Derivation of the Variational Bayes Equations

Alianna J. Maren
Themasis Technical Report TR-2019-01
alianna.maren@northwestern.edu
alianna@aliannajmaren.com
June 27, 2019

Abstract

The derivation of key equations for the variational Bayes approach is well-known in certain circles. However, translating the fundamental derivations (e.g., as found in Beal (2003)) to the notation of Friston (2013, 2015) is somewhat delicate. Further, the notion of using variational Bayes in the context of a system with Markov blankets requires special attention. This Technical Report presents the derivation in detail. It further illustrates how the variational Bayes method provides a framework for a new computational engine, incorporating the 2-D cluster variation method (CVM), which provides a necessary free energy equation that can be minimized across both the external and representational systems’ states, respectively.

“Do you understand this?” she demanded. ...
“It seems simple enough,” he said after a moment.
“I knew it,” she muttered, crossing her arms. “I knew it was written in male.”

Heir to the Shadows: Book 2 of the Black Jewels Trilogy
Anne Bishop (1999), p. 214 (Trade edition).

1 Introduction

Friston (2010, 2013, 2015)[1][2][3] has proposed that free energy minimization serves as a unifying theory for describing neural dynamics.
He further suggests that statistical thermodynamics can model neuronal systems, drawing on the dynamic properties of activated neuronal ensembles [3]. This elegant and fascinating notion depends on the use of the variational Bayes approach, together with the idea of a Markov blanket which separates an internal computational (“representational”) set of units from external ones.

While this approach is certainly attractive, it is potentially difficult for many readers to follow the translation from one of the earlier presentations of variational Bayes (by Beal (2003) [4]) to the equations used by Friston (op. cit.). This Technical Report serves as a minitutorial, carefully delineating how the free energy equations presented in Friston (op. cit.) correspond to the detailed derivations presented in Beal (2003), which were originally presented in Feynman [5] and in Hinton and van Camp [6]. It identifies how - although the equations may seem formally identical - there are certain key differences in the two presentations. Further, it addresses the notion of terminology, carefully identifying what certain terms actually mean, regardless of how they have been described by various authors in different works.

The intention of this work is to facilitate a new computational engine. Such an engine would make use of not only Friston’s notion of a set of computational (representational) units separated from an external system by a Markov blanket, but also follow the variational Bayes (free energy minimization) approach described by both Friston (op. cit.) and Beal (2003).

In brief, Friston proposes a computational system in which a Markov blanket separates the computational (representational) elements of the engine from external events, as shown in Figure 1. The communications between the external system elements (denoted $\tilde{\psi}$) with those of the representational system (denoted $\lambda$ or $\tilde{r}$) are mediated by two distinct layers or components of the Markov blanket; the sensing ($\tilde{s}$) elements and the action ($\tilde{a}$) ones. An explanatory comment offered by Friston (personal communication) is that “The distinction between active and sensory states is dictated by the definition of a Markov blanket; namely, active states influence but are not influenced by external system elements, while sensory states influence but are not influenced by the representational system. This ensures that the external states and the internal or representational states are conditionally independent.”

In the notions offered by Friston, both the external system ($\tilde{\psi}$
Figure 1: Illustration of a cluster variation method (CVM)-based computational engine, for which the Markov blanket of sensing and active units corresponds to input and output layers (see Friston [3]). Unique to the approach advanced here, the computational layer is composed as a 2-D CVM, for which the free energy equation can be explicitly written, and the free energy minimum can be found either analytically or computationally, depending on the parameters used. The CVM layer comprises the internal or representational units ($\tilde{r}$), and cannot communicate with the external field (shown in two parts for visualization purposes only). However, units within the representational layer can receive inputs from the sensory units ($\tilde{s}$) and send signals to the active ($\tilde{a}$) units. The sensory units can receive inputs from external stimulus, and send signals to the representational units. The active units can receive inputs from the representational units, and send signals to the external system. (In the notional view advanced by Friston (op. cit.), a broader set of interactions is allowed; for simplicity in this engine, the interaction pathways have been streamlined.)
units) and the representational system (˜r units) each independently come to a free energy minimum. The activation of ˜r units within the representational system can be mediated by certain parameters (θ), so that the representational system models the external one.

More appropriately, since we are specifying that both the external and representational systems come to free energy minima, we would state that the nature of the representational system, when it achieves free energy minimization, approximates that of the external system, which also comes to free energy minimization. The degree-of-closeness of the model approximation to the external system is mediated by the parameter(s) θ.

In order to create such a computational engine, we need a formalism that will actually allow this free energy minimization to take place, in both the external and representational components of the system. Maren [7] has suggested that a free energy formulation, known as the cluster variation method (CVM), can potentially serve in such a computational engine, as was shown in Fig. 1.

As a first step, we will use Friston’s framework for the variational Bayes approach, and this requires that we derive the basic variational Bayes equations.

2 The variational free energy

The goal of this section is to follow the Friston approach and express variational free energy as (what he refers to as) the “thermodynamic free energy” minus the relative entropy (of the model system) or, equivalently, as a free energy-like term plus the Kullback-Leibler divergence between the variational density (the density of the model system) and the posterior density (of the external conditioned on the model) over external states (p. 4, [3]), as shown in Eqn. 2.

Friston expresses the variational free energy of an ensemble as the following equations [2, 3] (where the exact notation is taken from Friston (2015) [3], Eqn. 3.2)

\[
\begin{align*}
    f_r(\tilde{s}, \tilde{a}, \tilde{r}) &= (Q_r - \Gamma_r)\nabla_{\tilde{r}} F \\
    f_a(\tilde{s}, \tilde{a}, \tilde{r}) &= (Q_a - \Gamma_a)\nabla_{\tilde{a}} F
\end{align*}
\]  

and
\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = E_q[L(\tilde{x})] - H[q(\tilde{\psi}|\tilde{r})] \]
\[ = L(\tilde{s}, \tilde{a}, \tilde{r}) + D_KL[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})]. \]

where the variables are defined in the following Table 2. The tilde notation (with variables \( \tilde{\psi}, \tilde{r}, \tilde{s}, \) and \( \tilde{a} \)) all refer to these as being “generalized” variables [3].

The flow of system states, represented by Eqn. 1, is particular to each type of unit, so that \( f_a(\tilde{s}, \tilde{a}, \tilde{r}) \) is the (gradient descent) change in the set of action units (\( \tilde{a} \)), and \( f_s(\tilde{s}, \tilde{a}, \tilde{r}) \) is the change in the set of action units (\( \tilde{s} \)). These system states are subject to random fluctuations denoted by \( \omega \). As noted by Friston (personal communication), “The amplitude of the random fluctuations is controlled by [the diffusion tensor] \( \Gamma \), while [the set of] \( Q \) are antisymmetric matrices that allow for solenoidal flow (which does not change free energy).” (See further discussion in Sect. 2 of Friston (2013) [2].)

This Technical Report focuses exclusively on the static Eqn. 2 and defers the dynamic Eqn. 1 to a different occasion.

Eqn. 2 expresses a free energy-like term \( F(\tilde{s}, \tilde{a}, \tilde{r}) \) in two ways. First, as the difference of the expectation of \( L(\tilde{x}) \) (a log-likelihood term) and the entropy of the model system, \( H[q] \). We will determine the exact meaning of \( L(\tilde{x}) \) in Section 5. The second expression for \( F(\tilde{s}, \tilde{a}, \tilde{r}) \) is the sum of \( L(\tilde{s}, \tilde{a}, \tilde{r}) \) and the Kullback-Leibler (K-L) divergence between the model and the external system. We will determine the meaning of \( L(\tilde{s}, \tilde{a}, \tilde{r}) \) in Section 4, and identify how (and why) it is different from the \( L(\tilde{x}) \) used in the first expression.

| Variable         | Meaning                                                                 |
|------------------|-------------------------------------------------------------------------|
| \( \tilde{s}, \tilde{a}, \tilde{r} \) | Generalized expressions for sensory (\( \tilde{s} \)), active (\( \tilde{a} \)), and internal (or representational) (\( \tilde{r} \)) states |
| \( \tilde{\psi} \) | States of the world (system being modeled) that cause sensory states, and which can be influenced by action |
| \( f_x(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \) | Flow of system’s states |
2.1 The final result in a nutshell

The essence of Eqn. 2 is that we are taking a single expression, and parsing and re-organizing it to achieve two different ways of re-expressing the same thing. The expression, sometimes called the “variational free energy” (see, e.g., Friston (op. cit.)) is given as

\[
F(\tilde{s}, \tilde{a}, \tilde{r}) = -\int_{\psi} q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})}{q(\tilde{\psi}|\tilde{r})} \right) d\psi. \quad (3)
\]

Diagrammatically, we can see this shown in Figure 2. The tilde notation is dropped in this figure and in the immediately-following subsection, which discusses this figure.

![Diagrammatic illustration of Eqn. 2](image)

Figure 2: Diagrammatic illustration of Eqn. 2

2.2 Quick Summary of Key Points

In Figure 2 we saw that the initiating expression, the “variational free energy,” was being reorganized in two different ways. On the
Left-Hand-Side (LHS), we saw that it led to a simple sum over the log-likelihood of the set of variables associated with the representational units and the Markov blanket, and also a Kullback-Leibler (K-L) divergence term expressing the difference between the model $q$ and the probability distribution of the external units $\psi$ as conditioned on the internal units and the Markov blanket. We achieve this result by re-expressing the probability of the joint co-occurrence, $p(\psi, s, a, r)$ as a conditional probability, using the formulation for a Bayesian posterior. As a supporting step, we'll see that the summation (or integration, as it depends on how the various systems are formed) over the model $q$ is independent of that over the probability $p(\psi, s, a, r)$, and so the summation over $q$ disappears, leaving us with a simpler term.

On the Right-Hand-Side (RHS), we see that the re-organization is much simpler, and can indeed be followed simply by examining the diagram itself. (There are a few subtleties, which are addressed in the following sections.) The first term on the RHS is weighted sum over the joint probability distribution $p(\psi, s, a, r)$. The second term is a term that looks remarkably like entropy.

We suspect that the structure of the equation on the RHS, and the “entropy-like” appearance of the last term, is what has given rise to expressing the whole equation as “variational free energy.” This is because there is an apparent similarity in the form of the resulting equations, and not because the equation results from the usual steps applied to a partition function, as is done more typically in statistical thermodynamics. (See Appendix A for a quick review of those basic statistical thermodynamics equations.)

### 2.3 Various interpretations

Various authors interpret Eqn. 2 with different notations and descriptive phrases. The purpose of this subsection is to identify a few of these interpretations, and to tease out exactly what is meant from exactly what is said. This should make it easier for those reading the source papers to understand what is actually being presented.

For example, Friston (2013) presents this Report’s Eqn. 2 as his Eqn. 2.7 in [2], using the notation

$$F(s, a, \lambda) = E_q[G(\psi, s, a, \lambda)] - H[q(\psi|\mu)].$$

Friston refers to $E_q[G]$ saying that “[T]he last equality just shows that free energy can be expressed as the expected Gibbs energy minus
the entropy of the variational density.”

