On the Impact of Kernel Approximation on Learning Accuracy

Corinna Cortes
Google Research
New York, NY
corinna@google.com

Mehryar Mohri
Courant Institute and Google Research
New York, NY
mohri@cs.nyu.edu

Ameet Talwalkar
Courant Institute
New York, NY
ameet@cs.nyu.edu

Abstract

Kernel approximation is commonly used to scale kernel-based algorithms to applications containing as many as several million instances. This paper analyzes the effect of such approximations in the kernel matrix on the hypothesis generated by several widely used learning algorithms. We give stability bounds based on the norm of the kernel approximation for these algorithms, including SVMs, KRR, and graph Laplacian-based regularization algorithms. These bounds help determine the degree of approximation that can be tolerated in the estimation of the kernel matrix. Our analysis is general and applies to arbitrary approximations of the kernel matrix. However, we also give a specific analysis of the Nyström low-rank approximation in this context and report the results of experiments evaluating the quality of the Nyström low-rank kernel approximation when used with ridge regression.

1 Introduction

The size of modern day learning problems found in computer vision, natural language processing, systems design and many other areas is often in the order of hundreds of thousands and can exceed several million. Scaling standard kernel-based algorithms such as support vector machines (SVMs) (Cortes and Vapnik, 1995), kernel ridge regression (KRR) (Saunders *et al.*, 1998), kernel principal component analysis (KPCA) (Schölkopf *et al.*, 1998) to such magnitudes is a serious issue since even storing the kernel matrix can be prohibitive at this size.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

One solution suggested for dealing with such large-scale problems consists of a low-rank approximation of the kernel matrix (Williams and Seeger, 2000). Other variants of these approximation techniques based on the Nyström method have also been recently presented and shown to be applicable to large-scale problems (Belabbas and Wolfe, 2009; Drineas and Mahoney, 2005; Kumar *et al.*, 2009a; Talwalkar *et al.*, 2008; Zhang *et al.*, 2008). Kernel approximations based on other techniques such as column sampling (Kumar *et al.*, 2009b), incomplete Cholesky decomposition (Bach and Jordan, 2002; Fine and Scheinberg, 2002) or Kernel Matching Pursuit (KMP) (Hussain and Shawe-Taylor, 2008; Vincent and Bengio, 2000) have also been widely used in large-scale learning applications. But, how does the kernel approximation affect the performance of the learning algorithm?

There exists some previous work on this subject. Spectral clustering with perturbed data was studied in a restrictive setting with several assumption by Huang *et al.* (2008). In Fine and Scheinberg (2002), the authors address this question in terms of the impact on the value of the *objective function* to be optimized by the learning algorithm. However, we strive to take the question one step further and directly analyze the effect of an approximation in the kernel matrix on the *hypothesis* generated by several widely used kernel-based learning algorithms.

We give stability bounds based on the norm of the kernel approximation for these algorithms, including SVMs, KRR, and graph Laplacian-based regularization algorithms (Belkin *et al.*, 2004). These bounds help determine the degree of approximation that can be tolerated in the estimation of the kernel matrix. Our analysis differs from previous applications of stability analysis as put forward by Bousquet and Elisseeff (2001). Instead of studying the effect of changing one training point, we study the effect of changing the kernel matrix. Our analysis is general and applies to arbitrary approximations of the kernel matrix. However, we also give a specific analysis of the Nyström

low-rank approximation given the recent interest in this method and the successful applications of this algorithm to large-scale applications. We also report the results of experiments evaluating the quality of this kernel approximation when used with ridge regression.

The remainder of this paper is organized as follows. Section 2 introduces the problem of kernel stability and gives a kernel stability analysis of several algorithms. Section 3 provides a brief review of the Nyström approximation method and gives error bounds that can be combined with the kernel stability bounds. Section 4 reports the results of experiments with kernel approximation combined with kernel ridge regression.

2 Kernel Stability Analysis

In this section we analyze the impact of kernel approximation on several common kernel-based learning algorithms: KRR, SVM and graph Laplacian-based regularization algorithms. Our stability analyses result in bounds on the hypotheses directly in terms of the quality of the kernel approximation. In our analysis we assume that the kernel approximation is only used during training where the kernel approximation may serve to reduce resource requirements. At testing time the true kernel function is used. This scenario that we are considering is standard for the Nyström method and other approximations.

We consider the standard supervised learning setting where the learning algorithm receives a sample of m labeled points $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$, where X is the input space and Y the set of labels, $Y = \mathbb{R}$ with $|y| \leq M$ in the regression case, and $Y = \{-1, +1\}$ in the classification case. Throughout the paper the kernel matrix $\mathbf{K}$ and its approximation $\mathbf{K}'$ are assumed to be symmetric, positive, and semi-definite (SPSD).

2.1 Kernel Ridge Regression

We first provide a stability analysis of kernel ridge regression. The following is the dual optimization problem solved by KRR (Saunders *et al.*, 1998):

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^m} \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha} \mathbf{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^\top \mathbf{y}, \tag{1}$$

where $\lambda = m\lambda_0 > 0$ is the ridge parameter. The problem admits the closed form solution $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}$. We denote by h the hypothesis returned by kernel ridge regression when using the exact kernel matrix.

Proposition 1. *Let h' denote the hypothesis returned by kernel ridge regression when using the approximate kernel matrix $\mathbf{K}' \in \mathbb{R}^{m \times m}$. Furthermore, define $\kappa > 0$ such that $K(x, x) \leq \kappa$ and $K'(x, x) \leq \kappa$ for all $x \in X$. This condition is verified with $\kappa = 1$ for Gaussian kernels for example. Then the following inequalities hold for all $x \in X$,*

$$|h'(x) - h(x)| \leq \frac{\kappa M}{\lambda_0^2 m} \|\mathbf{K}' - \mathbf{K}\|_2. \tag{2}$$

