Chapter 1

Unilateral hidden Markov model

This chapter concerns itself with modelling the unilateral lymphatic spread using the formalism of hidden Markov models (HMMs) that will also be introduced below.

The content of this chapter is largely based on our publication [1] with some modifications and additions to improve continuity with the next chapter, where the presented model will be extended.

1.1 Formulating lymphatic progression as HMM

We consider discrete time-steps $t \in \{0, 1, 2, \ldots, t_{\max}\}$. We will start by defining the hidden random variable for the state of the HMM at time t to be

$$\mathbf{X}[t] = (X_v[t]) \tag{1.1}$$

which represents the patient's state of lymph node level (LNL) involvement as in the Bayesian network (BN), but for each time-step we have an instance of it. For the diagnosis $\mathbf{Z}$ on the other hand, we do not need to differentiate between different times, since in practice we will only ever see one diagnosis. This is illustrated in **??**. The reason for this is that, if we diagnose a patient with cancer, treatment starts timely and we no longer observe the natural progression of the disease. From a modeling standpoint however, this is a problem that we will address later.

A hidden Markov model is fully described by the starting state $\mathbf{X}[0] \coloneqq \boldsymbol{\pi}$ and the two conditional probability functions that govern the progression from a state $X[t]$ at time t to a state $X[t+1]$ at the following time-step

$$P_{HMM}\left(\mathbf{X}[t+1] \mid \mathbf{X}[t]\right) \tag{1.2}$$

and the probability of a diagnostic observation given the true state of the patient

$$P_{HMM}\left(\mathbf{Z} \mid \mathbf{X}[t]\right) \tag{1.3}$$

Since both our state space and our observation space are discrete and finite, it is possible to enumerate all possible states and observations and collect them in a table or matrix. This so-called *transition matrix* would then be

$$\mathbf{A} = (A_{ij}) = \left(P_{HMM}\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j\right)\right) \tag{1.4}$$

and the *observation matrix*

$$\mathbf{B} = (B_{ij}) = \left(P_{HMM}\left(\mathbf{Z} = \boldsymbol{\zeta}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j\right)\right) \tag{1.5}$$

Here, $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_j$ are no new variables, but just $\mathbf{x}$ and $\mathbf{z}$ renamed and reordered. The indices i and j for one of the possible states or observations for the entire patient, not for an individual LNL. In total, there are $S = |\{0,1\}|^V$ different states and the same number of different possible observations per diagnostic modality. We order the hidden states from

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \tag{1.6}$$

to

$$\boldsymbol{\xi}_{16} = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \tag{1.7}$$

in this case of $V = 4$. The exact ordering does not matter, it is just a convenience for the notation. our ordering of the states can be seen in the axes of **??**. In analogy, we order the observations $\boldsymbol{\zeta}_j$ from 1 to 2^V. Note that for now we will not consider multiple diagnostic modalities and how to combine them. We will get back to that topic in **??**.

In our case, the starting state corresponds to a primary tumor being present but all LNLs are still in the healthy state. The observation matrix $\mathbf{B}$ is specified via sensitivity and specificity as described in eq. (1.7). The main task is to infer the transition matrix $\mathbf{A}$. Usually, it is inferred from a series of observations and there exist efficient algorithms for that, e.g. the sum-product algorithm, which is particularly efficient in chains. Unfortunately, these algorithms cannot be applied for our problem for two profound reasons:

1. We only have a single observation instead of a consecutive series of observations.

2. It is unclear how many time-steps it took from the starting state to the one observation we have at the time of diagnosis.

In the remainder of this section, we will detail the HMM step-by-step, starting with the parameterization of the transition matrix $\mathbf{A}$ in section 1.2. Afterwards, in section 1.3, I will tackle the aforementioned problems, followed up by explaining how we perform inference on this model (section 1.4), incorporate information about a patient's T-stage (section 1.5) and assess the risk of LNL involvement in a new patient (section 1.7). Lastly, we will introduce a way to incorporate incomplete observations in section 1.8.

1.2 Parametrization of the transition matrix

The square transition matrix $\mathbf{A}$ has $S = 2^{2V}$ entries and therefore $S(S-1) = 2^{2V} - 2^V$ degrees of freedom. Although searching the full space of viable transition matrices is possible via unparametrized sampling techniques, it is computationally challenging and hard to interpret. To achieve this reduction in degrees of freedom, and also preserve the anatomically and medically motivated structure of the Bayesian network from **??**, we can represent the transition probability from one state $\mathbf{x}[t]$ to another state $\mathbf{x}[t+1]$ using the conditional probabilities defined for

the BN. The difference is that the probability of observing a certain state of LNL v now depends on the state of the patient one time-step before. Note that from here on, we will mostly drop the probabilistically correct notation $P(X = x)$ and just write $P(x)$ for brevity

$$P_{HMM}\left(\mathbf{x}[t+1] \mid \mathbf{x}[t]\right) = \prod_{v \leq V} Q\left(x_v[t+1]; x_v[t]\right) \times \left[P_{BN}\left(x_v[t+1] \mid \left\{x_r[t]\,,\, \tilde{t}_{rv}\right\}_{r \in \mathrm{pa}(v)}, \tilde{b}_v\right)\right]^{1-x_v[t]} \tag{1.8}$$

