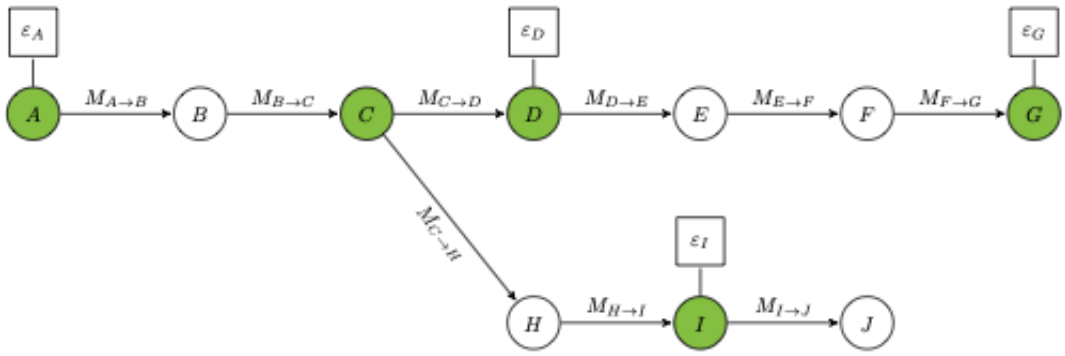
Complete derivation of belief update rules for trees

I. Formal definitions

Evidence

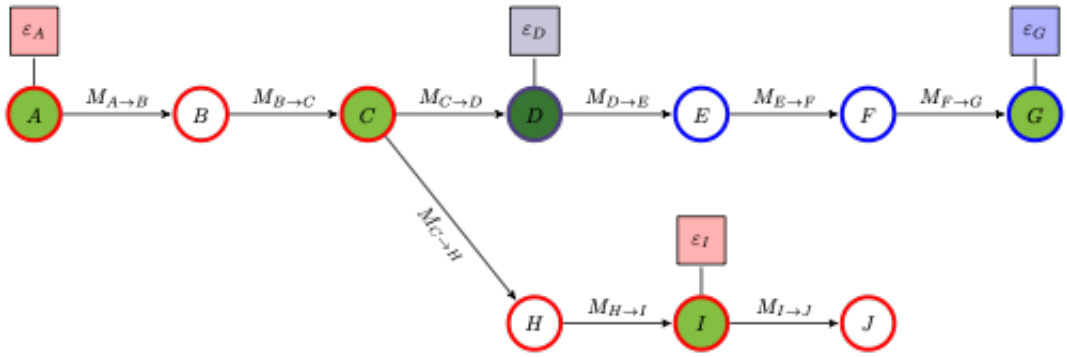


- **Direct evidence** ε_V for a node V is an observation $V = v_*$ that determines the **local likelihood** $\lambda^*_V(v) = [v = v_*]$.
- **Indirect evidence** ε_V for a node V is a partial observation that determines the **local likelihood** $\lambda^*_V(v) = \Pr[\varepsilon_V | V = v]$.

As indirect evidence ε_V can be modelled by adding a new successor V_* to the node V with the conditional distribution $\Pr[V_* = \varepsilon_V | V = v]$ determined by the local likelihood $\lambda^*_V(v)$, we do not consider indirect evidence in the further analysis. For the chains, we needed to analyse the effect of the indirect evidence separately as this extra node converts a chain into a tree.

Evidence partitioning

- **Evidence** evidence is the summary evidence of all nodes in the tree.
- **Upstream evidence** $\text{evidence}^+(V)$ is the evidence of all nodes reachable through a predecessor of V together with the evidence for V.
- **Downstream evidence** $\text{evidence}^-(V)$ is the evidence of all nodes succeeding V together with the evidence for V.

