Towards Indefinite Gaussian Processes

Murat Seckin Ayhan
Center for Advanced Computer Studies
University of Louisiana at Lafayette
msa4307@cacs.louisiana.edu

Chee-Hung Henry Chu
Center for Advanced Computer Studies
University of Louisiana at Lafayette
cice@cacs.louisiana.edu

Abstract

Gaussian processes (GPs) enable probabilistic kernel-machines with remarkable modeling efficacy and GPML toolbox facilitates a widespread use by practitioners and researchers. Many modern applications demand *non-metric* (dis)similarities. As a result, Mercer's condition for positive semidefiniteness is violated. Through a simple text categorization example that involves a KL-divergence based kernel function, we have demonstrated that, despite all the care taken for numerical stability, the current framework is *vulnerable to indefiniteness*. Learning a spectral transformation is an option to tackle the problem with. However, the need for a general and principled solution towards indefinite Gaussian processes is *urgent*.

1 Introduction

GPML toolbox [16] facilitates the plug-and-play use of many covariance (kernel) functions. As such, covariance functions must be *isotropic* and a kernel matrix corresponding to a covariance function is *positive semidefinite* (PSD) [17]. A PSD kernel guarantees the uniqueness of a *reproducing kernel Hilbert space* (RKHS). However, in practice, indefinite (non-PSD) kernels do occur. To utilize an indefinite kernel, one can reformulate an algorithm so that an indefinite kernel matrix can be effectively learned [9, 2, 19, 15]. Alternatively, spectral transformations [8, 7, 3, 4] suggest relatively easy corrections on indefinite kernel matrices so that existing algorithms can be used as usual.

To combine both generative and discriminative methods, a kernel that exploits the Kullback-Leibler (KL) divergence has been introduced for SVMs [12]. Language models are also generative. In [18], a language model-based kernel was proposed for an information retrieval framework. However, we only consider language model objects and their mappings to feature spaces for classification.

GPML toolbox [16] contains many careful steps in order to achieve numerical stability and computational efficiency. One stability measure particularly addresses the issue of indefiniteness, and effectively applies a spectral transformation. However, our experiments, which utilize a KL-divergence based kernel, show that GP learning can be easily hindered by indefiniteness. In this regard, our work focuses on reinforcing the stability of the toolbox with minimal sacrifice on computational efficiency. We achieve this through Bayesian model selection. Since spectral transformations are learned with respect to kernel hyperparameters, the existing algorithms remain unchanged.

2 Gaussian Processes and Covariance Functions

In GPs terminology, a kernel is a covariance function that estimates the covariance of two latent variables $f(\mathbf{x})$ and $f(\mathbf{z})$ in terms of input vectors $\mathbf{x}$ and $\mathbf{z}$, respectively. These functions are parameterized by *hyperparameters*. A well-known and widely used example is the *squared-exponential* (SE) covariance function [17]:

$$k_{SE}(\mathbf{x},\mathbf{z}) = \sigma_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{z})^T M(\mathbf{x}-\mathbf{z})}{2}\right), \tag{1}$$

where σ_f^2 is the signal variance. Depending on a diagonal matrix M, $k_{SE}(\mathbf{x}, \mathbf{z})$ can carry out a Gaussian-shaped kernel ($M = \ell^{-2}I$), or implement *Automatic Relevance Determination* (ARD) ($M = diag\,(\boldsymbol{\ell})^{-2}$) [17, 10, 13], where $\boldsymbol{\ell} = [\ell_1, \ell_2, \cdots, \ell_D]$ is a vector of length-scales.

Information processing capabilities of GPs are mostly determined by the choice of covariance function. Moreover, the impact of covariance function is larger for small to medium-sized datasets [6].

3 KL-divergence based Covariance Function

The KL-divergence measures the distance between probability distributions. It is reflexive, non-negative but *anisotropic*. We use the isotropic KL-divergence: $KL_{iso}(\mathbf{x}||\mathbf{z}) = KL(\mathbf{x}||\mathbf{z}) + KL(\mathbf{z}||\mathbf{x})$. It is *non-Euclidean*. Non-Euclidean measures can be informative and may enable the construction of good classifiers [4]. By choosing $M = \ell^{-2}I$ and rewriting the exponent in eq.1 with respect to $KL_{iso}(\mathbf{x}||\mathbf{z})$, we obtain an isotropic covariance function:

$$k_{KL_{iso}}(\mathbf{x}, \mathbf{z}) = \sigma_f^2 \exp\left(-\frac{KL_{iso}(\mathbf{x}||\mathbf{z})}{2\ell^2}\right). \tag{2}$$

However, this covariance function *does not* give rise to PSD kernel matrices.

