# APPENDIX: Truncated Message Passing

This appendix derives algorithm 2 from the main paper, which calculates the gradient of some loss function on the predicted marginals obtained after a fixed number of message-passing updates. The following derivation will be quite terse, as the fundamental idea of the algorithm follows that of reverse-mode automatic differentiation.

In this entire appendix, the dependence of factors $\psi$ and predicted marginals $\mu$ on the input $\mathbf{x}$ is suppressed for simplicity.

The basic iteration of TRW is [1, Theorem 7.2]

$$m_{t\to s}(y_s) \propto \sum_{y_t} \psi(y_t)\psi(y_t,y_s)^{\rho_{st}^{-1}} \frac{\prod_{v\in N(t)} m_{v\to t}(y_t)^{\rho_{vt}}}{m_{s\to t}(y_t)}.$$

After the messages have been iterated, one obtains predicted marginals via [1, Eqs. 7.13a & 7.13b]

$$\mu(y_s) \propto \psi(y_s) \prod_{v\in N(s)} m_{v\to s}(y_s)^{\rho_{vs}}.$$

$$\mu(y_s,y_t) \propto \psi(y_s)\psi(y_t)\psi(y_s,y_t)^{\rho_{ij}^{-1}} \frac{\prod_{v\in N(s)} m_{v\to s}(y_s)^{\rho_{vs}}}{m_{t\to s}(y_s)} \frac{\prod_{v\in N(t)} m_{v\to t}(y_t)^{\rho_{vt}}}{m_{s\to t}(y_t)}$$

We prefer to rewrite these rules in the following equivalent forms, which make the normalization steps explicit. Here, a superscript of "0" denotes a value before normalization.

$$\begin{aligned}
m^0_{t\to s}(y_s) &= \sum_{y_t} \psi(y_t)\psi(y_t,y_s)^{\rho_{st}^{-1}} \frac{\prod_{v\in N(t)} m_{v\to t}(y_t)^{\rho_{vt}}}{m_{s\to t}(y_t)} \\
m_{t\to s}(y_s) &= m^0_{t\to s}(y_s) / \sum_{y'_s} m^0_{t\to s}(y'_s) \\
\mu^0(y_s) &= \psi(y_s) \prod_{v\in N(s)} m_{v\to s}(y_s)^{\rho_{vs}} \\
\mu(y_s) &= \frac{\mu^0(y_s)}{\sum_{y'_s} \mu^0(y'_s)}
\end{aligned}$$

$$\mu^0(y_s,y_t) = \psi(y_s)\psi(y_t)\psi(y_s,y_t)^{\rho_{st}^{-1}} \frac{\prod_{v\in N(s)} m_{v\to s}(y_s)^{\rho_{vs}}}{m_{t\to s}(y_s)} \frac{\prod_{v\in N(t)} m_{v\to t}(y_t)^{\rho_{vt}}}{m_{s\to t}(y_t)}$$

$$\mu(y_s,y_t) = \frac{\mu^0(y_s,y_t)}{\sum_{y_s',y_t'}\mu^0(y_s',y_t')}$$

Before proceeding with the derivation, there are two lemmas that will be used repeatedly. First note the general rule for "back-propagating with respect to normalization":

**Lemma 1.** If $b_i = \frac{a_i}{\sum_j a_j}$, then

$$\frac{dL}{da_k} = \frac{dL}{db_k}\frac{1}{\sum_j a_j} - \sum_j \frac{dL}{db_j}\frac{a_j}{\left(\sum_j a_j\right)^2}. \tag{1}$$

Because this formula is somewhat awkward, we will simply make reference to it, rather than reproduce it in the algorithm.

A second lemma is

**Lemma 2.** If $y = \prod_i x_i^{a_i}$, then

$$\frac{dy}{dx_i} = (\prod_{j\neq i} x_j^{a_j}) a_i x_i^{a_i-1} = \frac{y}{x_i}a_i.$$

