Kernel methods for vessel trajectories

de Vries, G.K.D.

Citation for published version (APA):
EDIT DISTANCE SOFT-MAX KERNEL PROOF

To prove that the $\text{sim}_{\text{softmax,ED}}$ function is positive semi-definite, we construct it as a mapping kernel (see Definition 2.1.2 and the paper by Shin and Kuboyama (2008)). Because of the use of the mapping kernel, this proof is different than the one by Vert et al. (2004). The proof considers sequences in general, not specifically trajectories. Furthermore, we assume that we have a conditionally positive definite substitution function $\text{sub}$, e.g. the negative of the Euclidean distance as used in Chapter 4. We also assume $\beta > 0$ and a value for $g$. Because $\text{sub}$ is conditionally positive definite, $\exp(\beta \text{sub})$ is positive semi-definite (Berg et al. (1984)).

First we define the set of all non-contiguous subsequences of a sequence $S$ of length $n$:

$$\text{subseq}(S, n) = \{(S', x) | S' = S(i_1), \ldots, S(i_n), 1 \leq i_1 < \ldots < i_n \leq |S|, x = |S| - n\}.$$  \hspace{1cm} (B.1)

The integer $x$ represents the number of unused sequence elements, i.e. $|S| - n$. Thus, this set contains pairs of subsequences and integers.

For a set of sequences $\delta$ and $n \geq 0$ we define the kernel below on the set $\bigcup_{S \in \delta} \text{subseq}(S, n)$,

$$k_{\text{subseq}}^n((S', x), (T', y)) = \exp(\beta g \cdot (x + y)) \cdot \prod_{i=1}^{n} \exp(\beta \text{sub}(S'(i), T'(i))).$$ \hspace{1cm} (B.2)

The part $\exp(\beta g \cdot (x + y))$ is positive semi-definite, since it can be written as $\exp(\beta gx) \cdot \exp(\beta gy)$.\footnote{It is a basic string kernel defined by Vert et al. (2004).} We know that $\exp(\beta \text{sub})$ is positive semi-definite. Hence, $k_{\text{subseq}}^n$ is positive semi-definite, since it is the product of positive semi-definite kernels.

For the set $\bigcup_{S \in \delta} \text{subseq}(S, n)$ we create the straightforward transitive map:

$$M^n(S, T) = \text{subseq}(S, n) \times \text{subseq}(T, n).$$ \hspace{1cm} (B.3)

With the map $M^n$ we create the mapping kernel:

$$k_{\text{fixed}}^n(S, T) = \sum_{((S', x), (T', y)) \in M^n(S, T)} k_{\text{subseq}}^n((S', x), (T', y)).$$ \hspace{1cm} (B.4)

We create the following positive semi-definite kernel, by closure under pointwise limit:

$$k_{\text{all}}(S, T) = \sum_{i=0}^{\infty} k_{\text{fixed}}^i(S, T).$$ \hspace{1cm} (B.5)
This kernel considers all possible length subsequences of $S$ and $T$.

The kernel $k_{\text{all}}$ is equivalent to $\text{sim}_{\text{softmax,ED}}$. To see this we construct the map $M^n$ in another way, via the definition of the edit distance alignment. First, consider the set of all alignments of length $n$ for all edit distance alignments $\Pi(S, T)$ between two sequences $S$ and $T$ (see Definition 4.2.3):

$$\Pi^n(S, T) = \{\pi \in \Pi(S, T) \mid |\pi| = n\}.$$  \hspace{1cm} (B.6)

Using this set we define the map:

$$N^n(S, T) = \{(S', (x), [T', y]) \mid S' = S(\pi_1(1)), \ldots, S(\pi_1(n)),$$
$$T' = T(\pi_2(1)), \ldots, T(\pi_2(n)),$$
$$\pi \in \Pi^n(S, T), x = |S| - n, y = |T| - n\}.$$ \hspace{1cm} (B.7)

This map is equal to the map $M^n$ as in the following lemma.

