Observations on the Nyström Method for Gaussian Process Prediction

Christopher K. I. Williams
Division of Informatics
University of Edinburgh
c.k.i.williams@ed.ac.uk

Carl Edward Rasmussen
Gatsby Computational Neuroscience Unit
University College London
edward@gatsby.ucl.ac.uk

Anton Schwaighofer
Institute for Theoretical Computer Science
Technische Universität Graz
anton.schwaighofer@gmx.net

Volker Tresp
Siemens Corporate Technology
Department of Information and Communications
Volker.Tresp@mchp.siemens.de

July 17, 2002

Abstract

A number of methods for speeding up Gaussian Process (GP) prediction have been proposed, including the *Nyström method* of Williams and Seeger (2001). In this paper we focus on two issues (1) the relationship of the Nyström method to the Subset of Regressors method (Poggio and Girosi, 1990; Luo and Wahba, 1997) and (2) understanding in what circumstances the Nyström approximation would be expected to provide a good approximation to exact GP regression.

Over recent years kernel-based predictors such as Support Vector Machines (SVMs) (Vapnik, 1995), Gaussian process predictors (see, e.g. Williams and Rasmussen, 1996; Williams and Barber, 1998) and splines (Wahba, 1990) have become very popular. One of the main problems with such methods is that the computational complexity required to find the solution generally scales as $O(n^3)$, where n is the number of training examples[1]. This scaling has led to the proposal of a number of methods for the approximation of the exact solution with lower complexity. In this paper we focus on approximations to Gaussian Process (GP) regression. Examples of approximation methods are the Subset of Regressors (SR) (Poggio and Girosi, 1990; Luo and Wahba, 1997), the Bayesian Committee Machine (BCM) (Tresp, 2000), the Nyström method Williams and Seeger (2001), and work by Gibbs and MacKay (1997), Smola and Bartlett (2001) and Rasmussen (2002).

In section 1 we present an overview of Gaussian Process regression and the Nyström approximation. Section 2 focuses on the relationship of the Nyström method to the Subset of Regressors method and section 3 analyzes in what circumstances the Nyström approximation would be expected to yield a good approximation to exact GP regression.

[1] In certain cases the complexity can be better. For example for splines in 1-d, the computation required is $O(n)$ as the matrix concerned is banded. For SVMs the quadratic programming problem can be solved faster if the number of support vectors is small relative to n. Also, if the dimension of the feature space N_F corresponding to the kernel is less than n, then the complexity of the solution will scale at least as well as $O(N_F^3)$.

1 Gaussian process regression and the Nyström approximation

We follow the presentation of GP regression as in Williams and Rasmussen (1996) and Williams (1998). GP regression has a long history in various literatures, going back at least as far as Whittle (1963).

A Gaussian process prior is placed over random functions $y(\mathbf{x})$. This is achieved by specifying a mean function $\mu(\mathbf{x})$ (which we take to be identically zero) and a covariance function $k(\mathbf{x}, \mathbf{x}')$ which specifies $\langle y(\mathbf{x})y(\mathbf{x}')\rangle$. For n $\mathbf{x}$ locations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ the corresponding function values $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n))^T \stackrel{def}{=} (y_1, y_2, \ldots, y_n)^T$ are distributed as $N(\mathbf{0}, K)$, where K is the $n \times n$ covariance (or Gram) matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ with $i,\ j = 1, \ldots, n$.

We are given observations $\mathbf{t} = (t_1, \ldots, t_n)^T$ which are assumed to be noisy versions of the corresponding y's, so that $t_i \sim N(y_i, \sigma_i^2)$. Below we assume that $\sigma_i^2 = \sigma_\nu^2$ for all i. The prediction for some new input point $\mathbf{x}_*$ is $y(\mathbf{x}_*|\mathbf{t}) \sim N(\hat{y}(\mathbf{x}_*), \sigma^2(\mathbf{x}_*))$, where

$$\hat{y}(\mathbf{x}_*) = \mathbf{k}^T(\mathbf{x}_*)(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t}, \tag{1}$$

$$\sigma^2(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}^T(\mathbf{x}_*)(K + \sigma_\nu^2 I_n)^{-1}\mathbf{k}(\mathbf{x}_*) \tag{2}$$

