Belief Propagation and Bethe approximation part I

Caterina De Bacco

1 Introduction: $TrueSkill^{TM}$ Bayesian rating system

Consider the problem of rating chess players. The outcome of a game between k teams is $\mathbf{r} = (r_1, \dots, r_k)$, $r_i \in \{1, \dots, k\}$, where r_i is the rank of each team and $r_i = 1$ if i is the winner (best rank is the smallest one). Each team j is made of a subset $A_j \subset \{1, \dots, n\}$ of n players.

Let's assume that the probability $P(\mathbf{r}|\mathbf{s}, A)$ of a given outcome $\mathbf{r}$ depends on a set of hidden variables, the skills s_i of each player and the team assignments A.

Objective: infer the hidden variables $\mathbf{s}$ given the data $\mathbf{r}$ and A.

We can solve this inference problem by using a Bayesian approach as in Herbrich *et al.* (2007) and writing the posterior:

$$\begin{aligned} P(\mathbf{s}|\mathbf{r}, A) &= \frac{P(\mathbf{r}|\mathbf{s}, A)\, P(\mathbf{s})}{P(\mathbf{r}|A)} \qquad \text{Posterior of the skills} && (1) \\ P(\mathbf{s}) &= \prod_{i=1}^{n} \mathcal{N}(s_i|\mu_i, \sigma_i^2) \qquad \text{Prior over the skills} \quad . && (2) \end{aligned}$$

We further assume that the each player has a performance centered around its skill s_i:

$$p_i \sim \mathcal{N}(p_i; s_i, \beta^2) \quad . \tag{3}$$

A team's j performance is simply the sum over the players' performances:

$$t_j = \sum_{i \in A_j} p_i \quad . \tag{4}$$

Ordering the teams in ascending performance $r_1 \leq r_2 \leq \cdots \leq r_k$ and disregarding draws, the probability of an outcome is:

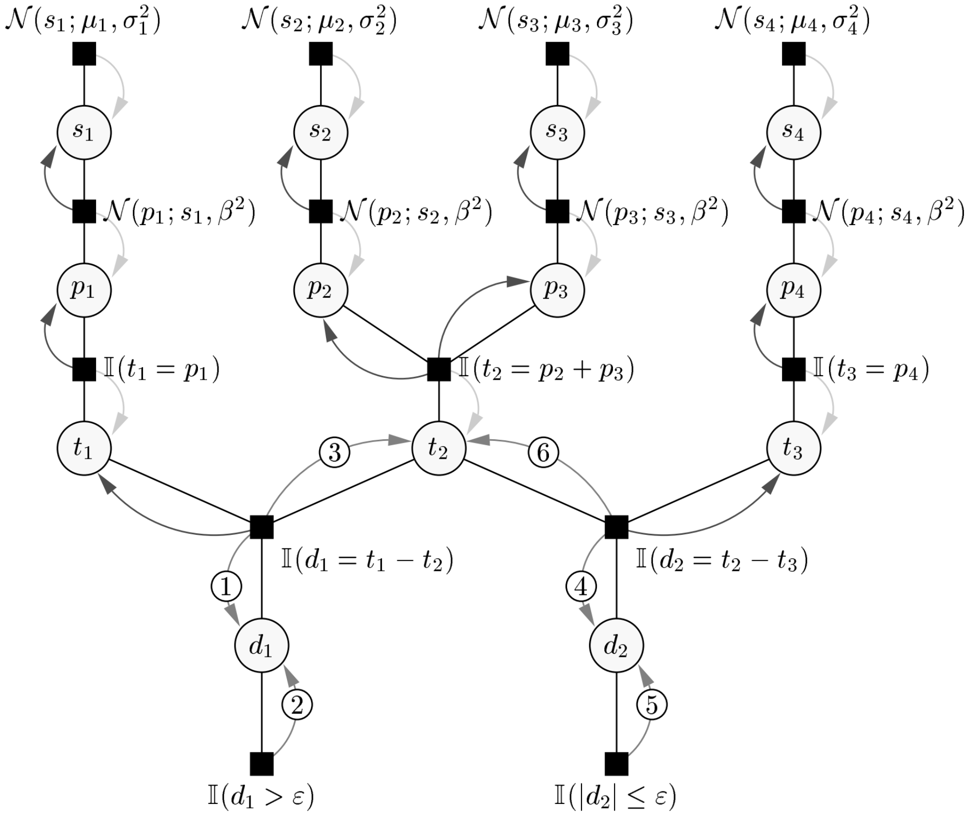
$$P(\mathbf{r}|\,\{t_1, \dots, t_k\}) = P(t_{r_1} > t_{r_2} > \cdots > t_{r_k}) \tag{5}$$

