished by information geometry are ideally suited to characterise and visualise inference in biological organisms as well as scoring its efficiency. This paves the way for further applications of information geometry to analyse information processing in biological systems.

8. Conclusions

In the first part, we showed how a generic account of brain function provided by active inference is consistent with the more biophysically detailed and empirically driven mean-field models of neural population dynamics.

We then demonstrated that state estimation under active inference approximates natural gradient descent on free energy. This suggests that the beliefs about states of the world of biological agents evolve approximately according to a steepest descent on free energy. Since the free energy landscape for discrete state-estimation is convex, the trajectories taken to the point of optimal inference are short, which incurs minimal computational and metabolic cost. This demonstrates the efficiency of neural dynamics under active inference, an important feature of the brain as we know it. Further testing of active inference as a process theory of brain function should focus on extending the empirical evaluation of simulated neurophysiological responses.

Author Contributions: Conceptualization, L.D.C., T.P., B.S. and K.F.; Formal analysis, L.D.C.; Software, L.D.C. and T.P.; Writing—original draft, L.D.C.; Writing—review and editing, T.P., B.S. and K.F. All authors have read and agreed to the published version of the manuscript.

Funding: L.D.C. is supported by the Fonds National de la Recherche, Luxembourg (Project code: 13568875). T.P. is supported by the Rosetrees Trust (Award number: 173346). K.F. is funded by a Wellcome Trust Principal Research Fellowship (Ref: 088130/Z/09/Z).

Data Availability Statement: This article has no additional data.

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

Software Note: The belief updating process described in this article can be implemented using the routine spm_MDP_VB_X.m available as Matlab code in the SPM academic software http://www.fil.ion.ucl.ac.uk/spm/ (accessed on 9 April 2021). The active inference simulations described in Figures 1 and 3 can be reproduced via a graphical user interface by typing DEM (DEM_demo_MDP_X.m for the T-Maze task, rule_learning.m for the abstract rule learning task). The natural gradient and information length simulations can be reproduced with code available at https://github.com/lancelotdacosta/Neural-Dynamics-AI (accessed on 9 April 2021).

Appendix A. Convexity of the Free Energy

This Appendix shows the convexity of the free energy employed in discrete state-estimation. On discrete state-space generative models (e.g., partially observable Markov decision processes), the free energy optimised during state-estimation can be expressed as [7]:

$$\mathrm{F}(\mathbf{s}_{\pi 1}, \ldots, \mathbf{s}_{\pi T}) = \sum_{\tau=1}^{T} \mathbf{s}_{\pi\tau} \cdot \ln \mathbf{s}_{\pi\tau} - \sum_{\tau=1}^{t} o_\tau \cdot \ln(A)\, \mathbf{s}_{\pi\tau} - \mathbf{s}_{\pi 1} \cdot \ln D - \sum_{\tau=2}^{T} \mathbf{s}_{\pi\tau} \cdot \ln(B_{\pi_{\tau-1}})\, \mathbf{s}_{\pi\tau-1}$$

In this context, the neuronal dynamics described in the paper are:

$$\begin{aligned} \dot{v}\,(\mathbf{s}_{\pi 1}, \ldots, \mathbf{s}_{\pi T}) &= -\nabla_{\mathbf{s}_{\pi\tau}} \mathrm{F}(\mathbf{s}_{\pi 1}, \ldots, \mathbf{s}_{\pi T}) \\ \mathbf{s}_{\pi\tau} &= \sigma(v) \end{aligned}$$

Here, τ corresponds to time (which is discretised), $\mathbf{s}_{\pi\tau}$ corresponds to the beliefs about states at timepoint τ, conditioned upon the fact that the agent is pursuing a certain sequence of actions π. The particular meaning of the other variables is not important for our purposes; the only important consideration is that $B_{\pi_{\tau-1}}$ are matrices, whose components are strictly contained between zero and one, and logarithms are taken component-wise.

Recall that a sum of convex functions is convex. Furthermore,

- $x \mapsto x \ln x$ is convex in the interval $[0, 1]$, which implies that $\sum_{\tau=1}^{T} \mathbf{s}_{\pi\tau} \cdot \ln \mathbf{s}_{\pi\tau}$ is convex.
- $-\sum_{\tau=1}^{t} o_\tau \cdot \ln(A)\, \mathbf{s}_{\pi\tau} - \mathbf{s}_{\pi 1} \cdot \ln D$ is a linear function, hence it is convex.
- $-\ln(B_{\pi_{\tau-1}})$ only has positive components, hence $-\sum_{\tau=2}^{T} \mathbf{s}_{\pi\tau} \cdot \ln(B_{\pi_{\tau-1}})\, \mathbf{s}_{\pi\tau-1}$ is a positive linear combination of polynomials of degree two, which is convex.

This implies that the free energy is convex.

Appendix B. Fisher Information Metric Tensor, Information Length and Information Distance

The Fisher information metric tensor is the canonical mathematical object that the enables computation of (a certain kind of) information distance on a statistical manifold. Technically, a metric tensor is a choice of symmetric, positive definite matrix at each point, that varies smoothly on the statistical manifold. This is equivalent to specifying an inner product at each point of the manifold and doing so smoothly.

