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 kernels for mixed continuous/discrete spaces and conditional 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 $\{v_{i,1}, \ldots, v_{i,m_i}\}$.

Associated with $\mathcal{X}$, there is a DAG structure $\mathcal{D}$, whose vertices are the dimensions $\{1, \ldots, D\}$. $\mathcal{X}$ 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 $\delta_i \colon \mathcal{X} \to \mathcal{B}$, for $i \in \{1, \ldots, D\}$, and where $\mathcal{B} = \{\text{true}, \text{false}\}$. We take

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

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

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

product,

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

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 $\delta_i(\underline{x})$,

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

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

Below we will construct pseudometrics $d_i \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$: 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

$$d_i(\underline{x}, \underline{x}') = d_{\mathrm{E}}\big(f_i(\underline{x}), f_i(\underline{x}')\big) \tag{5}$$

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 κ. We'll make this explicit in Proposition 2.

FH: I tried to make things a bit more formal here by using explicit definitions and referring back to them later. We might want to go one step further and do something similar for isometries/pseudometrics, once it is clear that we actually need these concepts in the formal proofs. What do you think?

MO: sure

.

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

$$K_{m,n} = \kappa\big(d_{\mathrm{E}}(\underline{y}_m, \underline{y}_n)\big), \quad \textit{for } \underline{y}_m, \underline{y}_n \in \mathbb{R}^P, \quad m, n = 1, \ldots, N,$$

is positive semi-definite for any $\underline{y}_1, \ldots, \underline{y}_N \in \mathbb{R}^P$.

A popular example of such a κ 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* κ *be a positive semi-definite covariance function over Euclidean space and let* d_i *satisfy Equation 5. Then,* $k_i \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$*, defined by*

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

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

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

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

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

$$K_{m,n} = \kappa\Big(d_{\mathrm{E}}(f_i(\underline{x}_m), f_i(\underline{x}_n))\Big) = \kappa\big(d_{\mathrm{E}}(\underline{y}_m, \underline{y}_n)\big)$$

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

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 define the 'difference' function d_i on $\mathcal{X}$ and the isometry f_i from $(\mathcal{X}, d_i)$ to $\mathbb{R}^2, d_{\mathrm{E}}$ for the case that the input $\mathcal{X}_i = [l_i, u_i]$ is bounded real; to emphasize that we're in this real case, we explicitly denote these functions as d_i^{r} and f_i^{r}. We first define d_i^{r}, recalling that $\delta_i(\underline{x})$ is true iff dimension i is active given the instantiation of i's ancestors in $\underline{x}$.

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

This pseudometric (see Proposition 3) is defined by two parameters: $\omega_i \in [0,1]$ and $\rho_i \in [0,1]$. We firstly define

$$\omega_i = \prod_{j \in \mathrm{anc}_i} \gamma_j, \tag{6}$$

where $\gamma_j \in [0,1]$. Hence the higher up i is in the hierarchy $\mathcal{D}$, the greater the distance due to i can be.

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 $|x_i - x_i'|$. ρ_i then 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 eaxctly 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 only one of the two had i active. For $\rho < 1/3$, the converse is true. [1]

FH: I don't fully understand why we actually need to show that d_i^{r} is a (pseudo)metric. I understand that you might argue that isometries are only defined between metric spaces, but we don't even need the fact that f_i is an isometry, other than by using the property $d_i(\underline{x}, \underline{x}') = d_{\mathrm{E}}(f_i(\underline{x}), f_i(\underline{x}'))$ from Equation 5, do we? I believe we could just drop Proposition 3, reword Proposition 4 to just say $d_{\mathrm{E}}(f_i^{\mathrm{r}}(\underline{x}), f_i^{\mathrm{r}}(\underline{x}')) = d_i^{\mathrm{r}}(\underline{x}, \underline{x}')$, and then aren't we done? (same for the categorical case)

MO: you're absolutely correct, for our purposes we can do exactly as you say. I just thought it might be nice to prove that d_i was a pseudometric and that f_i was a (pseudo-)isometry; perhaps someone else might want to use those properties

.

Proposition 3. *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 $d_i^{\mathrm{r}}(\underline{x}, \underline{x}') = 0$. Here, from non-negativity, clearly $d_i^{\mathrm{r}}(\underline{x}, \underline{x}') = 0 \leq d_i^{\mathrm{r}}(\underline{x}, \underline{x}'') + d_i^{\mathrm{r}}(\underline{x}', \underline{x}'')$.