Here we have reused the conditional probability from the BN for each LNL, but we take it to the power of one minus that node's previous value. This ensures that an involved node stays involved with probability 1. The parameters $\tilde{t}_{\mathrm{pa}(v)v}$ and $\tilde{b}_v$ take the same role as in the BN, but they are now probability *rates*, since they act per time-step. Lastly, the first term Q in the product formalizes the fact that a metastatic lymph node level cannot become healthy again once it was involved. This also means that several entries in the transition matrix $\mathbf{A}$ must be zero. In a table the values of $Q\left(x_v[t+1]; x_v[t]\right)$ can be written like this:

$$\begin{aligned} Q\left(X_v[t+1] = 0; X_v[t] = 0\right) &= 1 \\ Q\left(X_v[t+1] = 0; X_v[t] = 1\right) &= 0 \\ Q\left(X_v[t+1] = 1; X_v[t] = 0\right) &= 1 \\ Q\left(X_v[t+1] = 1; X_v[t] = 1\right) &= 1 \end{aligned} \tag{1.9}$$

which gives rise to a "mask" for $\mathbf{A}$ which can be seen in **??**.

To illustrate eq. (1.8), it helps to look at a specific example. E.g., the transition probability from state $\boldsymbol{\xi}_5 = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix}$ to state $\boldsymbol{\xi}_7 = \begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}$, which represents starting with involvement only in LNL II and asking for the probability that LNL III becomes involved as well over the next time-step:

$$\begin{aligned} &P_{HMM}\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_7 \mid \mathbf{X}[t] = \boldsymbol{\xi}_5\right) \\ &\quad = Q\left(X_1[t+1] = 0; X_1[t] = 0\right) P_{BN}\left(X_1[t+1] = 0 \mid \tilde{b}_1\right)^1 \\ &\quad \times Q\left(X_2[t+1] = 1; X_2[t] = 1\right) P_{BN}\left(X_2[t+1] = 1 \mid X_1[t] = 0, \tilde{t}_{12}, \tilde{b}_2\right)^0 \\ &\quad \times Q\left(X_3[t+1] = 1; X_3[t] = 0\right) P_{BN}\left(X_3[t+1] = 1 \mid X_2[t] = 1, \tilde{t}_{23}, \tilde{b}_3\right)^1 \\ &\quad \times Q\left(X_4[t+1] = 0; X_4[t] = 0\right) P_{BN}\left(X_4[t+1] = 0 \mid X_3[t] = 0, \tilde{t}_{34}, \tilde{b}_4\right)^1 \\ &\quad = \left(1 - \tilde{b}_1\right) \cdot 1 \cdot \left(\tilde{b}_3 + \tilde{t}_{23} - \tilde{b}_3 \tilde{t}_2 3\right) \cdot \left(1 - \tilde{b}_4\right) \end{aligned} \tag{1.10}$$

The interpretation of the last line is that this is the probability that LNL I and IV do not become involved, while LNL III gets infected through lymphatic drainage from either the main tumor or LNL II. The probability of LNL II remaining involved is 1, of course, which is why we take the respective term to the power of 0.

1.3 Marginalization

To calculate the likelihood function, we have to calculate the probability of a given diagnostic observation. To that end, we first calculate the probability of observing a given diagnosis $\mathbf{z} = \boldsymbol{\zeta}_j$ at a fixed time-step t. As depicted in **??**, we must consider every possible evolution of a patient's disease that leads to the observed diagnosis. Mathematically, this means that we need to marginalize over all such paths. And here is where the HMM-formalism comes in very useful, because this marginalization happens automatically when we multiply the transition matrix with itself and eventually with the observation matrix:

$$P\left(\mathbf{Z} = \boldsymbol{\zeta}_j \mid t\right) = \left[\boldsymbol{\pi}^\top \cdot (\mathbf{A})^t \cdot \mathbf{B}\right]_j \tag{1.11}$$

where the $\boldsymbol{\pi}$ is the column vector for the healthy starting state. $\mathbf{A}$ is multiplied with itself t times and thereby produces a matrix that describes the transition probability from the healthy state to all possible states $\mathbf{x}[t]$ in exactly t time-steps marginalized over the actual pathway of the patient's disease. The index $[\ldots]_j$ here means that from the resulting (row-)vector of probabilities we take the component that corresponds to the diagnose $\mathbf{z} = \boldsymbol{\zeta}_j$.

