1 The algorithm for variable-order linear-chain CRFs

In this section, we introduce three algorithms that constitute the main part of this paper. One is "Sum-Difference" algorithm, which is used to calculate the expectations of the feature functions to estimate parameters. This algorithm uses "sum" scores and "difference" ones when calculating forward and backward scores. This way, we can apply dynamic programming. The other two are decoding algorithm used to infer the best labels for an sequence of observations. One process the sequences from head to tail and the other from tail to head.

1.1 Notations

We introduce the notations used in this paper.

$|\cdot|$ denotes the length of a sequence. $\mathbf{x}$ and $\mathbf{y}$ represent respectively a sequence of observations and labels of length T and $T+2$. $\mathbf{y}$ is a sequence which has positions 0 and $T+1$, which represent respectively the beginning and ending of a sequence. They have the labels l_{BOS} and l_{EOS}. Suppose we have a set of labels $\mathcal{Y} = \{l_1, \cdots, l_L\}$.

We define the set of labels $\mathcal{Y}_t$ at a particular position t as follows.

$$\mathcal{Y}_t = \begin{cases} \{l_{BOS}\} & \text{if } t = 0 \\ \{l_{EOS}\} & \text{if } t = T+1 \\ \mathcal{Y} & \text{otherwise} \end{cases} \tag{1}$$

We denote $\mathbf{z} \in \mathcal{Y}_{n:m}$ when a label sequence $\mathbf{z}$ satisfies $|\mathbf{z}| = m - n + 1, \forall (t : 1 \leq t \leq m - n + 1)\mathbf{z}_n \in \mathcal{Y}_{n+t-1}$.

For a label sequence $\mathbf{z}$, we define $\mathbf{z}_{i:j} \stackrel{\text{def}}{=} \mathbf{z}_{i \cdots j}$. When $j < i$, let $\mathbf{z}_{i:j} = \epsilon$, where ϵ represents an empty string.

When two label sequences $\mathbf{z}^1$ and $\mathbf{z}^2$ are concatenated into $\mathbf{z}^3$, we denote it as $\mathbf{z}^3 = \mathbf{z}^1 + \mathbf{z}^2$.

When a label sequence $\mathbf{z}^1$ is a suffix of $\mathbf{z}^2$, we denote it as $\mathbf{z}^1 \leq_s \mathbf{z}^2$, otherwise $\mathbf{z}^1 \not\leq_s \mathbf{z}^2$. For any label sequence $\mathbf{z}$, $\epsilon \leq_s \mathbf{z}$. When $\mathbf{z}^1$ is a *proper* suffix of $\mathbf{z}^2$, we denote it as $\mathbf{z}^1 <_s \mathbf{z}^2$. The following two equations can be trivially derived:

$$\mathbf{z}^1 <_s \mathbf{z}^2, \mathbf{z}^1 \neq \epsilon \Rightarrow \mathbf{z}^1_{1:|\mathbf{z}^1|-1} <_s \mathbf{z}^2_{1:|\mathbf{z}^2|-1} \tag{2}$$

$$\mathbf{z}^1 <_s \mathbf{z}^3, \mathbf{z}^2 <_s \mathbf{z}^3 \Rightarrow \mathbf{z}^1 <_s \mathbf{z}^2 \text{ or } \mathbf{z}^2 \leq_s \mathbf{z}^1 \tag{3}$$

We define $s(\mathbf{z}^1, \mathcal{S})$ as follows:

$$\begin{aligned} &s(\mathbf{z}^1, \mathcal{S}) = \mathbf{z}^2 \text{ if and only if } \mathbf{z}^2 \in \mathcal{S} \text{ and } \mathbf{z}^2 <_s \mathbf{z}^1 \\ &\text{and } \forall (\mathbf{z} \in \mathcal{S})\mathbf{z} <_s \mathbf{z}^1 \Rightarrow \mathbf{z} \leq_s \mathbf{z}^2 \end{aligned} \tag{4}$$

and call it *longest proper suffix of* $\mathbf{z}^1$ *with respect to* $\mathcal{S}$. $\mathbf{z}^1$ may or may not be an element of $\mathcal{S}$.

