Raiders of the Lost Architecture: A Kernel for Conditional Parameter Spaces

Anonymous Author(s)
Affiliation
Address
email

Abstract

When performing model-based optimization, we must often search over structures with differing numbers of parameters. For instance, we may wish to search over neural network architectures with an unkown number of layers. To combine information between different architectures, we define a family of kernels for conditional parameter spaces.

1 Introduction

[?]

2 A Kernel for Conditional Parameter Spaces

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

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

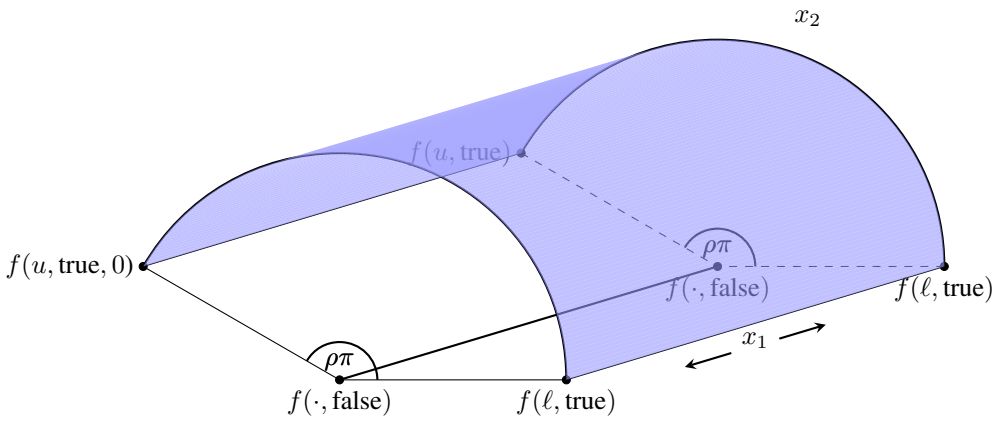

Figure 1: A demonstration of the embedding giving rise to the pseduo-metric: All points for which $\delta_i(x) = \text{false}$ are mapped onto a line depending only on x_1. Points for which $\delta_i(x) = \text{true}$ are mapped to the surfrace of a semicylinder, having an extra dimension x_2. This embedding gives a constant distance between pairs of points which have differing values of δ but the same values of x_1. The parameter ρ determines how much distance there is along the arc.

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 $d_i \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$: 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

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

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 **??**.

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

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 $\mathbb{R}^2, d_{\text{E}}$ 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 $\underline{x}$.

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

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

Although our formal arguments do not rely on this, Proposition **??** 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

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

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 ρ_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.

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 1. *Let κ be a positive semi-definite covariance function over Euclidean space. Then, $k_i \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, defined by*

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

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

A proof of proposition **??** can be found in the supplementary material.

Categorical Dimensions The pseduometric can also be extended to handle categorical dimensions - see the supplementary material.

3 Experiments

All the separate models split the data into 6 different datasets, one for each level, and build a separate model for each level. The gp-hierarchical model gets all the data together because it can handle it.

However, The gp-hierarchical model changes two things at once compared to the separate-gp-ard model: besides embedding the missing data in a different spot, it also has embeds the fully-observed data on semi-circles, and has a different parameterization. So, it could be the case that even when the data are fully observed, embedding the data on a semi-circle and using a different parameterization might cause better or worse performance than a standard squared-exp. To find out if this is the case, we compare separate-gp-ard and separate-hierarchical to find out if these two models have different performance even in the standard fully-observed case.

Table 1: Normalized Mean Squared Error

Method	NN	NN log	NN half	NN log half
Separate Linear	**0.968**	0.886	**1.039**	2.120
Separate GP	**0.925**	0.641	**0.860**	0.848
Poor Man's embedding Linear	**0.905**	0.763	0.996	0.851
Poor Man's embedding GP	**0.907**	0.518	**1.178**	**0.752**
Separate Hierarchical GP	**0.801**	0.627	**0.956**	0.950
Hierarchical GP	**0.801**	**0.441**	**0.993**	**0.674**

4 Conclusion

References

[1] C.E. Rasmussen and CKI Williams. Gaussian Processes for Machine Learning. *The MIT Press, Cambridge, MA, USA*, 2006.