So, essentially, eq. (1.11) first computes the probability vector of all possible true hidden states, given a time step t

$$P\left(\mathbf{X} = \boldsymbol{\xi}_i \mid t\right) = \left[\boldsymbol{\pi}^\top \cdot (\mathbf{A})^t\right]_i \tag{1.12}$$

and then multiplies it with the respective observation probability vector, which is a column of the $\mathbf{B}$ matrix, to finally marginalize over all possible true hidden states – effectively a sum over i in eq. (1.12) – at the time t of diagnosis.

The problem that the number of time-steps until diagnosis is unknown cannot be solved in such an elegant fashion. Therefore, we must resort to brute force marginalization and introduce a prior $p(t)$, which is a discrete distribution over a finite number of time-steps. It describes the prior probability that a patient's cancer is diagnosed at a particular time-step t. To get the probability of a diagnosis $\mathbf{z}$ we must compute

$$P\left(\mathbf{Z} = \boldsymbol{\zeta}_j\right) = \sum_{t=0}^{t_{\max}} p(t) \cdot P\left(\mathbf{Z} = \boldsymbol{\zeta}_j \mid t\right) = \left[\sum_{t=0}^{t_{\max}} p(t) \cdot \boldsymbol{\pi}^\top \cdot (\mathbf{A})^t \cdot \mathbf{B}\right]_j \tag{1.13}$$

While the choice of the time-prior may seem unclear at this point, its role for including T-stage into this model will be discussed in section 1.5.

1.4 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 eq. (1.13) 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 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

$$P(\boldsymbol{\mathcal{Z}} \mid \theta) = \prod_{i=1}^{2^V} P(\boldsymbol{\zeta}_i \mid \theta)^{f_i} \tag{1.14}$$

with the probability $P(\boldsymbol{\zeta}_i \mid \theta)$ specified by eq. (1.13). 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

$$P(\theta \mid \boldsymbol{\mathcal{Z}}) = \frac{P(\boldsymbol{\mathcal{Z}} \mid \theta)\, P(\theta)}{\int P(\boldsymbol{\mathcal{Z}} \mid \theta')\, P(\theta')\; d\theta'} \tag{1.15}$$

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

$$p(\theta) = \begin{cases} 1 & \text{if } \; \theta_r \in [0,1]\,; \forall r \leq E \\ 0 & \text{otherwise} \end{cases} \tag{1.16}$$

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. (1.15). Hence, we will use Markov-chain Monte Carlo (MCMC) sampling methods to estimate the parameters θ and their uncertainty.

1.5 Incorporation of T-stage

We have introduced the HMM with the promise that it could handle the concept of T-stages through its explicit modeling of dynamic processes. To keep up with that, we will now explain how this is achieved using the time-prior $p(t)$.

The core idea is to assume that early T-stage and late T-stage tumors share the same patterns of metastatic progression, except that late T-stage tumors are on average diagnosed at a later point in time, and thereby also show, on average, higher LNL involvement. Formally, this can be described by assuming a different time-prior $p_T(t)$ for every T-stage T. On the other hand, the transition matrix $\mathbf{A}$ is assumed to be the same for all T-stages.

For the inference of model parameters, the training data is split into subgroups according to T-stage. We now define a column-vector $\mathbf{f}_T$ separately for each T-stage, which counts the number of patients in the dataset that were diagnosed with one of the possible observational states and a given T-stage. The log-likelihood

from which we want to sample is then simply a sum of the likelihoods as above, where the essential difference is that we equip each marginalization over time with a different time-prior $p_T(t)$, according to its T-stage:

$$\log P\left(\boldsymbol{\mathcal{Z}} \mid \theta\right) = \sum_{T=1}^{4} \log \left[\sum_{t=0}^{t_{\max}} p_T(t) \cdot \boldsymbol{\pi}^\top \cdot (\mathbf{A})^t \cdot \mathbf{B} \right] \cdot \mathbf{f}_T \tag{1.17}$$

The logarithm must be taken element-wise for the resulting row-vector inside the square brackets. The only data-dependent term here is the vector $\mathbf{f}_T$ counting the occurrences of all possible observations. It is again important to note that the only difference between the part of the log-likelihood for the different T-stages is the exact shape or parametrization of the time-prior. The transition probabilities, and hence also the transition matrix $\mathbf{A}$, are the same for all T-stages. For this to work, we rely on the assumption that different typical patterns of nodal involvement for the same primary tumor location are caused mainly by different progression times

