entropy

MDPI

Article

A Technical Critique of Some Parts of the Free Energy Principle

Martin Biehl [1,*,†], Felix A. Pollock [2,*,†] and Ryota Kanai [1]

[1] Araya Inc., Tokyo 107-6024, Japan; kanair@araya.org
[2] School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia
* Correspondence: martin@araya.org (M.B.); felix.pollock@monash.edu (F.A.P.)
† These authors contributed equally to this work.

Abstract: We summarize the original formulation of the free energy principle and highlight some technical issues. We discuss how these issues affect related results involving generalised coordinates and, where appropriate, mention consequences for and reveal, up to now unacknowledged, differences from newer formulations of the free energy principle. In particular, we reveal that various definitions of the "Markov blanket" proposed in different works are not equivalent. We show that crucial steps in the free energy argument, which involve rewriting the equations of motion of systems with Markov blankets, are not generally correct without additional (previously unstated) assumptions. We prove by counterexamples that the original free energy lemma, when taken at face value, is wrong. We show further that this free energy lemma, when it does hold, implies the equality of variational density and ergodic conditional density. The interpretation in terms of Bayesian inference hinges on this point, and we hence conclude that it is not sufficiently justified. Additionally, we highlight that the variational densities presented in newer formulations of the free energy principle and lemma are parametrised by different variables than in older works, leading to a substantially different interpretation of the theory. Note that we only highlight some specific problems in the discussed publications. These problems do not rule out conclusively that the general ideas behind the free energy principle are worth pursuing.

Keywords: free energy principle; stochastic differential equations; Markov blanket

Citation: Biehl, M.; Pollock, F.A.; Kanai, R. A Technical Critique of Some Parts of the Free Energy Principle. *Entropy* **2021**, *23*, 293. https://doi.org/10.3390/e23030293

Academic Editor: Kevin H. Knuth

Received: 16 September 2020
Accepted: 24 February 2021
Published: 27 February 2021

1. Overview

In [1], it was argued that the internal coordinates of an ergodic random dynamical system with a Markov blanket necessarily appear to engage in active Bayesian inference. Here, we reproduce the argument supporting this interpretation in detail and highlight at which points it faces technical issues. In the course of our critique, we also mention issues of some closely related alternative arguments. In cases where our results have clear consequences for the more recent related publications [2,3], we also mention those. In particular, we point out a conceptual difference in these latter works that has not previously been acknowledged. However, our analysis thereof does not go beyond a few remarks. In an additional section, we discuss the effect of our argument on [4]. The logical structure of the present paper is depicted in Figure 1. We note that the technical issues presented here do not affect the validity of approaches where a (expected) free energy minimizing agent is assumed a priori, as presented in, e.g., [5]. None of [1–4] make this assumption; they instead aim to identify the conditions under which such agents will emerge within a given stochastic process. We criticize specific formal issues in the latter publications but leave open whether they can be fixed. We now briefly introduce the setting of [1] and then sketch the content of this paper. We now briefly introduce the setting of [1] and then sketch the content of this paper.

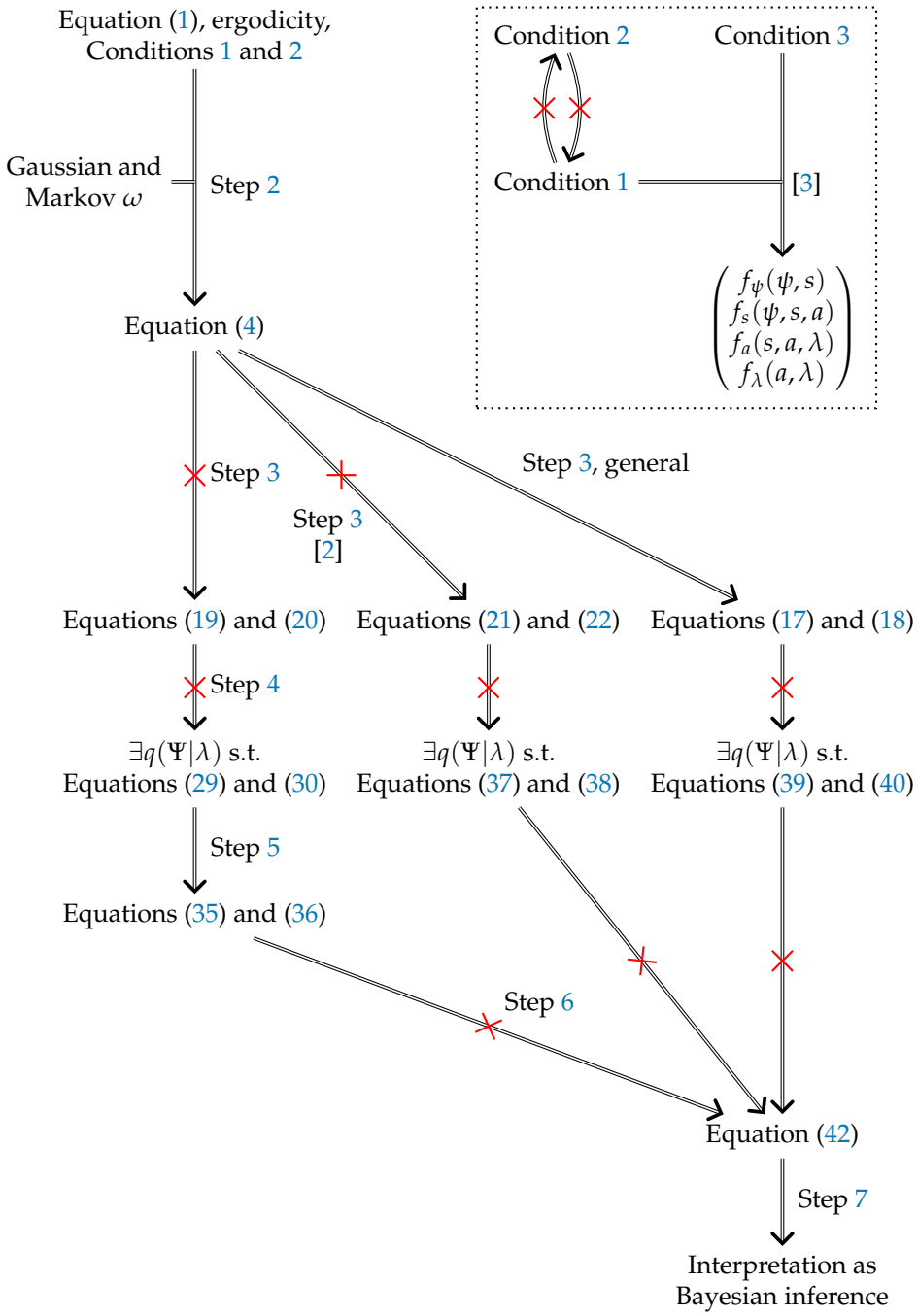

Figure 1. Argument visualization. Numbers labelling edges indicate corresponding steps in this paper. Struck out edges indicate implications that we prove incorrect. The main argument in [1] takes the left path. The box in the top right indicates the relations between Conditions 1 to 3 and their role in [3]. Merged edges indicate a logical AND combination of the parent nodes.

The starting point is a random dynamical system whose evolution is governed by the stochastic differential equation:

$$\dot{x} = f(x) + \omega, \tag{1}$$

where the system state x and vector field $f(x)$ are multi-dimensional and ω is a Gaussian noise term. There is an additional assumption that the system is ergodic, such that the steady state probability density $p^*(x)$ is well defined (In the original paper, the ergodic

density is simply denoted $p(x)$. We here add a star to highlight that it is a time independent probability density.). In this case, $-\ln p^*(x)$ plays the role of a potential function, in the sense that f can be formulated in terms of its gradients [6,7].

It is then assumed that there is a coordinate system $x = (\psi, s, a, \lambda)$ with $\psi = (\psi_1, ..., \psi_{n_\psi})$, $s = (s_1, ..., s_{n_s})$, $a = (a_1, ..., a_{n_a})$, and $\lambda = (\lambda_1, ..., \lambda_{n_\lambda})$, referred to as external, sensory, active, and internal coordinates (these are called "states" in [1]), respectively, such that the following condition holds:

Condition 1. *The function $f(x)$ can be written as:*

$$f(x) = \begin{pmatrix} f_\psi(\psi, s, a) \\ f_s(\psi, s, a) \\ f_a(s, a, \lambda) \\ f_\lambda(s, a, \lambda) \end{pmatrix}. \tag{2}$$

This particular structure is described as "[formalizing] the dependencies implied by the Markov blanket" [1]. In contrast, more recent works [2,3] formulated the Markov blanket in terms of the statistical dependencies of the ergodic density $p^*(x) = p^*(\psi, s, a, \lambda)$. Specifically, the following condition is presented:

Condition 2. *The ergodic density factorises as:*

$$p^*(\psi, s, a, \lambda) = p^*(\psi|s, a)p^*(\lambda|s, a)p^*(s, a). \tag{3}$$

In other words, the internal and external coordinates are independently distributed when conditioned on the sensory and active coordinates. This means we have two different formal expressions of what constitutes a Markov blanket in these publications, and their relationship has not previously been established.

Taking Condition 1 to hold, the argument of [1] then proceeds along the following steps:

Step 2 Rewrite the vector field $f(\psi, s, a, \lambda)$ describing the dynamics of the system in terms of the gradient of the negative logarithm of the ergodic density $p^*(\psi, s, a, \lambda)$ of that system.

Step 3 Rewrite the components $f_\lambda(s, a, \lambda)$ and $f_a(s, a, \lambda)$ of the vector field $f(\psi, s, a, \lambda)$ in terms of only partial gradients of the negative logarithm of $p^*(\psi, s, a, \lambda)$.

Step 4 Assert (in the free energy lemma) the existence of a density $q(\psi|\lambda)$ over the external coordinates ψ parameterized by the internal coordinates λ and that $f(\psi, s, a, \lambda)$ can again be rewritten, this time in terms of a free energy depending on $q(\Psi|\lambda)$ (here, and whenever it would otherwise be ambiguous, we use a capitalized Ψ to indicate full distributions, rather than the probability density for a specific value of ψ).

Step 5 Claim that the equivalence of the equations of motion in Step 3 and Step 4 implies that certain partial gradients of the KL divergence between $q(\Psi|\lambda)$ and the conditional ergodic density $p^*(\Psi|s, a, \lambda)$ must vanish.

Step 6 Claim that it follows from Step 5 that $q(\Psi|\lambda)$ and $p^*(\Psi|s, a, \lambda)$ are "rendered" equal.