Proof. Let $\boldsymbol{\alpha}'$ denote the solution obtained using the approximate kernel matrix $\mathbf{K}'$. We can write

$$\begin{aligned} \boldsymbol{\alpha}' - \boldsymbol{\alpha} &= (\mathbf{K}' + \lambda \mathbf{I})^{-1}\mathbf{y} - (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y} && (3) \\ &= -\big[(\mathbf{K}' + \lambda \mathbf{I})^{-1}(\mathbf{K}' - \mathbf{K})(\mathbf{K} + \lambda \mathbf{I})^{-1}\big]\mathbf{y}, && (4) \end{aligned}$$

where we used the identity $\mathbf{M}'^{-1} - \mathbf{M}^{-1} = -\mathbf{M}'^{-1}(\mathbf{M}' - \mathbf{M})\mathbf{M}^{-1}$ valid for any invertible matrices $\mathbf{M}, \mathbf{M}'$. Thus, $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|$ can be bounded as follows:

$$\begin{aligned} \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| &\leq \|(\mathbf{K}' + \lambda \mathbf{I})^{-1}\| \|\mathbf{K}' - \mathbf{K}\| \|(\mathbf{K} + \lambda \mathbf{I})^{-1}\| \|\mathbf{y}\| \\ &\leq \frac{\|\mathbf{K}' - \mathbf{K}\|_2 \, \|\mathbf{y}\|}{\lambda_{\min}(\mathbf{K}' + \lambda \mathbf{I}) \lambda_{\min}(\mathbf{K} + \lambda \mathbf{I})}, \end{aligned} \tag{5}$$

where $\lambda_{\min}(\mathbf{K}' + \lambda \mathbf{I})$ is the smallest eigenvalue of $\mathbf{K}' + \lambda \mathbf{I}$ and $\lambda_{\min}(\mathbf{K} + \lambda \mathbf{I})$ the smallest eigenvalue of $\mathbf{K} + \lambda \mathbf{I}$. The hypothesis h derived with the exact kernel matrix is defined by $h(x) = \sum_{i=1}^m \alpha_i K(x, x_i) = \boldsymbol{\alpha}^\top \mathbf{k}_x$, where $\mathbf{k}_x = (K(x, x_1), \ldots, K(x, x_m))^\top$. By assumption, no approximation affects $\mathbf{k}_x$, thus the approximate hypothesis h' is given by $h'(x) = \boldsymbol{\alpha}'^\top \mathbf{k}_x$ and

$$|h'(x) - h(x)| \leq \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| \|\mathbf{k}_x\| \leq \kappa \sqrt{m} \|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|. \tag{6}$$

Using the bound on $\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\|$ given by inequality (5), the fact that the eigenvalues of $(\mathbf{K}' + \lambda \mathbf{I})$ and $(\mathbf{K} + \lambda \mathbf{I})$ are larger than or equal to λ since $\mathbf{K}$ and $\mathbf{K}'$ are SPSD matrices, and $\|\mathbf{y}\| \leq \sqrt{m} M$ yields

$$\begin{aligned} |h'(x) - h(x)| &\leq \frac{\kappa m M \|\mathbf{K}' - \mathbf{K}\|_2}{\lambda_{\min}(\mathbf{K}' + \lambda \mathbf{I}) \lambda_{\min}(\mathbf{K} + \lambda \mathbf{I})} \\ &\leq \frac{\kappa M}{\lambda_0^2 m} \|\mathbf{K}' - \mathbf{K}\|_2. \end{aligned}$$

□

The generalization bounds for KRR, e.g., stability bounds (Bousquet and Elisseeff, 2001), are of the form $R(h) \leq \widehat{R}(h) + O(1/\sqrt{m})$, where $R(h)$ denotes the generalization error and $\widehat{R}(h)$ the empirical error of a hypothesis h with respect to the square loss. The proposition shows that $|h'(x) - h(x)|^2 = O(\|\mathbf{K}' - \mathbf{K}\|_2^2 / \lambda_0^4 m^2)$. Thus, it suggests that the kernel approximation tolerated should be such that $\|\mathbf{K}' - \mathbf{K}\|_2^2 / \lambda_0^4 m^2 \ll O(1/\sqrt{m})$, that is, such that $\|\mathbf{K}' - \mathbf{K}\|_2 \ll O(\lambda_0^2 m^{3/4})$.

Note that the main bound used in the proof of the theorem, inequality (5), is tight in the sense that it can be matched for some kernels K and K'. Indeed, let K and K' be the

kernel functions defined by $K(x,y)=\beta$ and $K'(x,y)=\beta'$ if $x=y$, $K'(x,y)=K(x,y)=0$ otherwise, with $\beta,\beta'\geq 0$. Then, the corresponding kernel matrices for a sample S are $\mathbf{K}=\beta\mathbf{I}$ and $\mathbf{K}'=\beta'\mathbf{I}$, and the dual parameter vectors are given by $\boldsymbol{\alpha}=\mathbf{y}/(\beta+\lambda)$ and $\boldsymbol{\alpha}'=\mathbf{y}/(\beta'+\lambda)$. Now, since $\lambda_{\min}(\mathbf{K}'+\lambda\mathbf{I})=\beta'+\lambda$ and $\lambda_{\min}(\mathbf{K}+\lambda\mathbf{I})=\beta+\lambda$, and $\|\mathbf{K}'-\mathbf{K}\|=\beta'-\beta$, the following equality holds:

$$\|\boldsymbol{\alpha}'-\boldsymbol{\alpha}\| = \frac{|\beta'-\beta|}{(\beta'+\lambda)(\beta+\lambda)}\|\mathbf{y}\| \tag{7}$$

$$= \frac{\|\mathbf{K}'-\mathbf{K}\|}{\lambda_{\min}(\mathbf{K}'+\lambda\mathbf{I})\lambda_{\min}(\mathbf{K}'+\lambda\mathbf{I})}\|\mathbf{y}\|. \tag{8}$$

This limits significant improvements of the bound of Proposition 1 using similar techniques.

2.2 Support Vector Machines

This section analyzes the kernel stability of SVMs. For simplicity, we shall consider the case where the classification function sought has no offset. In practice, this corresponds to using a constant feature. Let $\Phi\colon X\to F$ denote a feature mapping from the input space X to a Hilbert space F corresponding to some kernel K. The hypothesis set we consider is thus $H=\{h\colon \exists\mathbf{w}\in F|\forall x\in X, h(x)=\mathbf{w}^\top\Phi(x)\}$.

