1 DGM

We consider G, a DAG:

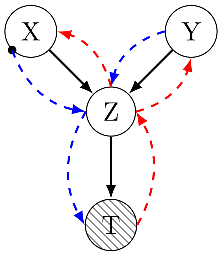


Figure 1: Graph G, where T is observed. An active path (blue dashed line) can lead from X to Y since this is an undirected path. The arrow is there to indicate the motion of the Bayes Ball. The starting point of the algorithm is represented as a black dot.

To prove that $X \perp\!\!\!\perp Y \mid T$, we use the Bayes Ball algorithm (see algorithm 1 in the appendix). We conclude that $\boxed{X \not\perp\!\!\!\perp Y \mid T}$ because there exists an undirected active path (blue dashed line in figure 1) from X to Y. This could also have been observed using the fact that the unobserved node Z with two parents has a descendant that is observed.

A distribution $p \in \mathcal{L}(G)$ satisfy the factorization

$$\mathcal{L}(G) = \left\{ p \mid p(x_V) = \prod_{i=1}^{n} p(x_i \mid x_{\pi_i}) \right\}$$

where π_i is the set of all parents of node i and V is the set of vertex in the graph. Therefore, $p \in \mathcal{L}(G)$ must satisfy

$$p(x_V) = p(x)p(y)p(z \mid x, y)p(t \mid z)$$

2 D-separation in DGM

We consider the graph G:

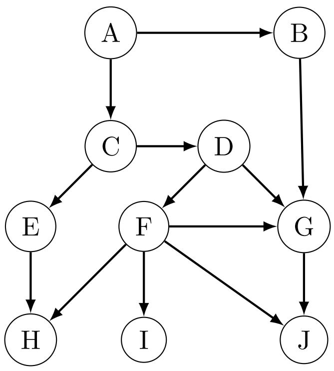


Figure 2: Complete graph G.

We are interested in the verification of several conditional independence relations. For each case, we will verify the relation using Bayes Ball algorithm (see algorithm 1). Here, we exploit the fact that unobserved node do not have the ability to bounce the ball when visited by a parent. Therefore, for each cases, we only consider a relevant sub-graph of G built by plucking away all the unobserved nodes until an observed node is found or a node of interest is found.

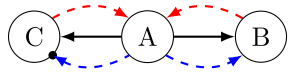
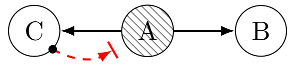
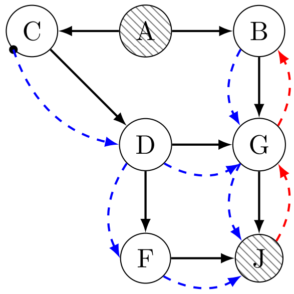
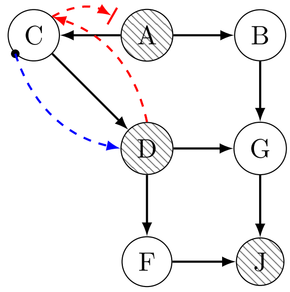
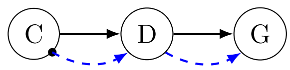
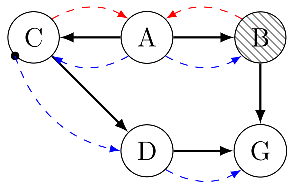
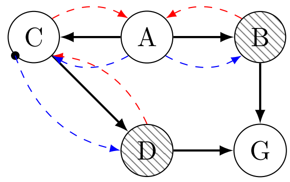
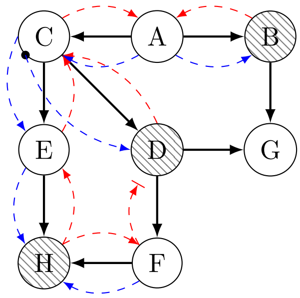
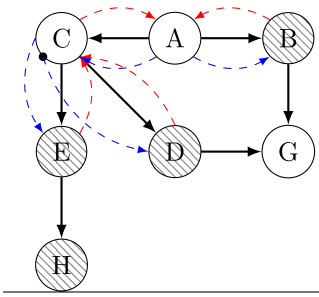
Relevant sub-graph	Conditional independence	True/False
	$C \perp\!\!\!\perp B \mid \emptyset$	False
	$C \perp\!\!\!\perp B \mid A$	True
	$C \perp\!\!\!\perp B \mid A, J$	False
	$C \perp\!\!\!\perp B \mid A, J, D$	True
	$C \perp\!\!\!\perp G \mid \emptyset$	False
	$C \perp\!\!\!\perp G \mid B$	False
	$C \perp\!\!\!\perp G \mid B, D$	True
	$C \perp\!\!\!\perp G \mid B, D, H$	True
	$C \perp\!\!\!\perp G \mid B, D, H, E$	True

