Natural Regularization in SVMs

Nuria Oliver[*,†]
nuria@media.mit.edu

Bernhard Schölkopf[*]
bsc@microsoft.com

Alex Smola[*]
Alex.Smola@anu.edu.au

[†]**Media Arts and Sciences Laboratory**
MIT, 20 Ames Street, E15-385
Cambridge, MA 02139
http://www.media.mit.edu/∼nuria

Microsoft Research Limited
St. George House, 1 Guildhall Street
Cambridge CB2 3NH, UK
http://www.research.microsoft.com/∼bsc/

Department of Engineering
Australian National University
Canberra 0200 ACT, Australia
http://spigot.anu.edu.au/∼smola/

Abstract

Recently the so called Fisher kernel was proposed by [6] to construct discriminative kernel techniques by using generative models. We provide a regularization-theoretic analysis of this approach and extend the set of kernels to a class of natural kernels, all based on generative models with density $p(\mathbf{x}|\theta)$, like the original Fisher kernel. This allows us to incorporate distribution dependent smoothness criteria in a general way.

As a result of this analyis we show that the Fisher kernel corresponds to a $L_2(p)$ norm regularization. Moreover it allows us to derive explicit representations of the eigensystem of the kernel, give an analysis of the spectrum of the integral operator, and give experimental evidence that this may be used for model selection purposes.

1 Introduction

There has been significant progress on Learning Theory using discriminative and generative models over the last decade. Generative techniques such as HMMs, dynamic graphical models, or mixtures of experts have provided a principled framework for dealing with missing and incomplete data, uncertainty or variable length sequences. On the other hand, discriminative models like SV Machines [2] and other kernel methods (Gaussian Processes [12], Regularization Networks [5], etc.) have become standard tools of applied machine learning technology, leading to record benchmark results in a variety of domains. However, until recently, these two strands have been largely separated.

A promising approach to combine the strengths of both worlds was made in the work of [6]. The main idea is to design kernels inspired by generative models. In particular, they propose to use a so called Fisher kernel to give a 'natural' similarity measure taking into account an underlying probability distribution.

Since defining a kernel function automatically implies assumptions about metric relations between the examples, they argue that these relations should be defined directly from a generative probability model $p(\mathbf{x}|\theta)$, where θ are the parameters of the model. Their choice is justified from two perspectives: that of improving the discriminative power of the model and from an attempt to find a 'natural' comparison between examples induced by the generative model.

While this is quite an abstract concept, it would be desirable to obtain a deeper understanding of the regularization properties of the resulting kernel. In other words, it would be instructive to see which sort of functions such a kernel favours, which degrees of smoothness are chosen, or how categorical data are treated. Many of these properties can be seen by deriving the regularization operator (with the associated prior) [9] to which such a kernel corresponds to.

The paper is structured as follows. In section 2 we introduce tools from information geometry and define a class of *natural kernels* to which also the two kernels proposed by [6] belong. A regularization theoretic analysis of natural kernels follows in section 3. In particular we show that the so called Fisher kernel corresponds to a prior distribution over the functions $f(\theta)$ taking the form $p(f) \propto \exp\left(-\frac{1}{2}\|f\|_p^2\right)$, where $\|\cdot\|_p^2$ is the norm of the $L_2(p)$ space of functions square integrable wrt. the measure corresponding to $p(x|\theta)$, i.e. the usual norm weighted by the underlying generative model. Finally, in section 4 we derive the decomposition of natural kernels into their eigensystem which allows to describe the image of input space in feature space. The shape of the latter has consequences for the generalization behavior of the associated kernel method (cf. e.g. [13]). Section 5 concludes the paper with some experiments and a discussion.