Let $p(x|\mathbf{s})$ be a probability distribution parameterised by $\mathbf{s}$. The set of all possible choices of $\mathbf{s}$ is the statistical manifold associated with p, which we will denote by M. This is (in the case of classical probability distributions, which includes the scope of this paper) a smooth manifold, where each point corresponds to a certain parameterisation of the

probability distribution, i.e., a (smooth) statistical manifold. We can then define the Fisher information metric tensor as

$$\mathbf{g}(\mathbf{s}) = \nabla^2_\theta D_{KL}[p(x|\mathbf{s})\|p(x|\theta)]\Big|_{\theta=\mathbf{s}}.$$

This is an n-by-n matrix where n is the dimensionality of $\mathbf{s}$ and θ. There exist other equivalent definitions [107,108].

This is beneficial, because a choice of an inner product at each point on the manifold makes it possible to compute the length of tangent vectors. Let v be such a tangent vector at a point $\mathbf{s}$, then its norm is given by

$$\|v\|_\mathbf{g} := \sqrt{v^T \mathbf{g}(\mathbf{s}) v}$$

This means that we can also compute the length of smooth curves. Let $\gamma : [0,1] \subset \mathbb{R} \to M$ be such a curve. Its information length is given by

$$\ell_\mathbf{g}(\gamma) := \int_0^1 \sqrt{\dot\gamma(t)^T \mathbf{g}(\gamma(t))\dot\gamma(t)}\, \mathrm{d}t,$$

where $\dot\gamma := \frac{d\gamma}{dt}$.

We can trivially extend this definition to compute the information distance between points, say $\mathbf{s}$ and $\mathbf{s'}$. This is simply the information length of the shortest curve connecting the two points

$$d_\mathbf{g}(\mathbf{s}, \mathbf{s}') = \inf_{\substack{\gamma:\,[0,1]\to M \\ \gamma(0)=\mathbf{s},\,\gamma(1)=\mathbf{s}'}} \ell(\gamma).$$

where, inf denotes the infimum of the quantity subject to the constraints in the subscript. Let us take a step back to see why these definitions are sensible.

Statistical manifolds are generally curved; therefore, it is only possible to compute distances locally, by deforming the small region of consideration into a portion of Euclidean space. This is impractical and does not solve the problem of computing distances over larger scales. Even if one did so, one would recover a deformed version of the Euclidean distance, which would, generally speaking, not measure distance in terms of information. The raison d'être of the metric tensor is to allow the computation of distances on the manifold in a consistent way, and, in our case, consistently with the difference in Shannon information.

If one replaced $\mathbf{g}$ in the definitions above by the identity matrix (i.e., the metric tensor that is used implicitly in Euclidean space), one recovers the classical notion of length of a vector (i.e., the square root of the inner product), the classical notion of the length of a curve, namely

$$\ell(\gamma) := \int_0^1 \|\dot\gamma(t)\|\, \mathrm{d}t.$$

The distance between two points is a little trickier as it involves proving that the shortest path between two points is the straight line when the metric tensor is the identity. This involves solving the geodesic equation (see Appendix D) for this metric tensor. Once this is done, inserting a straight line in the above equation returns the usual Euclidean distance.

Appendix C. Fisher Information Metric Tensor on the Simplex

Suppose there are $n+1$ states $S = \{s_0, \ldots, s_n\}$. Then, a categorical distribution $p(x|\mathbf{s})$ over those states is defined as $p(s_i|\mathbf{s}) := \mathbf{s}_i$. The statistical manifold of all possible parameters is the interior of the n-dimensional simplex, which is defined as

$$\Delta^n := \{\mathbf{s} = (\mathbf{s}_0, \ldots, \mathbf{s}_n) \in \mathbb{R}^{n+1} \mid \mathbf{s}_i > 0, \sum_i \mathbf{s}_i = 1\}$$

The Fisher information metric tensor can be defined as

$$\mathbf{g}(\mathbf{s}) = \nabla^2_\theta D_{KL}[p(x|\mathbf{s})\|p(x|\theta)]\Big|_{\theta=\mathbf{s}}.$$

The KL-divergence between two categorical distributions is given by

$$\begin{aligned} D_{KL}[p(x|\mathbf{s})\|p(x|\theta)] &= \sum_{x\in S} p(x|\mathbf{s}) \log \tfrac{p(x|\mathbf{s})}{p(x|\theta)} \\ &= \sum_{i=0}^{n} p(s_i|\mathbf{s}) \log \tfrac{p(s_i|\mathbf{s})}{p(s_i|\theta)} \\ &= \sum_{i=0}^{n} \mathbf{s}_i \log \tfrac{\mathbf{s}_i}{\theta_i} \end{aligned}$$

We can take second derivatives

$$\frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_k}\left(\sum_{i=0}^{n} \mathbf{s}_i \log\frac{\mathbf{s}_i}{\theta_i}\right) = \delta_{jk}\frac{\mathbf{s}_k}{\theta_k^2}.$$

where δ_{jk} is the Kronecker delta. Finally,

$$\mathbf{g}(\mathbf{s}) = \begin{bmatrix} \mathbf{s}_0^{-1} & 0 & \cdots & 0 \\ 0 & \mathbf{s}_1^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{s}_n^{-1} \end{bmatrix}.$$

Technical remark: since the statistical manifold of interest is n-dimensional, it is best to view this metric tensor as being defined on an $n+1$ dimensional neighbourhood of the simplex, e.g., the positive orthant of $\mathbb{R}^{n+1}$.