Table 1: Conditional independence statements and active trail (blue dashed paths) shown in the relevant path column for questions (a) to (i).

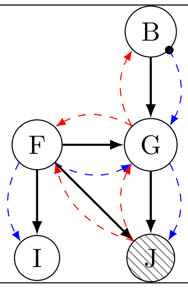
Relevant sub-graph	Conditional independence	True/False
	$B \perp\!\!\!\perp I \mid J$	False

Table 2: Conditional independence statement of question (j). In the relevant sub-graph we omitted the node D because we wanted to show the shortest active trail.

3 Positive interaction in V-structure

We let X, Y, Z be binary variables ($\in \{0, 1\}$) with a joint distribution parametrized according to the V-structure shown in figure 3.

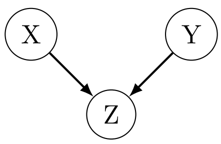


Figure 3: V-structure between the binary variables X, Y and Z.

We define the following constant

$$\begin{aligned}
\alpha &\equiv P(X = 1) \\
\beta &\equiv P(X = 1 \mid Z = 1) \\
\gamma &\equiv P(X = 1 \mid Z = 1, Y = 1)
\end{aligned}$$

In term of the marginal $P(X, Y)$ and the conditional $P(Z \mid X, Y)$, we can rewrite theses constants as follows

$$\alpha = \sum_y p(x = 1, y)$$

$$\beta = \frac{\sum_y P(X = 1, y) P(Z = 1 \mid X = 1, y)}{\sum_x \sum_y P(x, y) P(Z = 1 \mid x, y)}$$

$$\gamma = \frac{P(X = 1, Y = 1) P(Z = 1 \mid X = 1, Y = 1)}{\sum_x P(x, Y = 1) P(Z = 1 \mid x, Y = 1)}$$

a) Inequalities

The most general case for $p \in \mathcal{L}(G)$ we can consider where all variable are binary random variable is the following

$$X \sim \text{Bernoulli}(p)$$
$$Y \sim \text{Bernoulli}(q)$$

$$Z \mid X, Y \sim \begin{cases} \text{Bernoulli}(\pi_{11}), & \text{if } X = 1 \text{ and } Y = 1 \\ \text{Bernoulli}(\pi_{10}), & \text{if } X = 1 \text{ and } Y = 0 \\ \text{Bernoulli}(\pi_{01}), & \text{if } X = 0 \text{ and } Y = 1 \\ \text{Bernoulli}(\pi_{00}), & \text{if } X = 0 \text{ and } Y = 0 \end{cases}$$

To simplify the expressions for the constant, we will work with coin flip for X and Y. Therefore,

$$\alpha = pq + p(1-q) = 0.5 \quad \{\text{Since } X \perp\!\!\!\perp Y \mid \emptyset\} \tag{3.1}$$

$$\begin{aligned}\beta &= \frac{pq\pi_{11} + p(1-q)\pi_{10}}{pq\pi_{11} + p(1-q)\pi_{10} + (1-p)q\pi_{01} + (1-p)(1-q)\pi_{00}} \\ &= \frac{\pi_{11} + \pi_{10}}{\pi_{11} + \pi_{10} + \pi_{01} + \pi_{00}}\end{aligned} \tag{3.2}$$

$$\gamma = \frac{pq\pi_{11}}{pq\pi_{11} + (1-p)q\pi_{01}} = \frac{\pi_{11}}{\pi_{11} + \pi_{01}} \tag{3.3}$$

