Chapter 1

Hidden Markov model

This chapter concerns itself with the modeling part of this work and will introduce the hidden Markov model (HMM) based probabilistic models in the order they were developed.

1.1 Unilateral

1.1.1 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 1.1.2. Afterwards, in section 1.1.3, I will tackle the aforementioned problems, followed up by explaining how we perform inference on this model (section 1.1.4), incorporate information about a patient's T-stage (section 1.1.5) and assess the risk of LNL involvement in a new patient (section 1.1.7). Lastly, we will introduce a way to incorporate incomplete observations in section 1.1.8.

1.1.2 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.

1.1.3 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 1.1.5.

1.1.4 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 $\mathcal{Z}$ 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.

1.1.5 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:

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

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

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

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 $\mathbf{\Lambda}_{ti}$ 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:

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

with $p_T\left(\mathbf{t}\right) = \begin{pmatrix} p_T(0) & p_T(1) & \cdots & p_T(t_\text{max}) \end{pmatrix}$. The matrix $\mathbf{\Lambda}$ 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.

1.1.6 Sampling

For learning we employed the `python` implementation of an advanced ensemble sampler called `emcee` [2] based on an affine invariant ensemble sampler [3] to draw parameter samples from the likelihood in eq. (1.17). Although sampling is the slowest and least preferable option of inference it is also without doubt in a large number of cases the only available option and in our case even feasible; we get relatively short auto-correlation times (around a couple of hundred steps) and an average modern multi-core CPU can easily draw hundreds of thousands of samples within minutes. Many distributions in the form of histograms we show in this work are made by computing the respective quantity – e.g., the risk (see below) – for a subset of the sampled parameters. We typically randomly select between 1% and 2% of the 200,000 samples drawn after the so-called burn-in phase, when the sampling has already converged to the target distribution, as a subset. The learned parameter densities are depicted as a corner [1] plot (e.g. in **??**).

1.1.7 Risk assessment of microscopic involvement

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

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{z}, \theta\right) &= \frac{P\left(\mathbf{Z} = \mathbf{z} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{Z} = \mathbf{z} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{Z} = \mathbf{z} \mid \theta\right)} \end{aligned} \tag{1.20}$$

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

$$\mathbb{E}_{\boldsymbol{\theta}}\left[R\left(X_v = 1 \mid \mathbf{z}\right)\right] = \frac{1}{L}\sum_{k=1}^{L} R\left(X_v = 1 \mid \mathbf{z}, \theta_k\right) \tag{1.21}$$

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.

1.1.8 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

$$d_v \in \{0, 1, \emptyset\} \tag{1.22}$$

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

$$\mathrm{match}(\mathbf{d}, \mathbf{z}) \coloneqq \begin{cases} \text{true} & \text{if } d_v = z_v \vee d_v = \emptyset;\ \forall v \\ \text{false} & \text{else} \end{cases} \tag{1.23}$$

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. (1.20), we can compute the risk

for an incomplete observation as

$$\begin{aligned} R\left(X_v = 1 \mid \mathbf{d}, \theta\right) &= \frac{P\left(\mathbf{d} \mid X_v = 1, \theta\right) P\left(X_v = 1 \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \\ &= \sum_{i\,:\,\xi_{iv}=1} \frac{P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right)}{P\left(\mathbf{d} \mid \theta\right)} \end{aligned} \tag{1.24}$$

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

$$\begin{aligned} P\left(\mathbf{d} \mid \boldsymbol{\xi}_i, \theta\right) P\left(\boldsymbol{\xi}_i \mid \theta\right) &= \sum_{\left\{j\,:\,\mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_i, \theta\right) P\Big(\boldsymbol{\xi}_i \mid \theta\Big) \\ &= \sum_{\left\{j\,:\,\mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} B_{ij} \Big[p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda}\Big]_i \end{aligned} \tag{1.25}$$

