Assignment 1

Problem 1

Let $h \sim \mathcal{N}(0, I_N)$, $\Sigma \in \mathbb{R}^{N\times N} = U\Lambda U^T$ in the eigendecomposition, and $x = U\Lambda^{\frac{1}{2}}h + \mu$, for some $\mu \in \mathbb{R}^N$. We wish to show that $x \sim \mathcal{N}(\mu, \Sigma)$. Put another way, letting $f_h(h)$ and $f_x(x)$ represent the probability density functions of the distributions of h and x respectively, we wish to show that

$$f_x(x) = \frac{\exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]}{\sqrt{(2\pi)^N|\Sigma|}}.$$

Recall that when a random variable x is the result of a deterministic, bijective transformation g of another random variable h, $x = g(h)$, the probability density function of x takes the form

$$f_x(x) = f_h(g^{-1}(x))|\frac{dh}{dx}|.$$

In our case, $x = g(h) = U\Lambda^{\frac{1}{2}}h + \mu$, and so $h = g^{-1}(x) = (U\Lambda^{\frac{1}{2}})^{-1}(x-\mu)$. Then $\frac{dh}{dx} = (U\Lambda^{\frac{1}{2}})^{-1}$. The determinant of a matrix's inverse is just the inverse of the determinant, so $|\frac{dh}{dx}| = |U\Lambda^{\frac{1}{2}}|^{-1}$. Because $\Sigma = U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T$, we have that

$$\begin{aligned}
|U\Lambda^{\frac{1}{2}}| =& |\Lambda^{\frac{1}{2}}U^T| \\
=& \sqrt{|U\Lambda^{\frac{1}{2}}||\Lambda^{\frac{1}{2}}U^T|} \\
=& \sqrt{|U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^T|} \\
=& \sqrt{|\Sigma|}.
\end{aligned}$$

Then $|\frac{dh}{dx}| = \frac{1}{\sqrt{|\Sigma|}}$. Putting the pieces together, we have:

$$
\begin{aligned}
f_x(x) =& f_h(g^{-1}(x))|\frac{dh}{dx}| \\
=& f_h((U\Lambda^{\frac{1}{2}})^{-1}(x-\mu))|\frac{dh}{dx}| \\
=& \mathcal{N}((U\Lambda^{\frac{1}{2}})^{-1}(x-\mu); 0, I)|\frac{dh}{dx}| \\
=& \frac{\exp[-\frac{1}{2}((U\Lambda^{\frac{1}{2}})^{-1}(x-\mu))^T(U\Lambda^{\frac{1}{2}})^{-1}(x-\mu)]}{\sqrt{(2\pi)^N}}|\frac{dh}{dx}| \\
=& \frac{\exp[-\frac{1}{2}(x-\mu)^T(U\Lambda^{-\frac{1}{2}})(\Lambda^{-\frac{1}{2}}U^T)(x-\mu)]}{\sqrt{(2\pi)^N}}|\frac{dh}{dx}| \\
=& \frac{\exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]}{\sqrt{(2\pi)^N}}|\frac{dh}{dx}| \\
=& \frac{\exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]}{\sqrt{(2\pi)^N|\Sigma|}}
\end{aligned}
$$

as required. Here we used the fact that since U is orthonormal, its inverse is its transpose.

Problem 2

Once we know the joint distribution $p(x, h; \theta)$, the marginal distribution is simply $\sum_h p(x, h; \theta)$. So the bulk of our work will be in finding the joint distribution. Recall that a joint probability distribution can be decomposed as follows:

$$p(x, h; \theta) = p(x|h; \theta)p(h; \theta).$$

Intuitively, the probability of both x and h is the probability of first drawing h then, conditioned on h, drawing x.

In a Sigmoid belief network, each of the components of h is independent, and each component of x is independent when h is given. This allows us to write:

$$p(x, h; \theta) = \prod_{k=1}^{K_2} p(h_k; \theta) \prod_{j=1}^{K1} p(x_j|h; \theta).$$

