Supplemental Information: Proofs for 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, CA

E-mail: pande@stanford.edu

Abstract

This supplementary information contains, for the lemmas and theorems in the manuscript "Error Bounds on the SCISSORS Approximation Method", lengthy proofs that were left out of the main text.

Proof of Lemma 1

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 (resp. 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*

[*] To whom correspondence should be addressed
[†] Stanford Computer Science
[‡] Stanford Chemistry

inner product vector L, in d dimensions, is defined by the expression:

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

Proof. Let the result of equation 1 be denoted B, the full-dimension basis vector matrix. Let the restriction of B to d dimensions be denoted B'; this can be defined by $B' = VD^{1/2}R$ with the restriction matrix R defined by:

$$R = \begin{bmatrix} I_{d \times d} \\ 0_{N-d \times d} \end{bmatrix}$$

Where $I_{d \times d}$ is the $d \times d$ identity matrix, and 0 is a zero matrix of appropriate dimensions. The desired library vector w is then defined by the least-squares solution to the equation $B'w = L$. This can be solved analytically:

$$\begin{aligned} w &= \left(B'^T B'\right)^{-1} B'^T L \\ &= \left(R^T D^{1/2} V^T V D^{1/2} R\right)^{-1} R^T D^{1/2} V^T L \\ &= \left(R^T D R\right)^{-1} R^T D^{1/2} V^T L \end{aligned}$$

Solve for each part of this separately (with D_d and D_{N-d} denoting corresponding blocks of matrix D:

$$\begin{aligned} R^T D R &= \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} D_d & 0 \\ 0 & D_{N-d} \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} \\ &= D_d = \operatorname{diag}\left([\lambda_1, \lambda_2, \cdots, \lambda_d]\right) \\ \therefore \left(R^T D R\right)^{-1} &= D_d^{-1} = \operatorname{diag}\left(\left[\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_d^{-1}\right]\right) \end{aligned}$$

$$
\begin{aligned}
w &= D_d^{-1} \begin{bmatrix} I_d & 0 \end{bmatrix} \begin{bmatrix} D_d^{1/2} & 0 \\ 0 & D_{N-d}^{1/2} \end{bmatrix} V^T L \\
&= D_d^{-1} \begin{bmatrix} D_d^{1/2} & 0 \end{bmatrix} V^T L \\
&= \begin{bmatrix} D_d^{-1/2} & 0 \end{bmatrix} V^T L \\
w &= \begin{bmatrix} \lambda_1^{-1/2} \langle V_1, L \rangle \\ \lambda_2^{-1/2} \langle V_2, L \rangle \\ \vdots \\ \lambda_d^{-1/2} \langle V_d, L \rangle \end{bmatrix}
\end{aligned}
$$

□

Proof of Lemma 2

Lemma 2 (The pseudoinverse of W_k)**.** *$W_k^+ = \bar{V} D_k^{-1} \bar{V}^T$, where $\bar{V} = [V_1 V_2 \cdots V_k]$, the matrix formed from the first k columns of the basis matrix eigenvectors, and $D_k^{-1} = diag[\lambda_1^{-1}, \lambda_2^{-1}, \cdots, \lambda_k^{-1}]$, the diagonal matrix of the reciprocals of the first k eigenvalues of the basis matrix.*

Proof. The Moore-Penrose pseudoinverse of matrix A is defined to be a matrix X of dimension equal to that of A^T such that the following conditions hold:

$$AXA = A$$

$$XAX = X$$

AX is Hermitian

XA is Hermitian

Given that the matrix of basis row vectors in k dimensions is defined by:

$$B = \bar{V} D_k^{1/2} = \left[V_1 \, V_2 \, \cdots \, V_k \right] \text{diag} \left[\lambda_1^{1/2}, \lambda_2^{1/2}, \cdots, \lambda_k^{1/2} \right]$$

Then $W_k = BB^T = \bar{V} D_k^{1/2} D_k^{1/2} \bar{V}^T = \bar{V} D_k \bar{V}^T$. Define $X = \bar{V} D_k^{-1} \bar{V}^T$. Let $U = \bar{V}$ and $E = D_k$; note that columns of U are orthogonal and that all matrices are real, so that Hermitian can be interpreted as symmetric. Then:

$$\begin{aligned}
XW_kX &= UE^{-1}U^TUEU^TUE^{-1}U^T \\
&= UE^{-1}EE^{-1}U^T \\
&= UE^{-1}U^T \\
&= X \\
W_kXW_k &= UEU^TUE^{-1}U^TUEU^T \\
&= UEE^{-1}EU^T \\
&= UEU^T \\
&= W_k \\
XW_k &= UE^{-1}U^TUEU^T \\
&= UU^T = \left(UU^T\right)^T = (XW_k)^T \\
W_kX &= UEU^TUE^{-1}U^T \\
&= UU^T = \left(UU^T\right)^T = (W_kX)^T
\end{aligned}$$

Thus, the matrix $X = \bar{V} D_k^{-1} \bar{V}^T$ satisfies all the Moore-Penrose properties and can be used as the value of W_k^+. □

Proof of Theorem 1

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:

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

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.[1] 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.