Appendix D. Geodesics on the Simplex

The aim of this section is to find the expression of the shortest path (in information length) between two points on the simplex.

As shown in Appendix C, the metric tensor is given by

$$\mathbf{g}(\mathbf{s}) = \begin{bmatrix} \mathbf{s}_0^{-1} & 0 & \cdots & 0 \\ 0 & \mathbf{s}_1^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{s}_n^{-1} \end{bmatrix}.$$

Let $\mathbf{s}^{(0)}, \mathbf{s}^{(1)}$ be two points on the simplex. From standard differential geometry, the shortest path γ between two points satisfies the geodesic equation:

$$\ddot{\gamma}_k + \sum_{i,j=0}^{n} \Gamma^k_{ij}(\gamma)\,\dot{\gamma}_i\dot{\gamma}_j \equiv 0.$$

where Γ^k_{ij} are the Christoffel symbols of the Levi-Civita connection. These are real valued functions defined with respect to the metric:

$$\Gamma^k_{ij} := \frac{1}{2}\sum_{r=0}^{n} \mathbf{g}^{kr}\Big(\partial_j\mathbf{g}_{ri} + \partial_i\mathbf{g}_{rj} - \partial_r\mathbf{g}_{ij}\Big).$$

In this expression, $\mathbf{g}^{kr}$ is the (k,r) entry of the inverse metric tensor $\mathbf{g}^{-1}$ and ∂_j is a shorthand for $\frac{\partial}{\partial \mathbf{s}_j}$. In our case the only non-zero Christoffel symbols are given by

$$\Gamma^i_{ii}(\mathbf{s}) = -\frac{1}{2\mathbf{s}_i}.$$

This means that each component of the geodesic must satisfy the equation

$$2\gamma_i\ddot{\gamma}_i - \dot{\gamma}_i{}^2 \equiv 0.$$

By inspection, one can see that the differential equation admits a polynomial solution of degree two. Solving with the boundary conditions $\gamma(0) = \mathbf{s}^{(0)}, \gamma(1) = \mathbf{s}^{(1)}$ and discarding those solutions that leave the positive orthant of $\mathbb{R}^{n+1}$ (c.f., last remark Appendix C) yields the expression of the geodesic:

$$\gamma(t) = \left((1-t)\sqrt{\mathbf{s}^{(0)}} + t\sqrt{\mathbf{s}^{(1)}}\right)^2.$$

Appendix E. Information Distance on the Simplex

The distance between two points on a statistical manifold is given by the information length of the shortest path (i.e., the geodesic) between the two. Given two points $\mathbf{s}^{(0)}, \mathbf{s}^{(1)}$ on the simplex, we have seen in Appendix D that the geodesic between these points is

$$\gamma(t) = \left((1-t)\sqrt{\mathbf{s}^{(0)}} + t\sqrt{\mathbf{s}^{(1)}}\right)^2$$

Furthermore, from Appendix B, we have seen that the information distance between two points is the information length of the geodesic between them

$$d_{\mathbf{g}}(\mathbf{s}^{(0)}, \mathbf{s}^{(1)}) = \ell_{\mathbf{g}}(\gamma) = \int_0^1 \sqrt{\dot{\gamma}(t)^T \mathbf{g}(\gamma(t))\dot{\gamma}(t)}\, \mathrm{d}t$$

Lastly, from Appendix C, the Fisher information metric tensor on the simplex is

$$\mathbf{g}(\mathbf{s}) = \begin{pmatrix} \mathbf{s}_0^{-1} & 0 & \cdots & 0 \\ 0 & \mathbf{s}_1^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{s}_n^{-1} \end{pmatrix}$$

Therefore, expanding the expression inside the information distance

$$\dot{\gamma}(t)^T \mathbf{g}(\gamma(t))\dot{\gamma}(t) = \sum_{i=0}^{n} \frac{\dot{\gamma}_i(t)^2}{\gamma_i(t)}$$

it is possible to show that $\frac{\dot{\gamma}_i(t)^2}{\gamma_i(t)}$ is constant for each i. One can do this by taking the derivative with respect to t and noting that the result vanishes. This means that one can remove the integral and find a concise expression for the information distance:

$$\begin{aligned} d_{\mathbf{g}}(\mathbf{s}^{(0)}, \mathbf{s}^{(1)}) &= \sqrt{\dot{\gamma}(0)^T \mathbf{g}(\gamma(0))\dot{\gamma}(0)} \\ &= \sqrt{\sum_{i=0}^{n} \frac{\dot{\gamma}_i(0)^2}{\gamma_i(0)}} \\ &= \sqrt{4\sum_{i=0}^{n} \left(\sqrt{\mathbf{s}_i^{(1)}} - \sqrt{\mathbf{s}_i^{(0)}}\right)^2} \\ &= 2\|\sqrt{\mathbf{s}^{(1)}} - \sqrt{\mathbf{s}^{(0)}}\| \end{aligned}$$

This expression is compelling, since it relates the information distance on the simplex to the Euclidean distance on the n-dimensional sphere.