and $\mathbf{k}(\mathbf{x}_*) = (k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_n, \mathbf{x}_*))^T$. It is also easy to show that $P(\mathbf{y}|\mathbf{t}) \sim N(\hat{\mathbf{y}}, \sigma_\nu^2 K(K + \sigma_\nu^2 I_n)^{-1})$ where $\hat{\mathbf{y}} = K(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$, and that $\hat{y}(\mathbf{x}_*) = \mathbf{k}^T(\mathbf{x}_*)K^{-1}\hat{\mathbf{y}}$. In addition, by integrating out $\mathbf{y}$ we can show that the log marginal likelihood (or evidence) is given by

$$\log P(\mathbf{t}|\mathbf{x}_1^n) = -\frac{1}{2}\log|K + \sigma_\nu^2 I_n| - \frac{1}{2}\mathbf{t}^T(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t} - \frac{n}{2}\log(2\pi), \tag{3}$$

where $\mathbf{x}_1^n$ denotes $\{\mathbf{x}_i\}_{i=1}^n$. This quantity is useful in Bayesian approaches to model selection/averaging.

The major computational problem in GP regression is the need to invert $K + \sigma_\nu^2 I_n$ which takes $O(n^3)$. The idea of the Nyström approximation (Williams and Seeger, 2001) is to approximate K with a reduced-rank matrix $\tilde{K}$. This is constructed as $\tilde{K} = K_{nm}K_{mm}^{-1}K_{mn}$, where K_{nm} is the $n \times m$ block of the original matrix K, and with similar definitions for the other blocks[2]. (This approximation, presented in Williams and Seeger (2001), turned out to be an independent derivation of a special case of the results in Frieze et al. (1998); the same approximation was derived by yet another different route in Smola and Schölkopf (2000). Fowlkes et al. (2001) have applied the Nyström method to approximate the top few eigenvectors in a computer vision problem where the matrices in question are larger than $10^6 \times 10^6$ in size.). The m points are a subset of the total n points; for now we leave unspecified exactly how this subset is selected. The Nyström approximation then consists of replacing the matrix K by $\tilde{K}$ in equations 1-3. The complexity of the resulting computations is now $O(m^2 n)$; this is achieved by use of the Woodbury formula (see, e.g. Press et al. (1992)) and an analogous relationship for determinants. Thus the Nyström approximation for $\hat{y}(\mathbf{x}_*)$ is given by

$$\hat{y}_{Ny}(\mathbf{x}_*) = \beta\mathbf{k}^T(\mathbf{x}_*)\left(\mathbf{t} - K_{nm}(K_{mn}K_{nm} + \sigma_\nu^2 K_{mm})^{-1}K_{mn}\mathbf{t}\right). \tag{4}$$

Numerical stability may be improved by computing the SVD of K_{mm} and using equation (11) from Williams and Seeger (2001) to obtain $\hat{y}_{Ny}(\mathbf{x}_*)$.

[2]Here and below we assume without loss of generality and for simplicity of notation that the m chosen points occur first.

2 Relationship to SR method[3]

Silverman (1985, section 6.1) showed that the *mean* GP predictor can be obtained from a finite-dimensional generalized linear regression model $y(\mathbf{x}) = \sum_{i=1}^{n} c_i k(\mathbf{x}, \mathbf{x}_i)$ with a prior $\mathbf{c} \sim N(\mathbf{0}, K^{-1})$. In the subset of regressors method (Poggio and Girosi, 1990; Luo and Wahba, 1997) the sum over all n points is replaced by a sum over a subset $m < n$ of the points, setting the remaining coefficients to zero. This gives a finite-dimensional Gaussian process model with covariance function $k_{SR}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_m^T(\mathbf{x}) K_{mm}^{-1} \mathbf{k}_m(\mathbf{x}')$, where $\mathbf{k}_m(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_m))^T$. The subset of regressors method gives the following predictions:

$$\hat{y}_{SR}(\mathbf{x}_*) = \mathbf{k}_m^T(\mathbf{x}_*)(K_{mn}K_{nm} + \sigma_\nu^2 K_{mm})^{-1} K_{mn}\mathbf{t}, \tag{5}$$

$$\sigma_{SR}^2(\mathbf{x}_*) = \sigma_\nu^2 \mathbf{k}_m^T(\mathbf{x}_*)(K_{mn}K_{nm} + \sigma_\nu^2 K_{mm})^{-1} \mathbf{k}_m(\mathbf{x}_*). \tag{6}$$

