Problem 3.2

So we first need to prove that matrix $\Psi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$ projects any N-dimensional vector v onto the subspace spanned by M columns of Φ (lets denote this subspace as $S(\Phi)$). Here we just assume $(\Phi^\top\Phi)^{-1}$ exists (i.e. $\Phi^\top\Phi$ is invertible) since it is a part of the definition of the matrix given in the problem condition.

Let us consider any N-dimensional vector v. We need to prove that there exist $\alpha_1 \dots \alpha_M \in \mathbb{R}$ such that $\Psi v = \alpha_1\varphi_1(D) + \ldots + \alpha_M\varphi_M(D)$. If we denote $\alpha = (\alpha_1, \ldots, \alpha_M)^\top$, then we need to prove there exists M-dimensional vector α such that $\Phi \cdot \alpha = \Psi v$. We notice now that $\alpha = (\Phi^\top\Phi)^{-1}\Phi^\top v$ is the very vector we are looking for, so it exists, so we proved that Ψ indeed projects v onto the subspace of columns of Φ.

Lets us now consider $w_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top t$. We need to prove $y = \Phi w_{ML}$ is an orthogonal projection of t onto the subspace of columns of Φ. This means, we need to prove $y - t \perp S(\Phi)$. This is the same as proving that $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t \perp S(\Phi)$.

Consider left part of the statement and multiply it by $\Phi^\top$. This gives us $\Phi^\top(\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t) = (\Phi^\top\Phi)(\Phi^\top\Phi)^{-1}\Phi^\top t - \Phi^\top t = \mathbb{0}$. So, we see that all the columns of Φ are orthogonal with $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t - t$, which means $\Phi(\Phi^\top\Phi)^{-1}\Phi^\top t$ is an orthogonal projection of t onto $S(\Phi)$.

Problem 3.3

Since w^* is extremum, we can equate E_D gradient to zero:

$$\nabla E_D = -\sum_{n=1}^{N} r_n(t_n - w^\top\varphi(x_n))\varphi^\top(x_n) = 0 \tag{1}$$

Let us consider $R = \begin{pmatrix} r_1 & 0 & ... & 0 \\ 0 & r_2 & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & r_n \end{pmatrix}$

We can rewrite equation 1 in the following way:

$$\nabla E_D = -\Phi^\top Rt + \Phi^\top R\Phi w = 0 \tag{2}$$

Thus we obtain

$$w^* = (\Phi^\top R\Phi)^{-1}\Phi^\top Rt \tag{3}$$

We can consider the matrix R, one the one hand, as inverse data-dependent noise variance: different x_i will have different correspondent r_i, and the smaller r_i is, the smaller is the impact of i_{th} sample. So, r_i can be used as our confidence in the t_i value.

On the other hand, at least when r_i is integer, it can be considered as the number of times sample (x_i, t_i) was present in the dataset.

Problem 3.4

Let us average error function over all possible noise values, i.e. let us compute it's expected value with respect to added noise $\{\epsilon_i\}$. Lets us denote $x'_n = x_n + \epsilon_n$ - input variable with added noise.

$$\mathbb{E}[E_D] = \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\{y(x'_n, w) - t_n\}^2] = \tag{4}$$

$$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) - t_n\}^2 = \tag{5}$$

$$= \mathbb{E}[\frac{1}{2}\sum_{n=1}^{N}\{w_0 + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) - t_n\} = \tag{6}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[\{w_0 - t_n + \sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni})\}^2] = \tag{7}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\mathbb{E}[(w_0 - t_n)^2 + (w_0 - t_n)\sum_{i=1}^{D} w_i(x_{ni} + \epsilon_{ni}) + \sum_{i,j=1}^{D} w_i w_j (x_{ni} + \epsilon_{ni})(x_{nj} + \epsilon_{nj})] = \tag{8}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{(w_0 - t_n)^2 + (w_0 - t_n)\sum_{i=1}^{D} w_i x_{ni} + \sum_{i,j=1}^{D} w_i w_j (x_{ni} x_{nj} + \delta_{ij}\sigma^2)\} = \tag{9}$$

$$= \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, w) - t_n\}^2 + \frac{N\sigma^2}{2} w^\top w \tag{10}$$

So, as expected, we see that error function, averaged over noise values, gives us weight-decayed sum-of-squares error function over noise-free input variables with omitted bias in regularization term, so minimizing the latter gives the same result as minimizing the former.