Step 7 Interpret:

- $p^*(\Psi|s, a, \lambda)$ as a posterior over external coordinates given particular values of sensor, active, and internal coordinates,
- $q(\Psi|\lambda)$ as encoding Bayesian beliefs about the external coordinates by the internal coordinates, and
- their equality as the internal coordinates appearing to "solve the problem of Bayesian inference".

In the present paper, we make the following main observations:

- The re-expression of Equation (1) in the form chosen in Step 2 is derived under restrictive assumptions, including that the system is subject to Gaussian and Markov noise.
- Conditions 1 and 2 are independent of each other.
- Conditions 1 and 3 together lead to a system where the interpretation of s and a as sensory and active coordinates is questionable.
- Under both Conditions 1 and 2, the expressions of $f_\lambda(s,a,\lambda)$ and $f_a(s,a,\lambda)$ resulting from Step 3 are not as general as those contained in the result of Step 2. The more general alternative expression derived in [2] remains insufficiently general.
- Under both Conditions 1 and 2, the free energy lemma, when taken at face value, is wrong and cannot be salvaged by using alternatives in Step 3.
- Under both Conditions 1 and 2, contrary to Step 6, the vanishing of the gradient of the KL divergence does not imply the equality of $q(\Psi|\lambda)$ and $p^*(\Psi|s,a,\lambda)$.
- As a consequence, the basic preconditions for the interpretations in Step 7 are not implied by either of the two proposed Markov blankets Conditions 1 and 2.

The latter [4] presents an argument almost identical to the one in the original [1]. In Section 8, we discuss how our observations apply to this publication.

2. Expression via the Gradient of the Ergodic Density

Here, we introduce the expression of the system's dynamics Equation (1) in the form used for the free energy lemma (Lemma 2.1 in [1]). This form expresses the dynamics of the internal and active coordinates of the given ergodic random dynamical system in terms of the gradient of the ergodic density $p^*(x)$. In accordance with the results of [7], $f(x)$ is rewritten as (see Equation (2.5) in [1]):

$$f(x) = (\Gamma + R) \cdot \nabla \ln p^*(x), \tag{4}$$

where Γ is the diffusion matrix, which we will take to be block diagonal (in [1], and later work such as [2], Γ is taken to be proportional to the identity matrix), and R is an antisymmetric matrix, defined through the relation:

$$MR + RM^T = M\Gamma - \Gamma M^T, \tag{5}$$

with

$$M_{ij} = \nabla_j f_i(x). \tag{6}$$

Here, and in all of [1–4], both Γ and R are assumed constant. We emphasise here that, for general nonlinear models, these matrices can vary with the coordinates and Equation (5) holds only approximately [8,9] (the exact conditions under which these matrices can be chosen to be constant can be found in [9,10] and, for the discrete state case, [11]). Moreover, Equation (4) is derived in the literature under the explicit assumption that the fluctuations ω be Gaussian and Markov [6,7]. For the counterexamples we present here, we restrict ourselves to the class of Ornstein–Uhlenbeck processes, for which R and Γ are always constant, and the ergodic density $p^*(x) = p^*(\psi,s,a,\lambda)$ is necessarily a multivariate Gaussian with zero mean. Specifically, following [7],

$$p^*(\psi,s,a,\lambda) := \frac{1}{Z}\exp\left[-\frac{1}{2}(\psi,s,a,\lambda)U(\psi,s,a,\lambda)^\top\right], \tag{7}$$

where (ψ,s,a,λ) is a row vector and Z is a suitable normalisation constant. From Equation (4), it can be seen that,

$$U = -(\Gamma + R)^{-1}M; \tag{8}$$

though we emphasise here that strict relations between M and U can only be made because of the assumption that Γ and R are coordinate independent [12]. This concludes Step 2.

Before moving on to Step 3, we note that, under the assumptions implicit in Step 2, we can express Conditions 1 and 2 in terms of the matrices M and U (in the nonlinear case, these matrices can still be defined in terms of the derivatives of the force vector field and potential, respectively; however, they will be generally coordinate-dependent, even when Γ and R are not [8]). Firstly, since it effectively states that $\nabla_\psi f_a(x) = \nabla_\psi f_\lambda(x) = \nabla_\lambda f_s(x) = \nabla_\lambda f_\psi(x) = 0$,

$$\text{Condition 1} \quad \Leftrightarrow \quad M_{a\psi} = M_{\lambda\psi} = M_{s\lambda} = M_{\psi\lambda} = 0, \tag{9}$$

with $M_{\alpha\beta}$ a block sub-matrix of M in general. Secondly, because of the multivariate Gaussian nature of $p^*(\psi, s, a, \lambda)$, the dependencies of conditional distributions are encoded in the inverse U of the covariance matrix; we therefore have that:

$$\text{Condition 2} \quad \Leftrightarrow \quad U_{\psi\lambda} = U_{\lambda\psi} = 0, \tag{10}$$

where $U_{\alpha\beta}$ is a block sub-matrix of U. These implications bring us to our first observation:

Observation 1. *Neither Condition 1 (the vector field dependency structure) nor Condition 2 (conditional independence in the ergodic distribution) imply the other:*

$$\text{Condition 1} \quad \not\Rightarrow \quad \text{Condition 2} \tag{11}$$

$$\text{Condition 1} \quad \not\Leftarrow \quad \text{Condition 2}. \tag{12}$$

Proof. In Appendix A, we provide direct counterexamples, using the equivalent constraints on the matrices M and U in Equations (9) and (10), for the implication in either direction. That is, there exists a system obeying Condition 1 that does not obey Condition 2 (proving Equation (11)), and there exists one obeying Condition 2 that does not obey Condition 1 (proving Equation (12)). □

Henceforth, unless otherwise stated, we will assume both Conditions 1 and 2. Any implications that fail to hold in this special case cannot hold generally.

3. Re-Expression Using only Partial Gradients

For Step 3, we focus on the components $f_\lambda = (f_{\lambda_1}, ..., f_{n_\lambda})$ and $f_a = (f_{a_1}, ..., f_{n_a})$ of f. Without loss of generality, we can rewrite them from Equation (4) as:

$$\begin{aligned} f_a(s, a, \lambda) =& \big(R_{a\psi} \cdot \nabla_\psi + R_{as} \cdot \nabla_s + (\Gamma_{aa} + R_{aa}) \cdot \nabla_a \\ &+ R_{a\lambda} \cdot \nabla_\lambda\big) \ln p^*(\psi, s, a, \lambda), \end{aligned} \tag{13}$$

$$\begin{aligned} f_\lambda(s, a, \lambda) =& \big(R_{\lambda\psi} \cdot \nabla_\psi + R_{\lambda s} \cdot \nabla_s + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda \\ &+ R_{\lambda a} \cdot \nabla_a\big) \ln p^*(\psi, s, a, \lambda), \end{aligned} \tag{14}$$

where Γ_{nm} (R_{nm}) is the block of Γ (R) connecting derivatives with respect to the m coordinates to the time derivatives of the n coordinates. The expectation value with respect to $p^*(\psi|s, a, \lambda)$ leaves the left-hand side of these equations unchanged. A few manipulations ([2] cf. Equation (12.14), p. 129) reveal that, on the right-hand side, this leads to the ergodic density $p^*(\psi, s, a, \lambda)$ being replaced by the marginalised ergodic density $p^*(s, a, \lambda)$ so that we get:

$$\begin{aligned} f_a(s, a, \lambda) =& \big(R_{a\psi} \cdot \nabla_\psi + R_{as} \cdot \nabla_s + (\Gamma_{aa} + R_{aa}) \cdot \nabla_a \\ &+ R_{a\lambda} \cdot \nabla_\lambda\big) \ln p^*(s, a, \lambda) \end{aligned} \tag{15}$$

$$\begin{aligned} f_\lambda(s, a, \lambda) =& \big(R_{\lambda\psi} \cdot \nabla_\psi + R_{\lambda s} \cdot \nabla_s + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda \\ &+ R_{\lambda a} \cdot \nabla_a\big) \ln p^*(s, a, \lambda). \end{aligned} \tag{16}$$

Since $\nabla_\psi \ln p^*(s,a,\lambda) = 0$, the terms involving ∇_ψ drop out:

$$f_a(s,a,\lambda) = (R_{as} \cdot \nabla_s + (\Gamma_{aa} + R_{aa}) \cdot \nabla_a + R_{a\lambda} \cdot \nabla_\lambda) \ln p^*(s,a,\lambda), \tag{17}$$
$$f_\lambda(s,a,\lambda) = (R_{\lambda s} \cdot \nabla_s + R_{\lambda a} \cdot \nabla_a + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda) \ln p^*(s,a,\lambda). \tag{18}$$

We are not aware of how to further simplify this equation without additional assumptions. However, in (Equations (2.5) and (2.6) of [1]), all of the off-diagonal terms are implicitly assumed to vanish, i.e., Equation (4) is equated with:

$$f_a(s,a,\lambda) = (\Gamma_{aa} + R_{aa}) \cdot \nabla_a \ln p^*(s,a,\lambda), \tag{19}$$
$$f_\lambda(s,a,\lambda) = (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda \ln p^*(s,a,\lambda). \tag{20}$$

This equation is the result of Step 3.

More recently (Appendix B of [2]), a more detailed discussion of Equation (4) was presented, where it was claimed that Condition 1 implies Condition 2 (cf. our Observation 1) along with the following simplification of Equations (17) and (18) ([2], Equations (12.8)–(12.11), (12.15), pp. 126–129):

$$f_a(s,a,\lambda) = ((\Gamma_{aa} + R_{aa}) \cdot \nabla_a + R_{a\lambda} \cdot \nabla_\lambda) \ln p^*(s,a,\lambda), \tag{21}$$
$$f_\lambda(s,a,\lambda) = (R_{\lambda a} \cdot \nabla_a + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda) \ln p^*(s,a,\lambda). \tag{22}$$

However, Equations (21) and (22) are still provably less general than Equations (13) and (14), even when both Conditions 1 and 2 are satisfied.