Equation 5 is, in our notation, equation 25 of Poggio and Girosi (1990). As Poggio and Girosi were working in a regularization framework, they did not give an expression for the predictive variance σ_{SR}^2. Clearly the Nyström approximation is different to the SR approximation, not least in that the Nyström method predictor $\hat{y}_{Ny}(\mathbf{x}_*)$ is a linear combination of all n kernel functions, not just m. However, we note that both the Nyström method and the SR method give the same value for the marginal likelihood, namely

$$\log P_{SR}(\mathbf{t}|\mathbf{x}_1^n) = -\frac{1}{2}\log|\tilde{K} + \sigma_\nu^2 I_n| - \frac{1}{2}\mathbf{t}^T(\tilde{K} + \sigma_\nu^2 I_n)^{-1}\mathbf{t} - \frac{n}{2}\log(2\pi). \tag{7}$$

One aspect of Silverman's construction which carries through to the SR model is that if we choose decaying kernel functions (such as Gaussian kernels), then far from any datapoints the prior variance of the linear combination $y(\mathbf{x}) = \sum_{i=1}^{m} c_i k(\mathbf{x}, \mathbf{x}_i)$ will be very small; this seems to be a peculiar prior assumption. In contrast the true GP prior can have large variance far from the datapoints (e.g. for a stationary kernel), and typically this variance will also remain large in the posterior, reflecting that we haven't learned much about the function here. However, if we apply the Nyström approximation as stated above (by replacing K by $\tilde{K}$ in equation 2) it can happen that the predicted variance turns out embarrassingly to be negative.

In the Nyström method the kernel matrix K was approximated so that $K_{ij} \simeq \tilde{K}_{ij} = \mathbf{k}_m^T(\mathbf{x}_i) K_{mm}^{-1} \mathbf{k}_m^T(\mathbf{x}_j)$. (In fact the only approximation occurs in the block $K_{(n-m)(n-m)}$.) If we also apply this approximation $\tilde{k}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_m^T(\mathbf{x}) K_{mm}^{-1} \mathbf{k}_m(\mathbf{x}')$ generally to all appearances of the kernel function in equations 1 and 2 we obtain the SR predictors.

3 When does the Nyström method work well?

Consider the matrix eigenvalue equation $K\mathbf{e}_i = \lambda_i \mathbf{e}_i$, where the eigenvalues are ordered so that $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$, and $K = \sum_{i=1}^{n} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$. The Nyström method replaces K with $\tilde{K}$, a rank-m approximation to K. Let $L = \sum_{i=1}^{m} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$, a rank-$m$ approximation to K based on the first m eigenvectors. In an optimistic setting we might expect that $\tilde{K}$ would be close to L.

Consider the prediction $\hat{\mathbf{y}} = K(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$. Expanding $\mathbf{t}$ in the eigenbasis, so that $\mathbf{t} = \sum_{i=1}^{n} \gamma_i \mathbf{e}_i$, we obtain $\hat{\mathbf{y}} = \sum_{i=1}^{n} \gamma_i \frac{\lambda_i}{\lambda_i + \sigma_\nu^2} \mathbf{e}_i$. Notice that an eigenvector $\mathbf{e}_i$ with eigenvalue λ_i such that $\lambda_i \ll \sigma_\nu^2$ is effectively zeroed out, i.e. it does not matter if it is not represented in L. A similar conclusion can be obtained by analyzing the effect of approximating K by L on

[3] Much of the content of this section was initially set out in an email from Chris Williams to a number of kernel-methods researchers on January 22 2001, in response to an email query from Grace Wahba. Prof Wahba suggested the term subset of regressors.

m	Nyström	SR	just-m	m-eigenvectors
100	34.4430 ± 43.2918	0.1436 ± 0.0360	0.2267 ± 0.0656	0.6733
200	1.0266 ± 0.9009	0.1059 ± 0.0141	0.1446 ± 0.0329	0.0844
300	0.1335 ± 0.0536	0.0885 ± 0.0073	0.1171 ± 0.0222	0.0846
400	0.0871 ± 0.0071	0.0843 ± 0.0026	0.0922 ± 0.0193	0.0845