Opposition to Earlier KL-divergence based Kernels: In [12, 18], kernels are described as $\ddot{k}_{KL_{iso}}(\mathbf{x}, \mathbf{z}) = \exp\left(-A \times KL_{iso}(\mathbf{x}, \mathbf{z}) + B\right)$, where A and B are scale and shift factors. In [18], these are fixed; $A = 1$ and $B = 0$. Even then the indefiniteness is in effect and the fixation of parameters is strictly at odds with the essence of GP learning. Learning in GPs is equivalent to finding suitable parameters for the covariance function.

4 Pseudo-Euclidean Space and Spectral Transformations

A *pseudo-Euclidean* space $pE = \mathbb{R}^{(p,q)}$ is a vector space endowed with an indefinite inner product $\langle \cdot, \cdot \rangle_{pE}$. The signature of pE is (p, q), where $p, q \in \mathbb{N}_0$. The inner product $\langle \cdot, \cdot \rangle_{pE}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$. This gives rise to p positive and q negative eigenvalues [4].

Ref. [7] analyses four common spectral transformations: denoise, flip, diffusion and shift. These are essentially the mappings that can be written as $\lambda_{new} = f(\lambda)$, where λ is an eigenvalue. denoise treats negative eigenvalues as noise, $f(\lambda) = max(\lambda, 0)$. flip flips the sign of negative eigenvalues, $f(\lambda) = |\lambda|$. diffusion applies a matrix diffusion [2], $f(\lambda) = e^{\beta\lambda}$. shift shifts all eigenvalues by a positive constant, $f(\lambda) = \lambda + \eta$.

A negative space can be informative for class separation [7, 3, 4]. Denoise, flip and diffusion present higher risks at damaging the semantics of (dis)similarity [7]. On the other hand, shift is more conservative. It only changes the *diagonal* elements in an $N \times N$ kernel matrix K, and better generalizes to unseen data [7].

In regards to computational cost, denoise, flip and diffusion require the diagonalization of K. On the other hand, shift can benefit from a nice property (eq.3) that allows us to estimate η:

$$-\sqrt{\|K\|_F^2 - \frac{\max(\|K\|_{col}^2, \|K\|_{row}^2)}{N}} \leq \lambda_N \leq \max(0, \frac{\mathrm{tr}(K)}{N}), \tag{3}$$

where $\|K\|_{col}^2 = \max_{1 \leq j \leq N} \sum_{i=1}^{N} |k_{ij}|$ and $\|K\|_{row}^2 = \max_{1 \leq i \leq N} \sum_{j=1}^{N} |k_{ij}|$. Any η that is greater than $|\lambda_N|$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$, induces an Euclidean correction on K [7].

In [7], η is considered as a regularization term in the context of SVMs and is empirically shown to have an impact on support vectors and margins. Further, it is suggested that C can be fixed while performing cross validation, and η should be used, instead. As a result, shift does not incur any additional computational cost [7] and exact approximation is not needed. The intricate balance between semantics preservation and noise removal is ultimately set [7].

shift in GPML Toolbox: shift has already been utilized in GPML toolbox [16], such that an identity matrix I is added to the matrix of interest[1]. Even though this is pretty safe for many PSD covariance

[1] $B = I + \tilde{S}^{\frac{1}{2}} K \tilde{S}^{\frac{1}{2}}$ for Expectation Propagation (EP) algorithm. $\tilde{S}$ is induced by the algorithm itself.

functions [17], noisy eigenvalues can be magnified and poorly translated to the positive space. In addition, during the Bayesian model selection, the signal variance σ_f^2 can grow arbitrarily large and void this *additive* measure.

Learning to shift: Non-parametric methods, such as GPs, are usually criticized for their (in)ability to scale to big data. In this regard, the degree and computational burden of spectral transformations must be addressed simultaneously, which cannot be achieved by merely adding ones.

The basic complexity of GP learning is $O(N^3)$ and the computational overhead per hyperparameter is $O(N^2)$ [17]. Therefore, spectral transformations can be achieved with a little extra effort. To this end, we use a variant of the covariance function: $\widehat{k}_{KL_{iso}}(\mathbf{x}, \mathbf{z}) = k_{KL_{iso}}(\mathbf{x}, \mathbf{z}) - \eta\delta_{\mathbf{xz}}$, where η is a regularizer and $\delta_{\mathbf{xz}}$ is a Kronecker delta which is one iff $\mathbf{x} = \mathbf{z}$, and zero otherwise. Our formulation aims to recover the semantics damaged due to the addition of I, and penalizes large values of σ_f^2. Moreover, the Expectation Propagation (EP) implementation remains untouched, which has been a preference.