Similarly, we define $S(\mathbf{z}^1, \mathcal{S})$ as follows:

$$\begin{aligned} &S(\mathbf{z}^1, \mathcal{S}) = \mathbf{z}^2 \text{ if and only if } \mathbf{z}^2 \in \mathcal{S} \text{ and } \mathbf{z}^2 \leq_s \mathbf{z}^1 \\ &\text{and } \forall(\mathbf{z} \in \mathcal{S})\mathbf{z} \leq_s \mathbf{z}^1 \Rightarrow \mathbf{z} \leq_s \mathbf{z}^2 \end{aligned} \tag{5}$$

and call it the *longest suffix of* $\mathbf{z}^1$ *with respect to* $\mathcal{S}$. $S(\mathbf{z}, \mathcal{S}) = \mathbf{z}$ if and only if $\mathbf{z}$ is an element of $\mathcal{S}$.

For any set $\mathcal{S}$, let $s(\epsilon, \mathcal{S}) = \perp$. Let $\perp$ be an imaginary label sequence that never equals to any other label sequences ($\forall(\mathbf{z})\mathbf{z} \neq \perp$) and let $\perp^n$ be such a label sequence that has the length n.

Let the feature functions be $f_1, \cdots, f_m$. Each feature function is a binary function that takes three arguments (obervations, label sequence, position) and can be expressed as a product of two binary functions $b_i(\mathbf{x}, t)$ (we call it *observation function*) and $L_i(\mathbf{y}, t)$ (we call it *label function*).

$$f_i(\mathbf{x}, \mathbf{y}, t) = b_i(\mathbf{x}, t)L_i(\mathbf{y}, t) \tag{6}$$

Each L_i is associated with a label sequence $\mathbf{z}^i$ and is defined as follows:

$$L_i(\mathbf{y}, t) = \begin{cases} 1 & \text{if } \mathbf{y}_{t-|\mathbf{z}^i|+1:t} = \mathbf{z}^i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We also define $L'_i(\mathbf{y}^s, t)$ and $L''(\mathbf{z})$ as alternate representations of the label function, where $\mathbf{y}^s$ is a suffix of a label sequence $\mathbf{y}$ and $\mathbf{z}$ is an arbitrary label sequence. We use zero-based indexes for $\mathbf{y}^s$.

$$L'_i(\mathbf{y}_s, t) = \begin{cases} 1 & \text{if } t + |\mathbf{y}^s| - |\mathbf{y}| - |\mathbf{z}^i| + 1 \geq 0 \text{ and} \\ & \mathbf{y}_{t+|\mathbf{y}^s|-|\mathbf{y}|-|\mathbf{z}^i|+1:t+|\mathbf{y}^s|-|\mathbf{y}|} = \mathbf{z}^i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$L''_i(\mathbf{z}) = \begin{cases} 1 & \text{if } |\mathbf{z}| \geq |\mathbf{z}^i| \text{ and} \\ & \mathbf{z}_{|\mathbf{z}|-|\mathbf{z}^i|+1:|\mathbf{z}|} = \mathbf{z}^i \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We define f'_i and f''_i as products of b_i and L'_i, b_i and L''_i respectively.

$$f'_i(\mathbf{x}, \mathbf{y}^s, t) = b_i(\mathbf{x}, t)L'_i(\mathbf{y}^s, t) \tag{10}$$

$$f''_i(\mathbf{x}, \mathbf{z}, t) = b_i(\mathbf{x}, t)L''_i(\mathbf{z}) \tag{11}$$

$$\tag{12}$$

When we define f_i, f'_i and f''_i as above, the following can be said:

$$f_i(\mathbf{x}, \mathbf{y}, t) = f''_i(\mathbf{x}, \mathbf{y}_{t-|\mathbf{z}^i|+1:t}, t) \tag{13}$$

$$f'_i(\mathbf{x}, \mathbf{y}^s, t) = f''_i(\mathbf{x}, \mathbf{y}^s_{t+|\mathbf{y}^s|-|\mathbf{y}|-|\mathbf{z}^i|+1:t+|\mathbf{y}^s|-|\mathbf{y}|}, t) \tag{14}$$

1.2 Training of High-order CRF

Here is the expected sum of f_i based on current parameters:

$$E[f_i] = \sum_{(\mathbf{x},\tilde{\mathbf{y}})\in\mathcal{T}} \sum_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}) \sum_{t=|\mathbf{z}^i|-1}^{T+1} f_i(\mathbf{x}, \mathbf{y}, t) \tag{15}$$

We will need expectations of f_i later when we update the parameters. Here we introduce a way to calculate them.