this is the likelihood of the data in the numerator of (1). If there is a draw, then the winning outcome $r_j < r_{j+1}$ requires $t_j > t_{j+1} + \epsilon$ and the draw outcome $r_j = r_{j+1}$ requires $|t_j - t_{j+1}| \leq \epsilon$, where $\epsilon > 0$ is the draw margin.

Example. Consider a game with 3 teams, 4 players and assignments $A_1 = \{1\}$, $A_2 = \{2, 3\}$ and $A_3 = \{4\}$. Let's assume that team 1 wins, i.e. $r_1 = 1$ and the other two draw, i.e. $r_2 = r_3 = 2$. Then the joint distribution $P(\mathbf{s}, \mathbf{p}, \mathbf{t}|\mathbf{r}, A)$ is given by the factor graph in Figure 1.

Goal: model the marginal $P(s_i|\mathbf{r}, A)$ of the posterior $P(\mathbf{s}|\mathbf{r}, A)$. In other words, we want to infer the distribution of each of the single skills. The posterior is obtained by integrating out the individuals' and teams' performances:

$$P(\mathbf{s}|\mathbf{r}, A) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} P(\mathbf{s}, \mathbf{p}, \mathbf{t}|\mathbf{r}, A)\, d\mathbf{p}\, d\mathbf{t} \quad . \tag{6}$$



$\mathcal{N}(s_1;\mu_1,\sigma_1^2)$
$\mathcal{N}(s_2;\mu_2,\sigma_2^2)$
$\mathcal{N}(s_3;\mu_3,\sigma_3^2)$
$\mathcal{N}(s_4;\mu_4,\sigma_4^2)$
s_1
s_2
s_3
s_4
$\mathcal{N}(p_1;s_1,\beta^2)$
$\mathcal{N}(p_2;s_2,\beta^2)$
$\mathcal{N}(p_3;s_3,\beta^2)$
$\mathcal{N}(p_4;s_4,\beta^2)$
p_1
p_2
p_3
p_4
$\mathbb{I}(t_1=p_1)$
$\mathbb{I}(t_2=p_2+p_3)$
$\mathbb{I}(t_3=p_4)$
t_1
3
t_2
6
t_3
$\mathbb{I}(d_1=t_1-t_2)$
$\mathbb{I}(d_2=t_2-t_3)$
1
4
d_1
d_2
2
5
$\mathbb{I}(d_1>\varepsilon)$
$\mathbb{I}(|d_2|\leq\varepsilon)$

Figure 1: An example TrueSkill factor graph. There are four types of variables: s_i for the skills of all players, p_i for the performances of all players, t_i for the performances of all teams and d_i for the team performance difference. The first row of factors encode the (product) prior; the product of the remaining factors characterizes the likelihood for the game outcome Team 1 > Team 2 = Team 3. The arrows indicate the optimal message passing schedule: first, all light arrow messages are updated from top to bottom. In the following, the schedule over the team performance (difference) nodes are iterated in the order of the numbers. Finally, the posterior over the skills is computed by updating all the dark arrow messages from bottom to top. Figure taken from Herbrich *et al.* (2007).

Not only this is hard to calculate by itself, it might be even harder to then marginalize over the $\mathbf{s} \setminus j$ to get the marginal $P(s_i|\mathbf{r},A)$!