5 Experiments

We have demonstrated the impact of the proposed method for spectral transformation on text categorization task. Expectation Propagation (EP) [11] and probit likelihood are used for GP inference. Learning of hyperparameters is achieved by maximizing marginal likelihoods on training sets and using L-BFGS [14] for 100 iterations. Performances are evaluated via 10-fold cross validation.

We used a subset obtained from the well-known Reuters21578[2] collection and made available by [1]. Language models are obtained using two common methods: i) *additive* smoothing and ii) *Good-Turing* probability estimation. For simplicity, unigram models are used and it is assumed that words occur independently. Experiments include 4 categorization tasks as shown in Table 1.

Table 1: Summary of binary classification problems

	Coffee/Gold	Ship/Sugar	MoneyFX/Interest	Crude/Trade
Number of words	3629	4310	7608	7608
Size of negative class {-1}	110	142	245	321
Size of positive class {+1}	90	114	197	298

5.1 Experimental Results

Table 2 and Table 3 present the results obtained by additive smoothing and Good-Turing probability estimation, respectively. Results in columns are the averages of 10 classification tasks. N/A indicates that at least one task has failed. Thus, no (average) results are reported.

Table 2 shows that training procedures without the regularizer η severely suffer from the indefiniteness, especially for the first and third datasets. Even though the successful completions (2^{nd} and 4^{th} columns) of cross validation, the σ_f values are quite disturbing. With such large numbers, even the noisy eigenvalues around zero can be so magnified that they cancel I out. On the other hand, η keeps σ_f from growing large, as well as prevents the magnification of eigenvalues. Consequently, significantly more likely models (Table 2, in **bold**) are obtained and the classification performances are equally good. Further, η enables the successful completions of cross validation in all cases. Also note that, the use of η promotes model complexity since the induced length-scales (ℓ) are smaller.

Good-Turing estimates help in finding reasonable parameters. For instance, the large σ_f phenomenon, even in the absence of η, disappears (Table 3). However, Table 3 still indicates a failure of completion in its last column. On the other hand, η is still successful at penalizing σ_f and at promoting cross validation completions. There is no significant difference between corresponding[3] classification performances given in Table 3. However, η promotes model complexity, again.

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[3] Except the last column.

Table 2: Results of cross validation with additive smoothing

		Coffee / Gold	Ship / Sugar	MoneyFX / Interest	Crude / Trade
$k_{KL_{iso}}(\mathbf{x}, \mathbf{z})$	Hyperparameters	N/A N/A	$\ell = 0.61^a$ $\sigma_f = 1.38e{+}11$	N/A N/A	$\ell = 0.85$ $\sigma_f = 7.53e{+}09$
	Neg. log lik.	N/A	54.48±1.50	N/A	74.66±1.15
	Training acc.	N/A	100.00±0.00	N/A	100.00±0.00
	Test acc.	N/A	99.20±1.69	N/A	99.02±1.15
$\widehat{k}_{KL_{iso}}(\mathbf{x}, \mathbf{z})$	Hyperparameters	$\ell = 0.44$ $\sigma_f = 6.45$ $\eta = 1.3534$	$\ell = 0.41$ $\sigma_f = 5.98$ $\eta = 1.2372$	$\ell = 0.38$ $\sigma_f = 3.91$ $\eta = 0.9831$	$\ell = 0.45$ $\sigma_f = 7.10$ $\eta = 1.4411$
	Neg. log lik.	31.54±0.46	**51.04±0.36**	128.30±2.06	**65.20±1.18**
	Training acc.	100.00±0.00	100.00±0.00	99.72±0.14	100.00±0.00
	Test acc.	99.50±1.58	99.60±1.26	89.07±3.96	99.02±1.15

[a] The average out of first 9 trainings; the last length-scale ℓ was 148.02.

Table 3: Results of cross validation with Good-Turing probability estimation

		Coffee / Gold	Ship / Sugar	MoneyFX / Interest	Crude / Trade
$k_{KL_{iso}}(\mathbf{x}, \mathbf{z})$	Hyperparameters	$\ell = 4.86$ $\sigma_f = 7.68$	$\ell = 4.70$ $\sigma_f = 3.21$	$\ell = 3.63$ $\sigma_f = 2.73$	N/A N/A
	Neg. log lik.	51.96±4.10	108.79±3.39	167.17±5.57	N/A
	Training acc.	98.44±0.35	97.92±0.34	93.43±1.01	N/A
	Test acc.	96.00±3.94	96.80±2.53	86.74±5.15	N/A
$\widehat{k}_{KL_{iso}}(\mathbf{x}, \mathbf{z})$	Hyperparameters	$\ell = 4.07$ $\sigma_f = 2.72$ $\eta = 0.7908$	$\ell = 4.56$ $\sigma_f = 2.29$ $\eta = 0.4319$	$\ell = 2.88$ $\sigma_f = 1.89$ $\eta = 0.2838$	$\ell = 7.27$ $\sigma_f = 5.41$ $\eta = 0.7930$
	Neg. log lik.	52.01±4.12	108.88±3.42	167.23±5.55	94.74±8.37
	Training acc.	98.44±0.35	97.97±0.29	93.66±1.07	99.66±0.13
	Test acc.	96.00±3.94	97.20±1.93	86.98±5.39	97.70±1.76

