ARTICLE

Error Bounds on the SCISSORS Approximation Method

Imran S. Haque[†] and Vijay S. Pande[*,†,‡]

[†]Department of Computer Science and [‡]Department of Chemistry, Stanford University, Stanford, California, United States

S *Supporting Information*

ABSTRACT: The SCISSORS method for approximating chemical similarities has shown excellent empirical performance on a number of real-world chemical data sets but lacks theoretically proven bounds on its worst-case error performance. This paper first proves reductions showing SCISSORS to be equivalent to two previous kernel methods: kernel principal components analysis and the rank-k Nyström approximation of a Gram matrix. These reductions allow the use of generalization bounds on these techniques to show that the expected error in SCISSORS approximations of molecular similarity kernels is bounded in expected pairwise inner product error, in matrix 2-norm and Frobenius norm for full kernel matrix approximations and in root-mean-square deviation for approximated matrices. Finally, we show that the actual performance of SCISSORS is significantly better than these worst-case bounds, indicating that chemical space is well-structured for chemical sampling algorithms.

$$\mathrm{RMS}\{K - \tilde{K}\} \leq \mathrm{RMS}\{K - K_k\}$$

$$+ \left[\frac{64k}{\ell}\right]^{1/4} R^2 \left[1 + 2\sqrt{\frac{(n-\ell)^2}{(n-1/2)(n-\ell-1/2)} \log\frac{1}{\delta}}\right]^{1/2}$$

INTRODUCTION

The SCISSORS method is a technique for accelerating a chemical similarity search by transforming Tanimoto similarity scores to inner products, computing a metric embedding for a small "basis set" of molecules that optimally reconstructs the given inner products and then projecting remaining nonbasis "library" molecules into this vector space.[1] SCISSORS similarities are then computed as Tanimotos on these embedded vectors. Significant speedups can be achieved for certain similarity measures (those which are expensive to compute and have highly concentrated eigenvalue spectra) for repeated queries into a static database. The work done to compute vector projections for each database molecule can be amortized easily across a large number of queries. In the original SCISSORS paper, Haque and Pande report that for a database of approximately 57 000 molecules, a basis set of 1000 molecules and an embedding dimension of 100 was sufficient to accurately reproduce the shape similarity over the whole database.

The embedding used in SCISSORS is computed by first calculating the pairwise inner product matrix G between all pairs of basis molecules. G is then decomposed into eigenvectors V and eigenvalues along the diagonal of a matrix D; the vector embedding for the basis molecules lies along the rows of matrix B in the following equation:

$$G = BB^T = VDV^T = VD^{1/2}D^{1/2}V^T \quad \therefore \quad B = VD^{1/2} \quad (1)$$

The rank of the approximation can be controlled by ordering the eigenvalues in order of decreasing value, setting all eigenvalues below a certain desired count to zero, and truncating these zero dimensions in the resulting vectors.

Figure 1 shows an example of SCISSORS applied to a molecular similarity kernel. In this example, SCISSORS is used to approximate the intersection size (SMILES overlap), as used for the LINGO similarity measure,[2] between two molecules from the

Maybridge Screening Collection: (S)-mandelate (molecule 1) and (2R)-3-(4-chlorophenoxy)propane-1,2-diol (molecule 2). The true intersection sizes, as computed by the SIML implementation of LINGO,[3] are shown in the first row: between molecule 1 and itself, molecule 2 and itself, and the two molecules against each other. We then constructed a basis set of molecules from 3072 isomeric SMILES strings drawn at random from the Maybridge Screening Collection and embedded molecules 1 and 2 into SCISSORS vector spaces of varying dimensionalities: 64, 256, and 1024. The three SCISSORS data rows in the table show the approximated values of each intersection, as a function of embedding dimension. As the dimension grows, the approximation error (difference between LINGO true and SCISSORS approximated values) decreases. Our objective in this paper is to derive theoretical bounds on the magnitude of this error.

A number of methods used in chemical informatics are mathematically similar to SCISSORS. In particular, the "molecular basis set" approach taken by Raghavendra and Maggiora[4] (RM) is very similar. The RM method skips Tanimoto-to-inner product conversion (treating Tanimotos as inner products directly), does not restrict the dimensionality of the vector expansion, and is derived using a different justification but otherwise is very similar. In particular, both this method and SCISSORS are variants of kernel principal components analysis.

While the RM method and SCISSORS in particular seem to have good empirical performance, they lack theoretically rigorous guarantees on their approximations. In this paper, we derive theoretical guarantees on the SCISSORS approximation error by reducing SCISSORS to previously described kernel methods from machine learning.

