Minding Your P's and Q's: Notational Variations Expressing the Kullback-Leibler Divergence

Alianna J. Maren[1,2,†]

[1] Themesis, Inc.
3-2600 Kaumualii Hwy Ste 1300 PMB 188, Lihue, HI 96766

[2] Northwestern University School of Professional Studies
Master of Science in Data Science Program
633 Clark St, Evanston, IL 60208

† Address to which correspondence should be addressed:
themesis1@themesis.com

Themesis Technical Note THM TN2024-001 (ajm)

Date: 2024-08-11
Version 1.0

Abstract

The Kullback-Leibler (K-L) divergence has become foundational in the machine learning community, leading to various authors expressing the same formula with differing notations. This can cause difficulties for those attempting to trace the same (or similar) lines of thought across various research papers. This work addresses the different notation forms used by various authors, e.g., Matthew Beal and David Blei et al. working with variational inference, and with particular attention to the early works (2013-2015) of Karl Friston and colleagues. We briefly identify notation used by Diederik Kingma and Max Welling, working with variational autoencoders. We also address how the *reverse* K-L divergence plays a leading role in all generative methods. We briefly note on how, in Action Perception Divergence (an evolution of active inference), the reverse K-L divergence uses a "target" rather than a specific data representation, and compares how various action steps lead to distributions which can be compared against this target, in order to select a preferred action. In addition, we briefly summarize the K-M (Kikuchi-Maren) divergence, which useful when applying the cluster variation method (CVM) as a model in variational methods.

Keywords: Kullback-Leibler divergence, K-L divergence, KL divergence, Kikuchi-Maren divergence, K-M divergence, KM divergence, cluster variation method, variational inference, variational Bayes, notation, active inference, Action Perception Divergence.

1 Introduction and Overview

One challenge in reading both the classic and recent works in energy-based neural networks and machine learning (especially variational methods) is that concepts such as the Kullback-Leibler (K-L) divergence are often expressed using different notation. This requires the reader to track the precise meaning of each variable in each different work. This task can be more difficult when variable meanings are reversed.

A particular point of interest is that in the works by Blei et al. (2016, 2017) [1, 2] and Beal (in his 2003 dissertation; [3]), an important variable means one thing in one work, and something different in the other. Further, the notation used in early works by Karl Friston (2010, 2013) [4, 5] and in Friston et al. (2015) [6] is exceptionally complex, as it attempts to describe how a given system interacts with a representation of that system through a Markov blanket.

One of the most important things that we address in this work is the essential distinction between the "typical" K-L divergence and the *reverse* K-L divergence, in which we compare various models against a *representation* of the data or an observed systeml. To clarify this latter point, we give special attention to a subtle but significant shift in notation used in Action Perception Divergence (APD), devised by Hafner et al. (2020, rev. 2022) [7].

The reverse K-L divergence is important in all areas of generative artificial intelligence, including variational autoencoders, as developed by Kingma and Welling (2013, 2019) [8, 9]. Thus, to provide continuity in our treatment of the notation used for the reverse K-L divergence across applications, we briefly address that used by Kingma and Welling.

Finally, Maren (2022) [10] has developed a specific divergence measure (the Kikuchi-Maren divergence) that is akin to the KL divergence, but which is specific for use with Kikuchi's cluster variation method (CVM). This method, which underlies the CORTECON(R) (COntent-Retentive, TEmporally-CONnected neural network) computational engine, plays an essential role in enabling signal-to-symbol connections for artificial general intelligence (AGI).

Thus, this *Technical Note* serves three purposes:

- Review of how the ***reverse*** Kullback-Leibler divergence, rather than the simple Kullback-Leibler divergence, is commonly used in deriving the algorithms for energy-based neural networks and machine learning methods such as variational inference,

- Identify the notational differences used by various key authors, by presenting a table of notation that serves as a *Rosetta stone* for cross-comparisons, and
- Introduce the Kikuchi-Maren (K-M) divergence method proposed by Maren (2022) [10] that allows the cluster variation method to be used as a model within a variational context.

A further advantage to the reader is that, through extensive quotation from original sources, the reader's time can be effectively focused on cross-comparisons of material within this single document, rather than separately finding and reading each of the relevant source works.

The following Table 1 presents a glossary of the thermodynamic terms used in this Report.

Table 1: Thermodynamic Terms

Variable	**Meaning**
Activation enthalpy	Enthalpy ε_0 associated with a single unit (node) in the "on" or "active" state (**A**); influences configuration variables and is set to 0 (for this work) in order to achieve an analytic solution for the free energy equilibrium in the cluster variation method
Configuration variable(s)	Nearest neighbor, next-nearest neighbor, and triplet patterns
Degeneracy	Number of ways in which a configuration variable can appear
Enthalpy	Internal energy H results from both per unit and pairwise interactions; often denoted H in thermodynamic treatments
Entropy	The entropy S is the distribution over all possible states; often denoted S in thermodynamic treatments and H in information theory

2 Kullback-Leibler Notation Used by Various Authors

The following Table 2 presents a "Rosetta Stone" of the differing notations as used by Beal (2003) [3], Friston and colleagues ([5, 6]), and Blei et al. (2016, 2017) [1, 2], together with that of Kingma and Welling (2013, 2019) [8, 9].

An important note on the historical evolution of the Friston et al. notation (op. cit.) as compared with that of Beal is that Matthew Beal wrote his dissertation on variational inference in 2003. Friston referenced Beal in his own works, and heavily adopted Beal's notation ([5, 6]).

In order to understand Friston's works, the most relevant starting place is Beal's dissertation, since Beal presents derivations for the various equations, whereas Friston (and colleagues) simply present their final form.

Also, it is worth noting that in recent years, Friston (in concert with other authors) has changed both his emphasis and his notation, so the notation presented here corresponds to Friston's work in the 2013-2015 timeframe; variations appeared after that.

Friston and Beal both deal with the role of Markov blankets. In this work, we use only the simpler form of Beal's notation. However, it is central to Friston's work, and so we include it in our discussion here.

Table 2: The ***Rosetta Stone***: Notation from Beal, Friston, Blei et al., and Kingma and Welling

Variable / Notation	**Beal**	**Friston**	**Blei et al.**	**Kingma & Welling**
Observable Variable; *Dependent or (Friston) "Internal States"*	y_i	$\lambda, \tilde{r}$	x_i	x
Hidden Variable; *Independent, Latent, or (Friston) "External States"*	x_i	$\tilde{\Psi}$	z_i	z
Markov "sensing" units *(Friston)*	-	$\tilde{s}$	-	-
Markov "active" units *(Friston)*	-	$\tilde{a}$	-	-
Model parameters	θ	m	-	θ
Model distribution	$q(x)$ (1)	$q(\Psi\|\lambda)$ (2)	-	-
Observations distribution	$p(y\|\theta)$ (3)	$p(\Psi, s, a, r\|m)$ (4)	-	-
Variational free energy	-	$F(\tilde{s}, \tilde{a}, \tilde{r})$	-	-

The authors specifically identify their notation, according to the following enumerated points (corresponding to elements of Table 2):

1. **Observations distribution - Beal:** $p(y|\theta)$: " ... [the] generative model that produces a dataset $y = \{y_1, ..., y_n\}$ consisting of n independent and identically distributed (i.i.d.) items, generated using a set of hidden variables $x = \{x_1, ..., x_n\}$ such that the likelihood can be written as a function of θ ..." (Beal, 2003, p.46, Eqn. 2.9),
2. **Observations distribution - Friston:** $p(\Psi, s, a, r|m)$: "... ergodic density $p(\Psi, s, a, r|m)$ [is] a probability density function over external $\psi \in \Psi$, sensory $s \in S$, active $a \in A$ and internal states $\lambda \in \Lambda$ for a system denoted by m" (Friston, 2013, p. 2, Table 1),
3. **Model distribution - Beal:** $q_{x_i}(x_i)$: "we use a distinct distribution $q_{x_i}(x_i)$ over the hidden variables ..." (Beal, 2003, p. 47, just before Eqn. 2.12), and
4. **Model distribution - Friston:** $q(\Psi|\lambda)$: " ... a probability density over external states $q(\Psi|\lambda)$ that is encoded (parametrized) by internal states." (Friston, 2013, p. 4, just before Lemma 2.1).