Table 1: Comparison of the Nyström, SR, just-m and m-eigenvectors methods on the Boston housing data set for values of m of 100, 200, 300, 400. For the first three methods ten replications were used, with random choice of the $\mathbf{x}$ points; each entry shows the mean and standard deviation of the 10 MSE results.

the calculation of the coefficient vector $\mathbf{c}$, as described in Appendix A. Thus we would expect that the Nyström method might work well if $\lambda_{m+1} \ll \sigma_\nu^2$. This will occur if the eigenvalues of K decay fast enough, or if σ_ν^2 is relatively large. Note that for a covariance function like the Gaussian or squared exponential kernel (see equation 8), for fixed inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ the eigenspectrum will decay faster if the kernel is wider. The effect of the approximation on the estimation of $\hat{\mathbf{y}}$ is analyzed in Appendix B.

We provide an illustration of this using the Boston housing data set originally published by Harrison and Rubinfeld (1978). This data set is publicly available at the UCI database (Blake and Merz, 1998) and in DELVE `http://www.cs.utoronto.ca/~delve`. There are $D = 13$ predictor attributes. A split of 455 training points and 51 test points was used. The kernel function used is the "Gaussian" or "squared-exponential" kernel plus a linear regression model, of the form[4]

$$k(\mathbf{x}, \mathbf{x}') = \sum_{d=1}^{D} a_d x_d x'_d + v_0 \exp(-\frac{1}{2} \sum_{d=1}^{D} w_d (x_d - x'_d)^2). \tag{8}$$

For this data set and parameter settings there are 191 eigenvalues of K which are larger than σ_ν^2. The predictive mean squared error (MSE) using a GP predictor with all 455 examples is 0.0845. Table 1 shows the average MSE for the Nyström, SR and the just-m methods using 10 random choices of the $m = 100$, 200, 300, 400 points. (In the just-m method, a Gaussian process predictor using only the targets corresponding to the m chosen $\mathbf{x}$ points was used.) Also shown for comparison with the Nyström method is the m-eigenvectors method where we have carried out an eigendecomposition of K and used only the top m eigenvalues/vectors to approximate K (as discussed in Williams and Seeger (2000)). The results show that for smaller m, the Nyström method performs worse than the other methods. It is uniformly worse on average than the SR method, but does beat the just-m method for $m = 400$. It is close in performance to the SR method for $m = 400$, in fact it outperforms the SR method on 4 out of the 10 replications. We see that for $m = 100$ the m-eigenvectors method is inferior to the SR and just-m methods but by $m = 200$ it is very close to the optimal performance. However, the Nyström method performs much worse than the m-eigenvectors method for $m = 200$, 300 which suggests that important eigenvalues/vectors are not well approximated. Figure 1 shows the log eigenvalues of $K + \sigma_\nu^2 I_n$ and $\tilde{K} + \sigma_\nu^2 I_n$ for $m = 100$. (The log scale emphasizes the differences between

[4]The parameters $w_1, \ldots, w_{13}$ had values $(0.0124, 0.0008, 0.0022, 0.0509, 21.4585, 0.1914, 0.0418, 0.4933, 0.3645, 0.7684, 0.0180, 0.0059, 0.1321)$, $a_1, \ldots, a_{13}$ had values $(0.0083, 0.0006, 0.0028, 0.0015, 0.0268, 0.1394, 0.0347, 0.0920, 0.0720, 0.0396, 0.0277, 0.0061, 0.0520)$, v_0 was 0.8686 and σ_ν^2 was 0.0291. These values were obtained by maximizing the marginal likelihood with respect to the kernel parameters.

the small eigenvalues; in a plot with linear scaling the curves are superimposed.) If $\tilde{K}$ were exactly equal to L then the first 100 eigenvalues would be identical, and then there would be a sharp drop-off. In either case, eigenvalues significantly larger than σ_ν^2 are inaccurately estimated, leading to poor predictions. We also note that the spectrum of $\tilde{K}$ drops faster than that for K; this phenomenon is currently under study.

