Free Energy in Statistical Physics and Inference

Richard Turner

Random notes on the useage of the free energy in statistical physics and inference. Taken from David's book Chapter 31 on Ising models, Radford Neal's review of MCMC methods, and Yedidia, Freeman and Weiss's introduction to understanding belief propogation and its generalisations.

1 Statistical Physics

First of all we briefly summarise statistical physics:

micro-state = exact microscopic description of a physical system (precise velocity and position of each molecule in a gas for example) **macro-state** = description of a physical system sufficient to determine any macroscopic observable (the temperature, volume and mass of the gas, for example) The micro-state is generally unknowable and therefore must be handled using probabilites. One of the **goals of statistical physics is to related these two levels of description**

Every possible microstate s of the system has some definite energy $E(s)$. If the system is isolated then the energy is fixed at E_0, and the assumption is generally made that **all microstates with that energy are equally likely**: $P(s) = Z^{-1}\delta[E_0 - E(s)]$ (which is the microcanonical distribution).

If the system can exchange energy with a large reservoir that maintains the system at constant temperature, the system's energy can fluctuate and it is assumed: $P(s) = Z^{-1}\exp[-\beta E(s)]$, where the normalising constant is $Z = \sum_{\tilde{s}} \exp[-\beta E(\tilde{s})]$. This is called the canonical, **Boltzmann** or Gibbs distribution over microstates.

Intensive quantities are independent of system size (eg. temperature) **Extensive** quantities grow with system size eg. energy. If interactions are local then the growth will be linear for large systems and the values of extensive quantity per unit of system will reach limiting values. Extensive quantities are interesting at phase transitions.

Theory of universality: All systems with the same dimension and same symmetries have equivalent critical properties (ie. the scaling laws shown by their phase transitions are identical)

2 Ising Models

Let's consider the Ising models as an example:

$\mathbf{x}$ = a length N vector of spin up $+1$ and spin down -1 states the mth and nth spins may be neighbours in which case $J_{mn} = J$ otherwise $J_{m,n} = 0$. $J > 0$ ferromagnetic $J < 0$ antiferromagnetic.

$$\begin{aligned} P(\mathbf{x}|\beta, J, H) &= \frac{1}{Z(\beta, J, H)} \exp\left[-\beta E(\mathbf{x}, J, H)\right] & (1) \\ E(\mathbf{x}, J, H) &= -\left[\frac{1}{2}\sum_{m,n} J_{m,n} x_m x_n + \sum_n H x_n\right] & (2) \\ Z(\beta, J, H) &= \sum_{\mathbf{x}} E(\mathbf{x}, J, H) & (3) \\ \beta &= 1/K_B T & (4) \end{aligned}$$

Note the connection to Boltzmann machines when the couplings $J_{m,n}$ and applied fields h_n are free to take any value.

Here are several remarkable relationships from statistical physics:

The partition function ($Z(\beta)$) turns out to be a very useful function to know. For instance, here are some relations between the partition function and the heat capacity:

$$\begin{aligned} C &= \frac{\partial \langle E \rangle}{\partial T} & (5) \\ \langle E \rangle &= \sum_{\mathbf{x}} P(\mathbf{x}|\beta, J, H) E(\mathbf{x}, J, H) & (6) \\ \frac{\partial \ln Z}{\partial \beta} &= -\langle E \rangle & (7) \\ \frac{\partial^2 \ln Z}{\partial \beta^2} &= \langle E^2 \rangle - \langle E \rangle^2 & (8) \\ \frac{\partial \langle E \rangle}{\partial T} &= -\frac{\partial}{\partial T}\frac{\partial \ln Z}{\partial \beta} & (9) \\ &= -\frac{\partial \beta}{\partial T}\frac{\partial^2 \ln Z}{\partial \beta^2} & (10) \\ C &= \frac{var(E)}{K_B T^2} & (11) \end{aligned}$$

So the heat capacity is related to both the change in the energy of the system as we change its temperature AND to the fluctuations in the energy of a system at a particular temperature.

The free energy is another useful quantity, which is important whenever we use the Boltzmann distribution. Here are some relations between the partition function, the free energy and the entropy:

$$\begin{aligned} S &= -\sum_{\mathbf{x}} P(\mathbf{x}) \ln P(\mathbf{x}) && (12)\\ &= \ln Z(\beta) + \beta\langle E(\beta)\rangle && (13)\\ F &= -kT \ln Z && (14)\\ S &= -\frac{\partial F}{\partial \beta} && (15) \end{aligned}$$

Phase transitions are associated with a discontinuity in the derivative of the free energy. Note that the Boltzmann distribution can be shown to minimise the free energy: $F = U - TS$ using variational calculus.