2 Natural Kernels

Many learning algorithms can be formulated in terms of dot products between input patterns $x_i \cdot x_j$ in $\mathbb{R}^N$, such as separating hyperplanes for pattern recognition. Using a suitable kernel k instead of a dot product in $\mathbb{R}^N$ corresponds to mapping the data into a possibly high-dimensional dot product space F by a (usually nonlinear) feature map $\Phi : \mathbb{R}^N \to F$, and taking the dot product there (cf. e.g. [2])

$$k(x, x') = (\Phi(x) \cdot \Phi(x')). \tag{1}$$

Any linear algorithm which can be cast in terms of dot products can be made nonlinear by substituting an a priori chosen kernel for the dot product. Examples thereof are SV Machines and kernel PCA [2, 8]. The solutions of kernel algorithms are kernel expansions

$$f(x) = \sum_i \alpha_i k(x, x_i). \tag{2}$$

Since all the computations are done in terms of dot products, all information used for training, based on patterns $x_1, \ldots, x_\ell \in \mathcal{X}$ (usually $\mathcal{X} \subset \mathbb{R}^N$) resides in the Gram matrix

$$K_{ij} := k(x_i, x_j), \tag{3}$$

and in target values which might be provided additionally.

These conventional SV kernels ignore knowledge of the underlying distribution of the data $p(x)$ which could be provided by a generative model or additional information about the problem at hand. Instead, a general requirement of smoothness is imposed [4, 10]. This may not always be desirable, e.g. in the case of categorical data (attributes such as english, german, spanish, ...) and sometimes one may want to enforce a higher degree of smoothness where data is sparse, and less smoothness where data is abundant. Both issues will be addressed in the following.

To introduce a class of kernels derived from generative models, we need to introduce basic concepts of information geometry. Consider a family of generative models $p(\mathbf{x}|\theta)$ (i.e. probability measures) smoothly parametrized by θ. These models form a manifold (also called statistical manifold) in the space of all probability measures. The key idea introduced by [6] is to exploit the geometric structure on this manifold to obtain an (induced) metric for the training patterns $\mathbf{x}_i$. Rather than dealing with $p(\mathbf{x}|\theta)$ directly one uses the log-likelihood instead, i.e. $l(\mathbf{x}, \theta) := \ln p(\mathbf{x}|\theta)$.

- The derivative map of $l(\mathbf{x}|\theta)$ is usually called the **score map** $U_\theta : \mathcal{X} \to \mathbb{R}^r$ with

  $$U_\theta(\mathbf{x}) := (\partial_{\theta^1} l(\mathbf{x}, \theta), \ldots, \partial_{\theta^r} l(\mathbf{x}, \theta)) = \nabla_\theta l(\mathbf{x}, \theta) = \nabla_\theta \ln p(\mathbf{x}|\theta), \tag{4}$$

  whose coordinates are taken as a 'natural' basis of tangent vectors. Note that θ is the coordinate system for any parametrization of the probability density $p(\mathbf{x}|\theta)$.

  For example, if $p(\mathbf{x}|\theta)$ is a normal distribution, one possible parametrization would be $\theta = (\mu, \sigma)$, where μ is the mean vector and σ is the covariance matrix of the Gaussian. The basis given by the score map represents the direction in which the value of the ith coordinate increases while the others are fixed.

- Since the manifold of $\ln p(\mathbf{x}|\theta)$ is Riemannian, there is an inner product defined in its tangent space T_p whose metric tensor is given by the **Fisher information matrix**

  $$I(p) := E_p \left[U_\theta(\mathbf{x}) U_\theta(\mathbf{x})^\top \right] \text{ i.e. } I_{ij}(p) = E_p \left[\partial_{\theta^i} \ln p(\mathbf{x}|\theta) \partial_{\theta^j} \ln p(\mathbf{x}|\theta) \right]. \tag{5}$$

  Here E_p denotes the expectation with respect to the density p.

- This metric is called the **Fisher information metric** and induces a 'natural' distance in the manifold. It can be used to measure the difference in the generative process between a pair of examples $\mathbf{x}_i$ and $\mathbf{x}_j$ via the score map $U_\theta(\mathbf{x})$ and I^{-1}.