M1

(S)-Mandelate
c1ccc(cc1)[C@@H](C(=O)[O-])O

M2

(2R)-3-(4-chlorophenoxy)propane-1,2-diol
c1cc(ccc1OC[C@@H](CO)O)Cl

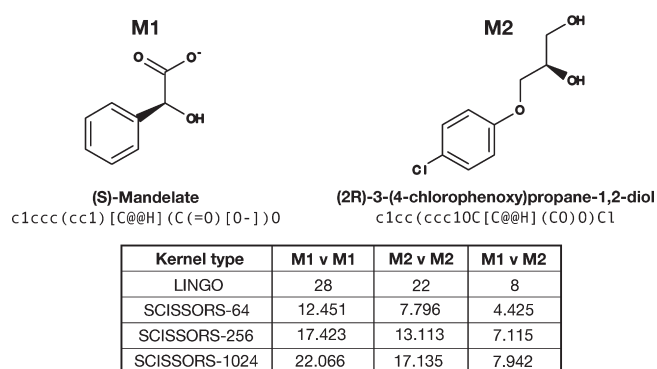| Kernel type	M1 v M1	M2 v M2	M1 v M2
LINGO	28	22	8
SCISSORS-64	12.451	7.796	4.425
SCISSORS-256	17.423	13.113	7.115
SCISSORS-1024	22.066	17.135	7.942

Figure 1. Example of SCISSORS applied to a molecular similarity kernel (LINGO intersection size). Table indicates LINGO true kernel and SCISSORS approximated kernel values for various dimensionalities.

■ PRELIMINARIES

SCISSORS as a Kernel Method. The key insight of the SCISSORS technique is that molecular similarity measures, after appropriate transformation, can be treated as "kernel functions" taking pairs of molecules to scalar values that can be interpreted as inner products. Kernels are mathematical objects widely used in machine learning which can be used to adapt linear machine learning models (e.g., support vector machines) to nonlinear spaces. Informally, a kernel function is one taking two "objects" (often vectors, but in the chemical context molecules, strings, or fingerprints) and returning a non-negative real scalar satisfying particular properties of the real dot product (including symmetry and positive semidefiniteness). While molecular similarity scores such as Tanimotos are not in themselves inner products or the result of kernel functions, they are often constructed from intermediate quantities which are. For example, the set intersection in LINGO[2] is a kernel function, and the shape overlap volume from Gaussian shape overlay[5] is approximately a kernel (non-negative and symmetric but not positive definite).

The advantage of interpreting SCISSORS as working on kernel functions or inner products is that it allows leveraging the body of machine learning literature on kernel methods. The SCISSORS pipeline can be roughly segmented into the following operations:

(1) Convert Tanimotos to inner products (basis vs basis or library vs basis).
(2) Compute a vector embedding on the inner products (by eigendecomposition or least-squares).
(3) Compute vector-space inner products (standard dot product in $\mathcal{R}^N$).
(4) Convert vector-space inner products to Tanimotos using a standard vector Tanimoto equation.

Steps 1 and 4 in this pipeline involve ratios of inner products (or kernel values) and, as such, introduce nonlinearities into the analysis. However, if one assumes that exact kernel values are given or easily obtained (as demonstrated for the shape overlap volume in ref 1) and that the goal is to directly approximate these kernel values rather than the Tanimoto, then SCISSORS directly resembles a typical kernel method. Therefore, in this paper, we will consider only the error in these inner product space stages, rather than error introduced at the Tanimoto stages. Accordingly, we replace the notion of a "molecular similarity function" with that of a "molecular similarity kernel", which can be thought of as the composition of a similarity

function with the Tanimoto-to-inner-product operation from SCISSORS.

The following lemma will be useful in demonstrating the equivalence of SCISSORS to various other kernel methods. Proof is provided in the Supporting Information.