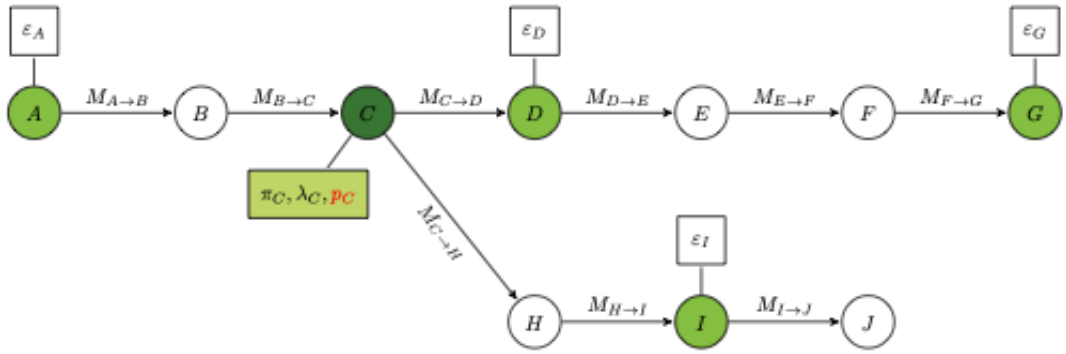


In this figure, upstream and downstream evidence for the node D are the following:

$$\text{evidence}^+(D) = \{\varepsilon_A, \varepsilon_D, \varepsilon_I\}$$
$$\text{evidence}^-(D) = \{\varepsilon_D, \varepsilon_G\} \quad .$$

II. Derivation of iterative update rules

Marginal posterior probabilities



Mechanical application of Bayes rule yields

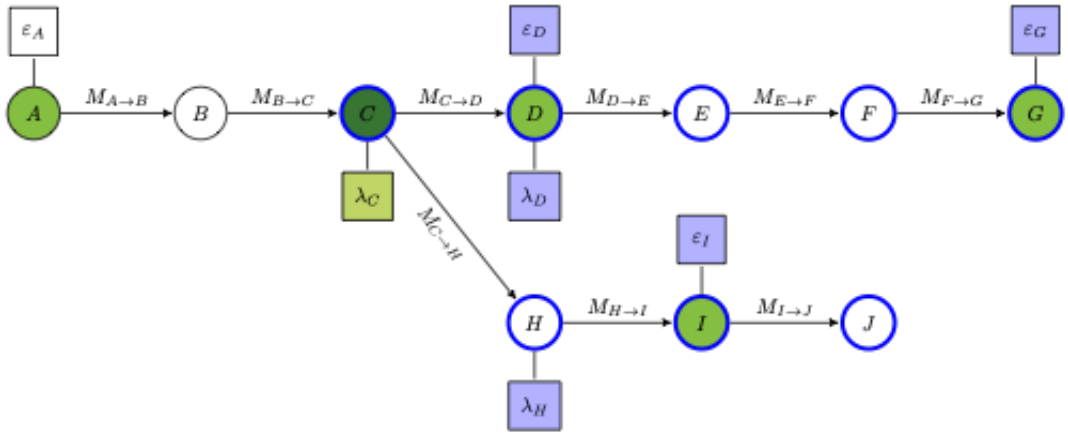
$$\begin{aligned} p_V(v) &= \Pr[V = v | \text{evidence}^+(V), \text{evidence}^-(V)] \\ &= \frac{\Pr[\text{evidence}^-(V) | V = v, \text{evidence}^+(V)] \cdot \Pr[V = v | \text{evidence}^+(V)]}{\Pr[\text{evidence}^-(V) | \text{evidence}^+(V)]} \\ &\propto \Pr[\text{evidence}^-(V) | V = v, \text{evidence}^+(V)] \cdot \Pr[V = v | \text{evidence}^+(V)] \quad . \end{aligned}$$

As direct knowledge of the state $V = v$ completely determines what happens with the next node, the knowledge $V = v, \text{evidence}^+(V)$ is equivalent to the knowledge $V = v$ and we can simplify:

$$\begin{aligned} p_V(v) &\propto \Pr[\text{evidence}^-(V) | V = v] \cdot \Pr[V = v | \text{evidence}^+(V)] \\ &\propto \lambda_V(v) \cdot \pi_V(v) \quad . \end{aligned}$$

As a result, if we know the likelihood $\lambda_V(\cdot)$ and posterior $\pi_V(\cdot)$ up to a constant then we can recover the marginal posterior $p_V(\cdot)$ through normalisation. Up to a constant in this context means that we can omit all factors that do not depend on the value v.

