A A Quart-sized Review of Continuous Linear Algebra

In this section we introduce continuous analogues of matrices and their factorisations. We only provide a brief quart-sized review for what is needed in this exposition. Chapters 3 and 4 of Townsend [6] contains a reservoir-sized review.

A *matrix* $F \in \mathbb{R}^{m\times n}$ is an $m \times n$ array of numbers where $F(i, j)$ denotes the entry in row i, column j. We will also look at cases where either m or n is infinite. A *column qmatrix* (quasi-matrix) $Q \in \mathbb{R}^{[a,b]\times m}$ is a collection of m functions defined on $[a, b]$ where the row index is continuous and column index is discrete. Writing $Q = [q_1, \dots, q_m]$ where $q_j : [a, b] \to \mathbb{R}$ is the j^{th} function, $Q(y, j) = q_j(y)$ denotes the value of the j^{th} function at $y \in [a, b]$. $Q^\top \in \mathbb{R}^{m\times[a,b]}$ denotes a row qmatrix with $Q^\top(j, y) = Q(y, j)$. A *cmatrix* (continous-matrix) $C \in \mathbb{R}^{[a,b]\times[c,d]}$ is a two dimensional function where both the row and column indices are continuous and $C(y, x)$ is value of the function at $(y, x) \in [a, b] \times [c, d]$. $C^\top \in \mathbb{R}^{[c,d]\times[a,b]}$ denotes its transpose with $C^\top(x, y) = C(y, x)$.

Qmatrices and cmatrices permit all matrix multiplications with suitably defined inner products. Let $F \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{[a,b]\times m}$, $P \in \mathbb{R}^{[a,b]\times n}$, $R \in \mathbb{R}^{[c,d]\times m}$ and $C \in \mathbb{R}^{[a,b]\times[c,d]}$. It follows that $F(:, j) \in \mathbb{R}^m$, $Q(y, :) \in \mathbb{R}^{1\times m}$, $Q(:, i) \in \mathbb{R}^{[a,b]}$, $C(y, :) \in \mathbb{R}^{1\times[c,d]}$ etc. Then the following hold:

- $QF = S \in \mathbb{R}^{[a,b]\times n}$ where $S(y, j) = Q(y, :)F(:, j) = \sum_{k=1}^m Q(y, k)F(i, k)$.
- $Q^\top P = H \in \mathbb{R}^{m\times n}$ where $H(i, j) = Q(:, j)^\top P(:, j) = \int_a^b Q^\top(i, s)P(s, j)\mathrm{d}s$.
- $QR^\top = D \in \mathbb{R}^{[a,b]\times[c,d]}$ where $D(y, x) = Q(y, :)R(x, :)^\top = \sum_1^m Q(y, k)R^\top(k, x)$.
- $CR = T \in \mathbb{R}^{[a,b]\times m}$ where $T(y, j) = C(y, :)R(:, j) = \int_c^d C(y, s)R(s, j)\mathrm{d}s$.

Here, the integrals are with respect to the Lebesgue measure.

A cmatrix has a singular value decomposition (SVD). If $C \in \mathbb{R}^{[a,b]\times[c,d]}$, an SVD of C is the sum $C(y, x) = \sum_{j=1}^\infty \sigma_j u_j(y)v_j(x)$, which converges in L^2. Here $\sigma_1 \geq \sigma_2 \geq \dots$. are the singular values of C. $\{u_j\}_{j\geq 1}$ and $\{v_j\}_{j\geq 1}$ are the left and right singular vectors and form orthonormal bases for $L^2([a, b])$ and $L^2([c, d])$ respectively, i.e. $\int_a^b u_j(s)u_k(s)\mathrm{d}s = \mathbb{1}(j = k)$. It is known that the SVD of a cmatrix exists uniquely with $\sigma_j \to 0$, and continuous singular vectors (Theorem 3.2, [6]). Further, if C is Lipshcitz continuous w.r.t both variables then the SVD is absolutely and uniformly convergent. Writing the singular vectors as infinite qmatrices $U = [u_1, u_2 \dots], V = [v_1, v_2 \dots]$, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2 \dots)$ we can write the SVD as,