We will examine how each author forms their respective notations in more detail later in this work.

3 The Kullback-Leibler Divergence: Original Work

The expression that interests us the most is taken directly from Kullback and Leibler's original work ([11]).

They state, "We shall denote by $I(1:2)$ the mean information for discrimination between H_1 and H_2 per observation from μ_1; i.e. *[in Eqn. 2.4 of their work]*

$$
\begin{aligned}
I(1:2) = I_{1:2}(X) = & \\
& = \int d\mu_1(x) \log \frac{f_1(x)}{f_2(x)} \\
& = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x)."
\end{aligned}
\tag{1}
$$

where the terms are originally defined by Halmos and Savage [12], specifically:

- **The data,** X: [There is a] " set X of objects x, to be thought of as possible outcomes of an experimental program, distributed according to an unknown one of a certain set of probability measures" (see Halmos and Savage, p. 225).
- **There exist hypotheses** $H_i; i = 1, 2$; where the hypothesis is that x was selected from the population whose probability measure is $\mu_i, i = 1, 2$.
- **There is a probability measure** λ such that $\lambda \equiv \{\mu_1, \mu_2\}$; i.e., λ is some linear combination of μ_1 and μ_2.

Returning to the original Kullback-Leibler work, we read:
"Further, we define the functions f_i as

$$\mu_i(E) = \int_E f_u(x) d\lambda(x), \tag{2}$$

where E is a subset of X."

While this expression is couched very abstractly, in succeeding works by others and in practical applications, μ_1 has come to be thought of as the probability of observing some actual real data.

4 Recent Notation for the Kullback-Leibler Divergence

The Kullback-Leibler divergence was designed to measure the difference, or "divergence," between any two sets that have the same support basis. These could be two data sets, taken over the same interval and with the same sampling. This could also be a divergence between a set of data observations and the predictions made by a model.

The notation proposed by Kullback and Leibler is abstract, and does not indicate a preference for either of these two uses for the divergence measure.

Very commonly in recent literature, the notation used is the p and q notation, where p represents the probability with regard to the actual data, and q represents the model-based probability of occurrence of an observation. Thus, for example, we see the Kullback-Leibler divergence is used in an autoencoder in Balesdent et al. (2016) [13]. In their work, we see the K-L divergence expressed as:

"Let P and Q be two probability distributions defined by their pdf *[probability distribution function]* p and q with support R^d. The Kullback-Leibler divergence between P and Q is defined by

$$\text{“}D_{KL}(P,Q) = \int_{R^d} ln\left(\frac{p(x)}{q(x)}\right) p(x)dx.\text{”} \tag{3}$$

Similarly, the K-L divergence is expressed by [14] (Eqn. 2.161) as

$$\text{“}KL(p||q) := \int_{-\infty}^{\infty} p(x)ln\left(\frac{p(x)}{q(x)}\right) dx.\text{”} \tag{4}$$

Even the Wikipedia (although not typically regarded as an acceptable source for citations) uses the same notation, as illustrated first for the continuous case [15]:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x)ln\left(\frac{p(x)}{q(x)}\right) dx. \tag{5}$$

Similarly, the notation presented for the discrete case is:

$$D_{KL}(P||Q) = \sum_{x\epsilon X} P(x)ln\left(\frac{P(x)}{Q(x)}\right) dx. \tag{6}$$

Thus, the casual (if voracious) reader might be forgiven for assuming that, throughout the literature, P referred to an actual data set, whereas Q referred to the model, and that the divergence was always taken of the data P against the model Q.

It may come as a surprise to the unwary that in the realm of energy-based neural networks and variational inference, instead of the typical Kullback-Leibler divergence, we use the ***reverse*** Kullback-Leibler divergence.

A typical form for expressing the *reverse* KL divergence is

$$D_{KL}(Q||P) = \sum_{x\epsilon X} Q(x)ln\left(\frac{Q(x)}{P(x)}\right) dx. \tag{7}$$

This means that the distribution that we are comparing against a given "reference" is the ***model***; we are testing various models to see how well they minimize the divergence. We will address this later in this work.

Occasionally, we will see that the numerator and denominator within the logarithmic term are interchanged, and there is a "minus" sign in front of the equation, so that

$$D_{KL}(Q||P) = -\sum_{x \epsilon X} Q(x) ln \left(\frac{P(x)}{Q(x)} \right) dx. \tag{8}$$

When this is done, it is to set up the reverse K-L divergence for later steps in which a subsequent resulting equation will formally resemble a free energy equation from statistical mechanics.

5 A Brief Mathematical Digression

This section is directed towards those who are relatively new to work with the K-L divergence. More experienced readers can skip this section completely.

It will be easier if we imagine that our various probabilities and model predictions are taken in discrete units, so we consider the discrete case equation for the K-L divergence, given previously as Eqn. 6.

We can ask ourselves: what would happen if the multiplier in front of the logarithm of the difference term were the model probability, and not the data probability?

We will compare the two different formulations, one where the multiplier is $p(x)$ and the other where it is $q(x)$. (The first is the K-L divergence, and the second is a made-up contrary equation, just to show the difference.)

In one sense, it might make sense to multiply by the *expected* data value, that is, the one given by the model. That would (presumably) avoid all sorts of bumps in the data distribution.

5.1 The KL Divergence as It Is: $p(x)$ Is Multiplier

We begin by looking at a graph of the logarithm function, as shown in Figure 1.

We are interested in the area around $x = 1$, and we know that $ln(1) = 0$.

Let's consider two cases, one where $p(x) > q(x)$, and the other where $p(x) < q(x)$.

Case 1: $p(x) > q(x)$

In this first case, the probability term $p(x)$ is a bit more than we are expecting; $p(x) > q(x)$.