Likelihood update for a node without evidence



Let $W_1, \ldots, W_k$ be direct successor nodes of V, then the downstream evidence decomposes into k classes

$$\text{evidence}^-(V) = \text{evidence}^-(W_1) \cup \ldots \cup \text{evidence}^-(W_k)$$

as the node V has no evidence. Moreover, these events are independent for a fixed $V = v$ value as they occur in different tree brances. Consequently,

$$\begin{aligned}\lambda_V(v) &= \Pr[\text{evidence}^-(V)|V = v] \\ &= \Pr[\text{evidence}^-(W_1) \wedge \ldots \wedge \text{evidence}^-(W_k)|V = v] \\ &= \Pr[\text{evidence}^-(W_1)|V = v] \cdots \Pr[\text{evidence}^-(W_k)|V = v] \quad .\end{aligned}$$

Mechanical application of marginalisation rule to one of the terms yields

$$\begin{aligned}\lambda_j(v) &= \Pr[\text{evidence}^-(W_j)|V = v] \\ &= \sum_{w_j \in W_j} \Pr[\text{evidence}^-(W_j) \wedge W_j = w_j|V = v] \\ &= \sum_{w_j \in W_j} \Pr[\text{evidence}^-(W_j)|W_j = w_j, V = v] \cdot \Pr[W_j = w_j|V = v] \quad .\end{aligned}$$

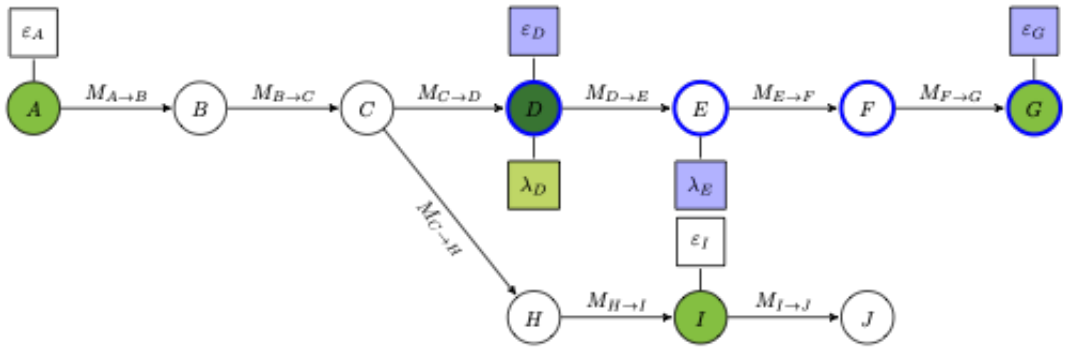
The Markov property assures that knowledge of $V = v$ is redundant when we know $W_j = w_j$. Consequently, we get

$$\begin{aligned}\lambda_j(v) &= \sum_{w_j \in W_j} \Pr[\text{evidence}^-(W_j)|W_j = w_j] \cdot \Pr[W_j = w_j|V = v] \\ &= \sum_{w_j \in W_j} \lambda_{W_j}(w_j) M_{V\to W_j}[v, w_j] \quad .\end{aligned}$$

Representing $\lambda_j(\cdot)$ and $\lambda_{W_j}(\cdot)$ as column vectors allows us to compact the equation in matrix notation:

$$\begin{aligned}\lambda_j &= \boldsymbol{M}_{V\to W_j} \lambda_{W_j} \\ \lambda_V &= \lambda_1 \otimes \ldots \otimes \lambda_k \quad .\end{aligned}$$

Likelihood update for a node with direct evidence



Let $V = v_*$ be a direct evidence associated with the node V and let

$$\text{evidence}^-_*(V) = \text{evidence}^-(V) \setminus \{V = v_*\}$$

be the evidence associated with the node V downstream of V. Then evidence decomposition yieds