I $\gamma < \alpha$

Here, we notice that π_{11} and π_{01} have a big impact on the relative importance of γ when compared. We can see that

$$\pi_{11} \gg \pi_{01} \leq 1 \implies \gamma \to 1$$
$$\pi_{01} \gg \pi_{11} \leq 1 \implies \gamma \to 0$$

Setting every other parameter to 0.5, $\pi_{11} = 0.01$ and $\pi_{01} = 0.9$, we get the following example for which $\gamma < \alpha$

X	Y	P(X, Y)
1	1	0.25
1	0	0.25
0	1	0.25
0	0	0.25

Table 3: Joint distribution of 2 coin flip.

X	Y	$P(Z=1 \mid X, Y)$
1	1	0.01
1	0	0.5
0	1	0.9
0	0	0.5

Table 4: Conditional on $Z = 1$ with $\gamma \ll 1$.

α	β	γ
0.5	0.28	0.01

Table 5: Results

II $\alpha < \gamma < \beta$

$\gamma > \alpha$ can be obtained by setting $\pi_{11} \gg \pi_{01}$. To get the RHS of the inequality, we must set $\pi_{00} \ll 1$. The following table works as intended:

X	Y	P(X, Y)
1	1	0.25
1	0	0.25
0	1	0.25
0	0	0.25

Table 6: Joint distribution of 2 coin flip.

X	Y	$P(Z = 1 \mid X, Y)$
1	1	0.9
1	0	0.5
0	1	0.5
0	0	0.001

Table 7: Conditional on $Z = 1$ with $\gamma \ll 1$.

α	β	γ
0.5	0.74	0.64

Table 8: Results

III $\beta < \alpha < \gamma$

Here, starting from equilibrium (all parameter equal), we can shift β to be smaller than the other by increasing π_{00}. Then we can increase γ slightly by increasing slightly π_{11}. Too harsh an increase will upset the lower inequality.

X	Y	P(X, Y)
1	1	0.25
1	0	0.25
0	1	0.25
0	0	0.25

Table 9: Joint distribution of 2 coin flip.

X	Y	$P(Z = 1 \mid X, Y)$
1	1	0.6
1	0	0.5
0	1	0.5
0	0	0.9

Table 10: Conditional on $Z = 1$ with $\gamma \ll 1$.

α	β	γ
0.5	0.44	0.54

Table 11: Results

b)

I $\gamma < \alpha$

In order for this inequality to hold, then the effect $Z = 1$ must contain information about a correlation between the causes $Y = 1$ and the possible values of X. This information is encoded through the parameters π_{11} and π_{01}. By lowering π_{01}, we naturally lower the possibility that $X = 0$ given $Z = Y = 1$. Conversely for π_{11}. Therefore, the condition $\pi_{01} \ll \pi_{11} \leq 1$, all other parameters being equal, means that the observation of the effect $Z = 1$ and the cause $Y = 1$ update strongly our belief about the probability of observing $X = 1$. Initially, we had no particular opinion $\alpha = 0.5$. After observation, we are now almost convinced that the second cause was $X = 1$.

II $\alpha < \gamma < \beta$

Here, we reversed the previous inequality first so that the observation of $Y = 1$ and the observation of the effect $Z = 1$ will increase the probability of observing $X = 1$. The RHS can function independently of the first one since π_{00} do not affect γ. Indeed, $Y = 1$ is observed for γ therefore the probability of observing $Y = 0$ do not impact γ.

