1 Notes

2 LSTM

The forward / backwards passes for LSTM, with variables mirroring our code

2.1 forward pass

2.1.1 input (I), forget (F), cell activations

$$\begin{aligned}
a_I^t &= W_{I,X}x^t + W_{I,H}h^{t-1} + W_{I,S}s^{t-1} \\
b_I^t &= f(a_I^t) \\
\\
a_F^t &= W_{F,X}x^t + W_{F,H}h^{t-1} + W_{F,S}s^{t-1} \\
b_F^t &= f(a_F^t) \\
\\
a_Z^t &= W_{Z,X}x_i^t + W_{Z,H}h^{t-1} \\
b_Z &= f(a_Z^t)
\end{aligned}$$

2.1.2 state cell value

$$s^t = b_F^t s^{t-1} + b_I^t b_Z^t$$

2.1.3 output gate (ω)

$$\begin{aligned}
a_O^t &= W_{O,X}x^t + W_{O,H}h^{t-1} + W_{O,S}s^t \\
b_O^t &= f(a_O)
\end{aligned}$$

2.1.4 hidden cells

$$h^t = b_O^t f(s^t)$$

2.2 backwards pass

2.2.1 Hidden Block Output

$$\frac{dE}{dh^t} = \frac{dE}{\circ}\frac{\circ}{dh^t} + \frac{dE}{da_I^{t+1}}\frac{da_I^{t+1}}{dh^t} + \frac{dE}{da_F^{t+1}}\frac{da_F^{t+1}}{dh^t} + \frac{dE}{da_Z^{t+1}}\frac{da_Z^{t+1}}{dh^t} + \frac{dE}{da_O^{t+1}}\frac{da_O^{t+1}}{dh^t}$$

$$= \quad \frac{dE}{\circ}\frac{\circ}{dh^t} + \frac{dE}{da_I^{t+1}}W_{I,H} + \frac{dE}{da_F^{t+1}}W_{F,H} + \frac{dE}{da_Z^{t+1}}W_{Z,H} + \frac{dE}{da_O^{t+1}}W_{O,H}$$

2.2.2 Output Gate

$$\begin{aligned}
\frac{dE}{db_O^t} &= \frac{dE}{dh^t}\frac{dh^t}{db_O^t} = \frac{dE}{dh^t}f(s^t) \\
\frac{dE}{da_O^t} &= \frac{dE}{db_O^t}\frac{db_O^t}{da_O^t} \\
&= \frac{dE}{db_O^t}f_O'(a_O^t) \\
&= \frac{dE}{dh^t}f(s^t)f_\omega'(a_\omega^t)
\end{aligned}$$

2.2.3 state cells

$$\begin{aligned}
\frac{dE}{ds^t} &= \frac{dE}{dh^t}\frac{dh^t}{ds_t} + \frac{dE}{da_O^t}\frac{da_O^t}{ds^t} + \frac{dE}{ds^{t+1}}\frac{ds^{t+1}}{ds^t} + \frac{dE}{da_F^{t+1}}\frac{da_F^{t+1}}{ds^t} + \frac{dE}{da_I^{t+1}}\frac{da_I^{t+1}}{ds^t} \\
&= \frac{dE}{dh^t}b_O^t f'(s^t) + \frac{dE}{da_O^t}W_{O,S} + \frac{dE}{ds^{t+1}}b_F^{t+1} + \frac{dE}{da_F^{t+1}}W_{F,S} + \frac{dE}{da_I^{t+1}}W_{I,S}
\end{aligned}$$

2.2.4 cell activations

$$\begin{aligned}
\frac{dE}{db_z^t} &= \frac{dE}{ds^t}\frac{ds^t}{db_z^t} \\
&= \frac{dE}{ds^t}b_I^t \\
\frac{dE}{da_z^t} &= \frac{dE}{db_z^t}\frac{db_z^t}{da_z^t} \\
&= \frac{dE}{ds^t}b_I^t f_s'(a_z^t)
\end{aligned}$$

