Derivation of the log prob gradient for sigmoid belief nets.

By Anne Paulson.

We want the log probability gradient for w_{ij}, the weight that goes from parent s_j to child s_i.

Let $p(R)$ be the probability that the nodes in the network other than s_i and its parents are whatever they are. Let $p(Pa)$ be the probability that the parents of s_i, including s_j, are whatever values they are, *given the rest of the network*. and let $p(s_i)$ be the probability that s_i has the value it has *given its parents*. (I should really write $P(Pa|R)$ and $P(s_i|Pa)$, but that would make a lot of notation into even more notation. And the focus of this writeup is the derivative.) Then the probability that the whole network N has the value it has is

$p(N) = p(Pa)p(R)p(s_i)$

The log of this is of course

$log(p(N)) = log(p(Pa)p(R)p(s_i)) = log(p(Pa)) + log(p(R)) + log(p(s_i))$

Let's take the derivative with respect to w_{ij}:

$\frac{d(log(p(N))}{dw_{ij}} = \frac{d}{dw_{ij}}(log(p(Pa)) + log(p(R)) + log(p(s_i)))$

Fortunately, the first two log terms are constants; they don't depend on w_{ij}. Goodbye! It doesn't matter a whit what the state of the rest of the network is, other than the parents and the child.

$\frac{d(log(p(N))}{dw_{ij}} = \frac{d}{dw_{ij}}(log(p(s_i)))$

Now I split the cases (there might be an easier way, but I'll do it this way). When $s_i = 1$, then $p(s_i) = p(s_i = 1) = \sigma(\Sigma w_{ik} s_k)$, where k ranges over the parents of our s_i.

$\frac{d(log(P(N))}{dw_{ij}} = \frac{d}{dw_{ij}}(log(\sigma(\Sigma w_{ik} s_k)))$ (when $s_i = 1$)

We do a simple chain rule. The derivative of log x is $\frac{1}{x}$; the derivative of sigmoid(x) is $\sigma(x)(1-\sigma(x))$; the derivative of $\Sigma w_{ik} s_k$ with respect to w_{ij} is s_j because all the other terms in the sum drop out. That gives us

$s_j \sigma(\Sigma w_{ik} s_k)(1 - \sigma(\Sigma w_{ik} s_k))\frac{1}{\sigma(\Sigma w_{ik} s_k)}$

The two sigmoids cancel, and we have:

$s_j(1 - \sigma(\Sigma w_{ik} s_k))$

Now $\sigma(\Sigma w_{ik} s_k) = p_i$; it's the probability that s_i would be on. So $(1 - \sigma(\Sigma w_{ik} s_k)) = 1 - p_i = s_i - p_i$. So we have

$\frac{d(log(P(N))}{dw_{ij}} = s_j(s_i - p_i)$ (when $s_i = 1$)

And, as I said, the case when $s_i = 0$ is almost identical. When $s_i = 0$, then $p(s_i) = p(s_i = 0) = 1 - \sigma(\Sigma w_{ik} s_k)$.

$\frac{d(log(P(N))}{dw_{ij}} = \frac{d}{dw_{ij}}(log(1 - \sigma(\Sigma w_{ik} s_k)))$ (when $s_i = 0$)

We do the chain rule as above, but this time there's a -1 factor because we're taking the derivative of minus sigmoid:

$s_j(-1)\sigma(\Sigma w_{ik} s_k)(1 - \sigma(\Sigma w_{ik} s_k))\frac{1}{1-\sigma(\Sigma w_{ik} s_k)}$

We cancel, and get:

$s_j(-\sigma(\Sigma w_{ik} s_k))$

which is the same as:

$s_j(0 - \sigma(\Sigma w_{ik} s_k))$

Now $\sigma(\Sigma w_{ik} s_k) = p_i$. So $(0 - \sigma(\Sigma w_{ik} s_k)) = 0 - p_i = s_i - p_i$. So we have $\frac{d(log(P(N))}{dw_{ij}} = s_j(s_i - p_i)$
and we're done. Whew.