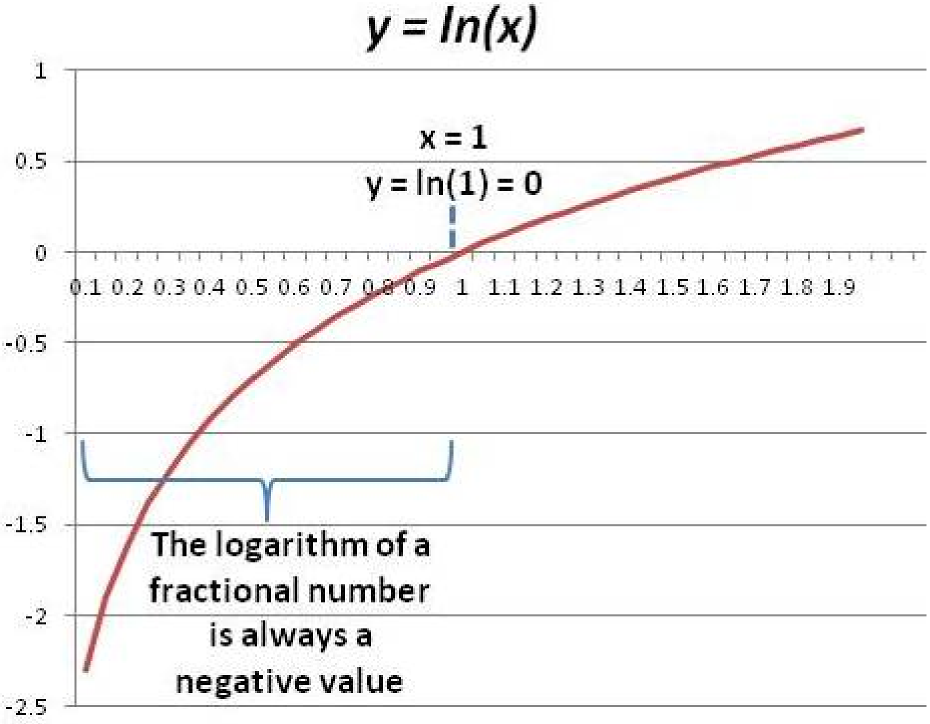


Figure 1: The logarithm function, plotted for the neighborhood where $x = 1$.

In this situation, the ratio of $p(x)/q(x)$ is a number greater than one; $p(x)/q(x) > 1$, which means that the logarithm of this ratio is both *positive* and *relatively small*, as the slope for the logarithm in the neighborhood of small numbers that are greater than one is relatively low.

Thus, we'll have a slightly-larger-than-expected number multiplying a logarithmic result that is not zero, but is also relatively small - given the small slope in that region of the curve.

Case 2: $p(x) < q(x)$

Let's consider the other case, where $p(x) < q(x)$. In this case, $p(x)$ is smaller than that anticipated by our model. The ratio of $p(x)/q(x) < 1$, so we take the logarithm of a fraction, which gives us a negative number.

Not only do we get a negative number, but as $p(x)$ diverges more and more from $q(x)$ ($p(x) << q(x)$), we get a smaller fraction value,

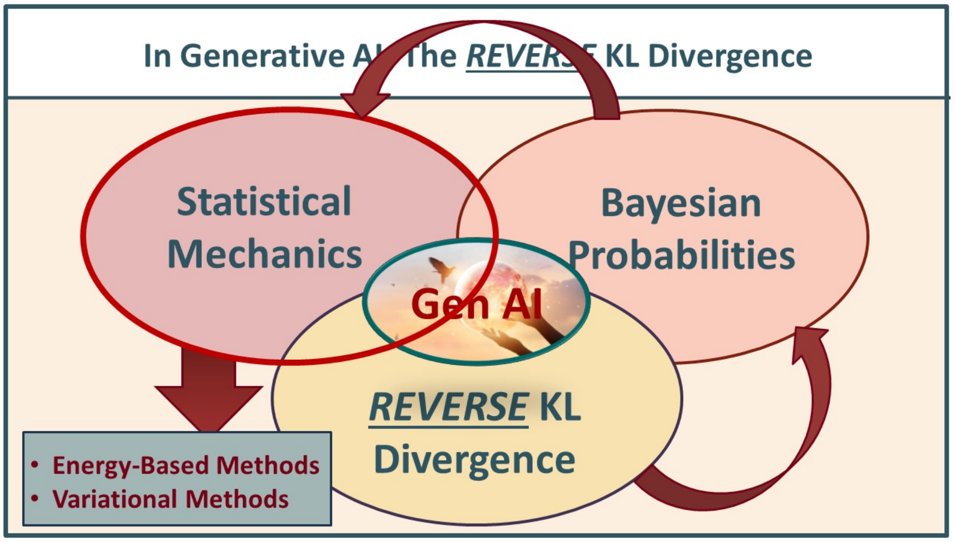


Figure 2: Three major disciplines or lines of thought converge in creating the framework for variational inference: the reverse K-L divergence, Bayesian probabilities, and statistical mechanics.

and the logarithm of that ratio becomes a progressively larger negative number; the slope increases as $p(x)$ gets smaller relative to $q(x)$.

The logarithm of that ratio may be getting larger, but it is being multiplied by $p(x)$, which is smaller - so the effects of the two numbers tends to cancel out.

In short, if $p(x)$ is greater than expected, its impact is minimized because it multiplies a value that is (relatively) small.

If $p(x)$ is less than expected, it multiplies a number that is larger (in magnitude) than what would be the case if $p(x)$ diverged from $q(x)$ in the other direction - but again, the effect is minimized because $p(x)$ is a smaller-than-expected value.

In sum, when $p(x)$ is the multiplier, the impacts of deviance from the expected $q(x)$ are minimized, either way.

5.2 If the KL Divergence Were Different: If $q(x)$ Was Multiplier

In contrast, if $q(x)$ were to be the multiplier, we would have exactly the opposite - the impact of divergence of the observed probability $p(x)$

from the expected or model probabiliyt $q(x)$ would be exaggerated. (Showing this is left as an exercise for the reader.)

In sum, the construction of the K-L divergence makes intuitively good sense.

6 Using the Reverse K-L Divergence

The goal of this section is to illustrate how three different authors - Matthew Beal, Karl Friston (and colleagues), and Blei et al. - use similar (but still distinctive) notation for their work in variational inference. (In the case of Friston and colleagues, the work is on active inference.)

To do this, we offer figures drawn from a 2024 Themesis YouTube video [16] that presents the reverse K-L divergence.

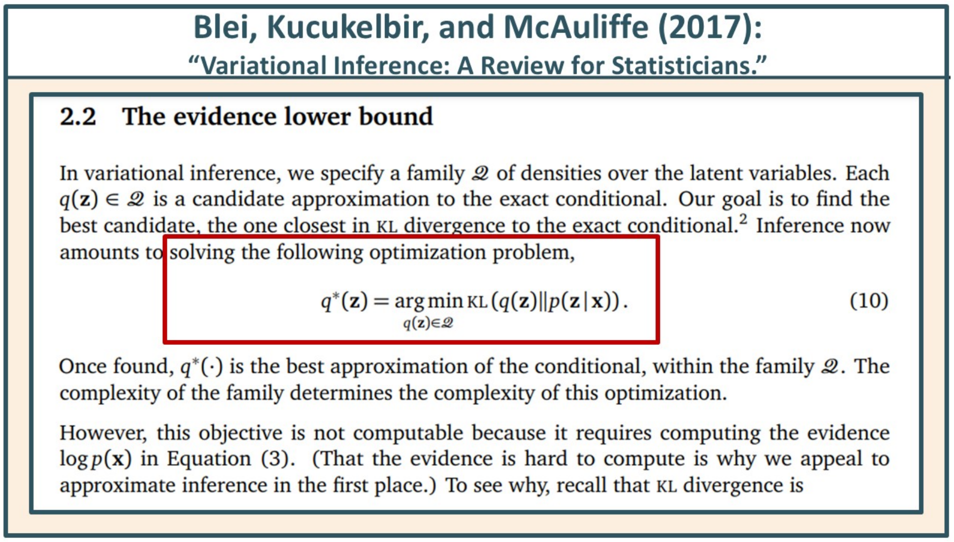
The primary organizing concept is that the reverse K-L divergence is the entry point to understanding energy-based methods and variational inference. Once we establish this entry point, we use conditional Bayes to rewrite the conditional probablity within the divergence equation. Once we've done that, mathematical manipulations give us a result that is isomorphic with a statistical mechanics-based free energy equation. This is shown in Figure 2.