Problem 3.5

Suppose we want to minimize $E_D(w)$ subject to $\sum\limits_{j=1}^{M} |w_j|^q \leq \eta$ This is equivalent to minimizing Lagrange function $L(w, \lambda) = E_D(w) + \lambda'(\sum\limits_{j=1}^{M}(|w_j|^q - \eta)$ under conditions that $\lambda' \geq 0$, $\sum\limits_{j=1}^{M}|w_j|^q \leq \eta$ and $\lambda'(\sum\limits_{j=1}^{M}|w_j|^q - \eta) = 0$. When we use regularization, we suppose we only vary $\{w_j\}$ while keeping λ fixed. So we can substitute $\lambda' = \frac{\lambda}{2}$ and pay no attention to η since it is constant, and try to minimize the Lagrangian which now takes form of $E_D(w) + \frac{\lambda}{2}(\sum\limits_{j=1}^{M}(|w_j|^q)$, and this is exactly the regularized least squares error function.

To find the dependence between η and w we note that for optimal solution $\{w_j^*\}$ we have $\eta = \sum\limits_{j=1}^{M}(|w_j^*|^q)$.

As of dependence between η and λ we can see that the greater λ is, the more is regularization effect and so the less will be η.

Problem 3.6

This problem looks simple, but in fact to understand it deeply one need to perform certain actions. First of all, we will write down the log-likelihood of the dataset with given W:

$$L(D) = \prod_{n=1}^{N}(2\pi)^{-\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(t_n - W^\top\varphi(X_n))^\top\Sigma^{-1}(t_n - W^\top\varphi(X_n)) \tag{11}$$

$$\log L(D) = -\frac{kN}{2}\log(2\pi) - \frac{N}{2}|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(t_n - W^\top\varphi(X_n))^\top\Sigma^{-1}(t_n - W^\top\varphi(X_n)) \tag{12}$$

We now want to find the maximum of the log-likelihood with respect to W to find W_{ML}. To achieve this, we will use matrix derivatives notation (see `https://en.wikipedia.org/wiki/Matrix_calculus`)

In fact, we see that taking derivative of scalar with respect to matrix gives us a matrix of the same size:

$$\frac{\partial \log L(D)}{\partial W} = \begin{pmatrix} \frac{\partial \log L(D)}{\partial W_{11}} & \frac{\partial \log L(D)}{\partial W_{12}} & \dots & \frac{\partial \log L(D)}{\partial W_{1K}} \\ \frac{\partial \log L(D)}{\partial W_{21}} & \frac{\partial \log L(D)}{\partial W_{22}} & \dots & \frac{\partial \log L(D)}{\partial W_{2K}} \\ \dots & \dots & \dots & \dots \\ \frac{\partial \log L(D)}{\partial W_{M1}} & \frac{\partial \log L(D)}{\partial W_{M2}} & \dots & \frac{\partial \log L(D)}{\partial W_{MK}} \end{pmatrix}$$

To get the work done, we first prove following identity:

$$\frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X} = (C + C^\top)(Xa+b)a^\top \tag{13}$$

Here, C is a square matrix of some size $N \times N$, b is a vector of size N, a is a vector of size M, and X is $M \times N$ matrix.

To prove it, let's use write down the numerator in non-matrix form:

$$(Xa+b)^\top C(Xa+b) = \sum_{l=1}^{N}(\sum_{k=1}^{N}(\sum_{\beta=1}^{M}X_{l\beta}a_\beta + b_l)C_{kl})(\sum_{\gamma=1}^{N}X_{k\gamma}a_\gamma + b_k) \tag{14}$$

Taking now derivative of this with respect to X_{ij} we get:

$$\frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X_{ij}} = \sum_{k=1}^{N}a_jC_{ki}(X_ka+b_k) + \sum_{l=1}^{N}(X_la+b_l)C_{il}a_j = \tag{15}$$

$$= a_j[{C^i}^\top(Xa+b) + (Xa+b)^\top C_i] = \tag{16}$$

$$= (C + C^\top)_i(Xa+b)a_j \tag{17}$$

So, we see that $\frac{\partial (Xa+b)^\top C(Xa+b)}{\partial X} = (C + C^\top)(Xa+b)a^\top$.

Now, returning to the original problem, we have

$$\frac{\partial \log L(D)}{\partial W} = -\frac{1}{2}\sum_{n=1}^{N}(\Sigma^{-1} + {\Sigma^{-1}}^\top)(t_n - W^\top\varphi(X_n))\varphi(X_n)^\top = \tag{18}$$

$$= \Sigma^{-1}(T^\top - W^\top\Phi^\top)\Phi = 0 \tag{19}$$

Just a reminder, T is a $N \times K$ matrix, $\Phi(X)$ is a $N \times M$ matrix, and W is $M \times K$ matrix.