Note that the metric tensor, i.e. I_p, depends on p and therefore on the parametrization θ. This is different to the conventional Euclidean metric on $\mathbb{R}^n$ where the metric tensor is simply the identity matrix. For the purposes of calculation it is often easier to compute I_{ij} as the Hessian of the scores:

$$I(p) = -E_p \left(\nabla_\theta \nabla_\theta^\top \ln p(\mathbf{x}|\theta) \right) \text{ with } I_{ij}(p) = -E_p \left(\partial_{\theta_i} \partial_{\theta_j} \ln p(\mathbf{x}|\theta) \right) \tag{6}$$

In summary, what we need is a family of probability measures for which the log-likelihood $l(\mathbf{x}, \theta) = \ln p(\mathbf{x}|\theta)$ is a differentiable map.

Definition 1 (Natural Kernel) *Denote by M a positie definite matrix and by $U_\theta(x)$ the score map defined above. Then the corresponding natural kernel is given by*

$$k_M^{\mathrm{nat}}(\mathbf{x}, \mathbf{x}') := U_\theta(\mathbf{x})^\top M^{-1} U_\theta(\mathbf{x}') = \nabla_\theta \ln p(\mathbf{x}|\theta)^\top M^{-1} \nabla_\theta \ln p(\mathbf{x}'|\theta) \tag{7}$$

In particular, if $M = I$, hence k_I^{nat}, the (7) reduces to the ***Fisher kernel*** *[6]. Moreover if $M = \mathbf{1}$ one obtains a kernel we will call the* ***plain kernel*** *which is often used for convenience if I is too difficult to compute.*[1]

In the next section, we will give a regularization theoretic analysis of the class of natural kernels, hence in particular of k_I^{nat} and $k_{\mathbf{1}}^{\mathrm{nat}}$. This answers the question to which type of smoothness (or rather 'simplicity') the kernels proposed in [6] correspond to.

3 The Natural Regularization Operator

In SV machines one minimizes a regularized risk functional $R_{reg}[f]$, which is the weighted sum of empirical risk functional $R_{emp}[f]$ and regularization or complexity term $Q[f]$ (8)

$$R_{reg}[f] = R_{emp} + \lambda Q[f] \tag{8}$$

where the complexity term can be written as $\frac{\lambda}{2}||w||^2$ in feature space notation, or as $\frac{\lambda}{2}||Pf||^2$ when considering the functions in input space directly. In particular, the connection between kernels k, feature spaces F and regularization operators P is given by (9):

$$k(\mathbf{x}_i, \mathbf{x}_j) = ((Pk)(\mathbf{x}_i, .) \cdot (Pk)(\mathbf{x}_j, .)) \, . \tag{9}$$

It states that if k is a Greens function of P^*P, minimizing $||w||$ in feature space is equivalent to minimizing the regularized risk functional given by $||Pf||^2$.

To analyze the properties of natural kernels k_I^{nat}, we exploit this connection between kernels and regularization operators by finding the operator P_M^{nat} such that (9) holds. To this end, we need to specify a dot product in (9). Note that this is part of the choice of the class of regularization operators that we are looking at — in particular, the choice is a choice of the dot product space that P maps into. We opt for the dot product in $L_2(p)$ space, i.e.

$$\langle f, g \rangle := \int f(\mathbf{x}) g(\mathbf{x}) p(\mathbf{x}|\theta) d\mathbf{x} \tag{10}$$

since this will lead to a simple form of the corresponding regularization operators. Other measures would also have been possible, leading to different formal representations of P.

