The EP energy function and minimization schemes

Thomas P. Minka

August 7, 2001

Abstract

This note discusses the EP energy function in both its primal and dual forms and the connection to the Bethe free energy. It gives a theoretical justification of damping algorithms as well as a simple representation of Yuille's double-loop algorithm.

1 The energy function

EP tries to approximate

$$p(\mathbf{x}|D) = p(\mathbf{x}) \prod_i t_i(\mathbf{x}) \quad (1)$$

$$\approx p(\mathbf{x}) \prod_i \tilde{t}_i(\mathbf{x}) = q(\mathbf{x}) \quad (2)$$

where the $\tilde{t}_i(\mathbf{x})$ are in an exponential family:

$$\tilde{t}_i(\mathbf{x}) = \exp(\sum_j f_j(\mathbf{x})\tau_j) \quad (3)$$

The $t_i(\mathbf{x})$ must be positive but they do not have to be proper densities. However $p(\mathbf{x})$, which is not being approximated, does need to be a proper density. It doesn't have to be the actual "prior" in the problem; you could choose it to be 1, for example, if $\mathbf{x}$ has finite domain. $p(\mathbf{x})$ does not have to be in an exponential family, but it will make your life easier if it is.

The EP (dual) energy function is

$$\min_\nu \max_\lambda \ (n-1) \log \int_{\mathbf{x}} p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\nu_j) d\mathbf{x} - \sum_{i=1}^n \log \int_{\mathbf{x}} t_i(\mathbf{x}) p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\lambda_{ij}) d\mathbf{x} \quad (4)$$

$$\text{such that } (n-1)\nu_j = \sum_i \lambda_{ij} \quad (5)$$

The EP primal energy function is

$$\min_{\hat{p}_i} \max_q \sum_i \int_{\mathbf{x}} \hat{p}_i(\mathbf{x}) \log \frac{\hat{p}_i(\mathbf{x})}{t_i(\mathbf{x})p(\mathbf{x})} - (n-1) \int_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \quad (6)$$

$$\text{such that } \int_{\mathbf{x}} f_j(\mathbf{x}) \hat{p}_i(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{x}} f_j(\mathbf{x}) q(\mathbf{x}) d\mathbf{x} \quad (7)$$

$$\int_{\mathbf{x}} \hat{p}_i(\mathbf{x}) d\mathbf{x} = 1 \quad (8)$$

$$\int_{\mathbf{x}} q(\mathbf{x}) d\mathbf{x} = 1 \quad (9)$$

The EP primal is not very useful since it is an optimization over functions instead of vectors. This duality relationship is similar to that in maximum entropy modeling. It is proven in section 2.2.

2 Relationship with BP

This section derives the BP energy function from the EP energy function. For BP, the expectations $f_j(\mathbf{x})$ are delta functions, with $j = (s, c)$:

$$f_j(\mathbf{x}) = \delta(x_s - c) \tag{10}$$

For a pairwise MRF, each term $t_i(\mathbf{x})$ is a clique potential, with $i = (s, t)$:

$$t_i(\mathbf{x}) = \Psi_{st}(x_s, x_t) \tag{11}$$

Let n_s be the number of neighbors of x_s, i.e. the number of clique potentials involving x_s. Furthermore the prior is disconnected:

$$p(\mathbf{x}) = \prod_s \Psi_s(x_s) \tag{12}$$

Define the shorthand

$$\begin{aligned} \Phi_{st}(x_s, x_t) &= \Psi_{st}(x_s, x_t)\Psi_s(x_s)\Psi_t(x_t) && (13)\\ \nu(x_s) &= \nu_j \text{ where } j = (s, x_s) && (14)\\ \lambda_{st}(x_u) &= \lambda_{ij} \text{ where } i = (s, t),\ j = (u, x_u) && (15) \end{aligned}$$

Then the energy function simplifies to

$$\min_{\nu} \max_{\lambda} \ (n-1) \log \sum_{\mathbf{x}} \prod_s \Psi_s(x_s) \exp(\nu(x_s))$$

$$- \sum_{st} \log \sum_{\mathbf{x}} \Phi_{st}(x_s, x_t) \exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \prod_{u \neq s, u \neq t} \Psi_u(x_u) \exp(\lambda_{st}(x_u)) \tag{16}$$

$$\text{such that } (n-1)\nu(x_u) = \sum_{st} \lambda_{st}(x_u) \tag{17}$$

The sum over st means all distinct cliques. Adding Lagrange multipliers $\mu(x_u)$ for the constraints, we find the stationary condition

$$\frac{\Psi_u(x_u) \exp(\nu(x_u))}{\sum_u \Psi_u(x_u) \exp(\nu(x_u))} = \mu(x_u) \tag{18}$$