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. (1.24) 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. (1.24) becomes a marginalization over all possible diagnoses that are not available to us or that we deem unimportant

$$P\left(\mathbf{d} \mid \theta\right) = \sum_{\left\{j\,:\,\mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_j)\right\}} \Big[p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda}\Big]_j \tag{1.26}$$

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

$$c_i^{\mathbf{d}} = \mathrm{match}(\mathbf{d}, \boldsymbol{\zeta}_i) \tag{1.27}$$

where every *true* corresponds to a 1 and every *false* to a 0. This way we can rewrite eq. (1.26) in the following way:

$$P\left(\mathbf{d} \mid \theta\right) = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \cdot \mathbf{B} \cdot \mathbf{c}^{\mathbf{d}} \tag{1.28}$$

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

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}^{\mathbf{d}_1} & \mathbf{c}^{\mathbf{d}_2} & \cdots & \mathbf{c}^{\mathbf{d}_N} \end{pmatrix} \tag{1.29}$$

so that the row-vector of likelihoods reads as

$$P\left(\boldsymbol{\mathcal{D}} \mid \theta\right) = \Big(P\left(\mathbf{d}_n \mid \theta\right)\Big)_{n \in [1,N]} = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \cdot \mathbf{B} \cdot \mathbf{C} \tag{1.30}$$

1.1.9 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 $\mathcal{O} = \{\text{CT}, \text{MRI}, \text{FNA}, \ldots\}$ be the set of modalities. Then we can extend the collection of observed binary random variables (RVs) $\mathbf{z}$ from a single modality

$$\mathbf{z} = (x_v)_{v\in[1,V]} = \begin{pmatrix} x_1 & \cdots & x_V \end{pmatrix} \tag{1.31}$$

to multiple diagnostic modalities

$$\mathbf{z} = \left(x_v^k\right)_{\substack{v\in[1,V] \\ k\in[1,|\mathcal{O}|]}} = \begin{pmatrix} x_1^1 & \cdots & x_V^1 & x_2^2 & \cdots & x_V^{|\mathcal{O}|} \end{pmatrix} \tag{1.32}$$

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:

$$B_{i*} = B_{i*}^1 \otimes B_{i*}^2 \otimes \cdots \otimes B_{i*}^{|\mathcal{O}|} \tag{1.33}$$

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.

1.1.10 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 $\boldsymbol{\Omega} \coloneqq \mathbf{B} \cdot \mathbf{C}$ directly, which saves us building up and multiplying matrices with potentially millions of entries.

To compute this matrix $\boldsymbol{\Omega}$, 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 $\boldsymbol{\zeta}_j^k$ and $\mathbf{d}_n^k$ respectively for each modality, where $n \in [1, N]$ enumerates the patients in the data.

$$\begin{aligned}\Omega_{mn} = P\left(\mathbf{d}_n \mid \boldsymbol{\xi}_m\right) &= \prod_{k=1}^{|\mathcal{O}|} P\left(\mathbf{d}_n^k \mid \boldsymbol{\xi}_m\right) \\ &= \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j\,:\,\mathrm{match}(\mathbf{d}_n^k, \boldsymbol{\zeta}_j^k)} P\left(\boldsymbol{\zeta}_j \mid \boldsymbol{\xi}_m\right) \right] = \prod_{k=1}^{|\mathcal{O}|} \left[\sum_{j\,:\,\mathrm{match}(\mathbf{d}_n^k, \boldsymbol{\zeta}_j^k)} B_{mj}^k \right]\end{aligned} \tag{1.34}$$

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

$$P\left(\boldsymbol{\mathcal{D}} \mid \theta\right) = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{\Omega} \tag{1.35}$$

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

1.2 Bilateral

In the previous chapter we have set up the formalism to deal with only one side of the neck. Implicitly, we have assumed that to be the ipsilateral side, i.e. the side of the sagittal plane where the primary tumor is located. This is because we assume lymphatic drainage to a process that is somewhat symmetric w.r.t. the sagittal plane, which means there can only be limited lymph flow across this plane. But depending on the tumor's location and lateralization, drainage and hence metastatic spread to the contralateral lymphatic system of the neck may also occur. In current clinical practice, a bilateral neck dissection or irradiation is often prescribed when the tumor is close to the mid-sagittal plane. So, ideally, we would like to model the risk for involvement in both sides of the neck at the same time.

