Viterbi Decoding

In this section we prove that for label sequence learning problems, decoding the top scoring output sequence

$$\hat{y} = \operatorname*{argmax}_{\bar{\mathbf{y}} \in \mathcal{Y}(x)} f(x, \bar{y})$$

can be performed by a Viterbi algorithm (Forney, 1973; Schwarz and Chow, 1990) in $\mathcal{O}(T|\Sigma|^2)$, where f is a generalized linear model

$$f(x, y) = \langle \mathbf{w}, \Phi(x, y) \rangle.$$

For notational convenience we will use the notation for $\Phi(x, y)$ introduced by Altun et al. (2003b). Given two labels $\sigma, \tau \in \Sigma$, we define

$$\begin{aligned} \phi^A_{\sigma,\tau}(y_{t-1}, y_t) &= [[y_{t-1} = \sigma \wedge y_t = \tau]] \\ \phi^B_{\sigma,j}(x_t, y_t) &= [[y_t = \sigma]]\psi_j(x_t) \end{aligned}$$

As described in Section 3.2.2, $\psi_j(x)$ extracts characteristics of x (e.g., feature $\psi_{234}(x) = 1$ if token x_t starts with a capital letter and 0 otherwise). We will refer to the vector $\psi(x) = (\ldots, \psi_j(x), \ldots)^\mathsf{T}$ and denote the inner product by means of $k(x, x') = \langle \psi(x), \psi(x') \rangle$.

Analogously to Equation 3.23, Altun et al. (2003b) define the joint feature representation $\Phi(x_i, y_i)$ of the i-th sequence as the sum of all feature vectors $\Phi(x_i, y_i|t) = (\ldots, \phi^A_{\sigma,\tau}(y_{i,t-1}, y_{i,t}), \ldots, \phi^B_{\sigma,j}(x_{i,t}, y_{i,t}), \ldots)^\mathsf{T}$ extracted at time t

$$\Phi(x_i, y_i) = \sum_{t=1}^{T_i} \Phi(x_i, y_i|t).$$

The feature map in Equation B.5 gives rise to the following inner product in input-output space that decomposes into a *label-label* and a *label-observation* part,

$$
\begin{aligned}
&\langle \Phi(\mathpzc{x}, \mathpzc{y}), \Phi(\mathpzc{x}', \mathpzc{y}') \rangle \\
&\qquad = \sum_{s,t} \sum_{\sigma,\tau} [[y_{s-1} = \sigma \wedge y_s = \tau]] \cdot [[y'_{t-1} = \sigma \wedge y'_t = \tau]] \\
&\qquad\quad + \sum_{s,t} \sum_{\sigma} [[y_s = \sigma]]\psi(x_s) \cdot [[y'_t = \sigma]]\psi(x'_t) \\
&\qquad = \sum_{s,t} [[y_{s-1} = y'_{t-1} \wedge y_s = y'_t]] + \sum_{s,t} [[y_s = y'_t]] k(x_s, x'_t).
\end{aligned}
$$

The Viterbi algorithm uses dynamic programming to maintain a trellis in which nodes correspond to hidden states versus times. Each entry stores the score of the most probable path leading to that node at a certain time, see Figure B.1. Once the computation reaches the end of the sequence, it backtracks the most likely path through the trellis and returns the highest scoring label sequence that generates the input sequence. The following Proposition B.1 shows that the argument of the maximum can be computed by a Viterbi algorithm.

Proposition *Given n input-output pairs of sequences of length T_i for $1 \leq i \leq n$, let Σ denote the output alphabet with $|\Sigma| < \infty$. Let f be defined as*

$$
f(\mathpzc{x}, \mathpzc{y}) = \sum_{i=1}^{n} \sum_{\substack{\bar{\mathpzc{y}} \in \mathcal{Y}(\mathpzc{x}_i) \\ \bar{\mathpzc{y}} \neq \mathpzc{y}_i}} \alpha_i(\bar{\mathpzc{y}}) \left(\langle \Phi(\mathpzc{x}_i, \mathpzc{y}_i), \Phi(\mathpzc{x}, \mathpzc{y}) \rangle - \langle \Phi(\mathpzc{x}_i, \bar{\mathpzc{y}}), \Phi(\mathpzc{x}, \mathpzc{y}) \rangle \right),
$$

with the joint feature map $\Phi(\mathpzc{x}, \mathpzc{y})$ as in Equation B.5. Then for all $\alpha_i(\bar{\mathpzc{y}}) \geq 0$ and any $\mathpzc{x} \in \mathcal{X}$,

$$
\hat{\mathpzc{y}} = \operatorname*{argmax}_{\bar{\mathpzc{y}} \in \mathcal{Y}(\mathpzc{x})} f(\mathpzc{x}, \bar{\mathpzc{y}})
$$

can be computed with a Viterbi algorithm in time $\mathcal{O}(T|\Sigma|^2)$.