However, Sengupta, Stemmler, and Friston [8] state that “$U(t) = -\ln p(s(t), \psi(t)|m)$ corresponds to an internal energy under a generative model of the world, described in terms of the density over sensory and hidden states $p(s,y|m)$.” (Author’s note: $U(t)$ corresponds to $E_q[G]$, from [8] and [2], respectively.) Moreover, Sengupta et al. state that “$F(t)$ is called free energy - by analogy with its thermodynamic homologue that is defined as internal energy minus entropy. However, it is important to note that variational free energy is not the Helmholtz free energy ... it is a functional of a probability distribution over hidden (fictive) states encoded by internal states $q(y|m)$, not the probability distribution over the (physical) internal states. This is why variational free energy pertains to information about hidden states that are represented, not the internal states that represent them.”

(Author’s note: For the benefit of those who wish to compare the information-theoretic approach of Beal, Friston, and others against a classic statistical thermodynamics formulation, Appendix A derives fundamental thermodynamic concepts, and Beal’s results compared with the corresponding statistical thermodynamic formalism are given in Appendix B. As stated by Sengupta et al., they are not precisely the same [8].)

It will be clear, in the succeeding derivations, that what is offered as $E_q[G]$ is not what we are familiar with as the Gibbs free energy from statistical thermodynamics. However, Sengupta et al. offer the following explanation as a Lemma:

“Lemma: (complexity minimisation) Minimising the complexity of a conditional distribution - whose sufficient statistics are (strictly increasing functions of) some unconstrained internal variables of a thermodynamic system - minimises the Helmholtz free energy of that system.”

As proof, they suggest that we can use standard results from Bayesian statistics [4] in order to express free energy as complexity minus accuracy. They are, in fact, referring to the derivation in Beal (2003) that will be the fundamental reference in this Technical Report. They conclude that “In sum, the internal states encoding prior beliefs about hidden states of the world are those that minimise Helmholtz free energy and the complexity defined by variational free energy.”

Friston cites Beal (2003) [4] for the derivation of Eqns. 1 and 2. Blei et al. [9] also provide a useful tutorial. The following sections walk
through the derivations as provided by Beal, using some of the material provided by Blei et al. to support and elucidate certain points. The goal is to make the match between the free energy equations as expressed by Beal (and in certain cases, by Blei et al.) with those used by Friston as clear and as transparent as possible.

3 Important distinction & clarification

Following the transition of the variational Bayes approach from Beal [4] to Friston (op. cit.), and from thence to an actual, computable model, requires some subtlety and attention to detail. The most crucial consideration at the outset is that with Beal (and with previous expostulations on the variational Bayes method), both the actual data points being modeled and the model itself are expressed with regard to some underlying variables, $x_i$. Thus, we have the external (or observable, or dependent) variables $y_i = y(x_i)$, and the model $q_{x_i}(x_i)$. An illustration of this is shown in Figure 3.

Figure 3: In the variational Bayes method described by Beal, the dependent variable $y(i)$ is the data taken over the “hidden” independent variable $x(i)$, and the model is a function of $x(i)$.

The illustration in Figure 3 serves as a depiction of how Gaussian models are typically used to illustrate the variational Bayes approach.
Later in this Report, I will introduce an alternative using the cluster variation method (CVM) for free energy.

In contrast to Beal’s description, with Friston, the external system and the representation are separated by a Markov blanket. Thus, the external system’s units, denoted $\psi$, are distinct from the representational system’s units. Friston’s approach is illustrated in Figure 4.

Friston’s Notation: $p(s,a,r)$ is separate and (somewhat) independent of $p(\psi)$; $q(r, \theta)$ is model

Figure 4: In the variational Bayes method described by Friston, the external system is denoted by $\psi$, which interacts with a separate representational system denoted by $r$. The two systems are separated by a Markov blanket composed of sensing ($s$) and action ($a$) units. (Note: tilde notation dropped from this figure.) The model $q$ is a model on the representational system; $q = q(r, \theta)$, where $\theta$ denotes the model parameters.

In a beautiful and intriguing illustration, Friston applies his formulation to the emergence of a Markov blanket, beginning with “an ensemble of elemental subsystems with (heuristically speaking) Newtonian and electrochemical dynamics ... One can think of these generalized states as describing the physical and electrochemical state of large macromolecules. Crucially, these states are coupled within and between the subsystems comprising an ensemble”[2]. In this illustration, the units comprising the Markov blanket and the internal, “representational” system emerge over time. Appropriately enough, Friston’s illustration focuses on the dynamic behaviors of the various units.
3.1 Primary distinction: Beal and Friston

Turning our attention back to the evolution of Friston’s formulation from that expressed by Beal, we investigate the correspondence between Beal’s expression of \( p(y_i|\theta) \) in comparison with Friston’s use of \( p(\tilde{s}, \tilde{a}, \tilde{r}) \).

In writing \( p(y_i|\theta) \), Beal is actually describing the external system; that which is being modeled. He is specifically referring to the (integration of) the joint probability distribution \( p(x_i, y_i|\theta) \), where the \( y_i \) are the dependent, or observable variables, and the \( x_i \) are the independent, or hidden, or latent variables. Specifically, Beal states, “Consider a model with hidden variables \( x \) and observed variables \( y \). The parameters describing the (potentially) stochastic dependencies between variables are given by \( \theta \).”

The key to understanding that Beal is describing the external (observable) system (also that which Friston refers to as being denoted by the units \( \psi \)) is that both Beal and Friston refer to the model as \( q(x_i) \) (see Eqn. 2 as an example). As Friston notes (personal communication), “The link between the two formulations rests upon associating the causes \( x_i \) and \( \theta \) in Beal’s notation with the external states that cause data \( (y_i) \) in Friston’s formulation (i.e., \( x \) and \( \theta \) correspond to \( \psi \)), where the data in Friston’s formulation become the sensory units states \( (\tilde{s}) \).”

In support of this, we note that Blei et al. [9] stated that “the approach is to posit a family of approximate densities \( Q \), which are defined over the set of latent variables ... Then, we try to find that member \( q^*(z) \) of the set \( Q \) that is the Kullback-Leibler divergence (KL) of \( q^*(z) \) with respect to the exact posterior \( p(z|x) \), which represents the probability distribution of the latent variables with regard to the observables, given as

\[
q^*(z) = \arg\min_{q(z) \in Q} KL(q(z)||p(z|x)); q(z) \in Q.\]

Note that in this explanation by Blei et al., the observable variable was denoted as \( z \) instead of \( y \), but the independent (and hidden) variable was still denoted \( x \).

In short, common to these different expostulations on the variational Bayesian approach, the model for the observable units is denoted by \( q \), where exact values for \( q \) are a function of certain parameters (\( \theta \), in the case of Beal, and \( m \) or sometimes \( \mu \), in the case of Friston).
The important thing to note, and the thing that differentiates Friston’s formulation from Beal’s, is that with Beal, both \( y \) and \( q \) are applied to the same underlying hidden or independent variable(s) \( x \). Essentially, \( p \) represents the exact observed data, and \( q \) is the model (subject to parameter fitting with \( \theta \)). In contrast, with Friston, we are dealing with entirely different sets of units; the external units (denoted by \( \psi \) and the internal, representational units (denoted by \( \tilde{r} \)). We can express a probability distribution over each, but they are not necessarily the respective observed data points and modeling functions taken over the same underlying base set of values.

3.2 Integrating over the model space

As we will note during subsequent derivations (presented in Subsection 4.1), the integration (in various equations) over \( q \) is shown as being with regard to the external system units, \( \psi \). More to the point, Friston uses the notation of integrating over \( \psi \), in order to show the comparison between his formulation and that of Beal.

Specifically, we will see that Beal uses the integration (Section 5)

\[
L(\theta) \geq \sum_{i=1}^{n} \int dx_i \ q_{\psi_i}(x_i) \ln \left( \frac{p(x_i, y_i | \theta)}{q_{\psi_i}(x_i)} \right),
\]

and correspondingly, Friston uses the integration (Subsection 4.3)

\[
F(\tilde{s}, \tilde{a}, \tilde{r}) = -\int_{\psi} q(\psi | \tilde{r}) \ln \left( \frac{p(\psi, \tilde{s}, \tilde{a}, \tilde{r})}{q(\psi | \tilde{r})} \right) d\psi.
\]

Friston is clearly making an effort to show the correspondence between his formulation and that of Beal. Also, as he envisions it, a summation or integration of the model over the external system may be possible. (His illustration in [2] showed the evolution of the internal states and the Markov blanket from an original “primordial soup” encompassing all the units.)

For the purposes of this Technical Report, though, we envision a system where the model system \( q \) refers strictly to the internal (representational) units \( \tilde{r} \). Over time, the goal is to adjust the units \( \tilde{r} \) so that the free energy of the model \( q \) approximates that of the external system with units \( \psi \).

Thus, the elements of our system that we’ve been considering so far consist of three things:
1. The external system which is composed of units $\tilde{\psi}$; we are trying to model this, and we operate under the presumption that we cannot always directly compute certain measures on this system, 

2. The internal system which is composed of units $\tilde{r}$; at any given moment we can determine certain measures on this system, yielding $L(\tilde{s}, \tilde{a}, \tilde{r})$ (we are temporarily ignoring $\tilde{s}$ and $\tilde{a}$), and 

3. A model of the external system expressed via the internal system, $q$, where the chief distinction is that when we take an actual value for $q$, we do so with the presumption that the internal system is brought to a free energy equilibrium for a given set of parameter values $\theta$. This means that the measures for a given distribution-in-the-moment, as represented by $L$, would be adjusted to represent what they would be if the internal system were brought to equilibrium, for a specific set of $\theta$. 

Thus, when it comes to the integration steps, we will consider that an integration of the model system $q$ over $\tilde{\psi}$ will be interpreted as integrating over the model units themselves $\tilde{r}$, but with consideration that the model will have come into a free energy equilibrium (subject to parameters $\theta$) that is a best approximation, for that set of $\theta$, to the external system.

4 The variational free energy: K-L divergence and log-likelihood

In Eqn. 2, the free energy is expressed in two different ways:

1. As the difference between an enthalpy-like term and the entropy of the model system, and

2. As the sum of a (potentially free energy-like) term and the K-L divergence between the probability distributions of the model and the external system.

In this section, we focus on the second half of Eqn. 2; the equality between the “variational free energy” and the sum of a different (presumably free energy-like) term and the K-L divergence. Specifically, we wish to show that

$$F(\tilde{s}, \tilde{a}, \tilde{r}) = L(\tilde{s}, \tilde{a}, \tilde{r}) + D_{KL}[q(\tilde{p} | \tilde{r}) || p(\tilde{p} | \tilde{s}, \tilde{a}, \tilde{r})].$$

(4)
In achieving this goal, we will also accomplish two other tasks, namely:

1. Obtain a precise mathematical formation for \( F(\tilde{s}, \tilde{a}, \tilde{r}) \), and
2. Interpret this mathematical formulation in a useful manner.