The stationary condition for $\lambda_{st}(x_u)$, $u \neq s, u \neq t$ is

$$\frac{\Psi_u(x_u) \exp(\lambda_{st}(x_u))}{\sum_{x_u} \Psi_u(x_u) \exp(\lambda_{st}(x_u))} = \mu(x_u) \tag{19}$$

This implies $\lambda_{st}(x_u) = \nu(x_u) + \text{const}$. Choosing the constant to be zero gives

$$\min_{\nu} \max_{\lambda} \sum_s (n_s - 1) \log \sum_{x_s} \Psi_s(x_s) \exp(\nu(x_s)) - \sum_{st} \log \sum_{x_s, x_t} \Phi_{st}(x_s, x_t) \exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \quad (20)$$

$$\text{such that } (n_s - 1)\nu(x_s) = \sum_t \lambda_{st}(x_s) \quad (21)$$

This is the BP energy function for an MRF.

The belief states are

$$b_s(x_s) \propto \Psi_s(x_s) \exp(\nu(x_s)) \quad (22)$$

$$q(\mathbf{x}) \propto p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\nu_j) \quad (23)$$

$$= \prod_s b_s(x_s) \quad (24)$$

$$b_{st}(x_s, x_t) \propto \Phi_{st}(x_s, x_t) \exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \quad (25)$$

$$\hat{p}_i(\mathbf{x}) \propto t_i(\mathbf{x}) p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\lambda_{ij}) \quad (26)$$

$$= b_{st}(x_s, x_t) \prod_{u \neq s, u \neq t} b_u(x_u) \quad (27)$$

2.1 Equivalence with Bethe free energy

This section shows that the BP energy function above is dual to the Bethe free energy of Yedidia et al:

$$\min_{b_{st}} \sum_{st} \sum_{x_s, x_t} b_{st}(x_s, x_t) \log \frac{b_{st}(x_s, x_t)}{\Phi_{st}(x_s, x_t)} - \sum_s (n_s - 1) \sum_{x_s} b_s(x_s) \log \frac{b_s(x_s)}{\Psi_s(x_s)} \quad (28)$$

$$\text{such that } \sum_{x_s} b_{st}(x_s, x_t) = b_t(x_t) \quad (29)$$

$$\sum_{x_t} b_{st}(x_s, x_t) = b_s(x_s) \quad (30)$$

$$\sum_{x_s} b_s(x_s) = 1 \quad (31)$$

The duality is based on the following representation of the KL-divergence:

$$\int_x p(x) \log \frac{p(x)}{q(x)} dx = \max_{\nu} \int_x p(x)\nu(x)dx - \log \int_x q(x) e^{\nu(x)} dx \quad (32)$$

To see that this representation is valid, the derivative wrt $\nu(x)$ is $p(x) - \frac{q(x)e^{\nu(x)}}{\int_x q(x)e^{\nu(x)}dx}$ and the second derivative is negative. Therefore the maximum is $\nu(x) = \log(p(x)/q(x)) + z$, for any z, at which point the two sides are equal.

Applying the duality to the first part gives

$$\sum_{x_s,x_t} b_{st}(x_s,x_t) \log \frac{b_{st}(x_s,x_t)}{\Phi_{st}(x_s,x_t)} = \max_{\lambda} \sum_{x_s,x_t} b_{st}(x_s,x_t)\lambda(x_s,x_t) - \log \sum_{x_s,x_t} \Phi_{st}(x_s,x_t) e^{\lambda(x_s,x_t)} \quad (33)$$

From the stationary conditions of the Bethe free energy, we know that $b_{st}(x_s,x_t)$ has the form

$$b_{st}(x_s,x_t) = \Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \quad (34)$$

Therefore without loss of generality we can restrict λ to decompose:

$$\lambda(x_s,x_t) = \lambda_{st}(x_s) + \lambda_{st}(x_t) \quad (35)$$

$$\begin{aligned}\sum_{x_s,x_t} b_{st}(x_s,x_t) \log \frac{b_{st}(x_s,x_t)}{\Phi_{st}(x_s,x_t)} &= \max_{\lambda} \sum_{x_s} b_s(x_s)\lambda_{st}(x_s) + \sum_{x_t} b_t(x_t)\lambda_{st}(x_t) \\ &\quad - \log \sum_{x_s,x_t} \Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \quad (36)\end{aligned}$$

With this transformation, b_{st} and its constraints disappear from the objective, and the minimization is over b_s instead.