One of the most straightforward starting points is with David Blei and colleagues, who have written an excellent tutorial on variational inference [1, 2]. Blei et al. offer a perspective that is a bit more clearly stated than found in Beal's dissertation. Thus, we include Blei et al. in our studies. (Note: Blei et al. was published to *arXiv* in 2017, although original journal publication date was earlier.)

Blei's notation for the reverse K-L divergence is shown in Figure 3.

Blei and Beal have contradictory notations. Specifically (see Table 2), Beal uses x_i for his hidden variables, and y_i for his observable variables. In contrast, Blei et al. use z_i for the hidden variables, and x_i for the observables. Keeping track of the meaning of x_i - from one paper to another - takes extra effort. Thus, this work serves as a simple cross-reference, or "Rosetta stone."

Friston's work is couched using the notation introduced by Beal, in his dissertation on variational Bayes [3].



Blei, Kucukelbir, and McAuliffe (2017):
"Variational Inference: A Review for Statisticians."

2.2 The evidence lower bound

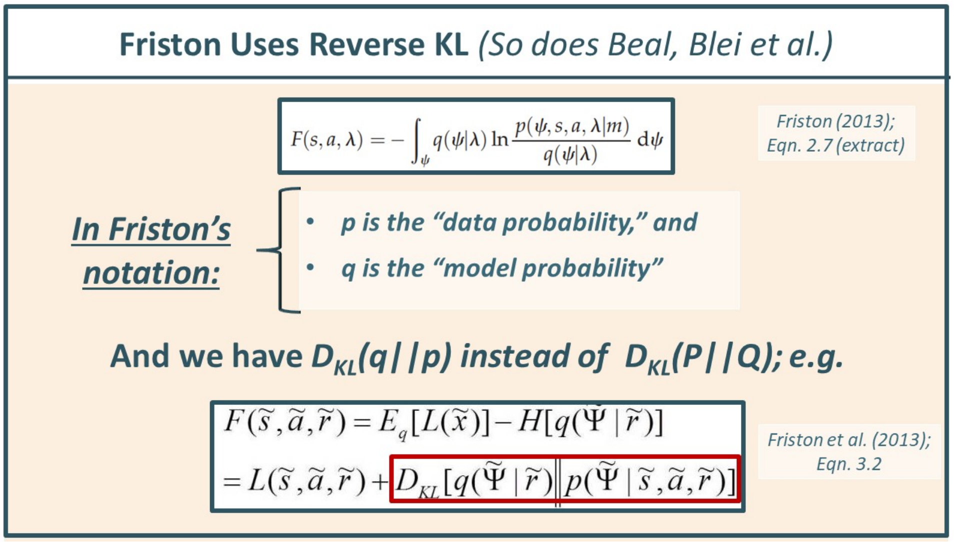
In variational inference, we specify a family $\mathscr{Q}$ of densities over the latent variables. Each $q(\mathbf{z}) \in \mathscr{Q}$ is a candidate approximation to the exact conditional. Our goal is to find the best candidate, the one closest in KL divergence to the exact conditional.[2] Inference now amounts to solving the following optimization problem,

$$\boxed{q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})\in\mathscr{Q}} \mathrm{KL}\left(q(\mathbf{z})\|p(\mathbf{z}\,|\,\mathbf{x})\right).} \tag{10}$$

Once found, $q^*(\cdot)$ is the best approximation of the conditional, within the family $\mathscr{Q}$. The complexity of the family determines the complexity of this optimization.

However, this objective is not computable because it requires computing the evidence $\log p(\mathbf{x})$ in Equation (3). (That the evidence is hard to compute is why we appeal to approximate inference in the first place.) To see why, recall that KL divergence is

Figure 3: Blei et al. offer a straightforward notation for the reverse K-L divergence. Figure taken from [16].



Friston Uses Reverse KL *(So does Beal, Blei et al.)*

$$F(s, a, \lambda) = -\int_\psi q(\psi|\lambda) \ln \frac{p(\psi, s, a, \lambda|m)}{q(\psi|\lambda)}\, \mathrm{d}\psi$$

Friston (2013); Eqn. 2.7 (extract)

In Friston's notation:

- ***p is the "data probability," and***
- ***q is the "model probability"***

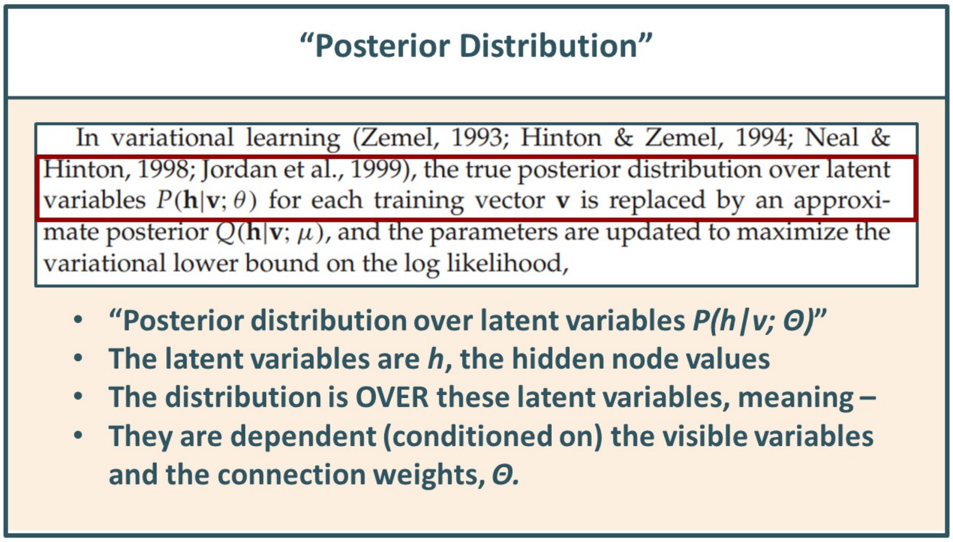
And we have $D_{KL}(q||p)$ instead of $D_{KL}(P||Q)$; e.g.

$$\begin{aligned} F(\tilde{s},\tilde{a},\tilde{r}) &= E_q[L(\tilde{x})] - H[q(\tilde{\Psi}\,|\,\tilde{r})] \\ &= L(\tilde{s},\tilde{a},\tilde{r}) + \boxed{D_{KL}[q(\tilde{\Psi}\,|\,\tilde{r})\,\|\,p(\tilde{\Psi}\,|\,\tilde{s},\tilde{a},\tilde{r})]} \end{aligned}$$

Friston et al. (2013); Eqn. 3.2

Figure 4: Friston uses a reverse K-L divergence in his formulation of active inference. Figure taken from [16].

As a brief comparison, we also look at the notation used by Salakhutdinov and Hinton (2012) [17] in their work on deep learning.