At this point, it makes sense to briefly introduce a notation of the above equation that is more suitable for the actual programmatic implementation of the inference and the extension we will discuss later. We can rewrite the term in the square brackets of eq. (1.17) by using the matrix

$$\boldsymbol{\Lambda} \coloneqq P\left(\mathbf{X} \mid \mathbf{t}\right) = \begin{pmatrix} \boldsymbol{\pi}^\top \cdot (\mathbf{A})^0 \\ \boldsymbol{\pi}^\top \cdot (\mathbf{A})^1 \\ \vdots \\ \boldsymbol{\pi}^\top \cdot (\mathbf{A})^{t_{\max}} \end{pmatrix} \tag{1.18}$$

were row number t corresponds to the vector $\boldsymbol{\pi}^\top \cdot (\mathbf{A})^t$, i.e. the probabilities for all possible hidden states, given the diagnose time. So, the element $\boldsymbol{\Lambda}_{ti}$ corresponds to the probability $P\left(\boldsymbol{\xi}_i \mid t\right)$ of a patient arriving in the ith state after t time steps. With this, we can rewrite the term in the square brackets of eq. (1.17) purely as a product of vectors and matrices:

$$\sum_{t=0}^{t_{\max}} p_T(t) \cdot \boldsymbol{\pi}^\top \cdot (\mathbf{A})^t = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \tag{1.19}$$

with $p_T\left(\mathbf{t}\right) = \begin{pmatrix} p_T(0) & p_T(1) & \cdots & p_T(t_{\max}) \end{pmatrix}$. The matrix $\boldsymbol{\Lambda}$ implicitly depends on the spread probabilities, while each of the $p_T(\mathbf{t})$ depends on the respective parametrization of the time prior. They are the only objects that depend on the parameters θ and they are independent of the data.

1.5.1 * Interpretation of time-steps and time-priors

To add more interpretability to the time-prior $p(t)$ introduced in section 1.3, we want to give some insights here to what we think the time-steps and the distribution over them is supposed to mean.

First, the time that passes in the real world between the abstract time-steps t and $t+1$ should not be seen as a somewhat arbitrarily chosen fixed time, measured in days or weeks. To how much real-world time that corresponds for a specific patient is irrelevant for our risk assessment, although it might prove very valuable

for other research on tumor growth. Also, the time between two time-steps does not need to be constant; the model makes no assumptions about this. It merely assumes the probability of transition between states to be the same from t to $t+1$ and for all t.

The time-prior $p(t)$ is essentially the probability that a patient is diagnosed after exactly t time-steps. If we knew how long a patient had cancer before getting diagnosed and we also knew how long a typical timestep for this patient and his/her type of cancer was, then we could just fix $p(t) = 1$ for the appropriate number of time-steps t and set $p(t') = 0, \forall t' \neq t$. Since it is likely almost never known, we need to spread the probability over a range of time-steps, reflecting the fact that the diagnose of cancer happens spontaneously, e.g. during a routine checkup.

1.5.2 * Impact of shape and length of the time-prior

It turns out that length and shape of $p(t)$ have almost no effect on the risk predictions as long as we are not concerned with different T-categories. So, if we learn our parameters from a dataset that only contains T1 patients and then compute risks for T1 patients only, the result will not differ almost regardless of the time-prior that was used for learning and risk assessment. Only too few time-steps may pose a problem, since then the system might not be able to spread to all LNLs via all pathways. And too many time-steps could introduce numerical problems, because the learned probability rates $\tilde{b}_v$ and $\tilde{t}_{\mathrm{pa}(v)v}$ become smaller for longer time-priors.

To understand the impact of the number of time-steps T on the results, we looked at a simple analytical model: Assume that there is a system with only one LNL that the primary tumor can spread to that is empirically involved with probability $p^\star = 0.4$. For this situation, we can now derive how the base probability rate $\tilde{b}_v$ changes for a uniform time-prior

$$p(t) = \frac{1}{T} \qquad \text{for} \quad t \in \{1, 2, \ldots, T\} \tag{1.20}$$

if we vary the total number of time-steps T. We can write $p^\star$ as

$$\begin{aligned} p^\star &= \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{bmatrix} (1-\tilde{b}_1) & \tilde{b}_1 \\ 0 & 1 \end{bmatrix}^t \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{T} \sum_{t=1}^{T} \begin{pmatrix} 1 & 0 \end{pmatrix} \cdot \begin{bmatrix} (1-\tilde{b}_1)^t & 1-(1-\tilde{b}_1)^t \\ 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left[1 - (1-\tilde{b}_1)^t \right] = 1 - \frac{1}{T} \sum_{t=1}^{T} (1-\tilde{b}_1)^t \end{aligned} \tag{1.21}$$

The right-hand side essentially contains the partial sum of the geometric series and can easily be computed to yield

$$p^\star = 1 - \frac{\left(1-\tilde{b}_1\right)\left(1-(1-\tilde{b}_1)^T\right)}{\tilde{b}_1 T} \tag{1.22}$$