References

1. Bogacz, R. A tutorial on the free-energy framework for modelling perception and learning. *J. Math. Psychol.* **2017**, *76*, 198–211. [CrossRef] [PubMed]
2. Friston, K. The free-energy principle: A rough guide to the brain? *Trends Cogn. Sci.* **2009**, *13*, 293–301. [CrossRef] [PubMed]
3. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *J. Physiol.* **2006**, *100*, 70–87. [CrossRef] [PubMed]
4. Friston, K.J. Life as we know it. *J. R. Soc. Interface* **2013**, *10*, 20130475. [CrossRef] [PubMed]
5. Friston, K. A Free Energy Principle for a Particular Physics. arXiv:190610184 [q-bio]. 2019. Available online: http://arxiv.org/abs/1906.10184 (accessed on 29 February 2020).
6. Parr, T.; Da Costa, L.; Friston, K. Markov blankets, information geometry and stochastic thermodynamics. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2020**, *378*, 20190159. [CrossRef]
7. Da Costa, L.; Parr, T.; Sajid, N.; Veselic, S.; Neacsu, V.; Friston, K. Active inference on discrete state-spaces: A synthesis. *J. Math. Psychol.* **2020**, *99*, 102447. [CrossRef]
8. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soc. B Biol. Sci.* **2005**, *360*, 815–836. [CrossRef] [PubMed]
9. Beal, M.J. Variational Algorithms for Approximate Bayesian Inference. Ph.D. Thesis, University of London, London, UK, May 2003; p. 281.
10. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An Introduction to Variational Methods for Graphical Models. In *Learning in Graphical Models*; Jordan, M.I., Ed.; Springer: Dordrecht, The Netherlands, 1998; pp. 105–161. [CrossRef]
11. Wainwright, M.J.; Jordan, M.I. Graphical Models, Exponential Families, and Variational Inference. *FNT Mach. Learn.* **2007**, *1*, 1–305. [CrossRef]
12. Buckley, C.L.; Kim, C.S.; McGregor, S.; Seth, A.K. The free energy principle for action and perception: A mathematical review. *J. Math. Psychol.* **2017**, *81*, 55–79. [CrossRef]
13. Friston, K. The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef]
14. Colombo, M.; Wright, C. First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese* **2018**. [CrossRef]
15. Kappen, H.J.; Gómez, V.; Opper, M. Optimal control as a graphical model inference problem. *Mach. Learn.* **2012**, *87*, 159–182. [CrossRef]
16. Da Costa, L.; Sajid, N.; Parr, T.; Friston, K.; Smith, R. The Relationship between Dynamic Programming and Active Inference: The Discrete, Finite-Horizon Case. arXiv:200908111 [cs, math, q-bio]. 2020. Available online: http://arxiv.org/abs/2009.08111 (accessed on 31 January 2021).
17. Millidge, B.; Tschantz, A.; Seth, A.K.; Buckley, C.L. On the Relationship between Active Inference and Control as Inference. arXiv:200612964 [cs, stat]. 2020. Available online: http://arxiv.org/abs/2006.12964 (accessed on 28 June 2020).
18. Watson, J.; Imohiosen, A.; Peters, J. Active Inference or Control as Inference? A Unifying View. arXiv:201000262 [cs, stat]. 2020. Available online: http://arxiv.org/abs/2010.00262 (accessed on 27 January 2021).
19. Aitchison, L.; Lengyel, M. With or without you: Predictive coding and Bayesian inference in the brain. *Curr. Opin. Neurobiol.* **2017**, *46*, 219–227. [CrossRef] [PubMed]
20. Knill, D.C.; Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **2004**, *27*, 712–719. [CrossRef]
21. Lake, B.M.; Ullman, T.D.; Tenenbaum, J.B.; Gershman, S.J. Building Machines That Learn and Think Like People. arXiv:160400289 [cs, stat]. 2016. Available online: http://arxiv.org/abs/1604.00289 (accessed on 11 August 2019).
22. Rao, R.P.N.; 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. [CrossRef]
23. 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. [CrossRef] [PubMed]
24. Friston, K.; Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 1211–1221. [CrossRef]
25. Kaplan, R.; Friston, K.J. Planning and navigation as active inference. *Biol. Cybern.* **2018**, *112*, 323–343. [CrossRef] [PubMed]
26. Pezzulo, G. An Active Inference view of cognitive control. *Front. Psychol.* **2012**, *3*, 478. [CrossRef]
27. Schwöbel, S.; Kiebel, S.; Marković, D. Active Inference, Belief Propagation, and the Bethe Approximation. *Neural Comput.* **2018**, *30*, 2530–2567. [CrossRef] [PubMed]
28. Matsumoto, T.; Tani, J. Goal-Directed Planning for Habituated Agents by Active Inference Using a Variational Recurrent Neural Network. *Entropy* **2020**, *22*, 564. [CrossRef]
29. Çatal, O.; Verbelen, T.; Nauta, J.; Boom, C.D.; Dhoedt, B. Learning Perception and Planning With Deep Active Inference. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 3952–3956. [CrossRef]
30. Çatal, O.; Wauthier, S.; Verbelen, T.; De Boom, C.; Dhoedt, B. Deep Active Inference for Autonomous Robot Navigation. arXiv:200303220. 2020. Available online: http://arxiv.org/abs/2003.03220 (accessed on 22 May 2020).