The formalism of **??** can easily be applied to the contralateral side and given respective training data for the sampling process, it would learn the appropriate spread probabilities to and among the contralateral LNLs just as it would learn the ones for the ipsilateral side. From clinical experience, the contralateral involvement is usually less severe than the ipsilateral one, and hence we would expect the contralateral spread to be less probable as well.

However, combining two such unilateral models naively would make the assumption that ipsi- and contralateral spread are independent, which seems unlikely: If we know a patient has advanced metastases in the contralateral neck nodes, the risk to find similarly or even more advanced disease in ipsilateral neck nodes should probably be higher than if the contralateral neck were healthy. In other words, we are now looking for the joint probability $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)$, where the superscripts i and c indicate the ipsi- and contralateral side respectively.

The following section will pick up the unilateral formalism, extend and modify it to come up with a less naive bilateral model.

1.2.1 Expanding the unilateral model

If we start by dissecting this joint conditional probability in the following way

$$P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid \mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right) = \frac{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}} \mid \mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) \cdot P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)}{P\left(\mathbf{Z}^{\mathrm{i}}, \mathbf{Z}^{\mathrm{c}}\right)} \tag{1.36}$$

we notice right away that the likelihood on the right factorizes: Given the true states of involvement in the two sides of the neck, their respective diagnoses must be independent. Furthermore, the two factors are already given by their corresponding observation matrices $\mathbf{B}^{\mathrm{i}}$ and $\mathbf{B}^{\mathrm{c}}$.

The joint probability of the hidden states $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)$ does not factorize in the same manner. But if we assume the lymphatic network to be symmetric and directed, there can be no direct connection between LNLs of the two sides of the neck, which means the probability for involvement of the ipsi- and contralateral side only correlate via the diagnose time t. Hence the joint probability is a sum of factorizing terms:

$$\begin{aligned} P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right) &= \sum_{t \in \mathbb{T}} p(t) \cdot P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}} \mid t\right) \\ &= \sum_{t \in \mathbb{T}} p(t) \cdot P\big(\mathbf{X}^{\mathrm{c}} \mid t\big) \cdot P^{\top}\big(\mathbf{X}^{\mathrm{i}} \mid t\big) \end{aligned} \tag{1.37}$$

Note that the two row vectors of probabilities in the second line are multiplied using an outer product. Using the notation from the last section, We can write this in an algebraic way to effectively factorize this sum as follows

$$P\left(\mathbf{X}^{\mathrm{c}} = \boldsymbol{\xi}_n, \mathbf{X}^{\mathrm{i}} = \boldsymbol{\xi}_m\right) = \left[\boldsymbol{\Lambda}_{\mathrm{c}}^{\top} \cdot \operatorname{diag} p(\mathbf{t}) \cdot \boldsymbol{\Lambda}_{\mathrm{i}}\right]_{n,m} \tag{1.38}$$

where the $\boldsymbol{\Lambda}$ are again matrices with rows of the conditional probabilities $P\left(\mathbf{X} \mid t\right)$ which can be computed as defined in eq. (1.18). Multiplying these two matrices – one for the contralateral side from the left and one for the ipsilateral side from the right – onto a diagonal matrix containing the time prior marginalizes over the diagnose time and results in a matrix where the value in row n and column m represents the probability to find the contralateral neck in state $\mathbf{X}^{\mathrm{c}} = \boldsymbol{\xi}_n$ and the ipsilateral lymphatic system in state $\mathbf{X}^{\mathrm{i}} = \boldsymbol{\xi}_m$.