1.2.1 Paths

We define $\mathcal{P}_t$, the *path set* at the position t, as follows:

$$\mathcal{P}_t = \mathcal{Y}_t \cup \{\epsilon\} \cup \bigcup_{t'=t}^{T} \{\mathbf{z}^k_{1:|\mathbf{z}^k|-(t'-t)} \mid b_k(\mathbf{x}, t) = 1, |\mathbf{z}^k| \geq t' - t\} \tag{16}$$

We call each element of $\mathcal{P}_t$ a *path* at the position t. A path is either one of the label sequences that are associated with the feature functions or a prefix of another path at the position $t' > t$, with $t' - t$ label(s) removed. When a label sequence $\mathbf{y}$ satisfies $\mathbf{y}_{t-|\mathbf{z}^p|+1:t} = \mathbf{z}^p$ for a path $\mathbf{z}^p$ at a position t, we say that $\mathbf{y}$ *contains* $\mathbf{z}^p$ at the position t.

When we define the path sets as in (16), the following can be said:

$$\mathbf{z} \in \mathcal{P}_t, \mathbf{z} \neq \epsilon, t > 0 \Rightarrow \mathbf{z}_{1:|\mathbf{z}|-1} \in \mathcal{P}_{t-1} \tag{17}$$

We can also say the following:

$$f''_i(\mathbf{x}, \mathbf{z}, t) = f''_i(\mathbf{x}, S(\mathbf{z}, \mathcal{P}_t), t) \tag{18}$$

Assume this is not true and that there is a feature function $f''_i(\mathbf{x}, S(\mathbf{z}, \mathcal{P}_t), t) = 0, f''_i(\mathbf{x}, \mathbf{z}', t) = 1$ for a $S(\mathbf{z}, \mathcal{P}_t) <_s \mathbf{z}' \leq_s \mathbf{z}$. This means that $\mathbf{z}' \in \mathcal{P}_t$ and this contradicts with the definition of longest suffix given in (5).

1.2.2 Scores of Paths

Considering positions 0 and $T + 1$ as the beginning and ending of a label sequence, conditional probability distributions of a linear-chan CRF is defined as $P(\mathbf{y}|\mathbf{x}) = Z_{\mathbf{x}}(\mathbf{y})/Z_{\mathbf{x}}$, where $Z_{\mathbf{x}}(\mathbf{y}) = \exp(\sum_{i=1}^{m} \sum_{t=|\mathbf{z}^i|-1}^{T+1} f_i(\mathbf{x}, \mathbf{y}, t)\lambda_i)$ and $Z_{\mathbf{x}} = \sum_{\mathbf{y}} Z_{\mathbf{x}}(\mathbf{y})$. We call $Z_{\mathbf{x}}(\mathbf{y})$ the *score* of $\mathbf{y}$. Let $\sigma(\mathbf{z}^p, t)$ denote the sum of scores of label sequences $\mathbf{y}$ that contain $\mathbf{z}^p$ at a position t. It can be expressed in the following way:

$$\sigma(\mathbf{z}^p, t) \stackrel{\text{def}}{=} \sum_{\mathbf{y}:\mathbf{y}\in\mathcal{Y}_{0:T}, \mathbf{y}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}} \exp\left(\sum_{i=1}^{m} \sum_{t'=1}^{T+1} f_i(\mathbf{x}, \mathbf{y}, t')\lambda_i\right) \tag{19}$$

Unlike first order linear-chain CRFs, we cannot directly decompose this using forward and backward scores. Since the fact that a label sequence $\mathbf{y}$ contains

a path $\mathbf{z}^p$ at a position t does not necessarily mean it is the longest path that $\mathbf{y}$ contains at that position, backward scores cannot be determined. So we introduce supplementary variables $\theta(\mathbf{z}^p, t)$ that denote the probability that a label sequence $\mathbf{y}$ contains a path $\mathbf{z}^p \in \mathcal{P}_t$ at a position t and does not contain any longer paths at the same position, i.e. does not contain any paths that have $\mathbf{z}^p$ as their longest proper suffix.