For the β probability, we are working with less information about the cause of $Z = 1$ (we do not know Y). Therefore, to have higher confidence of observing $X = 1$ without knowing the Y cause means we must have a strong belief that observing the effect $Z = 1$ means the cause $X = 0$ is unlikely. This is encoded in π_{00} and π_{01} (but we do not touch this one in order to keep γ intact). This is why setting π_{00} to a low value yielded the inequality we seeked.

III $\beta < \alpha < \gamma$

Here, we want to lower our confidence that observing $Z = 1$ was caused by $X = 1$. To do that, we bring π_{00} to a very probable value. This update our belief that an effect $Z = 1$ is most likely caused by $X = 0$ when $Y = 0$.

We do not increase π_{01} because we do not want to lower our confidence that observing $Z = 1$ and $Y = 1$ is related to an event with $X = 1$ (γ). This is why we increase only slightly π_{11} in order to increase γ without upsetting too much β.

In other words, observing the effect $Y = 1$ will increase slightly our confidence that the second cause for $Z = 1$ was $X = 1$ as opposed to no belief at all (α), updating our initial belief that the most probable cause was $X = 0$.

4 Equivalence of DGM with UGM

We consider a directed tree graphical model $G = (V, E)$. By definition, each nodes has at most one parent

$$|\pi_i| \leq 1 \; \forall \; i \in V$$

A distribution p is part of the family of distribution $\mathcal{L}(G)$ associated with the graph G if it has legal factors $\{f_i\}$ associated with the conditional independences of G

$$\mathcal{L}(G) = \left\{ \text{p is a distribution over } X_V \;\middle|\; p(X_V) = \prod_i f_i(x_i \mid x_{\pi_i}) \right\}$$

The moralized graph $\bar{G}$, which is the UGM associated with G will have cliques of size at most 2 since no edges will be added to the tree in the moralization procedure.

$$\max |C| \leq |\pi_i| + 1 \; \forall i \in V \text{ and } \forall \; C \; \in \mathscr{C}_{\max}$$

where $\mathscr{C}_{\max}$ is the set of maximal cliques in $\bar{G}$. This is to say that all cliques will have two nodes. Therefore, the probability distribution family associated with $\mathcal{L}(\bar{G})$ with legal factors (potentials) is defined over the exact same set of nodes as previously, making $\mathcal{L}(G) = \mathcal{L}(\bar{G})$.

$$\mathcal{L}(\bar{G}) = \left\{ \text{p is a distribution over } X_V \;\middle|\; p(X_V) = \frac{1}{Z} \prod_{C \in \mathscr{C}_{\max}} \psi_C(x_C) \right\}$$

With this idea in mind, we prove $\mathcal{L}(G) = \mathcal{L}(\bar{G})$ by induction.

Proof. Let $G_n = (V, E)$ be a directed tree of size $n = |V|$, the number of nodes in that tree. We first set $n = 1$. In this case,

$$p_{G_1}(X_V) = f_1(x_1)$$

and

$$p_{\bar{G}_1}(X_V) \propto \psi_1(x_i)$$

The two distributions have the same factorization and belongs to the same family. Therefore,

$$\mathcal{L}(G_1) = \mathcal{L}(\bar{G}_1)$$

For $n = 2$,

$$p_{G_2}(X_V) = f_1(x_1) f_2(x_2 \mid x_1)$$

and

$$p_{\bar{G}_2}(X_V) \propto \psi_1(x_1, x_2)$$

since there is only one clique of the maximal size in that set. We make the following map between the two parametrization:

$$\psi_1(x_1, x_2) \mapsto f_1(x_1) f_2(x_2 \mid x_1)$$

Since this is a bijective map between the two sets of factors, the set of distribution that can be represented by G_2 is equal to the set of distribution represented by $\bar{G}_2$

$$\mathcal{L}(G_2) = \mathcal{L}(\bar{G}_2)$$

Now, we assume that $\mathcal{L}(G_n) = \mathcal{L}(\bar{G}_n)$ is true. For $n + 1$, we have

$$p_{G_{n+1}}(X_V) = \prod_{i=1}^{n+1} f_i(x_i \mid p_{\pi_i})$$

Since $\pi_{n+1} = n$,

$$p_{G_{n+1}}(X_V) = f_{n+1}(x_{n+1} \mid x_n) \prod_{i=1}^{n} f_i(x_i \mid x_{\pi_i}) = f_{n+1}(x_{n+1} \mid x_n) p_{G_n}(X_V)$$

