

# A Kernel for Hierarchical Parameter Spaces

Frank Hutter and Michael A. Osborne

fh@informatik.uni-freiburg.de and mosb@robots.ox.ac.uk

September 6, 2013

## Abstract

We define a family of kernels for mixed continuous/discrete hierarchical parameter spaces and show that they are positive definite.

## 1 Introduction

We aim to do inference about some function  $g$  with domain (input space)  $\mathcal{X}$ .  $\mathcal{X} = \prod_{i=1}^D \mathcal{X}_i$  is a  $D$ -dimensional input space, where each individual dimension is either bounded real or categorical, that is,  $\mathcal{X}_i$  is either  $[l_i, u_i] \subset \mathbb{R}$  (with lower and upper bounds  $l_i$  and  $u_i$ , respectively) or  $\{v_{i,1}, \dots, v_{i,m_i}\}$ .

Associated with  $\mathcal{X}$ , there is a DAG structure  $\mathcal{D}$ , whose vertices are the dimensions  $\{1, \dots, D\}$ .  $\mathcal{X}$  will be restricted by  $\mathcal{D}$ : if vertex  $i$  has children under  $\mathcal{D}$ ,  $\mathcal{X}_i$  must be categorical.  $\mathcal{D}$  is also used to specify when each input is *active* (that is, relevant to inference about  $g$ ). In particular, we assume each input dimension is only active under some instantiations of its ancestor dimensions in  $\mathcal{D}$ . More precisely, we define  $D$  functions  $\delta_i: \mathcal{X} \rightarrow \mathcal{B}$ , for  $i \in \{1, \dots, D\}$ , and where  $\mathcal{B} = \{\text{true}, \text{false}\}$ . We take

$$\delta_i(\underline{x}) = \delta_i(\underline{x}(\text{anc}_i)), \quad (1)$$

where  $\text{anc}_i$  are the ancestor vertices of  $i$  in  $\mathcal{D}$ , such that  $\delta_i(\underline{x})$  is true only for appropriate values of those entries of  $\underline{x}$  corresponding to ancestors of  $i$  in  $\mathcal{D}$ . We say  $i$  is active for  $\underline{x}$  iff  $\delta_i(\underline{x})$ .

Our aim is to specify a kernel for  $\mathcal{X}$ , *i.e.*, a positive semi-definite function  $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ . We will first specify an individual kernel for each input dimension, *i.e.*, a positive semi-definite function  $k_i: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ .  $k$  can then be taken as either a sum,

$$k(\underline{x}, \underline{x}') = \sum_{i=1}^D k_i(\underline{x}, \underline{x}'), \quad (2)$$

product,

$$k(\underline{x}, \underline{x}') = \prod_{i=1}^D k_i(\underline{x}, \underline{x}'), \quad (3)$$

or any other permitted combination, of these individual kernels. Note that each individual kernel  $k_i$  will depend on an input vector  $\underline{x}$  only through dependence on  $x_i$  and  $\delta_i(\underline{x})$ ,

$$k_i(\underline{x}, \underline{x}') = \tilde{k}_i(x_i, \delta_i(\underline{x}), x'_i, \delta_i(\underline{x}')). \quad (4)$$

That is,  $x_j$  for  $j \neq i$  will influence  $k_i(\underline{x}, \underline{x}')$  only if  $j \in \text{anc}_i$ , and only by affecting whether  $i$  is active.

Below we will construct pseudometrics  $d_i: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ : that is,  $d_i$  satisfies the requirements of a metric aside from the identity of indiscernibles. As for  $k_i$ , these pseudometrics will depend on an input vector  $\underline{x}$  only through dependence on both  $x_i$  and  $\delta_i(\underline{x})$ .  $d_i(\underline{x}, \underline{x}')$  will be designed to provide an intuitive measure of how different  $g(\underline{x})$  is from  $g(\underline{x}')$ . For each  $i$ , we will then construct a (pseudo-)isometry  $f_i$  from  $\mathcal{X}$  to a Euclidean space ( $\mathbb{R}^2$  for bounded real parameters, and  $\mathbb{R}^m$  for categorical-valued parameters with  $m$  choices). That is, denoting the Euclidean metric on the appropriate space as  $d_E$ ,  $f_i$  will be such that

$$d_i(\underline{x}, \underline{x}') = d_E(f_i(\underline{x}), f_i(\underline{x}')) \quad (5)$$

for all  $\underline{x}, \underline{x}' \in \mathcal{X}$ . We can then use our transformed inputs,  $f_i(\underline{x})$ , within any standard Euclidean kernel  $\kappa$ . We'll make this explicit in Proposition 2.