So assigning this to 0 and reducing Σ^{-1} we get $T^\top\Phi = W_{ML}^\top\Phi^\top\Phi$ and so $W_{ML} = (\Phi^\top\Phi)^{-1}\Phi^\top T$ which shows us that indeed i-th column of W_{ML} has the same well known form of $W_{ML_i} = (\Phi^\top\Phi)^{-1}\Phi^\top t_i$ and W_{ML} is independent of Σ.

Now let us consider

$$\frac{\partial \log L(D)}{\partial \Sigma^{-1}} = -\frac{N}{2}\frac{\partial \log|\Sigma|}{\partial \Sigma^{-1}} - \frac{1}{2}\sum_{n=1}^{N}(t_n - W_{ML}^\top\varphi(X_n))(t_n - W_{ML}^\top\varphi(X_n))^\top = \tag{20}$$

$$= \frac{N}{2}\Sigma - \frac{1}{2}(T^\top - W^\top\Phi^\top)(T^\top - W^\top\Phi^\top)^\top \tag{21}$$

So we obtain $\Sigma = \frac{1}{N}\sum\limits_{n=1}^{N}(t_n - W_{ML}^\top\varphi(X_n))(t_n - W_{ML}^\top\varphi(X_n))^\top$.

Problem 8.1

The idea is to start integrating from x_K down to x_1. Since there is just one factor containing x_K, which is $p(x_K|pa_K)$ we can easily integrate it with respect to x_K and we obviously get 1. After that, we get a reduced problem with $K-1$ factors, and by the same logic we can integrate sequentially with respect to $x_{K-1}...x_1$. On every step, we get a 1 factor after integration, so in the end we will get 1, which means the distribution is normalized.

Problem 8.2

Suppose the opposite, that there is a directed cycle in such a graph. Then we can start from any node on this cycle and start traversing it from this node along cycle edges. It is obvious that each edge will connect current node with a one having a greater number. But graph has a finite amount of nodes (lets say K), and so the sequence of nodes can not be ascending for more than K nodes, after which we will have an edge connecting two nodes with descending numbers. This is a contraction (there should be no such edges in a graph), so our assumption is incorrect and there can not exist a cycle in such directed graph.

Problem 8.5

The following probabilistic graphical model describes relevance vector machine introduced in the 7th chapter

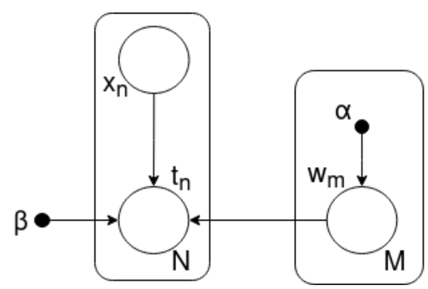

Figure 1: Relevance vector machine graphical model

Problem 8.6

As we can see, each variable x_i has an effect of multiplying negative term by $(1-\mu_i)^{x_i}$. If x_i is zero, this factor reduces to 1. But, if x_i is 1, this turns into $(1-\mu_i)$, which is the probability of x_i being zero. The greater this probability is (i.e. the smaller is μ_i) the greater will be the negative factor. So this accounts to the fact that x_i could be 1 because of noise with the probability $\mu_i - 1$. μ_0 is the probability of y being 1 when all $x_1..x_N$ are zeros.

Problem 8.8

From the statement of the problem we know that $p(a,b,c|d) = p(a|d)p(b,c|d)$, and we need to prove $p(a,b|d) = p(a|d)p(b|d)$.

To prove this, we just integrate over c the known equality: $\int\limits_c p(a,b,c|d) = \int\limits_c p(a|d)p(b,c|d) = p(a|d)\int\limits_c p(b,c|d) = p(a|d)p(b|d)$.

On the other hand, $\int\limits_c p(a,b,c|d) = p(a,b|d)$ and so the equality is proven.

Problem 8.9

Let us consider a node describing variable x and its Markov blanket. Let us consider any $y \notin Blanket(x)$ and show that any path from it to x is blocked. The last node on this path before x can be either:

- parent of x
- child of x

First suppose there is a path from y to x which includes some parent z of x as a last node in the path (except x). Such a path is blocked by the corresponding parent because the arrows on this path are head-to-tail or tail-to-tail (because z is a parent of x, the arrow goes from z to x).

Now let us consider the situation when the path from y to x ends with a child u of x. Then we need again to consider two options:

- the path enters u via arrow
- the path enters u via tail

In the first case, the path goes from v to u, where v is a co-parent of x. It can be seen that v blocks the path because it has a tail-to-tail or tail-to-head connection at v.

In the second case, the path has a tail-to-head connection at u and is blocked by u.

So, in any case, the path from y to x is blocked.

Problem 8.10

The only path from a to b goes through c. It enters c head-to-head and so when no variables are observed, this path is blocked by the second condition of D-separation.