5.2 Discussion

The desired degree and efficiency of spectral transformations can be achieved simultaneously. However, the problem remains ill-conditioned since we utilize the existing framework, which is valid for RKHS. The ultimate drawback is that a solution is *not guaranteed*. Consequently, we resort to heuristics[4], which may become tricky for larger matrices and different degrees of non-Euclidean influence. In such cases, where very large negative eigenvalues are involved, we conjecture that I matrix should be scaled up and η would handle the regularization.

A spectral transformation may induce a *many-to-one* mapping and increase the class overlap [3]. Non-Euclidean influence is usually greater when Good-Turing estimates are used. Compared to Table 2, class overlaps may be a reason for the decrease in classification performance in Table 3.

6 Conclusion

GPs enable powerful tools for solving regression and classification tasks [6, 17], and the GPML toolbox [16] facilitates a widespread use by practitioners and researchers. Many modern applications demand indefinite kernels [7, 9, 2], as they represent an opportunity to include specific application knowledge [5]. Our work has revealed the vulnerability of the current framework for GP learning in this regard. Learning a spectral transformation is handy. However, the need for a general and principled solution towards indefinite Gaussian processes is *urgent*.

[4]Initial hyperparameters: $\sigma_f = 1$, $\ell = 1$ and $\eta = 0.5 - |\lambda_N|$. λ_N is computed exactly once w.r.t. eq.2.

References

[1] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 105–112, 2009.

[2] J. Chen and J. Ye. Training svm with indefinite kernels. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 136–143, New York, NY, USA, 2008. ACM.

[3] R.P. Duin, E. Pękalska, A. Harol, W.J. Lee, and H. Bunke. On euclidean corrections for non-euclidean dissimilarities. In *Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, SSPR & SPR'08, pages 551–561, Berlin, Heidelberg, 2008. Springer-Verlag.

[4] R.P.W. Duin and E. Pękalska. Non-euclidean dissimilarities: causes and informativeness. In *Proceedings of the 2010 joint IAPR international conference on Structural, syntactic, and statistical pattern recognition*, SSPR & SPR'10, pages 324–333, Berlin, Heidelberg, 2010. Springer-Verlag.

[5] R.P.W. Duin and E. Pękalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recogn. Lett.*, 33(7):826–832, May 2012.

[6] D. Duvenaud, H. Nickisch, and C.E. Rasmussen. Additive Gaussian processes. In *Advances in Neural Information Processing Systems 25*, pages 1–8, Granada, Spain, December 2011.

[7] W. Gang, Y.C. Edward, and Z. Zhihua. An analysis of transformation on non-positive semidefinite similarity matrix for kernel machines. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[8] B. Haasdonk. Feature space interpretation of svms with indefinite kernels. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(4):482 –492, april 2005.

[9] R. Luss and A. D'Aspremont. Support vector machine classification with indefinite kernels. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 953–960. MIT Press, Cambridge, MA, 2008.

[10] D.J.C. MacKay. Bayesian methods for backpropagation networks. In J.L. van Hemmen, E. Domany, and K. Schulten, editors, *Models of Neural Networks II*, chapter 6. Springer, 1993.

[11] T.P. Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[12] P.J. Moreno, P.P. Ho, and N. Vasconcelos. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[13] R.M. Neal. Bayesian learning for neural networks. In *Lecture Notes in Statistics*. Springer, New York, 1996.

[14] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35:773–782, 1980.

[15] C.S. Ong, X. Mary, S. Canu, and A.J. Smola. Learning with non-positive kernels. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 81–, New York, NY, USA, 2004. ACM.

[16] C.E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research*, 11:3011–3015, 2010.

[17] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, second printing edition, 2006.

[18] Y. Xie and V.V. Raghavan. Language-modeling kernel based approach for information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 58:2353–2365, December 2007.

[19] Y Ying, Colin Campbell, and M Girolami. Analysis of svm with indefinite kernels. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2205–2213. 2009.