Lemma 1 (SCISSORS library vectors are projections onto eigenvectors of the basis inner product matrix). Given an $N \times N$ SCISSORS basis inner product matrix (that is, a similarity matrix post-Tanimoto-to-inner-product conversion) K. Let the eigenvalues (respectively eigenvectors) of K be denoted λ_i and V_i, with eigenvalues sorted in descending order of value. Let the matrix of all eigenvectors be named $V = [V_1 V_2 \cdots V_N]$. The SCISSORS vector w for a new molecule with library vs basis inner product vector L, in d dimensions, is defined by the expression:

$$
w = \begin{bmatrix} \lambda_1^{-1/2} \langle V_1, L \rangle \\ \lambda_2^{-1/2} \langle V_2, L \rangle \\ \vdots \\ \lambda_d^{-1/2} \langle V_d, L \rangle \end{bmatrix}
\tag{2}
$$

Note that lemma 1 suggests a method to compute SCISSORS vectors that is distinct from, but equivalent to, the least-squares calculation specified by Haque and Pande.[1] Given a vector L of basis vs library inner products and a matrix $M = VD^{1/2}$ of basis SCISSORS vectors, the original SCISSORS calculation suggested solving the least-squares equation $Mx = L$ for the SCISSORS vector x of the new molecule. This lemma shows that the same problem is solved by the matrix multiplication $D^{-1/2}V^T L$. This provides a computational shortcut for the projection of large numbers of library molecules: the projection matrix $D^{-1/2}V^T$ can be computed once for a basis set; after library vs basis Tanimotos have been computed and converted to inner products, the SCISSORS vector can be computed by a simple matrix multiplication rather than least-squares.

Assumptions. The analysis in this paper rests on the following assumptions:

(1) SCISSORS is given molecular similarity kernel values, not Tanimotos, to analyze. While the conversion from Tanimoto to inner product will introduce distortion (particularly if different molecules x and y have very different values of $\kappa(x,x)$ and $\kappa(y,y)$ for similarity kernel κ, we will not consider this distortion here.
(2) It is assumed that the similarity kernel κ is symmetric positive semidefinite (SPSD). Similarity kernels that are not SPSD are not Mercer kernels, and some proofs will fail in the presence of negative kernel eigenvalues. However, given non-SPSD κ, the results of this paper can still be applied to a modified kernel κ', the nearest SPSD approximation to κ. If κ is symmetric but indefinite, then certain divergence terms can be easily calculated between the kernel matrices K and K' induced by κ and κ':
  - $\|K - K'\|_2$ = absolute value of the negative eigenvalue with largest magnitude.
  - $\|K - K'\|_F = \sum \lambda_{<0}^2$, where $\lambda_{<0}$ are the negative eigenvalues.
(3) It is assumed that kernel values are exactly computable. In particular, the case in which kernel values themselves are subject to noise or inexactitude is not considered here.

Under these assumptions, it is possible to bound the additional error made by SCISSORS in choosing a small random basis rather than using the eigendecomposition of the full kernel

matrix over the entire library. Two different types of bounds will be shown in this paper, arising from reductions to two different kernel methods: kernel principal components analysis (PCA) and the rank-k Nyström approximation.

■ REDUCTION OF SCISSORS TO KERNEL PCA

Overview of Kernel PCA. Kernel PCA[6,7] is a generalization of traditional principal components analysis from the data space to a feature space defined by a Mercer kernel function κ. Given a sample of N data points, kernel PCA computes up to N directions of maximum variance of the data, in the kernel's feature space. Points can then be projected into this N-dimensional subspace by a projection of their kernel values against the original (training) data points; thus, kernel PCA can be considered to perform a metric embedding of data points into a subspace of the feature space defined by a given kernel.

Similar to traditional (linear) PCA, kernel PCA can be preceded by a centering step, in which the data are centered in feature space; this ensures that the data mean is not reflected in the recovered coordinates. However, the uncentered case has relevance to SCISSORS, so we now proceed to derive the kernel PCA algorithm without data centering (following the approach of Schölkopf).[6]

Derivation of kernel PCA. Given a set of data points x_i, $i \in [1, ..., \ell]$, and a Mercer kernel $\kappa(x,y)$, defined by $\kappa(x,y) = \langle \Phi(x), \Phi(y) \rangle$ for some feature-space projection Φ. Consider the feature covariance matrix $\overline{C}$:

$$\overline{C} = \frac{1}{\ell} \sum_{j=1}^{\ell} \Phi(x_j)\Phi(x_j)^T$$

Let the eigenvalues and eigenvectors of $\overline{C}$ be named λ_k and V_k, respectively, such that $\forall k \ \lambda V = \overline{C} V$. All such V_i must lie in the span of $\Phi(x_1) \cdots \Phi(x_\ell)$. Thus the following system is equivalent:

$$\lambda(\Phi(x_k) \cdot V) = (\Phi(x_k) \cdot \overline{C} V) \ \forall \ k$$

and there exist $a_1 \cdots a_\ell$ such that

$$V = \sum_{i=1}^{\ell} a_i \Phi(x_i)$$

Defining matrix $K_{ij} = \langle \Phi(x_i), \Phi(x_j) \rangle$ and vector $\alpha = [a_1 \cdots a_n]^T$ we get $\ell \lambda K\alpha = K^2\alpha$, so we solve the eigenvalue problem $\ell\lambda = K\alpha$. Solutions λ_k and α^k correspond to eigenvalues/vectors of the kernel matrix.