Applying the duality to the second part of the objective gives

$$-\sum_{x_s} b_s(x_s) \log \frac{b_s(x_s)}{\Psi_s(x_s)} = \min_{\nu} -\sum_{x_s} b_s(x_s)\nu(x_s) + \log \sum_{x_s} \Psi_s(x_s)\exp(\nu(x_s)) \quad (37)$$

The transformed objective is now

$$\begin{aligned}\min_{b_s}\min_{\nu}\max_{\lambda} \sum_s \sum_{x_s} b_s(x_s) \sum_t \lambda_{st}(x_s) - \sum_{st} \log \sum_{x_s,x_t} \Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \\ + \sum_s (n_s - 1)\left(-\sum_{x_s} b_s(x_s)\nu(x_s) + \log \sum_{x_s} \Psi_s(x_s)\exp(\nu(x_s))\right) \quad (38)\end{aligned}$$

$$\text{such that } \sum_{x_s} b_s(x_s) = 1 \quad (39)$$

where the ordering of $\min_{b_s}\min_{\nu}\max_{\lambda}$ is arbitrary. The optimality conditions for ν and λ are:

$$\frac{\Psi_s(x_s)\exp(\nu(x_s))}{\sum_{x_s}\Psi_s(x_s)\exp(\nu(x_s))} = b_s(x_s) \quad (40)$$

$$\frac{\sum_{x_t}\Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s) + \lambda_{st}(x_t))}{\sum_{x_s,x_t}\Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s) + \lambda_{st}(x_t))} = b_s(x_s) \quad (41)$$

The same optimality conditions follow from adding the constraint

$$(n_s - 1)\nu(x_s) = \sum_t \lambda_{st}(x_s) \quad (42)$$

as long as we use the ordering $\min_{\nu}\max_{\lambda}$. With this constraint, b_s drops out of the objective and we are left with the desired dual (20).

2.2 Proof of EP duality

We can apply the techniques of the last section to the EP primal:

$$\min_{\hat{p}_i} \max_{q} \sum_i \int_{\mathbf{x}} \hat{p}_i(\mathbf{x}) \log \frac{\hat{p}_i(\mathbf{x})}{t_i(\mathbf{x})p(\mathbf{x})} - (n-1) \int_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \tag{43}$$

$$\text{such that } \int_{\mathbf{x}} f_j(\mathbf{x})\hat{p}_i(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}} f_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} \tag{44}$$

$$\int_{\mathbf{x}} \hat{p}_i(\mathbf{x})d\mathbf{x} = 1 \tag{45}$$

$$\int_{\mathbf{x}} q(\mathbf{x})d\mathbf{x} = 1 \tag{46}$$

Applying the KL duality to the first term gives

$$\int_{\mathbf{x}} \hat{p}_i(\mathbf{x}) \log \frac{\hat{p}_i(\mathbf{x})}{t_i(\mathbf{x})p(\mathbf{x})} = \max_{\lambda} \int_{\mathbf{x}} \hat{p}_i(\mathbf{x})\lambda_i(\mathbf{x})d\mathbf{x} - \log \int_{\mathbf{x}} t_i(\mathbf{x})p(\mathbf{x}) \exp(\lambda_i(\mathbf{x}))d\mathbf{x} \tag{47}$$

From the stationary conditions we can assume w.l.o.g. that

$$\lambda_i(\mathbf{x}) = \sum_j f_j(\mathbf{x})\lambda_{ij} \tag{48}$$

$$\int_{\mathbf{x}} \hat{p}_i(\mathbf{x}) \log \frac{\hat{p}_i(\mathbf{x})}{t_i(\mathbf{x})p(\mathbf{x})} = \max_{\lambda} \sum_j \lambda_{ij} \int_{\mathbf{x}} f_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} - \log \int_{\mathbf{x}} t_i(\mathbf{x})p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\lambda_{ij})d\mathbf{x} \tag{49}$$