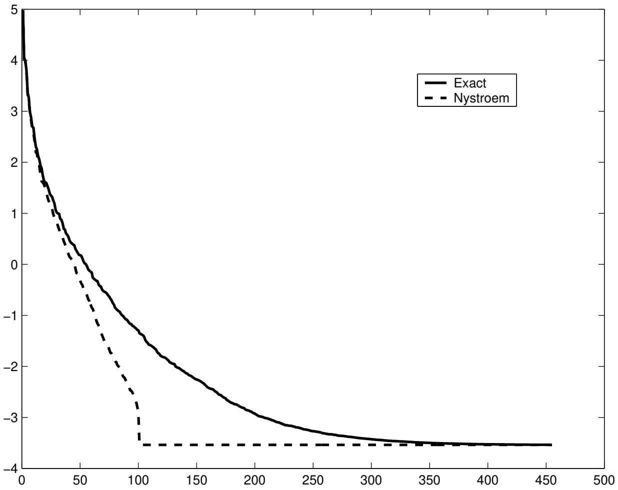

Figure 1: A plot of $\log_e$ eigenvalue against the index of the eigenvalue for both $K + \sigma_\nu^2 I_n$ (exact) and $\tilde{K} + \sigma_\nu^2 I_n$ (Nyström) for $m = 100$. Notice the drop at index 100 for the Nyström case, due to the rank-100 approximation. The horizontial line in the dashed plot is at $\log_e \sigma_\nu^2$.

The Nyström method was originally tested on the UCI abalone data set using $\sigma_\nu^2 = 0.05$. Analysis shows that for the kernel parameters used, 112 eigenvalues in K were larger than σ_ν^2. This is in good agreement with the experimental results that values of m of 250 or larger gave good results, but for $m = 125$ performance declined quite markedly.

Our conclusions are that the quality of the Nyström approximation for a given m will depend on the relative rate of decay of the eigenspectrum of K in relation to σ_ν^2. Note that Nyström theory provides an estimate of the first m eigenvalues of K by rescaling the eigenvalues of K_{mm} by a factor of n/m, and that these m eigenvalues can be computed in $O(m^3)$. Hence it should be possible to assess with reasonable efficiency when the Nyström approximation would be expected to hold, although the systematic under-estimation of small eigenvalues observed in Figure 1 means that this should be treated with some caution.

The results given above apply to regression problems. However, for GP classification problems it is common to add some "jitter" to the kernel matrix (i.e. to add on ϵI_n to K, Neal (1998)). In this case the analysis presented above also applies.

On the Boston Housing problem the SR method does better than the Nyström method. However, note that on a MNIST digit binary classification task (classifying 0-4 against 5-9) Tresp and Schwaighofer (2001) reported significantly better results for the Nyström than the SR method.

4 Summary

Section 2 above describes the relationship between the SR and Nyström methods; if we use the approximate kernel $k_{SR}(\mathbf{x}, \mathbf{x}') = \mathbf{k}_m^T(\mathbf{x}) K_{mm}^{-1} \mathbf{k}_m(\mathbf{x}')$ only when both points $\mathbf{x}$ and $\mathbf{x}'$ are in the training set we obtain the Nyström method, while if we use $\mathbf{k}_{SR}(\cdot, \cdot)$ everywhere in equations 1-3 we obtain the SR method. From the experimental evidence the SR method seems to superior to the Nyström method. For large m there is not much difference, but the Nyström approximation can be quite poor for small m. There is also a difference between the Nyström and SR methods in terms of the predictive variance $\sigma^2(\mathbf{x}_*)$, especially when $\mathbf{x}_*$ is far from any training point.

The intuition behind the Nyström method is that it will work well when K can be well-approximated by a rank-m matrix $\tilde{K}$. Section 3 makes this more precise, and shows that this largely explains the observed behaviour on the Boston housing and abalone data sets. Ongoing work focuses on variations of the Nyström approach which would produce more accurate approximations of the eigenvalues/vectors.

