Exploring the relation of variational inference and integrated information in a minimal model

Nadine Spychala[1,2,*] and Miguel Aguilera[3,4,1]

[1]University of Sussex, Falmer, Brighton. United Kingdom.
[2]Software Sustainability Institute. United Kingdom
[3]BCAM – Basque Center for Applied Mathematics, Bilbao, Spain
[4]IKERBASQUE, Basque Foundation for Science. Bilbao, Spain
[*]nadine.spychala@gmail.com

Abstract

Integrated information and variational inference provide influential mathematical frameworks in neuroscience. Yet, the understanding of the connection between the two is limited. Here, we study a minimal model to show how variational inference displays large integrated information for highly correlated target distributions, in contrast with alternative inference approaches like maximum likelihood estimation.

Many problems in different domains can be cast as the approximation of complicated probability densities. Variational inference (VI, also known as approximate Bayesian inference) is a method from machine learning used for approximating such difficult-to-compute probability densities (Jordan et al., 1999). VI approaches have been extensively explored in neuroscience to investigate the brain's capacity to operate in situations of uncertainty in a Bayes' optimal way (i.e., close to what Bayesian models predict) (Clark, 2013). In this context, it has also been hypothesized that the brain is facing hard-to-calculate posterior densities equally by exploiting approximate solutions such as VI.

Integrated information (normally symbolized by φ) denotes the idea that a system of interconnected elements, considered in as a "whole", can encode information that goes beyond the information encoded by the sum of individual "parts" (Barrett and Seth, 2011). This idea has been operationalized in mathematically different ways (Tegmark, 2016), and explored in different scientific contexts - particularly so in consciousness science, where an entire theoretical framework called *Integrated Information Theory* (IIT) has been developed (Oizumi et al., 2014), pursuing the core hypothesis that consciousness arises as a result of high information integration.

Both VI and integrated information have been influential and explored considerably – albeit separately from each other – in neuroscience. Thus, a link between the two so far is missing, begging the question of whether VI as an inference method and φ as a measure of dependency between a system's parts are related – i.e., will systems performing optimal Bayesian inference display integrated information? This question is also relevant in the field of Artificial Life where learning mechanisms and whole-parts relationships play important roles in the context of understanding mind and life processes. Here, we study this question by inspecting the evolution of φ during a minimal example of approximate Bayesian inference, and comparing it with the behaviour resulting from maximum likelihood estimation.

Variational inference and Maximum Likelihood We consider two unobserved (or latent) variables y_1, y_2, following a bivariate Gaussian distribution $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}^{-1})$.They are inferred by defining a mean field distribution $\mathbf{y} \sim \mathcal{N}(\mathbf{x}, \mathbf{B}^{-1})$, where $\mathbf{B}$ is a diagonal matrix (for simplicity, $B_{ii} = A_{ii}$). In this setup, VI prescribes minimizing the Kullback-Leibler (KL) divergence $D(q(\mathbf{y})||p(\mathbf{y}))$. Conversely, an alternative inference setup is given by a maximum likelihood estimation (MLE) strategy, which corresponds instead to maximizing the divergence $D(p(\mathbf{y})||q(\mathbf{y}))$. In both cases, we assume that we only can access a noisy estimation of the true mean values of the distribution $p(\mathbf{y})$, $\hat{\boldsymbol{\mu}}_t = \boldsymbol{\mu}_t + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t$ being Gaussian white noise with a diagonal covariance $\boldsymbol{\Gamma}$.

In both VI and MLE, inference is implemented by a gradient descent on the KL divergence, leading to the following stochastic process

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma \left(\mathbf{C}(\boldsymbol{\mu} - \mathbf{x}_t) + \mathbf{C}\boldsymbol{\epsilon}_t \right), \tag{1}$$

where γ is a learning rate, and $\mathbf{C} = \mathbf{A}$ for VI and $\mathbf{C} = \mathbf{B}$ for MLE. More generally, we can consider a mixed learning strategy with $\mathbf{C} = \alpha\mathbf{A} + (1-\alpha)\mathbf{B}$, with $\alpha \in [0, 1]$.

We study the behaviour of ensembles of learning dynamics, in which the stochastic process $\mathbf{x}_t$ takes the form of a multivariate normal distribution with statistical moments

$$\mathbf{m}_{t+1} = (\mathbf{I} - \gamma\mathbf{C})\mathbf{m}_t + \gamma\mathbf{C}\boldsymbol{\mu}, \tag{2}$$

$$\boldsymbol{\Sigma}_{t+1,t+1} = (\mathbf{I} - \gamma\mathbf{C})\boldsymbol{\Sigma}_{t,t}(\mathbf{I} - \gamma\mathbf{C}) + \gamma^2\mathbf{C}\boldsymbol{\Gamma}\mathbf{C}, \tag{3}$$

$$\boldsymbol{\Sigma}_{t+1,t} = (\mathbf{I} - \gamma\mathbf{C})\boldsymbol{\Sigma}_{t,t}. \tag{4}$$

As $\mathbf{C}$ is symmetric, the steady-state distribution of $\mathbf{x}_t$ is an equilibrium distribution $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*)$, with

$$\boldsymbol{\Sigma}^* = \gamma^2((\mathbf{C}\boldsymbol{\Gamma}\mathbf{C})^{-1} - (\mathbf{I} - \gamma\mathbf{C})(\mathbf{C}\boldsymbol{\Gamma}\mathbf{C})^{-1}(\mathbf{I} - \gamma\mathbf{C}))^{-1}. \tag{5}$$

Integrated information Following (Barrett and Seth, 2011), for a dynamical system of elements $\mathbf{x} = x_1, \ldots, x_n$ evolving over time $t = 1, \ldots, m$, integrated information is defined as:

$$\varphi \equiv I(\mathbf{x}_{t-\tau}, \mathbf{x}_t) - I(\mathbf{x}^A_{t-\tau}, \mathbf{x}^A_t) - I(\mathbf{x}^B_{t-\tau}, \mathbf{x}^B_t), \quad (6)$$

where I is the mutual information (MI), A and B refer to two partitions of the variables in the system, and τ is a time-lag. φ will be greater than zero if the system as a whole predicts itself better compared to the sum of predictive information its parts have.