It is not possible to analytically solve for $\tilde{b}_1$ in the case of arbitrary T, but numerical solutions are very easy to find and are plotted in fig. 1.1. This confirms

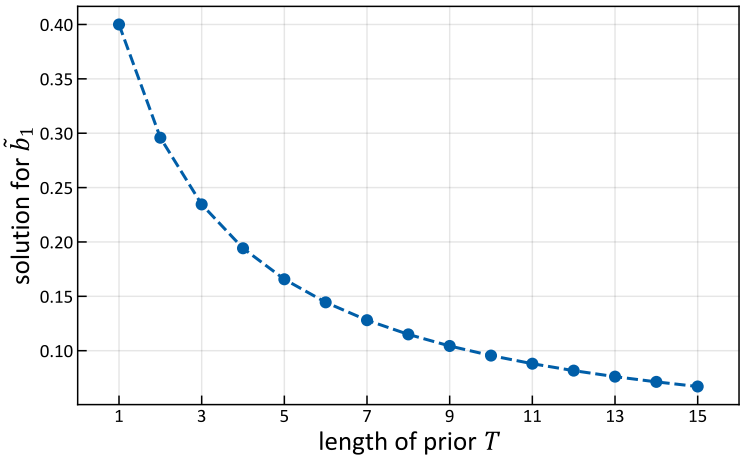

Figure 1.1: Solutions to eq. (1.22) for the base probability rate $\tilde{b}_1$ given a $p^\star$ of 0.4 and T increasing from 1 to 15.

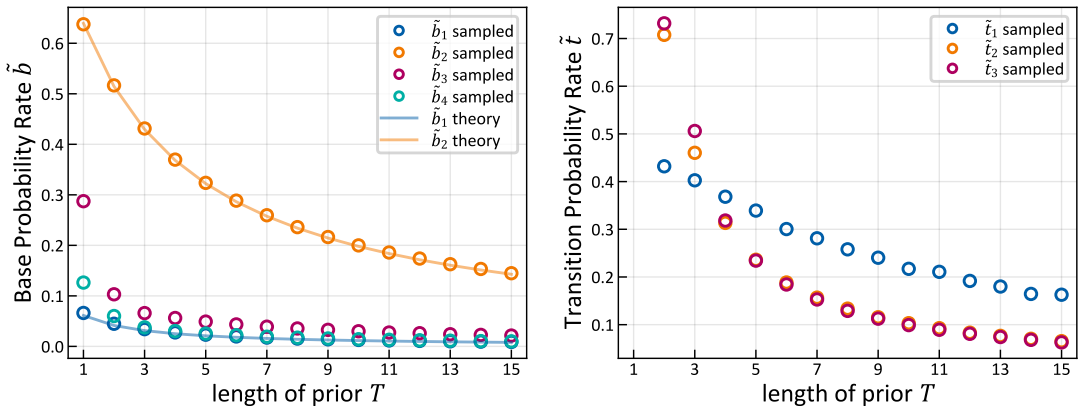

Figure 1.2: Decay of base probability rates as a function of the number of time-steps the time-prior has. Circles depict the results from learning the same dataset with different time-priors while solid lines show the analytical result starting with a $p^\star$ corresponding to the prevalence of involvement of LNL I and II respectively.

the intuition, that the base and transition probability rates become smaller when the total time over which the tumor spreads is divided into more but shorter time-steps.

Now we compare this idealized result to the decay of the probability rates for the full system. To that end, the model with LNLs I-IV was trained as in the same way as for the figures in the section above, but with differently long uniform time-priors instead of a Binomial prior. So, the probability for every time-step is $p(t) = 1/T$ for all $t \geq 1$, but zero for the starting state π. fig. 1.2 shows the expected value of the parameters as a function of T. It is important to stress again that the risk predicted by the models using all those different-length uniform time-priors was the same for $T \geq 2$. For the one-step model with $T = 1$ the risk prediction of a LNL does not depend on the diagnose. For example, we expect the risk in level III is higher, when level II is involved, due to the spread from LNL II to III. With such a short time-prior, however, the model cannot capture this, and