We normalize the resulting solutions by requiring that the feature-space eigenvectors (V_k) be unit magnitude. This implies

$$\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} a_i^k a_j^k \langle \Phi(x_i), \Phi(x_j) \rangle = \langle \alpha^k, K\alpha^k \rangle = \lambda_k \langle \alpha^k, \alpha^k \rangle = 1$$

We can compute the projection of a new data point x onto the feature-space correlation matrix eigenvectors V_k by

$$\langle V_k, \Phi(x) \rangle = \sum_{i=1}^{\ell} a_i^k \langle \Phi(x_i), \Phi(x) \rangle$$

So for d eigenvectors, the projected KPCA coordinate vector w is

$$w = \text{KPCA}\{x\} = \begin{bmatrix} \sum_i a_i^1 \langle \Phi(x_i), \Phi(x) \rangle \\ \sum_i a_i^2 \langle \Phi(x_i), \Phi(x) \rangle \\ \vdots \\ \sum_i a_i^d \langle \Phi(x_i), \Phi(x) \rangle \end{bmatrix}$$

Equivalently

$$L = [\langle \Phi(x_1), \Phi(x) \rangle , ..., \langle \Phi(x_\ell), \Phi(x) \rangle]^T$$
$$w = KPCA\{x\} = [\langle \alpha^1, L \rangle , ..., \langle \alpha^d, L \rangle]^T$$

Reduction Proof. We now demonstrate that SCISSORS is equivalent to kernel PCA performed without data centering.

As proven in lemma 1, the SCISSORS vector w corresponding to a library molecule is defined by weighted inner products between the eigenvectors of the kernel matrix and the library vs basis inner product vector L. Define new vectors $V'_i = \lambda_i^{-1/2} V_i$. Recall that the kernel matrix and vector L are already identical between methods, and both V'_i and α_i are defined to be eigenvectors of the kernel matrix. To prove equivalence, all that is left to prove is that the SCISSORS projection vectors V'_i have the same normalization as the KPCA α^i; KPCA requires $\lambda_k \langle \alpha^k, \alpha^k \rangle = 1$.

Proof: We hypothesize that $V'_k = \alpha^k$, then

$$\lambda_k \langle V'_k, V'_k \rangle = \lambda_k \langle \lambda_k^{-1/2} V_k, \lambda_k^{-1/2} V_k \rangle = \lambda_k \lambda_k^{-1} \langle V_k, V_k \rangle = 1$$

■ REDUCTION OF SCISSORS TO THE NYSTRÖM RANK-K APPROXIMATION

Overview of the Nyström Method. In many large-scale machine learning methods, the computation and eigendecomposition of very-large scale kernel matrices are bottlenecks as the time complexity of eigendecomposition scales as $O(N^3)$. Williams and Seeger introduced a method, based on the Nyström approximation from integral equation theory, to compute a low-rank approximation to a large kernel matrix, based on computing approximate eigenvectors for the entire matrix from a random sample of a small number of points.[8] Precisely, using notation from Drineas et al.,[9] given an $n \times n$ kernel matrix K, a desired rank k, and a number of basis elements ℓ, the Nyström approximation computes $\tilde{K}_k$, a rank-k approximation to K by the following procedure:

Algorithm sketch 1 (Nyström approximation): Given a kernel matrix $K \in \mathbb{R}^{n \times n}$, choose ℓ columns (equivalently, ℓ basis/landmark input points) $[b_1, b_2, ..., b_\ell]$ to obtain matrices C and W:

$$C = \begin{bmatrix} K_{1b_1} & K_{1b_2} & \cdots & K_{1b_\ell} \\ K_{2b_1} & K_{2b_2} & \cdots & K_{2b_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ K_{nb_1} & K_{nb_2} & \cdots & K_{nb_\ell} \end{bmatrix}$$

$$W = \begin{bmatrix} K_{b_1b_1} & K_{b_1b_2} & \cdots & K_{b_1b_\ell} \\ K_{b_2b_1} & K_{b_2b_2} & \cdots & K_{b_2b_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ K_{b_\ell b_1} & K_{b_\ell b_2} & \cdots & K_{b_\ell b_\ell} \end{bmatrix}$$

Let W_k be the best rank-k approximation to matrix W and W_k^+ be the Moore–Penrose pseudoinverse of W_k. Then the rank-k Nyström approximation to matrix K is defined by $\tilde{K}_k = CW_k^+C^T$.