On the other hand

$$p_{\bar{G}_{n+1}}(X_V) = \frac{1}{Z} \prod_{C \in \mathscr{C}_{\max}} \psi_C(x_C)$$

Since all cliques are of length 2, we can factor out the clique

$$C_{n+1} = \{x_{n+1}, x_n\}$$

$$p_{\bar{G}_{n+1}}(X_V) = \frac{1}{Z} \psi_{C_{n+1}}(x_{n+1}, x_n) \prod_{C \in \mathscr{C}_{\max} \setminus C_{n+1}} \psi_C(x_C)$$

Ignoring the normalization constant, this is

$$p_{\bar{G}_{n+1}}(X_V) \propto \psi_{C_{n+1}}(x_{n+1}, x_n) p_{\bar{G}_n}(X_V)$$

Since we know there exist a bijective map of the factors for the graph of n nodes, then we have found that G_{n+1} also have a bijective map between factors:

$$\psi_{C_{n+1}}(x_{n+1}, x_n) \mapsto f_{n+1}(x_{n+1} \mid x_n)$$

Therefore

$$\mathcal{L}(G_{n+1}) = \mathcal{L}(\bar{G}_{n+1})$$

□

5 Hammersley-Clifford Counter-Example

Hammersley-Clifford's Theorem. *Suppose $\mathcal{UI}$ is the set of positive distributions that satisfy the global Markov property of the Markov network $G = (V, E)$.*

$$p(X_V) \models (X \perp\!\!\!\perp Y \mid Z), \;\; s.t. \;\; sep_G(X; Y \mid Z)$$

for the disjoint sets X, Y and Z in V. Then, suppose that $\mathcal{UF}$ is the set of Gibbs distributions that can be expressed as a factorization of the form

$$p(X_V) = \frac{1}{Z} \prod_{C \in \mathscr{C}_{\max}} \psi_C(x_C)$$

Then

$$\mathcal{UI} = \mathcal{UF}$$

We can show that a probability distribution that is not strictly positive will not factorize according to the set of distributions $\mathcal{UF}$, even if it respect the global Markov property of the graph G.

We consider the following graph

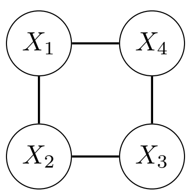


Figure 4: A Markov Network of binary variables.

And consider the following probability distribution

$$P(X_V) = \begin{cases} \frac{1}{8}, & \text{if } X_V \in \Xi \\ 0, & \text{otherwise} \end{cases}$$

Where we have created the set

$$\Xi = \begin{Bmatrix} (0,\ 0,\ 0,\ 0),\ (1,\ 0,\ 0,\ 0),\ (1,\ 1,\ 0,\ 0), (1,\ 1,\ 1,\ 0), \\ (0,\ 0,\ 0,\ 1),\ (0,\ 0,\ 1,\ 1),\ (0,\ 1,\ 1,\ 1),\ (1,\ 1,\ 1,\ 1) \end{Bmatrix}$$

We do not show that P satisfy the global Markov property, and take it for granted since it can be proved easily by showing that conditioning on any set Z that separates X and Y will make X and Y deterministically determined on Z, making X and Y independent.

Here, we prove that $P(X_V) \notin \mathcal{UF}$:

Proof. Suppose that $P(X_V) \in \mathcal{UF}$. Then, it can be factorized over the set of maximal cliques of G:

$$P(X_V) = \frac{1}{Z}\phi_1(X_1, X_2)\phi_2(X_1, X_4)\phi_3(X_2, X_3)\phi_4(X_3, X_4)$$

We consider the case $p(0, 1, 1, 0) = 0$. To have a null probability, we could have $Z = \infty$, but this would not be consistent with the distribution $P(X_V)$ since Z is common to all X_V.