When either c or d is observed, this path becomes unblocked, and so a starts to depend on b.

Problem 8.11

We have that

$$P(F, D, G, B) = P(F)P(B)P(G|F, B)P(D|G) \tag{22}$$

And we will use the following equation, inferred from the bayes' theorem application:

$$P(F|D) = \frac{P(F, D)}{P(D)} \tag{23}$$

$$\begin{aligned} P(F=0, D=0, B=0, G=0) &= 0.1 * 0.1 * 0.9 * 0.9 = 0.0081 \\ P(F=0, D=0, B=0, G=1) &= 0.1 * 0.1 * 0.1 * 0.1 = 0.0001 \\ P(F=0, D=0, B=1, G=0) &= 0.1 * 0.9 * 0.8 * 0.9 = 0.0648 \\ P(F=0, D=0, B=1, G=1) &= 0.1 * 0.9 * 0.2 * 0.1 = 0.0018 \end{aligned} \tag{24}$$

So, $P(F=0, D=0) = 0.0748$

$$
\begin{aligned}
&P(F=1, D=0, B=0, G=0) = 0.9 * 0.1 * 0.8 * 0.9 = 0.0648 \\
&P(F=1, D=0, B=0, G=1) = 0.9 * 0.1 * 0.2 * 0.1 = 0.0018 \\
&P(F=1, D=0, B=1, G=0) = 0.9 * 0.9 * 0.2 * 0.9 = 0.1458 \\
&P(F=1, D=0, B=1, G=1) = 0.9 * 0.9 * 0.8 * 0.1 = 0.0648
\end{aligned} \tag{25}
$$

So, $P(F=1, D=0) = 0.2772$

$$
P(F=0|D=0) = \frac{P(F=0, D=0)}{P(D=0)} = \frac{0.0748}{0.0748 + 0.2772} = 0.2125 \tag{26}
$$

Now let us suppose we also know, that the battery is flat ($B=0$):
$P(F=0, D=0, B=0) = 0.0082$
And thus we get

$$
P(F=0|D=0, B=0) = \frac{P(F=0, D=0, B=0)}{P(D=0, B=0)} = \frac{0.0082}{0.0666 + 0.0082} = 0.1096 \tag{27}
$$

So the probability of tank being empty in case we know battery is flat and driver reported the gauge shows empty is indeed lower than the same probability but when we do not know the battery is actually flat. The intuition behind is that if we know the battery is flat, it can be the reason why the gauge shows the tank is empty (whether it is really empty or not), and as driver's report is highly correlated with the gauge readings, we should have less belief in the driver's report results.

This means that $P(F|D, B) \neq P(F|D)$, so tank emptiness and battery flattness are not independent conditioned on driver's report, while they are independent inconditionally. This shows that conditioning on the decendants (D) of the collider (G) leads to unblocking the association path in graphical models.

Problem 8.12

For each pair of nodes, there either can be and edge or it can be absent in the graph. There are $\frac{M(M-1)}{2}$ pairs of nodes, which gives us $2^{\frac{M(M-1)}{2}}$ different graphs.

Problem 8.13

$$
E(\boldsymbol{x}, \boldsymbol{y}) = h \sum_i x_i - \beta \sum_{i,j} x_i x_j - \eta \sum_i x_i y_i \tag{28}
$$

Suppose all the variables are fixed except x_j. Let us denote $\boldsymbol{x_{-1}}$ the set of $\boldsymbol{x}$ variables with $x_j = -1$, and $\boldsymbol{x_1}$ the same set, but with $x_j = 1$. Then

$$
E(\boldsymbol{x_1}, \boldsymbol{y}) - E(\boldsymbol{x_{-1}}, \boldsymbol{y}) = h(1+1) - \beta \sum_{i:ne(j))} 2 \cdot x_i - \eta 2 \cdot y_j = 2h - 2\beta \sum_{i:ne(j))} x_i - 2\eta y_j \tag{29}
$$

As we see, this expression indeed depends only on the quantities, that are local to x_j in the graph: y_j and x_i which are x_j's neighbors.

Problem 8.14

Suppose $\beta = h = 0$. Then equation 28 turns into

$$E(\boldsymbol{x}, \boldsymbol{y}) = -\eta \sum_i x_i y_i \tag{30}$$

It is clear, that this value reaches its minimum when $x_i = y_i$ for all i, in which case the energy function value will be $-N$, where N is the number of pixels in the image. Since the probability of configuration is defined as $P(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z}\exp(-E(x, y))$, it reaches it's maximum when energy is minimal.