The following is the standard primal optimization problem for SVMs:

$$\min_{\mathbf{w}} F_K(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C_0\widehat{R}_K(\mathbf{w}), \tag{9}$$

where $\widehat{R}_K(\mathbf{w})=\frac{1}{m}\sum_{i=1}^m L(y_i\mathbf{w}^\top\Phi(x_i))$ is the empirical error, with $L(y_i\mathbf{w}^\top\Phi(x_i))=\max(0,1-y_i\mathbf{w}^\top\Phi(x_i))$ the hinge loss associated to the ith point.

In the following, we analyze the difference between the hypothesis h returned by SVMs when trained on the sample S of m points and using a kernel K, versus the hypothesis h' obtained when training on the same sample with the kernel K'. For a fixed $x\in X$, we shall compare more specifically $h(x)$ and $h'(x)$. Thus, we can work with the finite set $X_{m+1}=\{x_1,\ldots,x_m,x_{m+1}\}$, with $x_{m+1}=x$.

Different feature mappings Φ can be associated to the same kernel K. To compare the solutions $\mathbf{w}$ and $\mathbf{w}'$ of the optimization problems based on F_K and $F_{K'}$, we can choose the feature mappings Φ and Φ' associated to K and K' such that they both map to $\mathbb{R}^{m+1}$ as follows. Let $\mathbf{K}_{m+1}$ denote the Gram matrix associated to K and $\mathbf{K}'_{m+1}$ that of kernel K' for the set of points X_{m+1}. Then for all $u\in X_{m+1}$, Φ and Φ' can be defined by

$$\Phi(u) = \mathbf{K}_{m+1}^{*1/2}\begin{bmatrix}K(x_1,u)\\ \vdots \\ K(x_{m+1},u)\end{bmatrix} \tag{10}$$

$$\text{and}\quad \Phi'(u) = \mathbf{K}'^{*1/2}_{m+1}\begin{bmatrix}K'(x_1,u)\\ \vdots \\ K'(x_{m+1},u)\end{bmatrix}, \tag{11}$$

where $\mathbf{K}^*_{m+1}$ denotes the pseudo-inverse of $\mathbf{K}_{m+1}$ and $\mathbf{K}'^*_{m+1}$ that of $\mathbf{K}'_{m+1}$. It is not hard to see then that for all $u,v\in X_{m+1}$, $K(u,v)=\Phi(u)^\top\Phi(v)$ and $K'(u,v)=\Phi'(u)^\top\Phi'(v)$ (Schölkopf and Smola, 2002). Since the optimization problem depends only on the sample S, we can use the feature mappings just defined in the expression of F_K and $F_{K'}$. This does not affect in any way the standard SVMs optimization problem.

Let $\mathbf{w}\in\mathbb{R}^{m+1}$ denote the minimizer of F_K and $\mathbf{w}'$ that of $F_{K'}$. By definition, if we let $\Delta\mathbf{w}$ denote $\mathbf{w}'-\mathbf{w}$, for all $s\in[0,1]$, the following inequalities hold:

$$F_K(\mathbf{w}) \leq F_K(\mathbf{w}+s\Delta\mathbf{w}) \tag{12}$$

$$\text{and}\quad F_{K'}(\mathbf{w}') \leq F_{K'}(\mathbf{w}'-s\Delta\mathbf{w}). \tag{13}$$

Summing these two inequalities, rearranging terms, and using the identity $(\|\mathbf{w}+s\Delta\mathbf{w}\|^2-\|\mathbf{w}\|^2)+(\|\mathbf{w}'-s\Delta\mathbf{w}\|^2-\|\mathbf{w}'\|^2)=-2s(1-s)\|\Delta\mathbf{w}\|^2$, we obtain as in (Bousquet and Elisseeff, 2001):

$$s(1-s)\|\Delta\mathbf{w}\|^2 \leq C_0\Big[\big(\widehat{R}_K(\mathbf{w}+s\Delta\mathbf{w})-\widehat{R}_K(\mathbf{w})\big) + \big(\widehat{R}_{K'}(\mathbf{w}'-s\Delta\mathbf{w})-\widehat{R}_{K'}(\mathbf{w}')\big)\Big].$$

Note that $\mathbf{w}+s\Delta\mathbf{w}=s\mathbf{w}'+(1-s)\mathbf{w}$ and $\mathbf{w}'-s\Delta\mathbf{w}=s\mathbf{w}+(1-s)\mathbf{w}'$. Then, by the convexity of the hinge loss and thus $\widehat{R}_K$ and $\widehat{R}_{K'}$, the following inequalities hold:

$$\widehat{R}_K(\mathbf{w}+s\Delta\mathbf{w})-\widehat{R}_K(\mathbf{w})\leq s(\widehat{R}_K(\mathbf{w}')-\widehat{R}_K(\mathbf{w}))$$
$$\widehat{R}_{K'}(\mathbf{w}'-s\Delta\mathbf{w})-\widehat{R}_{K'}(\mathbf{w}')\leq -s(\widehat{R}_{K'}(\mathbf{w}')-\widehat{R}_{K'}(\mathbf{w})).$$

Plugging in these inequalities on the left-hand side, simplifying by s and taking the limit $s\to 0$ yields

$$\|\Delta\mathbf{w}\|^2 \leq C_0\Big[\big(\widehat{R}_K(\mathbf{w}')-\widehat{R}_{K'}(\mathbf{w}')\big)+\big(\widehat{R}_{K'}(\mathbf{w})-\widehat{R}_K(\mathbf{w})\big)\Big]$$
$$= \frac{C_0}{m}\sum_{i=1}^m\Big[\Big(L(y_i\mathbf{w}'^\top\Phi(x_i))-L(y_i\mathbf{w}'^\top\Phi'(x_i))\Big) + \Big(L(y_i\mathbf{w}^\top\Phi'(x_i))-L(y_i\mathbf{w}^\top\Phi(x_i))\Big)\Big],$$

where the last inequality results from the definition of the empirical error. Since the hinge loss is 1-Lipschitz, we can

bound the terms on the right-hand side as follows:

$$\|\Delta\mathbf{w}\|^2 \leq \frac{C_0}{m}\sum_{i=1}^{m}\Big[\|\mathbf{w}'\|\|\Phi'(x_i)-\Phi(x_i)\| + \|\mathbf{w}\|\|\Phi'(x_i)-\Phi(x_i)\|\Big] \quad (14)$$

$$= \frac{C_0}{m}\sum_{i=1}^{m}(\|\mathbf{w}'\|+\|\mathbf{w}\|)\,\|\Phi'(x_i)-\Phi(x_i)\|. \quad (15)$$