$$C = U\Sigma V^\top = \sum_{j=1}^\infty \sigma_j U(:, j)V(:, j)^\top.$$

If only $m < \infty$ singular values are nonzero then we say that C is of rank m. The SVD of a Qmatrix $Q \in \mathbb{R}^{[a,b]\times m}$ is, $Q = U\Sigma V^\top = \sum_{j=1}^m \sigma_j U(:, j)V(:, j)^\top$, where $U \in \mathbb{R}^{[a,b]\times m}$ and $V \in \mathbb{R}^{m\times m}$ have orthonormal columns and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m)$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m \geq 0$. The SVD of a qmatrix also exists uniquely (Theorem 4.1, [6]). The rank of a column qmatrix is the number of linearly independent columns (i.e. functions) and is equal to the number of nonzero singular values.

Finally, the pseudo inverse of the cmatrix C is $C^\dagger = V\Sigma^{-1}U^\top$ with $\Sigma^{-1} = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, \dots)$. The p-operator norm of a cmatrix, for $1 \leq p \leq \infty$ is $\|C\|_p = \sup_{\|x\|_p=1} \|Cx\|_p$ where $x \in \mathbb{R}^{[c,d]}$, $Cx \in \mathbb{R}^{[a,b]}$, $\|x\|_p^p = \int_c^d (x(s))^p \mathrm{d}s$ for $p < \infty$ and $\|x\|_\infty = \sup_{s\in[c,d]} x(s)$. The Frobenius norm of a cmatrix is $\|C\|_F = \left(\int_a^b \int_c^d C(y, x)^2 \mathrm{d}x\mathrm{d}y\right)^{1/2}$. It can be shown that $\|C\|_2 = \sigma_1$ and $\|C\|_F^2 = \sum_j \sigma_j^2$ where $\sigma_1 \geq \sigma_2 \geq \dots$ are its singular values. Note that analogous relationships hold with finite matrices. The pseudo inverse and norms of a qmatrix are similarly defined and similar relationships hold with its singular values.

Notation: In what follows we will use $\mathbf{1}_{[a,b]}$ to denote the function taking value 1 everywhere in $[a, b]$ and $\mathbf{1}_m$ to denote m-vectors of 1's. When we are dealing with L^p norms of a function we will explicitly use the subscript L^p to avoid confusion with the operator/Frobenius norms of qmatrices and cmatrices. For example, for a cmatrix $\|C\|_{L^2}^2 = \int\int C(\cdot, \cdot)^2 = \|C\|_F^2$. As we have already done, throughout the paper we will overload notation for inner products, multiplications and pseudo-inverses depending on whether they hold for matrices, qmatrices or cmatrices. E.g. when

$p, q \in \mathbb{R}^m, p^\top q = \sum_1^m p_i q_i$ and when $p, q \in \mathbb{R}^{[a,b]}$, $p^\top q = \int_a^b p(s)q(s)\mathrm{d}s$. $\mathbb{P}$ will be used to denote probabilities of events while p will denote probability density functions (pdf).

B Some Perturbation Theory Results for Continuous Linear Algebra

We recommend that readers unfamiliar with continuous linear algebra first read the review in Appendix A. Throughout this section $\mathcal{L}(\cdot)$ maps a matrix (including q/cmatrices) to its eigenvalues. Similarly, $\sigma(\cdot)$ maps a matrix to its singular values. When we are dealing with infinite sequences and qmatrices "=" refers to convergence in L^2. When dealing with infinite sequences and cmatrices, "=" refers to convergence in the operator norm. For all theorems, we follow the template of Stewart and Sun [27] for the matrix case and hence try to stick with their notation.

