Derivations

Computation of the posterior $p\left(r_t \mid x_{1:t+h}\right)$

Starting with $h=1$, the joint distribution $p(r_t, x_{1:(t+1)})$ is:

$$
\begin{aligned}
p(r_t, x_{1:(t+1)}) &= \sum_{r_{t+1}} p(r_t, r_{t+1}, x_{1:(t+1)}) && (1)\\
&= \sum_{r_{t+1}} p(r_t, x_{1:t}) p(r_{t+1}, x_{t+1} | r_t, x_{1:t}) && (2)\\
&= p(r_t, x_{1:t}) \sum_{r_{t+1}} p(x_{t+1} | r_t, r_{t+1}, x_{1:t}) p(r_{t+1} | r_t, \cancel{x_{1:t}}) && (3)\\
&= p(r_t, x_{1:t}) \sum_{r_{t+1}} p(x_{t+1} | r_t, x_{(t-r_{t+1}):t}) p(r_{t+1} | r_t) && (4)
\end{aligned}
$$

Explanations:

- $p(x_{t+1}|r_t, r_{t+1}, x_{1:t}) = p(x_{t+1}|r_t, x_{(t-r_{t+1}):t})$, because knowing r_{t+1} just selects the previous observations w.r.t. which we condition on.
- The transition probability $p(r_{t+1}|r_t, x_{1:t})$ from the run length at time t to the run length at time $t+1$ does not depend on the history of observations $x_{1:t}$. This is a model assumption in BOCD.

For $h=2$, the joint distribution $p(r_t, x_{1:t+2})$ is:

$$
\begin{aligned}
p(r_t, x_{1:t+2}) &= \sum_{r_{t+1}, r_{t+2}} p(r_t, r_{t+1}, r_{t+2}, x_{1:(t+1)}) && (5)\\
&= \sum_{r_{t+1}, r_{t+2}} p(r_t, r_{t+1}, x_{1:(t+1)}) p(r_{t+2}, x_{t+2} | r_t, r_{t+1}, x_{1:(t+1)}) && (6)\\
&= \sum_{r_{t+1}, r_{t+2}} p(r_t, r_{t+1}, x_{1:(t+1)}) p(x_{t+2} | \cancel{r_t}, r_{t+1}, r_{t+2}, x_{1:(t+1)}) p(r_{t+2} | \cancel{r_t}, r_{t+1}, \cancel{x_{1:(t+1)}}) && (7)\\
&= \sum_{r_{t+1}, r_{t+2}} p(r_t, r_{t+1}, x_{1:(t+1)}) p(x_{t+2} | r_{t+1}, x_{(t+1-r_{t+2}):(t+1)}) p(r_{t+2} | r_{t+1}) && (8)\\
&= p(r_t, x_{1:t}) \sum_{r_{t+1}, r_{t+2}} p(x_{t+1} | r_t, x_{(t-r_{t+1}):t}) p(r_{t+1} | r_t) p(x_{t+2} | r_{t+1}, x_{(t+1-r_{t+2}):(t+1)}) p(r_{t+2} | r_{t+1}) && (9)
\end{aligned}
$$

Explanations:

- The same reasoning as before applies to both, the predictive distribution as well as the transition probability.
- In the last step, the previous result was used.

By following this pattern, for arbitrary $h \geq 0$ the joint distribution $p\left(r_t, x_{1:t+h}\right)$ becomes:

$$
p(r_t, x_{1:t+h}) = p(r_t, x_{1:t}) \underbrace{\sum_{r_{t+1}, \ldots, r_{t+h}} \prod_{m=1}^{h} \left[p(x_{t+m} | r_{t+m-1}, x_{(t+m-r_{t+m}):(t+m-1)}) p(r_{t+m} | r_{t+m-1}) \right]}_{=:M} \qquad (10)
$$

The factor M is understood to be one, if $h=0$.

The joint distribution $p\left(r_t, x_{1:t+h}\right)$ can therefore be computed by the product of the joint distribution $p(r_t, x_{1:t})$, which BOCD computes anyway, and a factor M, which depends on the predictive and transition probabilities at future time steps. These objects, however, are components of the BOCD algorithm which are anyway computed in every time step and once computed can be re-used internally in the implementation.

The run length posterior can be computed from the joint distribution:

$$p(r_t|x_{1:t+h}) = \frac{p(r_t, x_{1:t+h})}{p(x_{1:t+h})} = \frac{p(r_t, x_{1:t+h})}{\sum_{r_t} p(r_t, x_{1:t+h})} \tag{11}$$

The factor M can be best understood as being the product of h matrices of the form M_{t+m}. Let's, e.g., assume that $t = 1$. The first product for $m = 1$, can be written as:

$$M_2 = \underbrace{\begin{bmatrix} p(x_2) & p(x_2|x_1) & 0 \\ p(x_2) & 0 & p(x_2|x_1) \end{bmatrix}}_{p(x_2|x_{(2-r_2):1})} \circ \underbrace{\begin{bmatrix} p(r_2 = 0|r_1 = 0) & p(r_2 = 1|r_1 = 0) & 0 \\ p(r_2 = 0|r_1 = 1) & 0 & p(r_2 = 2|r_1 = 1) \end{bmatrix}}_{p(r_2|r_1)}, \tag{12}$$

where $\circ$ denotes element-wise multiplication. Similarly, for $m = 2$ we have:

$$M_3 = \underbrace{\begin{bmatrix} p(x_3) & p(x_3|x_1) & 0 & 0 \\ p(x_3) & 0 & p(x_3|x_{1:2}) & 0 \\ p(x_3) & 0 & 0 & p(x_3|x_{1:2}) \end{bmatrix}}_{p(x_3|x_{(3-r_3):2})} \tag{13}$$

$$\circ \underbrace{\begin{bmatrix} p(r_3 = 0|r_2 = 0) & p(r_3 = 1|r_2 = 0) & 0 & 0 \\ p(r_3 = 0|r_2 = 1) & 0 & p(r_3 = 2|r_2 = 1) & 0 \\ p(r_3 = 0|r_2 = 2) & 0 & 0 & p(r_3 = 3|r_2 = 2) \end{bmatrix}}_{p(r_3|r_2)} \tag{14}$$

For all matrices M_{t+m}, r_{t+m-1} increases along the rows and r_{t+m} along the columns.

If h was just two, then $M = M_2 M_3$ in this example. The joint distribution can be written as a vector:

$$\begin{bmatrix} p(r_1 = 0, x_{1:3}) \\ p(r_1 = 1, x_{1:3}) \end{bmatrix} = \begin{bmatrix} p(r_1 = 0, x_1) \\ p(r_1 = 1, x_1) \end{bmatrix} \circ M \tag{15}$$

It is possible to use the particular structure of each matrix M_m (which has only elements in the first column and first upper diagonal) to implement this matrix multiplication efficiently (see `bocd._log_matmul_fast()`).