Proposition 2 (Regularization Operators for Natural Kernels) *Given a positive definite matrix M, a generative model $p(\mathbf{x}|\theta)$, and a corresponding natural*

[1] For the sake of correctness one would have to write $k_{M,p(\mathbf{x},\cdot)}^{\mathrm{nat}}$ rather than k_M^{nat} since k also depends on the generative model and the parameter θ chosen by some other procedure such as density estimation. Moreover note that rather than requiring M to be positive definite, semidefiniteness would be sufficient. However, then, we would have to replace M^{-1} by the pseudoinverse and the subsequent reasoning would be significantly more cumbersome.

kernel $k_M^{\text{nat}}(\mathbf{x},\mathbf{x}')$, P_M^{nat} is an equivalent regularization operator if it satisfies the following condition:

$$M = \int \left[P_M^{\text{nat}} \nabla_\theta \ln p(\mathbf{z}|\theta)\right] \left[P_M^{\text{nat}} \nabla_\theta \ln p(\mathbf{z}|\theta)\right]^\top p(\mathbf{z}|\theta) d\mathbf{z} \tag{11}$$

Proof Substituting (7) into (9) yields

$$\begin{aligned}
k_M^{\text{nat}}(\mathbf{x},\mathbf{x}') &\overset{\text{by\_def}}{=} \nabla_\theta \ln p(\mathbf{x}|\theta)^\top M^{-1} \nabla_\theta \ln p(\mathbf{x}'|\theta) && (12)\\
&\overset{(9)}{=} \left\langle P_M^{\text{nat}} k_M^{\text{nat}}(\mathbf{x},\mathbf{z}), P_M^{\text{nat}} k_M^{\text{nat}}(\mathbf{x}',\mathbf{z}) \right\rangle && (13)\\
&= \int \nabla_\theta \ln p(\mathbf{x}|\theta)^\top M^{-1} \left[P_M^{\text{nat}} \nabla_\theta \ln p(\mathbf{z}|\theta)\right] \times \\
&\qquad \left[P_M^{\text{nat}} \nabla_\theta \ln p(\mathbf{z}|\theta)^\top\right] M^{-1} \nabla_\theta \ln p(\mathbf{x}'|\theta) p(\mathbf{z}|\theta) d\mathbf{z} && (14)
\end{aligned}$$

Note that P_M^{nat} acts on p as a function of $\mathbf{z}$ only — the terms in $\mathbf{x}$ and $\mathbf{x}'$ are not affected which is why we may collect them outside. Thus the necessary condition (11) ensures that the rhs (13) equals (14) which completes the proof. $\blacksquare$

Let us consider the two special cases proposed by [6].

Corollary 3 (Fisher Kernel) *The Fisher Kernel ($M = I$) induced by a generative probability model with density p corresponds to a regularizer equal to the squared $L_2(p)$-norm of the estimated function. Therefore the regularization term is given by*

$$||Pf||^2 = ||f||^2_{L_2(p)}. \tag{15}$$

This can be seen by substituting in $P_I^{\text{nat}} = \mathbf{1}$ into the rhs of (11) which yields the definition of the Fisher information matrix.

To get an intuition about what this regularizer does, let us spell it out explicitly. The solution of SV regression using the Fisher kernel has the form $f(\mathbf{x}) = \sum_{i=1}^{\ell} \alpha_i k_I^{\text{nat}}(\mathbf{x},\mathbf{x}_i)$, where the $\mathbf{x}_i$ are the SVs, and $\boldsymbol{\alpha}$ is the solution of the SV programming problem. Applied to this function, we obtain

$$\begin{aligned}
||f(\theta)||^2_{L_2(p)} &= \int |f(\mathbf{x})|^2 p(\mathbf{x}|\theta) d\mathbf{x} && (16)\\
&= \int \left(\sum\nolimits_i \alpha_i \nabla_\theta \ln p(\mathbf{x}|\theta) I^{-1} \nabla_\theta \ln p(\mathbf{x}_i|\theta)\right)^2 p(\mathbf{x}|\theta) d\mathbf{x}.
\end{aligned}$$

To understand this term, first recall that what we actually minimize is the regularized risk $R_{\text{reg}}[f]$, the sum of (16) and the empirical risk given by the normalized negative log likelihood. The regularization term (16) prevents overfitting by favoring solutions with smaller $\nabla_\theta \ln p(\mathbf{x}|\theta)$. Consequently, the regularizer will favor the solution which is more stable (flat). Figure 1 illustrates this effect.