With this transformation, $\hat{p}_i(\mathbf{x})$ and its constraints disappear from the objective. Applying the KL duality to the second term gives

$$-\int_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} = \min_{\nu} -\int_{\mathbf{x}} q(\mathbf{x})\nu(\mathbf{x})d\mathbf{x} + \log \int_{\mathbf{x}} p(\mathbf{x}) \exp(\nu(\mathbf{x}))d\mathbf{x} \tag{50}$$

where again we assume that

$$\nu(\mathbf{x}) = \sum_j f_j(\mathbf{x})\nu_j \tag{51}$$

$$-\int_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} = \min_{\nu} -\sum_j \nu_j \int_{\mathbf{x}} f_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} + \log \int_{\mathbf{x}} p(\mathbf{x}) \exp(\sum_j f_j(\mathbf{x})\nu_j)d\mathbf{x} \tag{52}$$

To eliminate $q(\mathbf{x})$ we add the constraint

$$(n-1)\nu_j = \sum_i \lambda_{ij} \tag{53}$$

and we get the desired dual (4).

3 Minimizing the BP energy

Define the BP message

$$m_{t\to s}(x_s) = \sum_{x_t} \Psi_{st}(x_s, x_t)\Psi_t(x_t)\exp(\lambda_{st}(x_t)) \tag{54}$$

Then the optimality condition for $\lambda_{st}(x_s)$ can be written

$$\frac{\Psi_s(x_s)\exp(\lambda_{st}(x_s))m_{t\to s}(x_s)}{\sum_{x_s}\Psi_s(x_s)\exp(\lambda_{st}(x_s))m_{t\to s}(x_s)} = \frac{\Psi_s(x_s)\exp(\nu(x_s))}{\sum_{x_s}\Psi_s(x_s)\exp(\nu(x_s))} \tag{55}$$

whose solution is

$$\lambda_{st}(x_s) = \nu(x_s) - \log m_{t\to s}(x_s) + z \tag{56}$$

for any constant z which we can choose to be zero. Combining this with the constraint (42) gives

$$\nu(x_s) = \sum_t \log m_{t\to s}(x_s) \tag{57}$$

Belief Propagation is simply the repeated application of these updates.

3.1 Damped BP

By using a smarter optimization scheme, we can get a damped form of BP. The optimality condition (41) for $\lambda_{st}(x_s)$ can be written in terms of the messages as

$$\frac{\exp(\lambda_{st}(x_s))\Psi_s(x_s)m_{t\to s}(x_s)}{\sum_{x_s}\exp(\lambda_{st}(x_s))\Psi_s(x_s)m_{t\to s}(x_s)} = b_s(x_s) \tag{58}$$

whose solution is

$$\lambda_{st}(x_s) = \log\frac{b_s(x_s)}{\Psi_s(x_s)m_{t\to s}(x_s)} + z \tag{59}$$

for any constant z which we can choose to be zero. From the constraint $\sum_t \lambda_{st}(x_s) = (n_s - 1)\nu(x_s)$, we find

$$\log b_s(x_s) = \frac{(n_s-1)}{n_s}\nu(x_s) + \frac{1}{n_s}\sum_t \log m_{t\to s}(x_s) + \log\Psi_s(x_s) \tag{60}$$

$$\lambda_{st}(x_s) = \frac{(n_s-1)}{n_s}\nu(x_s) + \frac{1}{n_s}\left(\sum_t \log m_{t\to s}(x_s)\right) - \log m_{t\to s}(x_s) \tag{61}$$

This is a recursive equation, because $m_{t\to s}$ depends on other λ's. However, it does approximately capture the dependence of λ on ν. By substituting this equation into the objective we can derive an update for ν.

The update will minimize an upper bound. The second term of the objective can be upper bounded using Jensen's inequality:

$$-\sum_t \log \sum_{x_s,x_t} \Phi_{st}(x_s,x_t)\exp(\lambda_{st}(x_s)+\lambda_{st}(x_t)) \;\leq\; -\sum_t \sum_{x_s,x_t} b_{st}(x_s,x_t)\,(\lambda_{st}(x_s)+\lambda_{st}(x_t)) + \text{const.} \quad (62)$$

$$b_{st}(x_s,x_t) \;=\; \frac{\Phi_{st}(x_s,x_t)\exp(\lambda_{st}^{old}(x_s)+\lambda_{st}^{old}(x_t))}{\sum_{x_s,x_t}\Phi_{st}(x_s,x_t)\exp(\lambda_{st}^{old}(x_s)+\lambda_{st}^{old}(x_t))} \quad (63)$$