Observation 2. *Given a random dynamical system obeying Equation (1), ergodicity, and both Conditions 1 and 2, none of Equations (19)–(22) generally hold.*

Proof. By counterexample, see Appendix B. There, we show explicitly that a model satisfying the above assumptions does not satisfy the equations in question. □

In order to arrive at Equations (21) and (22) from Equations (17) and (18) in general, one must remove the offending "solenoidal flow" terms by fiat. That is, one assumes $R_{as} = R_{\lambda s} = 0$. In [2], Equation (12.4), the following, even stronger, condition was assumed as an alternative starting point (along with Condition 2):

Condition 3. *The blocks of the R matrix appearing in Equation (4) coupling (s,a) coordinates to λ and ψ coordinates and ψ coordinates to λ coordinates vanish, i.e.,*

$$R_{\psi s} = R_{\psi a} = R_{\psi\lambda} = R_{s\lambda} = R_{a\lambda} = 0. \tag{23}$$

This is claimed to imply $M_{\psi\lambda} = M_{\lambda\psi} = 0$, but not the full Condition 1. However, in [3], both Conditions 1 and 3 were assumed (along with $R_{as} = 0$). This prompts our next observation.

Observation 3. *In a system satisfying both Conditions 1 and 3, the internal coordinates cannot be directly influenced by the sensory coordinates: $f_\lambda(s,a,\lambda) = f_\lambda(a,\lambda)$, and the external coordinates cannot be directly influenced by the active coordinates: $f_\psi(\psi,s,a) = f_\psi(\psi,s)$.*

Proof. From Equation (5), it follows that:

$$M = (\Gamma + R)M^T(\Gamma - R)^{-1}, \tag{24}$$

with the inverse replaced by a pseudoinverse if $\Gamma - R$ is not invertible. Therefore, if $\Gamma_{\alpha\beta} = \delta_{\alpha\beta}\Gamma_{\alpha\alpha}$ and $R_{\alpha\beta} = \delta_{\alpha\beta}R_{\alpha\alpha}$ for blocks of coordinates labelled by α and β, then:

$$M_{\alpha\beta} = (\Gamma_{\alpha\alpha} + R_{\alpha\alpha})M_{\beta\alpha}^T(\Gamma_{\beta\beta} - R_{\beta\beta})^{-1}, \tag{25}$$

and $M_{\beta\alpha} = 0 \Rightarrow M_{\alpha\beta} = 0$.

Condition 3 implies that only the nonzero blocks of R are $R_{\psi\psi}$, R_{ss}, R_{sa}, R_{as}, R_{aa}, and $R_{\lambda\lambda}$, and Γ is assumed to be block diagonal. As noted in Equation (9), Condition 1 requires that $M_{a\psi} = M_{\lambda\psi} = M_{s\lambda} = M_{\psi\lambda} = 0$. Through Equation (25), these together imply that $M_{\lambda s} = M_{\psi a} = 0$, and hence that:

$$f(x) = \begin{pmatrix} f_\psi(\psi, s) \\ f_s(\psi, s, a) \\ f_a(s, a, \lambda) \\ f_\lambda(a, \lambda) \end{pmatrix}, \tag{26}$$

as shown. □

In this case, the four sets of coordinates interact in a chain, and it is questionable whether the s and a coordinates can be meaningfully interpreted, respectively, as sensory inputs to the internal coordinates or their boundary-mediated influence on the external coordinates.

4. Free Energy Lemma

The relation of the dynamics of the internal coordinates to Bayesian beliefs is made by introducing a density (called the variational density) $q(\Psi|\lambda)$ that is then interpreted as encoding a Bayesian belief. It is parameterized by the internal coordinates λ and claimed to be "arbitrary". We take this "at face value" and consider $q(\Psi|\lambda)$ to be parameterized only by λ and, therefore, to be independent of (s, a). (We note that there is a convention in the literature on variational Bayesian inference, e.g., in [13], to drop the observed variables/data in the variational density. It is possible that in [1], (s, a) was seen as observed variables and dropped from the variational density $q(\Psi|\lambda)$ as in this convention. However, the reason that dropping the observed variables is justified in the established convention is that those observed variables are fixed throughout the minimization of the variational free energy and the parameters of the variational density do not influence the observed data in any way. In other words, the variational density is optimized for a single data point. In [1], the data point was continuously changing and partially doing so with dependence on the parameter λ as $\dot{a} = f_a(s, a, \lambda)$. These differences and their consequences are non-trivial and beyond the scope of this paper, so we assume that the variational density does not depend on (s, a).) If $q(\Psi|\lambda)$ is allowed to depend on (s, a), Observation 4 does not apply, and the free energy lemma is made trivially true by setting $q(\psi|s, a, \lambda) := p^*(\psi|s, a, \lambda)$. The existence of the variational density $q(\Psi|\lambda)$ is asserted by the free energy lemma (see Lemma 2.1 in [1]) (Explicitly, the free energy lemma asserts the existence of a free energy $F(s, a, \lambda)$ in terms of which $f(\psi, s, a, \lambda)$ can be expressed and not the existence of $q(\Psi|\lambda)$. However, since the free energy is defined as a functional of $q(\Psi|\lambda)$, it exists if and only if a suitable $q(\Psi|\lambda)$ exists.).

More precisely, the free energy lemma (and Step 4) asserts that for every ergodic density (equivalently as expressed in [1], for every Gibbs energy $G(x) := -\ln p^*(\psi, s, a, \lambda)$) $p^*(\psi, s, a, \lambda)$ of a system obeying Equations (19) and (20), there is a free energy $F(s, a, \lambda)$, defined as:

$$F(s, a, \lambda) := -\ln p^*(s, a, \lambda) + \int q(\psi|\lambda) \ln \frac{q(\psi|\lambda)}{p^*(\psi|s, a, \lambda)} \, \mathrm{d}\psi \tag{27}$$

$$= -\ln p^*(s, a, \lambda) + \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s, a, \lambda)], \tag{28}$$

in terms of the "posterior density" $p^*(\Psi|s, a, \lambda)$ (here, we keep the conditioning argument λ, as in [1], and do not explicitly assume Condition 2, though our conclusions are unaffected by it), such that Equations (19) and (20) can be rewritten as:

$$f_a(s, a, \lambda) = -(\Gamma + R)_{aa} \cdot \nabla_a F(s, a, \lambda), \tag{29}$$

$$f_\lambda(s, a, \lambda) = -(\Gamma + R)_{\lambda\lambda} \cdot \nabla_\lambda F(s, a, \lambda). \tag{30}$$

It is worth considering what a proof of the free energy lemma could look like. A proof of the existence of a free energy (and therefore of the free energy lemma) would need to show that, for every system satisfying the given assumptions, there always exists a $q(\Psi|\lambda)$ such that the right-hand sides of Equations (29) and (30) are equal to the right-hand sides of Equations (19) and (20). Expanding Equations (29) and (30) using (28) leads to:

$$\begin{aligned} f_a(s,a,\lambda) &= (\Gamma+R)_{aa}\cdot\nabla_a \ln p^*(s,a,\lambda) \\ &\quad - (\Gamma+R)_{aa}\cdot\nabla_a \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)], \end{aligned} \tag{31}$$

$$\begin{aligned} f_\lambda(s,a,\lambda) &= (\Gamma+R)_{\lambda\lambda}\cdot\nabla_\lambda \ln p^*(s,a,\lambda) \\ &\quad - (\Gamma+R)_{\lambda\lambda}\cdot\nabla_\lambda \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)]. \end{aligned} \tag{32}$$

For the equality of the right-hand sides to those of Equations (19) and (20), we need:

$$(\Gamma+R)_{aa}\cdot\nabla_a \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0 \tag{33}$$

$$(\Gamma+R)_{\lambda\lambda}\cdot\nabla_\lambda \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0. \tag{34}$$

In other words, these equations say that the free energy lemma holds if any of the following three conditions (of strictly increasing strengths) are given:

1. There is a $q(\Psi|\lambda)$ such that the partial gradients ∇_a and ∇_λ of the KL divergence between the variational density and the conditional ergodic density are elements of the nullspaces of $(\Gamma+R)_{aa}$ and $(\Gamma+R)_{\lambda\lambda}$, respectively.
2. There is a $q(\Psi|\lambda)$ such that the gradients of the KL divergence to $p^*(\Psi|s,a,\lambda)$ are equal to the nullvector:

   $$\nabla_a \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0, \tag{35}$$

   $$\nabla_\lambda \mathrm{D_{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0, \tag{36}$$

   Then, they are always elements of the nullspaces of $(\Gamma+R)_{aa}$ and $(\Gamma+R)_{\lambda\lambda}$, respectively.
3. There is a $q(\Psi|\lambda)$ such that $q(\Psi|\lambda) = p^*(\Psi|s,a,\lambda)$ (and hence, $p^*(\Psi|s,a,\lambda) = p^*(\Psi|\lambda)$), which implies that the KL divergence to $p^*(\Psi|s,a,\lambda)$ vanishes for all a,λ and the two partial gradients are always nullvectors and therefore elements of the according nullspaces.

The free energy lemma can then be proven by showing that one of these three cases follows from the conditions of the lemma. However, no attempt was made in [1] to establish this. Instead, the given proof discusses the purported consequences of the existence of a suitable $q(\Psi|\lambda)$. These will be discussed in Steps 5 and 6.

Even if the free energy lemma does not hold for systems obeying Equations (19) and (20), one might expect that the systems instead only satisfy the more general Equations (21) and (22) or the most general Equations (17) and (18). For these systems, the free energy lemma would require that there is a $q(\Psi|\lambda)$ such that:

$$f_a(s,a,\lambda) = ((\Gamma_{aa}+R_{aa})\cdot\nabla_a + R_{a\lambda}\cdot\nabla_\lambda)F(s,a,\lambda), \tag{37}$$

$$f_\lambda(s,a,\lambda) = (R_{\lambda a}\cdot\nabla_a + (\Gamma_{\lambda\lambda}+R_{\lambda\lambda})\cdot\nabla_\lambda)F(s,a,\lambda). \tag{38}$$

or:

$$f_a(s,a,\lambda) = (R_{as}\cdot\nabla_s + (\Gamma_{aa}+R_{aa})\cdot\nabla_a + R_{a\lambda}\cdot\nabla_\lambda)F(s,a,\lambda), \tag{39}$$

$$f_\lambda(s,a,\lambda) = (R_{\lambda s}\cdot\nabla_s + R_{\lambda a}\cdot\nabla_a + (\Gamma_{\lambda\lambda}+R_{\lambda\lambda})\cdot\nabla_\lambda)F(s,a,\lambda), \tag{40}$$

hold, respectively. However, we find this not to be the case in general.