Before we proceed, we introduce the "cmatrix" $I_{[0,1]}$ on $[0, 1]$. For any $u \in \mathbb{R}^{[0,1]}$ this is the operator which satisfies $I_{[0,1]}u = u$. That is, $(I_{[0,1]}u)(y) = \int_0^1 I_{[0,1]}(y, x)u(x)\mathrm{d}x = u(y)$. Intuitively, it can be thought of as the Dirac delta function along the diagonal, $\delta(x - y)$. Let $Q = [q_1, q_2, \dots,] \in \mathbb{R}^{[0,1]\times\infty}$ be a qmatrix containing an orthonormal basis for $[0, 1]$ and $Q_k \in \mathbb{R}^{[0,1]\times k}$ denote the first k columns of Q. We make note of the following observation.

Theorem 8. *$Q_k Q_k^\top \to I_{[0,1]}$ as $k \to \infty$. Here convergence is in the operator norm.*

Proof. We need to show that for all $x \in \mathbb{R}^{[0,1]}$, $\|Q_k Q_k^\top x - x\|_2 \to 0$. Let $x = Q\boldsymbol{\alpha} = \sum_{k=1}^\infty \alpha_k q_k$ be the representation of x in the Q-basis. Here $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots)$ satisfies $\sum_k \alpha_k^2 < \infty$. We then have $\|Q_k Q_k^\top x - x\|_2^2 = \sum_{j=k+1}^\infty \alpha_j^2 \to 0$ by the properties of sequences in ℓ^2. □

We now proceed to our main theorems. We begin with a series of intermediary results.

Theorem 9. *Let $X \in \mathbb{R}^{[0,1]\times m}$. Define the **linear** operator $\mathbf{T}(X) = AX - XB$ where $A \in \mathbb{R}^{[0,1]\times[0,1]}$ and $B \in \mathbb{R}^{m\times m}$ are a **square** cmatrix and matrix, respectively. Then, T is nonsingular if and only if $\mathcal{L}(A) \cap \mathcal{L}(B) = \varnothing$.*

Proof. Assume $\lambda \in \mathcal{L}(A) \cup \mathcal{L}(B)$. Then, let $Ap = \lambda p$, $q^\top B = \lambda q^\top$ where $p \in \mathbb{R}^{[0,1]}$ and $q \in \mathbb{R}^m$. Then $\mathbf{T}(pq^\top) = \mathbf{0}$ and $\mathbf{T}$ is singular. This proves one side of the theorem.

Now, assume that $\mathcal{L}(A) \cap \mathcal{L}(B) = \varnothing$. As the operator is linear, it is sufficient to show that $AX - XB = C$ has a unique solution for any $C \in \mathbb{R}^{[0,1]\times m}$. Let the Schur decomposition of B be $Q = V^\top BV$ where V is orthogonal and Q is upper triangular. Writing $Y = XV$ and $D = CV$ it is sufficient to show that $AY - YQ = D$ has a unique solution. We write

$$Y = (y_1, y_2, \dots y_m) \in \mathbb{R}^{[0,1]\times m} \text{ and } D = (d_1, d_2, \dots, d_m) \in \mathbb{R}^{[0,1]\times m}$$

and use an inductive argument over the columns of Y.

The first column of Y is given by $Ay_1 - Q_{11}y_1 = (A - Q_{11}I_{[0,1]})y_1 = d_1$. Since $Q_{11} \in \mathcal{L}(B)$ and $\mathcal{L}(A) \cap \mathcal{L}(B)$ is empty $(A - Q_{11}I_{[0,1]})$ is nonsingular. Therefore y_1 is uniquely determined by inverting the cmatrix (see Appendix A). Assume $y_1, y_2 \dots, y_{k-1}$ are uniquely determined. Then, the k^{th} column is given by $(A - Q_{kk}I_{[0,1]})y_k = d_k + \sum_{i=1}^{k-1} Q_{ik}y_i$. Again, $(A - Q_{kk}I_{[0,1]})$ is nonsingular by assumption, and hence this uniquely determines y_k. □