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

$$d_i^{\mathrm{r}}(\underline{x}, \underline{x}'') + d_i^{\mathrm{r}}(\underline{x}', \underline{x}'') = d_i^{\mathrm{r}}(\underline{x}, \underline{x}'') + \omega_i \tag{7}$$

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

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

[1] 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 differences in the activity of dimension i are penalized both in distance d_i and in the distance for the ancestor dimension causing the difference in i's activity.

FH: note: if we wanted to, we could probably get a condition for the joint overall kernel that the distance will always be larger if configurations differ on a higher level in the dimensionality DAG, by multiplying ω_i by the weights $\omega_j \in [0, 1]$ of all of i's ancestors j (because at least one ancestor has to differ). We'd further have to divide ω_i by the maximal number of descendants a dimension has, in order to ensure that a difference at a higher level counts more than all the differences at the descendant-dimensions combined. But of course for this we wouldn't be able to do the proof of PSD-ness for each dimension by itself, so things would get a lot more hairy, and at least at this point there is no need for that.

MO: actually, we should absolutely be able to do that, without any problem. I don't think it'll get hairy. It's just a matter of replacing ω_i with a slightly different constant, dependent on i, but independent of $\underline{x}$. Let's do it!

$\delta_i(\underline{x}'') = \text{false}$,

$$d_i^{\mathrm{r}}(\underline{x}, \underline{x}'') + d_i^{\mathrm{r}}(\underline{x}', \underline{x}'') = 2\omega_i \geq \omega_i \sqrt{2} \sqrt{1 - \cos(\pi \rho_i \frac{x_i - x_i'}{u_i - l_i})} = d_i^{\mathrm{r}}(\underline{x}, \underline{x}'). \tag{8}$$

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 $d_i^{\mathrm{r}}(\underline{x}, \underline{x}') = 2\omega_i^2$. We define the abbreviation $\beta'' = \frac{x_i'' - l_i}{u_i - l_i}$, giving

$$\begin{aligned}
\left(d_i^{\mathrm{r}}(\underline{x}, \underline{x}'') + d_i^{\mathrm{r}}(\underline{x}', \underline{x}'')\right)^2 &= 2\omega_i^2 \left(\sqrt{1 - \cos(\pi \rho_i \beta'')} + \sqrt{1 - \cos\big(\pi \rho_i (1 - \beta'')\big)} \right)^2 \\
&= 2\omega_i^2 \Bigg(2 - \cos(\pi \rho_i \beta'') - \cos\big(\pi \rho_i (1 - \beta'')\big) \\
&\qquad + 2 \sqrt{\Big(1 - \cos(\pi \rho_i \beta'')\Big) \Big(1 - \cos\big(\pi \rho_i (1 - \beta'')\big)\Big)} \Bigg) \\
&= 2\omega_i^2 \left(2 + 2 \sqrt{1 + \cos(\pi \rho_i \beta'') \cos\big(\pi \rho_i (1 - \beta'')\big)} \right) \\
&= 4\omega_i^2 \big(1 + |\sin \pi \rho_i \beta''|\big) \\
&\geq 4\omega_i^2 = d_i^{\mathrm{r}}(\underline{x}, \underline{x}')^2.
\end{aligned} \tag{9}$$

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

Now we define an isometric embedding f_i^{r} of $(\mathcal{X}, d_i^{\mathrm{r}})$ into $(\mathbb{R}^2, d_{\mathrm{E}})$:

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

Proposition 4. *f_i^{r} is an isometry from $(\mathcal{X}, d_i^{\mathrm{r}})$ to $(\mathbb{R}^2, d_{\mathrm{E}})$.*

Proof. Consider two inputs $\underline{x}, \underline{x}' \in \mathcal{X}$. We need to show that $d_{\mathrm{E}}\big(f_i^{\mathrm{r}}(\underline{x}), f_i^{\mathrm{r}}(\underline{x}')\big) = d_i^{\mathrm{r}}(\underline{x}, \underline{x}')$. 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

$$d_{\mathrm{E}}(f_i^{\mathrm{r}}(\underline{x}), f_i^{\mathrm{r}}(\underline{x}')) = d_{\mathrm{E}}([0, 0]^\intercal, [0, 0]^\intercal) = 0 = d_i^{\mathrm{r}}(\underline{x}, \underline{x}').$$

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

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

and symmetrically for $d_{\mathrm{E}}([0, 0]^\intercal, [\sin \alpha, \cos \alpha]^\intercal)$.