Observation 4. *Given a random dynamical system obeying Equation (1), ergodicity, Conditions 1 and 2, there need not exist a free energy expressed in terms of a variational density $q(\Psi|\lambda)$ such that:*

(i) *Equations (29) and (30) hold if Equations (19) and (20) do;*

(ii) *Equations (37) and (38) hold if Equations (19) and (20) do not hold, but Equations (21) and (22) do;*

(iii) *Equations (39) and (40) hold if neither Equations (19) and (20) nor Equations (21) and (22) hold, but Equations (17) and (18) do.*

Proof. In Appendix C, we derive a set of conditions on the R and U matrices and on the putative variational density $q(\Psi|\lambda)$, which follow from each of the pairs of equations in Cases (i–iii). We show that, in general, each pair leads to a contradiction, and in each case, we provide a counterexample that falls into the according system class. □

Before proceeding, we note that later works presented an alternative version of the free energy lemma, where the conditioning argument of $q(\Psi|\lambda)$ was replaced by the most likely value of λ conditional on the (s, a) coordinates [2,3]. We here concern ourselves with the version apparent in [1], where $q(\Psi|\lambda)$ is parametrised by the internal states themselves, but we briefly comment on the interpretation of the alternative approach in Step 7.

5. Vanishing Gradients

As mentioned in Step 4, the proof of the free energy lemma in [1] only discussed its consequences. The first proposed consequence is that expressing the vector field in terms of a free energy as in Equations (29) and (30) "requires" that the gradients with respect to a and λ of the KL divergence vanish, i.e., that Equations (35) and (36) hold.

We mentioned in Step 4 that the implication in the opposite direction holds. This can be seen from Equations (33) and (34). However, if the nullspace of $(\Gamma + R)_{aa}$ or $(\Gamma + R)_{\lambda\lambda}$ is non-trivial, then the gradient may be a non-zero element of this subspace and Equations (29) and (30) will still hold. In that case, the vanishing gradients would not be necessary for the free energy lemma.

The conditions under which a non-trivial nullspace exists were discussed in [7]. In short, the nullspace is guaranteed to be trivial in the special case where Γ is positive definite. Whether or not ergodic systems with a Markov blanket can ever admit a non-trivial nullspace, and hence divergences in Equations (31) and (32) with non-vanishing gradients, is not immediately clear. However, in order to establish the necessity of Equations (35) and (36), this remains to be proven.

6. Equality of $Q(\Psi|\lambda)$ and $P^*(\Psi|s, a, \lambda)$

The proof of the free energy lemma in [1] also proposes that the vanishing of the gradients of the KL divergence, of the variational density $q(\Psi|\lambda)$ from the conditional ergodic density $p^*(\Psi|s, a, \lambda)$, implies the equality of these densities. We mentioned in Equations (5) that the implication in the opposite direction holds. This can also be seen from Equations (33) and (34). Concerning the implication in the direction proposed by [1], let us now assume that for a given system of Equations (19) and (20) holds, a variational density $q(\Psi|\lambda)$ does exist, and the gradients of the KL divergence of the variational and ergodic densities vanish, i.e., Equations (35) and (36) hold. Then, consider the argument by [1] in this direct quote (comments in square brackets by us):

> "However, Equation (2.6) [Equations (19) and (20) above] requires the gradients of the divergence to be zero [Equations (35) and (36)], which means the divergence must be minimized with respect to internal states. This means that the variational and posterior densities must be equal:
>
> $$q(\psi|\lambda) = p^{[*]}(\psi|s, a, \lambda) \Rightarrow \mathrm{D}_{\mathrm{KL}} = 0 \Rightarrow \begin{cases} (\Gamma + R) \cdot \nabla_\lambda \mathrm{D}_{\mathrm{KL}} = 0, \\ (\Gamma + R) \cdot \nabla_a \mathrm{D}_{\mathrm{KL}} = 0. \end{cases}$$
>
> In other words, the flow of internal and active states minimizes free energy, rendering the variational density equivalent to the posterior density over external states."

The first problem in the above quote is that the minimization of the divergence does not follow from the vanishing gradients. On the contrary, since Equations (35) and (36) must hold for all (s, a, λ), the KL divergence:

$$\mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s, a, \lambda)]$$

cannot depend on (λ, a); it therefore has no extremum (and thus no minimum) with respect to either of these coordinates.

The second problem pertains to the identification of the two distributions at a minimum. In general, if we try to find the minimum of a KL divergence between a given probability density $p_1(Y)$ and a family of densities $p_2(Y|\theta)$ parameterized by θ, then the lowest possible value of zero is achieved only if there is a parameter θ_1 such that $p_2(Y|\theta_1) = p_1(Y)$. If there is no such θ_1, then the minimum value will be larger than zero. Therefore, even if the divergence were minimized, it would not need to be zero. More generally, the divergence $K(s)$ need not be zero for any value of s.

There is therefore no satisfactory reason given why the variational density $q(\Psi|\lambda)$ and the posterior density $p^*(\Psi|s, a, \lambda)$ should be equal or have low KL divergence. In fact, they need not be (Note that, since any $q(\Psi|\lambda)$ that does not depend on (s, a) is an element of the set of those that do, Observation 5 remains true for the case where we allow this dependence. In that case, the free energy lemma holds because we can set $q(\Psi|s, a, \lambda) := p^*(\Psi|s, a, \lambda)$, and thus, a q exists for which the densities are actually equal. However, the claim here is that for every q that obeys the conditions in Observation 5, we must have equality.).

Observation 5. *Given a random dynamical system obeying Equation (1), ergodicity, Conditions 1 and 2. Then if, additionally,*

(i) *Equations (19) and (20) hold and the free energy lemma holds, i.e., there exists a probability density $q(\Psi|\lambda)$ such that Equations (29) and (30) hold, or*

(ii) *Equations (21) and (22) hold and there exists $q(\Psi|\lambda)$ such that Equations (37) and (38) hold, or*

(iii) *Equations (17) and (18) hold and there exists $q(\Psi|\lambda)$ such that Equations (39) and (40) hold,*

then there is no $c \geq 0$ for which it can be guaranteed that:

$$\mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s, a, \lambda)] < c. \tag{41}$$

In particular, it does not follow from these conditions that:

$$q(\Psi|\lambda) = p^*(\Psi|s, a, \lambda). \tag{42}$$

Proof. By example, see Appendix D. To show that the implication does not generally hold for a given system and densities $q(\Psi|\lambda)$ that obey Equations (19), (20), (29), and (30), Equations (21), (22), (37), and (38), or Equations (17), (18), (39), and (40), we only have to consider a system that obeys all three pairs of equations, Equations (19) and (20), Equations (21) and (22), and Equations (21) and (22), and for which a suitable $q(\Psi|\lambda)$ exist. For this system, we then need to show that the $q(\Psi|\lambda)$ that obey Equations (29) and (30) are not necessarily equal (or similar) to $p^*(\Psi|s, a, \lambda)$.

We use a variant of the model used in Appendix B as such a counterexample. This system obeys all three of Equations (19) and (20), Equations (21) and (22), and Equations (21) and (22), and the nullspace of the associated $\Gamma + R$ is trivial. We identify a set of possible $q(\Psi|\lambda)$ satisfying Equations (29) and (30), which implies that the gradients of the KL divergence between those $q(\Psi|\lambda)$ and $p^*(\Psi|s, a, \lambda)$ vanish, i.e., Equations (35) and (36) hold. We then demonstrate that for the $q(\Psi|\lambda)$ in this set, the value of the KL divergence to $p^*(\Psi|s, a, \lambda)$ can be arbitrarily large. □

7. Interpretation

Finally, we turn our attention to the interpretation in terms of Bayesian inference, i.e., Step 7. We again quote directly from [1]:

> Because (by Gibbs inequality) this divergence [$D_{KL}[q(\psi|\lambda)||p^*(\psi|s,a,\lambda)]$] cannot be less than zero, the internal flow will appear to have minimized the divergence between the variational and posterior density. In other words, the internal states will appear to have solved the problem of Bayesian inference by encoding posterior beliefs about hidden (external) states, under a generative model provided by the Gibbs energy.

We showed that, in general, there is no suitable variational density that is only parameterized by the internal coordinate λ. We then showed that, even if there is a suitable variational density (including those parameterized by all of (s,a,λ)), it can be arbitrarily different from the posterior density. Since the arguments for the internal flow appearing to minimize the divergence between variational and posterior density are therefore incorrect, there is no reason why the internal states should appear to have solved the problem of Bayesian inference.

As mentioned in Step 4, some newer works (e.g., [2,3]) formulated a different free energy principle, where the variational density of beliefs is parametrised not by the internal coordinates λ, but by $\bar{\lambda}(s,a) = \arg\max_\lambda p^*(\lambda|s,a)$, the most likely value of the internal coordinates given the sensory and active ones. In this case, Observations 4 and 5 do not apply. However, the new parameters $\bar{\lambda}(s,a)$ are strictly a function of the sensory and active coordinates. This means we have a Markov chain (with capitalisations indicating random variables associated with the corresponding lower case coordinates (or functions of coordinates)) $\Lambda \to (S,A) \to \bar{\Lambda}$ and, by the data processing inequality [14], the mutual information between the both sensory and active coordinates and the belief parameter $\bar{\lambda}$ upper bounds that are between the internal coordinates and the belief parameter. It is therefore not clear to what extent the internal coordinates λ, rather than the active and sensory coordinates (s,a) themselves, can be said to be encoding beliefs about the external coordinates. Note also that, on any given trajectory, unless the distribution $p^*(\lambda|s,a)$ is sufficiently peaked and unimodal, the internal coordinates are not guaranteed to spend most of their time close to their most likely conditional value, and (by definition if Condition 2 holds) they will not be better predictors of the external coordinates than those in the Markov blanket.

Generally, $\lambda \neq \bar{\lambda}$, and $\bar{\lambda}$ is the solution to an optimization problem that is assumed to be solved in these later works. Using this optimized variable to parametrise beliefs is therefore a considerable departure from [1]. Contrary to the impression created by the way it was referenced in [2,3], the older theory in [1] should be clearly distinguished from the newer ones in these more recent papers.

8. Consequences for Friston, K. et al. 2014

Reference [4] argued for the same interpretation as [1], but there were some differences in the argument.

The differences were the following:

- In [4], Equation (1) was formulated for "generalized states", which we refer to here as generalized coordinates. This means that the variable x is replaced by a multidimensional variable denoted $\tilde{x} = (x, x', x'', ...)$.
- The Markov blanket structure was not explicitly defined via Equation (2). Formally, it was introduced directly (see [4] Equation (10)) in a less general form corresponding to Equations (19) and (20) (at the same time, [1] is referenced in connection to the Markov blanket so there seems to be no intention to replace the original definition with the stronger one). Therefore, our observations concerning Steps 2 to 4 are not directly relevant to this paper.