$$\begin{aligned}\theta(\mathbf{z}^p, t) &\stackrel{\text{def}}{=} \sum_{\mathbf{y}:\mathbf{y}\in\mathcal{Y}_{0:T},\mathbf{y}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p,\forall(\mathbf{z}'\in\mathcal{P}_t,\mathbf{z}^p\leq_s\mathbf{z}')\mathbf{y}_{t-|\mathbf{z}'|+1:t}\neq\mathbf{z}'} \\ &\qquad \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{T+1} f_i(\mathbf{x},\mathbf{y},t')\lambda_i\right) \\ &= \sum_{\mathbf{y}:\mathbf{y}\in\mathcal{Y}_{0:T},\mathbf{y}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p,\forall(\mathbf{z}'\in\mathcal{P}_t,s(\mathbf{z}',\mathcal{P}_t)=\mathbf{z}^p)\mathbf{y}_{t-|\mathbf{z}'|+1:t}\neq\mathbf{z}'} \\ &\qquad \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{T+1} f_i(\mathbf{x},\mathbf{y},t')\lambda_i\right) \end{aligned} \tag{20}$$

When we define $\theta(\mathbf{z}^p, t)$ like this, we can express $\sigma(\mathbf{z}^p, t)$ as a sum of θ.

$$\sigma(\mathbf{z}^p, t) \stackrel{\text{def}}{=} \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{P}_t,\mathbf{z}^p\leq_s\mathbf{z}} \theta(\mathbf{z}, t) \tag{21}$$

By the definition of path given in (16), if a label sequence $\mathbf{y}$ contains $\mathbf{z}^p$ at a position t and does not contain any longer paths, we can know that there are no feature functions that are affected by the labels before $t - |\mathbf{z}| + 1$. So we can decompose $\theta(\mathbf{z}^p, t)$ in three parts, the *forward* part ($1 \leq t' \leq t$) and *backward* part ($t + 1 \leq t' \leq T + 1$). The forward part can also be represented in the form of substraction. Let $\alpha(\mathbf{z}^p, t)$ and $\beta(\mathbf{z}^p, t)$ denote the forward and backward parts respectively.

$$\begin{aligned}\theta(\mathbf{z}^p, t) &= \left(\sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{0:t},\mathbf{z}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p,\forall(\mathbf{z}'\in\mathcal{P}_t,s(\mathbf{z}',\mathcal{P}_t)=\mathbf{z}^p)\mathbf{z}_{t-|\mathbf{z}'|+1:t}\neq\mathbf{z}'} \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{t} f_i(\mathbf{x},\mathbf{z},t')\lambda_i\right)\right) \\ &\qquad \left(\sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{t+1:T+1}} \exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f'_i(\mathbf{x},\mathbf{z}^p+\mathbf{z},t')\lambda_i\right)\right) \\ &\stackrel{\text{def}}{=} \alpha(\mathbf{z}^p, t)\beta(\mathbf{z}^p, t) \end{aligned} \tag{22}$$

1.2.3 Path weights

For a path $\mathbf{z}^p \in \mathcal{P}_t$ and a position t, we define $w(\mathbf{z}^p, t)$ as follows:

$$w(\mathbf{z}^p, t) \stackrel{\text{def}}{=} \sum_{i:\mathbf{z}^i=\mathbf{z}^p,b_i(\mathbf{x},t)=1} \lambda_i \tag{23}$$

This is the sum of the weights of the feature functions associated with the path $\mathbf{z}^p$ at the position t.

We also define $W(\mathbf{z}^p, t)$ for a path $\mathbf{z}^p \in \mathcal{P}_t$ and a position t as follows:

$$W(\mathbf{z}^p, t) \stackrel{\text{def}}{=} \sum_{i:\mathbf{z}^i \leq_s \mathbf{z}^p, b_i(\mathbf{x},t)=1} \lambda_i = \sum_{i=1}^{m} f_i''(\mathbf{x}, \mathbf{z}^p, t)\lambda_i \tag{24}$$

This is the sum of the weigths of the feature functions associated with the path $\mathbf{z}^p$ or any of its suffixes at the position t. From (24) and (23), we can express W using w as follows:

$$W(\mathbf{z}^p, t) = \sum_{i:\mathbf{z}^i \in \mathcal{P}_t, \mathbf{z}^i \leq_s \mathbf{z}^p} w(\mathbf{z}^i, t) \tag{25}$$

When we define W and w as above, we can say that for any $\mathbf{z}^p \neq \epsilon$, $W(\mathbf{z}^p, t) = W(s(\mathbf{z}, \mathcal{P}_t), t) + w(\mathbf{z}^p, t)$.