Problem 8.15

We just need to rearrange summations carefully:

$$\begin{aligned}
&P(x_{n-1}, x_n) = \sum_{x_i, i \notin \{n-1,n\}} \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\ldots\psi_{n-1,n}(x_{n-1}, x_n))\ldots\psi_{N-1,N}(x_{N-1}, x_N) \\
&= \frac{1}{Z}\psi_{n-1,n}(x_{n-1}, x_n)) \sum_{x_i, i \notin \{n-1,n\}} \psi_{1,2}(x_1, x_2)\ldots\psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \\
&\psi_{n,n+1}(x_n, x_{n+1}\ldots\psi_{N-1,N}(x_{N-1}, x_N) \\
&= \frac{1}{Z}\psi_{n-1,n}(x_{n-1}, x_n) \sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1}) \sum_{x_{n-3}} \psi_{n-3,n-2}(x_{n-3}, x_{n-2})\ldots \sum_{x_1} \psi_{1,2}(x_1, x_2) \\
&\sum_{x_1} \psi_{n,n+1}(x_n, x_{n+1})\ldots \sum_{x_N} \psi_{N-1,N}(x_{N-1}, x_N) = \frac{1}{Z}\psi_{n-1,n}(x_{n-1}, x_n)\mu_\alpha(x_{n-1})\mu_\beta(x_n)
\end{aligned} \tag{31}$$

Problem 8.16

Let us start with writing down the joint probability of $(x_n,\ x_N)$:

$$\begin{aligned}
p(x_n, x_N) = \frac{1}{Z} &\sum_{x_{n-1}} \psi_{n-1,n}(x_{n-1}, x_n) \sum_{x_{n-2}} \psi_{n-2,n-1}(x_{n-2}, x_{n-1})\ldots \\
&\sum_{x_1} \psi_{1,2}(x_1, x_2) \sum_{x_{n+1}} \psi_{n,n+1}(x_n, x_{n+1}) \sum_{x_{n+1}} \psi_{n+1,n+2}(x_{n+1}, x_{n+2})\ldots \\
&\sum_{x_{N-1}} \psi_{N-2,N-1}(x_{N-2}, x_{N-1})\psi_{N-1,N}(x_{N-1}, x_N)
\end{aligned} \tag{32}$$

As we see, the first part is exactly $\mu_\alpha(x_n)$. The second part though is a bit different from $\mu_\beta(x_n)$, so we need to introduce some changes into message definitions. Let us define $\mu_\beta(x_k, X_N)$ in the following way:

$$\mu_\beta(x_{N-1}, x_N) = \psi_{N-1,N}(x_{N-1}, x_N) \tag{33}$$

$$\mu_\beta(x_{k-1}, x_N) = \sum_{x_k} \psi_{k-1,k}(x_{k-1}, x_k)\mu_\beta(x_k, x_N) \qquad k < N \tag{34}$$

Then, we can rewrite the equation 32 as

$$p(x_n, x_N) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n, x_N) \tag{35}$$

Finally, we can write the conditional distribution as

$$p(x_n|x_N) = \frac{\mu_\alpha(x_n)\mu_\beta(x_n, x_N)}{\mu_\alpha(x_N)} \tag{36}$$

Problem 8.17

The only path from x_2 to x_5 in undirected chain graph goes through x_3, so if we condition on x_3, we block this path, and so $x_2 \perp\!\!\!\perp x_5 \,|\, x_3$.

Now let us consider $p(x_2|x_3, x_5)$ and show that it is independent of x_5.

$$p(x_2|\hat{x}_3, \hat{x}_5) = \frac{1}{Z}\mu_\alpha(x_2)\mu'_\beta(x_2) \tag{37}$$

$$\begin{aligned}\mu'_\beta(x_2) &= \sum_{x_3}\psi_{2,3}(x_2, x_3)I(x_3, \hat{x}_3)\sum_{x_4}\psi_{3,4}(x_3, x_4)\sum_{x_5}\psi_{4,5}(x_4, x_5)I(x_5, \hat{x}_5) \\ &= \psi_{2,3}(x_2, \hat{x}_3)\sum_{x_4}\psi_{3,4}(\hat{x}_3, x_4)\psi_{4,5}(x_4, \hat{x}_5)\end{aligned} \tag{38}$$

So, it seems that $p(x_2|\hat{x}_3, \hat{x}_5)$ still depends on the value of $\hat{x}_5$. But we haven't considered the form of Z, which probably will help.

$$\begin{aligned}Z &= \sum_{x_1 \dots x_5}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\psi_{3,4}(x_3, x_4)\psi_{4,5}(x_4, x_5)I(x_3, \hat{x}_3)I(x_5, \hat{x}_5) \\ &= \sum_{x_1, x_2}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, \hat{x}_3)\sum_{x_4}\psi_{3,4}(\hat{x}_3, x_4)\psi_{4,5}(x_4, \hat{x}_5)\end{aligned} \tag{39}$$