- The internal coordinate λ was renamed to r, and the role of matrix R was played by the matrix $-Q$.
- The proof of the free energy lemma given in [4] was different. It (implicitly) suggested setting the variational density equal to the ergodic conditional posterior.
- The proof of the free energy lemma no longer contained the proposition that the gradient of the KL divergence of the variational density and the ergodic conditional density vanish, i.e., Step 5.
- The proof also no longer contained the claim that the vanishing gradients of the KL divergence of the variational density and the ergodic conditional density imply the equality of those densities, i.e., Step 6 was not present.

The interpretation in terms of Bayesian inference was unchanged and still relied on the equality of the variational and the ergodic conditional density.

Since there were no explicit generalized coordinate versions of Steps 2, 3, 5 and 6 in [4], we do not discuss those steps here. We only disprove the free energy lemma and the claim that when the free energy lemma holds, the variational and ergodic conditional density become equal. For this, we present a way to translate the counterexamples used in Observations 4 and 5 into counterexamples in generalized coordinates. The interpretation in terms of Bayesian inference given in [4] is therefore equally as unjustified as the one in [1].

For completeness, we first state the generalized coordinate versions of the stochastic differential Equation (1):

$$\dot{\tilde{x}} = f(\tilde{x}) + \tilde{\omega}, \tag{43}$$

the less general version of the Markov blanket structure Equation (2):

$$\begin{aligned} f_{\tilde{\psi}}(\tilde{\psi}, \tilde{s}, \tilde{a}) &= (\Gamma - Q)_{\tilde{\psi}\tilde{\psi}} \nabla_{\tilde{\psi}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ f_{\tilde{s}}(\tilde{\psi}, \tilde{s}, \tilde{a}) &= (\Gamma - Q)_{\tilde{s}\tilde{s}} \nabla_{\tilde{s}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ f_{\tilde{a}}(\tilde{s}, \tilde{a}, \tilde{r}) &= (\Gamma - Q)_{\tilde{a}\tilde{a}} \nabla_{\tilde{a}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ f_{\tilde{r}}(\tilde{s}, \tilde{a}, \tilde{r}) &= (\Gamma - Q)_{\tilde{r}\tilde{r}} \nabla_{\tilde{r}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}), \end{aligned} \tag{44}$$

the expression of the $\tilde{a}$ and $\tilde{r}$ components of the vector field in terms of the marginalised ergodic density Equations (19) and (20):

$$f_{\tilde{a}}(\tilde{s}, \tilde{a}, \tilde{r}) = (\Gamma - Q)_{\tilde{a}\tilde{a}} \cdot \nabla_{\tilde{a}} \ln p^*(\tilde{s}, \tilde{a}, \tilde{r}), \tag{45}$$

$$f_{\tilde{r}}(\tilde{s}, \tilde{a}, \tilde{r}) = (\Gamma - Q)_{\tilde{r}\tilde{r}} \cdot \nabla_{\tilde{r}} \ln p^*(\tilde{s}, \tilde{a}, \tilde{r}), \tag{46}$$

and in terms of free energy Equations (29) and (30):

$$f_{\tilde{a}}(\tilde{s}, \tilde{a}, \tilde{r}) = (Q - \Gamma)_{\tilde{a}\tilde{a}} \cdot \nabla_{\tilde{a}} F(\tilde{s}, \tilde{a}, \tilde{r}), \tag{47}$$

$$f_{\tilde{r}}(\tilde{s}, \tilde{a}, \tilde{r}) = (Q - \Gamma)_{\tilde{r}\tilde{r}} \cdot \nabla_{\tilde{r}} F(\tilde{s}, \tilde{a}, \tilde{r}). \tag{48}$$

The free energy lemma then requires that there exists $q(\tilde{\Psi}|\tilde{r})$ such that the KL divergence between $p^*(\tilde{\Psi}|\tilde{s}, \tilde{a}, \tilde{r})$ vanishes. Without going into further details of the difference between the proof in [4] and that in [1], we can prove the former wrong by translating the counterexample used for the latter into generalised coordinates.

Observation 6. *There is a general way to translate a system in ordinary coordinates into a system of generalised coordinates that corresponds to an infinite number of independent copies of the original system. This means all properties of the original system (e.g., linearity, ergodicity, the Gaussian and Markovian property of the noise, Conditions 1 and 2, the properties of* Γ, R, U*) are preserved during this translation.*

Proof. By construction, see Appendix E. □

This implies that the counterexamples used in proving Observations 4 and 5 directly translate to the setting of the generalised coordinates. The free energy lemma is therefore also wrong for generalised coordinates, and the variational density $q(\tilde{\Psi}|\tilde{r})$ is not "ensured" [4] to be equal to the conditional ergodic density $p^*(\tilde{\Psi}|\tilde{s},\tilde{a},\tilde{r})$.

9. Conclusions

We find that the two different Markov blanket conditions proposed in [1–3] are independent of each other. We then show that under both of those Markov blanket conditions, among the six steps contained in the argument in [1], three do not hold independently of each other. We also show that fixing the second of those steps (Step 3) does not provide a valid alternative. The line of reasoning of [1] therefore does not support its claim that the internal coordinates of a Markov blanket "appear to have solved the problem of Bayesian inference by encoding posterior beliefs about hidden (external) [coordinates], ...". We also show that using generalised coordinates as in [4] does not remedy the situation. Additionally, we identify a technical error in [2] and an interpretational issue resulting from possibly too strong assumptions (both Conditions 1 and 3) in [3]. We also highlight that the latter publications both argued that it is the most likely internal coordinates given sensory and active coordinates that encode posterior beliefs about external states instead of the internal coordinates themselves. The resulting free energy principle and lemma are therefore a different proposal. This is not subject to our technical critique.

Author Contributions: Conceptualization, M.B., F.A.P., and R.K.; formal analysis, M.B. and F.A.P.; funding acquisition, F.A.P. and R.K.; methodology, M.B., F.A.P., and R.K.; visualization, M.B. and F.A.P.; writing—original draft, M.B. and F.A.P.; writing—review and editing, M.B., F.A.P., and R.K. All authors read and agreed to the published version of the manuscript.

Funding: The work by Martin Biehl and Ryota Kanai on this publication was made possible through the support of a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation, Inc. Martin Biehl and Ryota Kanai are also funded by the Japan Science and Technology Agency (JST) CREST project. Felix A. Pollock acknowledges support from the Monash University Network of Excellence for Consciousness and Complexity in the Conscious Brain.

Acknowledgments: All authors are grateful to Karl Friston and Thomas Parr for constructive feedback on an earlier version of this work. We also want to thank Danijar Hafner for pointing us to [9]. Martin Biehl wants to thank Yen Yu for helpful discussions on generalized coordinates.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Appendix A. Counterexamples for Observation 1

Consider a four-dimensional linear system obeying Equation (1) for which there are coordinates $x = (\psi, s, a, \lambda)$ with $n_\psi = n_s = n_a = n_\lambda = 1$ and:

$$f(x) = Mx, \tag{A1}$$

with the parametrisation:

$$M = \begin{pmatrix} -1 & m_1 & m_2 & m_3 \\ m_2 & -1 & m_2 & m_3 \\ m_3 & m_2 & -1 & m_2 \\ m_3 & m_2 & m_1 & -1 \end{pmatrix}. \tag{A2}$$

From Equation (9), it is clear that the system obeys Condition 1 if $m_3 = 0$. In this case, taking Γ to be the identity matrix, it is possible to show that:

$$U_{\psi\lambda} = -\frac{m_2(m_1 - m_2 + 2)(m_2^3 + m_1^2 m_2 - 2m_1 m_2^2 - 4m_1 - 2)}{(m_1^2 + m_2^2 - 4m_2 + 4)(m_1^2 + 5m_2^2 - 4m_1 m_2 + 4m_2 + 4)}. \tag{A3}$$

For fixed, finite m_2, this is zero only for a few discrete values of m_1, such as $m_1 = m_2 - 2$; that it is generically non-zero proves Equation (11). As a concrete example, the following:

$$M = \begin{pmatrix} -1 & -2/3 & -2/3 & 0 \\ -2/3 & -1 & -2/3 & 0 \\ 0 & -2/3 & -1 & -2/3 \\ 0 & -2/3 & -2/3 & -1 \end{pmatrix}, \tag{A4}$$

has:

$$R = \begin{pmatrix} 0 & -1/8 & 3/8 & 0 \\ 1/8 & 0 & 0 & -3/8 \\ -3/8 & 0 & 0 & 1/8 \\ 0 & 3/8 & -1/8 & 0 \end{pmatrix} \tag{A5}$$

and (full rank and hence ergodic):

$$U = \begin{pmatrix} 236/255 & 127/255 & -31/85 & -12/85 \\ 127/255 & 274/255 & 206/255 & 31/85 \\ 31/85 & 206/255 & 274/255 & 127/255 \\ -12/85 & 31/85 & 127/255 & 236/255 \end{pmatrix}, \tag{A6}$$

and hence ergodic density:

$$p^*(\psi, s, a, \lambda) = \sqrt{\frac{28}{2295\pi^4}} \exp\left[-\frac{1}{255}\Big(137(a^2 + s^2) + 118(\psi^2 + \lambda^2) + 127(\psi s + a\lambda) + 93(\psi a + s\lambda) + 206as - 36\psi\lambda\Big)\right], \tag{A7}$$

which does not conditionally factorise.

Taking the same parametrisation as in Equation (A2) and fixing $m_1 = m_2 = -1/2$, we can search for a non-zero value of m_3 that leads to $U_{\psi\lambda} = 0$ (equivalent to Condition 2 through Equation (10)). We find such a value in the real root $c \simeq -0.08$ of the quintic equation $8c^5 - 4c^4 - 6c^3 + 31c^2 + 40c + 3 = 0$; that is, with:

$$M = \begin{pmatrix} -1 & -1/2 & -1/2 & c \\ -1/2 & -1 & -1/2 & c \\ c & -1/2 & -1 & -1/2 \\ c & -1/2 & -1/2 & -1 \end{pmatrix}, \tag{A8}$$

which does not satisfy Condition 1, we have:

$$R = \begin{pmatrix} 0 & -0.06\ldots & 0.22\ldots & 0 \\ 0.06\ldots & 0 & 0 & -0.22\ldots \\ -0.22\ldots & 0 & 0 & 0.06\ldots \\ 0 & 0.22\ldots & -0.06\ldots & 0 \end{pmatrix}, \tag{A9}$$

and:

$$U = \begin{pmatrix} 0.96\ldots & 0.43\ldots & 0.30\ldots & 0 \\ 0.43\ldots & 1.03\ldots & 0.58\ldots & 0.30\ldots \\ 0.30\ldots & 0.58\ldots & 1.03\ldots & 0.43\ldots \\ 0 & 0.30\ldots & 0.43\ldots & 0.96\ldots \end{pmatrix}, \tag{A10}$$

which has a non-zero determinant (i.e., the dynamics is ergodic) and an ergodic density satisfying Condition 2. This proves Equation (12).