A The influence of approximating K by L on the coefficient vector $\mathbf{c}$

The Gaussian process predictor has the form $\hat{y}(\mathbf{x}_*) = \mathbf{k}^T(\mathbf{x}_*)(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t} = \mathbf{k}^T(\mathbf{x}_*)\mathbf{c}$, where $(K + \sigma_\nu^2 I_n)\mathbf{c} = \mathbf{t}$. We consider the effect of replacing K by L, the rank-m matrix which has the leading m eigenvalues and eigenvectors identical to those in K. We denote the eigenvalues of L as $\{\lambda_i^L\}$, where $\lambda_i^L = \lambda_i$ for $i = 1, \ldots, m$ and $\lambda_i^L = 0$ for $i = m+1, \ldots, n$. Let the corresponding linear system be $(L + \sigma_\nu^2 I_n)\mathbf{c}^L = \mathbf{t}$.

We make use of the following theorem (Theorem III.2.11 in Stewart and Sun (1990))

Theorem 1 *Consider the linear system $A\mathbf{x} = \mathbf{b}$ for a non-singular matrix A. Let $\tilde{A} = A + E$ be a perturbation of A. If there is a vector $\tilde{\mathbf{x}}$ that solves the perturbed system $\tilde{A}\tilde{\mathbf{x}} = \mathbf{b}$ then*

$$\frac{||\mathbf{x} - \tilde{\mathbf{x}}||}{||\tilde{\mathbf{x}}||} \leq ||A^{-1}E||. \tag{9}$$

Here $||\cdot||$ denotes both a matrix and a vector norm, where the two norms must be consistent. For example $||\cdot||$ may be the Euclidean norm for vectors and the spectral norm for matrices.

Let $K + \sigma_\nu^2 I_n = A$ and $L + \sigma_\nu^2 I_n = \tilde{A}$, so that $E = L - K$. Using the Euclidean norm for vectors and the spectral norm for matrices, we obtain

$$||A^{-1}E|| = \max_i \frac{|\lambda_i - \lambda_i^L|}{\lambda_i + \sigma^2} \leq \max_i \frac{|\lambda_i - \lambda_i^L|}{\lambda_i^L + \sigma^2}, \tag{10}$$

and can therefore write

$$\frac{||\mathbf{c}^L - \mathbf{c}||}{||\mathbf{c}^L||} \leq \max_i \frac{|\lambda_i - \lambda_i^L|}{\lambda_i^L + \sigma^2}. \tag{11}$$

Using the properties of K and L, we see that this reduces to

$$\frac{||\mathbf{c}^L - \mathbf{c}||}{||\mathbf{c}^L||} \leq \frac{\lambda_{m+1}}{\sigma^2}. \tag{12}$$

As for the analysis for $\hat{\mathbf{y}}$ given in section 3, we see that we expect this approximation to be good when $\lambda_{m+1} \ll \sigma_\nu^2$.

B The influence of the approximations on computing $\hat{\mathbf{y}}$

As we have seen, using GP regression the prediction at the training points $\hat{\mathbf{y}}$ is given by $\hat{\mathbf{y}} = K(K + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$. If we use reduced rank matrix L in place of K we would obtain an approximation $\hat{\mathbf{y}}_{LL} = L(L + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$ to $\hat{\mathbf{y}}$. However, in the Nyström approximation we only use the reduced rank approximation within the factor $(K + \sigma_\nu^2 I_n)^{-1}$ and thus by analogy to predictions at new $\mathbf{x}$'s we would obtain $\hat{\mathbf{y}}_{KL} = K(L + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$.[5] Below we compare the errors introduced by these two approximations.

As in section 3 we write $\mathbf{t} = \sum_{i=1}^n \gamma_i \mathbf{e}_i$. Then we have

$$\hat{\mathbf{y}}_{KL} = \sum_{i=1}^{m} \frac{\gamma_i \lambda_i}{\lambda_i + \sigma_\nu^2}\mathbf{e}_i + \sum_{i=m+1}^{n} \frac{\gamma_i \lambda_i}{\sigma_\nu^2}\mathbf{e}_i \tag{13}$$

and

$$\hat{\mathbf{y}}_{LL} = \sum_{i=1}^{m} \frac{\gamma_i \lambda_i}{\lambda_i + \sigma_\nu^2}\mathbf{e}_i. \tag{14}$$