Preliminaries. Consider a SCISSORS computation of full pairwise similarity over some large set of molecules $\mathscr{M}$. Partition this set, by random selection without replacement, into a basis set $\mathscr{B}$ and a library set $\mathscr{L}$. Then, the matrix W in algorithm 1 corresponds to the SCISSORS basis inner-product matrix on $\mathscr{B}$; similarly, C is an aggregation of transposed library vs basis inner-product vectors. To prove the equivalence of SCISSORS and the Nyström method, we will demonstrate that the inner-product matrix computed by the SCISSORS-approximated vectors is identical to that computed by the Nyström method. It is sufficient to show (by lemma 1 that $CW_k^+C^T$, the Nyström-approximated Gram matrix, factorizes as $S_kS_k^T$ where

$$S_k^T = D_k^{1/2}\begin{bmatrix} V_1^T \\ V_2^T \\ \vdots \\ V_k^T \end{bmatrix} C^T \qquad (3)$$

S_k is the matrix with library (and basis) vectors along the rows, so $S_kS_k^T$ is the SCISSORS-approximated Gram matrix. The following lemma is helpful for the proof. Proof of the lemma is provided in the Supporting Information.

Lemma 2 (pseudoinverse of W_k): $W_k^+ = \overline{V}D_k^{-1}\overline{V}^T$, where $\overline{V} = [V_1V_2\cdots V_k]$, the matrix formed from the first k columns of the basis matrix eigenvectors, and $D_k^{-1} = \mathrm{diag}[\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_k^{-1}]$, the diagonal matrix of the reciprocals of the first k eigenvalues of the basis matrix.

Final Reduction. We must show that $CW_k^+C^T$ is equal to $S_kS_k^T$, where $S_k^T = D_k^{-1/2}\overline{V}^TC^T$.

Proof:

$$\begin{aligned} S_kS_k^T &= C\overline{V}D_k^{-1/2}D_k^{-1/2}\overline{V}^TC^T \quad \text{by definition of } S_k^T \\ &= C(\overline{V}D_k^{-1}\overline{V}^T)C^T \\ &= CW_k^+C^T \qquad\qquad \text{by lemma 2} \end{aligned}$$

EXPECTED ERROR IN INDIVIDUAL SCISSORS INNER PRODUCTS IS BOUNDED WITH HIGH PROBABILITY

Statement of the Theorem. *Theorem 1 (Bounded Expected Inner Product Error).* Given a chemical similarity kernel κ defined over pairs of molecules from some distribution D, such that $\kappa(x, x) < R^2$ for some positive real constant R for all $x \in D$. Construct a SCISSORS basis set from a random sample S of ℓ molecules drawn uniformly at random from D. Denote by κ_k^S the SCISSORS-approximated kernel of k dimensions from basis set S. Then, with probability at least $(1 - \delta)^2$, the expected error in SCISSORS approximation, over pairs of independently chosen molecules $x,y \in D$, is bounded:

$$\begin{aligned} 0 &\leq \mathbb{E}[\kappa(x, y) - \kappa_k^S(x, y)] \\ &\leq \left[\min_{1 \leq d \leq k} \left(\frac{1}{\ell}\hat{\lambda}^{>d}(S) + \frac{1 + \sqrt{d}}{\sqrt{\ell}}\sqrt{\frac{2}{\ell}\sum_{i=1}^{\ell}\kappa(s_i, s_i)2} \right) \right. \\ &\quad \left. + R^2\left(\frac{1}{4} + \sqrt{\frac{18}{\ell}\ln\left(\frac{2\ell}{\delta}\right)} \right) \right] \end{aligned} \qquad (4)$$

Where s_i are the basis molecules and $\hat{\lambda}^{>d}(S)$ is the sum of the eigenvalues of the basis matrix not used in SCISSORS:

$$\hat{\lambda}^{>d}(S) = \sum_{i=k+1}^{\ell}\lambda_i$$

Proof Overview. The proof of theorem 1 relies on a bound on the generalization error of kernel PCA projections due to Shawe-Taylor.[10] This theorem bounds the expected residual from projecting new data onto a sampled kernel PCA basis; we extend this proof to bound the expected error in inner products from projecting two points onto a kernel PCA basis. Then, the translation to SCISSORS follows trivially from the reduction of SCISSORS to kernel PCA. Because the full proof is lengthy, it has been included in the Supporting Information; this section presents a sketch.