1.2.4 The Forward Variables

As shown in (22), $\alpha(\mathbf{z}^p, t)$ is defined for a path $\mathbf{z}^p$ and a position $t \geq 1$ as follows:

$$\begin{aligned} \alpha(\mathbf{z}^p, t) \quad = \quad & \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{0:t}, \mathbf{z}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p, \forall(\mathbf{z}'\in\mathcal{P}_t, s(\mathbf{z}',\mathcal{P}_t)=\mathbf{z}^p)\mathbf{z}_{t-|\mathbf{z}'|+1:t}\neq\mathbf{z}'} \\ & \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{t} f_i(\mathbf{x}, \mathbf{z}, t')\lambda_i\right) \end{aligned} \tag{26}$$

Let $\alpha(\epsilon, t) = 0$ for any t. In order to calculate these using dynamic programming, we also define $\gamma(\mathbf{z}^p, t)$ for a path $\mathbf{z}^p \in \mathcal{P}_t$ and a position t as follows:

$$\gamma(\mathbf{z}^p, t) \stackrel{\text{def}}{=} \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{0:t}, \mathbf{z}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p} \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{t} f_i(\mathbf{x}, \mathbf{z}, t')\lambda_i\right) \tag{27}$$

When we define α and γ like this, we can say the following using (24):

$$\begin{aligned} \alpha(\mathbf{z}^p, t) \quad = \quad & \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{0:t}, \mathbf{z}_{t-|\mathbf{z}^p|+1:t}=\mathbf{z}^p, \forall(\mathbf{z}'\in\mathcal{P}_t, s(\mathbf{z}',\mathcal{P}_t)=\mathbf{z}^p)\mathbf{z}_{t-|\mathbf{z}'|+1:t}\neq\mathbf{z}'} \\ & \exp\left(\sum_{i=1}^{m}\sum_{t'=1}^{t-1} f_i(\mathbf{x}, \mathbf{z}_{1:|\mathbf{z}|-1,t'})\lambda_i\right) \exp\left(\sum_{i=1}^{m} f_i''(\mathbf{x}, \mathbf{z}^p, t)\lambda_i\right) \\ = \quad & \left(\gamma(\mathbf{z}^p_{1:|\mathbf{z}^p|-1}, t-1) - \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{P}_t, s(\mathbf{z},\mathcal{P}_t)=\mathbf{z}^p} \big(\gamma(\mathbf{z}_{1:|\mathbf{z}|-1}, t-1)\big)\right) \\ & \exp(W(\mathbf{z}^p, t)) \end{aligned} \tag{28}$$

$$\gamma(\mathbf{z}^p, t) \quad = \quad \sum_{\mathbf{z}\in\mathcal{P}_t, \mathbf{z}^p \leq_s \mathbf{z}} \alpha(\mathbf{z}, t) \tag{29}$$

1.2.5 The Backward Variables

Here we rewrite the definition of $\beta(\mathbf{z}^p, t)$ given in (22) using f'_i defined in (11):

$$\begin{aligned}\beta(\mathbf{z}^p, t) &\stackrel{\text{def}}{=} \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{t+1:T+1}} \exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f_i(\mathbf{x}, \perp^{t-|\mathbf{z}^p|+1} + \mathbf{z}^p + \mathbf{z}, t')\lambda_i\right) \\ &= \sum_{\mathbf{z}:\mathbf{z}\in\mathcal{Y}_{t+1:T+1}} \exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f'_i(\mathbf{x}, \mathbf{z}^p + \mathbf{z}, t')\lambda_i\right) \qquad (30)\end{aligned}$$