Similarily, we can now multiply the observation matrices $\mathbf{B}$ from the left and the right onto $P\left(\mathbf{X}^{\mathrm{i}}, \mathbf{X}^{\mathrm{c}}\right)$ to compute the bilateral equivalent of **??**:

$$P\left(\mathbf{Z}^{\mathrm{c}} = \boldsymbol{\zeta}_n, \mathbf{Z}^{\mathrm{i}} = \boldsymbol{\zeta}_m\right) = \left[\mathbf{B}^{\top} \cdot \boldsymbol{\Lambda}_{\mathrm{c}}^{\top} \cdot \operatorname{diag} p(\mathbf{t}) \cdot \boldsymbol{\Lambda}_{\mathrm{i}} \cdot \mathbf{B}\right]_{n,m} \tag{1.39}$$

Formally, all necessary terms can now be computed so that both inference and the subsequent risk prediction can be performed. However, in the next section we will go into more detail regarding how this was implemented.

1.2.2 Parameter symmetries and mid-line extension

Although it has been omitted, eqs. (1.36) to (1.39) are still functions of the same parameters as in the unilateral model, but each side now has their own set $\boldsymbol{\theta}^{\mathrm{c}}$ and $\boldsymbol{\theta}^{\mathrm{i}}$ of spread probabilities that are used to parameterize the transition matrices $\mathbf{A}^{\mathrm{c}}$ and $\mathbf{A}^{\mathrm{i}}$ respectively.

In principle, the spread probabilities of the two sides are entirely indepent, and a lateralized primary tumor certainly spreads to a different extend to the ipsi- versus the contralateral side. But the spread probabilities among the LNLs should be equal when assuming that the lymphatic network in the head and neck region is symmetric. This means

$$\begin{aligned} \tilde{b}_v^{\mathrm{c}} &\neq \tilde{b}_v^{\mathrm{i}} \\ \tilde{t}_{rv}^{\mathrm{c}} &= \tilde{t}_{rv}^{\mathrm{i}} \end{aligned} \quad \forall\, v \leq V\,,\, r \in \mathrm{pa}(v) \tag{1.40}$$

Due to the reasoable assumption of a symmetric neck anatomy, we may avoid doubling the spread parameters when we model the bilateral lymphatic spread.

However, there are cases in which the primary tumor lies almost or exactly on the mid-sagittal plane of the patient. In such cases, we cannot reasonably distinguish between the ipsi- and contralateral side. Consequently, we must assume the base probability rates as well to be symmetric: $\tilde{b}_v^{\mathrm{c}} = \tilde{b}_v^{\mathrm{i}}$
This means there must be a continuous increase in the spread probabilities from the primary tumor to the contralateral LNLs if we were to move a patient's tumor from a clearly lateralized location closer and closer to that patient's mid-sagittal plane. Ideally, we would like to factor information about the tumor's "degree of asymmetry" into our model, e.g. by considering a normalized perpendicular distance from the mid-sagittal plane to the tumor's center of mass or by considering the tumor volume on either side of this plane. Data like this, however, is rarely available. What is frequently reported and also clinically considered as a risk factor for contralateral involvement is whether or not the tumor touches or extends over the mid-sagittal plane. With this binary variable (and the information on whether the tumor is central/symmetric w.r.t. to the sagittal symmetry plane) we can now distinguish three degrees of lateralizations:

1. $\not{\mathrm{s}}, \not{\mathrm{e}}$: The tumor does not cross or touch the mid-sagittal plane and is thus clearly lateralized. The base spread probabilities are $\left\{\tilde{b}_v^{\mathrm{i}}\right\}$ and $\left\{\tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}}\right\}$.

2. $\not{\mathrm{s}}, \mathrm{e}$: The tumor is lateralized, but crosses or touches the mid-sagittal plane. We will discuss how to define the spread probabilities to the contralateral side below.