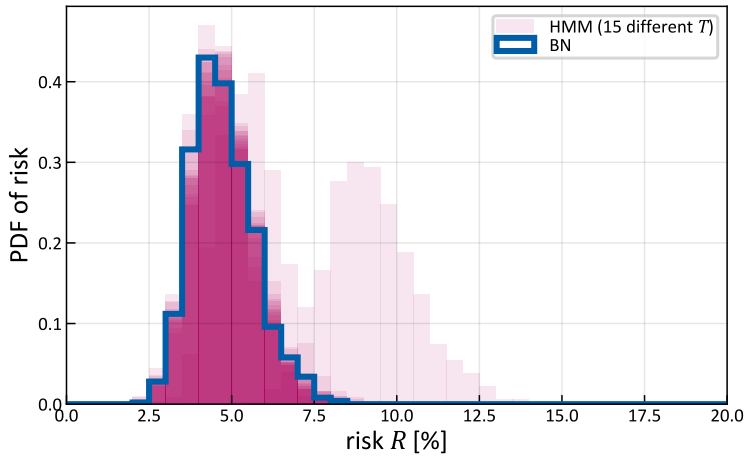

Figure 1.3: Prediction for the risk of involvement in LNL III, given that no other LNL was observed to be involved. Computed by training our hidden Markov model with 15 different-length uniform time-priors (transparent red) as well as the Bayesian network model (blue line). The one outlier is the HMM with a time-prior covering only one time-step. It is centered on the prevalence of involvement for LNL III, regardless of the given diagnose.

all risk predictions will just yield the prevalence of involvement. This effect can be seen in Figure A3 and shows what has been stated earlier: The support of the time-prior has little effect on the model's predictions, as long as it is sufficient to capture the spread through the lymphatic system.

The theoretical result in eq. (1.22) is applicable to the parameter $\tilde{b}_1$, and approximately to $\tilde{b}_2$ since involvement of level II is driven by direct infiltration from the primary tumor rather than transition from level I. For levels III and IV, the theory is not applicable as they have two relevant parent nodes. The solid lines in fig. 1.2 show agreement of the theoretical result with the sampling based training of the full model (circles), where the probabilities $p^\star$ were set to 6.1% and 63.9%, corresponding to the prevalence of level I and II involvement in the dataset, respectively.

This again shows that, while looking at one T-category only, the time-prior's parameters overdetermine the system. For any choice of T, the base and transition probability rates can be adjusted such that the Hidden Markov Model is equivalent to the Bayesian network's performance. Only if we want to distinguish between patients of different T-category the HMM can outperform the BN.

1.6 Sampling

With a parameter set $\theta = \left(\{\tilde{b}_v\}, \{\tilde{t}_{rv}\}_{r\in \mathrm{pa}(v)}\right)\ \forall v \leq V$, we can assess the risk of nodal involvement, given a diagnosis $\mathbf{z}$, of a new patient. Using Bayes' law, the

risk for a certain LNL v being involved is given by the conditional probability

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{z}, \theta\right) &= \frac{P\left(\mathbf{Z} = \mathbf{z} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \end{aligned} \tag{1.23}$$

Note that in the second line, we have explicitly written out the marginalization over all hidden states $\boldsymbol{\xi}_i$ that have LNL v involved. We have written the state of LNL v in the state $\boldsymbol{\xi}_i$ as ξ_{iv}. The denominator can be computed using eq. (1.13), which already includes the marginalization over all hidden states $\boldsymbol{\xi}_i$.

The process of sampling randomly generates L sets of parameters $\theta = \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_L\end{pmatrix}$. They are therefore random variables and so is the risk $R\left(X_v \mid \mathbf{z}, \theta\right)$ since it is a function of θ. Using the Monte Carlo estimator, we can therefore compute the moments of the distribution over the risk, including e.g. the expectation value

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R\left(X_v = 1 \mid \mathbf{z}\right)\right] = \frac{1}{L} \sum_{k=1}^{L} R\left(X_v = 1 \mid \mathbf{z}, \theta_k\right) \tag{1.24}$$

In the result sections below, we compute the individual risks for a large enough number L of sampled parameters. Thereby, we can compute histograms for the risk that will approach the real probability density of the respective risk for $L \to \infty$. This provides additional information on the uncertainty in the predicted risk resulting from uncertainty in the model parameters.

1.7 Risk assessment of microscopic involvement

With a parameter set $\theta = \left(\left\{\tilde{b}_v\right\}, \left\{\tilde{t}_{rv}\right\}_{r \in \mathrm{pa}(v)}\right)\ \forall v \leq V$, we can assess the risk of nodal involvement, given a diagnosis $\mathbf{z}$, of a new patient. Using Bayes' law, the risk for a certain LNL v being involved is given by the conditional probability

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{z}, \theta\right) &= \frac{P\left(\mathbf{Z} = \mathbf{z} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \end{aligned} \tag{1.25}$$

Note that in the second line, we have explicitly written out the marginalization over all hidden states $\boldsymbol{\xi}_i$ that have LNL v involved. We have written the state of LNL v in the state $\boldsymbol{\xi}_i$ as ξ_{iv}. The denominator can be computed using eq. (1.13), which already includes the marginalization over all hidden states $\boldsymbol{\xi}_i$.