**Definition 1.** A function  $\kappa: \mathbb{R}^+ \rightarrow \mathbb{R}$  is a positive semi-definite covariance function over Euclidean space if  $K \in \mathbb{R}^{N \times N}$ , defined by

$$K_{m,n} = \kappa(d_E(y_m, y_n)), \quad \text{for } y_m, y_n \in \mathbb{R}^P, \quad m, n = 1, \dots, N,$$

is positive semi-definite for any  $y_1, \dots, y_N \in \mathbb{R}^P$ .

A popular example of such a  $\kappa$  is the exponentiated quadratic, for which  $\kappa(\delta) = \sigma^2 \exp(-\frac{1}{2} \frac{\delta^2}{\lambda^2})$ ; another popular choice is the rational quadratic, for which  $\kappa(\delta) = \sigma^2 (1 + \frac{1}{2\alpha} \frac{\delta^2}{\lambda^2})^{-\alpha}$ .

**Proposition 2.** Let  $\kappa$  be a positive semi-definite covariance function over Euclidean space and let  $d_i$  satisfy Equation 5. Then,  $k_i: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ , defined by

$$k_i(\underline{x}, \underline{x}') = \kappa(d_i(\underline{x}, \underline{x}'))$$

is a positive semi-definite covariance function over input space  $\mathcal{X}$ .

*Proof.* We need to show that for any  $\underline{x}_1, \dots, \underline{x}_N \in \mathcal{X}$ ,  $K \in \mathbb{R}^{N \times N}$  defined by

$$K_{m,n} = \kappa(d_i(\underline{x}_m, \underline{x}_n)), \quad \text{for } \underline{x}_m, \underline{x}_n \in \mathcal{X}, \quad m, n = 1, \dots, N,$$

is positive semi-definite. Now, by the definition of  $d_i$ ,

$$K_{m,n} = \kappa(d_E(f_i(\underline{x}_m), f_i(\underline{x}_n))) = \kappa(d_E(y_m, y_n))$$

where  $y_m = f_i(\underline{x}_m)$  and  $y_n = f_i(\underline{x}_n)$  are elements of  $\mathbb{R}^P$ . Then, by assumption that  $\kappa$  is a positive semi-definite covariance function over Euclidean space,  $K$  is positive semi-definite.  $\square$

We'll now define pseudometrics  $d_i$  and associated isometries  $f_i$  for both the bounded real and categorical cases.

## 2 Bounded Real Dimensions

Let's first focus on a bounded real input dimension  $i$ , i.e.,  $\mathcal{X}_i = [l_i, u_i]$ . To emphasize that we're in this real case, we explicitly denote the pseudometric as  $d_i^r$  and the (pseudo-)isometry from  $(\mathcal{X}, d_i)$  to  $\mathbb{R}^2, d_E$  as  $f_i^r$ . For the definitions, recall that  $\delta_i(\underline{x})$  is true iff dimension  $i$  is active given the instantiation of  $i$ 's ancestors in  $\underline{x}$ .

$$d_i^r(\underline{x}, \underline{x}') = \begin{cases} 0 & \text{if } \delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{false} \\ \omega_i & \text{if } \delta_i(\underline{x}) \neq \delta_i(\underline{x}') \\ \omega_i \sqrt{2} \sqrt{1 - \cos(\pi \rho_i \frac{x_i - x'_i}{u_i - l_i})} & \text{if } \delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{true}. \end{cases}$$

$$f_i^r(\underline{x}) = \begin{cases} [0, 0]^\top & \text{if } \delta_i(\underline{x}) = \text{false} \\ \omega_i [\sin \pi \rho_i \frac{x_i - l_i}{u_i - l_i}, \cos \pi \rho_i \frac{x_i - l_i}{u_i - l_i}]^\top & \text{otherwise.} \end{cases}.$$

Although our formal arguments do not rely on this, Proposition 5 in the appendix shows that  $d_i^r$  is a pseudometric. This pseudometric is defined by two parameters:  $\omega_i \in [0, 1]$  and  $\rho_i \in [0, 1]$ . We firstly define

$$\omega_i = \prod_{j \in \text{anc}_i \cup \{i\}} \gamma_j, \quad (6)$$

where  $\gamma_j \in [0, 1]$ . This encodes the intuitive notion that differences on lower levels of the hierarchy count less than differences in their ancestors.

