Open problem: kernel methods on manifolds and metric spaces
what is the probability of a positive definite geodesic exponential kernel?
Feragen, Aasa; Hauberg, Søren

Published in:
29th Annual Conference on Learning Theory

Publication date:
2016

Document version
Publisher’s PDF, also known as Version of record

Document license:
Unspecified

Citation for published version (APA):
Open Problem: Kernel methods on manifolds and metric spaces
What is the probability of a positive definite geodesic exponential kernel?

Aasa Feragen
University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Søren Hauberg
Technical University of Denmark, Richard Petersens Plads, Bld. 321, 2800 Kgs. Lyngby, Denmark

Abstract
Radial kernels are well-suited for machine learning over general geodesic metric spaces, where pairwise distances are often the only computable quantity available. We have recently shown that geodesic exponential kernels are only positive definite for all bandwidths when the input space has strong linear properties. This negative result hints that radial kernel are perhaps not suitable over geodesic metric spaces after all. Here, however, we present evidence that large intervals of bandwidths exist where geodesic exponential kernels have high probability of being positive definite over finite datasets, while still having significant predictive power. From this we formulate conjectures on the probability of a positive definite kernel matrix for a finite random sample, depending on the geometry of the data space and the spread of the sample.

Keywords: Kernel methods, geodesic metric spaces, geodesic exponential kernel, positive definiteness, curvature, bandwidth selection.

1. Introduction
In a number of applications, learning can be improved by incorporating domain-specific knowledge that constrains the data to reside on a nonlinear subspace such as a Riemannian manifold. Examples include diffeomorphism groups in computational anatomy (Grenander and Miller, 1998) and probability distributions under the Fisher information metric (Amari and Nagaoka, 2000). In such nonlinear spaces, the geodesic distance is often the only computable quantity and methods that solely depend on pairwise distances are therefore practical. Manifold statistics (Fletcher and Joshi, 2004) are popular for such data, but in practice often suffer from poor numerical precision and do not even scale to medium-size data sets. Kernel methods (Schölkopf and Smola, 2002) provide an attractive alternative, where the most common kernel is the geodesic exponential kernel:

$$k(x, x') = \exp \left( -\lambda d^q(x, x') \right), \quad q \in \mathbb{R}_+,$$

where $d$ is the geodesic distance metric defined by shortest path length in $X$. These kernels can be defined on any geodesic data space, not just manifolds.

We have shown (Feragen et al., 2015) that Gaussian kernels ($q = 2$) are only PD for all $\lambda$ if the metric space is flat in the sense of Alexandrov (Bridson and Haefliger, 1999). Similar restrictions are shown for other $q$. This suggests that geodesic exponential kernels are not generally useful in nonlinear spaces. Below, however, we show that for finite datasets, there exist intervals of bandwidths $\lambda$ for Eq. 1 for which kernel matrices are PD. Moreover, in our simulations, these intervals contain bandwidths with good predictive power. This raises a series of open questions.
2. State-of-the-art

There is a vast literature on generalizing kernels to non-Euclidean data spaces, roughly falling into three approaches:

- The nonlinear data space $X$ is linearized and a kernel is designed over the linear approximation space (Courty et al., 2012; Jaakkola and Haussler, 1998). This discards the nonlinear structure of $X$ and thereby the domain specific knowledge it encodes.

- $X$ is embedded in a higher-dimensional Euclidean space, over which the kernel is designed (Harandi et al., 2014; Jayasumana et al., 2014). Since the distance measure in the embedding space can arbitrarily depart from the geodesic distance, this approach also discards the domain specific knowledge that $X$ was designed to capture.

- A kernel is designed directly on $X$ through the geodesic distance, e.g. using Eq. 1 (Chapelle et al., 1999; Jayasumana et al., 2015, 2013; Harandi and Salzmann, 2015). The fact that such kernels are not PD (Feragen et al., 2015) for all bandwidths is ignored, and the analysis proceeds as if the kernel were, in fact, PD. This strategy has recently attracted wide attention in computer vision.

The first two approaches discard the domain specific knowledge that the data space was supposed to encode, while the third approach violates the fundamental assumption made by kernel methods. As a consequence, the following statistical analysis is not guaranteed to be well-defined.