Let e_i denote the ith unit vector of $\mathbb{R}^{m+1}$, then $(K(x_1,x_i),\ldots,K(x_{m+1},x_i))^\top = \mathbf{K}_{m+1}e_i$. Thus, in view of the definition of Φ, for all $i\in[1,m+1]$,

$$\Phi(x_i) = \mathbf{K}_{m+1}^{*1/2}[K(x_1,x_i),\ldots,K(x_m,x_i),K(x,x_i)]^\top = \mathbf{K}_{m+1}^{*1/2}\mathbf{K}_{m+1}e_i = \mathbf{K}_{m+1}^{1/2}e_i, \quad (16)$$

and similarly $\Phi'(x_i)=\mathbf{K}_{m+1}'^{1/2}e_i$. $\mathbf{K}_{m+1}^{1/2}e_i$ is the ith column of $\mathbf{K}_{m+1}^{1/2}$ and similarly $\mathbf{K}'^{1/2}e_i$ the ith column of $\mathbf{K}_{m+1}'^{1/2}$. Thus, (15) can be rewritten as

$$\|\mathbf{w}'-\mathbf{w}\|^2 \leq \frac{C_0}{m}\sum_{i=1}^{m}\left(\|\mathbf{w}'\|+\|\mathbf{w}\|\right)\|(\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2})e_i\|.$$

As for the case of ridge regression, we shall assume that there exists $\kappa>0$ such that $K(x,x)\leq\kappa$ and $K'(x,x)\leq\kappa$ for all $x\in X_{m+1}$. Now, since $\mathbf{w}$ can be written in terms of the dual variables $0\leq\alpha_i\leq C$, $C=C_0/m$ as $\mathbf{w}=\sum_{i=1}^m\alpha_i K(x_i,\cdot)$, it can be bounded as $\|\mathbf{w}\|\leq mC_0/m\kappa^{1/2}=\kappa^{1/2}C_0$ and similarly $\|\mathbf{w}'\|\leq\kappa^{1/2}C_0$. Thus, we can write

$$\|\mathbf{w}'-\mathbf{w}\|^2 \leq \frac{2C_0^2\kappa^{1/2}}{m}\sum_{i=1}^{m}\|(\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2})e_i\| \leq \frac{2C_0^2\kappa^{1/2}}{m}\sum_{i=1}^{m}\|(\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2})\|\|e_i\| = 2C_0^2\kappa^{1/2}\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|. \quad (17)$$

Let $\mathbf{K}$ denote the Gram matrix associated to K and $\mathbf{K}'$ that of kernel K' for the sample S. Then, the following result holds.

Proposition 2. *Let h' denote the hypothesis returned by SVMs when using the approximate kernel matrix $\mathbf{K}'\in\mathbb{R}^{m\times m}$. Then, the following inequality holds for all $x\in\mathcal{X}$:*

$$|h'(x)-h(x)| \leq \sqrt{2}\kappa^{\frac{3}{4}}C_0\|\mathbf{K}'-\mathbf{K}\|_2^{\frac{1}{4}}\left[1+\left[\frac{\|\mathbf{K}'-\mathbf{K}\|_2}{4\kappa}\right]^{\frac{1}{4}}\right]. \quad (18)$$

Proof. In view of (16) and (17), the following holds:

$$\begin{aligned}
&|h'(x)-h(x)|\\
&= \|\mathbf{w}'^\top\Phi'(x)-\mathbf{w}^\top\Phi(x)\|\\
&= \|(\mathbf{w}'-\mathbf{w})^\top\Phi'(x)+\mathbf{w}^\top(\Phi'(x)-\Phi(x))\|\\
&\leq \|\mathbf{w}'-\mathbf{w}\|\|\Phi'(x)\|+\|\mathbf{w}\|\|\Phi'(x)-\Phi(x)\|\\
&= \|\mathbf{w}'-\mathbf{w}\|\|\Phi'(x)\|+\|\mathbf{w}\|\|\Phi'(x_{m+1})-\Phi(x_{m+1})\|\\
&\leq \left(2C_0^2\kappa^{1/2}\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|\right)^{1/2}\kappa^{1/2}\\
&\quad+\kappa^{1/2}C_0\|(\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2})e_{m+1}\|\\
&\leq \sqrt{2}\kappa^{3/4}C_0\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|^{1/2}\\
&\quad+\kappa^{1/2}C_0\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|.
\end{aligned}$$

Now, by Lemma 1 (see Appendix), $\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|_2\leq\|\mathbf{K}_{m+1}'-\mathbf{K}_{m+1}\|_2^{1/2}$. By assumption, the kernel approximation is only used at training time so $K(x,x_i)=K'(x,x_i)$, for all $i\in[1,m]$, and since by definition $x=x_{m+1}$, the last rows or the last columns of the matrices $\mathbf{K}_{m+1}'$ and $\mathbf{K}_{m+1}$ coincide. Therefore, the matrix $\mathbf{K}_{m+1}'-\mathbf{K}_{m+1}$ coincides with the matrix $\mathbf{K}'-\mathbf{K}$ bordered with a zero-column and zero-row and $\|\mathbf{K}_{m+1}'^{1/2}-\mathbf{K}_{m+1}^{1/2}\|_2\leq\|\mathbf{K}'-\mathbf{K}\|_2^{1/2}$. Thus,

$$|h'(x)-h(x)| \leq \sqrt{2}\kappa^{3/4}C_0\|\mathbf{K}'-\mathbf{K}\|^{1/4} + \kappa^{1/2}C_0\|\mathbf{K}'-\mathbf{K}\|^{1/2}, \quad (19)$$

which is exactly the statement of the proposition. □

Since the hinge loss l is 1-Lipschitz, Proposition 2 leads directly to the following bound on the pointwise difference of the hinge loss between the hypotheses h' and h.