So what do we do?

We could use the tools learned so far, but they might give bad results ... we can do better.

Hold on and will find out by the end of this lecture...

2 The Bethe Approximation: Variational approach

So far, we approximated the joint probability distribution $P(\mathbf{s})$ of a system of N variables with a factorized probability distribution $Q(\mathbf{s})$, a factorization of only one-variable function:

$$Q(\mathbf{s}) = \prod_{i=1}^{N} Q_i(s_i) \quad . \tag{7}$$

This was the main idea of the Mean Field variational approach (which was then further corrected using TAP).

The next natural step to improve this approximation, which was valid only in some restricted scenarios (e.g. weak couplings), is to consider also two-variable functions $b_{ij}(s_i,s_j)$ inside the factorization. We then consider a variational distribution of the form:

$$Q_{Bethe}(\mathbf{s}) = \prod_{ij} b_{ij}(s_i,s_j) \prod_{i=1}^{N} b_i(s_i)^{1-d_i} \quad , \tag{8}$$

where d_i is the degree of node i, i.e. the number of nodes with whom i is interacting.

As before, the goal is to find the best set of b_{ij} and b_i such that it minimizes the KL divergence $KL(Q||P)$. This is equivalent to minimizing the variational free energy:

$$G_{Bethe} := F[Q_{Bethe}] = E[Q_{Bethe}] - \frac{1}{\beta} S[Q_{Bethe}] \quad , \tag{9}$$

as defined in a previous lecture.

Before calculating an expression for G_{Bethe}, let's assume that the exact joint has a particular shape:

$$P(\mathbf{s}) = \prod_{ij} P_{ij}(s_i,s_j) \prod_{i=1}^{N} P_i(s_i)^{1-d_i} \quad , \tag{10}$$

where d_i is the degree of node i, i.e. the number of variables that i interacts with. This term is there to ensure that $P(\mathbf{s})$ is correctly normalized.
$P_{ij}(s_i,s_j)$ and $P_i(s_i)$ are two-variable and one-variable *marginals* respectively. This means that they are normalized to 1, and they agree, by definition of *marginal*, to a set of **consistency** equations of the type:

$$P_{ij}(s_i,s_j) = \sum_{\mathbf{s}\setminus\{i,j\}} P(\mathbf{s}) \tag{11}$$

$$P_i(s_i) = \sum_{s_j} P_{ij}(s_i,s_j) \quad . \tag{12}$$

Consider a generic Hamiltonian for the exact model:

$$H(\mathbf{s}) = -\sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i \quad , \tag{13}$$

from this we can derive the exact joint distribution $P(\mathbf{s}) = \frac{1}{Z} e^{-\beta H(\mathbf{s})}$ as usual.

Obs1: the expression $P(\mathbf{s})$ as in Equation (10) is an exact representation of $P(\mathbf{s}) = \frac{1}{Z} e^{-\beta H(\mathbf{s})}$ for tree structures, i.e. when there are no loops. It is a good approximation of that for *locally* tree-like structures, i.e. networks with no short loops, or, equivalently, with correlations between pairs of variables decaying fast enough.
Obs2: Equation (10) is similar to Q_{Bethe}, except the important difference that for Q_{Bethe} we did not assume that Q_{ij} and Q_i are consistent. The b_i and b_{ij} can be interpreted as *beliefs*, i.e. approximations for the marginals of the exact distribution.