Note, however, that the validity of this intuitive explanation is somewhat limited since some effects can compensate each other as the α_i come with different signs. Finally, we remark that the regularization operator of the conformal transformation [1] of the Fisher kernel k_I^{nat} into $\sqrt{p(\mathbf{x}|\theta)}\sqrt{p(\mathbf{x}'|\theta)} k_I^{\text{nat}}(\mathbf{x},\mathbf{x}')$ is the identity map in L_2 space.

In practice, [6] often use $M = \mathbf{1}$. In this case, proposition 2 specializes to the following result.

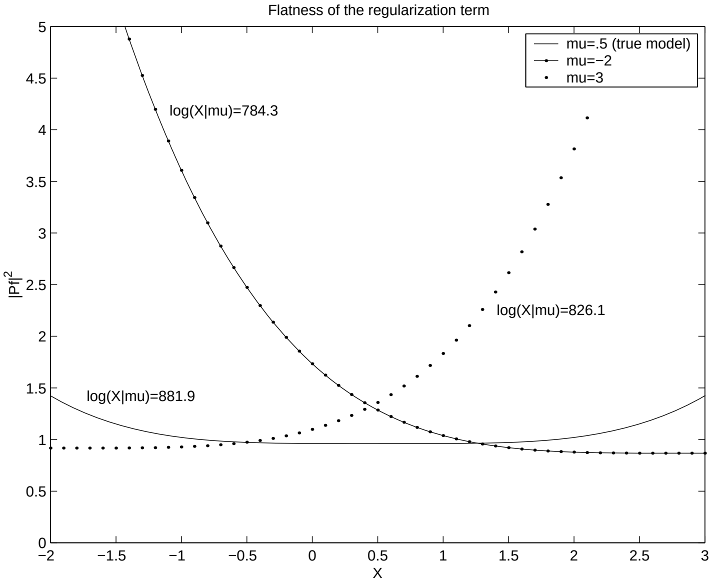

Figure 1: Flatness of the natural regularizer for a Gaussian generative pdf $\sim \mathcal{N}(0.5, 3)$, $\theta = (\mathbf{0.5}, \mathbf{3})$. Let us assume we are given two parameter vectors θ_1 and θ_2 which both lead to the same high likelihood. In this case, the regularizer will pick the parameter vector with the property that *perturbing* it will (on average) lead to a smaller change in the log likelihood, for in that case $\nabla_\theta \ln p(\mathbf{x}|\theta)$ will be smaller. Consequently, the regularizer will favor the solution which is more stable (flat).

Corollary 4 (Plain Kernel) *The regularization operator associated with the plain kernel* $k_{\mathbf{1}}^{\text{nat}}$ *is the gradient operator* $\nabla_{\mathbf{x}}$ *in the case where* $p(\mathbf{x}|\theta)$ *belongs to the exponential family of densities, i.e.* $\ln p(\mathbf{x}|\theta) = \theta \cdot \mathbf{x} - \pi(\mathbf{x}) + c_0$.