We begin our derivation of Eqn. 4 by first considering the definition for the K-L divergence in the context of the system that we are describing (and using the notation advanced by Friston (2015) [3]).

### 4.1 Interpreting the K-L divergence

For the discrete case, we write the Kullback-Leibler (K-L) divergence as

\[
D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = \sum_{i=1}^{I} q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})} \right). \tag{5}
\]

We briefly interpret the physical meaning of the terms in Eqn. 5. The K-L divergence measures the difference between the model (i.e., probability distribution over) of the external system, \( q \), and the external system itself, \( \tilde{\psi} \).

The previous Eqn. 5 includes a summation sign, which is typically found in expressions of the K-L divergence. This summation, however, refers to summing over all instances of data points in the system being modeled (here, denoted \( \tilde{\psi} \) as it occurs with a specific probability \( p \)) and the corresponding points in the model, denoted \( q(\tilde{\psi}) \). In our case, our use of the notation \( \tilde{\psi} \) refers to the full collection of elements being modeled, and the summation sign is not needed. Thus, without loss of meaning, we can rewrite Eqn. 5 as

\[
D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})} \right). \tag{6}
\]

The model \( q \) is a model of the external system, \( \tilde{\psi} \), which is why we write \( q = q(\tilde{\psi}) \). The key feature in computing \( q \) is that (for the application being considered here) we take it at the equilibrium state. That is, \( q \) corresponds to the equilibrium free energy of the external system, which can be computed (or approximated) if we have a suitable free energy equation. Thus, in Eqn. 5, we are looking at the
divergence between the model of the system at equilibrium and the probabilities of various components of the system, potentially in a not-yet-at-equilibrium state.

The parameter(s) $\theta$ can indeed influence $q$, but the notation for $\theta$ is suppressed throughout.

Thus, we can read the term $q(\tilde{\psi}|\tilde{r})$ as the “probability distribution of the model of the external system $\tilde{\psi}$, which is computed based solely on the value of the representational units $\tilde{r}$ that are isolated from the external system $\tilde{\psi}$ by a Markov blanket, but these representational units are to be considered with their at-equilibrium values.”

Next, we examine the term $p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})$, which expresses the probability distribution of units $\tilde{\psi}$ in the external system, conditioned on the Markov blanket sensory units $\tilde{s}$ and action units $\tilde{a}$, along with the representational units $\tilde{r}$. We recall, from the design of the entire system (external plus Markov blanket plus representational units), and also from figures given in Friston [2] and Friston et al. [3], and replicated in Figure 4 that the representational units do not communicate directly with the external units. Thus, the dependence of the $\tilde{\psi}$ is very much an implicit relationship; one that is at a distance because the direct interactions of the units in $\tilde{\psi}$ are exclusively with $\tilde{s}$ and $\tilde{a}$. Further, the system design is that the sensory units receive inputs from the external units $\tilde{\psi}$, but do not directly influence the $\tilde{\psi}$ themselves.

Thus, the conditional relationship expressed in $p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})$ seems a little forced. However, it is the basis for our next steps in the derivation, and we will think of it simply as stating that the external system can indeed be influenced by the evolving values for the representational system $\tilde{r}$.

With this in mind, we go back to Eqn. 5 and interpret the K-L divergence on the Right-Hand-Side (RHS) of the equation. It states that the K-L divergence is the sum, over all possible states in which the system can possibly find itself, of the probability distribution of the model $q$ for a given specific state, multiplying the natural log of the probability distribution of the model over the probability distribution of the actual, external system, both for that given specific state. In this expression, the probability distribution of the external system is conditioned by the states of the Markov blanket and the representational system.

Thinking ahead, we consider how the traditional formulation for how the relationship between the model and the external system is expressed, specifically as a K-L divergence. Typically, the actual “ex-
ternal” system is some set of values, \( p(x) \), and the model is given as \( q(x) \). The model \( q \) is adjusted via certain parameters \( \theta \) to become close to \( p \), as illustrated in Figure 3. The summation over values \( i = 1..I \) thus runs through the gamut of possible \( x \) values.

In our situation, though, we will be looking at both an external system and an internal model that will each, separately, come to their respective free energy equilibrium points. That is, there will not be a sum over all possible values of some distribution over \( i \); there will instead be a single probability distribution \( p \) and a single probability distribution \( q \), after each has reached free energy minimization.

### 4.2 Rewriting the Bayesian posterior distribution

Before we rewrite the K-L divergence term of Eqn. (5), we first recall how the Bayesian posterior probability density can be rewritten, as framed in Blei et al. [9].

Consider a system that has a set of observable variables \( v = v_{1..V} \) and a set of latent or “hidden” variables \( w = w_{1..W} \). In a feedforward neural network, for example, the observable variables \( v \) would be the values of the output layer neurons, and the latent (hidden) variables would be the associated values of the hidden layer \( w \) neurons.

Similarly, we can envision many other situations in which we can identify an observation that is a function of multiple input factors. Sometimes, not all of those input factors can be directly observed.

In the Bayesian formalism, the prior density of the (set of) latent variables \( w \) is defined as \( p(w) \). A Bayesian model relates these latent variables to the observations \( v \) through the likelihood \( p(v|w) \). The interpretation is straightforward; it speaks to the likelihood of observing an outcome or observable variable \( v \) given the hidden variables \( w \). This is called the prior distribution.

Sometimes, though, we don’t have an accurate means of establishing the values for the latent or hidden variables \( w \). Thus, we use approximate inference to determine the posterior distribution, \( p(w|v) \). This means that we are trying to estimate the values of the hidden variables, seeing only the values for the observable variables.

To rewrite the probability density, we first consider a system that can be described in terms of a joint density of latent variables \( w = w_{1..W} \) and observations \( v = v_{1..V} \), where the conditional density function is given as
\begin{equation}
    p(w|v) = p(w,v)/p(v).
\end{equation}

Conversely, we also have

\begin{equation}
    p(w,v) = p(w|v)p(v).
\end{equation}

### 4.3 Rewriting the K-L divergence

We wish now to rewrite the probability density of the external states \( \tilde{\psi} \) that is conditional on the Markov blanket and internal states, so that the probability density becomes a joint distribution.

Applying the transformation of Eqn. 7, we identify the conditional distribution from Eqn. 5 in terms of the joint probability distribution \( p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) \), together with the simple probability distribution over the model states.

\begin{equation}
    p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r}) = p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})/p(\tilde{s}, \tilde{a}, \tilde{r}).
\end{equation}

Substituting this result into Eqn. 6, we have

\begin{equation}
    D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})p(\tilde{s}, \tilde{a}, \tilde{r})}{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})} \right),
\end{equation}

which we can reorganize to write as

\begin{equation}
    D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = q(\tilde{\psi}|\tilde{r}) \ln (p(\tilde{s}, \tilde{a}, \tilde{r}))
    + q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\psi, \tilde{s}, \tilde{a}, \tilde{r})} \right).
\end{equation}

Following Beal [4] (Eqns. 2.32 - 2.34), we note that the (implicit) sum over the model terms \( q \) in the first term on the RHS comes to 1 (as \( q \) is independent of \( p \)), so that we have

\begin{equation}
    D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = \ln (p(\tilde{s}, \tilde{a}, \tilde{r}))
    + q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\psi, \tilde{s}, \tilde{a}, \tilde{r})} \right),
\end{equation}

keeping in mind that the \( q \) are taken over the hidden or latent variables.

We rearrange terms in Eqn. 12 to obtain
\[
q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})} \right) = -\ln (p(\tilde{s}, \tilde{a}, \tilde{r})) \\
+ D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})]
\]

We notice that Eqn. 13 is similar to form of Eqn. 4; sufficiently so that we can establish the identities for \(F(\tilde{s}, \tilde{a}, \tilde{r})\) and \(L(\tilde{s}, \tilde{a}, \tilde{r})\).

For the Left-Hand-Side (LHS) of Eqn. 13 we create the identity for \(F(\tilde{s}, \tilde{a}, \tilde{r})\) as

\[
F(\tilde{s}, \tilde{a}, \tilde{r}) = q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})} \right) \\
= -q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})}{q(\tilde{\psi}|\tilde{r})} \right).
\]

This gives us the precise form for \(F(\tilde{s}, \tilde{a}, \tilde{r})\); the variational free energy. We note the specific difference between this Eqn. 14 and Eqn. 5; both have the form of a K-L divergence. However, in Eqn. 14, the divergence is between the model and the joint distribution of the model and the external system; in Eqn. 6, the divergence is between the model and the external system as conditioned on the representational system and the Markov blanket.

We notice also (in the second part of Eqn. 14) that Friston et al. prefer to represent the variational free energy as the negative of a divergence-like term; it is now between the joint probability distribution (of the external system and the model) against the model, although the multiplying factor is still that of the model distribution.

Similarly, for the first term on the RHS of Eqn. 13 we also take note of the interpretation offered by Friston (2015) [3], which give us

\[
L(\tilde{s}, \tilde{a}, \tilde{r}) = -\ln p(\tilde{s}, \tilde{a}, \tilde{r}),
\]

so that \(L\) is defined as the negative of the logarithm of the probability of internal (representational) units, together with the Markov blanket units.

If we were to substitute these two expressions into Eqn. 13 we would obtain
\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = -q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})}{q(\tilde{\psi}|\tilde{r})} \right) \] (16)

which is identical with the second part of Eqn. 4, and also with Friston (2015), Eqn. 3.2 [3], and also with Friston (2013) Eqn. 2.8 [2].

Once again, it may be useful to consider the distinction between \( q \) and \( L \). More precisely, we need to ask ourselves exactly what it is that we mean when we speak of \( p(\tilde{s}, \tilde{a}, \tilde{r}) \). This is presumably the probability distribution of the internal and Markov blanket states. However, we have been representing the distribution of the internal and Markov blanket states as \( q \); the probability distribution of the model system that is encoded by the internal states \( \tilde{r} \).

A useful interpretation is that we may take \( L(\tilde{s}, \tilde{a}, \tilde{r}) \) to be the actual distribution of the model (as computed directly over its various components), and \( q \) to be the computational distribution of the model. As noted by Friston (personal communication), “In other words, the negative log likelihood of sensory states (and active plus internal states) pertains to the actual state of affairs, while the free energy corresponds to the equivalent measure that would be obtained if the sensory units were caused by the latent or hidden states encoded by the internal states. By minimizing free energy, the two become close but (in general) will never be exactly the same.”

The physical implication here could be that we will obtain \( q \) as the probability distribution for the model in a free energy-minimized state. In contrast, the values of specific elements in the distribution over \( (\tilde{s}, \tilde{a}, \tilde{r}) \) may, at a certain point, not be in a free energy-minimized state.

