Hierarchical variational models for statistical physics

Jaan Altosaar
Princeton University
altosaar@princeton.edu

Rajesh Ranganath
New York University
rajeshr@cims.nyu.edu

Kyle Cranmer
New York University
kyle.cranmer@nyu.edu

Abstract

The Boltzmann distribution of a statistical physics model can be studied using the variational principle. However, the variational principle requires deriving model-specific approximations to the Boltzmann distribution. We review variational inference (VI), a machine learning framework for inferring distributions, and show it is equivalent to the Gibbs-Bogoliubov-Feynman variational principle in physics. The VI perspective can be useful, as recent VI methods do not require model-specific derivations. For example, variational autoregressive networks (VANs) are generic variational approximations for statistical physics models [19]. But VANs were developed using a reinforcement learning approach, so framing Wu et al. [19] as VI allows comparison to a plethora of variational approximations from the VI literature. As one example, we test hierarchical variational models (HVMs) as variational approximations [15] for statistical physics models. HVMs allow efficient sampling of system configurations, and we show that HVMs therefore scale to larger systems than VANs [19].

In statistical physics, the normalization constant of the Boltzmann distribution is a central quantity. The partition function can be used to derive properties of physics models, such as specific heat or magnetization. Model predictions can be compared to experimental values, which can inform how a model might be improved.

Calculating the partition function can be difficult, and there are many ways around computing the partition function, such as sampling methods and variational methods. Markov chain Monte and Carlo [11] allows sampling system configurations from the Boltzmann distribution of a model; these samples can be used to approximate physical quantities. Variational inference relies on optimizing (varying) functionals to derive approximations of observables, and originated via mean-field methods in physics [17, 6, 2].

We show that the machine learning framework of VI is equivalent to the Gibbs-Bogoliubov-Feynman variational principle. This can allow practitioners to study statistical physics models using many variational approximations, including VANs, which were developed in Wu et al. [19] using reinforcement learning. As an example of a variational method enabled by VI, we study HVMs as approximations to the Boltzmann distribution. We find that HVMs scale to larger systems sizes than VANs in Sherrington-Kirkpatrick and Ising models. Testing the feasibility of VI methods in statistical physics is a twofold opportunity. Statistical physics problems might serve as benchmarks for VI, and using VI for these problems can lead to improved computational methods in statistical physics.

Variational inference is the Gibbs-Bogoliubov-Feynman (GBF) variational principle. The Gibbs-Bogoliubov-Feynman (GBF) inequality, or GBF variational principle, is used to study physics models with intractable partition functions [3, 4]. A mean-field family of energy functions (with tractable partition functions) is used to find an approximation to a target model of interest. For a

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

model with energy function E and partition function Z, consider an approximating mean-field energy function E_{MF} (and partition function Z_{MF}). Then the GBF inequality reads

$$Z \geq Z_{\text{MF}} \exp\left(-\beta \left\langle E - E_{\text{MF}} \right\rangle_{\text{MF}}\right) . \tag{1}$$

The right-hand side of the inequality is tractable: the mean-field family can be chosen such that Z_{MF} can be computed. The parameters of the mean-field family can be varied to maximize the right-hand side and find the best approximation to the partition function Z. The GBF variational principle enables the study of models whose partition functions Z cannot be computed [3, 4].

The variational principle in machine learning is called variational inference (VI), and the analog of the GBF inequality is the evidence lower bound (ELBO). To derive the equivalence of these inequalities, we recall concepts from probability models and posterior inference. Consider a generative model of data $\mathbf{x}$ with latent variables $\mathbf{z}$ and joint probability $p(\mathbf{x}, \mathbf{z})$. One goal of probabilistic modeling is to infer the posterior distribution of latent variables given observed data, $p(\mathbf{z} \mid \mathbf{x})$. The partition function of the posterior distribution can be read from Bayes' rule,

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{z}) p(\mathbf{z})}{p(\mathbf{x})} .$$