Also note that, as desired, if  $i$  is inactive for both  $\underline{x}$  and  $\underline{x}'$ ,  $d_i^r$  specifies that  $g(\underline{x})$  and  $g(\underline{x}')$  should not differ owing to differences between  $x_i$  and  $x'_i$ . Secondly, if  $i$  is active for both  $\underline{x}$  and  $\underline{x}'$ , the difference between  $g(\underline{x})$  and  $g(\underline{x}')$  due to  $x_i$  and  $x'_i$  increases monotonically with increasing  $|x_i - x'_i|$ . Parameter  $\rho_i$  controls whether differing in the activity of  $i$  contributes more or less to the distance than differing in  $x_i$  should  $i$  be active. If  $\rho = 1/3$ , and if  $i$  is inactive for exactly one of  $\underline{x}$  and  $\underline{x}'$ ,  $g(\underline{x})$  and  $g(\underline{x}')$  are as different as is possible due to dimension  $i$ ; that is,  $g(\underline{x})$  and  $g(\underline{x}')$  are exactly as different in that case as if  $x_i = l_i$  and  $x'_i = u_i$ . For  $\rho > 1/3$ ,  $i$  being active for both  $\underline{x}$  and  $\underline{x}'$  means that  $g(\underline{x})$  and  $g(\underline{x}')$  could potentially be more different than if  $i$  was active in only one of them. For  $\rho < 1/3$ , the converse is true.<sup>1</sup>

We now show that  $d_i^r$  and  $f_i^r$  can be plugged into a positive semi-definite kernel over Euclidean space to define a valid kernel over space  $\mathcal{X}$ .

**Proposition 3.** *Let  $\kappa$  be a positive semi-definite covariance function over Euclidean space. Then,  $k_i: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ , defined by*

$$k_i(\underline{x}, \underline{x}') = \kappa(d_i^r(\underline{x}, \underline{x}'))$$

*is a positive semi-definite covariance function over input space  $\mathcal{X}$ .*

<sup>1</sup>Note that  $\underline{x}$  and  $\underline{x}'$  must differ in at least one ancestor dimension of  $i$  in order for  $\delta_i(\underline{x}) \neq \delta_i(\underline{x}')$  to hold, such that in the final kernel combining kernels  $k_i$  due to each dimension  $i$ , differences in the activity of dimension  $i$  are penalized both in kernel  $k_i$  and in the distance for the kernel of the ancestor dimension causing the difference in  $i$ 's activity.

*Proof.* Due to Proposition 2, we only need to show that, for any two inputs  $\underline{x}, \underline{x}' \in \mathcal{X}$ , the isometry condition  $d_E(f_i^r(\underline{x}), f_i^r(\underline{x}')) = d_i^r(\underline{x}, \underline{x}')$  holds.

We use the abbreviation  $\alpha = \pi \rho_i \frac{x_i}{u_i - l_i}$  and  $\alpha' = \pi \rho_i \frac{x'_i}{u_i - l_i}$  and consider the following three possible cases of dimension  $i$  being active or inactive in  $\underline{x}$  and  $\underline{x}'$ .

Case 1:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{false}$ . In this case, we trivially have

$$d_E(f_i^r(\underline{x}), f_i^r(\underline{x}')) = d_E([0, 0]^\top, [0, 0]^\top) = 0 = d_i^r(\underline{x}, \underline{x}').$$

Case 2:  $\delta_i(\underline{x}) \neq \delta_i(\underline{x}')$ . In this case, we have

$$d_E(f_i^r(\underline{x}), f_i^r(\underline{x}')) = d_E([\sin \alpha, \cos \alpha]^\top, [0, 0]^\top) = \sqrt{\omega_i^2(\sin^2 \alpha + \cos^2 \alpha)} = \omega_i = d_i^r(\underline{x}, \underline{x}'),$$

and symmetrically for  $d_E([0, 0]^\top, [\sin \alpha, \cos \alpha]^\top)$ .