We have thus accomplished half of our goal, in deriving one of the equalities of Eqn. 2, involving the negative log-likelihood of the probability of states within the model, added to the K-L divergence of the probability distribution of the model with respect to the probability distribution of the external units.

Examining Eqn. 14, we note the correspondence between the expression for \( F(\tilde{s}, \tilde{a}, \tilde{r}) \) given there and the corollary expression given for the free energy of a system in Friston (2013, see Eqn. 2.7) [2], as

\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = -\int_{\psi} q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})}{q(\tilde{\psi}|\tilde{r})} \right) d\psi. \] (17)
The distinction between the (implicit) summation (recall the transition from Eqn. 5 to Eqn. 6) and integral form (just presented in Eqn. 17) is not of great matter. However, we note that the integration in Eqn. 17 is over the external units $\tilde{\psi}$, reinforcing our understanding that the free energy $F(\tilde{s}, \tilde{a}, \tilde{r})$ is really the variational free energy of the external system, with the probability distribution in the numerator of the logarithmic term being taken over the joint distribution of external units $\psi$ together with the internal or representational units $\tilde{r}$ as well as the Markov blanket units.

We notice the correspondence between Eqn. 16 above and Eqn. 2.34 of Beal, in that $F(\tilde{s}, \tilde{a}, \tilde{r}) = -F(q_x(x), \theta)$, where the former notation for the free energy is Friston’s, and the second is Beal’s. (A minor technical note, provided by Friston (personal communication), is that “… the free energy of a probability distribution ($q$) is a functional, while the free energy of its sufficient statistics ($\tilde{r}$) becomes a function.”) Correspondingly, Eqn. 16 is precisely the negative of Eqn. 2.34 in Beal.

Now, we wish to show the first part of Eqn. 2, which expresses the same free energy in terms of an expectation of log-likelihood of a certain term (whose precise nature will be clarified) minus the entropy of the model as applied to the external units (which will also need to be verified).

5 The variational free energy: log-likelihood expectation and entropy

The previous Subsection 4.3 presented a derivation for the second half of Eqn. 2, giving an expression for the variational free energy in terms of the log-likelihood (over the representational or model system) and the K-L divergence (of the model vis-à-vis the external system). In this section, we show how the other equality expressed in Eqn. 2 can be derived, giving the free energy in terms of what Friston calls a “thermodynamic free energy” and an entropy term, the exact natures of which will be determined as we proceed [3].

Several sources remark that the while the first term in the second line of Eqn. 2 is computable, the second term (the K-L divergence) is not [4, 9]. This then motivates the expression given on the first line of the equation, with the intention of giving an alternative - and computable - formulation for the variational free energy $F$.  


The derivation of the first part of Eqn. 2 is found in Beal (2003) [4], Eqns. 2.15 and 2.16.

For convenience, Eqn. 2 (Eqn. 3.2 in Friston et al. (2015) [3]) is presented again, as

\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = E_q[L(\tilde{x})] - H[q(\tilde{\psi}|\tilde{r})] = L(\tilde{s}, \tilde{a}, \tilde{r}) + D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})]. \]

Our goal is to verify the first equality presented in this equation. To accomplish this, we follow a line of reasoning presented in Beal (2003) [4], who introduced a formulation for the log likelihood equation.

We begin with Beal’s Eqn. 2.10, given as

\[ L(\theta) \equiv \ln p(y|\theta) = \sum_{i=1}^{n} \ln p(y_i|\theta) = \sum_{i=1}^{n} \ln \int dx_i p(x_i, y_i|\theta). \]  

Beal’s Eqns. 2.12 - 2.16 are reproduced here as

\[ L(\theta) = \sum_{i=1}^{n} \ln \int dx_i p(x_i, y_i|\theta) \]
\[ = \sum_{i=1}^{n} \ln \int dx_i q_{x_i}(x_i) \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \]
\[ \geq \sum_{i=1}^{n} \int dx_i q_{x_i}(x_i) \ln \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \]
\[ = \sum_{i=1}^{n} \left( \int dx_i q_{x_i}(x_i) \ln p(x_i, y_i|\theta) - \int dx_i q_{x_i}(x_i) \ln q_{x_i}(x_i) \right) \]
\[ = \sum_{i=1}^{n} \left( \int dx_i q_{x_i}(x_i) \ln p(x_i, y_i|\theta) \right) - \int dx_i q_{x_i}(x_i) \ln q_{x_i}(x_i) \]
\[ \equiv F(q_{x_1}(x_1), ..., q_{x_n}(x_n), \theta). \]  

In the preceding Eqn. 19, we note that the last two lines are those that interest us; we see there a formal similarity between those terms and those in the first equality expression of Eqn. 2. Specifically, we desire to show a correspondence between
The “greater-than-or-equal” relation in Eqn. 20 is due to Jensen’s inequality, and is essential to one of the steps shown in Eqn. 19; this is a minor omission in Friston’s phrasing, and does not substantially impact our translation.

We will note, as we pursue our investigation, that Friston’s identification of $F$ is the negative of that used by Beal; this will change the direction of the inequality, but will again not impact our work.

As a precursor step, we identify the three separate equivalences that we will wish to make, specifically that

$$-F(\tilde{s}, \tilde{a}, \tilde{r}) = \sum_{i=1}^{n} \ln \int dx_i q_{x_i}(x_i) \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)},$$

and

$$E_q[L(\tilde{x})] = -\sum_{i=1}^{n} \left( \int dx_i q_{x_i}(x_i) \ln p(x_i, y_i|\theta) \right),$$

and

$$H[q(\tilde{\psi}|\tilde{r})] = \int dx_i q_{x_i}(x_i) \ln q_{x_i}(x_i).$$

\section*{5.1 Equivalence of the variational free energy expressions}

We pause to recollect, from Subsection 4.3 Eqn. 14 the term that Friston has identified for variational free energy. We state this again here for reference as
\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})} \right) = -q(\tilde{\psi}|\tilde{r}) \ln \left( \frac{p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r})}{q(\tilde{\psi}|\tilde{r})} \right). \]  

(25)

We compare this with Beal’s expression from Eqn. 20 in that

\[ L(\theta) = \sum_{i=1}^{\infty} \ln \int dx_i q_{x_i}(x_i) \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \geq F(q_{x_1}(x_1), ..., q_{x_n}(x_n), \theta). \]  

(26)

We note that there is indeed the desired resemblance, however (as previously noted), Friston’s expression is the negative of Beal’s. Further, Beal’s expression integrates over the \( q(x_i) \). More precisely, Beal gives an integration over the hidden or latent elements \( x_i \), and a summation over the units \( y_i \). Friston’s approach simply specifies a distribution \( q \) associated with each specific probabilistic state \( p \).

This is a reasonable transition, as in the transition to the system being described by Friston, we no longer are assessing the values of \( p \) and \( q \) over the same underlying hidden variables \( x \). Rather, the \( p \) corresponds to the external system, and the \( q \) corresponds to the internal (representational) system, which we are bringing into alignment with \( p \). Thus, for every step along the summation of various distinct instances of \( p \), we have a single associated model state \( q \), and not an integration. Thus, this translation is acceptable. We can state that

\[ L(\theta) = -F(\tilde{s}, \tilde{a}, \tilde{r}) \geq F(q_{x_1}(x_1), ..., q_{x_n}(x_n), \theta). \]  

(27)

### 5.2 Equivalence of the entropy expressions

For ease in flow, we next address the equivalence of the two entropy terms, as this is relatively straightforward.

Friston identifies an entropy term \( H \) (using this notation, common to information theory, rather than the more classic thermodynamic notation \( S \)), and we desire that it be equivalent to Beal’s term, as expressed previously in Eqn. 24, and that it also be the negative of the classic entropy formation, which we restate here (using Beal’s notation) as

\[ \sum_{i=1}^{\infty} \ln \int dx_i q_{x_i}(x_i) \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \geq F(q_{x_1}(x_1), ..., q_{x_n}(x_n), \theta). \]
\[ H[q(\tilde{\psi}|\tilde{r})] = \int dx \, q_x(x) \ln q_x(x), \]

which we can also write (using Friston’s notation) as

\[ H[q(\tilde{\psi}|\tilde{r})] = \int d(\tilde{\psi}|\tilde{r}) \, q(\tilde{\psi}|\tilde{r}) \ln q(\tilde{\psi}|\tilde{r}). \]

We recall that the fundamental definition for the entropy of a system is given (see Appendix B) as

\[ S = -k \sum_n P_n \ln P_n, \quad (28) \]

where \( P_n \) refers to the probability of a unit being in energy state \( n \). This is a classic entropy formulation, and we see it replicated in Eqn. (24). The thing that we wish to carefully note is that in Eqn. (24), even though the entropy is being expressed as a function of \( q(x_i) \) (using Beal’s notation) or \( q(\tilde{\psi}|\tilde{r}) \) (using Friston’s notation), the units that are being summed (or integrated) are those in the model, not in the external system. Thus, \( H \) is a function of the model system represented by \( q \).

Very specifically, when it comes to evaluating this function, we would not need a distribution over all possible states in the model. Rather, the computational engine which this Technical Report envisions is one in which the external and internal systems separately come to free energy minima. Thus, when that minimum point is achieved, there would be a single existent value for \( q \); the one which represents the free energy-minimized state for a given set of parameters \( \theta \). Friston (personal communication) notes that “A complementary perspective on this computational saving follows from Feynman’s original motivation; namely, that we have converted a very difficult integration problem into an easy optimization problem. Here, the optimization problem simply entails minimizing variational free energy.”

### 5.3 Equivalence of the enthalpy expressions

Finally, we wish to show the equivalence between the enthalpy terms. The word “enthalpy” may be a misnomer here, but is being used in the classic sense of thermodynamics, in which (see Appendix A) the free energy \( F \) is equal to the enthalpy \( H \) (which is the classic notation for enthalpy, although \( U \) is sometimes used, depending on the version
of free energy being described) minus the temperature $T$ times the entropy $S$, or (in the case of this manuscript as well as in Friston’s work and most information theoretic works) $H$.

Thus, the most classic equation in thermodynamics is

$$ F = H - TS, $$

which states that the free energy is the enthalpy minus temperature times entropy.

The equations presented by Beal and Friston have the same formal structure as the classic free energy equation from statistical thermodynamics, as stated in Eqn. 21, and the version offered by Friston is replicated here as

$$ F(\tilde{s}, \tilde{a}, \tilde{r}) = E_q[L(\tilde{x})] - H[q(\tilde{\psi}|\tilde{r})]. $$

The temperature $T$ has been absorbed in the derivations presented in this work; we are dealing with something called a reduced free energy (and also reduced entropy and reduced enthalpy), which are dimensionless quantities. (Note also that this “reduction” also normalizes the thermodynamic variables with regard to the total number of units in the system. See Appendix A for a review of basic thermodynamics.)