**Lemma B.0.1.** For two sequences $S$ and $T$:

$$M^n(S, T) = N^n(S, T).$$

**Proof.** To prove Lemma B.0.1 we assume towards contradiction that the equality does not hold. This gives two cases:

$$\exists((S', |S'| - n), (T', |T'| - n)) \in M^n(S, T)$$
$$\wedge ((S', |S'| - n), (T', |T'| - n)) \notin N^n(S, T) \hspace{1cm} (1)$$

and (2)

$$\exists((S', |S'| - n), (T', |T'| - n)) \in N^n(S, T)$$
$$\wedge ((S', |S'| - n), (T', |T'| - n)) \notin M^n(S, T). \hspace{1cm} (2)$$

(1) The pair $((S', |S'| - n), (T', |T'| - n))$ is not in $N^n(S, T)$. Thus there is no alignment $\pi \in \Pi^n(S, T)$ for which $S' = S(\pi_1(1)), \ldots, S(\pi_1(n))$ and $T' = T(\pi_2(1)), \ldots, T(\pi_2(n))$. Thus one of the constraints of Definition 4.2.3 does not hold. Without loss of generality we assume that this is the first constraint: $1 \leq \pi_1(1) < \pi_1(2) < \ldots < \pi_1(n) \leq |S|$. Thus either $\pi_1(1) < 1, \pi_1(n) > |S'|$, or $\pi_1(i + 1) \leq \pi_1(i)$, all of which violate Equation B.1. Hence we arrive at a contradiction.

(2) We assume that the pair $((S', |S'| - n), (T', |T'| - n))$ is not in $M^n(S, T)$. Since $M^n(S, T)$ is defined as the full cross product, we can assume without loss of generality that $(S', |S'| - n) \notin \text{subseq}(S', n)$. Thus the constraint $1 \leq i_1 < \ldots < i_n \leq |S|$ is violated. Like in case (1), this leads to a contradiction with the definition of an edit distance alignment (Definition 4.2.3). \hfill \square
Thus, we can rewrite the kernel $k_{\text{all}}$, using the shorthand $S_{\pi_1}$ for $S(\pi_1(1)), \ldots, S(\pi_1(|\pi|))$ and $T_{\pi_2}$ for $T(\pi_2(1)), \ldots, T(\pi_2(|\pi|))$:

$$k_{\text{all}}(S, T) = \sum_{i=0}^{\infty} \sum_{(S', x), (T', y) \in N^i(S, T)} k^i_{\text{subseq}}((S', x), (T', y)) \tag{B.8}$$

$$= \sum_{i=0}^{\infty} \sum_{\pi \in \Pi^i(S, T)} k^i_{\text{subseq}}(\{S_{\pi_1}(S| - i), (T_{\pi_2}|T| - i)\}) \tag{B.9}$$

$$= \sum_{\pi \in \Pi(S, T)} \prod_{i=1}^{\pi} \exp(\beta_{\text{sub}}(S(\pi_1(i)), T(\pi_2(i)))) \cdot \exp(\beta g \cdot ((|S| - |\pi|) + (|T| - |\pi|))) \tag{B.10}$$

$$= \sum_{\pi \in \Pi(S, T)} \exp(\beta \sum_{i=1}^{\pi} \text{sub}(S(\pi_1(i)), T(\pi_2(i))))$$

$$+ g \cdot ((|S| - |\pi|) + (|T| - |\pi|))) \tag{B.11}$$

$$= \text{sim}_{\text{softmax,ED}}(S, T). \tag{B.12}$$

For line B.8 we use Equation B.4 and plug in the map $N^i$. We use Equation B.7 to get line B.9. With Equation B.2 and the fact that $\bigcup_{i=0}^{\infty} \Pi^i = \Pi$ we derive line B.10. Finally we put the product in the exponent to get line B.11. This line is equal to the definition of $\text{sim}_{\text{softmax,ED}}$. Thus, we have proven that $\text{sim}_{\text{softmax,ED}}$ is positive semi-definite.