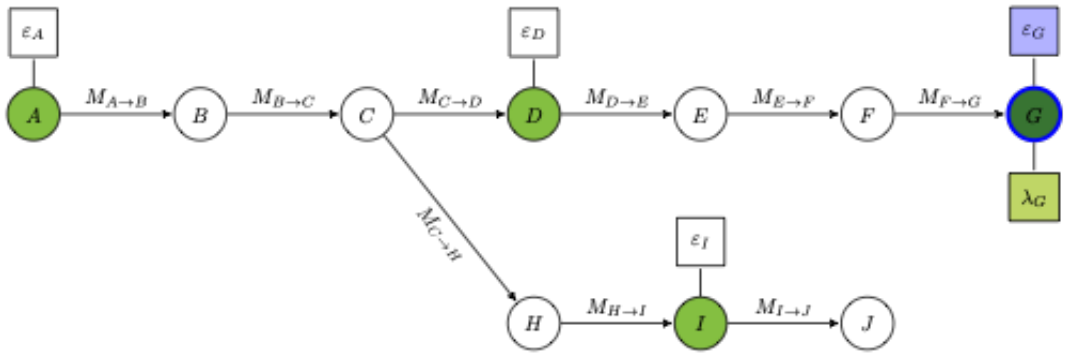
$$\begin{aligned}
\lambda_V(v) &= \Pr[\text{evidence}^-(V)|V = v] \\
&= \Pr[\text{evidence}^-_*(V) \wedge V = v_*|V = v] \\
&= \Pr[\text{evidence}^-_*(V)|V = v_*, V = v] \cdot \Pr[V = v_*|V = v] \\
&= \Pr[\text{evidence}^-_*(V)|V = v_*] \cdot [v_* = v] \quad .
\end{aligned}$$

Note that $\lambda_V(v)$ is nonzero only for a single value v_*. Thus by multiplying $\lambda_V(v)$ with a constant value $\lambda_V(v_*)^{-1}$, we get an indicator:

$$\lambda_V(v) \propto [v = v_*] \quad .$$

Note that $\lambda_V(v_*)^{-1}$ depends on v_* but remains constant if we consider different values of $v \in V$.

Likelihood update for a node without successors



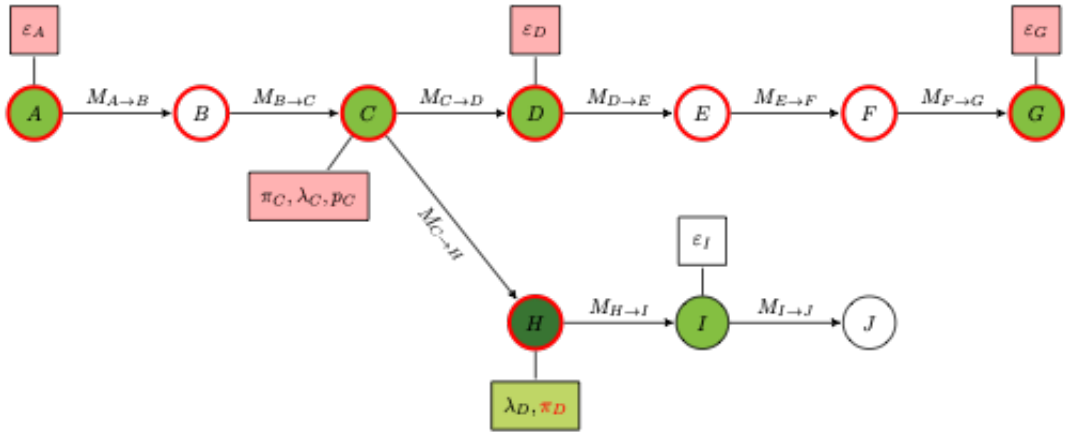
The rules for updating the likelihood are applicable for nodes that do have successors. Hence, we must address nodes without successors explicitly. If the node has direct evidence $V = v_*$ then it is the entire downstream evidence evidence$^-(V)$ and thus

$$\begin{aligned}\lambda_V(v) &= \Pr[\text{evidence}^-(V)|V = v] \\ &= \Pr[V = v_*|V = v] \\ &= [v = v_*] \quad .\end{aligned}$$