We have already established the correspondence between the Friston’s variational free energy $F(\tilde{s}, \tilde{a}, \tilde{r})$ and (the negative of) that used by Beal, and also identified it as corresponding (in position and form) to the classical free energy. We have also established the correspondence of Friston’s entropy term $H$ with (the negative of) that used by Beal, and identified it as corresponding to the classical entropy term. (In this case the actual expressions are very much aligned.)

We now seek to identify the correspondence between Friston’s enthalpy-like term, $E_q[L(\tilde{x})]$, and the negative of that used by Beal. Specifically, we want to establish Eqn. 23, restating it here for convenience as

$$ E_q[L(\tilde{x})] = -\sum_{i=1} \left( \int dx_i \ q_{x_i}(x_i) \ln p(x_i, y_i|\theta) \right). \quad (29) $$

We also take note of the interpretation offered by Friston (2015) [3], which states that $L(\tilde{x}) = -\ln p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}|m)$ (see Lemma 3.1 and also Eqn. 3.2), or more specifically
\[ L(\tilde{x}) = L(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) = -\ln p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}), \] 
so that \( L \) is defined as the negative of the sum of the logarithm of the joint probability of the external and internal (representational) units, together with the Markov blanket units.

We have previously addressed the nature of \( L \) in Subsection 4.3 and thus will just briefly recapitulate cogent arguments here.

First, we examine the term \( \ln p(x_i, y_i|\theta) \). The joint probability of the dependent variables \( y_i \), co-occurring with the independent variables \( x_i \), as conditioned by the model system parameters \( \theta \), is consistent with Friston’s notation involving a joint probability distribution.

Second, we consider the integration over the \( q_{x_i}(x_i) \) times the logarithm of the probability. As noted previously, the \( q \) and the \( p \) address the distributions over different systems, and thus are independent (to a first order). Thus, we can separate out the integration of the \( q \).

The agreement with Eqn. 30 becomes self-evident, if we associate the external states with \( x \), and the sensory states (plus active and internal states) with \( y \).

5.4 Recapitulation and summary

We now recast Eqn. 19 using Friston’s notation, and keeping in mind that we do not need the internal integral over \( dx_i \), as we are not requiring that the external units and the model units be taken over the same underlying generative source. Further, since \( L(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) = -\ln(p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}|m)) \), the signs on the terms on the RHS of Eqn. 19 have been changed throughout, along with the direction of the inequality.

A key feature in the following Eqn. 31 is that Friston is taking the integration over the units \( \tilde{\psi} \) in the external system, similar to how Beal is summing over the observable units \( y \). However, we will conduct our integration (or summation, which is more literally the case) over the units in the model system, as discussed in Subsection 3.2.

Specifically, Friston’s starting point is Eqn. 2.7 in [2], given as

\[
F(s, a, r) = -\int d\psi q(\psi|r) \ln \left( \frac{p(\psi, s, a, r)}{q(\psi|r)} \right) = E_q[L(\psi, s, a, r)] - H[q(\psi|\mu)].
\]

Note that the tilde notation, indicating generalized variables, is dropped, conforming with the notation that Friston uses in [2], where Friston uses \( G \) for \( L \) in [2].
Friston’s interpretation is that “Here, free energy is a functional of an arbitrary (variational) density \( q(\psi|r) \) [\( q(\psi|\lambda) \) in the original article] that is parametrized by internal states. The last equality just shows that free energy can be expressed as the expected Gibbs energy minus the entropy of the variational density.”

The corresponding expressions, from Eqns. 2.12 - 2.16 in Beal [4] are given as

\[
L(\theta) \geq \sum_{i=1}^{n} \int dx_i \ q_{x_i}(x_i) \ \ln \left( \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \right)
= \sum_{i=1}^{n} \left( \int dx_i \ q_{x_i}(x_i) \ \ln p(x_i, y_i|\theta) \right) - \int dx_i \ q_{x_i}(x_i) \ \ln q_{x_i}(x_i)
= F(q_{x_1}(x_1), \ldots, q_{x_n}(x_n), \theta).
\]

The full derivation, using Friston’s notation, can be found as

\[
L(s, a, r) = -\ln(p(\psi, s, a, r|m))
= -\int d\psi \ \ln (p(\psi, s, a, r))
= -\int d\psi \ \ln \left( q(\psi|r) \frac{p(\psi, s, a, r)}{q(\psi|r)} \right)
\leq -\int d\psi \ q(\psi|r) \ln \left( \frac{p(\psi, s, a, r)}{q(\psi|r)} \right)
= -\int d\psi \ q(\psi|r) \ln (p(\psi, s, a, r)) + \int d\psi \ q(\psi|r) \ln (q(\psi|r))
= E_q[G(\psi, s, a, r)] - H[q(\psi|\mu)]
\equiv F(s, a, r).
\]

(31)

As a small note, Friston shifts notation between the third-to-last and the second-to-last lines of Eqn. 6.2. In the third-to-last equation, he has the expression involving \( q(\tilde{\psi}|\tilde{r}) \). In the second-to-last equation, he uses \( q(\tilde{\psi}|\mu) \). A rationale is that after the integration (where the units that are being considered in the model \( q \) are dependent on the actual representational units \( \tilde{r} \), the dependence of \( q \) on \( \tilde{r} \) no longer needs to be explicitly stated. The introduction of \( \mu \) is simply noting that the computation for the model \( q \) was done with reference to sufficient statistics or parameters \( \mu \), which are associated with the internal states \( (\mu = \tilde{r}) \).
6 Discussion

Now that we’ve done a detailed derivation for both of the equalities expressed in Eqn. 2, it is useful to step back and ascertain exactly what is meant by these paired statements, which are reproduced below for convenience.

\[ F(\tilde{s}, \tilde{a}, \tilde{r}) = E_q[L(\tilde{x})] - H[q(\tilde{\psi}|\tilde{r})] = L(\tilde{s}, \tilde{a}, \tilde{r}) + D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})]. \]

The first expression for the variational free energy puts the influence of the external units in the first free energy term \( E_q[L(\tilde{x})] \). By stating that we desire the expectation of \( L(\tilde{x}) \), we are pushing to identify \( L \) at the point at which we “expect” the system to come to a stable state, i.e., a free energy minimum.

The influence of the “variation” or the perturbation to the system is expressed in terms of the “entropy of the variational density,” \( H[q(\tilde{\psi}|\tilde{r})] \). This puts the variation or perturbation of the external units in the context of the expected values for the internal and Markov blanket units.

The second expression for the variational free energy simply identifies a free energy-like term that involves only internal and Markov blanket units; \( L(\tilde{s}, \tilde{a}, \tilde{r}) \). (Friston notes that “in statistical terms, this is known as the marginal likelihood; having integrated out dependencies on the external states or causes of sensory states. In Bayesian statistics, this is also known as the (negative) log model evidence (see below).”)

The extracted influence of the expected external units (those typically associated with a specific state of internal and Markov blanket units) is now combined with the influence of the variational (or perturbed) external units, within the Kullback-Leibler divergence term, \( D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] \).

(Friston further notes that “in terms of Bayesian statistics, this is the divergence between the approximate and true posterior. This means that minimizing free energy is equivalent to approximate Bayesian inference.”)

6.1 Free Energy Physical Interpretation

We consider that the free energy formulation that we have been developing describes a system with external units \( \psi \), together with a model
system that contains internal units that encode “latent” or “hidden” states, in terms of their sufficient statistics $\tilde{r}$, that are separated from the external system by a Markov blanket comprising sensory units $\tilde{s}$ and action units $\tilde{a}$, as illustrated in Figure 1 of [2]. In other words, internal states encode probability distributions over latent states that ‘could have’ caused the sensory states.

Eqn. 32 gives us the free energy of the system, where the elements of $F(\tilde{s}, \tilde{a}, \tilde{r})$ are formulated in terms of the probability distribution over $\tilde{\psi}$ in terms of $q(\tilde{\psi}|\tilde{r})$, which is the density distribution posed by the model. In contrast, $L(\tilde{s}, \tilde{a}, \tilde{r})$ is a function strictly of the units associated with the model, where the elements include the representational (“latent”) units $\tilde{r}$ along with the Markov blanket units $\tilde{s}$ and $\tilde{a}$. Finally, the K-L divergence term (the final term on the RHS of Eqn. 32) expresses the divergence between the model (expressed as $q$) and the external system (expressed as $p$), given the Markov blanket.

Thus, we can understand that Eqn. 32 expresses the free energy of the system ($F(\tilde{s}, \tilde{a}, \tilde{r})$) as the sum of the free energy of the model ($L(\tilde{s}, \tilde{a}, \tilde{r})$) plus the extent to which the model diverges from the external system.

### 6.2 Free Energy as a Lower Bound

Beal notes that $F(q_x(x), \theta)$ is a lower bound on $L(\theta)$ and is a functional of the free distributions $q_x(x_i)$ and of $\theta$ (the dependence on $y$ is left implicit). The inequality introduced in the third expression makes use of *Jensen’s inequality*.

Beal notes: “Defining the energy of a global configuration $(x, y)$ to be $ln(p(x, y|\theta))$, the lower bound $F(q_x(x), \theta) \leq L(\theta)$ is the negative of a quantity known in statistical physics as the free energy: the expected energy under $q_x(x)$ minus the entropy of $q_x(x)$ (Feynman, 1972; Neal and Hinton, 1998).”

Beal further notes that $F(q_x(x), \theta)$ is the negative of what is known, in statistical thermodynamics, as the free energy of a system, which is the expected energy ($H$) under $q_x(x)$ minus the entropy of $q_x(x)$. Thus, when we shift to the notation of Friston (op.cit.), we reverse the signs on all of the terms on the right-hand-side of Eqn. 31.

As is often noted [2, 4, 9], since the $D_{KL} >= 0$, the free energy for the model is a lower bound for the free energy of the external system. As the model is brought closer to alignment with the external system (the K-L divergence decreases), the free energy of the model
approaches that of the external system \( L(\hat{s}, \hat{a}, \hat{r}) \Rightarrow F(\hat{s}, \hat{a}, \hat{r}) \)).

7 The variational free energy in a new computational engine

7.1 2-D cluster variation method overview

This Technical Report is more aligned with Friston’s approach, and yet has some key differences. We envision the formulation of a representational system whose component elements are pre-specified, and which is distinct from the external system that is being represented. Further, we envision a total system (external together with representation) in which both the external and representational systems can, and indeed do, separately achieve free energy minimization. Their ability to do this requires, of course, that a free energy equation exists for each of these respective systems.

One way in which we can have a system that allows for both free energy minimization and suitable modeling richness is to use a 2-D system constructed as a grid of bistate nodes, as is shown in the following Fig. 5. In such a system, we can use the cluster variation method (CVM) to compute a free energy, for which the enthalpy term is more complex than is typically used. The theoretical basis for this was first developed by Kikuchi [10], and then jointly by Kikuchi and Brush [11]. A more recent description is provided in Maren [7].