In order to calculate these using dynamic programming, we also define $\delta(\mathbf{z}^p, t)$ for a path $\mathbf{z}^p \in \mathcal{P}_t$ and a position t as follows:

$$\delta(\mathbf{z}^p, t) = \begin{cases} \sum_{\mathbf{z}\in\mathcal{Y}_{t+1:T+1}} \exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f'_i(\mathbf{x}, \mathbf{z}, t')\lambda_i\right) \text{ if } \mathbf{z}^p = \epsilon \\ \sum_{\mathbf{z}\in\mathcal{Y}_{t+1:T+1}} \left(\exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f'_i(\mathbf{x}, \mathbf{z}^p + \mathbf{z}, t')\lambda_i\right) - \right. \\ \left. \exp\left(\sum_{i=1}^{m}\sum_{t'=t+1}^{T+1} f'_i(\mathbf{x}, s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}, t')\lambda_i\right)\right) \text{ otherwise} \end{cases} \qquad (31)$$

When we define β and δ like this, we can say the following:

$$\delta(\mathbf{z}^p, t) = \begin{cases} \sum_{\mathbf{z}\in\mathcal{P}_{t+1}, \mathbf{z}_{1:|\mathbf{z}|-1}=\mathbf{z}^p} (\beta(\mathbf{z}, t+1)\exp(W(\mathbf{z}, t+1)) \text{ if } \mathbf{z}^p = \epsilon \\ \sum_{\mathbf{z}\in\mathcal{P}_{t+1}, \mathbf{z}_{1:|\mathbf{z}|-1}=\mathbf{z}^p} (\beta(\mathbf{z}, t+1)\exp(W(\mathbf{z}, t+1)) - \\ \beta(s(\mathbf{z}, \mathcal{P}_{t+1}), t+1)a\exp(W(s(\mathbf{z}, \mathcal{P}_{t+1}), t+1))) \text{ otherwise} \end{cases} \qquad (32)$$

$$\beta(\mathbf{z}^p, t) = \sum_{\mathbf{z}\in\mathcal{P}_t, \mathbf{z}\leq_s \mathbf{z}^p} \delta(\mathbf{z}, t) \qquad (33)$$

Here we present the derivation of (32) for $\mathbf{z}^p \neq \epsilon$, which is not straightforward. First, we rewrite the right hand side of (31) as follows:

$$\sum_{\mathbf{z}\in\mathcal{Y}_{t+2:T+1}} \sum_{\mathbf{z}'\in\mathcal{Y}_{t+1}} \left(\exp\left(\sum_{i=1}^{m}\sum_{t'=t+2}^{T+1} f'_i(\mathbf{x}, \mathbf{z}^p + \mathbf{z}' + \mathbf{z}, t')\lambda_i\right)\exp\left(\sum_{i=1}^{m} f''_i(\mathbf{x}, \mathbf{z}^p + \mathbf{z}', t+1)\lambda_i\right) - \right.$$
$$\left.\exp\left(\sum_{i=1}^{m}\sum_{t'=t+2}^{T+1} f'_i(\mathbf{x}, s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}' + \mathbf{z}, t')\lambda_i\right)\exp\left(\sum_{i=1}^{m} f''_i(\mathbf{x}, s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}' + \mathbf{z}, t)\lambda_i\right)\right) \qquad (34)$$

Using the equation (18), we can rewrite (34) as follows:

$$\sum_{\mathbf{z}\in\mathcal{Y}_{t+2:T+1}} \sum_{\mathbf{z}'\in\mathcal{Y}_{t+1:t+1}} \left(\exp\left(\sum_{t'=t+2}^{T+1}\sum_{i=1}^{m} f'_i(\mathbf{x}, S(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) + \mathbf{z}, t')\lambda_i\right)\exp\left(\sum_{i=1}^{m} f''_i(\mathbf{x}, S(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}), t+1)\lambda_i\right) - \right.$$
$$\left.\exp\left(\sum_{t'=t+2}^{T+1}\sum_{i=1}^{m} f'_i(\mathbf{x}, S(s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}', \mathcal{P}_{t+1}) + \mathbf{z}, t')\lambda_i\right)\exp\left(\sum_{i=1}^{m} f''_i(\mathbf{x}, S(s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}', \mathcal{P}_{t+1}), t)\lambda_i\right)\right) \qquad (35)$$

Here we can say the following:

$$s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) \leq_s s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}' \tag{36}$$