Corollary 10. *Let $\mathbf{T}$ be as defined in Theorem 9. Then*

$$\mathcal{L}(\mathbf{T}) = \mathcal{L}(A) - \mathcal{L}(B) = \{\alpha - \beta : \alpha \in \mathcal{L}(A), \beta \in \mathcal{L}(B)\}.$$

Proof. If $\lambda \in \mathcal{L}(\mathbf{T})$ there exists X such that $(A - \lambda I_{[0,1]})X - XB = \mathbf{0}$. Therefore, by Theorem 9 there exists $\alpha \in \mathcal{L}(A)$ and $\beta \in \mathcal{L}(B)$ such that $\lambda = \alpha - \beta$. Therefore, $\mathcal{L}(\mathbf{T}) \subset \mathcal{L}(A) - \mathcal{L}(B)$.

Conversely, consider any $\alpha \in \mathcal{L}(A)$ and $\beta \in \mathcal{L}(B)$. Then there exists $a \in \mathbb{R}^{[0,1]}$, $b \in \mathbb{R}^m$ such that $Aa = \alpha a$ and $b^\top B = \beta b^\top$. Writing $X = ab^\top$ we have $AX - XB = (\alpha - \beta)ab^\top$. Therefore, $\mathcal{L}(A) - \mathcal{L}(B) \subset \mathcal{L}(\mathbf{T})$. □

Theorem 11. *Let* $\mathbf{T}$ *be as defined in Theorem 9. Then*

$$\inf_{\|X\|_{\mathrm{F}}=1} \|\mathbf{T}(X)\|_{\mathrm{F}} = \min \mathcal{L}(\mathbf{T}) = \min |\mathcal{L}(A) - \mathcal{L}(B)|. \tag{7}$$

Proof. For any qmatrix $P = (p_1, p_2, \ldots, p_m) \in \mathbb{R}^{[0,1]\times m}$ let $\mathrm{vec}(P) = [p_1^\top, p_2^\top, \ldots, p_m^\top]^\top \in \mathbb{R}^{[0,m]\times 1}$ be the concatenation of all functions. Then $\mathrm{vec}(XB) = \vec{B}\mathrm{vec}(X)$ where,

$$\vec{B} = \begin{bmatrix} B_{11}I_{[0,1]} & B_{21}I_{[0,1]} & \cdots & B_{m1}I_{[0,1]} \\ B_{12}I_{[0,1]} & B_{22}I_{[0,1]} & \cdots & B_{m2}I_{[0,1]} \\ \vdots & \vdots & \ddots & \vdots \\ B_{1m}I_{[0,1]} & B_{2m}I_{[0,1]} & \cdots & B_{mm}I_{[0,1]} \end{bmatrix} \in \mathbb{R}^{[0,m]\times[0,m]}.$$

Here $I_{[0,1]}$ have been translated and should be interpreted as being a dirac delta function on that block. Similarly, $\mathrm{vec}(AX) = \vec{A}\mathrm{vec}(X)$ where $\vec{A} = \mathrm{diag}(A, A, \ldots, A) \in \mathbb{R}^{[0,m]\times[0,m]}$. Therefore $\mathrm{vec}(\mathbf{T}(X)) = (\vec{A} - \vec{B})\vec{X}$. Now noting that $\|X\|_{\mathrm{F}} = \|\mathrm{vec}(X)\|_2$ we have,

$$\inf_{\|X\|_{\mathrm{F}}=1} \|\mathbf{T}(X)\|_{\mathrm{F}} = \inf_{\|\mathrm{vec}(X)\|_2=1} \|\mathrm{vec}(\mathbf{T}(X))\|_2 = \min |\mathcal{L}(\vec{A} - \vec{B})|.$$

The theorem follows by noting that the eigenvalues of $(\vec{A} - \vec{B})$ are the same as those of $\mathcal{L}(\mathbf{T})$. □