Proof We substitute $\ln p(\mathbf{x}|\theta)$ into the condition (11). This yields

$$\begin{aligned} & \int \left[\nabla_{\mathbf{z}} \nabla_\theta \ln p(\mathbf{z}|\theta)\right]^\top \left[\nabla_{\mathbf{z}} \nabla_\theta \ln p(\mathbf{z}|\theta)\right] p(\mathbf{z}|\theta) d\mathbf{z} \\ = & \int \left[\nabla_{\mathbf{z}} (\mathbf{z} - \nabla_\theta \pi(\mathbf{x}))\right]^\top \left[\nabla_{\mathbf{z}} (\mathbf{z} - \nabla_\theta \pi(\mathbf{x}))\right] p(\mathbf{z}|\theta) d\mathbf{z} = \mathbf{1}. \end{aligned} \tag{17}$$

since the terms depending only on $\mathbf{z}$ vanish after application ∇_θ. ∎

This means that the regularization term can be written as (note $\nabla_{\mathbf{x}} f(\mathbf{x})$ is a vector)

$$||Pf||^2 = ||\nabla_{\mathbf{x}} f(\mathbf{x})||_p^2 = \int ||\nabla_{\mathbf{x}} f(\mathbf{x})||^2 p(\mathbf{x}|\theta) d\mathbf{x} \tag{18}$$

thus favouring smooth functions via flatness in the first derivative. Often one is facing the opposite problem of identifying a kernel k_M^{nat} from its corresponding regularization operator P. This can be solved by evaluating (11) for the appropriate class of operators. A possible choice would be Radon-Nikodym derivatives, i.e. $p^{-1}(\mathbf{x})\nabla_{\mathbf{x}}$ [3] or powers thereof. In this regard (11) is particularly useful, since methods such as the probability integral transform which can be used to obtain Greens functions for Radon-Nikodym operators in $\mathbb{R}$ by mapping $\mathbb{R}$ into $[0, 1]$ with density 1, cannot be extended to $\mathbb{R}^n$.

4 The Feature Map of Natural Kernel

Given a regularization operator P with an expansion P^*P into a discrete eigensystem (λ_n, ψ_n), where λ are the eigenvalues and ψ the eigenvectors, and given a kernel k with

$$k(\mathbf{x}_i, \mathbf{x}_j) := \sum_n \frac{d_n}{\lambda_n} \psi_n(\mathbf{x}_i)\psi_n(\mathbf{x}_j) \tag{19}$$

where $d_n \in 0, 1$ for all m, and $\sum_n \frac{d_n}{\lambda_n}$ convergent. Then k satisfies the self-consistency property stated in equation (9) [10]. For the purpose of designing a kernel with regularization properties given by P, eq. (19) is a constructive version of Mercer's Theorem (Th. ??).

The eigenvalues of the Gram Matrix of the training set are used to bound the generalization error or a linear classifier [7]. By linear algebra we may explicitly construct such an expansion (19).

Proposition 5 (Map into Feature Space) *Denote by I the Fischer information matrix, by M the kernel matrix, and by s_i, Λ_i the eigensystem of $M^{-\frac{1}{2}} I M^{-\frac{1}{2}}$. The kernel $k_M^{\mathrm{nat}}(\mathbf{x}, \mathbf{x}')$ can be decomposed into an eigensystem*

$$\psi_i(\mathbf{x}) = \frac{1}{\sqrt{\Lambda_i}} s_i^\top M^{-\frac{1}{2}} \nabla_\theta \ln p(\mathbf{x}|\theta) \textit{ and } \lambda_i = \Lambda_i. \tag{20}$$

Note that if $M = I$ we have $\lambda_i = \Lambda_i = 1$.

Proof It can be seen immediately that (19) is satisfied. This follows from the fact that s_i is an orthonormal basis, $(\mathbf{1} = \sum_i s_i s_i^\top)$ and the definition of k_M^{nat}. The terms depending on Λ_i cancel out mutually.