The model evidence $p(\mathbf{x})$ is the partition function of the posterior. Calculating the partition function is what makes posterior inference difficult, as it requires integration over the (possibly high-dimensional) latent variables $\mathbf{z}$,

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} .$$

Variational inference (VI) is an approximate inference method that can be used to approximate this partition function of the posterior distribution. In VI, the posterior distribution of a model is approximated by a variational distribution $q(\mathbf{z}; \boldsymbol{\nu})$ indexed by parameters $\boldsymbol{\nu}$. The ELBO can be derived via the Kullback-Leibler (KL) divergence between the variational approximation and the posterior distribution [2]:

$$\begin{aligned}
\text{KL}\left(q(\mathbf{z}; \boldsymbol{\nu}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) &= \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\nu})] - \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] + \log p(\mathbf{x}) \\
&\Rightarrow \log p(\mathbf{x}) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\nu})] + \text{KL}\left(q(\mathbf{z}; \boldsymbol{\nu}) \parallel p(\mathbf{z} \mid \mathbf{x})\right) \\
&\Rightarrow \log p(\mathbf{x}) \geq \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\nu})] ,
\end{aligned}$$

where the last line uses the positivity of the KL divergence. The ELBO, $\mathcal{L}(\boldsymbol{\nu})$, is this evidence lower bound:

$$\mathcal{L}(\boldsymbol{\nu}) = \mathbb{E}_q[\log p(\mathbf{x}, \mathbf{z})] - \mathbb{E}_q[\log q(\mathbf{z}; \boldsymbol{\nu})] . \tag{2}$$

Maximizing the ELBO by varying the variational parameters $\boldsymbol{\nu}$ finds the member of the variational family closest to the posterior in terms of the KL divergence. The output of the VI algorithm is $\boldsymbol{\nu}^*$, the variational parameters indexing the approximate posterior $q(\mathbf{z}; \boldsymbol{\nu}^*)$. The partition function can be approximated by the numeric value of the ELBO, $\mathcal{L}(\boldsymbol{\nu}^*)$.

To show that variational inference is the GBF variational principle, consider a latent variable model without data like the Ising model (i.e., where the data is an empty set $\mathbf{x} = \{\}$). Such a model has an energy function E and its latent variables (or spins) $\mathbf{z}$ obey the Boltzmann distribution at reciprocal thermodynamic temperature β: $p(\mathbf{z}; \beta) = \exp(-\beta E(\mathbf{z}))/Z$. The 'posterior' is the latent variable distribution $p(\mathbf{z}; \beta)$, and the partition function Z is its normalizing constant. With a mean-field variational family $q(\mathbf{z}; \beta, \boldsymbol{\nu}) = \exp(-\beta E_{\text{MF}})/Z_{\text{MF}}$, the 'evidence' lower bound becomes the GBF bound on the partition function:

$$\begin{aligned}
&\log Z \geq \mathbb{E}_q[-\beta E] - \mathbb{E}_q[-\beta E_{\text{MF}}] + \log Z_{\text{MF}} \\
&\Rightarrow Z \geq Z_{\text{MF}} \exp\left(-\beta \mathbb{E}_q[E - E_{\text{MF}}]\right) ,
\end{aligned}$$

where we used that the unnormalized posterior distribution, or log joint $p(\mathbf{x}, \mathbf{z})$, corresponds to the Boltzmann factor $\exp(-\beta E)$. This is the GBF inequality in Equation (1); variational inference is the variational principle tailored for probabilistic modeling.