Keeping in mind Equation (10), we now give the computations of the entropy and the average energy of P. First let's compute the entropy $S[P]$:

$$S[P] = -\sum_{\mathbf{s}} P(\mathbf{s}) \log(P(\mathbf{s})) \tag{14}$$

$$= -\mathbb{E}_P\left[\sum_{i<j} \log(P_{ij}(s_i,s_j)) + \sum_i (1-d_i)\log(P_i(s_i))\right] \tag{15}$$

$$= -\sum_{i<j}\sum_{s_i,s_j} P_{ij}(s_i,s_j)\log(P_{ij}(s_i,s_j)) - \sum_i (1-d_i)\sum_{s_i} P_i(s_i)\log(P_i(s_i)) \quad . \tag{16}$$

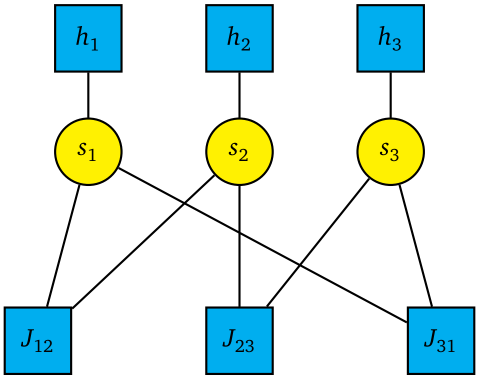


Figure 2: Factor graph example for a pairwise model with 3 variables. Function nodes a are the interaction couplings $J_{ij}(s_i, s_j)$ and external fields h_i.

Second, let's compute the internal energy $E[P]$:

$$\begin{aligned}
E[P] &= \sum_{\mathbf{s}} P(\mathbf{s})H(\mathbf{s}) && (17)\\
&= \mathbb{E}_P\Big[-\sum_{i<j} J_{ij}s_is_j - \sum_i h_is_i\Big] && (18)\\
&= -\sum_{i<j}\sum_{s_i,s_j} P_{ij}(s_i,s_j)\,J_{ij}s_is_j - \sum_i\sum_{s_i} P_i(s_i)\,h_is_i && (19)\\
&= -\sum_{i<j}\sum_{s_i,s_j} P_{ij}(s_i,s_j)\,\big[J_{ij}\,s_is_j + h_i\,s_i + h_j\,s_j\big] - \sum_i (1-d_i)\sum_{s_i} P_i(s_i)\,h_is_i && (20)\\
&= \sum_{i<j}\sum_{s_i,s_j} P_{ij}(s_i,s_j)\,E_{ij}(s_i,s_j) + \sum_i (1-d_i)\sum_{s_i} P_i(s_i)\,E_i(s_i) \quad , && (21)
\end{aligned}$$

where we defined:

$$\begin{aligned}
E_{ij}(s_i,s_j) &:= -(J_{ij}\,s_is_j + h_i\,s_i + h_j\,s_j) && (22)\\
E_i(s_i) &:= -h_i\,s_i \quad . && (23)
\end{aligned}$$

We developed the last step in order to write the internal energy in a more convenient way, as it will become clear later. The first term of eq. (21) is the average energy of each link, and the second term is the correction for the fact that the evidence is counted $d_i - 1$ times too many.

We now derive these functions for the above factorization (8).

$$\begin{aligned}
G_{bethe} &= E[Q_{Bethe}] - \frac{1}{\beta}S[Q_{Bethe}] && (24)\\
&= \sum_{ij}\sum_{s_i,s_j} b_{ij}(s_i,s_j)\Big[\frac{1}{\beta}\log(b_{ij}(s_i,s_j)) + E_{ij}(s_i,s_j)\Big] + \sum_i (1-d_i)\sum_{s_i} b_i(s_i)\Big[\frac{1}{\beta}\log(b_i(s_i)) + E_i(s_i)\Big] \quad .
\end{aligned}$$

2.1 Locally consistent marginals

Now, in order to define precisely the Bethe free energy G_{bethe}, we must consider a space of "possible" marginals, as the $P_i(s_i)$ and $P_{ij}(s_i, s_j)$ are. The starting choice is to restrict to the *locally consistent marginals*. For a factor graph $G(V, F, E)$ (see Figure 2 for a simple example) these are denoted as $LOC(G)$ and they are a collection of distributions $b_i(\cdot)$ over $\mathcal{X}$ for each variable node $i \in V$ and $b_a(\cdot)$ over $\mathcal{X}^{|\partial a|}$ for each functional node $a \in F$.