Theorem 12. *Let* $X_1, Y_1 \in \mathbb{R}^{[0,1]\times\ell}$ *have orthonormal columns. Then, there exist* $Q \in \mathbb{R}^{\infty\times[0,1]}$ *and* $U_{11}, V_{11} \in \mathbb{R}^{\ell\times\ell}$ *such that the following holds,*

$$QX_1U_{11} = \begin{bmatrix} I_\ell \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\infty\times\ell}, \qquad QY_1V_{11} = \begin{bmatrix} \Gamma \\ \Sigma \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{\infty\times\ell}.$$

Here $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_\ell)$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_\ell)$ *and they satisfy*

$$0 \le \gamma_1 \le \cdots \le \gamma_\ell,\ \sigma_1 \ge \cdots \ge \sigma_\ell \ge 0, \text{ and } \gamma_i^2 + \sigma_i^2 = 1,\ i = 1, \ldots, \ell.$$

Proof. Let $X_2, Y_2 \in \mathbb{R}^{[0,1]\times\infty}$ be orthonormal bases for the complementary subspaces of $\mathcal{R}(X_1), \mathcal{R}(Y_1)$, respectively. Denote $X = [X_1, X_2]$, $Y = [Y_1, Y_2]$ and

$$W = X^\top Y = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \in \mathbb{R}^{\infty\times\infty},$$

where $W_{11} = X_1^\top Y_1 \in \mathbb{R}^{\ell\times\ell}$ and the rest are defined accordingly. Now, using Theorem 5.1 from [27] there exist orthogonal matrices $U = \mathrm{diag}(U_{11}, U_{22}), V = \mathrm{diag}(V_{11}, V_{22})$ where $U_{11}, V_{11} \in \mathbb{R}^{\ell\times\ell}$ and $U_{22}, V_{22} \in \mathbb{R}^{\infty\times\infty}$ such that the following holds,

$$U^\top WV = \begin{pmatrix} \Gamma & -\Sigma & \mathbf{0} \\ \Sigma & \Gamma & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_\infty \end{pmatrix} \in \mathbb{R}^{\infty\times\infty}.$$

Here Γ, Σ satisfy the conditions of the theorem. Now set $\widehat{X} = [\widehat{X}_1, \widehat{X}_2]$, $\widehat{Y} = [\widehat{Y}_1, \widehat{Y}_2]$ where $\widehat{X}_1 = X_1U_{11}$, $\widehat{X}_2 = X_2U_{11}$, $\widehat{Y}_1 = Y_1V_{11}$, $\widehat{Y}_2 = Y_2V_{11}$. Then, $\widehat{X}^\top Y = U^\top WV$. Setting $Q = \widehat{X}^\top$ and setting U_{11}, V_{11} as above yields,

$$QX_1U_{11} = \begin{pmatrix} U_{11}^\top X_1^\top \\ U_{22}^\top X_2^\top \end{pmatrix} X_1U_{11} = \begin{bmatrix} I_\ell \\ \mathbf{0} \end{bmatrix}, \quad QY_1V_{11} = \begin{pmatrix} U_{11}^\top X_1^\top \\ U_{22}^\top X_2^\top \end{pmatrix} Y_1V_{11} = \begin{bmatrix} \Gamma \\ \Sigma \\ \mathbf{0} \end{bmatrix}$$

where $U_{11}^\top X_1^\top Y_1 U_{11} = \Gamma$, $U_{22}^\top X_2^\top Y_1 U_{11} = [\Sigma^\top, \mathbf{0}^\top]^\top$ from the decomposition of $U^\top WV$. □

Remark 13. *Stewart and Sun [27] prove Theorem 5.1 for a finite unitary* W*. However, it is straightforward to verify that the same holds if* W *is a unitary operator on the* ℓ^2 *sequence space, i.e., Theorem 5.1 is valid for (countably) infinite matrices.*

