7 Appendix A: supplement definitions

In this section we give some detailed definitions of variables used in section 3. Recall that in the ambient space to characterize the alignment between any two sequences is defined through the graph $G = (V, E)$ where $V = [T] \times [L] \times \hat{\Sigma}$. The edges $E = E_I \cup E_D \cup E_M$ are defined through V. In particular:

$$E_I = \left\{ (t,l,d) \to (t+1,l,d) \middle| \begin{array}{c} t \in [T-1], \\ l \in [L], \\ d \in \hat{\Sigma} \end{array} \right\}$$

$$E_D = \left\{ (t,l,d_1) \to (t,l+1,d_2) \middle| \begin{array}{c} t \in [T], \\ l \in [L-1], \\ d_1, d_2 \in \hat{\Sigma} \end{array} \right\}$$

$$E_M = \left\{ (t,l,d_1) \to (t+1,l+1,d_2) \middle| \begin{array}{c} t \in [T-1], \\ l \in [L-1], \\ d_1, d_2 \in \hat{\Sigma} \end{array} \right\}$$

These edges represent the action of insertion, deletion and matching while we are aligning two sequences. By assigning $\{0, 1\}$ to each edge we are actually saying if we are taking this action in alignment process or not. Similarly, by assigning penalty weights to each edge we can add our preference to each action. Given a sequence $x_n \in \hat{\Sigma}_n^T$, the penalty associated with edge e is defined as:

$$d(e;x) = \begin{cases} d_I & , e \in E_I \\ d_D & , e \in E_D \\ d_M & , e \in E_M, x_n[t+1] \neq d_2 \\ 0 & , otherwise \end{cases} \tag{26}$$

Then we can define the penalty variable associated with sequence x_n $D_{x_n} \in \{0, d_I, d_D, d_M\}^{|E|}$, where $D_{x_n}[e] = d(e; x_n)$. Now consider all data sequences $\{x_n\}_{n=1}^N$, then the whole penalty variable is defined by simply stacking all D_{x_n} together in a new dimension.