3 Variational Free Energy

From $Z(\beta, J)$ any thermodynamic quantity can be derived. However, evaluating the normalising constant $Z(\beta, J)$ is difficult. Variational free energy minimisation is a method for *approximating* the complex distribution $P(\mathbf{x})$ by a simpler distribution $Q(\mathbf{x}, \theta)$.

$$\begin{aligned} \beta\tilde{F}(\theta) &= \sum_{\mathbf{x}} Q(\mathbf{x}; \theta) \ln \frac{Q(\mathbf{x}; \theta)}{\exp[-\beta E(\mathbf{x}; J)]} && (16)\\ &= \beta\langle E(\mathbf{x}; J)\rangle_Q - S_Q && (17)\\ &= D_{KL}(Q||P) + \beta F && (18) \end{aligned}$$

Via **Gibbs inequality** the variational free energy is bounded below by F and only attains this value when $Q(\mathbf{x}; \theta) = P(\mathbf{x})$ We choose Q st the sums over $\mathbf{x}$ are tractable and efficient. The simplest approximation, that Q is fully factored, leads to the **mean field approximation**.

4 Bethe Free Energy

The Bethe free enegy is an alternative way to approximate the free energy. To recap, in general:

$$F = -\log Z = \langle E\rangle_{P(x_1,x_2,x_3,...)} - S[P(x_1, x_2, x_3, ...)] \qquad (19)$$

where

$$P(x_1, x_2, x_3, ...) = \frac{1}{Z}\exp[-E(x_1, x_2, x_3, ...)] = \frac{1}{Z}P^* \tag{20}$$

We have met one interesting class of energy function for the Ising or Potts model:

$$E(x_1, x_2, x_3, ...) = \sum_{pairs} E(x_i, x_j) + \sum_i E(x_i) \tag{21}$$

In the image modelling community these models are called **Markov random fields**, P^* is called a **non-negative potential function** and a marginal like $P(x_1)$ is called the **belief**. The model can be represented graphically by an **undirected graph**. Each variable x_i corresponds to a **node** on the graph and an **edge** is placed between any two nodes which share an energy term $E(x_i, x_j)$. Marginalisation is handeled efficiently by a **local message passing algorithm** called **belief propogation**.

The basic idea behind the Bethe approximation is to **derive an estimate for the free energy $\tilde{F}_B$ that is a function of the marginals** $P(x_i) = 1/Z_i \exp[-E_i]$ **and pairwise marginals** $P(x_i, x_j) = 1/Z_{ij}\exp[-E_{ij} - E_i - E_j]$ **only**.

For Ising models (Markov Random Fields) the marginals are sufficient to determine the average energy exactly:

$$\begin{aligned}
\langle E \rangle &= \left\langle \sum_{pairs} E(x_i, x_j) + \sum_i E(x_i) \right\rangle && (22)\\
&= \sum_{pairs}\sum_{x_i,x_j} P(x_i, x_j)E(x_i, x_j) + \sum_i \sum_{x_i} P(x_i)E(x_i) && (23)\\
&= \sum_{pairs}\sum_{x_i,x_j} P(x_i, x_j)[-\log P(x_i, x_j)^* - E_i - E_j] + \sum_i \sum_{x_i} P(x_i)\log P(x_i)^* && (24)\\
&= -\sum_{pairs}\sum_{x_i,x_j} P(x_i, x_j)\log P(x_i, x_j)^* - [ne(x_i) - 1]\sum_i \sum_{x_i} P(x_i)\log P(x_i)^* && (25)
\end{aligned}$$

Where $ne(x_i)$ is the number of neighbours that node x_i has in the undirected graph.

The entropy is not so easy to obtain and this is where the approximation comes in. We could compute the entropy exactly if the joint could be expressed in terms of the marginals and pair-wise marginals. For **singly connected graphs** this is possible:

$$P(\mathbf{x}) = \frac{\prod_{pairs} P(x_i, x_j)}{\prod_i P(x_i)^{ne(x_i)-1}} \tag{26}$$

This relation can be understood as follows:

1. pick up the tree by a randomly chosen node
2. let all the other nodes hang-down
3. make the graph directed; the node you are holding is the root
4. write out the joint interms of the conditionals
5. convert the conditionals into marginals and pair-wise marginals: the result is of the form 26

Pluggin 26 into the expression for the entropy:

$$\begin{aligned} S_B &= -\langle \sum_{pairs} \log P(x_i, x_j) + \sum_i [ne(x_i) - 1] \log P(x_i) \rangle_{P(\mathbf{x})} && (27) \\ &= -\sum_{pairs} \sum_{x_i, x_j} P(x_i, x_j) \log P(x_i, x_j) - \sum_i [ne(x_i) - 1] \sum_{x_i} P(x_i) \log P(x_i) && (28) \\ & && (29) \end{aligned}$$

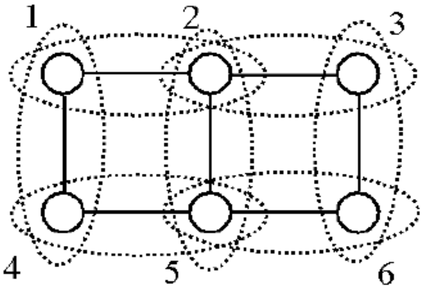
This expression is exact for singly connected graphs, and an approximation in graphs where there are loops. In contrast to variational free energies, **the Bethe free energy is not generally an upperbound on the true free energy**. Although it 'feels' a bit like a variational approach it is not. For instance, if we find the marginals and pair-wise marginals by minimising $\tilde{F}_B$ **we may recover a set of marginals and pair-wise marginals which are not consistent with any joint distribution**. It can be shown that belief propagation converges to stationary points of the Bethe free energy.