Corollary 1. *Let h' denote the hypothesis returned by SVMs when using the approximate kernel matrix $\mathbf{K}'\in\mathbb{R}^{m\times m}$. Then, the following inequality holds for all $x\in\mathcal{X}$ and $y\in\mathcal{Y}$:*

$$\left|L\big(yh'(x)\big)-L\big(yh(x)\big)\right| \leq \sqrt{2}\kappa^{\frac{3}{4}}C_0\|\mathbf{K}'-\mathbf{K}\|_2^{\frac{1}{4}}\left[1+\left[\frac{\|\mathbf{K}'-\mathbf{K}\|_2}{4\kappa}\right]^{\frac{1}{4}}\right]. \quad (20)$$

The bounds we obtain for SVMs are weaker than our bound for KRR. This is due mainly to the different loss functions defining the optimization problems of these algorithms.

2.3 Graph Laplacian regularization algorithms

We lastly study the kernel stability of graph-Laplacian regularization algorithms such as that of Belkin *et al.* (2004). Given a connected weighted graph $G=(X,E)$ in which

edge weights can be interpreted as similarities between vertices, the task consists of predicting the vertex labels of u vertices using a labeled training sample S of m vertices. The input space $\mathcal{X}$ is thus reduced to the set of vertices, and a hypothesis $h\colon \mathcal{X} \to \mathbb{R}$ can be identified with the finite-dimensional vector $\mathbf{h}$ of its predictions $\mathbf{h} = [h(x_1), \ldots, h(x_{m+u})]^\top$. The hypothesis set H can thus be identified with $\mathbb{R}^{m+u}$ here. Let $\mathbf{h}_S$ denote the restriction of $\mathbf{h}$ to the training points, $[h(x_1), \ldots, h(x_m)]^\top \in \mathbb{R}^m$, and similarly let $\mathbf{y}_S$ denote $[y_1, \ldots, y_m]^\top \in \mathbb{R}^m$. Then, the following is the optimization problem corresponding to this problem:

$$\begin{aligned} \min_{\mathbf{h} \in H} \quad & \mathbf{h}^\top \mathbf{L} \mathbf{h} + \frac{C_0}{m} (\mathbf{h}_S - \mathbf{y}_S)^\top (\mathbf{h}_S - \mathbf{y}_S) \\ \text{subject to} \quad & \mathbf{h}^\top \mathbf{1} = 0, \end{aligned} \tag{21}$$

where $\mathbf{L}$ is the graph Laplacian and $\mathbf{1}$ the column vector with all entries equal to 1. Thus, $\mathbf{h}^\top \mathbf{L} \mathbf{h} = \sum_{ij=1}^m w_{ij} (h(x_i) - h(x_j))^2$, for some weight matrix (w_{ij}). The label vector $\mathbf{y}$ is assumed to be centered, which implies that $\mathbf{1}^\top \mathbf{y} = 0$. Since the graph is connected, the eigenvalue zero of the Laplacian has multiplicity one.

Define $\mathbf{I}_S \in \mathbb{R}^{(m+u)\times(m+u)}$ to be the diagonal matrix with $[\mathbf{I}_S]_{i,i} = 1$ if $i \leq m$ and 0 otherwise. Maintaining the notation used in Belkin *et al.* (2004), we let $\mathbf{P}_H$ denote the projection on the hyperplane H orthogonal to $\mathbf{1}$ and let $\mathbf{M} = \mathbf{P}_H \left(\frac{m}{C_0} \mathbf{L} + \mathbf{I}_S \right)$ and $\mathbf{M}' = \mathbf{P}_H \left(\frac{m}{C_0} \mathbf{L}' + \mathbf{I}_S \right)$. We denote by $\mathbf{h}$ the hypothesis returned by the algorithm when using the exact kernel matrix $\mathbf{L}$ and by $\mathbf{L}'$ an approximate graph Laplacian such that $\mathbf{h}^\top \mathbf{L}' \mathbf{h} = \sum_{ij=1}^m w'_{ij} (h(x_i) - h(x_j))^2$, based on matrix (w'_{ij}) instead of (w_{ij}). We shall assume that there exist $M > 0$ such that $y_i \leq M$ for $i \in [1, m]$.

Proposition 3. *Let $\mathbf{h}'$ denote the hypothesis returned by the graph-Laplacian regularization algorithm when using an approximate Laplacian $\mathbf{L}' \in \mathbb{R}^{m\times m}$. Then, the following inequality holds:*

$$\|\mathbf{h}' - \mathbf{h}\| \leq \frac{m^{3/2} M / C_0}{(\frac{m}{C_0} \widehat{\lambda}_2 - 1)^2} \|\mathbf{L}' - \mathbf{L}\|, \tag{22}$$

where $\widehat{\lambda}_2 = \max\{\lambda_2, \lambda'_2\}$ with λ_2 denoting the second smallest eigenvalue of the kernel matrix $\mathbf{L}$ and λ'_2 the second smallest eigenvalue of $\mathbf{L}'$.

Proof. The closed-form solution of Problem 21 is given by Belkin *et al.* (2004): $\mathbf{h} = \left(\mathbf{P}_H \left(\frac{m}{C_0} \mathbf{L} + \mathbf{I}_S \right) \right)^{-1} \mathbf{y}_S$. Thus, we can use that expression and the matrix identity for $(\mathbf{M}^{-1} - \mathbf{M}'^{-1})$ we already used in the analysis of KRR to write

$$\begin{aligned} \|\mathbf{h} - \mathbf{h}'\| &= \|\mathbf{M}^{-1} \mathbf{y}_S - \mathbf{M}'^{-1} \mathbf{y}_S\| && (23)\\ &= \|(\mathbf{M}^{-1} - \mathbf{M}'^{-1}) \mathbf{y}_S\| && (24)\\ &= \|-\mathbf{M}^{-1} (\mathbf{M} - \mathbf{M}') \mathbf{M}'^{-1} \mathbf{y}_S\| && (25)\\ &\leq \frac{m}{C_0} \|-\mathbf{M}^{-1} (\mathbf{L} - \mathbf{L}') \mathbf{M}'^{-1} \mathbf{y}_S\| && (26)\\ &\leq \frac{m}{C_0} \|\mathbf{M}^{-1}\| \, \|\mathbf{M}'^{-1}\| \, \|\mathbf{y}_S\| \, \|\mathbf{L}' - \mathbf{L}\|. && (27) \end{aligned}$$