Related work. The GBF variational principle has been used to study Markov random fields [7] and the connection between variational inference and statistical physics has been well-documented [2, 6, 10]. But this equivalence between VI and the GBF inequality might serve as an introduction to VI for physicists. Wu et al. [19] implicitly use VI, by developing VANs and a reinforcement

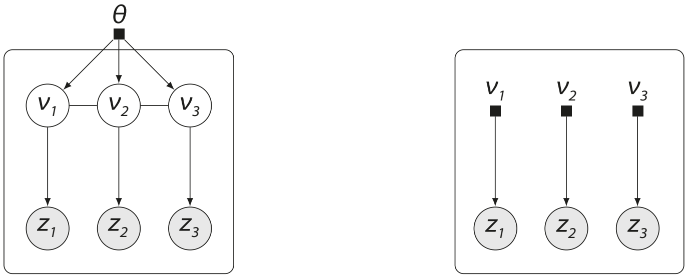


Figure 1: Hierarchical variational models (HVMs, left) capture dependencies between latent variables, compared to the mean-field variational family with independent variables (right).

learning policy gradient algorithm. Further, for a system of size L, autoregressive neural networks require $\mathcal{O}(L^2)$ forward passes to sample a system configuration, making VANs intractable in larger systems. A clear connection between statistical physics and VI in machine learning allows the use of VI techniques such as HVMs that can sample from a system in $\mathcal{O}(L)$ time and thus yield results for large systems.

Variational inference. VI is equivalent to the GBF variational principle and requires similar choices of a practitioner. The variational family $q(\mathbf{z}; \boldsymbol{\nu})$ to approximate a model must be chosen, in addition to a method to maximize the variational lower bound in Equations (1) and (2).

The VI literature provides several choices of variational family, such as a mean-field, factorized variational distribution with independent latent variables. Another choice of variational family is the Bethe approximation, which constrains the variational distribution to the polytope of mean parameters that are consistent for any two latent variables [18]. Some machine learning research focuses on developing variational approximations that capture correlations between latent variables [5, 8, 9, 19]. An example of a variational family that can model correlations between latent variables is the VAN family [19], which uses autoregressive neural networks to parameterize the variational distribution $q(\mathbf{z}_i \mid \mathbf{z}_1, \ldots, \mathbf{z}_{i-1})$. We explore the HVM class of variational approximations [16].

The second choice required to employ VI is how to optimize the variational lower bound in Equation (2). The choice of variational family can limit the available optimization techniques. For a simple variational family like the mean-field approximation, it may be possible to analytically evaluate the expectations in Equation (2). Then derivatives of the variational bound with respect to the variational parameters $\boldsymbol{\nu}$ and manual calculation can maximize the lower bound. If more expressive variational families are used (e.g. VANs with thousands or millions of variational parameters), the analytic approach is infeasible. Stochastic optimization and automatic differentiation software have been used to develop several approaches to computing gradients of the variational lower bound, such as black box variational inference [16, 12].

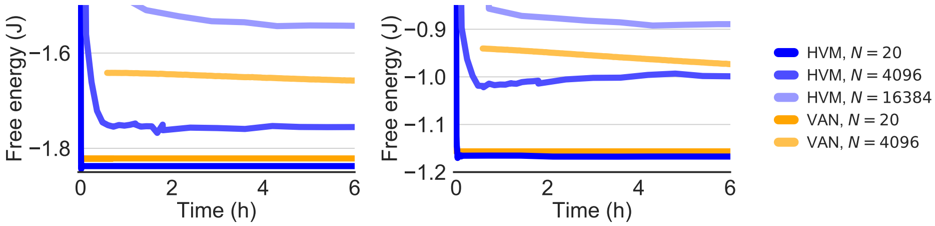
The choice of variational family $q(\mathbf{z}; \boldsymbol{\nu})$ and optimization method for maximizing the variational lower bound leads to a trade-off intrinsic to VI. Simple variational approximations such as the mean-field family may be computationally feasible but inaccurate. The cost of increased accuracy, say by using a structured variational approximation, is increased computation. We illustrate the use of VI in statistical physics by comparing two choices of variational approximation, HVMs [16] and VANs [19]. Many other variational approximations can be explored in future work.