Appendix B. Counterexample for Step 3

Here, we consider a linear system, as in the previous Appendix. We again assume Γ equal to the identity matrix and choose a force matrix of the form:

$$M = \begin{pmatrix} -1 & -\frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & 0 \\ -\frac{1}{2\sqrt{2}} & -1 & -\frac{1}{16} & 0 \\ 0 & \frac{1}{16} & -1 & -\frac{1}{2\sqrt{2}} \\ 0 & \frac{1}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & -1 \end{pmatrix} \tag{A11}$$

which explicitly satisfies Condition 1 and has full rank such that the system is ergodic. Using Equation (5), this leads to:

$$U = \begin{pmatrix} \frac{1023}{1057} & \frac{260\sqrt{2}}{1057} & \frac{136\sqrt{2}}{1057} & 0 \\ \frac{260\sqrt{2}}{1057} & \frac{1091}{1057} & 0 & -\frac{136\sqrt{2}}{1057} \\ \frac{136\sqrt{2}}{1057} & 0 & \frac{1091}{1057} & \frac{260\sqrt{2}}{1057} \\ 0 & -\frac{136\sqrt{2}}{1057} & \frac{260\sqrt{2}}{1057} & \frac{1023}{1057} \end{pmatrix} \tag{A12}$$

which shows that this system also satisfies Condition 2 since $U_{\psi\lambda} = U_{\lambda\psi} = 0$. We also find:

$$R = \begin{pmatrix} 0 & -\frac{17}{1786\sqrt{2}} & \frac{479}{1786\sqrt{2}} & -\frac{62}{893} \\ \frac{17}{1786\sqrt{2}} & 0 & -\frac{65}{14288} & \frac{479}{1786\sqrt{2}} \\ -\frac{479}{1786\sqrt{2}} & \frac{65}{14288} & 0 & \frac{17}{1786\sqrt{2}} \\ \frac{62}{893} & -\frac{479}{1786\sqrt{2}} & -\frac{17}{1786\sqrt{2}} & 0 \end{pmatrix}, \tag{A13}$$

which shows that all entries or R that can be non-zero for an anti-symmetric matrix are non-zero. For the marginal ergodic density, we find:

$$p^*(s,a,\lambda) = \frac{239}{16\sqrt{2415}\pi^{3/2}} \exp\left[-\frac{69a^2}{140} - \frac{37as}{70\sqrt{2}} + \frac{1}{35}\sqrt{22}a\psi - \frac{4867s^2}{9660} + \frac{74s\psi}{2415} - \frac{8429\psi^2}{19320} \right] \tag{A14}$$

The difference between the right-hand sides of Equations (17) and (19) is:

$$R_{as}\nabla_s \ln p^*(s,a,\lambda) + R_{a\lambda}\nabla_\lambda \ln p^*(s,a,\lambda) = \frac{37a + 69\sqrt{2}\lambda - 2563s}{4830} \neq 0, \tag{A15}$$

which shows that Equation (19) is wrong in this example and therefore not generally equivalent to Equation (17). Similarly, computing the difference between the right-hand sides of Equations (18) and (20), one finds:

$$R_{\lambda s}\nabla_s \ln p^*(s,a,\lambda) + R_{\lambda a}\nabla_a \ln p^*(s,a,\lambda) = \frac{2a - \sqrt{2}\lambda + 27s}{70\sqrt{2}} \neq 0, \tag{A16}$$

and hence, Equation (20) is also incorrect in general.

Performing the same comparison for the difference between the general expression in Equations (17) and (18) and the expressions taken from [2], one finds:

$$R_{a_i s} \nabla_s \ln p^*(s,a,\lambda) = \frac{73\left(296a + 552\sqrt{2}\lambda - 8429s\right)}{1,154,370} \neq 0 \tag{A17}$$

for the difference between the right-hand sides of Equations (17) and (21), and:

$$R_{\lambda s} \nabla_s \ln p^*(s,a,\lambda) = -\frac{53\left(296a + 552\sqrt{2}\lambda - 8429s\right)}{1,154,370\sqrt{2}} \neq 0, \tag{A18}$$

for the difference between the right-hand sides of Equations (18) and (22). Therefore, Equations (21) and (22) are also incorrect in general, even when Conditions 1 and 2 both hold.

Appendix C. Counterexamples for Step 4

We saw in Appendix B that Equations (19) and (20) are not generally equivalent to Equation (4), even when Conditions 1 and 2 hold simultaneously. We now show that if we instead use Equations (17) and (18), which are generally equivalent to Equation (4), the free energy lemma does not hold in general.

The original free energy lemma requires that (see Equations (31) and (32)):

$$(\Gamma + R)_{aa} \cdot \nabla_a \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0 \tag{A19}$$

$$(\Gamma + R)_{\lambda\lambda} \cdot \nabla_\lambda \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0. \tag{A20}$$

Replacing the partial gradient in Equations (29) and (30) with the full gradient and including the entire matrix $(\Gamma + R)$ lead to the corresponding requirement for the more general case:

$$(R_{as} \cdot \nabla_s + (\Gamma_{aa} + R_{aa}) \cdot \nabla_a + R_{a\lambda} \cdot \nabla_\lambda) \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0 \tag{A21}$$

$$(R_{\lambda s} \cdot \nabla_s + R_{\lambda a} \cdot \nabla_a + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda) \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0. \tag{A22}$$

Similarly, the version based on the equations taken from [2] implies:

$$((\Gamma_{aa} + R_{aa}) \cdot \nabla_a + R_{a\lambda} \cdot \nabla_\lambda) \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0 \tag{A23}$$

$$(R_{\lambda a} \cdot \nabla_a + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \cdot \nabla_\lambda) \, \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\Psi|s,a,\lambda)] = 0. \tag{A24}$$

Using the rules of Gaussian integration, we can write the logarithm of the conditional ergodic density as:

$$\ln p^*(\psi|s,a,\lambda) = -\frac{1}{2}|U_{\psi\psi}^{\frac{1}{2}}\psi + U_{\psi\psi}^{-\frac{1}{2}}U_{\psi s}s + U_{\psi\psi}^{-\frac{1}{2}}U_{\psi a}a + U_{\psi\psi}^{-\frac{1}{2}}U_{\psi\lambda}\lambda|^2 + C, \tag{A25}$$

with C a constant (and remembering each of ψ, s, a, and λ is a vector of coordinates in general). We can then expand the derivatives of the KL divergence to express them in terms of the coordinates:

$$\begin{aligned}
\nabla_s \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\psi|s,a,\lambda)] &= -\int \mathrm{d}\psi\, q(\psi|\lambda)\nabla_s \ln p^*(\psi|s,a,\lambda) \\
&= U_{s\psi}U_{\psi\psi}^{-1}\big(U_{\psi s}s + U_{\psi a}a \\
&\qquad + U_{\psi\lambda}\lambda + U_{\psi\psi}\langle\psi\rangle_{q(\Psi|\lambda)}\big), \qquad \text{(A26)} \\
\nabla_a \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\psi|s,a,\lambda)] &= -\int \mathrm{d}\psi\, q(\psi|\lambda)\nabla_a \ln p^*(\psi|s,a,\lambda) \\
&= U_{a\psi}U_{\psi\psi}^{-1}\big(U_{\psi s}s + U_{\psi a}a \\
&\qquad + U_{\psi\lambda}\lambda + U_{\psi\psi}\langle\psi\rangle_{q(\Psi|\lambda)}\big), \qquad \text{(A27)} \\
\nabla_\lambda \mathrm{D}_{\mathrm{KL}}[q(\Psi|\lambda)||p^*(\psi|s,a,\lambda)] &= \int \mathrm{d}\psi \Big[(\ln q(\psi|\lambda) - \ln p^*(\psi|s,a,\lambda) + 1)\nabla_\lambda q(\psi|\lambda) \\
&\qquad - q(\psi|\lambda)\nabla_\lambda \ln p^*(\psi|s,a,\lambda)\Big] \\
&= U_{\lambda\psi}U_{\psi\psi}^{-1}\big(U_{\psi s}s + U_{\psi a}a \\
&\qquad + U_{\psi\lambda}\lambda + U_{\psi\psi}\langle\psi\rangle_{q(\Psi|\lambda)}\big) \\
&\quad + \nabla_\lambda\langle\psi\rangle_{q(\Psi|\lambda)}\big(U_{\psi s}s + U_{\psi a}a + U_{\psi\lambda}\lambda\big) \\
&\quad + \nabla_\lambda\Big(\langle\psi^T U_{\psi\psi}\psi\rangle_{q(\Psi|\lambda)} - H[q(\Psi|\lambda)]\Big), \qquad \text{(A28)}
\end{aligned}$$

with $\langle g(\psi)\rangle_{q(\Psi|\lambda)} := \int \mathrm{d}\psi\, q(\psi|\lambda)g(\psi)$ and H the Shannon entropy.