The process of sampling randomly generates L sets of parameters $\theta = \begin{pmatrix}\theta_1 & \theta_2 & \dots & \theta_L\end{pmatrix}$. They are therefore random variables and so is the risk $R\left(X_v \mid \mathbf{z}, \theta\right)$ since it is a function of θ. Using the Monte Carlo estimator, we can therefore compute the moments of the distribution over the risk, including e.g. the expectation value

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R\left(X_v = 1 \mid \mathbf{z}\right)\right] = \frac{1}{L} \sum_{k=1}^{L} R\left(X_v = 1 \mid \mathbf{z}, \theta_k\right) \tag{1.26}$$

In the result sections below, we compute the individual risks for a large enough number L of sampled parameters. Thereby, we can compute histograms for the risk that will approach the real probability density of the respective risk for $L \to \infty$. This provides additional information on the uncertainty in the predicted risk resulting from uncertainty in the model parameters.

1.8 Inference and risk assessment for incomplete diagnoses

A diagnosis is often not complete, meaning that not all LNLs might have been checked with a diagnostic modality. E.g., fine needle aspiration (FNA) is usually only performed in a subset of LNLs. Hence, we must be able to deal with "incomplete" observations for some LNLs. To do so, we first introduce a new observation variable

$$d_v \in \{0, 1, \emptyset\} \tag{1.27}$$

where $\emptyset$ indicates *unobserved*. Furthermore, we define a *match function*

$$\text{match}(\mathbf{d}, \mathbf{z}) \coloneqq \begin{cases} \text{true} & \text{if } d_v = z_v \vee d_v = \emptyset;\ \forall v \\ \text{false} & \text{else} \end{cases} \tag{1.28}$$

which returns *true* if a - potentially incomplete - diagnosis $\mathbf{d}$ is consistent with a complete observation $\mathbf{z}$. We will use this function for conveniently marginalizing over the missing observations. In analogy to eq. (1.25), we can compute the risk for an incomplete observation as

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{d}, \theta\right) &= \frac{P\left(\mathbf{d} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \end{aligned} \tag{1.29}$$

where the enumerator of the second line can now be rewritten using the match function:

$$\begin{aligned} P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right) &= \sum_{\left\{j\,:\,\text{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_i, \theta\right) P\Big(\boldsymbol{\xi}_i \mid \theta\Big) \\ &= \sum_{\left\{j\,:\,\text{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} B_{ij} \Big[p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda}\Big]_i \end{aligned} \tag{1.30}$$

In this case B_{ij} denotes the element of the observation matrix that corresponds to state $\boldsymbol{\xi}_i$ and observation $\boldsymbol{\zeta}_j$. Again, the indices $\{i : \xi_{iv} = 1\}$ in eq. (1.29) correspond to all possible states with a positive involvement in lymph node level X_v. Essentially, the whole term is the likelihood of an observation $\mathbf{d}$ where we have removed all entries that correspond to states with $X_v \neq 1$ both from the observation matrix and the resulting probability vector of the evolution. It can therefore be easily computed algebraically, too.

The evidence in the denominator of eq. (1.29) becomes a marginalization over all possible diagnoses that are not available to us or that we deem unimportant

$$P(\mathbf{d} \mid \theta) = \sum_{\{j \,:\, \mathrm{match}(\mathbf{d},\boldsymbol{\zeta}_j)\}} \Big[p_T(\mathbf{t}) \cdot \boldsymbol{\Lambda} \Big]_j \tag{1.31}$$

We can make this summation a bit more elegant using a column vector $\mathbf{c}^{\mathbf{d}}$ that has entries corresponding to the match-function

$$c_i^{\mathbf{d}} = \mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_i) \tag{1.32}$$

where every *true* corresponds to a 1 and every *false* to a 0. This way we can rewrite eq. (1.31) in the following way:

$$P(\mathbf{d} \mid \theta) = p_T(\mathbf{t}) \cdot \boldsymbol{\Lambda} \cdot \mathbf{B} \cdot \mathbf{c}^{\mathbf{d}} \tag{1.33}$$

Using this notation for marginalizing over unknown or incomplete observations also allows us to encode entire datasets $\boldsymbol{\mathcal{D}} = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \cdots & \mathbf{d}_N \end{pmatrix}$ of (potentially incomplete) observations in the form of a matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}^{\mathbf{d}_1} & \mathbf{c}^{\mathbf{d}_2} & \cdots & \mathbf{c}^{\mathbf{d}_N} \end{pmatrix} \tag{1.34}$$

so that the row-vector of likelihoods reads as

$$P(\boldsymbol{\mathcal{D}} \mid \theta) = \big(P(\mathbf{d}_n \mid \theta) \big)_{n \in [1,N]} = p_T(\mathbf{t}) \cdot \boldsymbol{\Lambda} \cdot \mathbf{B} \cdot \mathbf{C} \tag{1.35}$$

1.9 Multiple diagnostic modalities

Throughout the last sections, we have only dealt with diagnoses from a single modality. In practice, however, most patients undergo screening for metastases using different modalities, like computed tomography (CT), magnetic resonance imaging (MRI) or FNA. The sensitivities and specificities of these might vary greatly and by combining them in a probabilistically rigorous way, we may gain a additional information.