Note that the marginal $\sum_{x_t} b_{st}(x_s,x_t)$ does not depend on t:

$$\sum_{x_t} b_{st}(x_s,x_t) \;\propto\; \Psi_s(x_s)\exp(\lambda_{st}^{old}(x_s))m_{t\to s}(x_t) = \Psi_s(x_s)\exp(u(x_s)) \quad (64)$$

$$u(x_s) \;=\; \frac{(n_s-1)}{n_s}\nu^{old}(x_s) + \frac{1}{n_s}\sum_t \log m_{t\to s}(x_s) \quad (65)$$

The derivative of the bound with respect to $\nu(x_s)$ is

$$(n_s-1)\frac{\Psi_s(x_s)\exp(\nu(x_s))}{\sum_{x_s}\Psi_s(x_s)\exp(\nu(x_s))} - \frac{(n_s-1)}{n_s}\sum_t\sum_{x_t} b_{st}(x_s,x_t) \;=\; 0 \quad (66)$$

$$(n_s-1)\frac{\Psi_s(x_s)\exp(\nu(x_s))}{\sum_{x_s}\Psi_s(x_s)\exp(\nu(x_s))} - (n_s-1)\frac{\Psi_s(x_s)\exp(u(x_s))}{\sum_{x_s}\Psi_s(x_s)\exp(u(x_s))} \;=\; 0 \quad (67)$$

So the update is

$$\nu(x_s) \;=\; u(x_s) + z = \frac{n_s-1}{n_s}\nu^{old}(x_s) + \frac{1}{n_s}\sum_t \log m_{t\to s}(x_s) + z \quad (68)$$

where z is any constant. In practice, it is good to choose $z = -\sum_{x_s} u(x_s)$ or something similar so that ν stays within floating point limits. Normalizing the messages is a good idea too. The new BP algorithm is:

loop nodes s:

1. Collect messages $m_{t\to s}$ into s. (Initial messages are all 1.)
2. Update ν_s according to (68). This is the belief state for x_s.
3. Recompute $\lambda_{st}(x_s)$ for all neighbors t, according to (61). This is the partial belief state for x_s excluding t.
4. Send messages out to neighbors (54).

Note the similarity with the BP updates (57) and (56). The difference is that the new algorithm is "damped"—it changes the belief state only part of the way toward the messages. Note that the damping is done in the log domain.

We can obtain a general family of BP algorithms by considering an arbitrary damping factor β:

$$\nu(x_s) = \frac{\beta-1}{\beta}\nu^{old}(x_s) + \frac{1}{\beta}\sum_t \log m_{t\to s}(x_s) + z \tag{69}$$

$$\lambda_{st}(x_s) = \frac{\beta-1}{\beta}\nu(x_s) + \frac{1}{\beta}\left(\sum_t \log m_{t\to s}(x_s)\right) - \log m_{t\to s}(x_s) \tag{70}$$

Regular BP has $\beta = 1$ and the new algorithm has $\beta = n_s$.

In practice, the drawbacks of this method seem to be:

1. It requires a large number of iterations (exponential in β).
2. If the damping level is too high, it will seek a maximum of the free energy instead of a minimum.

3.2 Yuille's algorithm

Yuille's algorithm is variational bound optimization applied to the BP primal (the Bethe free energy itself). We break the primal into a convex and concave part and then upper bound the concave part with a line:

$$E_{vex} = \sum_{st}\sum_{x_s,x_t} b_{st}(x_s,x_t)\log\frac{b_{st}(x_s,x_t)}{\Phi_{st}(x_s,x_t)} + \sum_s\sum_{x_s} b_s(x_s)\log\frac{b_s(x_s)}{\Psi_s(x_s)} \tag{71}$$

$$E_{cave} = -\sum_s n_s\sum_{x_s} b_s(x_s)\log\frac{b_s(x_s)}{\Psi_s(x_s)} \leq -\sum_s n_s\sum_{x_s} b_s(x_s)\nu(x_s) \tag{72}$$

$$\nu(x_s) = \log\frac{b_s^{old}(x_s)}{\Psi_s(x_s)} \tag{73}$$

Adding Lagrange multipliers λ_{st} for the marginal constraints, we have the modified objective