Substituting Equations (A27) and (A28) into Equations (A19) and (A20) leads to:

$$(\Gamma_{aa} + R_{aa})U_{a\psi}U_{\psi\psi}^{-1}\Big(U_{\psi s}s + U_{\psi a}a + U_{\psi\lambda}\lambda + U_{\psi\psi}\langle\psi\rangle_{q(\Psi|\lambda)}\Big) = 0, \qquad \text{(A29)}$$

and:

$$\begin{aligned}
0 =& (\Gamma_{\lambda\lambda} + R_{\lambda\lambda})U_{\lambda\psi}U_{\psi\psi}^{-1}\big(U_{\psi s}s + U_{\psi a}a + U_{\psi\lambda}\lambda + U_{\psi\psi}\langle\psi\rangle_{q(\Psi|\lambda)}\big) \\
&+ (\Gamma_{\lambda\lambda} + R_{\lambda\lambda})\nabla_\lambda\langle\psi\rangle_{q(\Psi|\lambda)}\big(U_{\psi s}s + U_{\psi a}a + U_{\psi\lambda}\lambda\big) \\
&+ (\Gamma_{\lambda\lambda} + R_{\lambda\lambda})\nabla_\lambda\Big(\langle\psi^T U_{\psi\psi}\psi\rangle_{q(\Psi|\lambda)} - H[q(\Psi|\lambda)]\Big). \qquad \text{(A30)}
\end{aligned}$$

Since these must hold for all values of the coordinates, they put strong requirements on the U and R matrices. Specifically,

$$(\Gamma_{aa} + R_{aa})U_{a\psi}U_{\psi\psi}^{-1}U_{\psi s} = 0, \qquad \text{(A31)}$$

$$(\Gamma_{aa} + R_{aa})U_{a\psi}U_{\psi\psi}^{-1}U_{\psi a} = 0, \qquad \text{(A32)}$$

$$(\Gamma_{aa} + R_{aa})U_{a\psi}U_{\psi\psi}^{-1}U_{\psi\lambda} = 0. \qquad \text{(A33)}$$

In other words, since $U_{\psi\psi}$ and Γ_{aa} must be nonzero for the dynamics to be ergodic, it must be that $U_{\psi a} = 0$ (This is equivalent to $p^*(\Psi|s,a,\lambda) = p^*(\Psi|s,\lambda)$. Therefore, if Condition 2 also holds, we must have $p^*(\Psi|s,a,\lambda) = p^*(\Psi|s)$ in order for there to be a suitable $q(\Psi|\lambda)$.). Specifically, consider the system specified by the force matrix:

$$M = \begin{pmatrix} -1 & 0 & \frac{1}{2} & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \qquad \text{(A34)}$$

which leads to:

$$R = \begin{pmatrix} 0 & 0 & -\frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{A35}$$

and:

$$U = \begin{pmatrix} \frac{16}{17} & 0 & -\frac{4}{17} & 0 \\ 0 & 1 & 0 & 0 \\ -\frac{4}{17} & 0 & \frac{18}{17} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{A36}$$

Here, M is full rank, so the system is ergodic; clearly, it also satisfies Condition 1 due to the structure of M. Since $R_{as} = R_{a\lambda} = R_{\lambda s} = 0$, it obeys Equations (19) and (20), and since $Q_{\psi\lambda} = 0$, it also obeys Condition 2. Additionally, we find $U_{\psi a} = -4/17$, which is a contradiction.

For the more general version, substituting Equations (A26)–(A28) into Equation (A21), one finds:

$$\begin{aligned} 0 = &\Big(\big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)U_{\psi\psi}^{-1} + R_{a\lambda}\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi s} s \\ &+ \Big(\big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)U_{\psi\psi}^{-1} + R_{a\lambda}\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi a} a \\ &+ \Big(\big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)U_{\psi\psi}^{-1} + R_{a\lambda}\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi\lambda} \lambda \\ &+ \big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)\langle \psi \rangle_{q(\Psi|\lambda)} \\ &+ R_{a\lambda}\nabla_\lambda \Big(\langle \psi^T U_{\psi\psi} \psi \rangle_{q(\Psi|\lambda)} - H[q(\Psi|\lambda)] \Big), \end{aligned} \tag{A37}$$

which, considering that the coordinates can take any values, implies that:

$$\big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)U_{\psi\psi}^{-1} + R_{a\lambda}\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \tag{A38}$$

lies in a common (left) nullspace of $U_{\psi s}$, $U_{\psi a}$, and $U_{\psi\lambda}$. However, the existence of such a nontrivial nullspace would imply that the corresponding subspace of ψ coordinates is independent of the s, a, and λ coordinates (to see this, consider marginalising over their complement in Equation (A25)). In other words, if only ψ coordinates that play a nontrivial role in the dynamics are considered, then Equation (A21) must imply that the quantity in Equation (A38) is zero and hence that:

$$R_{a\lambda}\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = -\big(R_{as}U_{s\psi} + (R_{aa} + \Gamma_{aa})U_{a\psi} + R_{a\lambda}U_{\lambda\psi}\big)U_{\psi\psi}^{-1}. \tag{A39}$$

However, through a similar procedure, one finds that Equation (A22) is equivalent to:

$$\begin{aligned}
0 = &\Big(\big(R_{\lambda s} U_{s\psi} + R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) U_{\psi\psi}^{-1} \\
&\quad + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi s} s \\
&+ \Big(\big(R_{\lambda s} U_{s\psi} + R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) U_{\psi\psi}^{-1} \\
&\quad + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi a} a \\
&+ \Big(\big(R_{\lambda s} U_{s\psi} + R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) U_{\psi\psi}^{-1} \\
&\quad + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} \Big) U_{\psi\lambda} \lambda \\
&+ \big(R_{\lambda s} U_{s\psi} + R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) \langle \psi \rangle_{q(\Psi|\lambda)} \\
&+ (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \nabla_\lambda \Big(\langle \psi^T U_{\psi\psi} \psi \rangle_{q(\Psi|\lambda)} - H[q(\Psi|\lambda)] \Big),
\end{aligned} \tag{A40}$$

implying that:

$$(\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = -\big(R_{\lambda s} U_{s\psi} + R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) U_{\psi\psi}^{-1}. \tag{A41}$$

Unless $R_{a\lambda}$ and $(\Gamma_{\lambda\lambda} + R_{\lambda\lambda})$ share a common nullspace or the U and R matrices are finely tuned, then Equations (A39) and (A41) contradict one another. In this case, there cannot exist a $q(\Psi|\lambda)$ that satisfies both Equations (A21) and (A22), and hence, the modified free energy lemma is invalid in general. In particular, using the example from Appendix B, if we solve Equation (A39) for $\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)}$, we find:

$$\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = -\frac{53}{5}, \tag{A42}$$

and from Equation (A41), we get:

$$\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = \frac{29}{239}, \tag{A43}$$

which is a contradiction.

If we now perform the same procedure for Equations (A23) and (A24), we arrive at the following conditions on the gradient of the variational density:

$$R_{a\lambda} \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = -\big((R_{aa} + \Gamma_{aa}) U_{a\psi} + R_{a\lambda} U_{\lambda\psi}\big) U_{\psi\psi}^{-1}. \tag{A44}$$

and:

$$R_{a\lambda} \nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = -\big(R_{\lambda a} U_{a\psi} + (\Gamma_{\lambda\lambda} + R_{\lambda\lambda}) U_{\lambda\psi}\big) U_{\psi\psi}^{-1}. \tag{A45}$$

Even when Condition 2 holds and $U_{\psi\lambda} = 0$, these will be inconsistent in general. As a specific counterexample, take the system with force matrix:

$$M = \begin{pmatrix} -1 & 0 & -\frac{1}{2} & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & \frac{\sqrt{3}}{2} \\ 0 & 0 & 0 & -\frac{1}{2} \end{pmatrix}, \tag{A46}$$

with corresponding:

$$R = \begin{pmatrix} 0 & 0 & \frac{1}{3} & \frac{1}{3\sqrt{3}} \\ 0 & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & -\frac{1}{\sqrt{3}} \\ -\frac{1}{3\sqrt{3}} & 0 & \frac{1}{\sqrt{3}} & 0 \end{pmatrix}, \tag{A47}$$

and:

$$U = \begin{pmatrix} \frac{9}{10} & 0 & \frac{3}{10} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{3}{10} & 0 & \frac{17}{20} & -\frac{\sqrt{3}}{4} \\ 0 & 0 & -\frac{\sqrt{3}}{4} & \frac{3}{4} \end{pmatrix}. \tag{A48}$$

This model is ergodic (full rank U), and it satisfies both Conditions 1 and 2. Moreover, the forces satisfy Equations (21) and (22). However, substituting the relevant elements of the U and R matrices into Equation (A44), we find:

$$\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = \frac{1}{\sqrt{3}}, \tag{A49}$$

but doing the same for Equation (A45) gives:

$$\nabla_\lambda \langle \psi \rangle_{q(\Psi|\lambda)} = \frac{1}{3}, \tag{A50}$$

which is a contradiction.

Appendix D. Counterexample for Step 6

Here, we provide an example system for which Conditions 1 and 2, as well as Steps 2 to 5 are valid, but Step 6 fails. We use a system with:

$$f(x) = Mx \tag{A51}$$

where:

$$M := \begin{pmatrix} -1 & 1/2 & 0 & 0 \\ 1/2 & -1 & 1/2 & 0 \\ 0 & 1/2 & -1 & 1/2 \\ 0 & 0 & 1/2 & -1 \end{pmatrix}. \tag{A52}$$

This system is ergodic, satisfies Condition 1, and as we will will see, satisfies Equations (19) and (20) as well. Using Equation (5), we find:

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{A53}$$

and from Equation (8):

$$U = -M \tag{A54}$$

which means that Condition 2 is also satisfied.

This leads to the ergodic density:

$$p^*(\psi, s, a, \lambda) = \frac{\sqrt{5}}{16\pi^2} e^{-\frac{1}{2}\left(\psi^2 - \psi s + s^2 - sa + a^2 - a\lambda + \lambda^2\right)} \tag{A55}$$

which can be used to check that Equations (19) and (20) hold for this example. The conditional ergodic density is:

$$p^*(\psi|s, a, \lambda) = p^*(\psi|s) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\psi - \frac{1}{2}s\right)^2}. \tag{A56}$$

If we now define $q(\psi|\lambda) = q(\psi) = \exp(-(\psi - \mu)^2/2)/\sqrt{2\pi}$ as a Gaussian distribution with mean μ and variance one, we can compute the KL divergence to get:

$$\mathrm{D}_{\mathrm{KL}}[q(\Psi)||p^*(\Psi|s, a, \lambda)] = K(s) = \frac{1}{2}\left(\mu - \frac{1}{2}s\right)^2. \tag{A57}$$

Clearly, for this choice of $q(\psi|\lambda)$, the gradients with respect to a and λ of the KL divergence vanish everywhere (Equations (35) and (36) hold). This also means we can express f_a, f_λ in terms of a free energy, i.e., the free energy lemma holds for this system. However, for any proposed bound $c \geq 0$ on the KL divergence, there is a value of s for which it is exceeded, whatever the choice of μ. Moreover, we can choose a μ such that the KL divergence is larger than any given c, even when $s = 0$.

Appendix E. Translating Systems into Generalized Coordinate Systems

We show how to get a generalized coordinate system from a finite-dimensional system. By definition, the generalized coordinates are infinite-dimensional. For all $n \in \mathbb{N}$ and a coordinate x, they also include the n-th time derivative of x.