The key measurable variables within a CVM systems are the configuration variables. In addition to the simple identification of proportional numbers of “on” (A, black) and “off” (B, white) nodes, these configuration variables also account for the various kinds of nearest-neighbor \( (y_i) \) and next-nearest-neighbor \( (w_i) \) pairs, as well as the six different kinds of triplets (denoted \( z_i \)).

In using the CVM method for describing the free energy, the equilibrium distribution of nodes is governed by two enthalpy parameters, \( h_0 \) and \( h_1 \). The parameter(s) \( h \) are the only “tunable” parameters available in the CVM formulation, and thus are identified with \( \theta \), as used by Friston (op. cit.) and Beal [4].

For the case where the distribution of units into the two states A and B is equiprobable, the activation enthalpy parameter \( h_0 \) is by definition zero. This leaves only a single “tunable” parameter, \( h_1 \); the interaction enthalpy parameter. For this specific case, where the frac-
tions of A and B nodes are equal, there is an analytic solution that provides the relative equilibrium fractions of the different configuration variables as a function of $h_1$ (which is referred to by this author as the h-value). This analytic solution is not particularly accurate for larger h-values (e.g., $h > 1.6$), but it provides a starting point for computational free energy minimization, for a given system and a corresponding given h-value.

7.2 CVM illustration

The following Fig. 5 illustrates initial prototype designs for two systems; (a) corresponds to the external ($\tilde{\psi}$) system, and (b) corresponds to the representational system ($\tilde{r}$). (This follows notation introduced by Friston [3].) The grid pattern of active (A) and inactive (B) units is suitable for modeling using a 2-D cluster variation method (CVM) free energy equation, as described in Maren [7].

Neither of the two systems depicted in Fig. 5 are at equilibrium; both are hand-crafted with the intent of embodying a scale-free type of system. Thus, the first step was to identify a likely h-value candidate (for the external, or $\tilde{\psi}$, system), and then to computationally bring that system into a free energy minimum for that particular h-value, which was determined to be approximately 1.2. (Detailed experimental results are in Maren [12].)

In actuality, we would most likely not know the h-values (plural because we would not be assured of the equiprobability of A and B nodes) for the external $\tilde{\psi}$ system. We would likely be able to sample the configuration variables at various locations, and for various degrees of granularity, for the external system. These would become the inputs ($\tilde{s}$) to the representational system $\tilde{r}$.

The two systems shown in Fig. 5 have similar pattern configurations, with the exception that the larger system shown in (a) has clusters that are proportionately larger than the clusters in the smaller-scale representational system (b). The natures of the patterns within each, though, are much the same. Thus, to a first order, they should have similar (reduced) free energies. More to the point, we envision that the representational system shown in (b) can be brought into alignment with that of (a), or (more specifically) be brought to a free energy minimum with the same (or similar) h-value as with the h-value corresponding to the external system of (a).
Figure 5: Illustration of two systems, arranged so that a 2-D CVM-based free energy can be directly computed for each. (a) The external system, with units denoted \( \tilde{\psi} \). (b) The representational system, showing only the representational units, \( \tilde{r} \). The Markov blanket around the grid of representational units is not shown in this figure. The dark and light-shaded grey and mottled units to the upper and right edges of each system illustrate the wrap-around from the left and bottom edges, used to compute the configuration variables leading to the free energies of each system. Both systems show an approximate scale-free distribution of islands of dark (A) units in a sea of white (B) units. The systems are designed with equiprobable distribution of units into states A and B \( x_A = x_B = 0.5 \), so that the (reduced) free energies of each can be computed directly, using the analytic solution provided in Maren [7]. Details of the corresponding thermodynamic calculations are found in Maren [12]. The systems shown in this figure have been hand-designed to illustrate a potential scale-free configuration; they have not yet been brought into free energy minimization.
7.3 Interpreting the CVM in the variational Bayes framework

The variational Bayes method provides a framework for a new computational engine, and the first step towards this is illustrated in Figure 6. As with the previous Fig. 5, the larger-scale system on the left (a) corresponds to the external system. In the case of Fig. 6 though, the system has been brought to a free energy minimum, for the case where \( h_0 = 0 \) (since the system has been designed with equiprobable distribution of \( A \) and \( B \) nodes) and \( h_1 = 1.2 \). The selection of \( h = h_1 = 1.2 \) was done by first counting the distribution of all the different nearest-neighbor and next-nearest-neighbor pairs, as well as the different triplets, and then computing their relative fractions as configuration variables.

Using the analytic solution for equilibrium values of the configuration values as functions of \( h \), it was possible to estimate a range of possible values for \( h \). (Since the original system of Fig. 5 was not at equilibrium, the various configuration values corresponded to different analytic \( h \)-values. For simplicity, the next step was done using \( h = 1.2 \), which was the \( h \)-value corresponding to the nearest-neighbor pairs for unlike nodes. For details, see Maren [12].)

As the phase space map for various equilibrium configuration values versus different \( h \)-values becomes worked out, it will be possible to find a corresponding set of \( h \)-values given an initial set of configuration values. It will then be possible to perform free energy minimization on the system \( \tilde{r} \), using different candidate \( h \)-values, to obtain a resultant model \( \tilde{q} \) that provides an acceptable fit to the representational system \( \tilde{r} \).

This mechanism can be used to model the external system \( \tilde{\psi} \) as it moves through different states, with various corresponding \( h \)-values. This means that we would potentially have a means for modeling evolving system dynamics over time. Moreover, the model would be encapsulated into an \( h \)-values trajectory, which would be a relatively simple \( \theta \) model.
Figure 6: (a) The external system $\tilde{\psi}$ has been brought to a free energy minimum for the case where $h = h_1 = 1.2$. Sampling this system provides different inputs to the representational system $\tilde{r}$. We would not directly know the $h$-values corresponding to $\tilde{\psi}$. However, we would trust that the system $\tilde{r}$, taking its configuration values from sensing applied to $\tilde{\psi}$, would also be at equilibrium. Finding the $h$-values for $\tilde{r}$ would give us the parameters for the model $\tilde{q}$, shown in (b). In this particular case, as the full set of $h$-values corresponding to different configuration values still needs to be developed, the system $\tilde{q}$ was devised for illustration purposes by performing free energy minimization on $\tilde{r}$, shown in the previous Fig. 5 for $h = 1.2$. The equilibrium configuration values for $\tilde{\psi}$ and $\tilde{q}$ are shown in (c).
8 Conclusions

“‘What?’ Female was an alien language, but he usually could translate it well enough to understand what was being said. But this was ...”

_Tangled Webs_

Anne Bishop (2008), p. 159 (Hardcover edition).

This Technical Report has served three purposes:

1. Perform a “Rosetta Stone” translation,
2. Describe how the external and representational systems begin as separate entities, each of which can (separately) come to free energy equilibrium, and
3. Introduce a method for system representation that could, within itself, undergo free energy minimization in order to yield a resulting model which could be described using only one or two parameters (the $\theta$ elements of the model $q$).

There has been substantial grumbling within the research community about how difficult it has been to read and understand Karl Friston’s various articles. (See Freed [13] as just one example.) Some of this is notational; Friston has changed his notation subtly - just enough to be difficult for (and perhaps maddening to) the reader throughout his various articles. Yet, a growing sense that he’s presenting a very useful approach is driving more and more researchers to attempt to read his works. (See, e.g., Raviv [14].)

Thus, the first intention of this work has been to perform a “Rosetta Stone” translation between the variational Bayes derivation as given by Beal [4] and the subsequent ones given by Friston [1, 2, 3], with particular attention to how the shifts in notation reflect a move from an envisioning where both the external and representational systems are predicated on the same underlying variables $x_i$ to one in which the external and model states can be separated. This should make Friston’s works more readable to the broader scientific community.

The second intention has been envisioning how this formulation can be used for a scenario in which the external and representational systems begin as separate entities, each of which can (separately) come to free energy equilibrium. (As an example, see Friston and Frith [15] for an example cast in terms of communication and birdsong.) This is important, because we have not typically thought about how various
systems (whether external or representational) need to be expressed in a way that allows free energy minimization. As one of Friston's key points is that free energy minimization underlies crucial processes (including brain processes), we need to have a better understanding of this premise. Some work, such as that done by a team led by Moran and published by Cullen et al. \[16\], has already shown the validity of this approach. That work, however, deals with a very limited and specific state-space model. Further capabilities are needed.

The third intention has been to introduce a means for representing a system, that is, the $\tilde{r}$ component of the system model, using a 2-D cluster variation method approach. This gives us a representation for which we can write a free energy equation, and thus carry out explicit free energy minimization, leading to parameter identification for the free energy-minimized state. This approach has only been briefly sketched; greater expostulation is provided elsewhere and further developmental work is underway.

It is possibly this last intention that will prove the most valuable over time. There is currently a paucity of useful models for which free energy minimization is an inherently appropriate method. Specifically, the well-known Ising model (which has become dearly beloved within deep learning circles) does not offer sufficient richness for more complex system modeling. The 2-D CVM approach allows both for richness in expression and a simplicity in terms of the two parameters that govern this expression.

The impediment thus far has been that the CVM approach has been theoretically obscure, and its practical capabilities so far unknown. In fact, the phase space behavior of this model - in terms of identifying how the activation and interaction enthalpy parameters impact the resulting free energy-minimized states - has not yet been fully mapped out. This is largely a computational problem, somewhat aided and abetted by (limited) analytic solutions. Work on the 2-D CVM is underway, which should make this model available for use in the near future as a means for implementing model systems that can, within their own nature, be free energy-minimized. This will introduce a new kind of modeling capability for a wide range of applications.
Acknowledgments

I am enormously indebted to Karl Friston for a careful, detailed, and thoughtful review, together with several very useful comments that I have included in their entirety, and which I’ve referenced as “personal communications.”

Declaration of No Conflicts

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Code Availability

The Python code for computing the 2-D CVM configurations, together with the corresponding entropy, enthalpy, and free energy values, will available from the author (as soon as the summary paper is published) on GitHub, and also on the author’s website at www.aliannajmaren.com/patents-publications/.

Those who want early (and very rough) code drafts may contact the author at: alianna@aliannajmaren.com. An alternative address is alianna.maren@northwestern.edu, although emails to this address tend to get lost in the swarm of student emails. Those who “Opt-In” at the author’s website (www.aliannajmaren.com) will receive word when code is released, along with word on new experimental results with the 2-D cluster variation method.

A Appendix: Fundamental thermodynamic concepts

In various commentaries, researchers note that the term “thermodynamic free energy,” as used by Friston (op. cit.) does not really correspond to a a true thermodynamic free energy. Similarly, there is a difference in the enthalpy term, as computed and used in Friston’s work (and in others using the variational Bayes method) and in the notion of enthalpy as it is found in statistical thermodynamics.
This Appendix briefly overviews some of the key concepts in statistical (and classical) thermodynamics, so that it is easier to compare the formalisms resulting from the variational Bayes method described in the body of this Report with the corresponding formalisms from statistical thermodynamics, as is commonly known.