1 Hierarchical variational models

For studying correlated models such as frustrated spin systems, unstructured variational families such as the mean-field are insufficient. Hierarchical variational models (HVMs) are one way to model correlated latent variables. An HVM is defined by placing a 'variational prior' on the variational parameters $\boldsymbol{\nu}$ of the mean-field variational family, in analogy to hierarchical probabilistic models. By leveraging neural networks to parameterize the variational prior, HVMs can capture complex dependencies between random variables [16].

For studying a model $p(\mathbf{x}, \mathbf{z})$, the variational family defined by an HVM is defined as

$$q_{\mathrm{HVM}}(\mathbf{z}; \boldsymbol{\theta}) = \int q(\boldsymbol{\nu}; \boldsymbol{\theta}) \prod_i q(\mathbf{z}_i \mid \boldsymbol{\nu}_i) d\boldsymbol{\nu} \,, \tag{3}$$



(a) $\beta = 0.4$

(b) $\beta = 0.7$

Figure 2: Hierarchical variational models (HVMs) scale to larger systems than variational autoregressive network (VAN) models [19] when fit to the Sherrington-Kirkpatrick model using variational inference. (Lower is better, as the variational lower bound yields an upper bound on the free energy.) For systems of size $N = 16,384$, the VAN method did not complete a single iteration.

where $q_{\mathrm{MF}}(\mathbf{z} \mid \boldsymbol{\nu}) = \prod_i q(\mathbf{z}_i \mid \boldsymbol{\nu}_i)$ is the mean-field 'variational likelihood', and $q(\boldsymbol{\nu}; \boldsymbol{\theta})$ is the variational prior with parameters $\boldsymbol{\theta}$. Figure 1 shows the graphical model for HVMs as compared to the mean-field family graphical model.

To use an HVM in VI, the variational lower bound must be optimized. But the variational lower bound in Equation (2) requires calculating the entropy of the variational distribution, and such integration in high dimensions can be intractable. As detailed in Ranganth [16], the entropy can be lower-bounded by introducing an auxiliary 'variational posterior' distribution $r(\boldsymbol{\nu} \mid \mathbf{z}; \boldsymbol{\phi})$ with parameters $\boldsymbol{\phi}$. This leads to the hierarchical evidence lower bound,

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\nu}; \boldsymbol{\theta})}[\log p(\mathbf{x}, \mathbf{z}) + \log r(\boldsymbol{\nu} \mid \mathbf{z}; \boldsymbol{\phi}) - \log q(\mathbf{z} \mid \boldsymbol{\nu}) - \log q(\boldsymbol{\nu}; \boldsymbol{\theta})] \,, \tag{4}$$

and a stochastic optimization algorithm for this objective is developed in [16]. VI with an HVM requires specifying the variational prior $q(\boldsymbol{\nu}; \boldsymbol{\theta})$ and the variational posterior $r(\boldsymbol{\nu} \mid \mathbf{z}; \boldsymbol{\phi})$, then optimizing the hierarchical ELBO in Equation (4).

2 Empirical study

To study the utility of VI tools in physics, we compare HVMs to VAN approximations [19]. We use the same benchmarks as in [19]: the Sherrington-Kirkpatrick and Ising models. For the HVM, the variational prior $q(\boldsymbol{\nu}; \boldsymbol{\theta})$ is specified as an inverse autoregressive flow [8] and the variational posterior $r(\boldsymbol{\nu} \mid \mathbf{z}; \boldsymbol{\phi})$ is a masked autoregressive flow [14]. These choices lead to a complexity of $\mathcal{O}(L)$ for sampling latent variables in a system of size L. (This is because the noise used to sample from the variational prior can be drawn in parallel.) HVMs should therefore outperform VAN approximations in large systems, as the autoregressive requirement in VANs leads to a complexity of $\mathcal{O}(L^2)$.