The proof relies on the following definitions from the Shawe-Taylor work:[1]

- For a sample of ℓ vectors $S = s_1, s_2, \cdots, s_\ell$ and a kernel function κ, the sample correlation

matrix $\mathbf{C}(\mathbf{S})$ is an $\ell \times \ell$ matrix with $C(S)_{ij} = \kappa(s_i, s_j)$.

- $\hat{\mathbf{V}}_\mathbf{k}$ is the space spanned by the first k eigenvectors of $C(S)$.
- $\hat{\mathbf{V}}_\mathbf{k}^\mathbf{T}$ is the orthogonal complement to space $\hat{V}_k$.
- λ_k is the kth process eigenvalue (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).
- $\lambda^{>\mathbf{k}}$ is the sum $\sum_{i>k} \lambda_k$, and similarly for $\hat{\lambda}^{>\mathbf{k}}$.
- The residual $\mathbf{P}^\mathbf{T}_{\hat{\mathbf{V}}_\mathbf{k}}(\mathbf{x})$ is the projection of x onto the space $\hat{V}_k^T$.

We make use of the following theorem:

Theorem 2 (Theorem 1 from [1])

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 $1-\delta$:

$$\begin{aligned}\lambda^{>k} &\leq \mathbb{E}\left[\left|\left|P^T_{\hat{V}_k}(\Phi(x))\right|\right|^2\right] \\ &\leq \min_{1\leq d\leq k}\left[\frac{1}{\ell}\hat{\lambda}^{>d}(S) + \frac{1+\sqrt{d}}{\sqrt{\ell}}\sqrt{\frac{2}{\ell}\sum_{i=1}^{\ell}\kappa(x_i,x_i)^2}\right] + R^2\sqrt{\frac{18}{\ell}\ln\left(\frac{2\ell}{\delta}\right)} \qquad (3)\end{aligned}$$

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

The Proof

Given two data vectors $\vec{x}$ and $\vec{y}$ chosen independently from a distribution D and a kernel κ. By Mercer's theorem, there exists a function Φ, the feature-space projection, such that $\langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle =$

$\kappa(\vec{x},\vec{y})$. Let $X = \Phi(\vec{x})$ and $Y = \Phi(\vec{y})$. Assume $||X||^2, ||Y||^2 \leq R^2$ for some positive real constant R (i.e., support of the feature-space distribution is bounded in a ball of radius R around the origin). Given a random ℓ-sample of vectors from D, construct the feature-space eigenvectors/eigenvalues from kernel PCA in k dimensions. Define $X_{\|}$ to be the projection of X onto $\hat{V}_k$, the eigenspace of chosen dimension from KPCA and $X_{\perp}$ to be the projection of X onto $\hat{V}_k^T$, the orthogonal eigenspace. Likewise define $Y_{\|}$ and $Y_{\perp}$.

We would like a bound on the error of inner products in the parallel eigenspace (the KPCA space) with respect to the true inner product. We will compute this in the form $L \leq \mathbb{E}\left[X \cdot Y - X_{\|} \cdot Y_{\|}\right] \leq U$.

$$\begin{aligned}
\kappa(\vec{x},\vec{y}) = X \cdot Y &= (X_{\|} + X_{\perp}) \cdot (Y_{\|} + Y_{\perp}) \\
&= X_{\|} \cdot Y_{\|} + X_{\perp} \cdot Y_{\perp} + X_{\|} \cdot Y_{\perp} + X_{\perp} \cdot Y_{\|} \\
&= X_{\|} \cdot Y_{\|} + X_{\perp} \cdot Y_{\perp} \\
&\therefore \mathbb{E}\left[X \cdot Y - X_{\|} \cdot Y_{\|}\right] = \mathbb{E}\left[X_{\perp} \cdot Y_{\perp}\right]
\end{aligned}$$

With step 3 following because dot products between orthogonal eigenspaces are zero by definition. From this, we know that $L \geq 0$, since inner products are positive. It is possible to prove a tighter lower bound, but we are here interested in the upper bound of the error only: $\mathbb{E}[X_{\perp} \cdot Y_{\perp}] \leq U$. This expression can further be bounded above:

$$\begin{aligned}
\mathbb{E}\left[X_{\perp} \cdot X_{\perp}\right] &\leq \mathbb{E}\left[||X_{\perp}||||Y_{\perp}||\right] && \text{Cauchy-Schwarz inequality} \\
&= \mathbb{E}\left[||X_{\perp}||\right]\mathbb{E}\left[||Y_{\perp}||\right] + \mathrm{Cov}\left(||X_{\perp}||, ||Y_{\perp}||\right) && (4)
\end{aligned}$$

From Theorem 2, we know that for every $\delta \in [0,1]$, sample size ℓ, and feature-space radius bound R, there exists some constant α such that $E\left[X_{\perp} \cdot X_{\perp}\right] \leq \alpha$ and $E\left[Y_{\perp} \cdot Y_{\perp}\right] \leq \alpha$ with probability greater than $(1-\delta)^2$ (since we have two independent events each of probability $\geq (1-\delta)$).