Another solution is to set one of the factor ϕ_i to 0. Suppose we set

$$\phi_1(0, 1) = 0$$

Then,

$$p(0, 1, 1, 1) = \frac{1}{Z}\phi_1(0, 1)\phi_2(0, 1)\phi_3(1, 1)\phi_4(1, 1) = 0$$

But, the distribution we started with assumes $P(0, 1, 1, 1) = \frac{1}{8}$ so we must have $\phi_1(0, 1) \neq 0$. Going over all other possibilities, we find that no factor can be set to 0.

$$\begin{aligned} p(0, 0, 0, 0) &= \frac{1}{8} \implies \phi_2(0, 0) \neq 0 \\ p(0, 1, 1, 1) &= \frac{1}{8} \implies \phi_3(1, 1) \neq 0 \\ p(1, 1, 1, 0) &= \frac{1}{8} \implies \phi_4(1, 0) \neq 0 \end{aligned}$$

We find that there is no way to have $p(0, 1, 1, 0) = 0$ given the factorisation in $\mathcal{UF}$. This is a contradiction, and we conclude that $P(X_V) \notin \mathcal{UF}$.

□

6 Bizarre Conditional Independence Properties

7 EM and Gaussian Mixture

I Primer

The Gaussian Mixture Model is a graphical model with a random categorical latent variable $\mathbf{z}_i$ and deterministic parameters

$$\theta = (\pi_{1,\ldots,K}, (\mu_i)_{i=1}^K, (\Sigma_i)_{i=1}^K)^T.$$

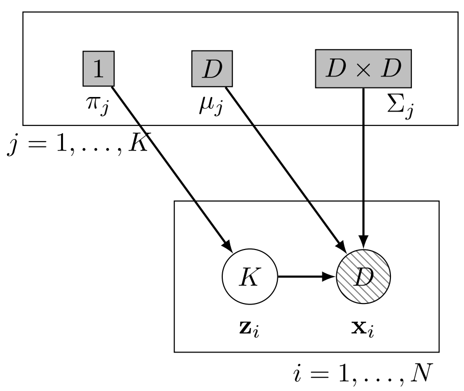


Figure 5: Gaussian Mixture Model with plate notation. The label inside the circles are references to the vector dimension. $\mathbf{z}_i$ is a K-vector, etc.

There is a latent variable $\mathbf{z}_i$ for each observations ($i = 1, \dots, N$) and latent variables μ_j and Σ_j for each cluster $j = 1, \dots, K$. The z_i are distributed with a K-Categorical distribution. The Multinoulli is a natural choice.

$$\mathbf{z}_i \sim \text{Multinoulli}(\pi_{1,\dots,K})$$

We then assume the conditional over an observation $\mathbf{x}_i$ assigned to the cluster j ($z_{i,j} = 1$) to follow a multivariate normal distribution

$$\mathbf{x}_i \mid z_{i,j} = 1 \sim \mathcal{N}(\mu_j, \Sigma_j)$$

where $\mathbf{x}$ and μ_j are D-vectors, and the covariance matrix is a positive semi-definite, symmetric $D \times D$ matrix. We define the complete log-likelihood (a lower bound on the log-likelihood with missing information):

$$\mathcal{L}_C = \log p(\mathbf{x}, \mathbf{z} \mid \theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_{i,j} \log \mathcal{N}(\mathbf{x}_i \mid \mu_j, \Sigma_j) + z_{i,j} \log \pi_j + (1 - z_{i,j}) \log(1 - \pi_j))$$