Definition 14 (Canonical Angles)**.** *Let $\mathcal{X}, \mathcal{Y}$ be ℓ dimensional subspaces of the same dimension for functions on $[0,1]$ and $X_1, Y_1 \in \mathbb{R}^{[0,1]\times\ell}$ be orthonormal functions spanning these subspaces. Then the canonical angles between $\mathcal{X}$ and $\mathcal{Y}$ are the diagonals of the matrix $\Theta[\mathcal{X},\mathcal{Y}] \triangleq \sin^{-1}(\Sigma)$ where Σ is from Theorem 12. It follows that $\cos\Theta[\mathcal{X},\mathcal{Y}] = \Gamma$ where $\sin$ and $\cos$ are in the usual trigonometric sense and satisfy $\cos^2(x) + \sin^2(x) = 1$.*

Corollary 15. *Let $\mathcal{X}, \mathcal{Y}, X_1, Y_1$ be as in Definition 14 and X_2, Y_2 be orthonormal functions for their complementary spaces. Then, the nonzero singular values of $X_2^\top Y_1$ are the sines of the nonzero canonical angles between $\mathcal{X}, \mathcal{Y}$. The singular values of $X_1^\top Y_1$ are the cosines of the nonzero canonical angles.*

Proof. From the proof of Theorem 12,

$$X_2^\top Y_1 = U_{22}\begin{pmatrix}\Sigma\\ \mathbf{0}\end{pmatrix}U_{11}^\top, \qquad X_1^\top Y_1 = U_{11}\Gamma U_{11}^\top.$$

Since U_{11}, U_{22} are orthogonal, the above are the SVDs of $X_2^\top Y_1$ and $X_1^\top Y_1$. □

Theorem 16. *Let $\mathcal{X}, \mathcal{Y}$ be ℓ dimensional subspaces of functions on $[0,1]$ and $X_1, Y_1 \in \mathbb{R}^{[0,1]\times l}$ be an orthonormal bases. Let $\sin\Theta[\mathcal{X},\mathcal{Y}] = \operatorname{diag}(\sigma_1,\ldots,\sigma_\ell)$. Denote $P_\mathcal{X} = X_1X_1^\top$ and $P_\mathcal{Y} = Y_1Y_1^\top$. Then, the singular values of $P_\mathcal{X}(I_{[0,1]} - P_\mathcal{Y})$ are $\sigma_1, \sigma_2, \ldots, \sigma_\ell, 0, 0, \ldots$.*

Proof. By Theorem 12, there exists $Q \in \mathbb{R}^{\infty\times[0,1]}$, $U_{11}, V_{11} \in \mathbb{R}^{\ell\times\ell}$, such that

$$\begin{aligned} QP_\mathcal{X}(I_{[0,1]} - P_\mathcal{Y})Q^\top &= QX_1X_1^\top Q^\top Q(I_{[0,1]} - Y_1Y_1^\top)Q^\top \\ &= (QX_1U_1)(U_1^\top X_1^\top Q^\top)(I_{[0,1]} - QY_1V_{11}(V_{11}^\top Y_1^\top Q^\top)) = \begin{bmatrix}\Sigma\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}[\Sigma \quad -\Gamma \quad \mathbf{0}] \end{aligned}$$

Here we have used $I_{[0,1]} = Q^\top Q$. The proof of this uses a technical argument involving the dual space of the class of operators described by cmatrices. (In the discrete matrix case this is similar to how the outer product of a complete orthonormal basis results in the identity $UU^\top = I$.) The last step follows from Theorem 12 and some algebra. Noting that $[\Sigma \quad -\Gamma \quad \mathbf{0}]$ has orthonormal rows, it follows that the singular values of $P_\mathcal{X}(I_{[0,1]} - P_\mathcal{Y})$ are Σ. □