If the node does not have evidence then the entire downstream evidence evidence$^-(V)$ is empty and thus

$$\begin{aligned}\lambda_V(v) &= \Pr[\text{evidence}^-(V)|V = v] \\ &= \Pr[\text{True}|V = v] \\ &= 1 \quad .\end{aligned}$$

Prior update if a node and its predecessor are without evidence



Let U be a predecessor node of V then mechanical application of marginalisation rule yields

$$\begin{aligned}\pi_V(v) &= \Pr[V = v|\text{evidence}^+(V)] \\ &= \sum_{u\in U} \Pr[V = v \wedge U = u|\text{evidence}^+(V)] \\ &= \sum_{u\in U} \Pr[V = v|U = u, \text{evidence}^+(V)] \cdot \Pr[U = u|\text{evidence}^+(V)] \quad .\end{aligned}$$

As the node V has no evidence, the upstream evidence must be reachable through the predecessor node U. However, this evidence is not only $\text{evidence}^+(U)$ if U has more child nodes than just V. Let $W_1, \dots, W_k$ denote the children of U so that $W_k = V$. Then the upstream evidence of V decomposes into up- and downstream evidence:

$$\text{evidence}^+(V) = \text{evidence}^+(U) \cup \text{evidence}^-(W_1) \cup \dots \cup \text{evidence}^-(W_{k-1}) \quad .$$

The Markov property assures that knowledge of $\text{evidence}^+(U)$ is redundant when we know $U = u$. Consequently, we get

$$\begin{aligned}\pi_V(v) &= \sum_{u\in U} \Pr[V = v|U = u, \text{evidence}^+(V)] \cdot \Pr[U = u|\text{evidence}^+(V)] \\ &= \sum_{u\in U} \Pr[V = v|U = u] \cdot \Pr[U = u|\text{evidence}^+(U), \text{evidence}^-(W_1), \dots, \text{evide} \\ &= \sum_{u\in U} M_{U\to V}[u, v] \cdot \frac{\Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_{k-1})|U = u, \text{evider}}{\Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_k} \\ &\propto \sum_{u\in U} M_{U\to V}[u, v] \cdot \Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_{k-1})|U = u, \text{eviden}\end{aligned}$$

where the last line follows from the fact that the denominator is a constant that does not depend on the values of $u \in U$ and $v \in V$.

The Markov property assures that knowledge of $\text{evidence}^+(U)$ is redundant when we know $U = u$ and thus we can express

$$\begin{aligned}\pi_V(v) &\propto \sum_{u\in U} M_{U\to V}[u, v] \cdot \Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_{k-1})|U = u] \cdot \Pr[U \\ &\propto \sum_{u\in U} M_{U\to V}[u, v] \cdot \Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_{k-1})|U = u] \cdot \pi_U(u \\ &\propto \sum_{u\in U} M_{U\to V}[u, v] \cdot \pi_U(u) \cdot \Pr[\text{evidence}^-(W_1), \dots, \text{evidence}^-(W_{k-1})|U = u\end{aligned}$$

Let us multiply and divide the summation term by the factor $\Pr[\text{evidence}^-(W_k)|U = u]$ to simplify the derivation. Then

$$\pi_V(v) \propto \sum_{u \in U} \frac{M_{U \to V}[u, v]}{\Pr[\text{evidence}^-(W_k)|U = u]} \cdot \pi_U(u) \cdot \Pr[\text{evidence}^-(W_1), \ldots, \text{evidence}^-$$

$$\propto \sum_{u \in U} \frac{M_{U \to V}[u, v]}{\Pr[\text{evidence}^-(W_k)|U = u]} \cdot \pi_U(u) \cdot \Pr[\text{evidence}^-(W_1), \ldots, \text{evidence}^-$$

where the last equation follows from the fact that branches starting from $W_1, \ldots, W_k$ are independent given the value $U = u$.