When a random variable y is distributed $\text{Bern}(\sigma(b))$, we have that $p(y)$ is $\frac{1}{1+e^b}$ when $y = 0$, and $\frac{e^b}{1+e^b}$ when $y = 1$; in both cases, the numerator is equal to $\exp(yb)$. Then we can write for each component h_k of h and x_j of x:

$$
\begin{aligned}
p(h_k; \theta) =& \frac{\exp(h_k b_k^{(2)})}{1 + \exp(b_k^{(2)})} \\
p(x_j|h; \theta) =& \frac{\exp(x_j z_j)}{1 + \exp(z_j)}
\end{aligned}
$$

where $z_j = W_{j:}h + b_j^{(1)}$.
Plugging these in to the product from above, we get that

$$
\begin{aligned}
p(x, h; \theta) &= \prod_{k=1}^{K_2} \frac{\exp(h_k b_k^{(2)})}{1 + \exp(b_k^{(2)})} \prod_{j=1}^{K1} \frac{\exp(x_j z_j)}{1 + \exp(z_j)} \\
&= \frac{\exp(\sum_{k=1}^{K_2} h_k b_k^{(2)})}{\prod_{k=1}^{K_2}[1 + \exp(b_k^{(2)})]} \frac{\exp(\sum_{j=1}^{K_1} x_j z_j)}{\prod_{j=1}^{K_1}[1 + \exp(z_j)]} \\
&= \frac{\exp(h^T b^{(2)} + x^T z)}{\prod_{k=1}^{K_2}[1 + \exp(b_k^{(2)})] \prod_{j=1}^{K_1}[1 + \exp(z_j)]} \\
&= \frac{\exp(h^T b^{(2)} + x^T W h + x^T b^{(1)})}{\prod_{k=1}^{K_2}[1 + \exp(b_k^{(2)})] \prod_{j=1}^{K_1}[1 + \exp(W_{j:}h + b_j^{(1)})]}
\end{aligned}
$$

as required. Stick a $\sum_h$ in front to get the marginal distribution.

Problem 3

We consider $p(x_k|h)$; the formulation for $p(h_k|x)$ is almost identical.

Since x_k is a binary random variable, we know that it is Bernoulli-distributed: $p(x_k|h) = \text{Bern}(p(x_k = 1|h))$. We just need to calculate $p(x_k = 1|h)$.

We begin by observing that

$$
p(x_k = 1|h) = \frac{p(x_k = 1, h)}{p(x_k = 1, h) + p(x_k = 0, h)}.
$$

To calculate this, we need to know the marginal distribution $p(x_k, h)$ (marginalizing out all $x_{i \neq k}$):

$$
\begin{aligned}
p(x_k, h) &= \sum_{x_{i,i \neq k}} p(x_k, x_{-k}, h) \\
&= \sum_{x_{i,i \neq k}} \frac{\exp(\sum_j W_{j:}h + x_j b_j^{(1)}) \exp(h^T b^{(2)})}{Z(\theta)} \\
&= \sum_{x_{i,i \neq k}} \frac{\prod_j \exp(W_{j:}h + x_j b_j^{(1)}) \exp(h^T b^{(2)})}{Z(\theta)} \\
&= \sum_{x_{i,i \neq k}} \frac{\exp(W_{k:}h + x_k b_k^{(1)}) \prod_{j \neq k} \exp(W_{j:}h + x_j b_j^{(1)}) \exp(h^T b^{(2)})}{Z(\theta)} \\
&= \exp(W_{k:}h + x_k b_k^{(1)}) \sum_{x_{i,i \neq k}} \frac{\prod_{j \neq k} \exp(W_{j:}h + x_j b_j^{(1)}) \exp(h^T b^{(2)})}{Z(\theta)}
\end{aligned}
$$

We can pack all the terms that don't depend on x_k (*i.e.*, that entire sum) into a function $C(h)$, which allows us to write

$$p(x_k, h) = \exp(x_k W_{k:} h + x_k b_k^{(1)}) C(h).$$

We can now finally calculate $p(x_k = 1|h)$:

$$\begin{aligned}
p(x_k = 1|h) =& \frac{p(x_k = 1, h)}{p(x_k = 1, h) + p(x_k = 0, h)} \\
=& \frac{\exp(W_{k:} h + b_k^{(1)}) C(h)}{\exp(0) C(h) + \exp(W_{k:} h + b_k^{(1)}) C(h)} \\
=& \frac{C(h)}{C(h)} \frac{\exp(W_{k:} h + b_k^{(1)})}{1 + \exp(W_{k:} h + b_k^{(1)})} \\
=& \sigma(W_{k:} h + b_k^{(1)})
\end{aligned}$$

$p(x_k|h) = \mathrm{Bern}(p(x_k = 1|h)) = Bern(\sigma(W_{k:} h + b_k^{(1)}))$ as required. $p(h_k|x)$ can be derived similarly.

Problem 4

The expectation of a Bernoulli random variable $y \sim \mathrm{Bern}(p)$ is just p, the probability that it equals 1. In this case, since $h|x \sim \mathrm{Bern}(\sigma(W^T x + b^{(2)}))$, we have that $\mathbb{E}_{p(h|x)}[h] = \sigma(W^T x + b^{(2)})$.