Assume as given an ergodic, linear, random dynamical system described by:

$$\dot{x} = Mx + \omega \tag{A58}$$

where $x = (x_1, ..., x_k)$ is a k-dimensional vector, M is a $k \times k$ real-valued matrix, and $\dot{x} := \frac{d}{dt}x$. We can look at the second time derivative of the state by differentiating both sides:

$$\frac{d}{dt}\dot{x} = \frac{d}{dt}(Mx + \omega) \tag{A59}$$

$$\ddot{x} = M\dot{x} + \dot{\omega} \tag{A60}$$

Similarly for the third time derivative:

$$\frac{d}{dt}\ddot{x} = \frac{d}{dt}(M\dot{x} + \dot{\omega}) \tag{A61}$$

$$= M\ddot{x} + \ddot{\omega} \tag{A62}$$

Similarly for all higher derivatives:

$$\frac{d^n}{dt^n}x = M\frac{d^{n-1}}{dt^{n-1}}x + \frac{d^n}{dt^n}\omega. \tag{A63}$$

Now, define the generalized coordinates $\tilde{x} = (x, x', x'', ...)$ as:

$$x = x \tag{A64}$$

$$x' = \frac{d}{dt}x \tag{A65}$$

$$x'' = \frac{d^2}{dt^2}x \tag{A66}$$

$$\vdots \tag{A67}$$

$$x^{(n)} = \frac{d^n}{dt^n}x \tag{A68}$$

$$\vdots \tag{A69}$$

Define also:

$$\tilde{\omega} := (\omega, \frac{d}{dt}\omega, \frac{d^2}{dt^2}\omega, ..., \frac{d^n}{dt^n}\omega, ...). \tag{A70}$$

Without further clarification, the derivatives of ω are not well defined when the latter is a Gaussian white noise process, as explicitly assumed in writing the vector field $f(x)$ in terms of the ergodic density [6–8]. As discussed in [15], delta-correlated Markovian noise is always a limiting approximation of noise with a finite correlation time. Meaningfully taking the derivatives requires first choosing a functional form for the (co)variance whose limit is a delta function (another, more direct approach would be in terms of generalized functions, but here too, additional information is required to specify the derivatives [16]). However, different choices can lead to vastly different central moments of the generalized noise distribution, including those that vanish or diverge at all orders. In the former case, the process in terms of generalized coordinates may not be ergodic [17]; in the latter case, the process is not well defined. In general, it is not clear that Equation (4) holds in the non-Markovian case, since the standard derivations in [6,7] and related works rely on delta-correlated noise.

Here, we can therefore assume that the noise is such that the derivatives in Equation (A70) can be treated as Markov and Gaussian. We also assume that $\frac{d^n}{dt^n}\omega$ is independently and identically distributed to $\frac{d^{n-1}}{dt^{n-1}}\omega$ for all n. Finally, we can then define the (infinite) matrix $\bar{M}$ as the block diagonal matrix with all blocks equal to M:

$$\bar{M} := \begin{pmatrix} M & 0 & \dots \\ 0 & M & \\ \vdots & & \ddots \end{pmatrix} \tag{A71}$$

The time derivative of ω is independent of ω, as the changes are independent of the value of ω. Therefore, we actually get an infinite number of independent and identically distributed systems. Using these definitions, we have:

$$\dot{\tilde{x}} = \bar{M}\tilde{x} + \tilde{\omega}. \tag{A72}$$

These equations describe a random dynamical system composed of an infinite number of independent linear random dynamical systems, all governed by the same matrix M and driven by independently and identically distributed noise. Since the first of these systems (for the variables x) is ergodic by assumption, all of the subsystems are also ergodic, and therefore, the whole system is ergodic with the ergodic density equal to a product of the original ergodic density:

$$\bar{p}^*(\tilde{x}) = p^*(x)p^*(x')p^*(x'')\cdots p^*(x^{(n)})\cdots. \tag{A73}$$

Additionally, if M is such that:

$$\begin{aligned} M_\psi \cdot (\psi, s, a, r)^\top &= f_\psi(\psi, s, a) = (\Gamma - Q)_{\psi\psi} \nabla_\psi \ln p^*(\psi, s, a, r) \\ M_s \cdot (\psi, s, a, r)^\top &= f_s(\psi, s, a) = (\Gamma - Q)_{ss} \nabla_s \ln p^*(\psi, s, a, r) \\ M_a \cdot (\psi, s, a, r)^\top &= f_a(s, a, r) = (\Gamma - Q)_{aa} \nabla_a \ln p^*(\psi, s, a, r) \\ M_r \cdot (\psi, s, a, r)^\top &= f_r(s, a, r) = (\Gamma - Q)_{rr} \nabla_r \ln p^*(\psi, s, a, r), \end{aligned} \tag{A74}$$

(which is the case for the M in the counterexample to Step 6) then for:

$$\begin{aligned} (x_1, x_2, x_3, x_4) &:= (\psi, s, a, r) \\ (x_1', x_2', x_3', x_4') &:= (\psi', s', a', r') \\ (x_1'', x_2'', x_3'', x_4'') &:= (\psi'', s'', a'', r'') \\ &\vdots \\ (x_1^{(n)}, x_2^{(n)}, x_3^{(n)}, x_4^{(n)}) &:= (\psi^{(n)}, s^{(n)}, a^{(n)}, r^{(n)}) \\ &\vdots, \end{aligned} \tag{A75}$$

$$\bar{Q} := \begin{pmatrix} Q & 0 & \cdots \\ 0 & Q & \\ \vdots & & \ddots \end{pmatrix}, \tag{A76}$$

$$\bar{\Gamma} := \begin{pmatrix} \Gamma & 0 & \cdots \\ 0 & \Gamma & \\ \vdots & & \ddots \end{pmatrix}, \tag{A77}$$

and using Equation (8) and that the inverse of a block diagonal matrix is block diagonal:

$$\bar{U} := \begin{pmatrix} U & 0 & \cdots \\ 0 & U & \\ \vdots & & \ddots \end{pmatrix}, \tag{A78}$$

we also have:

$$\begin{aligned} \bar{M}_{\tilde{\psi}} \cdot (\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})^\top &= f_{\tilde{\psi}}(\tilde{\psi}, \tilde{s}, \tilde{a}) = (\bar{\Gamma} - \bar{Q})_{\tilde{\psi}\tilde{\psi}} \nabla_{\tilde{\psi}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ \bar{M}_{\tilde{s}} \cdot (\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})^\top &= f_{\tilde{s}}(\tilde{\psi}, \tilde{s}, \tilde{a}) = (\bar{\Gamma} - \bar{Q})_{\tilde{s}\tilde{s}} \nabla_{\tilde{s}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ \bar{M}_{\tilde{a}} \cdot (\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})^\top &= f_{\tilde{a}}(\tilde{s}, \tilde{a}, \tilde{r}) = (\bar{\Gamma} - \bar{Q})_{\tilde{a}\tilde{a}} \nabla_{\tilde{a}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \\ \bar{M}_{\tilde{r}} \cdot (\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})^\top &= f_{\tilde{r}}(\tilde{s}, \tilde{a}, \tilde{r}) = (\bar{\Gamma} - \bar{Q})_{\tilde{r}\tilde{r}} \nabla_{\tilde{r}} \ln p^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}). \end{aligned} \tag{A79}$$

The ergodic density of such a system is a product of the ergodic densities of the original system Equation (A56):

$$\bar{p}^*(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) = p^*(\psi, s, a, r) p^*(\psi', s', a', r') p^*(\psi'', s'', a'', r'') \cdots . \tag{A80}$$

Thus, any property of the original system is also a property of the generalized coordinate system.

References

1. Friston, K. Life as we know it. *J. R. Soc. Interface* **2013**, *10*, 2013.0475. [CrossRef] [PubMed]
2. Friston, K. A free energy principle for a particular physics. *arXiv* **2019**, arXiv:1906.10184.

3. Parr, T.; Da Costa, L.; Friston, K. Markov blankets, information geometry and stochastic thermodynamics. *Philos. Trans. R. Soc. A* **2019**, *378*, 2019.0159. [CrossRef] [PubMed]
4. Friston, K.; Sengupta, B.; Auletta, G. Cognitive Dynamics: From Attractors to Active Inference. *Proc. IEEE* **2014**, *102*, 427–445. [CrossRef]
5. Friston, K.; Rigoli, F.; Ognibene, D.; Mathys, C.; Fitzgerald, T.; Pezzulo, G. Active inference and epistemic value. *Cogn. Neurosci.* **2015**, *6*, 187–214. [CrossRef] [PubMed]
6. Ao, P. Potential in stochastic differential equations: novel construction. *J. Phys. A* **2004**, *37*, L25–L30. [CrossRef]
7. Kwon, C.; Ao, P.; Thouless, D.J. Structure of stochastic dynamics near fixed points. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13029–13033. [CrossRef] [PubMed]
8. Kwon, C.; Ao, P. Nonequilibrium steady state of a stochastic system driven by a nonlinear drift force. *Phys. Rev. E* **2011**, *84*, 061106. [CrossRef] [PubMed]
9. Ma, Y.A.; Chen, T.; Fox, E.B. A complete recipe for stochastic gradient MCMC. In *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2*; MIT Press: Montreal, QC, Canada, 2015; pp. 2917–2925.
10. Yuan, R.; Tang, Y.; Ao, P. SDE decomposition and A-type stochastic interpretation in nonequilibrium processes. *Front. Phys.* **2017**, *12*, 120201. [CrossRef]
11. Ao, P.; Chen, T.Q.; Shi, J.H. Dynamical Decomposition of Markov Processes without Detailed Balance. *Chin. Phys. Lett.* **2013**, *30*, 070201. [CrossRef]
12. Yuan, R.S.; Ma, Y.A.; Yuan, B.; Ao, P. Lyapunov function as potential function: A dynamical equivalence. *Chin. Phys. B* **2014**, *23*, 010505. [CrossRef]
13. Bishop, C. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer: New York, NY, USA, 2006.
14. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: Hoboken, NJ, USA, 2006.
15. van Kampen, N.G. *Stochastic Processes in Physics and Chemistry*; North-Holland: Amsterdam, The Netherlands, 1981.
16. Oberguggenberger, M. Generalized Functions and Stochastic Processes. In *Seminar on Stochastic Analysis, Random Fields and Applications. Progress in Probability*; Bolthausen, E., Dozzi, M., Russo, F., Eds.; Birkhäuser: Basel, Switzerland, 1995; Volume 36, pp. 215–230.
17. Cornfeld, I.P.; Fomin, S.V.; Sinai, Y.G. *Ergodic Theory*; Springer: New York, NY, USA, 1982.