As the node $U = u$ does not have a direct evidence $U = u_*$ linked to it, the last factor is the likelihood of U by definition and we get

$$\pi_V(v) \propto \sum_{u \in U} M_{U \to V}[u, v] \cdot \pi_U(u) \cdot \frac{\lambda_U(u)}{\Pr[\text{evidence}^-(V)|U = u]} \quad .$$

Recall that the likelihood $\lambda_U(u)$ splits into the product $\lambda_1(u) \ldots, \lambda_k(u)$ by the likelihood update rule and thus

$$\pi_V(v) \propto \sum_{u \in U} M_{U \to V}[u, v] \cdot \frac{\pi_U(u)\lambda_U(u)}{\lambda_k(u)}$$

$$\propto \sum_{u \in U} M_{U \to V}[u, v] \cdot \frac{p_U(u)}{\lambda_k(u)}$$

where

$$\lambda_k(u) = \sum_{v \in V} \lambda_V(v) M_{U \to V}[u, v] \quad .$$

Representing $\pi_V(\cdot)$ and $p_U(\cdot)$ as row vectors allows us to compact the equation in matrix notation:

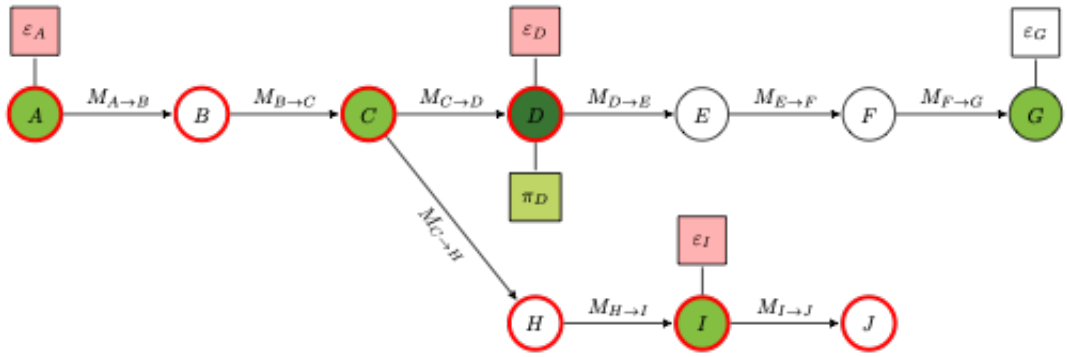
$$\pi_V \propto \frac{p_U}{\lambda_k} M_{U \to V}$$

where the division line represents element-wise division of vectors. If the predecessor has only one child node then the expression simplifies to

$$\pi_V \propto \pi_U M_{U \to V}$$

as expected (this is the prior update formula for chains).

Prior update for a node with direct evidence



Let U be the predecessor node of V and let $V = v_*$ be the direct evidence associated with the node V. Then the evidence decoposition yields

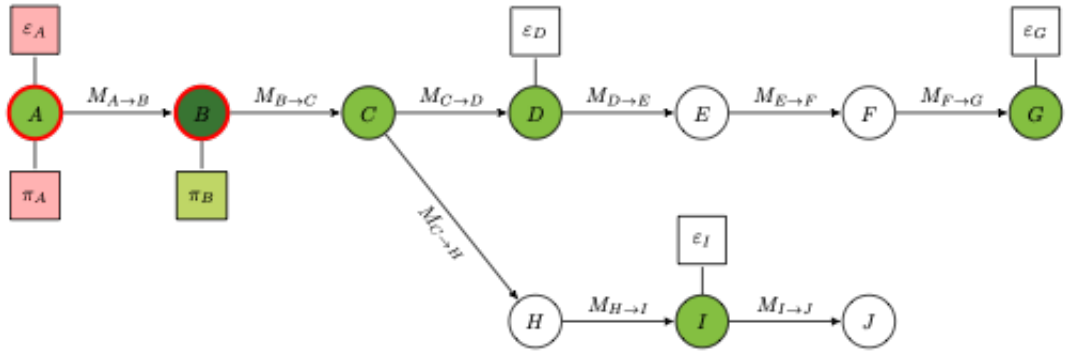
$$\begin{aligned}\pi_V(v) &= \Pr[V = v | \text{evidence}^+(V)] \\ &= \Pr[V = v | V = v_*, \text{evidence}^+_*(V)]\end{aligned}$$

where $\text{evidence}^+_*(V) = \text{evidence}^+(V) \setminus \{V = v_*\}$ denotes the remaining evidence upstream of V. Again the evidence $V = v_*$ is the most direct information about V, the remaining evidence $\text{evidence}^+_*(V)$ is irrelevant unless $\text{evidence}^+_*(V)$ directly contradicts $V = v_*$. In this case, nothing can be done and prior is not defined at all.