$$\begin{aligned} \tilde{F}_B &= \sum_{pairs} \sum_{x_i, x_j} P(x_i, x_j) [\log P(x_i, x_j) - \log P(x_i, x_j)^*] && (30) \\ & \quad - \sum_i [ne(x_i) - 1] \sum_{x_i} P(x_i) [\log P(x_i) - \log P(x_i)^*] && (31) \\ & && (32) \end{aligned}$$

5 Kikuchi Free Energy

Sometimes called the '**cluster variational**' method, the Kikuchi free energy approximation is a generalisation of the Bethe approximation. $\tilde{F}_K = \sum$ free energies of regions of N nodes. Looking again at the expression for $\tilde{F}_B$ we see it is equal to the sum of the free energies for each of the pair-wise regions, minus the free energies of the over counted single nodes (located where the pair-wise regions intersected). The Kikuchi approximation is constructed in an exactly analogous manner. It is best illustrated via some examples.

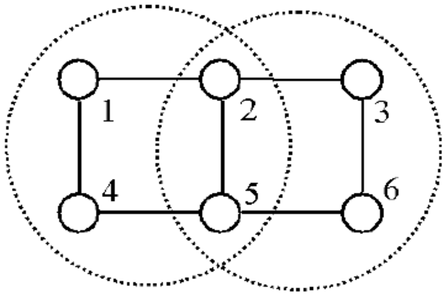
First, here's an example where we construct the Bethe approximation:



We have seven pairs (clusters) and this over-counts nodes 1, 3, 4 and 6 once and nodes 2 and 5 twice.

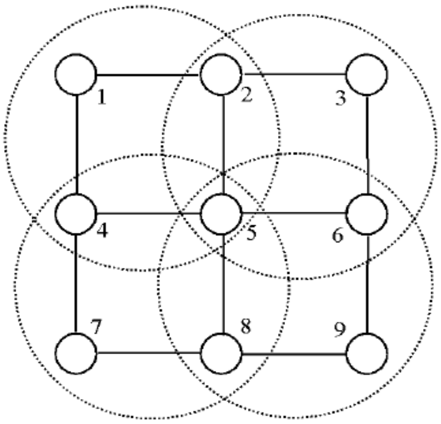
$$\begin{aligned}\tilde{F}_B &= F_{12} + F_{23} + F_{45} + F_{56} + F_{14} + F_{25} + F_{36} \quad (33)\\ &\quad -F_1 - 2F_2 - F_3 - F_4 - 2F_5 - F_6 \quad (34)\end{aligned}$$

And for the same model the Kikuchi approximation with clusters of size 4:



$$\tilde{F}_K = F_{1245} + F_{2356} - F_{25} \quad (35)$$

Finally for a more commplex model:



$$\tilde{F}_K = F_{1245} + F_{2356} + F_{4578} + F_{5869} - F_{25} - F_{45} - F_{56} - F_{85} + F_5 \quad (36)$$

In general, increasing the size of the clusters improves the approximation one obtains by minimising $\tilde{F}_K$. As we saw earlier, the Bethe free energy has an exact energy term for the Potts model and obviously this is true for the Kikuchi free energy too. The improvement comes from the treatment of the entropy which becomes increasingly accurate as the clusters become larger. When the size of the clusters is equal to the size of the largest loop in the model, the Kikuchi free energy is exact.

6 Correspondence of statistical thermodynamics with probabilistic inference

We can interpret the joint values of the random variables in a probabilistic inference problem as possible microstates of an **imaginary physical system**. Any probability distribution can then be regarded a canonical distribution with an energy: $E(s) = -\beta^{-1}[\log P(s) - \log(Z)]$ for any convenient choice for β (an arbitrary multiplication) and Z (an arbitrary offset). Usually, we set $\beta = 1$ and forget about it. eg.

Bayesian inference for model parameters:

$$\begin{aligned}
E(\theta) &= -\log\left[P(\theta)\prod_i P(x_i|\theta)\right] && (37)\\
Z &= \int d\theta\,\exp(-E(\theta)) && (38)\\
&= P(x_1...x_I) && (39)\\
P(\theta|x_1...x_I) &= \frac{P(\theta)\prod_i P(x_i|\theta)}{P(x_1...x_I)} && (40)\\
P(\theta|x_1...x_I) &= \frac{1}{Z}\exp[-E(\theta)] && (41)
\end{aligned}$$

The partition function is therefore the probability of the training data given the particular model being considered. Bayesian model comparison amounts to finding the partition function or, equivalently, the free energy or the entropy.