The second part (orthonormality of ψ_i) can be seen as follows.

$$\langle \psi_i, \psi_j \rangle \tag{21}$$

$$= \int \left(\frac{1}{\sqrt{\Lambda_i}} s_i^\top M^{-\frac{1}{2}} \nabla_\theta \ln p(\mathbf{x}|\theta) \right) \left(\frac{1}{\sqrt{\Lambda_j}} \nabla_\theta^\top \ln p(\mathbf{x}|\theta) M^{-\frac{1}{2}} s_j \right) p(\mathbf{x}|\theta) d\mathbf{x}$$

$$= \frac{1}{\sqrt{\Lambda_i \Lambda_j}} s_i^\top M^{-\frac{1}{2}} I M^{-\frac{1}{2}} s_i = \delta_{ij} \tag{22}$$

This completes the proof. ■

The eigenvalues λ_i of k_I^{nat} are all 1, reflecting the fact that the matrix I whitens the scores $\nabla_\theta \ln(p(x|\theta))$. It also can be seen from $P_I = \mathbf{1}$ that (20) becomes $\psi_i(x) = \frac{1}{\sqrt{\lambda_i^I}} s_i \cdot \nabla_\theta \ln(p(x|\theta))$, $1 \leq i \leq r$.

What are the consequences of the fact that all eigenvalues are equal? Standard VC dimension bounds [11] state that the capacity of a linear classifier or regression algorithm is essentially given by $R^2 \cdot \Lambda^2$. Here, R is the radius of the smallest sphere containing the data (in feature space), and Λ is the maximal allowed length of the weight vector. Recently, it has been shown that both the spectrum of an associated integral operator [13] and the spectrum of the Gram matrix k_{ij} [7] can be used to formulate generalization error bounds. This was done by exploiting the fact that since $C := \sup_j \|\psi_j\|_{L_\infty}$ exists, (20) implies that $|\Phi_i(x)| = \sqrt{\lambda_i}|\psi_i(\mathbf{x})| \leq \sqrt{\lambda_i} C$, i.e. the mapped data live in some parallelepiped whose sidelengths are given by the square roots of the eigenvalues. New bounds improved upon the generic VC dimension bounds by taking into account this fact: due to the decay of the eigenvalues, the mapped data are not distributed isotropically. Therefore capturing the shape of the mapped data only by the radius of a sphere should be a rather rough approximation. On the other hand, taking into account the rate of decay of the eigenvalues allows one to formulate kernel-dependent bounds which are much more accurate than the standard VC-bounds.

In our case all λ_i are 1, therefore $|\Phi_i(\mathbf{x})| = |\psi_i(\mathbf{x})|$. Hence the upper bound simply states that the mapped data is contained in some box with equal sidelengths (hypercube). Moreover, the $L_2(p)$ normalization of the eigenfunctions ψ_i means that $\int \psi_i(x)^2 p(x|\theta)\ dx = 1$. Therefore, the squared averaged size of the feature map's ith coordinate is independent of i, implying that the the mapped data have the same range in all directions. This isotropy of the Fisher kernels suggests that the standard 'isotropic' VC bounds should be fairly precise in this case.

5 Experiments

The flat eigenspectrum of the Fisher kernel suggests a way of comparing different models: we compute the Gram matrix for a set of $\mathcal{K}$ models $p(\mathbf{x}|\theta^j)$ with $j = 1 \ldots \mathcal{K}$. In the case of the true model, we expect $\lambda_i = 1$ for all i. Therefore one might select the model j such that its spectrum is the flattest. As a sanity check for the theory developed, Figure 5 illustrates the selection of the sufficient statistics (μ, σ) of a one-dimensional normal pdf $p(\mathbf{x}|\theta) = \mathcal{N}(\mu, \sigma)$ with 10 training data points sampled from $\mathcal{N}(0.5, 3)$. We computed the eigendecomposition of the empirical Gram matrices, using the Fisher kernels of a set of different models. The figure contains the error bar plots of the ratio of its 2 largest eigenvalues (note that in this case the parameter space is two-dimensional). The minimum corresponds to the model to be selected.