31. Sancaktar, C.; van Gerven, M.; Lanillos, P. End-to-End Pixel-Based Deep Active Inference for Body Perception and Action. arXiv:200105847 [cs, q-bio]. 2020 [cited 18 Sep 2020]. Available online: http://arxiv.org/abs/2001.05847 (accessed on 18 September 2020).
32. Tschantz, A.; Seth, A.K.; Buckley, C.L. Learning action-oriented models through active inference. *PLoS Comput. Biol.* **2020**, *16*, e1007805. [CrossRef] [PubMed]
33. Schwartenbeck, P.; Passecker, J.; Hauser, T.U.; Fitzgerald, T.H.; Kronbichler, M.; Friston, K.J. Computational mechanisms of curiosity and goal-directed exploration. *eLife* **2019**, *8*, 45. [CrossRef]
34. Tschantz, A.; Millidge, B.; Seth, A.K.; Buckley, C.L. Reinforcement Learning through Active Inference. ICLR. 2020. Available online: http://arxiv.org/abs/2002.12636 (accessed on 9 April 2021).
35. Marković, D.; Goschke, T.; Kiebel, S.J. Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cogn. Affect. Behav. Neurosci.* **2020**. [CrossRef]
36. Friston, K.; Da Costa, L.; Hafner, D.; Hesp, C.; Parr, T. Sophisticated Inference. *Neural Comput.* **2021**, *33*, 713–763. [CrossRef]
37. Lanillos, P.G. Adaptive Robot Body Learning and Estimation Through Predictive Coding. In Proceedings of the2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 4083–4090. [CrossRef]
38. Lanillos, P.; Pages, J.; Cheng, G. Robot Self/Other Distinction: Active Inference Meets Neural Networks Learning in a Mirror. arXiv:200405473 [cs]. 2020. Available online: http://arxiv.org/abs/2004.05473 (accessed on 18 September 2020).
39. Friston, K.J.; Lin, M.; Frith, C.D.; Pezzulo, G.; Hobson, J.A.; Ondobaka, S. Active Inference, Curiosity and Insight. *Neural Comput.* **2017**, *29*, 2633–2683. [CrossRef]
40. Smith, R.; Schwartenbeck, P.; Parr, T.; Friston, K.J. An Active Inference Approach to Modeling Structure Learning: Concept Learning as an Example Case. *Front. Comput. Neurosci.* **2020**, *14*. [CrossRef]
41. Wauthier, S.T.; Çatal, O.; Verbelen, T.; Dhoedt, B. Sleep: Model Reduction. In *Deep Active Inference*; Springer: Cham, Switzerland, 2020; p. 13.
42. Cullen, M.; Davey, B.; Friston, K.J.; Moran, R.J. Active Inference in OpenAI Gym: A Paradigm for Computational Investigations into Psychiatric Illness. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **2018**, *3*, 809–818. [CrossRef]
43. Fountas, Z.; Sajid, N.; Mediano, P.A.M.; Friston, K. Deep Active Inference Agents Using Monte-Carlo Methods. arXiv:200604176 [cs, q-bio, stat]. 2020. Available online: http://arxiv.org/abs/2006.04176 (accessed on 16 July 2020).
44. Ueltzhöffer, K. Deep active inference. *Biol. Cybern.* **2018**, *112*, 547–573. [CrossRef]
45. Marković, D.; Reiter, A.M.F.; Kiebel, S.J. Predicting change: Approximate inference under explicit representation of temporal structure in changing environments. *PLoS Comput. Biol.* **2019**, *15*, e1006707. [CrossRef] [PubMed]