Case 3:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{true}$ . We have:

$$\begin{aligned} d_E(f_i^r(\underline{x}), f_i^r(\underline{x}')) &= d_E(\omega_i[\sin \alpha, \cos \alpha]^\top, \omega_i[\sin \alpha', \cos \alpha']^\top) \\ &= \omega_i \sqrt{(\sin \alpha - \sin \alpha')^2 + (\cos \alpha - \cos \alpha')^2} \\ &= \omega_i \sqrt{\sin^2 \alpha - 2 \sin \alpha \sin \alpha' + \sin^2 \alpha' + \cos^2 \alpha - 2 \cos \alpha \cos \alpha' + \cos^2 \alpha'} \\ &= \omega_i \sqrt{(\sin^2 \alpha + \cos^2 \alpha) + (\sin^2 \alpha' + \cos^2 \alpha') - 2(\sin \alpha \sin \alpha' + \cos \alpha \cos \alpha')} \\ &= \omega_i \sqrt{1 + 1 - 2 \cos(\alpha - \alpha')} \\ &= \omega_i \sqrt{2} \sqrt{1 - \cos(\pi \rho_i \frac{x_i - x'_i}{u_i - l_i})} = d_i^r(\underline{x}, \underline{x}'), \end{aligned} \tag{7}$$

where (7) follows from the previous line by using the identity

$$\cos(a - b) = \cos a \cos b + \sin a \sin b.$$

□

### 3 Categorical Dimensions

Now let's define  $f_i^c$  and  $d_i^c$  for the case that the input  $\mathcal{X}_i = \{v_{i,1}, \dots, v_{i,m_i}\}$  is categorical with  $m_i$  possible values. Proceeding as above, we define a pseudometric  $d_i^c$  on  $\mathcal{X}$  and an isometry from  $(\mathcal{X}, d_i^c)$  to  $(\mathbb{R}^{m_i}, d_E^{m_i})$ , and show that we can combine these with a kernel over Euclidean space to construct a valid kernel over space  $\mathcal{X}$ .

$$d_i^c(\underline{x}, \underline{x}') = \begin{cases} 0 & \text{if } \delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{false} \\ w_i & \text{if } \delta_i(\underline{x}) \neq \delta_i(\underline{x}') \\ w_i \sqrt{2} \mathbb{I}_{x_i \neq x'_i} & \text{if } \delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{true}. \end{cases}$$

$$f_i^c(\underline{x}) = \begin{cases} \underline{0} \in \mathbb{R}^{m_i} & \text{if } \delta_i(\underline{x}) = \text{false} \\ w_i \underline{e}_j & \delta_i(\underline{x}) = \text{true and } x_i = v_{i,j}, \end{cases}$$

where  $\underline{e}_j \in \mathbb{R}^{m_i}$  is zero in all dimensions except  $j$ , where it is 1.

**Proposition 4.** *Let  $\kappa$  be a positive semi-definite covariance function over Euclidean space. Then,  $k_i: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ , defined by*

$$k_i(\underline{x}, \underline{x}') = \kappa(d_i^c(\underline{x}, \underline{x}'))$$

*is a positive semi-definite covariance function over input space  $\mathcal{X}$ .*

*Proof.* We proceed as in the proof of Proposition 3 to show that, for any two inputs  $\underline{x}, \underline{x}' \in \mathcal{X}$ , the isometry condition  $d_E^{m_i}(f_i^c(\underline{x}), f_i^c(\underline{x}')) = d_i^c(\underline{x}, \underline{x}')$  holds.

Case 1:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{false}$ . In this case, we trivially have

$$d_E^{m_i}(f_i^r(\underline{x}), f_i^r(\underline{x}')) = d_E^{m_i}(\underline{0}, \underline{0}) = 0 = d_i^r(\underline{x}, \underline{x}').$$

Case 2:  $\delta_i(\underline{x}) \neq \delta_i(\underline{x}')$ . In this case, we have

$$d_E^{m_i}(f_i^c(\underline{x}), f_i^c(\underline{x}')) = d_E^{m_i}(w_i \underline{e}_j, \underline{0}) = w_i = d_i(\underline{x}, \underline{x}'),$$

and symmetrically for  $d_E(\underline{0}, w_i \underline{e}_j)$ .