So we see, that the last factor, containing $\hat{x}_5$ reduces:

$$p(x_2|\hat{x}_3, \hat{x}_5) = \frac{\mu_\alpha(x_2)\psi_{2,3}(x_2, \hat{x}_3)}{\sum_{x_2}\mu_\alpha(x_2)\psi_{2,3}(x_2, \hat{x}_3)} \tag{40}$$

Problem 8.18

For directed trees, we have the following probablity distribution:

$$p(x) = \prod_{x_i} p(x_i|par(x_i)) \tag{41}$$

This also can be represented as undirected tree with the same edges (but with orientation abandoned), with potentials $\psi_{i,j}(x_j, x_i) = p(x_j|x_i)$ and normalizing constant $Z = 1$, where $x_i = parent(x_i)$ in the directed tree.

Let us consider the reverse process of moving from undirected tree to a directed tree. We can introduce orientation of every edge by selecting a single node as a root node, which will have no parents. After that, we can introduce directions over the edges in a bfs-order, strting from the selected root. For every pair of connected nodes x_i, x_j, $i < j$ we set $p(x_j|x_i) = \frac{\psi_{i,j}(x_i,x_j)}{\sum\limits_{x_j} \psi_{i,j}(x_i,x_j)}$ which gives us the conditional probabilities defining corresponsing directed graph edge.

Now, how many distinct directed trees one can get from a single undirected tree? It is obvious, that every root selection gives us a distinct oriented tree (since these trees will at least have different roots). On the other hand, after we have chosen the root, all the orientations are determined using bfs algorithm uniquely, so each root selection introduces just a single directed tree. So, we have a total of N distinct directed trees which can be geenrated from one undirected tree.

Problem 8.19

It's easy to see that we can get the same message functions $\mu_\alpha(x)$ and $\mu_\beta(x)$ using sum-product algorithm. Suppose we consider marginal distribution over x_n, thus making x_n to be the root of the factor-graph. The graph will have two branches: the left, where the messages will be passed from the leaf x_1 up to x_n, and the right, where the messages will be passed from the leaf x_N again to the root x_n.

For the left branch, we get

$$\begin{aligned}
\mu_{x_1 \to f_1}(x_1) &= 1 \\
\mu_{f_1 \to x_2}(x_2) &= \sum_{x_1} f_1(x_1, x_2)\mu_{x_1 \to f_1}(x_1) = \sum_{x_1} f_1(x_1, x_2) = \mu_\alpha(x_2) \\
\mu_{x_2 \to f_3}(x_2) &= \mu_{f_1 \to x_2} \\
&\dots \\
\mu_{f_{n-1} \to x_n}(x_n) &= \sum_{x_{n-1}} f_{n-1}(x_{n-1}, x_n)\mu_{x_{n-1} \to f_{n-1}}(x_{n-1}) = \mu_\alpha(x_n)
\end{aligned} \tag{42}$$

For the right branch, essentially, everything is similar:

$$\begin{aligned}
\mu_{x_N \to f_{N-1}}(x_N) &= 1 \\
\mu_{f_{N-1} \to x_{N-1}}(x_{N-1}) &= \sum_{x_N} f_{N-1}(x_{N-1}, x_N)\mu_{x_N \to f_{N-1}}(x_N) = \sum_{x_N} f_{N-1}(x_{N-1}, x_N) = \mu_\beta(x_{N-1}) \\
\mu_{x_{N-1} \to f_{N-2}}(x_{N-1}) &= \mu_{f_{N-1} \to x_{N-1}}(x_{N-1}) \\
&\dots \\
\mu_{f_n \to x_n}(x_n) &= \sum_{x_{n+1}} f_n(x_n, x_{n+1})\mu_{x_{n+1} \to f_n}(x_{n+1}) = \mu_\beta(x_n)
\end{aligned} \tag{43}$$

So, we obtain the same formula for marginal distribution: $p(x) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n)$

Problem 8.21

We can write the full distribution as

$$p(x) = f_s(x_{s_1}, ..., x_{s_m}) \prod_{i=1}^{m} G_{s_i}(x_{s_i}, X_{s_i}) \tag{44}$$

Applying marginalization over all the variables, except the ones contained in f_s, we get the marginal distribution we are interested in:

$$
\begin{aligned}
p(x_{s_1}, ..., x_{s_m}) &= \sum_{x \notin ne(f_s)} f_s(x_{s_1}, ..., x_{s_m}) \prod_{i=1}^{m} G_{s_i}(x_{s_i}, X_{s_i}) = f_s(x_{s_1}, ..., x_{s_m}) \prod_{i=1}^{m} \sum_{X_{s_i}} G_{s_i}(x_{s_i}, X_{s_i}) \\
&= f_s(x_{s_1}, ..., x_{s_m}) \prod_{i=1}^{m} \mu_{x_{s_i} \to f_s}(x_{s_i})
\end{aligned} \tag{45}
$$

Where the last transition is done, using formula 8.67.