"Posterior Distribution"

In variational learning (Zemel, 1993; Hinton & Zemel, 1994; Neal & Hinton, 1998; Jordan et al., 1999), the true posterior distribution over latent variables $P(\mathbf{h}|\mathbf{v};\theta)$ for each training vector $\mathbf{v}$ is replaced by an approximate posterior $Q(\mathbf{h}|\mathbf{v};\mu)$, and the parameters are updated to maximize the variational lower bound on the log likelihood,

- **"Posterior distribution over latent variables *P(h|v; Θ)*"**
- **The latent variables are *h*, the hidden node values**
- **The distribution is OVER these latent variables, meaning –**
- **They are dependent (conditioned on) the visible variables and the connection weights, *Θ*.**

Figure 5: Salakhutdinov and Hinton (2012) also use the reverse K-L divergence in their formulation of energy-based methods; this figure shows their notation for the probability of the hidden (latent) nodes conditioned on the visible, as well as the approximating model Q. Figure taken from [16].

The Salakhutdinov and Hinton (2012) work is a cornerstone for energy-based methods, and by including them in this notational study, we can see how using the reverse Kullback-Leibler divergence is consistent throughout both energy-based methods and variational inference. We see their notation in Figure 5.

7 Transition to Variational Inference: Rewriting the Bayesian Posterior Distribution

Using the reverse K-L divergence is the set-up for variational inference as well as for energy-based methods (e.g., the restricted Boltzmann machine).

The next step, after we establish the reverse K-L divergence, is that we rewrite the Bayesian conditional dependence term $(p(x))$ within the logarithmic term $(log(q(x)/p(x)))$.

Once we do this, we can rewrite the equation in two different ways.

One is useful, the other is not. Variational inference proceeds by working with the "useful" version of the resulting two equations.

Before we rewrite the K-L divergence term of Eqn. 7, we first recall how the Bayesian posterior probability density can be rewritten, as framed in Blei et al. (2016) [1].

Consider a system that has a set of observable variables $\mathbf{v} = v_{1..V}$ and a set of latent or "hidden" variables $\mathbf{w} = w_{1..W}$. In a feedforward neural network, for example, the observable variables $\mathbf{v}$ would be the values of the input and output layer neurons, and the latent (hidden) variables would be the associated values of the hidden layer $\mathbf{w}$ 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 $\mathbf{w} = w_{1..W}$ and observations (visible variables) $\mathbf{v} = v_{1..V}$, where the conditional density function is given as

$$p(w|v) = p(w,v)/p(v). \tag{9}$$

Conversely, we also have

$$p(w,v) = p(w|v)p(v). \tag{10}$$

With this brief recollection of Bayesian probabilities, we address how Friston (op. cit.) evolves his primary notation for active inference, using the reverse K-L divergence as a cornerstone.

8 Friston's Formulation of the Free Energy Function

This section very briefly reviews how Friston approaches active inference, beginning with the reverse Kullback-Leibler divergence.

8.1 Interpreting Friston's Use of the K-L divergence

Friston (combining [5] (Eqns. 2.7 & 2.8) and [6] (Eqn. 3.2)) writes the reverse Kullback-Leibler (K-L) divergence as

$$D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s},\tilde{a},\tilde{r})] = \sum_{i=1}^{I} q(\tilde{\psi}|\tilde{r}) \ln \left(\frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}|\tilde{s},\tilde{a},\tilde{r})} \right). \qquad (11)$$

We briefly interpret the physical meaning of the terms in Eqn. 11. The K-L divergence measures the divergence between the model-distribution q of (i.e., probability distribution over) the external system, as conditioned on the reprsentation $\tilde{r}$, and the actual representation of the external system itself $p(\tilde{\psi}|\tilde{s},\tilde{a},\tilde{r})$.

Friston (2013) [5] describes his approach in saying "This means that the internal states $[\tilde{r}]$ will appear to respond to sensory fluctuations based on posterior beliefs about underlying fluctuations in external states. We can formalize this notion by associating these beliefs with a probability density over external states $q(\Psi|\lambda)$ that is encoded (parametrized) by internal states." (In this description, the tilde was removed for simplicity.)

The overall intention of [5] appears to be focused on how we build an internal representation of an external world Ψ, mediated via *sensing* and *action* units that communicate information over a Markov blanket. This is illustrated in Figure 6.

Friston's formulation frames how the external world Ψ can be conceptualized as a Bayesian dependence on the internal representation. Thus, we have a model of how the external world depends on the representation, framed as $q(\Psi|\lambda)$ (both λ and r are used by Friston, in different works), and we have an actual set of observations that represent a constructed model of how Ψ depends on the *sensing* and *action* units, together with the representation r, expressed as $p(\Psi|s,a,r)$.

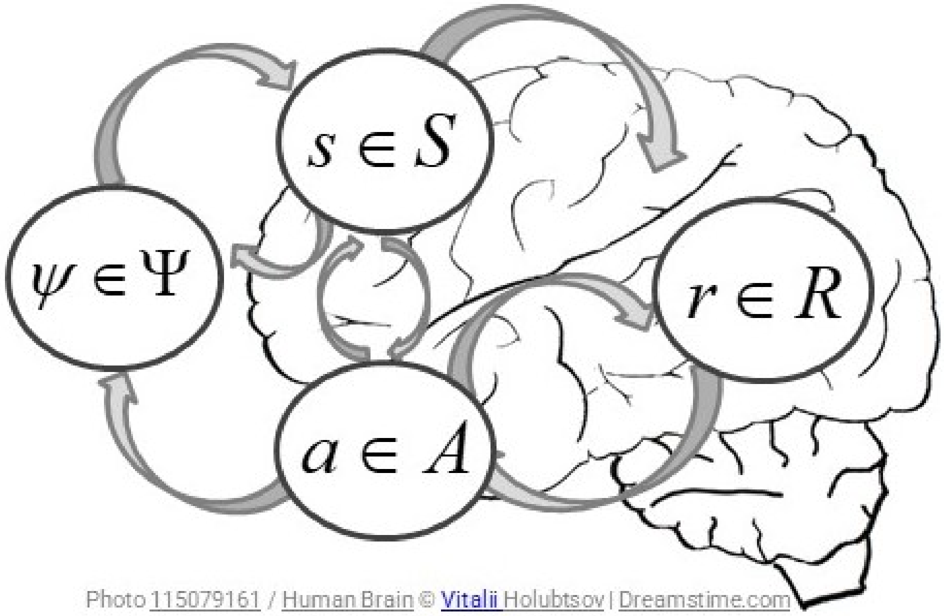


Photo 115079161 / Human Brain © Vitalii Holubtsov | Dreamstime.com

Figure 6: An Interpretation of Friston's Notion of Modeling the External System Ψ.

The model-distribution q is a model of the external system, $\tilde{\psi}$, which is why we write $q = q(\tilde{\psi}|\tilde{r})$. 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. 11, we are looking at the divergence between the model-distribution of the system at equilibrium and the probabilities p of various components of the *internal representation* of the external system, potentially in a not-yet-at-equilibrium state.

The parameter(s) θ directly influence p, but the notation for θ is suppressed in this section. We note that any time we write $p(x)$, we are implicitly writing $p(x|m)$, because we are using p to represent the notion of a model that uses a certain parameter set θ.

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 [5] and [6], and replicated in Figure 6, 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 $\tilde{r}$.

