0.1 Incorporation of T-stage

We have introduced the hidden Markov model (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 lymph node level (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:

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

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) by using the matrix

$$\boldsymbol{\Lambda} \coloneqq P\left(\mathbf{X} \mid \mathbf{t}\right) = \begin{pmatrix} \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^0 \\ \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^1 \\ \vdots \\ \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^{t_{\max}} \end{pmatrix} \tag{2}$$

were row number t corresponds to the vector $\boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^t$, 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) purely as a product of vectors and matrices:

$$\sum_{t=0}^{t_{\max}} p_T(t) \cdot \boldsymbol{\pi}^\top \cdot \left(\mathbf{A}\right)^t = p_T\left(\mathbf{t}\right) \cdot \boldsymbol{\Lambda} \tag{3}$$

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

0.1.1 * Interpretation of time-steps and time-priors

To add more interpretability to the time-prior $p(t)$ introduced in **??**, 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.

0.1.2 * 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

$$p(t) = \frac{1}{T} \qquad \text{for} \quad t \in \{1, 2, \dots, T\} \tag{4}$$

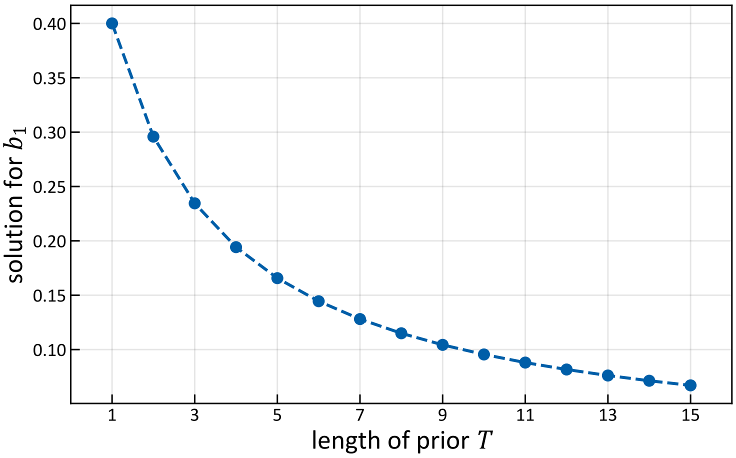

Figure 1: Solutions to eq. (6) for the base probability rate $\tilde{b}_1$ given a $p^\star$ of 0.4 and T increasing from 1 to 15.

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

$$\begin{aligned} p^\star &= \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1 & 0\end{pmatrix} \cdot \begin{bmatrix}(1-\tilde{b}_1) & \tilde{b}_1 \\ 0 & 1\end{bmatrix}^t \cdot \begin{pmatrix}0\\1\end{pmatrix} \\ &= \frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix}1 & 0\end{pmatrix} \cdot \begin{bmatrix}(1-\tilde{b}_1)^t & 1-(1-\tilde{b}_1)^t \\ 0 & 1\end{bmatrix} \cdot \begin{pmatrix}0\\1\end{pmatrix} \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[1-(1-\tilde{b}_1)^t\right] = 1 - \frac{1}{T}\sum_{t=1}^{T}(1-\tilde{b}_1)^t \end{aligned} \tag{5}$$

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

$$p^\star = 1 - \frac{\left(1-\tilde{b}_1\right)\left(1-(1-\tilde{b}_1)^T\right)}{\tilde{b}_1 T} \tag{6}$$

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. This confirms 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. 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

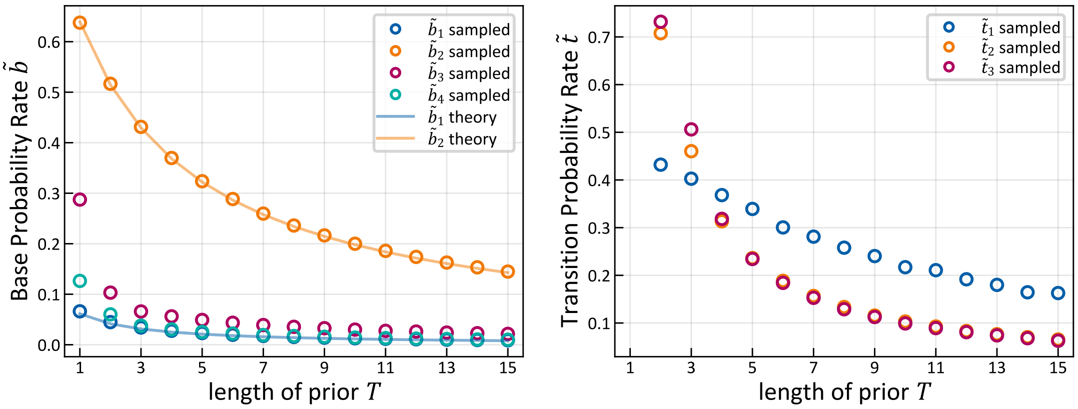

Figure 2: Decay of base probability rates as a function of the number of time-steps the time-prior has. Circles depict the results from learning the same dataset with different time-priors while solid lines show the analytical result starting with a $p^\star$ corresponding to the prevalence of involvement of LNL I and II respectively.

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 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. (6) 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. 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.

,,