We take the expectation of that likelihood

$$\mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z} \mid \theta)\right] = \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{E}_q[z_{i,j}](\log \mathcal{N}(\mathbf{x}_i \mid \mu_j, \Sigma_j) + \log \pi_j) + (1 - \mathbb{E}_q[z_{i,j}]) \log(1 - \pi_j)$$

The expectation is taken with respect to the marginal distribution of the cluster assignement variable z_i:

$$q^{(t+1)}(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x}_i, \theta^{(t)})$$

For a single observation, this is

$$q^{(t+1)}(\mathbf{z}_i) = p(\mathbf{z}_i \mid \mathbf{x}_i, \theta^{(t)})$$

Using Bayes theorem,

$$q^{(t+1)}(\mathbf{z}_i) = \frac{p(\mathbf{x}_i \mid \mathbf{z}_i, \theta^{(t)}) p(\mathbf{z}_i \mid \pi^{(t)})}{p(\mathbf{x}_i \mid \theta^{(t)})}$$

We define the weight $\Upsilon_{i,j}$ to be the probability $\Upsilon_{i,j}^{(t+1)} = q^{(t+1)}(z_{i,j} = 1)$. Using the expressions for the conditional and the prior on the latent variable $z_{i,j} = 1$, we can write

$$\mathbb{E}_{q^{(t+1)}}(z_{i,j}) = \Upsilon_{i,j}^{(t+1)} = \frac{\pi_j^{(t)} \mathcal{N}(\mathbf{x}_i \mid \mu_j^{(t)}, \Sigma_j^{(t)})}{\displaystyle\sum_{\ell=1}^{K} \pi_\ell^{(t)} \mathcal{N}(\mathbf{x}_i \mid \mu_\ell^{(t)}, \Sigma_\ell^{(t)})}$$

At the maximization step, we optimize for the deterministic parameters $\theta^{(t+1)}$

$$\theta^{(t+1)} \triangleq \underset{\theta}{\text{argmax}}\ \mathbb{E}_{q^{(t+1)}}[\log p(\mathbf{x}, \mathbf{z} \mid \theta)]$$

Writing out all the terms except the constant term,

$$\theta^{(t+1)} = \underset{\theta}{\text{argmax}} \sum_{i=1}^{N} \sum_{j=1}^{K} \Upsilon_{i,j}^{(t+1)} \left(\frac{1}{2} \log \det \Sigma^{-1} - \frac{1}{2} (\mathbf{x}_i - \mu_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mu_j) + \log \pi_j \right) + \left(1 - \Upsilon_{i,j}^{(t+1)}\right) \log(1 - \pi_j)$$

II M-step for π_j

$$\partial_{\pi_\ell} \mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{K} \Upsilon_{i,j}^{(t+1)} \frac{\delta_{j\ell}}{\pi_j} - (1 - \Upsilon_{i,j}^{(t+1)}) \frac{\delta_{j\ell}}{1 - \pi_j}$$

where $\delta_{j\ell}$ is the Kronecker delta coming from the partial derivate. The MLE solution is found where this derivative is 0:

$$\frac{1}{\hat{\pi}_\ell} \sum_{i=1}^{N} \Upsilon_{i,j}^{(t+1)} = \frac{1}{1 - \hat{\pi}_\ell} (N - \sum_{i=1}^{N} \Upsilon_{i,j}^{(t+1)})$$

By noticing the symmetry between the parameter π_ℓ^* and the sum over the soft labels, we conclude that

$$\boxed{\hat{\pi}_\ell = \frac{1}{N} \sum_{i=1}^{N} \Upsilon_{i,j}^{(t+1)}}$$

III M-step for μ_j

$$\partial_{\mu_\ell}\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{K}\Upsilon_{i,j}^{(t+1)}\Sigma_j^{-1}(\mathbf{x}_i - \mu_j)\delta_{j,\ell}$$

At the stationary point, we can simplify the covariance matrix and we get

$$0 = \sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}\mathbf{x}_i - \hat{\mu}_\ell\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}$$