$$\min_{b_{st},b_s}\overline{J} = \sum_{st}\sum_{x_s,x_t} b_{st}(x_s,x_t)\log\frac{b_{st}(x_s,x_t)}{\Phi_{st}(x_s,x_t)} + \sum_s\sum_{x_s} b_s(x_s)\log\frac{b_s(x_s)}{\Psi_s(x_s)} \tag{74}$$

$$-\sum_s n_s\sum_{x_s} b_s(x_s)\nu(x_s) + \sum_{st}\sum_{x_s}\lambda_{st}(x_s)\left(b_s(x_s) - \sum_{x_t} b_{st}(x_s,x_t)\right) \tag{75}$$

$$\text{such that } \sum_{x_s} b_s(x_s) = 1 \tag{76}$$

Solving for b_{st}, b_s gives

$$b_s(x_s) \propto \Psi_s(x_s) \exp(n_s \nu(x_s) - \sum_t \lambda_{st}(x_s)) \tag{77}$$

$$b_{st}(x_s, x_t) \propto \Phi_{st}(x_s, x_t) \exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) \tag{78}$$

We still need to solve for λ. From the marginal constraints, we have

$$\sum_{x_t} \Phi_{st}(x_s, x_t) \exp(\lambda_{st}(x_s) + \lambda_{st}(x_t)) = \Psi_s(x_s) \exp(n_s \nu(x_s) - \sum_j \lambda_{sj}(x_s)) \tag{79}$$

$$\exp(\lambda_{st}(x_s)) m_{t \to s}(x_s) = \exp(n_s \nu(x_s) - \lambda_{st}(x_s) - \sum_{j \neq t} \lambda_{sj}(x_s)) \tag{80}$$

$$2\lambda_{st}(x_s) = n_s \nu(x_s) - \sum_{j \neq t} \lambda_{sj}(x_s) - \log m_{t \to s}(x_s) \tag{81}$$

Equation (80) comes from definition (54). The last equation is not a closed-form solution but only a consistency condition that the λ's must satisfy. We can find the optimal λ's by iterating this equation.

Now we have Yuille's algorithm:

Outer loop $\nu^{new}(x_s) = n_s \nu(x_s) - \sum_t \lambda_{st}(x_s)$

Inner loop $2\lambda_{st}^{new}(x_s) = n_s \nu(x_s) - \sum_{j \neq t} \lambda_{sj}(x_s) - \log m_{t \to s}(x_s)$

It isn't necessary to normalize ν at each step, but in order to avoid numerical overflow I subtract a constant so that $\sum_{x_s} \nu(x_s) = 0$. Unlike BP and damped BP, Yuille's algorithm satisfies the constraints during its search.

The drawback of this algorithm in practice is that it requires even more iterations than damping does (about 10 times more).

4 Gaussian EP

This section considers the energy function in the Gaussian case. For spherical Gaussians, ν_1 and λ_{i1} are vectors, ν_2 and λ_{i2} are scalars.

$$f_1(\mathbf{x}) = \mathbf{x} \qquad f_2(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{x} \tag{82}$$

$$p(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}_0, v_0\mathbf{I}) \tag{83}$$

$$q(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}_x, v_x\mathbf{I}) \propto p(\mathbf{x})\exp(\nu_1^{\mathrm{T}}\mathbf{x} + \nu_2\mathbf{x}^{\mathrm{T}}\mathbf{x}) \tag{84}$$

$$\nu_1 = \frac{\mathbf{m}_x}{v_x} - \frac{\mathbf{m}_0}{v_0} \tag{85}$$

$$-2\nu_2 = v_x^{-1} - v_0^{-1} \tag{86}$$

$$q^{\backslash i}(\mathbf{x}) \sim \mathcal{N}(\mathbf{m}_x^{\backslash i}, v_x^{\backslash i}\mathbf{I}) \propto p(\mathbf{x})\exp(\lambda_{i1}^{\mathrm{T}}\mathbf{x} + \lambda_{i2}\mathbf{x}^{\mathrm{T}}\mathbf{x}) \tag{87}$$

$$\lambda_{i1} = \frac{\mathbf{m}_x^{\backslash i}}{v_x^{\backslash i}} - \frac{\mathbf{m}_0}{v_0} \tag{88}$$

$$-2\lambda_{i2} = (v_x^{\backslash i})^{-1} - v_0^{-1} \tag{89}$$

Let's use the more convenient parameterization $(\mathbf{m}_x, v_x, \mathbf{m}_x^{\backslash i}, v_x^{\backslash i})$ instead of $(\nu_1, \nu_2, \lambda_{i1}, \lambda_{i2})$. The dual energy is