Note that the thermodynamic quantities of free energy ($F$), enthalpy ($H$), and entropy ($S$) are all extensive variables; their values are subject to the number of total units in a given system. However, we will work throughout with the reduced variables, $(\bar{F})$, the enthalpy $(\bar{H})$, and the entropy $(\bar{S})$, for which the previous values have been divided through by $Nk\beta T$, where $N$ is the total number of units in the system, $k\beta$ is Boltzmann’s constant, and $T$ is the temperature. This reduces all values to dimensionless quantities which are independent of the size of the systems under consideration. For the remainder of this work, the overhead bar notation on the reduced thermodynamic variables will be dropped; all quantities are understood to be reduced.

\[ F = H - TS, \]  

where $F$, called the free energy, is the energy available to do work, $H$ is the enthalpy (also called the internal energy) of a system, $T$ is the (absolute) temperature, and $S$ is the entropy.

From a statistical thermodynamics perspective, in order to calculate any of these terms, we first need the probabilistic distribution of units in the system among available states. (Several of the following equations are drawn from Salzman [17], who has provided a lucid and cogent introduction. Another useful source, especially for describing how microstates are computed, is found in Maren [18].)

A.1 The classic partition function

To introduce the statistical thermodynamics approach to describing the free energy of a system, we begin with a simple (and classical) expression for the probability of finding a system in the quantum state $i$, characterized by energy $E_i$, as

\[ P_i \propto \exp(-E_i/k\beta T), \]  

We note that the energy $E_i$ describes the energy of the entire system of $n$ units, and that there very well may be degeneracy in the ways in which various units can be assembled to yield a certain energy.
As is true for any probability distribution,

\[ \sum_{i=1}^{n} P_i = 1. \]  \hspace{1cm} (A-3)

In Eqn. [A-2] we stated a proportionality. Now, to find the proportional constant \( c \), we state that

\[ P_i = c \exp\left(-\frac{E_i}{k\beta T}\right), \]  \hspace{1cm} (A-4)

which gives us

\[ \sum_{i=1}^{n} P_i = \sum_{i=1}^{n} c \exp\left(-\frac{E_i}{k\beta T}\right) = c \sum_{i=1}^{n} \exp\left(-\frac{E_i}{k\beta T}\right) = 1, \]  \hspace{1cm} (A-5)

so that

\[ c = \frac{1}{\sum_{i=1}^{n} \exp\left(-\frac{E_i}{k\beta T}\right)}. \]  \hspace{1cm} (A-6)

This sum of probabilities becomes a valuable quantity in and of itself; we refer to it as the partition function, \( Q \) (and in some sources referred to as \( Z \)), so that

\[ Q = \sum_{i=1}^{n} \exp\left(-\frac{E_i}{k\beta T}\right). \]  \hspace{1cm} (A-7)

which allows us to phrase a distinct probability as

\[ P = \frac{\exp\left(-\frac{E_i}{k\beta T}\right)}{Q}. \]  \hspace{1cm} (A-8)

### A.2 The classic enthalpy formulation

Now, we define the average energy \( H \), or enthalpy, of a system to be the expectation for the energy of the system, which can be described as the average of the sum of all the energies of the system,
\[ H = \langle \langle E_i \rangle \rangle \]
\[ = \sum_{i=1}^{n} E_i P_i \]
\[ = \sum_{i=1}^{n} E_i \exp(-E_i/k\beta T)/Q \quad \text{(A-9)} \]
\[ = \frac{1}{Q} \sum_{i=1}^{n} E_i \exp(-E_i/k\beta T). \]

For simplicity, we introduce the notation that \( \beta = 1/k\beta T \), so that

\[ H = \frac{1}{Q} \sum_{i=1}^{n} E_i \exp(-\beta E_i). \quad \text{(A-10)} \]

Now, we notice that we can make a further simplification, by observing that

\[ E_i \exp(-\beta E_i)/Q = -\frac{1}{Q} \frac{\partial}{\partial \beta} \exp(-\beta E_i) \quad \text{(A-11)} \]

so that

\[ E_i \exp(-\beta E_i) = -\frac{\partial}{\partial \beta} \exp(-\beta E_i) \quad \text{(A-12)} \]

Thus, we can rewrite Eqn. (A-10) as

\[ H = \sum_{i=1}^{n} \frac{E_i \exp(-\beta E_i)}{Q} \]
\[ = -\frac{1}{Q} \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \exp(-\beta E_i) \quad \text{(A-13)} \]
\[ = -\frac{1}{Q} \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \exp(-\beta E_i). \]

We notice, from Eqn. (A-7) that the entire of the summation in Eqn. (A-13) is exactly \( Q \) itself. Thus, we can rewrite the expression for \( H \) in Eqn. (A-13) as

\[ H = -\frac{1}{Q} \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \exp(-\beta E_i) = -\frac{1}{Q} \frac{\partial}{\partial \beta} Q. \quad \text{(A-14)} \]
We recall the expression for the derivative of a logarithmic function, that
\[
\frac{\partial}{\partial x} \ln(y(x)) = \frac{1}{y} \frac{\partial y(x)}{\partial x}. \quad (A-15)
\]

Thus, noticing the correspondence between \( Q = Q(\beta) = y(x) \), we can identify that
\[
\frac{\partial}{\partial \beta} \ln(Q(\beta)) = \frac{1}{Q} \frac{\partial Q(\beta)}{\partial \beta}. \quad (A-16)
\]

This allows us to rewrite Eqn. A-14 as
\[
H = -\frac{1}{Q} \frac{\partial}{\partial \beta} Q = -\frac{\partial}{\partial \beta} \ln(Q(\beta)) = -\frac{\partial}{\partial \beta} \ln(Q), \quad (A-17)
\]
dropping the functional dependence of \( Q \) on \( \beta \) in the last expression.

This gives us a very powerful means for expression the enthalpy, or internal energy, of a system in terms of \( \ln(Q) \). However, it is not a completely general expression for \( H \). Instead, we note that by taking the (negative of) the derivative of \( \ln(Q) \) with respect to \( \beta \), we are moving to a value of \( \ln(Q) \) that maximizes (minimizes) the function in terms of temperature (actually, \( 1/T \)). This then leads us to the value of enthalpy, \( H \), that helps to minimize the overall value for the free energy, \( F \). The value of \( H \) that would occur in this case (taking into account the impact of entropy), would then be the expected value of \( H \) when the system is at equilibrium.

What we can generalize from this is not so much the specific formulation for \( H \) (as a partial derivative of \( \ln(Q) \) with respect to \( \beta \), but rather, that we are looking for an expected value of \( H \). Depending on how the system is formulated, it could be a different resulting expression.

\section*{B Appendix: Variational Free Energy: Enthalpy and Entropy}

This Appendix provides more detail on the derivation of the free energy equation as given in Friston \cite{2, 3}, using the formulation established by Beal \cite{4}, and making the correspondence between the notations of each. It yields the expression of the free energy in terms...
of what Friston refers to as “thermodynamic free energy” plus an entropy term. Note that what Friston refers to as “thermodynamic free energy” is actually the expectation for the log-likelihood of a certain distribution, as given in Eqn. 30, which is restated here as

\[ L(\tilde{x}) = L(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}) = -\ln p(\tilde{\psi}, \tilde{s}, \tilde{a}, \tilde{r}), \]

Note that if we take this equation directly from Beal, we will have a summation. Friston, however (personal communication), suggests that we “would delete the summation unless you specify the factorization over the (continuous variables). Note that these variables are necessarily continuous because they are in generalized coordinates of motion.”

We will, very shortly, follow a line of argument introduced by Beal. One of the steps that he makes at a certain conclusion gives a form of the enthalpy equation, which is fundamentally the same as used by Friston (op. cit.), and is the one that we have shown in Section 5. It will be that

\[ H = \sum_{i=1}^{\infty} \int dx_i q_{x_i}(x_i) \ln p(x_i, y_i | \theta). \tag{B-1} \]

In the next few paragraphs, we will show that Eqn. [B-1] serves analogously to Eqn. [A-17]. It fits the role of enthalpy in what is being described as a “free energy” formalism. However, it is not a derivative equation; it is the expectation for the natural logarithm of the probability of certain variables.

By examining the correspondence between the two expressions (the variational Bayes in comparison with the thermodynamic), when we come to expressing the full free energy for the representational system (using the formulation expressed by Beal), the free energy derivation will be much more lucid.

Beal [4] (Sect. 2.2.1) introduces the scenario for parameter learning with the notion of a generative model with hidden variables \(x\) and observed variables \(y\), where the dataset of observed variables \(y = \{y_1, ..., y_n\}\) are generated by the set of hidden variables \(x = \{x_1, ..., x_n\}\), and where the \(n\) items in each case are independent and identically distributed (i.i.d.). He identifies the parameters describing the (potentially) stochastic dependencies between variables as \(\theta\). The probability distribution for observing \(y\) is given as
\[ p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \int dx_i p(x_i, y_i|\theta). \quad (B-2) \]

This tells us that the probability distribution of the observable variables \( y_i \) is conditioned on the parameters \( \theta \).

We compare this with our earlier expression for the probability that a system would be in the \( i^{th} \) energy state, where we stated in Eqn. A-7 that the partition function \( Q \) (which served as a normalizing factor) is given as

\[ Q = \sum_{i=1}^{n} \exp(-E_i/k_\beta T), \]

and from Eqn. A-8 that probability is given as

\[ P = P_i = \exp(-E_i/k_\beta T)/Q. \]

The difference in the two formulations is that, in Eqn. A-8, we are dealing with the probability of finding a system in a given \( i^{th} \) energy state, where there could indeed be multiplicity of units inhabiting various components of that overall system’s energy state, and there can also be degeneracy in the units inhabiting the various states. This means, various units, indistinguishable from each other, can inhabit the energy states so that there are multiple means of counting the units in various states and coming up with the same overall \( i^{th} \) system energy.

In contrast, the formulation given in Eqn. B-2 deals with a full set of observed variables, \( y = \{y_1, ..., y_n\} \).

As we know from the formulation of joint probabilities, the joint probability of observing variable \( y_1 \) in a given state \( i \), \( p(y_1 = i) \), together with the probability of observing variable \( y_2 \) in its own given state \( j \), \( p(y_2 = j) \), is \( p(y_1 = i)p(y_2 = j) \). Thus, for a set of such states, \( y = \{y_1, ..., y_n\} \), we need the product of each of the unique variable’s probability, which yields the first part of Eqn. B-3

\[ p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta). \quad (B-3) \]

As a second step in making the correlation between Eqn. B-2 and the combination of Eqns. A-7 and A-8, we examine the formulation (from Eqn. B-2)
\[ p(y_i|\theta) = \int dx_i p(x_i = y_i|\theta). \] (B-4)

We note that Eqn. (B-4) essentially states that the observable variable(s) \( y_i \) are dependent on a hidden, unobservable set of variables \( x_i \). For each distinct value (or set of values) \( y_i \), there are potentially multiple contributions, drawn over the complete set of \( x_i \). The notation here (keeping to that used by Beal) is just a little ambiguous; it does not mean that the set of countable variables \( x_j \) is exactly the same as the set of countable variables \( y_i \). Rather, it is saying that for a given instance of a set of variables \( y_i \), there is a corresponding set of variables \( x_i, x_i \equiv \{ X_i \} = \{ x_{i,1}, x_{i,2}, ..., x_{i,J} \} \), where \( J \) is the total number of elements in the set of variables \( \{ X_i \} \) contributing to values for \( y_i \).