Using this fact we will now bound each term in equation 4.

Bound on $E\left[||X_\perp||\right]$

From Theorem 2, we know that $E\left[||X_\perp||^2\right] \leq \alpha$. By Jensen's inequality ($f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$ for convex f):

$$\begin{aligned} \mathbb{E}^2\left[||X_\perp||\right] &\leq \mathbb{E}\left[||X_\perp||^2\right] \leq \alpha \\ \therefore \mathbb{E}\left[||X_\perp||\right] &\leq \sqrt{\alpha} \end{aligned} \tag{5}$$

Bound on $\mathrm{Cov}\left(||X_\perp||, ||Y_\perp||\right)$

By the Cauchy-Schwarz inequality,

$$\mathrm{Cov}\left(||X_\perp||, ||Y_\perp||\right) \leq \sqrt{\mathbb{V}\left[||X_\perp|\right]\mathbb{V}\left[||Y_\perp|\right]}$$

By symmetry, $\mathbb{V}\left[||X_\perp|\right] = \mathbb{V}\left[||Y_\perp|\right]$, so:

$$\mathrm{Cov}\left(||X_\perp||, ||Y_\perp||\right) \leq \mathbb{V}\left[||X_\perp|\right] \tag{6}$$

Since $||X_\perp|| \in [0, R]$ by the assumption on support of feature distribution, $\mathbb{E}\left[||X_\perp||\right]$ must exist; let $E\left[||X_\perp||\right] = \gamma$ for some γ. Under this constraint, the variance of the distribution of $||X_\perp||$ is maximized by a scaled Bernoulli random variable ν with probability density f:

$$\begin{aligned} f(x) &= \delta(x)\left(1 - \frac{\gamma}{R}\right) + \delta(x-R)\frac{\gamma}{R} \\ \mathbb{E}[\nu] &= 0 \times \left(1 - \frac{\gamma}{R}\right) + R \times \frac{\gamma}{R} = \gamma \\ \mathbb{V}[\nu] &= \mathbb{E}\left[\nu^2\right] - \mathbb{E}^2[\nu] \\ &= \left(0 + R^2\frac{\gamma}{R}\right) - \gamma^2 \\ &= \gamma(R - \gamma) \end{aligned}$$

This quadratic expression is maximized by $\gamma = \frac{R}{2}$, so:

$$\mathbb{V}[||X_\perp||] \leq \frac{R^2}{4} \tag{7}$$

Final steps

From Theorem 2: for every $\delta \in [0,1]$, sample size ℓ, and feature-space radius bound R, there exists some constant α such that $E[X_\perp \cdot X_\perp] \leq \alpha$ and $E[Y_\perp \cdot Y_\perp] \leq \alpha$ with probability greater than $(1-\delta)^2$ for all x and y sampled independently from the distribution.

$$\begin{aligned}
\mathbb{E}[X_\perp \cdot X_\perp] &\leq \mathbb{E}[||X_\perp||]\,\mathbb{E}[||Y_\perp||] + \mathrm{Cov}(||X_\perp||, ||Y_\perp||) && \text{Equation 4} \\
&\leq \alpha + \mathrm{Cov}(||X_\perp||, ||Y_\perp||) && \text{Equation 5} \\
&\leq \alpha + \mathbb{V}[||X_\perp||] && \text{Equation 6} \\
&\leq \alpha + \frac{R^2}{4} && \text{Equation 7}
\end{aligned}$$

Therefore, by substitution from Theorem 2, with probability greater than $(1-\delta)^2$ on ℓ-samples S, for any vectors X and Y independently sampled from D, the expected kernel approximation error $\mathbb{E}\left[X \cdot Y - X_\parallel \cdot Y_\parallel\right]$ is bounded above by:

$$\begin{aligned}
&\alpha + \frac{R^2}{4} \\
&= \min_{1 \leq d \leq k} \left[\frac{1}{\ell} \hat{\lambda}^{>d}(S) + \frac{1+\sqrt{d}}{\sqrt{\ell}} \sqrt{\frac{2}{\ell} \sum_{i=1}^{\ell} \kappa(x_i, x_i)^2} \right] + R^2 \sqrt{\frac{18}{\ell} \ln\left(\frac{2\ell}{\delta}\right)} + \frac{R^2}{4} \\
&= \min_{1 \leq d \leq k} \left[\frac{1}{\ell} \hat{\lambda}^{>d}(S) + \frac{1+\sqrt{d}}{\sqrt{\ell}} \sqrt{\frac{2}{\ell} \sum_{i=1}^{\ell} \kappa(x_i, x_i)^2} \right] + R^2 \left(\frac{1}{4} + \sqrt{\frac{18}{\ell} \ln\left(\frac{2\ell}{\delta}\right)} \right)
\end{aligned}$$

Since computing approximate inner products using SCISSORS is equivalent to computing inner products using kernel PCA (section), this bound also holds for SCISSORS.

References

(1) 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.