The proof sketch relies on the following definitions from the Shawe-Taylor work:[10]

- $\hat{\mathbf{V}}_{\mathbf{k}}$ is the space spanned by the first k eigenvectors of the sample correlation matrix of a sample of vectors S; $\hat{\mathbf{V}}_{\mathbf{k}}^{\mathbf{T}}$ is the orthogonal complement to this space.
- λ_k is the kth true eigenvalue of the kernel operator κ, computed over the entire distribution generating our data.
- $\hat{\lambda}_{\mathbf{k}}$ is the kth empirical eigenvalue (i.e., the kth eigenvalue, in descending order of value, of the kernel matrix on S).
- $\lambda^{>k}$ is the sum $\Sigma_{i>k}\lambda_k$ and similarly for $\hat{\lambda}^{>k}$.
- The residual $\mathbf{P}_{\hat{V}_k}^{\mathbf{T}}(\mathbf{x})$ is the projection of x onto the space $\hat{V}_k^T$.

We make use of the following theorem:

Theorem 2 [Theorem 1 from ref 10). If we perform PCA in the feature space defined by kernel κ, then over random samples of points S s.t. $|S| = \ell$ (ℓ-samples), for all $1 \leq k \leq \ell$, if we project new data onto the space $\hat{V}_k$, the expected squared residual is bounded by the following, with probability greater than $1 - \delta$:

$$\begin{aligned} \lambda^{>k} &\leq \mathbb{E}[\|P_{\hat{V}_k}^T(\Phi(x))\|^2] \\ &\leq \min_{1 \leq d \leq k}\left[\frac{1}{\ell}\hat{\lambda}^{>d}(S) + \frac{1 + \sqrt{d}}{\sqrt{\ell}}\sqrt{\frac{2}{\ell}\sum_{i=1}^{\ell}\kappa(x_i, x_i)^2} \right] \\ &\quad + R^2\sqrt{\frac{18}{\ell}\ln\left(\frac{2\ell}{\delta}\right)} \end{aligned} \qquad (5)$$

where the support of the distribution is in a ball of radius R in feature space.

Using theorem 2, it is possible to compute a bound on the projection error for each of the two points. The proof then bounds the variance of the resulting inner product error and uses this to bound the overall error.

ERROR IN SCISSORS-APPROXIMATED GRAM MATRICES IS BOUNDED IN 2-NORM, FROBENIUS NORM, AND RMS DEVIATION

Statement of Theorems. Given a chemical similarity kernel κ and a set of n input molecules drawn from some probability distribution such that the $\kappa(x,x) < R^2$ for all molecules x. Let the true kernel matrix be denoted K, and the best possible rank-k approximation to K be denoted K_k. Compute a SCISSORS-approximated kernel matrix $\tilde{K}$ based on a size-ℓ

2251

dx.doi.org/10.1021/ci200251a |*J. Chem. Inf. Model.* 2011, 51, 2248–2253

uniform random sample of these vectors and a k-dimensional vector expansion. Then the following three theorems hold:

Theorem 3 (Bounded Error 2-Norm). With probability at least $1 - \delta$, the error in the SCISSORS kernel matrix is worse than the lowest possible error from a rank k-approximated kernel matrix by at most a bounded amount in 2-norm:

$$\left\| K - \tilde{K} \right\|_2 \le \left\| K - K_k \right\|_2$$
$$+ \frac{2n}{\sqrt{\ell}} R^2 \left[1 + 2\sqrt{\frac{(n-\ell)^2}{(n-1/2)(n-\ell-1/2)}\log\frac{1}{\delta}} \right]$$

Theorem 4 (Bounded Error Frobenius Norm). With probability at least $1 - \delta$, the error in the SCISSORS kernel matrix is worse than the lowest possible error from a rank k-approximated kernel matrix by at most a bounded amount in Frobenius norm:

$$\left\| K - \tilde{K} \right\|_F \le \left\| K - K_k \right\|_F$$
$$+ \left[\frac{64k}{\ell}\right]^{1/4} nR^2 \left[1 + 2\sqrt{\frac{(n-\ell)^2}{(n-1/2)(n-\ell-1/2)}\log\frac{1}{\delta}} \right]^{1/2}$$