For any column matrix $\mathbf{z} \in \mathbb{R}^{(m+u)\times 1}$, by the triangle inequality and the projection property $\|\mathbf{P}_H \mathbf{z}\| \leq \|\mathbf{z}\|$, the following inequalities hold:

$$\begin{aligned} \|\frac{m}{C_0} \mathbf{P}_H \mathbf{L}\| &= \|\frac{m}{C_0} \mathbf{P}_H \mathbf{L} + \mathbf{P}_H \mathbf{I}_S \mathbf{z} - P_H \mathbf{I}_S \mathbf{z}\| && (28)\\ &\leq \|\frac{m}{C_0} \mathbf{P}_H \mathbf{L} + \mathbf{P}_H \mathbf{I}_S \mathbf{z}\| + \|\mathbf{P}_H \mathbf{I}_S \mathbf{z}\| && (29)\\ &\leq \|\mathbf{P}_H \Big(\frac{m}{C_0} \mathbf{L} + \mathbf{I}_S \Big) \mathbf{z}\| + \|\mathbf{I}_S \mathbf{z}\|. && (30) \end{aligned}$$

This yields the lower bound:

$$\begin{aligned} \|\mathbf{M}\mathbf{z}\| &= \|\mathbf{P}_H \left(\frac{m}{C_0} \mathbf{L} + \mathbf{I}_S \right) \mathbf{z}\| && (31)\\ &\geq \frac{m}{C_0} \|\mathbf{P}_H \mathbf{L}\| - \|\mathbf{I}_S \mathbf{z}\| && (32)\\ &\geq \left(\frac{m}{C_0} \lambda_2 - 1 \right) \|\mathbf{z}\|, && (33) \end{aligned}$$

which gives the following upper bounds on $\|\mathbf{M}^{-1}\|$ and $\|\mathbf{M}'^{-1}\|$:

$$\|\mathbf{M}^{-1}\| \leq \frac{1}{\frac{m}{C_0} \lambda_2 - 1} \quad \text{and} \quad \|\mathbf{M}'^{-1}\| \leq \frac{1}{\frac{m}{C_0} \lambda'_2 - 1}.$$

Plugging in these inequalities in (27) and using $\|\mathbf{y}_S\| \leq m^{1/2} M$ lead to

$$\|\mathbf{h} - \mathbf{h}'\| \leq \frac{m^{3/2} M / C_0}{(\frac{m}{C_0} \lambda_2 - 1)(\frac{m}{C_0} \lambda'_2 - 1)} \|\mathbf{L}' - \mathbf{L}\|. \qquad \square$$

The generalization bounds for the graph-Laplacian algorithm are of the form $R(h) \leq \widehat{R}(h) + O(\frac{m}{(\frac{m}{C_0} \lambda_2 - 1)^2})$ (Belkin *et al.*, 2004). In view of the bound given by the theorem, this suggests that the approximation tolerated should verify $\|\mathbf{L}' - \mathbf{L}\| \ll O(1/\sqrt{m})$.

3 Application to Nyström method

The previous section provided stability analyses for several common learning algorithms studying the effect of using an approximate kernel matrix instead of the true one. The difference in hypothesis value is expressed simply in terms of the difference between the kernels measured by

Dataset	Description	# Points (m)	# Features (d)	Kernel	Largest label (M)
ABALONE	physical attributes of abalones	4177	8	RBF	29
KIN-8nm	kinematics of robot arm	4000	8	RBF	1.5

Table 1: Summary of datasets used in our experiments (Asuncion and Newman, 2007; Ghahramani, 1996).

some norm. Although these bounds are general bounds that are independent of how the approximation is obtained (so long as $\mathbf{K}'$ remains SPSD), one relevant application of these bounds involves the Nyström method.

As shown by Williams and Seeger (2000), later by Drineas and Mahoney (2005); Talwalkar *et al.* (2008); Zhang *et al.* (2008), low-rank approximations of the kernel matrix via the Nyström method can provide an effective technique for tackling large-scale data sets. However, all previous theoretical work analyzing the performance of the Nyström method has focused on the quality of the low-rank approximations, rather than the performance of the kernel learning algorithms used in conjunction with these approximations. In this section, we first provide a brief review of the Nyström method and then show how we can leverage the analysis of Section 2 to present novel performance guarantees for the Nyström method in the context of kernel learning algorithms.

3.1 Nyström method

The Nyström approximation of a symmetric positive semidefinite (SPSD) matrix $\mathbf{K}$ is based on a sample of $n \ll m$ columns of $\mathbf{K}$ (Drineas and Mahoney, 2005; Williams and Seeger, 2000). Let $\mathbf{C}$ denote the $m \times n$ matrix formed by these columns and $\mathbf{W}$ the $n \times n$ matrix consisting of the intersection of these n columns with the corresponding n rows of $\mathbf{K}$. The columns and rows of $\mathbf{K}$ can be rearranged based on this sampling so that $\mathbf{K}$ and $\mathbf{C}$ be written as follows:

$$\mathbf{K} = \begin{bmatrix} \mathbf{W} & \mathbf{K}_{21}^\top \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{K}_{21} \end{bmatrix}. \tag{34}$$

Note that $\mathbf{W}$ is also SPSD since $\mathbf{K}$ is SPSD. For a uniform sampling of the columns, the Nyström method generates a rank-k approximation $\widetilde{\mathbf{K}}$ of $\mathbf{K}$ for $k \leq n$ defined by:

$$\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}_k^+\mathbf{C}^\top \approx \mathbf{K}, \tag{35}$$

where $\mathbf{W}_k$ is the best k-rank approximation of $\mathbf{W}$ for the Frobenius norm, that is $\mathbf{W}_k = \operatorname{argmin}_{\operatorname{rank}(\mathbf{V})=k} \|\mathbf{W} - \mathbf{V}\|_F$ and $\mathbf{W}_k^+$ denotes the pseudo-inverse of $\mathbf{W}_k$. $\mathbf{W}_k^+$ can be derived from the singular value decomposition (SVD) of $\mathbf{W}$, $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$, where $\mathbf{U}$ is orthonormal and $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$ is a real diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$. For $k \leq \operatorname{rank}(\mathbf{W})$, it is given by $\mathbf{W}_k^+ = \sum_{i=1}^k \sigma_i^{-1}\mathbf{U}^i{\mathbf{U}^i}^\top$, where $\mathbf{U}^i$ denotes the ith column of $\mathbf{U}$. Since the running time complexity of SVD is $O(n^3)$ and $O(nmk)$ is required for multiplication with $\mathbf{C}$, the total complexity of the Nyström approximation computation is $O(n^3 + nmk)$.