φ can get negative when redundancy between the two partitions is large. A refined definition based on Partial (Williams and Beer, 2010) and Integrated Information Decomposition (Mediano et al., 2019) alleviates this problem by adding double-redundancy (capured by a minimum information $I_{\min}$) to φ:

$$\varphi^{\mathrm{R}} = \varphi + \underbrace{\min_{i,j \in \{A,B\}} I(\mathbf{x}^i_{t-\tau}, \mathbf{x}^j_t)}_{I_{\min}}. \quad (7)$$

Results We explore different scenarios considering different covariances between latent variables (off-diagonals in $\mathbf{A}$, which we denote by ρ), as well as how much those covariances are taken into account (via α). In this setup, we study the steady-state solutions for φ, φ^{R}, as well as $I_{\min}$ (Fig. 1). We first observe that φ increases generally with both the values of ρ and α (except for a small decrease for very large ρ and $\alpha < 1$ (Fig. 1.a). This implies that information is larger when the reference system is strongly correlated and the inference strategy is closer to VI. In turn, φ^R peaks at a very large value of ρ, except for the case of $\alpha = 1$, which increases monotonically. This further supports that, even when large correlations (high ρ) introduce redundancies, behaving in a manner that is closer to VI increases information integration. Regarding $I_{\min}$, we observe a clear peak at at a very large value of ρ (around 0.9) in all cases, then decreasing to zero at $\rho = 1$.

Conclusions It seems that approximate Bayesian inference picks a linear re-combination of variables, allowing large values of integrated information that keep increasing the more correlated the latent variables are, whereas for very large correlations, integrated information is smaller and eventually goes down. Both low and very high correlations (around 0.9) yield smaller values of $I_{\min}$, suggesting that either very small or very large covariances prevent redundant information when learning the two variables. It is surprising that the behaviour of integrated information in this simple setup is neither trivial nor monotonic, which calls for further exploration of how these effects are generated.

This study provides, to our knowledge, the first evidence linking integrated information and variational inference formally and tractably in a simulation study. While the generality as well as neuroscientific relevance of these findings remain open questions, these preliminary results may spark further ideas on connection points that may be interesting to researchers from both fields.

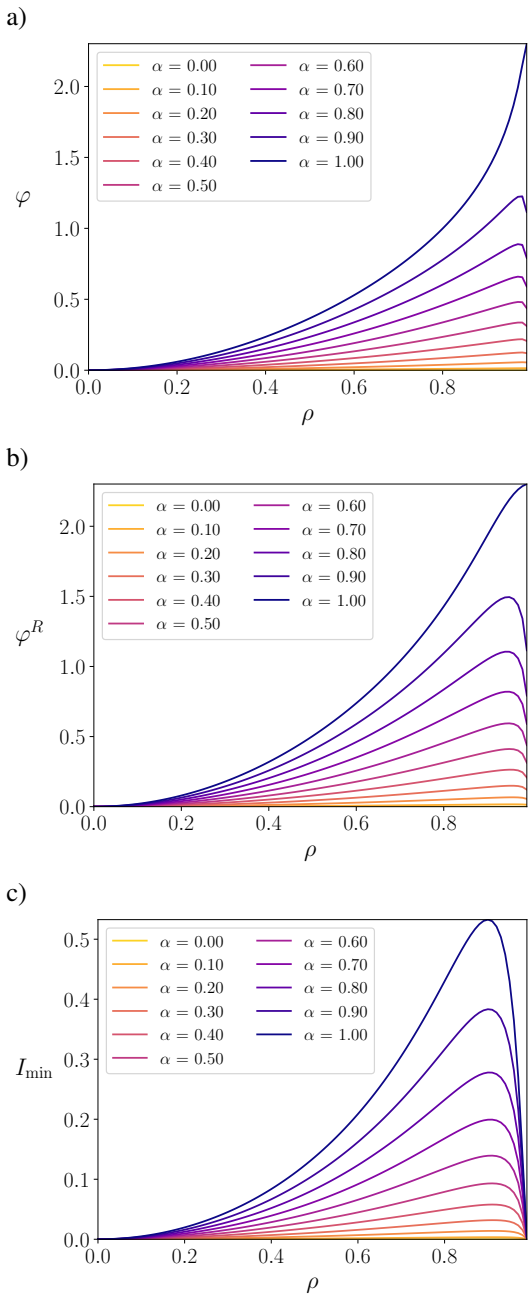

Figure 1: Steady-state values for a) φ, b) φ^{R} and c) $I_{\min}$.

References

Barrett, A. B. and Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS computational biology*, 7(1):e1001052.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Mediano, P. A., Rosas, F., Carhart-Harris, R. L., Seth, A. K., and Barrett, A. B. (2019). Beyond integrated information: A taxonomy of information dynamics phenomena. *arXiv preprint arXiv:1909.02297*.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, 10(5):e1003588.

Tegmark, M. (2016). Improved measures of integrated information. *PLoS computational biology*, 12(11):e1005123.

Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.