Theorem 17. *Let $A \in \mathbb{R}^{[0,1]\times[0,1]}$ satisfy,*

$$A = [X_1 \quad X_2]\begin{bmatrix}L_1 & \mathbf{0}\\ \mathbf{0} & L_2\end{bmatrix}\begin{bmatrix}X_1^\top\\ X_2^\top\end{bmatrix}$$

where $X_1 \in \mathbb{R}^{[0,1]\times\ell}$ and $[X_1, X_2]$ is unitary. Let $Z \in \mathbb{R}^{[0,1]\times m}$ and $T = AZ - ZB$ where $B \in \mathbb{R}^{m\times m}$. Let $\delta = \min|\mathcal{L}(L_2) - \mathcal{L}(B)| > 0$. Then,

$$\left\|\sin\Theta[\mathcal{R}(X_1),\mathcal{R}(Z)]\right\|_\mathrm{F} \;\le\; \frac{\|T\|_\mathrm{F}}{\delta}.$$

Proof. First note that $X_2^\top T = L_2X_2^\top Z - X_2^\top ZB$. The claim follows from Theorems 11 and 15.

$$\left\|\sin\Theta[\mathcal{R}(X_1),\mathcal{R}(Z)]\right\|_\mathrm{F} \;=\; \|X_2^\top Z\|_\mathrm{F} \le \frac{\|X_2^\top T\|_\mathrm{F}}{\min|\mathcal{L}(L_2) - \mathcal{L}(B)|} \le \frac{\|T\|_\mathrm{F}}{\delta}.$$

□

Theorem 18 (Wedin's Sine Theorem for cmatrices – Frobenius form). *Let $A, \tilde{A}, E \in \mathbb{R}^{[0,1]\times[0,1]}$ with $\tilde{A} = A + E$. Let $A, \tilde{A}$ have the following conformal partitions,*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix}, \qquad \tilde{A} = \begin{bmatrix} \tilde{U}_1 & \tilde{U}_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \tilde{\Sigma}_2 \end{bmatrix} \begin{bmatrix} \tilde{V}_1^\top \\ \tilde{V}_2^\top \end{bmatrix}.$$

where $U_1, \tilde{U}_1 \in \mathbb{R}^{[0,1]\times m}$, $V_1, \tilde{V}_1 \in \mathbb{R}^{[0,1]\times m}$ and $U_2, \tilde{U}_2 \in \mathbb{R}^{[0,1]\times\infty}$, $V_2, \tilde{V}_2 \in \mathbb{R}^{[0,1]\times\infty}$. Let $R = A\tilde{V}_1 - \tilde{U}_1\tilde{\Sigma}_1 \in \mathbb{R}^{[0,1]\times m}$ and $S = A^\top\tilde{U}_1 - \tilde{V}_1\tilde{\Sigma}_1 \in \mathbb{R}^{[0,1]\times m}$. Assume there exists $\delta > 0$ such that, $\min|\sigma(\tilde{\Sigma}_1) - \sigma(\Sigma_2)| \geq \delta$ and $\min|\sigma(\tilde{\Sigma}_1)| \geq \delta$. Let Φ_1, Φ_2 denote the canonical angles between $(\mathcal{R}(U_1), \mathcal{R}(\tilde{U}_1))$ and $(\mathcal{R}(V_1), \mathcal{R}(\tilde{V}_1))$ respectively. Then,

$$\sqrt{\|\sin\Phi_1\|_{\mathrm{F}}^2 + \|\sin\Phi_2\|_{\mathrm{F}}^2} \;\leq\; \frac{\sqrt{\|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2}}{\delta}.$$

Remark 19. *The two conditions on δ are needed because the theorem doesn't require $\Sigma_1, \Sigma_2, \tilde{\Sigma}_1, \tilde{\Sigma}_2$ to be ordered. If they were ordered, then it reduces to $\delta = \min|\sigma(\tilde{\Sigma}_1) - \sigma(\Sigma_2)| > 0$.*