3.2 Nyström kernel ridge regression

The accuracy of low-rank Nyström approximations has been theoretically analyzed by Drineas and Mahoney (2005); Kumar *et al.* (2009c). The following theorem, adapted from Drineas and Mahoney (2005) for the case of uniform sampling, gives an upper bound on the norm-2 error of the Nyström approximation of the form $\|\mathbf{K} - \widetilde{\mathbf{K}}\|_2 / \|\mathbf{K}\|_2 \leq \|\mathbf{K} - \mathbf{K}_k\|_2 / \|\mathbf{K}\|_2 + O(1/\sqrt{n})$. We denote by $\mathbf{K}_{\max}$ the maximum diagonal entry of $\mathbf{K}$.

Theorem 1. *Let $\widetilde{\mathbf{K}}$ denote the rank-k Nyström approximation of $\mathbf{K}$ based on n columns sampled uniformly at random with replacement from $\mathbf{K}$, and $\mathbf{K}_k$ the best rank-k approximation of $\mathbf{K}$. Then, with probability at least $1 - \delta$, the following inequalities hold for any sample of size n:*

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\|_2 \leq \|\mathbf{K} - \mathbf{K}_k\|_2 + \tfrac{m}{\sqrt{n}} \mathbf{K}_{\max} \big(2 + \log \tfrac{1}{\delta}\big).$$

Theorem 1 focuses on the quality of low-rank approximations. Combining the analysis from Section 2 with this theorem enables us to bound the relative performance of the kernel learning algorithms when the Nyström method is used as a means of scaling kernel learning algorithms. To illustrate this point, Theorem 2 uses Proposition 1 along with Theorem 1 to upper bound the relative performance of kernel ridge regression as a function of the approximation accuracy of the Nyström method.

Theorem 2. *Let h' denote the hypothesis returned by kernel ridge regression when using the approximate rank-k kernel $\widetilde{\mathbf{K}} \in \mathbb{R}^{n \times n}$ generated using the Nyström method. Then, with probability at least $1 - \delta$, the following inequality holds for all $x \in X$,*

$$|h'(x) - h(x)| \leq \frac{\kappa M}{\lambda_0^2 m} \Big[\|\mathbf{K} - \mathbf{K}_k\|_2 + \tfrac{m}{\sqrt{n}} \mathbf{K}_{\max} \big(2 + \log \tfrac{1}{\delta}\big) \Big].$$

A similar technique can be used to bound the error of the Nyström approximation when used with the other algorithms discussed in Section 2. The results are omitted due to space constraints.

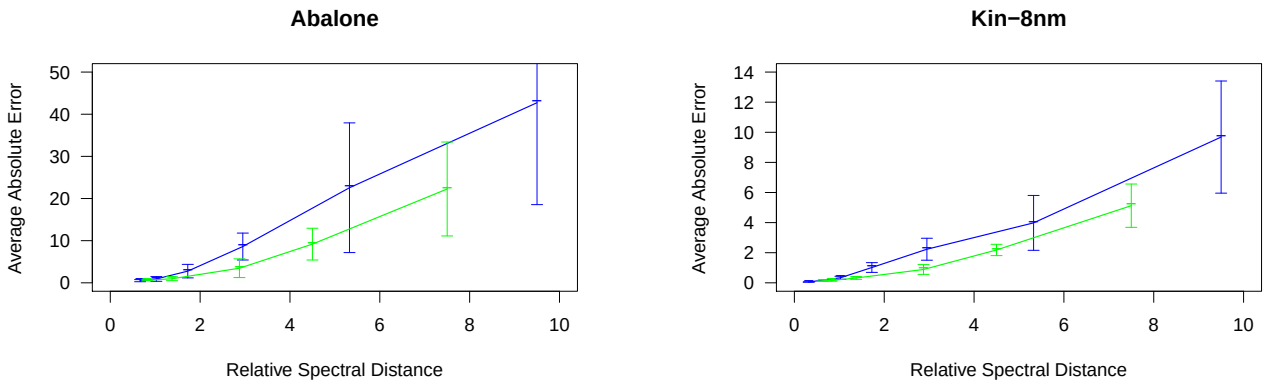

Figure 1: Average absolute error of the kernel ridge regression hypothesis, $h'(\cdot)$, generated from the Nyström approximation, $\widetilde{\mathbf{K}}$, as a function of relative spectral distance $\|\widetilde{\mathbf{K}} - \mathbf{K}\|_2/\|\mathbf{K}\|_2$. For each dataset, the reported results show the average absolute error as a function of relative spectral distance for both the full dataset and for a subset of the data containing $m = 2000$ points. Results for the same value of m are connected with a line. The different points along the lines correspond to various numbers of sampled columns, i.e., n ranging from 1% to 50% of m.

3.3 Nyström Woodbury Approximation

The Nyström method provides an effective algorithm for obtaining a rank-k approximation for the kernel matrix. As suggested by Williams and Seeger (2000) in the context of Gaussian Processes, this approximation can be combined with the Woodbury inversion lemma to derive an efficient algorithm for inverting the kernel matrix. The Woodbury inversion Lemma states that the inverse of a rank-k correction of some matrix can be computed by doing a rank-k correction to the inverse of the original matrix. In the context of KRR, using the rank-k approximation $\widetilde{\mathbf{K}}$ given by the Nyström method, instead of $\mathbf{K}$, and applying the inversion lemma yields

$$(\lambda \mathbf{I} + \mathbf{K})^{-1} \tag{36}$$

$$\approx \left(\lambda \mathbf{I} + \mathbf{C}\mathbf{W}_k^{+}\mathbf{C}^\top\right)^{-1} \tag{37}$$

$$= \frac{1}{\lambda}\Big(\mathbf{I} - \mathbf{C}\big[\lambda \mathbf{I}_k + \mathbf{W}_k^{+}\mathbf{C}^\top\mathbf{C}\big]^{-1}\mathbf{W}_k^{+}\mathbf{C}^\top\Big). \tag{38}$$

Thus, only an inversion of a matrix of size k is needed as opposed to the original problem of size m.