To assess whether HVMs outperform VANs in large systems, the computational budget for the VI algorithm using both variational approximations was set to 6 hours. All experiments were performed on NVIDIA Tesla P100 GPUs, and the reference implementation of VANs released in Wu et al. [19] was used. VAN models were unable to complete many iterations in the allocated compute time, so all experiments were run without annealing the temperature of the system. For calculating the free energy using HVMs, importance sampling [13] was used with an HVMs as the proposal (for VANs, the increased cost of sampling prohibited drawing enough samples for low-variance importance sampling estimates).

Sherrington-Kirkpatrick model. The free energy estimates using VI with either HVM or VAN approximations are plotted in Figure 2 for the Sherrington-Kirkpatrick model. HVM approximations outperformed VAN approximations, and scaled to larger systems where the $\mathcal{O}(L^2)$ cost of sampling from a VAN prohibited even a single iteration.

Ising model. For small systems, HVMs were more accurate than VAN models at lower temperatures; at higher temperatures (such as the critical temperature), VAN models were slightly more accurate. This could be because annealing was not used to fit VAN models, and the randomness of the hierarchical latent variables in HVMs obviates the need for annealing. In large systems (e.g. $L = 128$), VAN models failed to complete a single iteration, while HVMs were able to complete many iterations (at the cost of some accuracy).

Discussion. The GBF inequality holds for quantum systems [4], and applying VI and HVMs to quantum systems is a direction for future work. Physics tools (such as VI in its original incarnation) have been useful in machine learning [1], and we hope the reverse holds—that tools from machine learning such as VI and HVMs continue to find use in statistical physics.

References

[1] Robert Bamler, Cheng Zhang, Manfred Opper, and Stephan Mandt. 2017. Perturbative Black Box Variational Inference. In *Neural Information Processing Systems*.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational Inference: A Review for Statisticians. *J. Amer. Statist. Assoc.*

[3] D. Chandler and D. Wu. 1987. *Introduction to Modern Statistical Mechanics*. Oxford University Press.

[4] R.P. Feynman. 2018. *Statistical Mechanics: A Set Of Lectures*. CRC Press.

[5] Matthew Hoffman and David Blei. 2015. Stochastic Structured Variational Inference. In *International Conference on Artifical Intelligence and Statistics*.

[6] M. Hoffman, D. Blei, C. Wang, and J. Paisley. 2013. Stochastic Variational Inference. *Journal of Machine Learning Research*.

[7] Jun Zhang. 1996. The application of the Gibbs-Bogoliubov-Feynman inequality in mean field calculations for Markov random fields. *IEEE Transactions on Image Processing*.

[8] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. 2016. Improved Variational Inference with Inverse Autoregressive Flow. In *Neural Information Processing Systems*.

[9] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. 2016. Auxiliary Deep Generative Models. In *International Conference on Machine Learning*.

[10] David MacKay. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

[11] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* 21, 6, 1087–1092.

[12] Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. 2019. Monte Carlo Gradient Estimation in Machine Learning. *arXiv:1906.10652*.

[13] Art B. Owen. 2013. *Monte Carlo theory, methods and examples*.

[14] George Papamakarios, Theo Pavlakou, and Iain Murray. 2017. Masked Autoregressive Flow for Density Estimation. *Advances in Neural Information Processing Systems 30*.

[15] Rajesh Ranganath, Dustin Tran, and David M Blei. 2016. Hierarchical variational models. In *International Conference on Machine Learning*.

[16] Rajesh Ranganth. 2018. *Black Box Variational Inference: Scalable, Generic Bayesian Computation and its Applications*. Ph.D. Dissertation. Princeton University.

[17] L. Saul, T. Jaakkola, and M. Jordan. 1996. Mean Field Theory for Sigmoid Belief Networks. *Journal of Artificial Intelligence Research* 4, 61–76.

[18] M. Wainwright and M. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1, 1–2, 1–305.

[19] Dian Wu, Lei Wang, and Pan Zhang. 2019. Solving Statistical Mechanics Using Variational Autoregressive Networks. *Phys. Rev. Lett.* 122, 080602.