3. s, e: The tumor is symmetric w.r.t. to the sagittal plane, thus $\tilde{b}_v^{\mathrm{c,s}} = \tilde{b}_v^{\mathrm{i}}$

Note that $\mathrm{s}\,(\not{\mathrm{s}})$ and $\mathrm{e}\,(\not{\mathrm{e}})$ denote the two binary variables *symmetric* (or *not symmetric*) and *extending* (or *not extending*) over the mid-sagittal plane.
We can infer that in case 2 the spread probabilities to the contralateral LNLs must be between the ones for the clearly lateralized (1) and the symmetric (3) case. Hence, we introduce a new "mixing" parameter α that defines the contralateral spread from tumor to the LNLs as a linear superposition between the two extremes:

$$\tilde{b}_v^{\mathrm{c,e}} = \alpha \cdot \tilde{b}_v^{\mathrm{i}} + (1-\alpha) \cdot \tilde{b}_v^{\mathrm{c},\not{\mathrm{e}}} \tag{1.41}$$

This new mixing parameter must be inferred from data just like the other spread probabilities and the parametrization of the time prior.
When using the learned parameters to predict the risk of a new patient g, the set of parameters for the risk computation $\hat{\boldsymbol{\theta}}_g$ is compiled from the total set of

inferred parameters $\hat{\boldsymbol{\theta}} = \left\{\tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c},\not{\mathrm{c}}}, \alpha, \tilde{t}_{rv}, p_T\right\}$, depending on the risk factors the patient presents with at the time of diagnosis. As always, for $\hat{\boldsymbol{\theta}}$ we have $v \leq V$, $r \in \mathrm{pa}(v)$ and the T-stage $T \in \{1, 2, 3, 4\}$. For example, if patient g has a T1 tumor that is clearly lateralized, their $\hat{\boldsymbol{\theta}}_g$ may be computed in the following way:

$$\hat{\boldsymbol{\theta}}_g = \left\{\tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c}} = \tilde{b}_v^{\mathrm{c},\not{\mathrm{c}}}, \tilde{t}_{rv}, p_1\right\} \tag{1.42}$$

while another patient m with a T3 tumor that clearly crosses the mid-sagittal plane would have the following set of parameters used for their risk prediction:

$$\hat{\boldsymbol{\theta}}_m = \left\{\tilde{b}_v^{\mathrm{i}}, \tilde{b}_v^{\mathrm{c}} = \alpha \cdot \tilde{b}_v^{\mathrm{i}} + (1 - \alpha) \cdot \tilde{b}_v^{\mathrm{c},\not{\mathrm{c}}}, \tilde{t}_{rv}, p_3\right\} \tag{1.43}$$

In the actual computational implementation of this model, we essentially compute three different matrices $\boldsymbol{\Lambda}$ which are functions of different parameters:

$$\begin{aligned} \boldsymbol{\Lambda}_{\mathrm{i}} &= \boldsymbol{\Lambda}\left(\tilde{b}_v^{\mathrm{i}}, \tilde{t}_{rv}\right) \\ \boldsymbol{\Lambda}_{\mathrm{c},\not{\mathrm{c}}} &= \boldsymbol{\Lambda}\left(\tilde{b}_v^{\mathrm{c},\not{\mathrm{c}}}, \tilde{t}_{rv}\right) \\ \boldsymbol{\Lambda}_{\mathrm{c,e}} &= \boldsymbol{\Lambda}\left(\alpha, \tilde{b}_v^{\mathrm{c},\not{\mathrm{c}}}, \tilde{b}_v^{\mathrm{i}}, \tilde{t}_{rv}\right) \end{aligned} \tag{1.44}$$

From those, the likelihoods of all patients in the training data can be computed when used with the respective p_T – that gives rise to the corresponding $\operatorname{diag} p(\mathbf{t})$ – as in eq. (1.39).