Thus

$$E_{KL} = \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{KL}\|^2 = \sum_{i=m+1}^{n} \gamma_i^2 \left(\frac{\lambda_i}{\lambda_i + \sigma_\nu^2} - \frac{\lambda_i}{\sigma_\nu^2}\right)^2 = \sum_{i=m+1}^{n} \gamma_i^2 \frac{\lambda_i^4}{\sigma_\nu^4(\lambda_i + \sigma_\nu^2)^2} \tag{15}$$

and

$$E_{LL} = \|\hat{\mathbf{y}} - \hat{\mathbf{y}}_{LL}\|^2 = \sum_{i=m+1}^{n} \gamma_i^2 \frac{\lambda_i^2}{(\lambda_i + \sigma^2)^2}. \tag{16}$$

The ratio of the ith terms in these error expansions $(i = m+1, \ldots, n)$ is

$$\frac{E_{KL}^i}{E_{LL}^i} = \left(\frac{\lambda_i}{\sigma_\nu^2}\right)^2. \tag{17}$$

Thus we see that if $\lambda_{m+1} < \sigma_\nu^2$ then the Nyström-type approximation will give a more accurate estimate of $\hat{\mathbf{y}}$ than the simple reduced rank approximation.

References

Blake, C. L. and Merz, C. J. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Fowlkes, C., Belongie, S., and Malik, J. (2001). Efficient Spatiotemporal Grouping Using the Nyström Method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2001.*

[5] Williams and Seeger (2001) did not define exactly how predictions would be made at the training $\mathbf{x}$'s. The equivalence of the SR marginal likelihood (7) with Nyström marginal likelihood would imply that the Nyström predictions at the training points would be equivalent to $\hat{\mathbf{y}}_{SR}$. However, the form $K(\tilde{K} + \sigma_\nu^2 I_n)^{-1}\mathbf{t}$ is analagous to predictions at new test $\mathbf{x}$'s.

Frieze, A., Kannan, R., and Vempala, S. (1998). Fast Monte-Carlo Algorithms for finding low-rank approximations. In *39th Conference on the Foundations of Computer Science*, pages 370–378.

Gibbs, M. and MacKay, D. J. C. (1997). Efficient Implementation of Gaussian Processes. Unpublished manuscript. Cavendish Laboratory, Cambridge, UK. Available from `http://www.inference.phy.cam.ac.uk/mackay/`.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Luo, Z. and Wahba, G. (1997). Hybrid Adaptive Splines. *J. Amer. Statist. Assoc.*, 92:107–116.

Neal, R. M. (1998). Regression and classification using Gaussian process priors (with discussion). In Bernardo, J. M. et al., editors, *Bayesian statistics 6*, pages 475–501. Oxford University Press.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of IEEE*, 78:1481–1497.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, Second edition.

Rasmussen, C. E. (2002). Reduced Rank Gaussian Process Learning. Unpublished manuscript.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Stat. Soc. B*, 47(1):1–52.

Smola, A. J. and Bartlett, P. (2001). Sparse Greedy Gaussian Process Regression. In Leen, T. K., Diettrich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 619–625. MIT Press.

Smola, A. J. and Schölkopf, B. (2000). Sparse Greedy Matrix Approximation for Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann.

Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory*. Academic Press, San Diego, CA.

Tresp, V. (2000). A Bayesian Committee Machine. *Neural Computation*, 12(11):2719–2741.

Tresp, V. and Schwaighofer, A. (2001). Scalable Kernel Systems. Presentation at NIPS 2001 workshop on New Directions in Kernel-Based Learning Methods.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer Verlag, New York.

Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics, Philadelphia, PA. CBMS-NSF Regional Conference series in applied mathematics.

Whittle, P. (1963). *Prediction and regulation by linear least-square methods*. English Universities Press.

Williams, C. K. I. (1998). Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 599–621. Kluwer Academic.

Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351.

Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 514–520. MIT Press.

Williams, C. K. I. and Seeger, M. (2000). The Effect of the Input Density Distribution on Kernel-based Classifiers. In Langley, P., editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000).* Morgan Kaufmann.

Williams, C. K. I. and Seeger, M. (2001). Using the Nyström Method to Speed Up Kernel Machines. In Leen, T. K., Diettrich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press.