Problem 8.22

Suppose we are interested in $p(x_{i_1}, ..., x_{i_k})$ where $x_{i_1} ... x_{i_k}$ belong to a connected subgraph. We start with reconsidering 8.66:

$$
\mu_{f_s \to x}(x, x_{i_1}, ..., x_{i_k}) = \sum_{\boldsymbol{x_s} \setminus \{x, x_{i_1}, ..., x_{i_k}\}} f_s(x, x_1, ..., x_M) \prod_{m \in ne(f_s) \setminus x} \mu_{x_m \to f_s}(x_m, x_{i_1}, ..., x_{i_k}) \tag{46}
$$

In the same way, we reconsider 8.69:

$$
\mu_{x_m \to f_s}(x_m, x_{i_1}, ..., x_{i_k}) = \prod_{l \in ne(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m, x_{i_1}, ..., x_{i_k}) \tag{47}
$$

The initial conditions are still the same, so no other changes should be done to the algorithm.

Problem 8.23

We start with rewriting 8.63, selecting one of the factors from the product (say having index q) and writing it separately:

$$
p(x) = \prod_{s \in ne(x)} \mu_{f_s \to x}(x) = \mu_{f_q \to x}(x) \prod_{s \in ne(x) \setminus q} \mu_{f_s \to x}(x) \tag{48}
$$

We now can apply 8.69, which says that $\mu_{x \to f_q}(x) = \prod_{s \in ne(x) \setminus q} \mu_{f_s \to x}(x)$. Combining these observations, we get

$$
p(x) = \mu_{f_q \to x}(x) \mu_{x \to f_q}(x) \tag{49}
$$

Side note: while solving the problem, I have figured out I did not understand completely how we obtained 8.66 formula, so I think it will be nice to write it down. We will start with substituting 8.65 into 8.64, but doing it a bit more explicitely than in the original book:

$$
\mu_{f_s \to x}(x) = \sum_{X_s} F_s(x, X_s) = \sum_{X_s} f_s(x, x_1, ..., x_M) G_1(x_1, X_{s_1}) ... G_M(x_M, X_{s_M}) \tag{50}
$$

The idea is now to look carefully at the picture 8.47 and figure out that we can present X_s as a union of non-overlapping sets: $X_s = \{x_1, x_2, \ldots, x_M\} \cup X_{s_1} \cup X_{s_2} \cup \ldots \cup X_{s_M}$. This means, that we can view the summation over X_s as a summation over all of this distinct sets, and this will be convenient since f_s depends only on $\{x_1, \ldots, x_M\}$, and G_m depends only on X_{s_m}.

$$\mu_{f_s \to x}(x) = \sum_{X_s} f_s(x, x_1, ..., x_M) G_1(x_1, X_{s_1}) ... G_M(x_M, X_{s_M}) = \\ = \sum_{x_1 ... x_m} f_s(x_1 ... x_m) \prod_m \sum_{X_{s_m}} G_m(x_m, X_{s_m}) \qquad w \tag{51}$$

Problem 8.25

Let us check that after applying sum-product algorithm we get correct marginal distribution for x_1:

$$\begin{aligned} \tilde{p}(x_1) &= \mu_{f_a \to x_1}(x_1) = \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2) = \sum_{x_2} f_a(x_1, x_2) \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\ &= \sum_{x_2} f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) \\ &= \sum_{x_2} \sum_{x_3} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \end{aligned} \tag{52}$$

The same goes for x_3:

$$\begin{aligned} \tilde{p}(x_3) &= \mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \to f_b}(x_2) = \sum_{x_2} f_b(x_2, x_3) \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\ &= \sum_{x_2} f_b(x_2, x_3) \sum_{x_1} f_a(x_1, x_2) \sum_{x_4} f_c(x_2, x_4) \\ &= \sum_{x_1} \sum_{x_2} \sum_{x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) \end{aligned} \tag{53}$$

$$\begin{aligned} p(x_1, x_2) &= \sum_{x_3, x_4} f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4) = f_a(x_1, x_2) \sum_{x_3, x_4} f_b(x_2, x_3) f_c(x_2, x_4) \\ &= f_a(x_1, x_2) \sum_{x_3} f_b(x_2, x_3) \sum_{x_4} f_c(x_2, x_4) = f_a(x_1, x_2) \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2) \\ &= f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2) \cdot 1 = f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2) \mu_{x_1 \to f_a}(x_1) \\ &= f_a(x_1, x_2) \prod_{i \in ne(f_a)} \mu_{x_i \to f_a}(x_s) \end{aligned} \tag{54}$$