Theorem 5 (Bounded RMS Error). With probability at least $1 - \delta$, the elementwise root-mean-square (RMS) error in the SCISSORS kernel matrix is worse than the lowest possible RMS error from a rank k-approximated kernel matrix by at most a bounded amount:

$$\mathrm{RMS}\{K - \tilde{K}\} \le \mathrm{RMS}\{K - K_k\}$$
$$+ \left[\frac{64k}{\ell}\right]^{1/4} R^2 \left[1 + 2\sqrt{\frac{(n-\ell)^2}{(n-1/2)(n-\ell-1/2)}\log\frac{1}{\delta}} \right]^{1/2}$$

Proof Overview. The proofs of theorems 3, 4, and 5 rely on the following theorem, due to Talwalkar[11] bounding the error of the rank-k Nyström approximation of a Gram matrix:

Theorem 6 (Theorem 5.2 from ref 11). Let $\tilde{K}$ denote the rank-k Nyström approximation of an $n \times n$ Gram matrix K based on ℓ columns sampled uniformly at random without replacement from K, and K_k the best rank-k approximation of K. Then, with probability at least $1 - \delta$, the following inequalities hold for any sample of size ℓ:

$$\left\| K - \tilde{K} \right\|_2 \le \left\| K - K_k \right\|_2$$
$$+ \frac{2n}{\sqrt{\ell}} K_{\max} \left[1 + \sqrt{\frac{n-\ell}{n-1/2}\frac{1}{\beta(\ell,n)}\log\frac{1}{\delta}\frac{d_{\max}^K}{K_{\max}^{1/2}}} \right]$$

$$\left\| K - \tilde{K} \right\|_F \le \left\| K - K_k \right\|_F$$
$$+ \left[\frac{64k}{\ell}\right]^{1/4} nK_{\max} \left[1 + \sqrt{\frac{n-\ell}{n-1/2}\frac{1}{\beta(\ell,n)}\log\frac{1}{\delta}\frac{d_{\max}^K}{K_{\max}^{1/2}}} \right]^{1/2}$$

where:

- $K_{\max} = \max_i K_{ii}$
- $d_{\max}^K$ is the maximum distance implied in $K = \max_{i,j} (K_{ii}+K_{jj}-K_{ij})^{1/2}$
- $\beta(\ell,n) = 1 - (2\max\{\ell, n - \ell\})^{-1}$.

For SCISSORS, we are particularly interested in the case in which $\ell \ll n$, so $\beta(\ell,n) = 1 - 1/(2n - 2\ell)$ and $1/\beta(\ell,n) = (n - \ell)/(n - \ell - 1/2)$.

Proof of Theorems 3, 4, and 5. Given a kernel κ and a distribution of input vectors such that their distribution in the feature space implied by κ is D and that the support of D is contained within a ball of radius R in feature space. Then, $K_{\max}$ in the above equations is bounded above by R^2 and $d_{\max}^K \le 2R$. Note that this boundedness assumption holds for any finite sample of vectors from D, as we can construct an empirical distribution of vectors from the sample, which will be guaranteed to be of bounded radius.

Theorems 3 and 4 immediately follow from theorem 6 by applying the reduction of SCISSORS to the Nyström method, the definitions of $K_{\max}$ and $d_{\max}^K$, and the assumption above that $\ell \ll n$. Theorem 5 requires one additional step:

Lemma 3: Given an $n \times n$ matrix M, the root-mean-square value of each element of M, $\mathrm{RMS}\{M\}$ is related to the Frobenius norm of M, $\|M\|_F$ by the relationship:

$$\mathrm{RMS}\{M\} = \frac{1}{n}\|M\|_F$$

Proof:

$$\|M\|_F = \sqrt{\sum_{i,j} M_{ij}}$$

$$\mathrm{RMS}\{M\} = \sqrt{\frac{1}{n^2}\sum_{i,j} M_{ij}} = \frac{1}{n}\sqrt{\sum_{i,j} M_{ij}} = \frac{1}{n}\|M\|_F$$

Then theorem 5 follows by multiplying each term of theorem 4 by $1/n$.