3. Open Problems

Losing the ”for all $\lambda > 0$” condition appears detrimental, because we lose the ability to freely train the bandwidth parameter $\lambda$. In practice, however, we find that there often exist large intervals of $\lambda$ parameters that give PD kernel matrices for concrete finite datasets. Fig. 1 shows two simulation studies of Gaussian kernels ($q = 2$ in Eq. 1) on the unit sphere. On the left, we test PD’ness by plotting the minimum eigenvalue of the kernel matrix $K = [k(x_i, x_j)]_{ij}$, which is positive if and only if $K$ is PD (Schölkopf and Smola, 2002). We observe a ”PD interval” of $\lambda$ parameters. This can be explained theoretically: When $\lambda \to \infty$, the kernel matrix approaches the identity matrix, whose minimum eigenvalue is 1. As the ”minimum eigenvalue” function is continuous, there must be some $\lambda'$ such that the kernel matrix $K$ is PD for $\lambda > \lambda'$. On the right panel of Fig. 1, the blue curve estimates how reliable this interval is by sampling 200 points from two distributions on the unit sphere, computing their Gaussian kernel matrix, and checking for PD’ness for a range of $\lambda$ parameters. We repeat this 300 times and plot the percentage of PD matrices for each $\lambda$ value. Note that the percentage quickly approaches 100.

These simulations suggest that our current theory must be refined, and we conjecture:

**Conjecture 1** There exist conditions on the geometry of the data space $X$, the spread of the data, the exponent $q \leq 2$, the PD range of $\lambda$ parameters and the sample size $N$ such that for a random sample $\{x_1, \ldots, x_N\} \subset X$, and a fixed $\varepsilon > 0$, the kernel matrix $[\exp(-\lambda d^q(x_i, x_j))]_{ij}$ is PD with probability $1-\varepsilon$.

Conjecture 1 would be very powerful as it allows optimizing $\lambda$ for specific tasks, and gives insight into the expected generalizability of a PD interval to an unseen test set. Moreover, we conjecture:
Fig. 1 We show two simulation experiments on the unit sphere. **Left:** The minimal eigenvalue of the geodesic Gaussian kernel matrix for 100 uniformly sampled data points. **Right:** We compute the kernel matrix for a dataset with 100 points drawn from each of the Gaussian distributions $\mathcal{N}(\{1/2, 1/2, 1/2\}, I)$ and $\mathcal{N}(\{-1/2, 1/2, 1/2\}, I)$ projected onto the unit sphere. This experiment is repeated 300 times for each of a range of bandwidths $\lambda$, and we report the percentage of PD geodesic Gaussian kernel matrices (blue), and the average cross-validated classification accuracy with a support vector machine among those kernel matrices that are PD (yellow). For reference, the expected classification accuracy for the optimal separating hyperplane of the generating distributions are shown in red.

**Conjecture 2** The PD range of $\lambda$ parameters depends on the geometry of the data space $X$:

a) The PD range depends on the curvature of $X$, defined in the $\text{CAT}(\kappa)$ sense (Bridson and Haefliger, 1999).

b) The PD range depends on the distortion of the metric of the nonlinear data space under a low-distortion embedding into Euclidean spaces.

Conjecture 2, along with simple tests for curvature and recent results on embeddability into Euclidean spaces (Sidiropoulos and Wang, 2015; Matousek and Sidiropoulos, 2010), would lead to simple, testable conditions for positive definiteness, or for the size of the positive definite range of $\lambda$ parameters. In particular, Conjecture 2b may lead to simple tests for whether lacking PD’ness is a general property of the data, or caused by a select few outliers.

If the bounding $\lambda'$ is too large, then all PD kernel matrices are very close to $I$ and therefore non-informative. However, our second simulation shown by the yellow curve in Fig. 1 shows that there are bandwidth parameters that simultaneously lead to a high probability of a PD kernel matrix, and a close to optimal classification accuracy.

**Conjecture 3** The PD range of $\lambda$ parameters is useful, and its usefulness depends on the power $q$.

Clearly there is a dependence on $q$ because, as when $q \to 0$, the kernel matrix approaches a matrix of ones, which is non-informative. Conjecture 3 seeks to quantify this dependence.

What sets these conjectures apart from the current mindset is the formulation in terms of probability of positive definiteness in place of a deterministic worst-case analysis. We believe this probabilistic analysis will be intimately connected with the geometry of the data space.

**Acknowledgments**

We thank Oswin Krause for insightful discussions. S.H. is funded by the Danish Council for Independent Research, Natural Sciences.
References