Luckily, the introduced formalism requires very little changes to be able to incorporate multiple diagnostic modalities. Let $\mathcal{O} = \{\mathrm{CT}, \mathrm{MRI}, \mathrm{FNA}, \ldots\}$ be the set of modalities. Then we can extend the collection of observed binary random variables (RVs) $\mathbf{z}$ from a single modality

$$\mathbf{z} = (x_v)_{v \in [1,V]} = \begin{pmatrix} x_1 & \cdots & x_V \end{pmatrix} \tag{1.36}$$

to multiple diagnostic modalities

$$\mathbf{z} = \big(x_v^k\big)_{\substack{v \in [1,V] \\ k \in [1,|\mathcal{O}|]}} = \begin{pmatrix} x_1^1 & \cdots & x_V^1 & x_2^2 & \cdots & x_V^{|\mathcal{O}|} \end{pmatrix} \tag{1.37}$$

where k enumerates the elements in the set $\mathcal{O}$. We can use $\boldsymbol{\zeta}_j$ again and this time the counting variable j goes from 1 to $2^{V \cdot |\mathcal{O}|}$. Notice that this means the observation matrix $\mathbf{B}$ is not square anymore. Also, it now contains the sensitivities and specificities of all the modalities in $\mathcal{O}$. If we had separate square observation

matrices $\mathbf{B}^k$ for each diagnostic modality, the new total matrix' rows B_{i*} would be the outer products of the individual observation matrices:

$$B_{i*} = B^1_{i*} \otimes B^2_{i*} \otimes \cdots \otimes B^{|\mathcal{O}|}_{i*} \tag{1.38}$$

Completely analogous to how we enlarged the vector of binary RVs $\mathbf{z}$, we can also extend the vectors $\mathbf{c}$ and $\mathbf{d}$ and then immediately use the entire formalism of the section before to model lymphatic progression with potentially incomplete diagnoses from multiple modalities. However, we will drop this way of continuously enumerating the observations in the next section again, because there is a slightly more efficient and elegant way to do it. This section only served to show that it is naturally possible to extend the formalism to combine findings from different diagnostic modalities.

1.10 Combining modalities and data

Note that the matrix $\mathbf{B}$ – and also the matrix $\mathbf{C}$ – can get very large very quickly: The former is of size $2^V \times 2^{V \cdot |\mathcal{O}|}$ and the latter has dimensions $2^{V \cdot |\mathcal{O}|} \times N$, meaning both grow exponentially with the number of LNLs *and* diagnostic modalities. And although neither $\mathbf{B}$ not $\mathbf{C}$ depend on the parameters θ, meaning their product can be precomputed, we can simply iterate over all patients, possible hidden states and available diagnostic modalities to compute $\mathbf{\Omega} \coloneqq \mathbf{B} \cdot \mathbf{C}$ directly, which saves us building up and multiplying matrices with potentially millions of entries.

To compute this matrix $\mathbf{\Omega}$, we first abandon the just-introduced way of combining diagnoses for all modalities into one large vector and separate them again, so that we have complete and incomplete observations $\boldsymbol{\zeta}^k_j$ and $\mathbf{d}^k_n$ respectively for each modality, where $n \in [1, N]$ enumerates the patients in the data.

$$\begin{aligned} \Omega_{mn} = P\left(\mathbf{d}_n \mid \boldsymbol{\xi}_m\right) &= \prod_{k=1}^{|\mathcal{O}|} P\left(\mathbf{d}^k_n \mid \boldsymbol{\xi}_m\right) \\ &= \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j \,:\, \text{match}(\mathbf{d}^k_n, \boldsymbol{\zeta}^k_j)} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_m\right) \right] = \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j \,:\, \text{match}(\mathbf{d}^k_n, \boldsymbol{\zeta}^k_j)} B^k_{mj} \right] \end{aligned} \tag{1.39}$$

Now, the elements Ω_{mn} encode the observation likelihood of patient n's diagnose $\mathbf{d}_n$ given their true state of involvement is $\boldsymbol{\xi}_m$. Finally, with this the row-vector of likelihoods of a cohort of patients, given the model's spread parameters, becomes

$$P\left(\boldsymbol{\mathcal{D}} \mid \theta\right) = p_T\left(\mathbf{t}\right) \cdot \mathbf{\Lambda} \cdot \mathbf{\Omega} \tag{1.40}$$

Again, the objects $p_T(\mathbf{t})$ and $\mathbf{\Lambda}$ depend on the parameters and hence need to be recalculated for every sample drawn during MCMC inference. $\mathbf{\Omega}$ depends only on the patient data $\mathcal{D}$ and must therefore only be computed once at the beginning of the learning round.