Proof The model f has the form

$$
\begin{aligned}
f(\mathpzc{x}, \mathpzc{y}) =& \sum_{i=1}^{n} \sum_{\substack{\bar{\mathpzc{y}} \in \mathcal{Y}(\mathpzc{x}_i) \\ \bar{\mathpzc{y}} \neq \mathpzc{y}_i}} \alpha_i(\bar{\mathpzc{y}}) \left(\langle \Phi(\mathpzc{x}_i, \mathpzc{y}_i), \Phi(\mathpzc{x}, \mathpzc{y}) \rangle - \langle \Phi(\mathpzc{x}_i, \bar{\mathpzc{y}}), \Phi(\mathpzc{x}, \mathpzc{y}) \rangle \right) \\
=& \sum_{i=1}^{n} \sum_{\substack{\bar{\mathpzc{y}} \in \mathcal{Y}(\mathpzc{x}_i) \\ \bar{\mathpzc{y}} \neq \mathpzc{y}_i}} \alpha_i(\bar{\mathpzc{y}}) \left(\sum_{s,t} \left([[y_{i,s} = y_t]] - [[\bar{y}_s = y_t]] \right) k(x_{i,s}, x_t) \right. \\
&\left. + \sum_{s,t} [[y_{i,s-1} = y_{t-1} \wedge y_{i,s} = y_t]] - [[\bar{y}_{s-1} = y_{t-1} \wedge \bar{y}_s = y_t]] \right).
\end{aligned}
$$

We make the dependency on labels $\sigma, \tau \in \Sigma$ explicit by summing over all transitions and observation states

$$f(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\sigma,\tau\in\Sigma} \sum_{i} \sum_{\substack{\bar{\boldsymbol{y}}\in\mathcal{Y}(\boldsymbol{x}_i)\\ \bar{\boldsymbol{y}}\neq\boldsymbol{y}_i}} \alpha_i(\bar{\boldsymbol{y}}) \Big(\sum_{s,t} \left([[y_{i,s} = \sigma]] - [[\bar{y}_s = \sigma]]\right) [[y_t = \tau]] k(x_{i,s}, x_t)$$

$$+ \sum_{s,t} \Big([[y_{i,s-1} = \sigma \wedge y_{i,s} = \tau]] - [[\bar{y}_{s-1} = \sigma \wedge \bar{y}_s = \tau]] \Big) [[y_{t-1} = \sigma \wedge y_t = \tau]] \Big) .$$

The transition scores from label σ to label τ are now given by

$$a(\sigma, \tau) = \sum_{i=1}^{n} \sum_{\substack{\bar{\boldsymbol{y}}\in\mathcal{Y}(\boldsymbol{x}_i)\\ \bar{\boldsymbol{y}}\neq\boldsymbol{y}_i}} \alpha_i(\bar{\boldsymbol{y}}) \Big(\sum_{t=1}^{T_i} [[y_{i,t-1} = \sigma \wedge y_{i,t} = \tau]] - [[\bar{y}_{t-1} = \sigma \wedge \bar{y}_t = \tau]] \Big)$$

and observation scores for label $y_s = \sigma$ and observation x_s by

$$b(\sigma, x) = \sum_{i=1}^{n} \sum_{t=1}^{T_i} \sum_{\substack{\bar{\boldsymbol{y}}\in\mathcal{Y}(\boldsymbol{x}_i)\\ \bar{\boldsymbol{y}}\neq\boldsymbol{y}_i}} \alpha_i(\bar{\boldsymbol{y}}) \left([[y_{i,t} = \sigma]] - [[\bar{y}_t = \sigma]]\right) k(x_{i,t}, x).$$

The hypothesis $f(\boldsymbol{x}, \boldsymbol{y})$ can be rewritten in terms of transition scores $a(\sigma, \tau)$ and observation scores $b(\sigma, x)$

$$f(\boldsymbol{x}, \boldsymbol{y}) = \underbrace{\sum_{\sigma,\tau\in\Sigma} a(\sigma, \tau) \sum_{s=1}^{T} [[y_{s-1} = \sigma \wedge y_s = \tau]]}_{=: f_a(\boldsymbol{x}, \boldsymbol{y})} + \underbrace{\sum_{s=1}^{T} \sum_{\sigma\in\Sigma} [[y_s = \sigma]] b(\sigma, x_s)}_{=: f_b(\boldsymbol{x}, \boldsymbol{y})} .$$

where f_a weights the occurences of neighboring labels in $\boldsymbol{y}$ by corresponding scores of the model and f_b determines how well observations x_s fit to their labels y_s given the model. To decode the top scoring sequence we define

$$\delta_t(\sigma) = \max_{y_1,\ldots,y_{t-1}} f(\boldsymbol{x}, y_1, \ldots, y_{t-1}, y_t = \sigma),$$

that is, $\delta_t(\sigma)$ denotes the top scoring partial sequence up to position $t-1$ where $y_t = \sigma$. We first show by induction that

$$\delta_{t+1}(\sigma) = \max_{\tau\in\Sigma} \left[\delta_t(\tau) + a(\tau, \sigma)\right] + b(\sigma, x_{t+1})$$

holds. The initialization is simply given by

$$\begin{aligned} \delta_0(\sigma) &= 0, \quad \forall \sigma \in \Sigma \\ \delta_1(\sigma) &= \max_{\tau\in\Sigma} \left[\delta_t(\tau) + a(\tau, \sigma)\right] + b(\sigma, x_{t+1}) \\ &= a(\epsilon, \sigma) + b(\sigma, x_1). \end{aligned}$$

