0.1 Inference of model parameters

In the formalism of the last sections, the P_{HMM} depends implicitly through P_{BN} on parameters $\theta = \left\{ \tilde{b}_v, \tilde{t}_{pv} \mid v \leq V, p \in \mathrm{pa}(v) \right\}$, which – as mentioned – are now probability rates and have therefore a slightly different interpretation. Due to the marginalization over time-steps in **??** the likelihood function additionally depends on the choice and parametrization of the prior $p(t)$. The parameters are to be inferred from a dataset of lymphatic progression patterns in a cohort of patients. We still assume that for each patient we record for every lymph node level (LNL) v whether it is involved according to only one diagnostic modality. In other words, for each patient we observe one of the 2^V possible diagnoses. As mentioned before, we will expand this to multiple diagnostic modalities furhter down in **??**.

Formally, we can then express the dataset $\boldsymbol{\mathcal{Z}}$ of N patients as vector $\mathbf{f}$ of the number of patients f_i for which the diagnosis corresponds to the observational state $\boldsymbol{\zeta}_i$. The likelihood $P(\boldsymbol{\mathcal{Z}} \mid \theta)$ of observing this dataset, given a particular choice of parameters, is then given by

$$P(\boldsymbol{\mathcal{Z}} \mid \theta) = \prod_{i=1}^{2^V} P(\boldsymbol{\zeta}_i \mid \theta)^{f_i} \tag{1}$$

with the probability $P(\boldsymbol{\zeta}_i \mid \theta)$ specified by **??**. The product runs formally over all possible observational states. In reality, f_i will likely be zero for a number of rare or implausible states that are not in the dataset. Note that $\sum_i f_i = N$.

By Bayes' rule, the posterior distribution of those parameters is

$$P(\theta \mid \boldsymbol{\mathcal{Z}}) = \frac{P(\boldsymbol{\mathcal{Z}} \mid \theta) \, P(\theta)}{\int P(\boldsymbol{\mathcal{Z}} \mid \theta') \, P(\theta') \, d\theta'} \tag{2}$$

where $P(\theta)$ is the prior over these parameters. Since they are exclusively probability rates, they must all come from the interval $[0,1] \in \mathbb{R}$. In this work we will choose the most uninformative prior

$$p(\theta) = \begin{cases} 1 & \text{if } \theta_r \in [0,1] \, ; \forall r \leq E \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where E is the number of edges in the directed acyclic graph (DAG) we use to represent the lymphatic system. While it is easy to compute the likelihood, it is not feasible to efficiently calculate the normalization constant in the denominator of eq. (2). Hence, we will use Markov-chain Monte Carlo (MCMC) sampling methods to estimate the parameters θ and their uncertainty.