Since we want them to be distributions, they are non-negative, i.e. $b_i(s_i) \geq 0$ and $b_a(s_{\partial a}) \geq 0$ and they must satisfy the normalization conditions:

$$\sum_{s_i} b_i(s_i) = 1 \quad \forall i \in V \tag{25}$$

$$\sum_{s_{\partial a}} b_a(s_{\partial a}) = 1 \quad \forall a \in F \quad . \tag{26}$$

To be *locally consistent*, the must satisfy the following consistency condition for marginalization:

$$\sum_{s_{\partial a \setminus i}} b_a(s_{\partial a}) = b_i(s_i) \qquad \forall a \in F\,, \forall i \in \partial a\,. \tag{27}$$

Obs1: the marginals $P_i(s_i)$ and $P_{ij}(x_i, x_j)$ of any probability distribution are locally consistent. However, the converse is not true: one can find a set of locally consistent marginals that do not correspond to any probability distribution. To emphasize this point, locally consistent marginals are sometimes called **beliefs**.

Example. Consider the factor model of figure 3 on binary variables (s_1, s_2, s_3), $s_i \in \{0, 1\}$. They are locally consistent (exercise: check this), but they do not represent the marginals of any probability distribution. In other words, the beliefs are “unrealizable”. Why?
Any joint probability function for this example is a $P(s_1, s_2, s_3)$ over 8 possible configuration. If we calculate, for instance, $\sum_{s_3 \in \{0,1\}} P(s_1 = 0, s_2 = 1, s_3) = b_{12}(0, 1) = 0.01$.
This implies that $P(0, 1, 0) \leq 0.01$ and $P(0, 1, 1) \leq 0.01$. Similarly for $P(1, 0, 0) \leq 0.01$ and $P(1, 0, 1) \leq 0.01$.
Repeating for the term $b_{23}(s_2, s_3)$ yields: $P(0, 0, 1) \leq 0.01$ and $P(1, 0, 1) \leq 0.01$; $P(0, 1, 0) \leq 0.01$ and $P(1, 1, 0) \leq 0.01$.
Finally for $b_{31}(s_1, s_3)$ we get: $P(0, 0, 0) \leq 0.01$ and $P(0, 0, 1) \leq 0.01$; $P(1, 1, 0) \leq 0.01$ and $P(1, 1, 1) \leq 0.01$.
This means that for every one among the 8 possible configurations $P(s_1, s_2, s_3) \leq 0.01$, which means that $\sum_{\mathbf{s}} P(\mathbf{s}) \leq 0.08 \neq 1$. This answers to our question.

Obs2: in general, the b_i and b_{ij} are unrealizable unless the network is a tree.

b

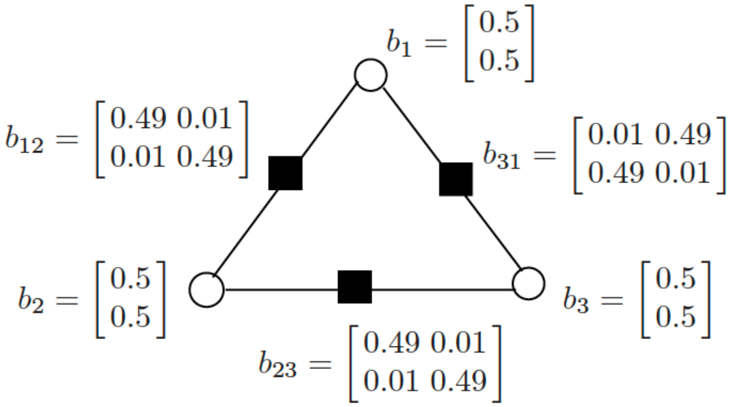

$b_1 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$
$b_{12} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix}$
$b_{31} = \begin{bmatrix} 0.01 & 0.49 \\ 0.49 & 0.01 \end{bmatrix}$
$b_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$
$b_3 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$
$b_{23} = \begin{bmatrix} 0.49 & 0.01 \\ 0.01 & 0.49 \end{bmatrix}$

Figure 3: A set of LOC(G) that cannot arise as marginals of any global probability distribution. Figure taken from Mezard and Montanari (2009).