■ DISCUSSION

Reduction to existing kernel methods makes it possible to prove rigorous probabilistic bounds on the approximation error made by SCISSORS under fairly mild restrictions on the input molecule distribution. However, because very few assumptions are made about the input distribution, the resulting bounds end up being very loose. For example, consider the added RMS error from basis-sampling (Theorem 5 under conditions similar to those in Figure 1, if we were to approximate 50 000 molecules in Maybridge rather than just two. Specifically, consider 256 dimensions, 3000 basis molecules, and a desired confidence of $1 - e^{-3} \approx 95\%$: $n = 50\,000$, $k = 256$, $l = 3000$, $\delta = e^{-3}$:

$$\left[\frac{64\cdot256}{3000}\right]^{1/4} K_{\max} \left[1 + 2\sqrt{\frac{(50\,000-3000)^2}{(50\,000-1/2)(50\,000-3000-1/2)}\log e^{-3}} \right]^{1/2}$$
$$\approx 1.8 K_{\max}$$

So with 95% confidence, the RMS kernel error will be less than 1.8 times the maximum value of the kernel. Looking back at Figure 1 shows that this is clearly a very loose result: 1.8 times the largest kernel value in the source data is an RMS error of 50.4 (1.8×28), whereas we achieve much smaller errors on the (randomly chosen) molecules given. However, it is notable that this result holds with no assumptions about the distribution of input molecules, except boundedness in the kernel values. The performance of SCISSORS on real-world data sets is significantly better than this worst-case estimate (see for example, the statistics on the full Maybridge data set in the original SCISSORS

2252

dx.doi.org/10.1021/ci200251a |*J. Chem. Inf. Model.* 2011, 51, 2248–2253

paper),[1] indicating that the distribution of molecules in the similarity space considered is somehow friendly to sampling-based algorithms.

■ CONCLUSION

Sampling algorithms, both kernel PCA-based like SCISSORS and the Raghavendra/Maggiora method[4] and non-PCA based diversity selection and clustering methods, are widespread in chemical informatics. In this paper we have provided theoretical performance guarantees on the approximation error arising from data set sampling and rank reduction of chemical kernels. Our results relate chemical dimensionality reduction algorithms to well-known methods in machine learning. In particular, the fact that the worst-case bounds are significantly looser than the real-world performance of sampling algorithms suggests that, in practice, many chemical kernels are representable in few dimensions and that chemical space is well-structured, such that sampling is a viable strategy.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Detailed proofs for Lemmas 1 and 2 and Theorem 1. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: pande@stanford.edu.

■ ACKNOWLEDGMENT

■ REFERENCES

(1) Haque, I. S.; Pande, V. S. SCISSORS: A Linear-Algebraical Technique to Rapidly Approximate Chemical Similarities. *J. Chem. Inf. Model.* **2010**, *50*, 1075–1088.

(2) Vidal, D.; Thormann, M.; Pons, M. LINGO, an Efficient Holographic Text Based Method To Calculate Biophysical Properties and Intermolecular Similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386–393.

(3) Haque, I. S.; Pande, V. S.; Walters, W. P. SIML: A Fast SIMD Algorithm for Calculating LINGO Chemical Similarities on GPUs and CPUs. *J. Chem. Inf. Model.* **2010**, *50*, 560–564.

(4) Raghavendra, A. S.; Maggiora, G. M. Molecular Basis Sets — A General Similarity-Based Approach for Representing Chemical Spaces. *J. Chem. Inf. Model.* **2007**, *47*, 1328–1340.

(5) Grant, J. A.; Pickup, B. T. A Gaussian Description of Molecular Shape. *J. Phys. Chem.* **1995**, *99*, 3503–3510.

(6) Schölkopf, B.; Smola, A.; Müller, K.-R.Nonlinear Component Analysis as a Kernel Eigenvalue ProblemNonlinear Component Analysis as a Kernel Eigenvalue Problem. *Technical Report 44*; Max-Planck-Institut für biologische Kybernetik: Tuebingen, Germany, 1996.

(7) Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319.

(8) Williams, C.; Seeger, M. Using the Nyström Method to Speed Up Kernel Machines. *Advances in Neural Information Processing Systems 13*; MIT Press: Cambridge, MA, 2001; pp 682−688.

(9) Drineas, P.; Mahoney, M. W. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *J. Mach. Learn. Res.* **2005**, *6*, 2153–2175.

(10) Shawe-Taylor, J.; Williams, C. K. I.; Cristianini, N.; Kandola, J. On the Eigenspectrum of the Gram Matrix and the Generalization Error of Kernel-PCA. *IEEE Trans. Info. Theory* **2005**, *51*, 2510–2522.

(11) Talwalkar, A. Ph.D. Thesis, Courant Institute of Mathematical Sciences, Courant Institute of Mathematical Sciences, New York University: New York, NY, 2010.