Therefore,

$$\boxed{\hat{\mu}_\ell = \frac{1}{\sum_i \Upsilon_{i,\ell}^{(t+1)}}\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}\mathbf{x}_i}$$

IV M-step for Σ_j

We will derive the full form of the update, and give the special result for a diagonal covariance matrix at the end. To simplify the derivative, we derive with respect to $\Lambda \equiv \Sigma^{-1}$:

$$\partial_{\Lambda_\ell}\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{K}\Upsilon_{i,j}^{(t+1)}\left(\frac{1}{2}\Sigma_j - \frac{1}{2}(\mathbf{x}_i - \hat{\mu}_j)(\mathbf{x}_i - \hat{\mu}_j)^T\right)\delta_{j\ell}$$

At the stationary point, this gives

$$\hat{\Sigma}_\ell = \frac{1}{\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}}\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}(\mathbf{x}_i - \hat{\mu}_\ell)(\mathbf{x}_i - \hat{\mu}_\ell)^T$$

Assuming a spherical covariance matrix $\Sigma_j = \sigma_j^2\mathbf{1}$, then the terms involving the covariance become scalars. The gradient will also be a scalar:

$$\partial_{\sigma_\ell^{-2}}\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{K}\Upsilon_{i,j}^{(t+1)}\left(\frac{1}{2}\sigma_j^2 - \frac{1}{2}(\mathbf{x}_i - \hat{\mu}_j)^T(\mathbf{x}_i - \hat{\mu}_j)\right)\delta_{j\ell}$$

At the stationary point, we get a similar expression to the previous one, but this time scalar:

$$\boxed{\hat{\sigma}_\ell^2 = \frac{1}{\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}}\sum_{i=1}^{N}\Upsilon_{i,\ell}^{(t+1)}(\mathbf{x}_i - \hat{\mu}_\ell)^T(\mathbf{x}_i - \hat{\mu}_\ell)}$$

8 Appendix

a) Bayes Ball algorithm

Here we focus our attention to the Bayes Ball algorithm for probabilistic node only. In that case, conditional independence $X_J \perp\!\!\!\perp X_L \mid X_K$, with $J, K, L \subseteq V$ requires the notion of d-separation:

Active path An active path from J to L given K is an undirected path between $\ell \in L$ and $j \in J$ such that every node i with two parents in this chain is observed ($i \in K$) or has a descendant in K

D-separation X_j is said to be conditionally independent to X_L (or *d-separate* X_L) given X_K if there is no active path from J to L given K.

With this description, we can devise a simple algorithm that will find all active path in the graph in linear time [1]. We will use the following convention

- $\dashrightarrow$ Dashed blue arrows indicate the ball drop from parent to child;
- $\dashrightarrow$ Dashed red arrow indicate the ball bounce back to the parent;

- Barred arrows mean the node will block the ball (neither bounce it back nor let it pass through). These can be used as an indicator and are not necessary for the algorithm.

With this convention, we describe an algorithm that will find all active path from a node $\ell \in L$ to a node $j \in J$ given K if they exists:

Algorithm 1: Bayes Ball

Result: active paths in the graph $\dashrightarrow$
initialize a schedule with all $j \in J$ as though they were visited by a child;
while *schedule not empty* **do**
  pick a node j and remove it from the schedule;
  if $j \notin K$ ***and*** *j is visited from a child* ***and*** *there is no* $\dashrightarrow$ *linking j to previous node* **then**
    draw $\dashrightarrow$ going into j;
    schedule its parents to be visited;
    schedule its children to be visited;
  else if *j is visited from a parent* ***and*** *there is no* $\dashrightarrow$ *linking j to previous node* **then**
    draw $\dashrightarrow$ going into j;
    if $j \in K$ **then**
      schedule its parents to be visited;
    else
      schedule its children to be visited;

References

[1] Shachter, R. D. (2013). Bayes-Ball: The Rational Pastime (for Determining Irrelevance and Requisite Information in Belief Networks and Influence Diagrams). http://arxiv.org/abs/1301.7412