Author's note: This entire argument, and the somewhat forced depiction of how the external states $\tilde{\Psi}$ depend on internal states $\tilde{r}$, is notation that typified Friston's early works on active inference, circa 2010 - 2015 [4, 5, 6]. Beginning in about 2016 - 2017, Friston's notation changed, so that the depiction of external states become more implicit [18, 19].

8.2 The Kullback-Leibler Divergence - As Interpreted by Friston et al.

Replicating the notation and description found in Maren (2022) [20], Subsection 3.2, "Integrating over the model space," "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 $\tilde{\psi}$."

The following is extracted from Maren (2022) [20], which provides a detailed discussion of the variational Bayes method, especially in the context used by Friston [5].

"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 θ. 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 θ.

“The previous Eqn. 11 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. 11 as

$$D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})] = q(\tilde{\psi}|\tilde{r}) \ln \left(\frac{q(\tilde{\psi}|\tilde{r})}{p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})} \right). \tag{12}$$

“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. 12, 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.”

8.3 Friston’s Free Energy Expression

In this subsection, we briefly identify how Friston uses the reverse K-L divergence (from the previous subsection in this work) in the second half of Eqn. 13; the equality between the "variational free energy" and the sum of the pooled negative log probabilities of sensory states (and their accompanying representational and active states) and the K-L divergence.

Specifically, [5] (Eqns. 2.7 & 2.8) and [6] (Eqn. 3.2) formulate this as

$$F(\tilde{s}, \tilde{a}, \tilde{r}) = L(\tilde{s}, \tilde{a}, \tilde{r}) + D_{KL}[q(\tilde{\psi}|\tilde{r})||p(\tilde{\psi}|\tilde{s}, \tilde{a}, \tilde{r})]. \tag{13}$$

In a related work, we discuss Friston's free energy in more detail [20]. Specifically, in that work, we:

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

The objective in this paper has been limited in scope to introducing Friston's notation and cross-comparing it with that used by others, so that Friston's work can be read in context.

9 Kullback-Leibler Notation in Action Perception Divergence

Recently, Hafner et al. (2020, 2022) evolved "Action Perception Divergence" (APD) [7], which is an extension of active inference. The APD notation is a bit different from prior Friston notation.

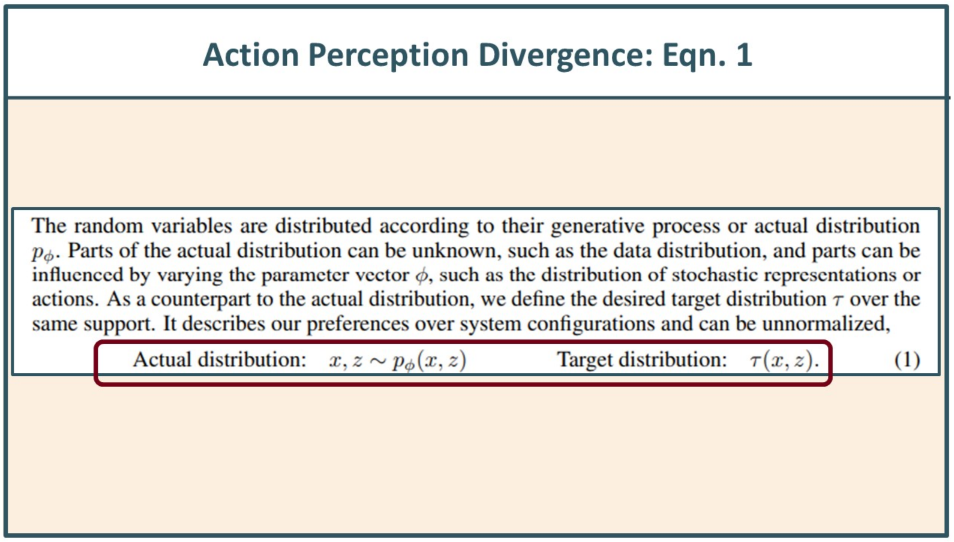
While still using a reverse K-L divergence, the divergence is between a set of potential probability distributions that can be assessed against a target T, as illustrated in Fig. 7.

APD is a more broadly-conceptualized evolution of active inference, so the target state T is not constrained to be a specific representation.

10 A New Divergence Measure for Using the 2-D Cluster Variation Method

This entire work, until now, has been devoted to the Kullback-Leibler divergence.

There are occasions on which an alternative divegence measure is needed.



Action Perception Divergence: Eqn. 1

The random variables are distributed according to their generative process or actual distribution p_ϕ. Parts of the actual distribution can be unknown, such as the data distribution, and parts can be influenced by varying the parameter vector ϕ, such as the distribution of stochastic representations or actions. As a counterpart to the actual distribution, we define the desired target distribution τ over the same support. It describes our preferences over system configurations and can be unnormalized,

$$\boxed{\text{Actual distribution:} \quad x, z \sim p_\phi(x, z) \qquad \text{Target distribution:} \quad \tau(x, z).} \tag{1}$$

Figure 7: Hafner et al. (2020, rev 2022) [7] use a slightly different expression for the reverse K-L divergence. Figure taken from Themesis YouTube video, "AGI: Action Perception Divergence (APD) - A Notation Review" [21].

One such occasion is the case where we use the 2-D cluster variation method (2-D CVM) as a *model* within a variational situation. An expanded role for the CVM is as a "computational engine," which allows temporal persistence of latent variable activations. This is embodied in the CORTECON(R) (COntent-Retentive, TEmporally-CONnected neural network) architecture. CORTECONs(R) are a new class of neural network, going beyond the original six classes or distinct "neural network topologies" identified in Maren (1991, Maren et al. 1990) [22, 23].

The 2-D CVM was introduced first by Kikuchi in 1951 [24], and expanded by Kikuchi and Brush in 1967 [25]. Maren (Maren et al.,1992, Maren, 1993, and Schwartz and Maren, 1993) devised the original CORTECON(R) method in 1992, using a 1-D CVM as the internal computational layer [26, 28, 27].

Maren and Szu (2015, Maren, 2016) [29, 30] described the 1-D CVM as a method for neurophysiologically-based modeling, and then Maren broadened the scope to the 2-D CVM between 2018 and 2021 [32, 31].

The unique advantage of using a (1-D or 2-D) CVM model is that the ***model can itself be brought to equilibrium***, as shown in [32]. However, this early work revealed the need for a more precise means of identifying the enthalpy parameters $(\varepsilon_0, \varepsilon_1)$ that best characterized a given data distribution. (Note: we have found it more useful to use h, or the *h-value*, instead of the interaction enthalpy parameter ε_1, with the relation that $h = exp(2\varepsilon_1)$ for the 2-D CVM.)

The reason that simple application of the Kullback-Leibler divergence is not applicable is that it does not address the full suite of (configuration) variables used in constructing the 2-D CVM grid and the equilibrium-based model of that grid.

As an example, initial experiments with the 2-D CVM as a model use the case where the probabilities of a grid node being in state **A** or state **B** are equiprobable ($x_1 = x_2 = 0.5$ for the case where $\varepsilon_0 = 0$), so that the logarithm of their ratio is zero. ($x_1/x_2 = 1$, so $log(x_1/x_2) = 0$.) If the divergence were simply limited to the notion of whether a given node was "on" (state **A**) or "off" (state **B**), then we wouldn't be able to compare topographries for the case where $x_1 = x_2 = 0.5$.