Problem 8.26

We just need to update the formulas for μ functions a little bit, so that they now become functions of two variables. We start with setting x_a as a root, and keeping in mind that now we do not marginalize over x_b. Now consider $\mu_{f_s \to x}$: there are two possibilities here. The first is that $x_b \in \{x_1, ..., x_m\}$, i.e. it is one of the variables connected to f_s, the other is that x_b is not in this set:

$$\mu_{f_s \to x}(x, x_b) = \sum_{X_s \setminus x_b} f_s(x, x_1, ..., x_M) G_1(x_1, X_{s_1}) ... G_M(x_M, X_{s_M}) = \\ = \sum_{x_1 ... x_M \setminus x_b} f_s(x_1 ... x_M) \prod_m \mu_{x_m \to f_s}(x, x_b) \qquad x_b \in \{x_1 ... x_M\} \tag{55}$$

$$\mu_{f_s \to x}(x, x_b) = \sum_{X_s \setminus x_b} f_s(x, x_1, ..., x_M) G_1(x_1, X_{s_1}) ... G_M(x_M, X_{s_M}) = \\ = \sum_{x_1 ... x_M} f_s(x_1 ... x_m)) \prod_m \mu_{x_m \to f_s}(x, x_b) \qquad x_b \notin \{x_1 ... x_M\} \tag{56}$$

Let us now consider $\mu_{x_m \to f_s}$: we have not much difference in here

$$\mu_{x_m \to f_s}(x, x_b) = \prod_{l \in ne(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m, x_b) \tag{57}$$

How do we use these formulas? We now perform the computations using them for each of the values x_b can take: unfortunately though, it means the algorithm becomes less efficient, since we need to runs standard sum-product algorithm multiple times.

Problem 8.27

Let us suppose $\hat{x} = 0$ and $\hat{y} = 0$. We want to construct such a distribution, that $p(\hat{x}, \hat{y}) = 0$, with $\hat{x} = \underset{x}{\operatorname{argmax}} \sum_y p(x, y)$, $\hat{y} = \underset{y}{\operatorname{argmax}} \sum_x p(x, y)$

We have a set of inequalities we need to satisfy (remember $p(0, 0) = 0$ so we omit it):

$$p(1,0) + p(2,0) > p(0,1) + p(1,1) + p(2,1) \tag{58}$$
$$p(1,0) + p(2,0) > p(0,2) + p(1,2) + p(2,2) \tag{59}$$
$$p(0,1) + p(0,2) > p(1,0) + p(1,1) + p(1,2) \tag{60}$$
$$p(0,1) + p(0,2) > p(2,0) + p(2,1) + p(2,2) \tag{61}$$

We can start by trying to set all the probabilities on the left to be equal, and to sum up to 1, getting the following table:

x	y	p(x, y)
0	0	0
0	1	0.25
0	2	0.25
1	0	0.25
1	1	0
1	2	0
2	0	0.25
2	1	0
2	2	0

In short, we get $p(x, y) = 0.25$ if either x or y is zero, but not both, and 0 otherwise. With this definition, $p(\hat{x}) = 0.5$ and $p(\hat{y}) = 0.5$, $p(y = 1) = p(y = 2) = p(x = 1) = p(x = 2) = 0.25$.

Problem 8.28

We can simply prove using induction that at every step of the algorithm, there will exist a variable or a factor belonging to a loop, which has a pending message.

The induction base is obvious: when we initialize an algorithm, all the nodes have pending messages, since we suppose unit message was sent over all the links.

Now, suppose after step n there is a pending message from node a to node b, where a and b lie on the loop. If step $n+1$ does not send this message, we still have this pending message after the $n+1$ step. Otherwise, node b has received a new message, and so it now has pending message which it should send further along the loop, so we again have a pending message after the $n+1$ step.

Problem 8.29

This can be done using induction as well. Suppose the statement is true for all tree-structured graph containing $\leq n$ nodes. To prove the statement for graph with n nodes, let us consider any of it's leaves, and compute it's message towards the parent. This can be easily done using 8.70, 8.71. After that, all the messages are passed in a graph containing $n-1$ nodes, so there will eventually be no pending messages, except the message over the link to the selected leaf node. After we pass this message over the link, no new messages from the node will emerge (since formulas 8.70, 8.71 do not depend on messages which the leaf node has received), and so there will be no more pending messages in the graph.