2.2.5 Forget Gates

$$\begin{aligned}
\frac{dE}{db_F^t} &= \frac{dE}{ds^t}\frac{ds^t}{db_F^t} \\
&= \frac{dE}{ds^t}s^{t-1} \\
\frac{dE}{da_F^t} &= \frac{dE}{db_F^t}\frac{db_F^t}{da_F^t}
\end{aligned}$$

$$\begin{aligned} &= \frac{dE}{db_F^t} f'(a_F^t) \\ &= \frac{dE}{ds^t} f'(a_F^t) s^{t-1} \end{aligned}$$

2.2.6 In Gates

$$\begin{aligned} \frac{dE}{db_I^t} &= \frac{dE}{ds^t}\frac{ds^t}{db_I^t} \\ &= \frac{dE}{ds^t} b_z^t \\ \frac{dE}{da_I^t} &= \frac{dE}{db_I^t}\frac{db_I^t}{da_I^t} \\ &= \frac{dE}{db_I^t} f'(a_I^t) \\ &= \frac{dE}{ds^t} b_z^t f'(a_I^t) \end{aligned}$$

3 HF-LSTM

3.1 f1 pass

3.1.1 input (in), forget (ϕ), cell

$$\begin{aligned} Ra_{in}^t &= V_{in,x}x^t + W_{in,h}Rb_h^{t-1} + V_{in,h}b_h^{t-1} + V_{in,s}s^{t-1} + W_{in,s}Rs^{t-1} \\ Rb_{in}^t &= f'_{in}(a_{in}^t)(Ra_{in}^t) \end{aligned}$$

$$\begin{aligned} Ra_\phi^t &= V_{\phi,x}x^t + W_{\phi,h}Rb_h^{t-1} + V_{\phi,h}b_h^{t-1} + V_{\phi,s}s^{t-1} + W_{\phi,s}Rs^{t-1} \\ Rb_\phi^t &= f'_\phi(a_\phi^t)(Ra_\phi^t) \end{aligned}$$

$$\begin{aligned} Ra_s^t &= V_{s,x}x_i^t + V_{s,h}b_h^{t-1} + W_{s,h}Rb_h^{t-1} \\ Rs^t &= Rb_\phi^t s^{t-1} + b_\phi^t Rs^{t-1} + Rb_{in}^t f(a_s^t) + b_{in}^t f'(a_s^t)(Ra_s^t) \end{aligned}$$

3.1.2 output gate (ω)

$$\begin{aligned} Ra_\omega^t &= V_{\omega,x}x^t + V_{\omega,h}b_h^{t-1} + W_{\omega,h}Rb_h^{t-1} + V_{\omega,s}s^t + W_{\omega,s}Rs^t \\ Rb_\omega^t &= f'_\omega(a_\omega)(Ra_\omega) \end{aligned}$$

3.1.3 hidden cells (trivial, but nonetheless)

$$\begin{aligned} Ra_h^t &= Rb_\omega^t f(s^t) + b_\omega^t f'(s^t)(Rs^t) \\ Rb_h^t &= Ra_h^t \end{aligned}$$

3.1.4 output cells

$$\begin{aligned} Ra_y^t &= V_{y,h} b_h^t + W_{y,h} Rb_h^t \\ Ry^t &= f_y'(a_y^t)(Ra_y^t) \end{aligned}$$

3.2 backwards pass

For gauss-newton method, take output of f1 pass, Ry^t, and push that through the normal backwards pass, rather than $\frac{dE}{dy^t} = d^t - y^t$. Use $\frac{RdE}{d*}$ variables instead of $\frac{dE}{d*}$

3.3 Pseudo-code

Algorithm 1 Hessian-Free LSTM

```
procedure HF-LSTM($a, b$)
    for $i \leftarrow 1, epochs$ do
        $g \leftarrow FullGradient$
        $x \leftarrow selectedHFBatch$
        procedure CONJ-GRAD($g, x, V, \lambda$)
            LATER
        end procedure
    end for
    procedure GDOTV($g, x, V$)
        $states \leftarrow f0pass(x)$
        $\delta' s \leftarrow bptt(x, states)$
        $Rs \leftarrow f1pass(x, states)$
        $GV \leftarrow bptt(Rs)$
    end procedure
end procedure
```