The recursion step is given for $2 \leq t \leq T$ by

$$\begin{aligned}
\delta_t(\sigma) = &\max_{y_1,\dots,y_{t-1}} f(\chi, y_1, \dots, y_{t-1}, y_t = \sigma) \\
= &\max_{y_1,\dots,y_{t-1}} \sum_{\tau,\bar{\tau}\in\mathcal{Y}} a(\tau,\bar{\tau}) \sum_{s=2}^{t-1} [[y_{s-1} = \tau \wedge y_s = \bar{\tau}]] \\
&+ \sum_{\tau\in\Sigma} a(\tau,\sigma)[[y_{t-1} = \tau \wedge y_t = \sigma]] \\
&+ \sum_{s=1}^{t-1} \sum_{\tau\in\Sigma} [[y_s = \tau]] b(\tau, x_s) + [[y_t = \sigma]] b(\sigma, x_t) \\
= &\max_{\sigma^\star} \max_{y_1,\dots,y_{t-2}} \sum_{\tau,\bar{\tau}\in\mathcal{Y}} a(\tau,\bar{\tau}) \sum_{s=2}^{t-2} [[y_{s-1} = \tau \wedge y_s = \bar{\tau}]] \\
&+ \sum_{\tau\in\Sigma} a(\tau,\sigma^\star)[[y_{t-2} = \tau \wedge y_{t-1} = \sigma^\star]] \\
&+ a(\sigma^\star,\sigma)[[y^{t-1} = \sigma^\star \wedge y^t = \sigma]] \\
&+ \sum_{s=1}^{t-2} \sum_{\tau\in\Sigma} [[y_s = \tau]] b(\tau, x_s) + b(\sigma^\star, x_{t-1}) + b(\sigma, x_t) \\
= &\max_{\sigma^\star} \left[\max_{y_1,\dots,y_{t-2}} f(\chi, y_1, \dots, y_{t-2}, y_{t-1} = \sigma^\star) + a(\sigma^\star,\sigma) \right] + b(\sigma, x_t) \\
= &\max_{\sigma^\star} \left[\delta_{t-1}(\sigma^\star) + a(\sigma^\star,\sigma) \right] + b(\sigma, x_t).
\end{aligned}$$

Thus, the top scoring sequence has the score

$$\max f(\chi, y) = \max_{\sigma\in\Sigma} \delta_T(\sigma).$$

We only sketch the extension to the argument of the maximum since it is analoguous to the regular Viterbi algorithm. We introduce path variables $\varphi_t(\sigma)$ that are initialized by $\varphi_1(\sigma) = \epsilon$ for all $\sigma \in \Sigma$. The sequence $\varphi_t(\sigma)$ is then defined recursively for $2 \leq t \leq T$ by

$$\varphi_t(\sigma) = \operatorname*{argmax}_{\sigma^\star\in\Sigma} \left[\delta_{t-1}(\sigma^\star) + a(\sigma^\star,\sigma) \right].$$

Once the $\delta_t(\sigma)$ of Proposition are fixed, the optimal label sequence can be found by backtracking

$$\begin{aligned}
y_T^\star &= \operatorname*{argmax}_{\sigma\in\Sigma} \delta_T(\sigma) \\
y_t^\star &= \varphi_{t+1}(y_{t+1}^\star) \quad \text{for} \quad t = T-1, \dots, 1.
\end{aligned}$$

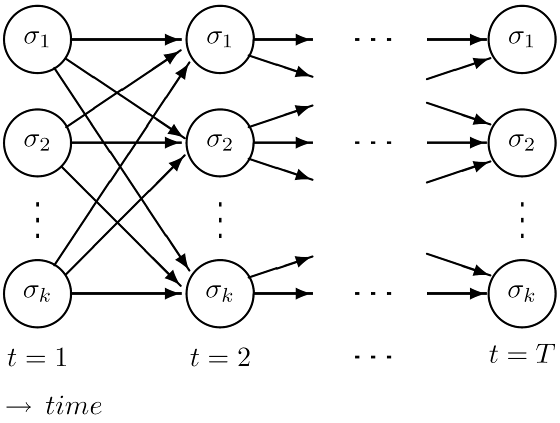

Figure B.1: *Visualization of a trellis over the alphabet* $\Sigma = \{\sigma_1, \ldots, \sigma_k\}$.

Given the transition matrix $[\mathbf{A}]_{\sigma,\tau} = a(\sigma, \tau)$ and the observation matrix $[\mathbf{B}_\chi]_{\sigma,t} = b(\sigma, x_t)$ for input χ, the computation of δ and φ for a fixed t and $\sigma \in \Sigma$ involves visiting $|\Sigma|$ predecessors; thus, for a sequence of length T the time needed is in $\mathcal{O}(T|\Sigma|^2)$. This concludes the proof. □