46. Markovic, D.; Stojic, H.; Schwoebel, S.; Kiebel, S.J. An Empirical Evaluation of Active Inference in Multi-Armed Bandits. arXiv:210108699 [cs]. 2021. Available online: http://arxiv.org/abs/2101.08699 (accessed on 26 January 2021).
47. Sajid, N.; Ball, P.J.; Friston, K.J. Active Inference: Demystified and Compared. arXiv:190910863 [cs, q-bio]. 2020. Available online: http://arxiv.org/abs/1909.10863 (accessed on 30 April 2020).
48. Parr, T.; Friston, K.J. Active inference and the anatomy of oculomotion. *Neuropsychologia* **2018**, *111*, 334–343. [CrossRef] [PubMed]
49. Parr, T. The Computational Neurology of Active Vision. Ph.D. Thesis, University College London, London, UK, 2019.
50. Mirza, M.B.; Adams, R.A.; Mathys, C.D.; Friston, K.J. Scene Construction, Visual Foraging, and Active Inference. *Front. Comput. Neurosci.* **2016**, *10*, 56. [CrossRef]
51. Parr, T.; Friston, K.J. Uncertainty, epistemics and active inference. *J. R. Soc. Interface* **2017**, *14*, 20170376. [CrossRef]
52. Adams, R.A.; Stephan, K.E.; Brown, H.R.; Frith, C.D.; Friston, K.J. The Computational Anatomy of Psychosis. *Front. Psychiatry* **2013**, *4*. [CrossRef]
53. Smith, R.; Kirlic, N.; Stewart, J.L.; Touthang, J.; Kuplicki, R.; Khalsa, S.S.; Feinstein, J.; Paulus, M.P.; Aupperle, R.L. Greater decision uncertainty characterizes a transdiagnostic patient sample during approach-avoidance conflict: A computational modeling approach. *PsyArXiv* **2020**. [CrossRef]
54. Smith, R.; Schwartenbeck, P.; Stewart, J.L.; Kuplicki, R.; Ekhtiari, H.; Investigators, T.; Martin, P. Imprecise Action Selection in Substance Use Disorder: Evidence for Active Learning Impairments When Solving the Explore-Exploit Dilemma. *PsyArXiv* **2020**. [CrossRef]
55. Smith, R.; Kuplicki, R.; Feinstein, J.; Forthman, K.L.; Stewart, J.L.; Paulus, M.P.; Tulsa 1000 Investigators; Khalsa, S.S. A Bayesian computational model reveals a failure to adapt interoceptive precision estimates across depression, anxiety, eating, and substance use disorders. *PLoS Comput. Biol.* **2020**, *16*, e1008484. [CrossRef]
56. Millidge, B. Deep active inference as variational policy gradients. *J. Math. Psychol.* **2020**, *96*, 102348. [CrossRef]
57. Friston, K.J.; Fitzgerald, T.H.B.; Rigoli, F.; Schwartenbeck, P.; Pezzulo, G. Active Inference: A Process Theory. *Neural Comput.* **2017**, *29*, 1–49. [CrossRef]
58. Parr, T.; Friston, K.J. The Anatomy of Inference: Generative Models and Brain Structure. *Front. Comput. Neurosci.* **2018**, *12*, 90. [CrossRef] [PubMed]
59. Schwartenbeck, P.; Fitzgerald, T.H.B.; Mathys, C.; Dolan, R.; Friston, K. The Dopaminergic Midbrain Encodes the Expected Certainty about Desired Outcomes. *Cereb. Cortex* **2015**, *25*, 3434–3445. [CrossRef] [PubMed]
60. Fitzgerald, T.H.B.; Dolan, R.J.; Friston, K. Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* **2015**, *9*. [CrossRef] [PubMed]