Case 3:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{true}$ . If  $x_i = x'_i = v_{i,j}$ , we have

$$d_E^{m_i}(f_i^c(\underline{x}), f_i^c(\underline{x}')) = d_E^{m_i}(w_i \underline{e}_j, w_i \underline{e}_j) = 0 = d_i^c(\underline{x}, \underline{x}')$$

If  $x_i = v_{i,j} \neq v_{i,j'} = x'_i$ , we have

$$d_E(f_i^c(\underline{x}), f_i^c(\underline{x}')) = d_E^{m_i}(w_i \underline{e}_j, w_i \underline{e}_{j'}) = w_i \sqrt{2} = d_i^c(\underline{x}, \underline{x}')$$

□

## A Proof of pseudometric properties

**Proposition 5.**  $d_i^r$  is a pseudometric on  $\mathcal{X}$ .

*Proof.* The non-negativity and symmetry of  $d_i^r$  are trivially proven. To prove the triangle inequality, consider  $\underline{x}, \underline{x}', \underline{x}'' \in \mathcal{X}$ .

Case 1:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{false}$ , such that  $d_i^r(\underline{x}, \underline{x}') = 0$ . Here, from non-negativity, clearly  $d_i^r(\underline{x}, \underline{x}') = 0 \leq d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}'')$ .

Case 2:  $\delta_i(\underline{x}) \neq \delta_i(\underline{x}')$ , such that  $d_i^r(\underline{x}, \underline{x}') = \omega_i$ . Without loss of generality, assume  $\delta_i(\underline{x}) = \text{true}$ ,  $\delta_i(\underline{x}') = \text{false}$  and  $\delta_i(\underline{x}'') = \text{true}$ .

$$d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}'') = d_i^r(\underline{x}, \underline{x}') + \omega_i \quad (8)$$

Hence  $d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}'') \geq \omega_i = d_i^r(\underline{x}, \underline{x}')$  by non-negativity.

Case 3:  $\delta_i(\underline{x}) = \delta_i(\underline{x}') = \text{true}$ , such that  $d_i^r(\underline{x}, \underline{x}') = \omega_i \sqrt{2} \sqrt{1 - \cos(\pi \rho_i \frac{x_i - x'_i}{u_i - l_i})}$ . If  $\delta_i(\underline{x}'') = \text{false}$ ,

$$d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}'') = 2\omega_i \geq \omega_i \sqrt{2} \sqrt{1 - \cos(\pi \rho_i \frac{x_i - x'_i}{u_i - l_i})} = d_i^r(\underline{x}, \underline{x}'). \quad (9)$$

If  $\delta_i(\underline{x}'') = \text{true}$ , consider the worst possible case in which, without loss of generality,  $x_i = l_i$  and  $x'_i = u_i$ , such that  $d_i^r(\underline{x}, \underline{x}') = 2\omega_i^2$ . We define the abbreviation  $\beta'' = \frac{x''_i - l_i}{u_i - l_i}$ , giving

$$\begin{aligned} (d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}''))^2 &= 2\omega_i^2 \left( \sqrt{1 - \cos(\pi \rho_i \beta'')} + \sqrt{1 - \cos(\pi \rho_i (1 - \beta''))} \right)^2 \\ &= 2\omega_i^2 \left( 2 - \cos(\pi \rho_i \beta'') - \cos(\pi \rho_i (1 - \beta'')) \right. \\ &\quad \left. + 2\sqrt{(1 - \cos(\pi \rho_i \beta''))(1 - \cos(\pi \rho_i (1 - \beta'')))} \right) \\ &= 2\omega_i^2 \left( 2 + 2\sqrt{1 + \cos(\pi \rho_i \beta'') \cos(\pi \rho_i (1 - \beta''))} \right) \\ &= 4\omega_i^2 (1 + |\sin \pi \rho_i \beta''|) \\ &\geq 4\omega_i^2 = d_i^r(\underline{x}, \underline{x}')^2. \end{aligned} \quad (10)$$

Hence, from non-negativity, we have  $d_i^r(\underline{x}, \underline{x}'') + d_i^r(\underline{x}', \underline{x}'') \geq d_i^r(\underline{x}, \underline{x}')$ .  $\square$

**Proposition 6.**  $d_i^{\mathcal{E}}$  is a pseudometric on  $\mathcal{X}$ .

*Proof.* A trivial modification of the proof to Proposition 5.  $\square$