However, this particular case (of equiprobable values for x_1 and x_2) is important. For this case, there is an analytic solution for the remaining configuration variable values (the nearest-neighbors y_i, the next-nearest-neighbors w_i, and the triplets z_i) in terms of h. This gives us a means to investigate a range of behaviors for the equiprobable distribution of "on" (state **A**) and "off" (state **B**) nodes when there is a range of different interaction enthalpy parameters, or *h-values.* In order to characterize these topographies, we need to find an *h-value* that "best fits" an observed set of configuration variable values (y_i, w_i, and z_i) associated with a given topography.

There may be different *h-values* associated with each configuration variable value, as a given topography may not be at equilibrium - that is, have a very non-equlibrium distribution of configuration variable values. (A distinct and different *h-value* would then correspond to the analytic solution for each of the different configuration variable values y_i, w_i, and z_i.) Thus, we have needed a new divergence measure that could be coupled with a variational approach to find the *h-value* that provides the "best fit" according to a minimized divergence.

Maren (2022) [10] proposed a divergence suitable for working with CVM models where there was a need to identify the *h-value* that provided a "best fit," according to this new divergence measure, for characterizing a 2-D CVM grid. This approach is particularly needed when there are equal numbers of nodes that are "on" (state **A**) or "off" (state **B**) ($x_1 = x_2 = 0.5$). This would be the case for *both* the original observed system p as well as any constructed (free-energy-minimized) system q, because one of the criteria for free energy minimization would be to keep the same number of units in states **A** and **B**. (That is, we would conduct free energy minimization for each of a different element in a range of *h-values*, subject to the constraint that $x_1 = x_2 = 0.5$.)

We introduce this new divergence measure by comparing it with what would be the case if we were *not* considering the full range of configuration variables y_i, w_i, and z_i.

We take the imaginary case where we are comparing a range of potential CVM *models* (denoting any given model as q, for a specific *h-value*) against an observed or *representational* system (denoting it as p). Any given q corresponds to a free-energy-minimized system, brought to equilibrium for the parameters $(\varepsilon_0, \varepsilon_1)$.

Our measure is a form of a *reverse* divergence.

If we were concerned only with the distribution of nodes in "on" and "off" states; i.e., measuring only x_1 and x_2, then expressing the K-L divergence would give us

$$D_{KL}[q(r)||p(r)] = \sum_{i=1}^{2} x_{i,q} \ln \left(\frac{x_{i,q}}{x_{i,p}} \right). \tag{14}$$

In this case, the summation would be over two states, and we would have $p(r) = x_1$ in the topography that we are modeling, and $q(r) = x_1$ in the resultant, free-energy-minimized topography. For clarity, we could identify these as $p(r_1) = x_{1,p}$ and $q(r_1) = x_{1,q}$, and $p(r_2) = x_{2,p}$ and $q(r_2) = x_{2,q}$ The associated parameter set is given as $\theta = \{\varepsilon_0, \varepsilon_1\}$. When $x_1 = x_2 = 0.5$, then $\varepsilon_0 = 0$.

If we were to apply this to the natural topographies that were selected for this work, the divergence value that would be found by applying Eqn. 14 would yield a value of zero, regardless of the *h-value* (or correspondingly, the ε_1) used. This is because by selecting an equiprobable distribution of units, we are constraining that $p(r) = q(r) = 0.5$ for both the "on" and ''off" states.

Clearly, this Eqn. 14 would be neither sufficient nor appropriate for our needs.

To work with natural topographies and their represenations, or with any 2-D system where the interest is in local topographies, we need to include terms indicative of relations between the remaining configuration variables.

To do this, we introduce a new divergence measure, expressed here for the 2-D CVM, as

$$\begin{aligned} D_{2D-CVM}[q(r)||p(r)] = \\ & 2\sum_{i=1}^{3}\beta_i y_{i,q}\ln\left(\frac{y_{i,q}}{y_{i,p}}\right) + \sum_{i=1}^{3}\beta_i w_{i,q}\ln\left(\frac{w_{i,q}}{w_{i,p}}\right) \\ & -\sum_{i=1}^{2}x_{i,q}\ln\left(\frac{x_{i,q}}{x_{i,p}}\right) - 2\sum_{i=1}^{6}\gamma_i z_{i,q}\ln\left(\frac{z_{i,q}}{z_{i,p}}\right) \end{aligned} \tag{15}$$

We refer to this as the Kikuchi-Maren divergence.

Maren (2022) [10] illustrates how this divergence measure works as a means of comparing four distinct naturally-occurring topographies.

11 Summary and Conclusions

One of the biggest challenges, for someone attempting to study how the (reverse) K-L divergence is used in variational and active inference, has been the subtle differences in notation used by different authors. This work has attempted to provide some clarity into these notational varients.

In addition, there is new divergence measure, the Kikuchi-Maren (K-M) divergence, which is conceptually similar to the K-L divergence, but is specifically devised for use with the cluster variation method. This new K-M divergence makes it possible to identify parameter sets $(\varepsilon_0, \varepsilon_1)$ that best characterize a given 1-D or 2-D grid pattern. This then makes it possible to use the 1-D or 2-D CVM to model existing topographies, and to become a computational engine within a larger (CORTECON(R)) system.

References

[1] Blei, David M., Alp Kucukelbir and Jon D. McAuliffe. 2016. "Variational Inference: A Review for Statisticians." *arXiv*:1601.00670v4 [stat:CO] (2 Nov 2016).

[2] Blei, D. M. , A. Kucukelbir and J. D. McAuliffe. 2017. “Variational Inference: A Review for Statisticians.” *arXiv*:1601.00670v4

[stat:CO], 2 Nov 2016. *J. American Statistical Association* **112** (58):859-877. doi:10.1080/01621459.2017.1285773.

[3] Beal, M. J. 2003. *Variational Algorithms for Approximate Bayesian Inference.* PhD Thesis, University College London. (PDF available online at: http://www.cse.buffalo.edu/faculty/mbeal/papers/beal03.pdf.)

[4] Friston, K. 2010. “The Free-Energy Principle: A Unified Brain Theory?” *Nat. Rev. Neurosci.*, **11**:127–138. doi:10.1038/nrn2787.

[5] Friston, K. 2013. “Life as We Know It.” *Journal of The Royal Society Interface*, **10**(86).

[6] Friston, K., M. Levin, B. Sengupta, and G. Pezzulo. 2015. “Knowing One’s Place: A Free-Energy Approach to Pattern Regulation.” *J. R. Soc. Interface*, **12**:20141383. doi:10.1098/rsif.2014.1383. (Available online at: http://dx.doi.org/10.1098/rsif.2014.1383.)

[7] Hafner, Danijar, Pedro A. Ortega, Jimmy Ba, Thomas Parr, Karl Friston and Nicolas Heess. 2020, rev. 2022. “Action and Perception as Divergence Minimization.” *arXiv*:2009.01791v3 [cs.AI] (13 Feb 2022). doi:10.48550/arXiv.2009.01791. (Accessed Aug. 10, 2024; available online at https://arxiv.org/pdf/2009.01791.)

[8] Kingma, Diederik P. and Max Welling. 2013. “Auto-Encoding Variational Bayes." *arXiv*:1312.6114v11 [stat.ML] (10 Dec 2022).