61. Schwartenbeck, P.; Fitzgerald, T.; Dolan, R.J.; Friston, K.J. Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* **2013**, *4*, 710. [CrossRef]
62. Schwartenbeck, P.; Fitzgerald, T.H.B.; Mathys, C.; Dolan, R.; Kronbichler, M.; Friston, K. Evidence for surprise minimization over value maximization in choice behavior. *Sci. Rep.* **2015**, *5*, 16575. [CrossRef]
63. Sengupta, B.; Stemmler, M.B.; Friston, K.J. Information and Efficiency in the Nervous System—A Synthesis. *PLoS Comput. Biol.* **2013**, *9*, e1003157. [CrossRef]
64. Levy, W.B.; Baxter, R.A. Energy Efficient Neural Codes. *Neural Comput.* **1996**, *8*, 531–543. [CrossRef]
65. Dan, Y.; Atick, J.J.; Reid, R.C. Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. *J. Neurosci.* **1996**, *16*, 3351–3362. [CrossRef]
66. Lewicki, M.S. Efficient coding of natural sounds. *Nat. Neurosci.* **2002**, *5*, 356–363. [CrossRef]
67. Chen, B.L.; Hall, D.H.; Chklovskii, D.B. Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 4723–4728. [CrossRef]
68. Raj, A.; Chen, Y.-H. The Wiring Economy Principle: Connectivity Determines Anatomy in the Human Brain. *PLoS ONE* **2011**, *6*, e14832. [CrossRef]
69. Barlow, H. Redundancy reduction revisited. *Comput. Neural Syst.* **2001**, *12*, 241–253. [CrossRef]
70. Barlow, H.B. *Possible Principles Underlying the Transformations of Sensory Messages*; The MIT Press: 1961. Available online: https://www.universitypressscholarship.com/view/10.7551/mitpress/9780262518420.001.0001/upso-9780262518420-chapter-13 (accessed on 9 April 2021).
71. Binder, M.D.; Hirokawa, N.; Windhorst, U. (Eds.) *Efficient Coding Hypothesis. Encyclopedia of Neuroscience*; Springer: Berlin/Heidelberg, Germany, 2009; p. 1037. [CrossRef]
72. Denève, S.; Machens, C.K. Efficient codes and balanced networks. *Nat. Neurosci.* **2016**, *19*, 375–382. [CrossRef]
73. Chelaru, M.I.; Dragoi, V. Efficient coding in heterogeneous neuronal populations. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 16344–16349. [CrossRef] [PubMed]
74. Kostal, L.; Lánský, P.; Rospars, J.-P. Efficient Olfactory Coding in the Pheromone Receptor Neuron of a Moth. *PLoS Comput. Biol.* **2008**, *4*, e1000053. [CrossRef]
75. Olshausen, B.A.; Field, D.J. Natural image statistics and efficient coding. *Net. Comput. Neural Syst.* **1996**, *7*, 333–339. [CrossRef]
76. Olshausen, B.A.; O'Connor, K.N. A new window on sound. *Nat. Neurosci.* **2002**, *5*, 292–294. [CrossRef]
77. Simoncelli, E.P.; Olshausen, B.A. Natural Image Statistics and Neural Representation. *Annu. Rev. Neurosci.* **2001**, *24*, 1193–1216. [CrossRef]
78. Karklin, Y.; Simoncelli, E.P. Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons. In *Advances in Neural Information Processing Systems 24*; Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q., Eds.; Curran Associates, Inc.: New York, NY, USA, 2011; pp. 999–1007.
79. Olshausen, B.A.; Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nat. Cell Biol.* **1996**, *381*, 607–609. [CrossRef] [PubMed]
80. Olshausen, B.A.; Field, D.J. Vision and the Coding of Natural Images: The human brain may hold the secrets to the best image-compression algorithms. *Am. Sci.* **2000**, *88*, 238–245. [CrossRef]
81. Bennett, C.H. Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Stud. Hist. Philos. Sci. Part B: Stud. Hist. Philos. Mod. Phys.* **2003**, *34*, 501–510. [CrossRef]
82. Landauer, R. Irreversibility and Heat Generation in the Computing Process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
83. Ito, S. Stochastic Thermodynamic Interpretation of Information Geometry. *Phys. Rev. Lett.* **2018**, *121*, 030605. [CrossRef] [PubMed]
84. Crooks, G.E. Measuring Thermodynamic Length. *Phys. Rev. Lett.* **2007**, *99*, 100602. [CrossRef]
85. Amari, S.-I. Natural Gradient Works Efficiently in Learning. *Neural Comput.* **1998**, *10*, 251–276. [CrossRef]
86. Wilson, H.R.; Cowan, J.D. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Biol. Cybern.* **1973**, *13*, 55–80. [CrossRef]
87. Fricker, D.; Verheugen, J.A.H.; Miles, R. Cell-attached measurements of the firing threshold of rat hippocampal neurones. *J. Physiol.* **1999**, *517*, 791–804. [CrossRef]
88. Marreiros, A.C.; Daunizeau, J.; Kiebel, S.J.; Friston, K.J. Population dynamics: Variance and the sigmoid activation function. *NeuroImage* **2008**, *42*, 147–157. [CrossRef]
89. Marreiros, A.C.; Kiebel, S.J.; Daunizeau, J.; Harrison, L.M.; Friston, K.J. Population dynamics under the Laplace assumption. *NeuroImage* **2009**, *44*, 701–714. [CrossRef]
90. Friston, K.J.; Harrison, L.; Penny, W. Dynamic causal modelling. *NeuroImage* **2003**, *19*, 1273–1302. [CrossRef]
91. Deco, G.; Jirsa, V.K.; Robinson, P.A.; Breakspear, M.; Friston, K. The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields. *PLoS Comput. Biol.* **2008**, *4*, e1000092. [CrossRef]
92. Moran, R.J.; Pinotsis, D.A.; Friston, K.J. Neural masses and fields in dynamic causal modeling. *Front. Comput. Neurosci.* **2013**, *7*, 57. [CrossRef]
93. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
94. Stone, J.V. *Artificial Intelligence Engines: A Tutorial Introduction to the Mathematics of Deep Learning*; Sebtel Press: Warszawa, Poland, 2019.
95. Engel, J. Polytomous logistic regression. *Stat. Neerlandica* **1988**, *42*, 233–252. [CrossRef]