Assume (36) is not true. By the definition of longest proper suffix in (4), $s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) <_s \mathbf{z}^p + \mathbf{z}'$ and $s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}' <_s \mathbf{z}^p + \mathbf{z}'$. From (3) and the assumption that (36) is not true, $s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}' <_s s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1})$. Let $\mathbf{z}'' = s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1})$. Since $\mathbf{z}'' \in \mathcal{P}_{t+1}$ and $\mathbf{z}'' \neq \epsilon$, we can derive $\mathbf{z}''_{1:|\mathbf{z}''|-1} \in \mathcal{P}_t$ from (17) and $s(\mathbf{z}^p, \mathcal{P}_t) <_s \mathbf{z}''_{1:|\mathbf{z}''|-1} <_s \mathbf{z}^p$ from (2). This contradicts the definition of longest proper suffix in (4).

From the definition of longest proper suffix given in (4), we can also say that there is no such $\mathbf{z}'' \in \mathcal{P}_{t+1}$ that satisfies $s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) <_s \mathbf{z}'' <_s \mathbf{z}^p + \mathbf{z}'$.

From the above argument, we can say the following:

$$S(s(\mathbf{z}^p, \mathcal{P}_t) + \mathbf{z}', \mathcal{P}_{t+1}) = s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) \tag{37}$$

When $\mathbf{z}^p + \mathbf{z}' \notin \mathcal{P}_{t+1}$, we can say the following from (4) and (5).

$$S(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) = s(\mathbf{z}^p + \mathbf{z}', \mathcal{P}_{t+1}) \tag{38}$$

So in (35), for $\mathbf{z}^p + \mathbf{z}' \notin \mathcal{P}_{t+1}$, we can cancel out the right and left sides of the subtraction and rewrite it as follows:

$$\sum_{\mathbf{z} \in \mathcal{Y}_{t+2:T+1}} \sum_{\mathbf{z}' \in \mathcal{P}_{t+1}, \mathbf{z}'_{|\mathbf{z}'|-1} = \mathbf{z}^p} \left(\exp\left(\sum_{t'=t+2}^{T+1} \sum_{i=1}^{m} f''_i(\mathbf{x}, \mathbf{z}' + \mathbf{z}, t') \lambda_i \right) \exp\left(\sum_{i=1}^{m} f_i(\mathbf{x}, \mathbf{z}', t+1) \lambda_i \right) - \right.$$
$$\left. \exp\left(\sum_{t'=t+2}^{T+1} \sum_{i=1}^{m} f_i(\mathbf{x}, s(\mathbf{z}', \mathcal{P}_{t+1}) + \mathbf{z}, t') \lambda_i \right) \exp\left(\sum_{i=1}^{m} f_i(\mathbf{x}, s(\mathbf{z}', \mathcal{P}_{t+1}) + \mathbf{z}, \mathcal{P}_{t+1}), t) \lambda_i \right) \right) \tag{39}$$

Applying the definition of β given in (30) and that of W given in (24), we can derive the equation (32).

Now that α's at a position t can be calculated using γ's at the position $t-1$ and γ's using α's at the same position, β's at a position t using δ's at the position $t+1$ and δ's using β's at the same position and $\alpha(BOS, 0) = 1, \delta(\epsilon, T+1) = 1$, we can calculate all the α's, β's, γ's and δ's using dynamic programming. We can then calculate θ's using α's and β's, and then, σ's using θ's.

The normalization factor $Z_{\mathbf{x}}$ can be expressed as $\gamma(\epsilon, T+1)$, because:

$$\gamma(\epsilon, T+1) = \sum_{\mathbf{z}: \mathbf{z} \in \mathcal{Y}_{0:T+1}, \mathbf{z}_{T+2:T+1} = \epsilon} \exp\left(\sum_{i=1}^{m} \sum_{t'=1}^{t} f_i(\mathbf{x}, \mathbf{z}, t') \lambda_i \right) = \sum_{\mathbf{y}} Z_{\mathbf{x}}(\mathbf{y}) = Z_{\mathbf{x}} \tag{40}$$

So we can now calculate the expectations of each feature function as follows:

$$P(f_i \mid \mathbf{x}) = \sum_{t=1}^{T+1} \sigma(\mathbf{z}^i, t) / Z(\mathbf{x}) \tag{41}$$