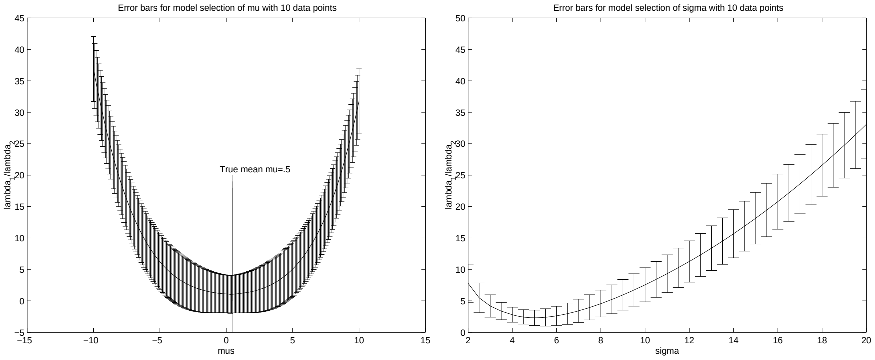

Figure 2: Model selection using the ratio of the two largest eigenvalues of the empirical Gram Matrix. Right: selecting the standard deviation. Left: selecting the mean

6 Discussion

In this paper we provided a regularization-theoretic analysis of a class of SV kernels — called natural kernels — based on generative models with density $p(\mathbf{x}|\theta)$, such as the Fisher kernel. In particular, we have shown that the latter corresponds to a regularization operator (prior) penalizing the $L_2(p)$-norm of the estimated function. Comparing this result to the regularization-theoretic analysis of SV kernels [9], where common SV kernels such as the Gaussian have been shown to correspond to a sum over differential operators of different orders, the question arises whether it is possible to find a modified natural kernel which uses higher order derivatives in the regularization term, such as

$$\|Pf\|^2 = \sum_{n=0}^{\infty} c_n \left\|\nabla^n f\right\|_{L_2(p)}^2 . \tag{23}$$

Second, we derived the feature map corresponding to natural kernels. It turned out that the Fisher natural kernel corresponding to a r-parameter generative model maps the input data into a r-dimensional feature space where the data are distributed isotropically (in the sense that the covariance matrix is the identity). This reflects the fact that all parameters are considered equally important, and that the Fisher kernel is invariant with respect to parameter rescaling; it automatically scales feature space in a principled way. Our analysis provides some understanding for the impressive empirical results obtained using the Fisher kernel.

Acknowledgments

Thanks to S.-I. Amari, A. Elisseeff, K.-R. Müller, and S. Wu for helpful discussions. Parts of this work were done while AS, BS, and NO were at GMD FIRST and University of Madison.

References

[1] S. Amari and S. Wu. Improving support vector machines by modifying kernel functions. Technical report, RIKEN, 1999.

[2] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, July 1992. ACM Press.

[3] S. Canu and A. Elisseeff. Unpublished manuscript.

[4] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6):1455–1480, 1998.

[5] F. Girosi, M. Jones, and T. Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[6] T. S. Jaakkola and D. Haussler. Probabilistic kernel regression models. In *Proceedings of the 1999 Conference on AI and Statistics*, 1999.

[7] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Generalization bounds via eigenvalues of the Gram matrix. Submitted to COLT99, February 1999.

[8] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[9] A. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[10] A. Smola, B. Schölkopf, and K.-R. Müller. General cost functions for support vector regression. In T. Downs, M. Frean, and M. Gallagher, editors, *Proc. of the Ninth Australian Conf. on Neural Networks*, pages 79 – 83, Brisbane, Australia, 1998. University of Queensland.

[11] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[12] C. K. I. Williams. Prediction with gaussian processes: From linear regression to linear prediction and beyond. In M. I. Jordan, editor, *Learning and Inference in Graphical Models.* Kluwer, 1998. To appear. Also: Technical Report NCRG/97/012, Aston University.

[13] R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. NeuroCOLT Technical Report NC-TR-98-019, Royal Holloway College, University of London, UK, 1998.