Now, suppose we have run message-passing for a fixed number of iterations. Now, we will calculate some loss function $L(\mu)$ of the beliefs, along with the partial derivatives

$$\frac{dL}{d\mu(y_s)} \text{ and } \frac{dL}{d\mu(y_s,y_t)}.$$

By application of Lemma 1, we can obtain

$$\frac{dL}{d\mu^0(y_s)} \text{ and } \frac{dL}{d\mu^0(y_s,y_t)}.$$

Holding the messages $m$ constant, we can recover the initial partial derivatives with respect to parameters

$$\begin{aligned}\frac{\partial L}{\partial\psi(y_s,y_t)} &= \frac{\partial L}{\partial\mu^0(y_s,y_t)}\frac{d\mu^0(y_s,y_t)}{d\psi(y_s,y_t)} \\ &= \frac{\partial L}{\partial\mu^0(y_s,y_t)}\frac{\mu^0(y_s,y_t)}{\psi(y_s,y_t)}\frac{1}{\rho_{st}}\end{aligned}$$

$$\begin{aligned}\frac{\partial L}{\partial\psi(y_s)} &= \frac{\partial L}{\partial\mu^0(y_s)}\frac{\partial\mu^0(y_s)}{\partial\psi(y_s)} + \sum_{v\in N(s)}\sum_{y_v}\frac{\partial L}{\partial\mu^0(y_s,y_v)}\frac{\partial\mu^0(y_s,y_v)}{\partial\psi(y_s)} \\ &= \frac{\partial L}{\partial\mu^0(y_s)}\frac{\mu^0(y_s)}{\psi(y_s)} + \sum_{v\in N(s)}\sum_{y_v}\frac{\partial L}{\partial\mu^0(y_s,y_v)}\frac{\mu^0(y_s,y_v)}{\psi(y_s)}\end{aligned}$$

Now, holding $\psi$ constant, we can initialize derivatives with respect to messages.

$$
\begin{aligned}
\frac{\partial L}{\partial m^0_{v\to s}(y_s)} &= \frac{\partial L}{\partial \mu^0(y_s)}\frac{\partial \mu^0(y_s)}{\partial m_{v\to s}(y_s)} + \sum_{t\in N(s)}\sum_{y_t}\frac{\partial L}{\partial \mu^0(y_s,y_t)}\frac{\partial \mu^0(y_s,y_t)}{\partial m_{v\to s}(y_s)} \\
&= \frac{\partial L}{\partial \mu^0(y_s)}\frac{\mu^0(y_s)}{m_{v\to s}(y_s)}\rho_{vs} + \sum_{t\in N(s)}\sum_{y_t}\frac{\partial L}{\partial \mu^0(y_s,y_t)}\frac{\mu^0(y_s,y_t)}{m_{v\to s}(y_s)}\big(\rho_{vs} - I[v=t]\big)
\end{aligned}
$$

Now, consider the single update of the messages from node $t$ to node $s$:

$$
\begin{aligned}
m^0_{t\to s}(y_s) &= \sum_{y_t}\psi(y_t)\psi(y_t,y_s)^{\rho_{st}^{-1}}\frac{\prod_{v\in N(t)} m_{v\to t}(y_t)^{\rho_{vt}}}{m_{s\to t}(y_t)} \\
m_{t\to s}(y_s) &= m^0_{t\to s}(y_s)/\sum_{y'_s} m^0_{t\to s}(y'_s)
\end{aligned}
$$

When reverse-propagating derivatives over this update, we must consider three "inputs": 1) Messages $m_{v\to t}$ into node $t$, 2) The univariate potentials $\psi(y_t)$, and 3) The bivariate potentials $\psi(y_t,y_s)$. We calculate these three derivatives separately.

$$
\begin{aligned}
\frac{\partial L}{\partial m_{v\to t}(y_t)} &= \sum_{y_s}\frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{\partial m^0_{t\to s}(y_s)}{\partial m^0_{v\to t}(y_t)} \\
&= \sum_{y_s}\frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{m^0_{t\to s}(y_s)}{m^0_{v\to t}(y_t)}\big(\rho_{vt} - I[v=s]\big)
\end{aligned}
$$

$$
\frac{\partial L}{\partial \psi(y_t)} = \sum_{y_s}\frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{\partial m^0_{t\to s}(y_s)}{\partial \psi(y_t)} = \frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{m^0_{t\to s}(y_s)}{\psi(y_t)}
$$

$$
\frac{\partial L}{\partial \psi(y_t,y_s)} = \frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{\partial m^0_{t\to s}(y_s)}{\partial \psi(y_t,y_s)} = \frac{\partial L}{\partial m^0_{t\to s}(y_s)}\frac{m^0_{t\to s}(y_s)}{\psi(y_t,y_s)}\frac{1}{\rho_{st}}
$$

## References

[1] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.