$$\int_{\mathbf{x}} p(\mathbf{x})\exp(\nu_1^{\mathrm{T}}\mathbf{x} + \nu_2\mathbf{x}^{\mathrm{T}}\mathbf{x})d\mathbf{x} = \frac{\mathcal{N}(0; \mathbf{m}_0, v_0\mathbf{I})}{\mathcal{N}(0; \mathbf{m}_x, v_x\mathbf{I})} \tag{90}$$

$$\int_{\mathbf{x}} t_i(\mathbf{x})p(\mathbf{x})\exp(\lambda_{i1}^{\mathrm{T}}\mathbf{x} + \lambda_{i2}\mathbf{x}^{\mathrm{T}}\mathbf{x})d\mathbf{x} = Z_i(\lambda)\frac{\mathcal{N}(0; \mathbf{m}_0, v_0\mathbf{I})}{\mathcal{N}(0; \mathbf{m}_x^{\backslash i}, v_x^{\backslash i}\mathbf{I})} \tag{91}$$

$$Z_i(\lambda) = \int_{\mathbf{x}} t_i(x)q^{\backslash i}(\mathbf{x})d\mathbf{x} \tag{92}$$

We can drop terms in $(\mathbf{m}_0, v_0)$ because the prior is fixed. Now "$\min_\nu \max_\lambda$" means "$\min_{\mathbf{m}_x, v_x} \max_{\mathbf{m}_x^{\backslash i}, v_x^{\backslash i}}$":

$$J = \min_\nu \max_\lambda \; -(n-1)\log\mathcal{N}(0; \mathbf{m}_x, v_x\mathbf{I}) \tag{93}$$

$$-\sum_{i=1}^n \log Z_i(\lambda) + \sum_{i=1}^n \log\mathcal{N}(0; \mathbf{m}_x^{\backslash i}, v_x^{\backslash i}\mathbf{I}) \tag{94}$$

$$J = \min_\nu \max_\lambda \; (n-1)\left(\frac{d}{2}\log v_x + \frac{\mathbf{m}_x^{\mathrm{T}}\mathbf{m}_x}{2v_x}\right) \tag{95}$$

$$-\sum_{i=1}^n \log Z_i(\lambda) - \sum_{i=1}^n \left(\frac{d}{2}\log v_x^{\backslash i} + \frac{(\mathbf{m}_x^{\backslash i})^{\mathrm{T}}\mathbf{m}_x^{\backslash i}}{2v_x^{\backslash i}}\right) \tag{96}$$

subject to

$$(n-1)\frac{\mathbf{m}_x}{v_x} + \frac{\mathbf{m}_0}{v_0} = \sum_{i=1}^n \frac{\mathbf{m}_x^{\backslash i}}{v_x^{\backslash i}} \tag{97}$$

$$(n-1)\frac{1}{v_x} + \frac{1}{v_0} = \sum_{i=1}^n \frac{1}{v_x^{\backslash i}} \tag{98}$$

4.1 Full Gaussian EP

For full Gaussians, ν_1 and λ_{i1} are vectors, ν_2 and λ_{i2} are matrices.

$$
\begin{aligned}
f_1(\mathbf{x}) = \mathbf{x} \qquad & f_2(\mathbf{x}) = \mathbf{x} \otimes \mathbf{x} && (99)\\
p(\mathbf{x}) &\sim \mathcal{N}(\mathbf{m}_0, \mathbf{V}_0) && (100)\\
q(\mathbf{x}) &\sim \mathcal{N}(\mathbf{m}_x, \mathbf{V}_x) \propto p(\mathbf{x}) \exp(\nu_1^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\nu_2\mathbf{x}) && (101)\\
\nu_1 &= \mathbf{V}_x^{-1}\mathbf{m}_x - \mathbf{V}_0^{-1}\mathbf{m}_0 && (102)\\
-2\nu_2 &= \mathbf{V}_x^{-1} - \mathbf{V}_0^{-1} && (103)\\
q^{\backslash i}(\mathbf{x}) &\sim \mathcal{N}(\mathbf{m}_x^{\backslash i}, \mathbf{V}_x^{\backslash i}) \propto p(\mathbf{x}) \exp(\lambda_{i1}^{\mathrm{T}}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\lambda_{i2}\mathbf{x}) && (104)\\
\lambda_{i1} &= (\mathbf{V}_x^{\backslash i})^{-1}\mathbf{m}_x^{\backslash i} - \mathbf{V}_0^{-1}\mathbf{m}_0 && (105)\\
-2\lambda_{i2} &= (\mathbf{V}_x^{\backslash i})^{-1} - \mathbf{V}_0^{-1} && (106)
\end{aligned}
$$