Thus, we can simplify and get an indicator prior:

$$\begin{aligned}\pi_V(v) &= \Pr[V = v | V = v_*] \\ &= [v = v_*] \quad .\end{aligned}$$

Prior update if a node is without evidence while its predecessor is with direct evidence



In the analysis above we obtained the formula

$$\pi_V(v) \propto \sum_{u\in U} M_{U\to V}[u,v]\cdot\pi_U(u)\cdot \Pr[\text{evidence}^-(W_1),\ldots,\text{evidence}^-(W_{k-1})|U=u$$

that holds for any predecessor U, provided that V is without direct evidence. If the node U has direct evidence $U=u_*$ then $\pi_U(u)\propto[u=u_*]$ and consequently the sum reduces to a single term:

$$\pi_V(v) \propto M_{U\to V}[u_*,v]\cdot \Pr[\text{evidence}^-(W_1),\ldots,\text{evidence}^-(W_{k-1})|U=u_*] \quad .$$

Morover, the second factor does not depend on the value of v and thus we can further simplify:

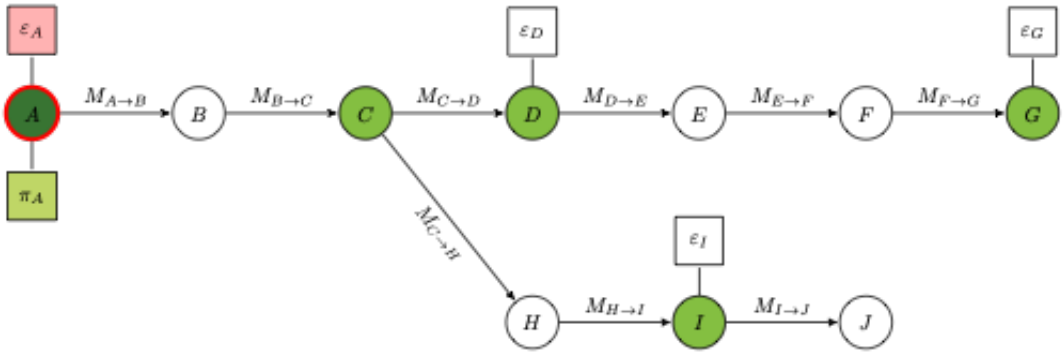
$$\begin{aligned}\pi_V(v) &\propto M_{U\to V}[u_*,v]\\ &\propto \sum_{u\in U}\pi_U(u)M_{U\to V}[u,v] \quad .\end{aligned}$$

The corresponding matrix algebra formulation is

$$\begin{aligned}\pi_V &\propto \pi_U M_{U\to V}\\ &\propto \frac{\pi_U}{\lambda_k}M_{U\to V}\end{aligned}$$

which is formally the same as for the non-exceptional case, although the special case formula is clearer and easier to understand.

Prior update for a node without a predecessor



The rules for updating the prior are applicable for nodes that do have predecessors. Hence, we must address nodes without predecessors explicitly. If there is a direct evidence $V = v_*$ then obviously

$$\begin{aligned}\pi_V(v) &= \Pr[V = v|\text{evidence}^+(V)] \\ &= \Pr[V = v|V = v_*] \\ &= [v = v_*]\end{aligned}$$

and thus

$$\pi_V \propto [v = v_*] \quad .$$

If there is no evidence then by definition

$$\begin{aligned}\pi_V(v) &= \Pr[V = v|\text{evidence}^+(V)] \\ &= \Pr[V = v] \\ &= M_V[v]\end{aligned}$$

where M_V is the vector of initial probabilities. The corresponding matrix formulation is

$$\pi_V \propto M_V \quad .$$