0.1 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 hidden Markov model (HMM) at time t to be

$$\mathbf{X}[t] = (X_v[t]) \tag{1}$$

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

$$P_{HMM}\left(\mathbf{X}[t+1] \mid \mathbf{X}[t]\right) \tag{2}$$

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

$$P_{HMM}\left(\mathbf{Z} \mid \mathbf{X}[t]\right) \tag{3}$$

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

$$\mathbf{A} = (A_{ij}) = \left(P_{HMM}\left(\mathbf{X}[t+1] = \boldsymbol{\xi}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j\right)\right) \tag{4}$$

and the *observation matrix*

$$\mathbf{B} = (B_{ij}) = \left(P_{HMM}\left(\mathbf{Z} = \boldsymbol{\zeta}_i \mid \mathbf{X}[t] = \boldsymbol{\xi}_j\right)\right) \tag{5}$$

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

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \tag{6}$$

to

$$\boldsymbol{\xi}_{16} = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \tag{7}$$

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. (7). The main task is to infer

the transition matrix **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 **A** in **??**. Afterwards, in **??**, I will tackle the aforementioned problems, followed up by explaining how we perform inference on this model (**??**), incorporate information about a patient's T-stage (**??**) and assess the risk of LNL involvement in a new patient (**??**). Lastly, we will introduce a way to incorporate incomplete observations in **??**.