This means, since each of these contribute to the value for \( y_i \), that summation (rather than multiplication) is needed. Since the realm of variables constituting \( \{ X_i \} \) is assumed here to be large, integration rather than summation is used. This, then, gives us a plausibility argument in support of Eqn. B-2.

Before moving on, we need one more step. Eqn. (B-1) involves more than the probability distribution of Eqn. B-2. We note that in Eqn. B-1 we have a summation rather than a product, the natural logarithm of \( p(x_i,y_i|\theta) \) as opposed to just a simple statement of the probability, and an additional modulating term, \( q_{x_i}(x_i) \).

To see how these occur, we go back to our original definition of the enthalpy, \( H \), as given in Eqn. (A-17):

\[ H = -\frac{1}{Q} \frac{\partial}{\partial \beta} Q = -\frac{\partial}{\partial \beta} \ln(Q(\beta)) = -\frac{\partial}{\partial \beta} \ln(Q). \]

Suppose that we have a probability distribution given as

\[ p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta). \] (B-5)

Then, the natural logarithm of this distribution is given as

\[ \ln p(y|\theta) = \ln \prod_{i=1}^{n} p(y_i|\theta) = \sum_{i=1}^{n} \ln p(y_i|\theta). \] (B-6)

We do not have agreement between the formulation given in Eqn. B-1 and that given in Eqn. (A-13) together with Eqn. (A-14).
Moving on, we now turn our attention to the derivation for the free energy as introduced in Beal [4].

Beal introduces a formulation for the log likelihood in his Eqns. 2.12 - 2.16, reproduced here:

\[
L(\theta) = \sum_{i=1}^{n} \ln \int dx_i \, p(x_i, y_i|\theta) \\
= \sum_{i=1}^{n} \ln \int dx_i \, q_{x_i}(x_i) \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \\
\geq \sum_{i=1}^{n} \left( \int dx_i \, q_{x_i}(x_i) \ln \frac{p(x_i, y_i|\theta)}{q_{x_i}(x_i)} \right) \\
= \sum_{i=1}^{n} \left( \int dx_i \, q_{x_i}(x_i) \ln p(x_i, y_i|\theta) - \int dx_i \, q_{x_i}(x_i) \ln q_{x_i}(x_i) \right) \\
= \sum_{i=1}^{n} \left( \int dx_i \, q_{x_i}(x_i) \ln p(x_i, y_i|\theta) - \int dx_i \, q_{x_i}(x_i) \ln q_{x_i}(x_i) \right) \\
\equiv F(q_{x_1}(x_1), ..., q_{x_n}(x_n), \theta).
\]

Beal notes that \(F(q_{x}(x), \theta)\) is a lower bound on \(L(\theta)\) and is a functional of the free distributions \(q_{x_i}(x_i)\) and of \(\theta\) (the dependence on \(y\) is left implicit). The inequality introduced in the third expression makes use of Jensen’s inequality.

Beal notes: “Defining the energy of a global configuration \((x, y)\) to be \(\ln(p(x, y|\theta))\), the lower bound \(F(q_{x}(x), \theta) \leq L(\theta)\) is the negative of a quantity known in statistical physics as the free energy: the expected energy under \(q_{x}(x)\) minus the entropy of \(q_{x}(x)\) (Feynman, 1972; Neal and Hinton, 1998).”

Thus, both \(F(q_{x}(x), \theta)\) and \(L(\theta)\) refer to free energy formalization, however, \(L(\theta)\) can be greater than \(F(q_{x}(x), \theta)\). Alternatively, we can say that \(F(q_{x}(x), \theta)\) is a lower bound on \(L(\theta)\).

Essentially (in Beal’s conception; Friston’s is reversed), we’re saying that the free energy of the model is going to be lower than the free energy of the system being modeled; \(L \geq F\). If we are going to improve the accuracy of the model, we will be bringing \(F(q_{x}(x), \theta)\) towards \(L(\theta)\).

Beal further notes that \(F(q_{x}(x), \theta)\) is the negative of what is known, in statistical thermodynamics, as the free energy of a system, which is the expected energy \((H)\) under \(q_{x}(x)\) minus the entropy of
Thus, when we shift to the notation of Friston (op.cit.), we will reverse the signs on all of the terms on the right-hand-side of Eqn. (second-to-last line of the equation), leading up to Beal’s definition of $F(q_x(x), \theta)$.

According to this understanding, and changing the signs of the terms to deal with Beal’s use of $F(q_x(x), \theta)$ as the negative of the free energy, we have the expected energy (or enthalpy) of a system $H$ is given as

$$H = - \sum_{i=1} dx_i q_{x_i}(x_i) \ln p(x_i, y_i|\theta), \quad \text{(B-8)}$$

and

$$S = - \sum_{i=1} dx_i q_{x_i}(x_i) \ln q_{x_i}(x_i). \quad \text{(B-9)}$$

The definition for the entropy of a system is given as

$$S = -k_B \ln \Omega. \quad \text{(B-10)}$$

An alternative formulation is

$$S = -k_B \langle \ln P \rangle, \quad \text{(B-11)}$$

where $P$ is the probability distribution over all states. Thus, $S$ is the expectation of the probability of the variable for which various states are possible; i.e., this correlates directly with Eqn. (B-9) and where $\Omega$ is the grand partition function, and $k_B$ is Boltzmann’s constant.

From Feynmann (1972), Eqn. 1.1, that

$$Q = \sum_n \exp(-E_n/kT) \equiv \exp(-F/kT). \quad \text{(B-12)}$$

From Feynmann (1972) we have the definition for $F$, the Helmholtz free energy, as

$$F = -kT \ln Q = -kT \ln \left( \sum_n \exp(-E_n/kT) \right) \quad \text{(B-13)}$$

and also (Eqn. 1.3)

$$S = -k \sum_n p_n \ln p_n, \quad \text{(B-14)}$$
where (Eqn. 1.4)

\[ P_n = \frac{1}{Q} \exp(-E_n/kT). \]  

(B-15)

Further, the average energy per unit, \( U \) (also known as the enthalpy), is given as (Eqn. 1.7)

\[ U = \frac{1}{Q} \sum_n E_n \exp(-E_n/kT), \]  

(B-16)

which can be rewritten as

\[ U = \sum_n E_n P_n. \]  

(B-17)

Essentially, Eqn. B-17 states that the enthalpy of a system (the total energy of the system), is simply the sum of the energy-per-state, multiplied by the probability that a unit will be in that state, where the whole is summed over all the units in the system.

This derivation by Feynmann is perhaps a bit more intuitive than that put forth in Appendix A, and may be preferred.

References

[1] K. Friston, “The free-energy principle: a unified brain theory?,” Nat. Rev. Neurosci., vol. 11, p. 127138, 2010. doi:10.1038/nrn2787.

[2] K. Friston, “Life as we know it,” Journal of The Royal Society Interface, vol. 10, no. 86, 2013.

[3] K. Friston, M. Levin, B. Sengupta, and G. Pezzulo, “Knowing ones place: a free-energy approach to pattern regulation,” J. R. Soc. Interface, vol. 12, p. 20141383, 2015. doi:10.1098/rsif.2014.1383; available online at: [http://dx.doi.org/10.1098/rsif.2014.1383](http://dx.doi.org/10.1098/rsif.2014.1383)

[4] M. J. Beal, Variational algorithms for approximate Bayesian inference. PhD thesis, University College London, 2003. PDF: [http://www.cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf](http://www.cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf).

[5] R. P. Feynman, Statistical Mechanics. Reading, MA: Benjamin, 1972, 1998.

[6] G. E. Hinton and D. van Camp, “Keeping neural networks simple by minimizing the description length of weights,” in Proc. of COLT-93, p. 513, 1993. doi:10. 1145/168304.168306.
[7] A.J. Maren, “The cluster variation method: a primer for neuroscientists,” Brain Sciences, vol. 6, no. 4, p. 44, 2016.

[8] B. Sengupta, M. B. Stemmler, and K. J. Friston, “Information and efficiency in the nervous system—a synthesis,” PLoS Comput Biol., vol. 9, no. 7:e1003157, 2013. doi:10.1371/journal.pcbi.1003157. PMID:23935475. PMCID:PMC3723496. Epub 2013 Jul 25.

[9] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians.” arXiv:1601.00670v4 [stat:CO], 2 Nov 2016.

[10] R. Kikuchi, “A theory of cooperative phenomena,” Phys. Rev., vol. 988, no. 81, p. 127138, 1951.

[11] R. Kikuchi and S.G. Brush, “Improvement of the cluster variation method,” J. Chem. Phys., vol. 47, p. 195, 1967.

[12] A.J. Maren, “2-D cluster variation method: Comparison of Computational and Analytic Results,” June, 2010. GitHub: github.com/ajmaren/2D-Cluster-Variation-Method; MS (TM) PPT Slidedeck: 2D-CVM-Expts-vary-eps0-and-eps1-computational-2019-06-17v4.pptx.

[13] P. Freed, “Research Digest,” Neuropsychoanalysis, vol. 12, no. 1, p. 103-106, 2010. doi:10.1080/15294145.2010.10773634; https://doi.org/10.1080/15294145.2010.10773634.

[14] S. Raviv, “The genius neuroscientist who might hold the key to true AI,” Wired, vol. 11, p. 127138, November, 2018. https://www.wired.com/story/karl-friston-free-energy-principle-artificial-intelligence/.

[15] K.L. Friston and C. Frith, “A duet for one,” Consciuosness and Cognition, vol. 36, p. 390-405, Nov., 2015. doi:10.1016/j.concog.2014.12.003.

[16] M. Cullen, B. Davey, K.J. Friston, and R.J. Moran, “Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness,” Biological Psychiatry CNNI, vol. 3, no. 9, p. 809-818, September, 2018. doi:10.1016/j.bpsc.2018.06.010

[17] W. Salzman, Course Materials for Chemistry 480B - Physical Chemistry, ch. An introduction to partition functions. 2001. Published online at: [http://cbc.arizona.edu/~salzmanr/480b/480b.html](http://cbc.arizona.edu/~salzmanr/480b/480b.html)
[18] A.J. Maren, *Statistical Mechanics, Neural Networks, and Artificial Intelligence: The Précis*, In progress. Available for download at: [http://www.aliannajmaren.com/book/](http://www.aliannajmaren.com/book/).