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

$$\begin{aligned}
d_{\mathrm{E}}(f_i^{\mathrm{r}}(\underline{x}), f_i^{\mathrm{r}}(\underline{x}')) &= d_{\mathrm{E}}(\omega_i[\sin\alpha, \cos\alpha]^{\intercal}, \omega_i[\sin\alpha', \cos\alpha']^{\intercal}) \\
&= \omega_i\sqrt{(\sin\alpha - \sin\alpha')^2 + (\cos\alpha - \cos\alpha')^2} \\
&= \omega_i\sqrt{\sin^2\alpha - 2\sin\alpha\sin\alpha' + \sin^2\alpha' + \cos^2\alpha - 2\cos\alpha\cos\alpha' + \cos^2\alpha'} \\
&= \omega_i\sqrt{(\sin^2\alpha + \cos^2\alpha) + (\sin^2\alpha' + \cos^2\alpha') - 2(\sin\alpha\sin\alpha' + \cos\alpha\cos\alpha')} \\
&= \omega_i\sqrt{1 + 1 - 2\cos(\alpha - \alpha')} \\
&= \omega_i\sqrt{2}\sqrt{1 - \cos(\pi\rho_i\frac{x_i - x_i'}{u_i - l_i})} = d_i^{\mathrm{r}}(\underline{x}, \underline{x}'), \qquad (10)
\end{aligned}$$

where (10) 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 $\mathcal{X}_i = \{v_{i,1}, \ldots, v_{i,m_i}\}$ is categorical with m_i possible values. Proceeding as above, we define a pseudometric d_i^{c} on $\mathcal{X}$ and an isometry from $(\mathcal{X}, d_i^{\mathrm{c}})$ to $(\mathbb{R}^{m_i}, d_{\mathrm{E}}^{m_i})$.

Firstly,

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

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

Proof. A trivial modification of the proof to Proposition 3. □

Secondly,

$$f_i^{\mathrm{c}}(\underline{x}) = \begin{cases} \underline{0} \in \mathbb{R}^{m_i} & \text{if } \delta_i(\underline{x}) = \text{ false} \\ \omega_i\left(\underline{e_j} + (1 - \rho)\sum_{l \neq j}\underline{e_l}\right) & \delta_i(\underline{x}) = \text{ true and } x_i = v_{i,j}, \end{cases}$$

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

Proposition 6. *f_i^{c} is an isometry from $(\mathcal{X}, d_i^{\mathrm{c}})$ to $(\mathbb{R}^{m_i}, d_E^{m_i})$.*

Proof. Consider two inputs $\underline{x}, \underline{x}' \in \mathcal{X}$. As in the proof of Proposition 4, we need to show that $d_i^{\mathrm{c}}(\underline{x}, \underline{x}') = d_{\mathrm{E}}^{m_i}(f_i^{\mathrm{c}}(\underline{x}), f_i^{\mathrm{c}}(\underline{x}'))$ and consider the following cases.

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

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

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

$$d_{\mathrm{E}}^{m_i}(f_i^{\mathrm{c}}(\underline{x}), f_i^{\mathrm{c}}(\underline{x}')) = d_{\mathrm{E}}^{m_i}(\omega_i\, \underline{e_j}, \underline{0}) = \omega_i = d_i(\underline{x}, \underline{x}'),$$

and symmetrically for $d_{\mathrm{E}}(\underline{0}, \omega_i\, \underline{e_j})$.

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

$$d_{\mathrm{E}}^{m_i}(f_i^{\mathrm{c}}(\underline{x}), f_i^{\mathrm{c}}(\underline{x}')) \quad = \quad d_{\mathrm{E}}^{m_i}(\omega_i\, \underline{e_j}, \omega_i\, \underline{e_j}) = 0 = d_i^{\mathrm{c}}(\underline{x}, \underline{x}')$$

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

$$d_{\mathrm{E}}(f_i^{\mathrm{c}}(\underline{x}), f_i^{\mathrm{c}}(\underline{x}')) \quad = \quad d_{\mathrm{E}}^{m_i}(\omega_i\, \underline{e_j}, \omega_i \underline{e_{j'}}) = \omega_i\sqrt{2} = d_i^{\mathrm{c}}(\underline{x}, \underline{x}')$$

$\square$