$$
\begin{aligned}
J = \min_{\nu} \max_{\lambda} \; & (n-1)\left(\frac{1}{2}\log|\mathbf{V}_x| + \frac{1}{2}\mathbf{m}_x^{\mathrm{T}}\mathbf{V}_x^{-1}\mathbf{m}_x\right) && (107)\\
& - \sum_{i=1}^{n} \log Z_i(\lambda) - \sum_{i=1}^{n}\left(\frac{1}{2}\log\left|\mathbf{V}_x^{\backslash i}\right| + \frac{1}{2}(\mathbf{m}_x^{\backslash i})^{\mathrm{T}}(\mathbf{V}_x^{\backslash i})^{-1}\mathbf{m}_x^{\backslash i}\right) && (108)
\end{aligned}
$$

$$
\text{subject to} \qquad (n-1)\mathbf{V}_x^{-1}\mathbf{m}_x + \mathbf{V}_0^{-1}\mathbf{m}_0 = \sum_{i=1}^{n}(\mathbf{V}_x^{\backslash i})^{-1}\mathbf{m}_x^{\backslash i} \qquad (109)
$$

$$
(n-1)\mathbf{V}_x^{-1} + \mathbf{V}_0^{-1} = \sum_{i=1}^{n}(\mathbf{V}_x^{\backslash i})^{-1} \qquad (110)
$$

4.2 Damped algorithm

As in section 3.1, define message variables:

$$
\begin{aligned}
\mathbf{m}_i &= \mathbf{m}_x^{\backslash i} + (v_x^{\backslash i} + v_i)\nabla_m \log Z_i && (111)\\
v_i &= (\nabla_m \nabla_m^{\mathrm{T}} - 2\nabla_v \log Z_i)^{-1} - v_x^{\backslash i} && (112)
\end{aligned}
$$

Holding $(\mathbf{m}_i, v_i)$ fixed and solving for λ gives

$$
\begin{aligned}
(v_x^{\backslash i})^{-1} &= v_x^{-1} - v_i^{-1} && (113)\\
\frac{\mathbf{m}_x^{\backslash i}}{v_x^{\backslash i}} &= \frac{\mathbf{m}_x}{v_x} - \frac{\mathbf{m}_i}{v_i} && (114)
\end{aligned}
$$

The constraints are now

$$
\begin{aligned}
(n-1)\nu_1 &= \sum_i \left(\frac{\mathbf{m}_x}{v_x} - \frac{\mathbf{m}_i}{v_i} - \frac{\mathbf{m}_0}{v_0}\right) && (115)\\
(n-1)(-2\nu_2) &= \sum_i \left(v_x^{-1} - v_i^{-1} - v_0^{-1}\right) && (116)
\end{aligned}
$$

Hold ν fixed and solve for $(\frac{\mathbf{m}_x}{v_x}, v_x^{-1})$ to get

$$(v_x^{\backslash i})^{-1} - v_0^{-1} = \frac{n-1}{n}(-2\nu_2) + \frac{1}{n}(\sum_i v_i^{-1}) - v_i^{-1} \quad (117)$$

$$\frac{\mathbf{m}_x^{\backslash i}}{v_x^{\backslash i}} - \frac{\mathbf{m}_0}{v_0} = \frac{n-1}{n}\nu_1 + \frac{1}{n}(\sum_i \frac{\mathbf{m}_i}{v_i}) - \frac{\mathbf{m}_i}{v_i} \quad (118)$$

Substitute this into the objective and solve for ν to get

$$\nu_1 = \frac{n-1}{n}\nu_1^{old} + \frac{1}{n}\sum_i \frac{\mathbf{m}_i}{v_i} \quad (119)$$

$$(-2\nu_2) = \frac{n-1}{n}(-2\nu_2^{old}) + \frac{1}{n}\sum_i v_i^{-1} \quad (120)$$

According to this derivation, damping of EP should be done in the natural parameterization, not the (mean,variance) parameterization. This algorithm has been tested on the clutter problem and it is effective in achieving convergence on difficult posteriors. However, the resulting approximation isn't very good anyway.