For the pairwise model considered above, we get that the constraints on the beliefs are:

$$\begin{aligned}
\sum_{s_i} b_i(s_i) &= 1 \quad \forall i \in V \qquad (28)\\
\sum_{s_i,s_j} b_{ij}(s_i,s_j) &= 1 \quad \forall ij \qquad (29)\\
\sum_{s_j} b_{ij}(s_i,s_j) &= b_i(s_i) \quad \forall s_i \qquad (30)\\
\sum_{s_i} b_{ij}(s_i,s_j) &= b_j(s_j) \quad \forall s_j \quad . \qquad (31)
\end{aligned}$$

The minimization of the Eq. (24) becomes a constrained minimization problem that can be solved by introducing Lagrange multipliers γ_i and γ_{ij} which enforce the constraints:

$$\begin{aligned}
L \quad := \quad & G_{bethe} + \sum_i \gamma_i \Big(1 - \sum_{s_i} b_i(s_i)\Big) + \sum_{ij} \gamma_{ij} \Big(1 - \sum_{s_i,s_j} b_{ij}(s_i,s_j)\Big) \qquad (32)\\
& + \sum_{ij} \sum_{s_j} \lambda_{ji}(s_j) \Big(b_i(s_i) - \sum_{s_j} b_{ij}(s_i,s_j)\Big) + \sum_{ij} \sum_{s_i} \lambda_{ij}(s_i) \Big(b_j(s_j) - \sum_{s_i} b_{ij}(s_i,s_j)\Big) \quad , \qquad (33)
\end{aligned}$$

which we aim at minimizing.
This means solving the following equations:

$$\begin{aligned}
\frac{\partial L}{\partial b_i(s_i)} &= (1-d_i)\frac{1}{\beta}[1 + \log(b_i(s_i)] + (1-d_i)E_i(s_i) - \gamma_i + \sum_j \lambda_{ji}(s_i) \qquad (34)\\
\frac{\partial L}{\partial b_i(s_i)} &\equiv 0 \Longrightarrow b_i(s_i) = \exp\left[\beta \frac{-(1-d_i)E_i(s_i) + \gamma_i - \sum_j \lambda_{ji}(s_i)}{(1-d_i)} - 1\right]. \qquad (35)
\end{aligned}$$

Adding a normalization constant for enforcing the normalization condition and keeping only the Lagrange multipliers that depends on both indexes, we obtain:

$$b_i(s_i) = \frac{1}{Z_i} \exp\left[-\beta\, E_i(s_i) + \beta\, \frac{\sum_j \lambda_{ji}(s_i)}{(d_i - 1)} \right]. \tag{36}$$

With the same procedure for the derivative over $b_{ij}(s_i,s_j)$ we obtain:

$$b_{ij}(s_i,s_j) = \frac{1}{Z_{ij}} \exp\left\{ -\beta \left[E_{ij}(s_i,s_j) - \lambda_{ji}(s_i) - \lambda_{ij}(s_j) \right] \right\}. \tag{37}$$

The equations (36) and (37) depend on the Lagrange multipliers. To find an expression for them, we would need to enforce the consistency equations (30) and (31), but this increases the computational time for solving this problem.

Question: how do we minimize the Bethe free energy efficiently?
In order to answer this question, we present an algorithmic tool (*Belief Propagation algorithm*) that provides a possible solution.

A main reference for this lecture is Chapter 14 of Mezard and Montanari (2009).
The TrueSkill model is presented in Herbrich *et al.* (2007).

References

R. Herbrich, T. Minka, and T. Graepel, in *Advances in neural information processing systems* (2007) pp. 569–576.

M. Mezard and A. Montanari, *Information, physics, and computation* (Oxford University Press, 2009).

J. S. Yedidia, W. T. Freeman, and Y. Weiss, Exploring artificial intelligence in the new millennium **8**, 236 (2003).