96. Huang, F.-L.; Hsieh, C.-J.; Chang, K.-W.; Lin, C.-J. Iterative scaling and coordinate descent methods for maximum entropy. *ACL-IJCNLP 2009 Conf. Short Pap.* **2009**. [CrossRef]
97. Rodríguez, G. Lecture Notes on Generalized Linear Models. Available online: https://data.princeton.edu/wws509/notes/ (accessed on 9 April 2021).
98. Åström, K. Optimal control of Markov processes with incomplete state information. *J. Math. Anal. Appl.* **1965**, *10*, 174–205. [CrossRef]
99. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann Math Statist.* **1951**, *22*, 79–86. [CrossRef]
100. Joyce, J.M. *Kullback-Leibler Divergence. International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 720–722. [CrossRef]
101. Rezende, D.J. Short Notes on Divergence Measures. 2018. Available online: https://danilorezende.com/wp-content/uploads/2018/07/divergences.pdf (accessed on 9 April 2021).
102. Stachenfeld, K.L.; Botvinick, M.M.; Gershman, S.J. The hippocampus as a predictive map. *Nat. Neurosci.* **2017**, *20*, 1643–1653. [CrossRef]
103. Hafting, T.; Fyhn, M.; Molden, S.; Moser, M.-B.; Moser, E.I. Microstructure of a spatial map in the entorhinal cortex. *Nat. Cell Biol.* **2005**, *436*, 801–806. [CrossRef]
104. Chen, L.L.; Lin, L.-H.; Green, E.J.; Barnes, C.A.; McNaughton, L.A. Head-direction cells in the rat posterior cortex. *Exp. Brain Res.* **1994**, *16*, 8–23. [CrossRef]
105. Taube, J.; Muller, R.; Ranck, J. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **1990**, *10*, 420–435. [CrossRef]
106. Stein, R.B.; Weber, D.J.; Aoyagi, Y.; Prochazka, A.; Wagenaar, J.B.M.; Shoham, S.; Normann, R.A. Coding of position by simultaneously recorded sensory neurones in the cat dorsal root ganglion. *J. Physiol.* **2004**, *560*, 883–896. [CrossRef]
107. Wagenaar, J.B.; Ventura, V.; Weber, D.J. State-space decoding of primary afferent neuron firing rates. *J. Neural Eng.* **2011**, *8*, 016002. [CrossRef]
108. Weber, D.; Stein, R.; Everaert, D.; Prochazka, A. Decoding Sensory Feedback From Firing Rates of Afferent Ensembles Recorded in Cat Dorsal Root Ganglia in Normal Locomotion. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2006**, *14*, 240–243. [CrossRef] [PubMed]
109. Hubel, D.H.; Wiesel, T.N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **1959**, *148*, 574–591. [CrossRef]
110. Amari, S. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016.
111. Ay, N.; Jost, J.; Lê, H.V.; Schwachhöfer, L. *Information Geometry*; Springer: Cham, Switzerland, 2017. [CrossRef]
112. Nielsen, F. An Elementary Introduction to Information Geometry. arXiv:180808271 [cs, math, stat]. 2018. Available online: http://arxiv.org/abs/1808.08271 (accessed on 11 August 2019).
113. Cencov, N.N. *Statistical Decision Rules and Optimal Inference*; American Mathematical Society: Providence, RI, USA, 1982.
114. Liang, T.; Poggio, T.; Rakhlin, A.; Stokes, J. Fisher-Rao Metric, Geometry, and Complexity of Neural Networks. arXiv:171101530 [cs, stat]. 2017. Available online: http://arxiv.org/abs/1711.01530 (accessed on 11 August 2019).
115. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New Jersey, NJ, USA, 2006.
116. Amari, S.; Douglas, S.C. Why natural gradient? In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat No98CH36181), Seattle, WA, USA, 12–15 May 1998; pp. 1213–1216. [CrossRef]
117. Bernacchia, A.; Lengyel, M.; Hennequin, G. Exact natural gradient in deep linear networks and its application to the nonlinear case. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 5941–5950.
118. Zonghai, S.; Buhai, S. The Projection Adaptive Natural Gradient Online Algorithm for SVM. Available online: https://www.infona.pl//resource/bwmeta1.element.ieee-art-000005573523 (accessed on 9 April 2021).
119. Zhang, L.; Cichocki, A.; Amari, S. Natural gradient algorithm for blind separation of overdetermined mixture with additive noise. *IEEE Signal Process. Lett.* **1999**, *6*, 293–295. [CrossRef]
120. Zhang, Z.; Sun, H.; Zhong, F. Natural gradient-projection algorithm for distribution control. *Optim. Control. Appl. Methods* **2008**, *30*, 495–504. [CrossRef]
121. Duan, T.; Anand, A.; Ding, D.Y.; Thai, K.K.; Basu, S.; Ng, A.; Schuler, A. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. Available online: http://proceedings.mlr.press/v119/duan20a.html (accessed on 9 April 2021).
122. Loeliger, H.-A. An Introduction to factor graphs. *IEEE Signal Process. Mag.* **2004**, *21*, 28–41. [CrossRef]
123. Mirza, M.B.; Adams, R.A.; Mathys, C.; Friston, K.J. Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* **2018**, *13*, e0190429. [CrossRef]
124. Parr, T.; Markovic, D.; Kiebel, S.J.; Friston, K.J. Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Sci. Rep.* **2019**, *9*, 1889. [CrossRef]
125. Yedidia, J.S.; Freeman, W.T.; Weiss, Y. Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms. *IEEE Trans. Inf. Theory* **2005**, *51*, 2282–2312. [CrossRef]
126. Dauwels, J. On Variational Message Passing on Factor Graphs. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 2546–2550. [CrossRef]
127. Winn, J.; Bishop, C.M. Variational Message Passing. *J. Mach. Learn. Res.* **2005**, *34*, 661–694.