[9] Kingma, Diederik P. and Max Welling. 2019. “An Introduction to Variational Autoencoders." *Foundations and Trends (R) in Machine Learning* **12** (4, Nov.):307-392. doi:10.1561/2200000056; published also as *arXiv*:1906.02691v3 [cs.LG] 11 Dec 2019.

[10] Maren, A.J. 2022. “A Variational Approach to Parameter Estimation for Characterizing 2-D Cluster Variation Method Topographies.” *Themesis Technical Report THM 2022-001 (ajm).* *arXiv*:2209.04087v1 [cs.NE] (9 Sep 2022). doi:10.48550/arXiv.2209.04087. (Accessed Aug. 12, 2024; available online at https://www.mdpi.com/1099-4300/23/3/319.)

[11] Kullback, S. and R.A. Leibler. 1951. “On Information and Sufficiency." *Ann. Math. Statist.* **22**(1):79-86.

[12] Halmos, Paul R. and L.J. Savage. 1949. “Application of the Radon-Nikodym Theorem to the Theory of Sufficient Statistics.” *Ann. Math. Statist.* **2** (June):225-241. doi:10.1214/aoms/1177730032.

[13] Balesdent, M., J. Morio, C. Verge and R. Paste. 2016. “Simulation Techniques,” in Jerome Morio and Mathieu Balesdent (Eds.), *Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems*, Ch. 5. (Frisco, CO: Elsevier.)

[14] Theodoridis, Sergios. 2020. *Machine Learning: A Bayesian and Optimization Perspective, 2nd Ed.* (Cambridge, MA: Academic Press) doi:10.1016/C2019-0-03772-7.

[15] Wikipedia-Kullback-Leibler-Divergence. (No date.) “Kullback-Leibler Divergence." (Accessed Jan 28, 2024, available online at https://en.wikipedia.org/wiki/Kullback-Leibler_divergence.)

[16] Maren, Alianna J. 2024. “How to Understand Generative AI Using the Reverse Kullback-Leibler Divergence.” *Themesis, Inc. YouTube Channel* (January 23, 2024). (Running time 15.14 min, https://www.youtube.com/watch?v=4m2WUaBZxu8&t=16s, last accessed Jan. 29, 2024.)

[17] Salakhutdinov, Ruslan and Geofrrey Hinton. 2012. “An Efficient Learning Procedure for Deep Boltzmann Machines.” *Neural Computation* **24**(8; August):1967-2006.

[18] Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, John O’Dohertye and Giovanni Pezzulo. 2016. “Active Inference and Learning." *Neuroscience & Biobehavioral Reviews* **68** (September 2016): 862-879. doi:10.1016/j.neubiorev.2016.06.022. (Accessed Aug. 10, 2024; available online at https://www.sciencedirect.com/science/article/pii/S0149763416301336?via%3Dihub.)

[19] Friston, Karl, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck and Giovanni Pezzulo. 2017. “Active Inference: A Process Theory.” *Neural Comput* **29**(1; Jan, 2017)):1-49. doi:10.1162/NECO_a_00912. (Accessed Aug. 10, 2024; available online at https://activeinference.github.io/papers/process_theory.pdf.)

[20] Maren, Alianna J. 2019, rev. 2022. “Derivation of the Variational Bayes Equations.” *arXiv*:1906.08804v5 cs[NE] (4 Nov 2022).

[21] Maren, Alianna J. 2024. “AGI: Action Perception Divergence (APD) - Notation Review.” *Themesis, Inc. YouTube Channel* (July 17, 2024). (Running time 15.48 min, https://www.youtube.com/watch?v=436l0cmnEi0&list=PLQ7kdul7PF0dNb7PBYNlBb23NQmH last accessed Aug. 11, 2024.)

[22] Maren, Alianna J. 1991. “A Logical Toplogy of Neural Networks,” in *Proceedings of the Second Workshop in Neural Networks – Academia, Industry, NASA, & Defense (WNN-AIND 91)* (Feb., 1991; Auburn, GA). Extended version published as *Themesis Inc. Technical Report THM-TR 1991 (ajm)* (1991). (Accessed Aug. 25 2022; available online at https://aliannajmaren.com/Downloads/Logical-topology-neural-networks.pdf.)

[23] Maren, A.J., C.Harston and R.Pap. 1990. *Handbook for Neural Computing Applications* (Cambridge, MA: Academic Press). doi: 10.13140/2.1.2917.5364.

[24] Kikuchi, R. 1951.“A Theory of Cooperative Phenomena,” *Phys. Rev.* **988**(81):127–138.

[25] Kikuchi, R. and S.G. Brush. 1967. “Improvement of the Cluster Variation Method.” *J. Chem. Phys.* **47**:195.

[26] Maren, A.J., E. Schwartz and J. Seyfried. 1992. “Configurational Entropy Stabilizes Pattern Formation in a Hetero-Associative Neural Network,” in *IEEE Int’l Conf. SMC* (October, 1992, Chicago, IL) pp. 89-93. doi:10.1109/ICSMC.1992.271796.

[27] Maren, A.J. 1993. “Free Energy as Driving Function in Neural Networks,” in *Symposium on Nonlinear Theory and Its Applications* (Dec. 5-10, 1993: Hawaii). pp.89-93. doi:10.13140/2.1.1621.1529 (Accessed Aug. 12, 2024; available online at https://www.aliannajmaren.com/Downloads/Free_Energy_Driving_Fn_NN_AJM_12-5-93.pdf.)

[28] Schwartz, E. and A.J. Maren. 1993. “Domains of Interacting Neurons: A Statistical Mechanical Model,” in *Proc. World Congress on Neural Networks (WCNN’93 – Portland)*, (Portland, OR: July 11-15, 1993), I-577 – I-580. (Available online at https://www.aliannajmaren.com/Downloads/WCNN-93-Presentation.pdf.)

[29] Maren, Alianna J. and Harold H. Szu. 2015. “A New EEG Measure Using the 1-D Cluster Variation Method,” in *Proceedings SPIE STA (Sensing Technology + Applications) Conference: Independent Component Analyses, Compressive Sampling, Large Data Analyses (LDA), Neural Networks, Biosystems, and Nanoengineering XIII* (April, 2015; Baltimore, MD). doi:10.13140/2.1.4042.6560. (Accessed Aug. 12 2024; available

online at https://www.aliannajmaren.com/Downloads/SPIE-2015-9496-26.pdf.)

[30] Maren, A.J. 2016. "The Cluster Variation Method: A Primer for Neuroscientists," *Brain Sciences*, **6**(4):44. doi:10.3390/brainsci6040044. (Available online at: https://doi.org/10.3390/brainsci6040044.)

[31] A.J. Maren, "Free Energy Minimization Using the 2-D Cluster Variation Method: Initial Code Verification and Validation," *Themesis Technical Report 2018-001v2 (ajm)*. v1: 2018; v2: 2019. arXiv:1801.08113v2 [cs.NE] (25 Jun 2019). (Accessed Aug. 12, 2024; available online at https://arxiv.org/pdf/1801.08113.)

[32] Maren, A.J. 2021. "The 2-D Cluster Variation Method: Topography Illustrations and Their Entropy Parameter Correlations," *Entropy.* **23**(3):319. doi:10.3390/e23030319. (Accessed Aug. 12, 2024; available online at https://www.mdpi.com/1099-4300/23/3/319.)