4 Experiments

For our experimental results, we focused on the kernel stability of kernel ridge regression, generating approximate kernel matrices using the Nyström method. We worked with the datasets listed in Table 1, and for each dataset, we randomly selected 80% of the points to generate $\mathbf{K}$ and used the remaining 20% as the test set, $\mathcal{T}$. For each test-train split, we first performed grid search to determine the optimal ridge for $\mathbf{K}$, as well as the associated optimal hypothesis, $h(\cdot)$. Next, using this optimal ridge, we generated a set of Nyström approximations, using various numbers of sampled columns, i.e., n ranging from 1% to 50% of m. For each Nyström approximation, $\widetilde{\mathbf{K}}$, we computed the associated hypothesis $h'(\cdot)$ using the same ridge and measured the distance between h and h' as follows:

$$\text{average absolute error} = \frac{\sum_{x \in \mathcal{T}} |h'(x) - h(x)|}{|\mathcal{T}|}. \tag{39}$$

We measured the distance between $\widetilde{\mathbf{K}}$ and $\mathbf{K}$ as follows:

$$\text{relative spectral distance} = \frac{\|\widetilde{\mathbf{K}} - \mathbf{K}\|_2}{\|\mathbf{K}\|_2} \times 100. \tag{40}$$

Figure 1 presents results for each dataset using all m points and a subset of 2000 points. The plots show the average absolute error of $h(\cdot)$ as a function of relative spectral distance. Proposition 1 predicts a linear relationship between kernel approximation and relative error which is corroborated by the experiments, as both datasets display this behavior for different sizes of training data.

5 Conclusion

Kernel approximation is used in a variety of contexts and its use is crucial for scaling many learning algorithms to very large tasks. We presented a stability-based analysis of the effect of kernel approximation on the hypotheses returned by several common learning algorithms. Our analysis is independent of how the approximation is obtained and simply expresses the change in hypothesis value in terms of the difference between the approximate kernel matrix and the true one measured by some norm. We also provided a specific analysis of the Nyström low-rank approximation in

this context and reported experimental results that support our theoretical analysis.

In practice, the two steps of kernel matrix approximation and training of a kernel-based algorithm are typically executed separately. Work by Bach and Jordan (2005) suggested one possible method for combining these two steps. Perhaps more accurate results could be obtained by combining these two stages using the bounds we presented or other similar ones based on our analysis.

A Lemma 1

The proof of Lemma 1 is given for completeness.

Lemma 1. *Let* $\mathbf{M}$ *and* $\mathbf{M}'$ *be two* $n \times n$ *SPSD matrices. Then, the following bound holds for the difference of the square root matrices:* $\|\mathbf{M}'^{1/2} - \mathbf{M}^{1/2}\|_2 \leq \|\mathbf{M}' - \mathbf{M}\|_2^{1/2}$.

Proof. By definition of the spectral norm, $\mathbf{M}' - \mathbf{M} \preceq \|\mathbf{M}' - \mathbf{M}\|_2 \mathbf{I}$ where $\mathbf{I}$ is the $n \times n$ identity matrix. Hence, $\mathbf{M}' \preceq \mathbf{M} + \|\mathbf{M}' - \mathbf{M}\|_2 \mathbf{I}$ and $\mathbf{M}'^{1/2} \preceq (\mathbf{M} + \lambda \mathbf{I})^{1/2}$, with $\lambda = \|\mathbf{M}' - \mathbf{M}\|_2$. Thus, $\mathbf{M}'^{1/2} \preceq (\mathbf{M} + \lambda \mathbf{I})^{1/2} \preceq \mathbf{M}^{1/2} + \lambda^{1/2} \mathbf{I}$, by sub-additivity of $\sqrt{\cdot}$. This shows that $\mathbf{M}'^{1/2} - \mathbf{M}^{1/2} \preceq \lambda^{1/2} \mathbf{I}$ and by symmetry $\mathbf{M}^{1/2} - \mathbf{M}'^{1/2} \preceq \lambda^{1/2} \mathbf{I}$, thus $\|\mathbf{M}'^{1/2} - \mathbf{M}^{1/2}\|_2 \leq \|\mathbf{M}' - \mathbf{M}\|_2^{1/2}$, which proves the statement of the lemma. □

References

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

Francis R. Bach and Michael I. Jordan. Predictive low-rank decomposition for kernel methods. In *International Conference on Machine Learning*, 2005.

M. A. Belabbas and P. J. Wolfe. Spectral methods in machine learning and new strategies for very large datasets. *Proceedings of the National Academy of Sciences of the United States of America*, 106(2):369–374, January 2009.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *Conference on Learning Theory*, 2004.

Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In *Advances in Neural Information Processing Systems*, 2001.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

Petros Drineas and Michael W. Mahoney. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, 6:2153–2175, 2005.

Shai Fine and Katya Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2002.

Zoubin Ghahramani. The kin datasets, 1996.

Ling Huang, Donghui Yan, Michael Jordan, and Nina Taft. Spectral clustering with perturbed data. In *Advances in Neural Information Processing Systems*, 2008.

Zakria Hussain and John Shawe-Taylor. Theory of matching pursuit. In *Advances in Neural Information Processing Systems*, 2008.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Ensemble Nyström method. In *Advances in Neural Information Processing Systems*, 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. On sampling-based approximate spectral decomposition. In *International Conference on Machine Learning*, 2009.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the Nyström method. In *Conference on Artificial Intelligence and Statistics*, 2009.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning*, 1998.

Bernhard Schölkopf and Alex Smola. *Learning with Kernels*. MIT Press: Cambridge, MA, 2002.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

Ameet Talwalkar, Sanjiv Kumar, and Henry Rowley. Large-scale manifold learning. In *Conference on Vision and Pattern Recognition*, 2008.

Pascal Vincent and Yoshua Bengio. Kernel Matching Pursuit. Technical Report 1179, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, 2000.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, 2000.

Kai Zhang, Ivor Tsang, and James Kwok. Improved Nyström low-rank approximation and error analysis. In *International Conference on Machine Learning*, 2008.