Proof. First define $Q \in \mathbb{R}^{[0,2]\times[0,2]}$,

$$Q = \begin{bmatrix} \mathbf{0} & A \\ A^\top & \mathbf{0} \end{bmatrix}.$$

It can be verified that if $u_i \in \mathbb{R}^{[0,1]}, v_i \in \mathbb{R}^{[0,1]}$ are a left/right singular vector pair with singular value σ_i, then $(u_i, v_i) \in \mathbb{R}^{[0,2]}$ is an eigenvector with eigenvalue σ_i and $(u_i, -v_i) \in \mathbb{R}^{[0,2]}$ is an eigenvector with eigenvalue $-\sigma_i$. Writing,

$$X = \frac{1}{\sqrt{2}}\begin{pmatrix} U_1 & U_1 \\ V_1 & -V_1 \end{pmatrix}, \qquad Y = \frac{1}{\sqrt{2}}\begin{pmatrix} U_2 & U_2 \\ V_2 & -V_2 \end{pmatrix},$$

we have,

$$Q = \begin{bmatrix} X & Y \end{bmatrix} \begin{bmatrix} \Sigma_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\Sigma_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & -\Sigma_2 \end{bmatrix} \begin{bmatrix} X^\top \\ Y^\top \end{bmatrix}.$$

We similarly define $\tilde{Q}, \tilde{X}, \tilde{Y}$ for $\tilde{A}$. Now let $T = Q\tilde{X} - \tilde{X}\,\mathrm{diag}(\tilde{\Sigma}_1, -\tilde{\Sigma}_1)$. We will apply Theorem 17 with $L_1 = \mathrm{diag}(\Sigma_1, -\Sigma_1)$, $L_2 = \mathrm{diag}(\Sigma_2, -\Sigma_2)$, $Z = \tilde{X}$, $B = \mathrm{diag}(\tilde{\Sigma}_1, -\tilde{\Sigma}_1)$. Then, using the conditions on δ gives us,

$$\left\|\sin\Theta[\mathcal{R}(X), \mathcal{R}(\tilde{X})]\right\|_{\mathrm{F}} \;\leq\; \frac{\|T\|_{\mathrm{F}}}{\delta}.$$

It is straightforward to verify that $\|T\|_{\mathrm{F}}^2 = \|R\|_{\mathrm{F}}^2 + \|S\|_{\mathrm{F}}^2$. To conclude the proof, first note that

$$XX^\top(I_{[0,2]} - YY^\top) = \begin{bmatrix} (U_1U_1^\top)(I_{[0,1]} - \tilde{U}_1\tilde{U}_1^\top) & \mathbf{0} \\ \mathbf{0} & (V_1V_1^\top)(I_{[0,1]} - \tilde{V}_1\tilde{V}_1^\top) \end{bmatrix}$$

Now, using Theorem 16 we have $\|\sin\Theta[\mathcal{R}(X), \mathcal{R}(\tilde{X})]\|_{\mathrm{F}}^2 = \|\sin\Phi_1^2\|_{\mathrm{F}}^2 + \|\sin\Phi_2^2\|_{\mathrm{F}}^2$. □

We can now prove Lemma 6 which follows directly from Theorem 18.

Proof of Lemma 6. Let $\tilde{U}_\perp \in \mathbb{R}^{[0,1]\times m}$ be an orthonormal basis for the complementary subspace of $\mathcal{R}(\tilde{U})$. Then, by Corollary 15, $\|\tilde{U}_\perp^\top U\|_{\mathrm{F}}^2 = \|\sin\Theta[\mathcal{R}(\tilde{U}), \mathcal{R}(U)]\|_{\mathrm{F}}^2$, $\|\tilde{V}_\perp^\top V\|_{\mathrm{F}}^2 = \|\sin\Theta[\mathcal{R}(\tilde{V}), \mathcal{R}(V)]\|_{\mathrm{F}}^2$. For R, S as defined in Theorem 18, we have. $\|R\|_{\mathrm{F}}^2, \|S\|_{\mathrm{F}}^2 < \|E\|_{\mathrm{F}}^2$. The lemma follows via the sin–cos relationships for canonical angles,

$$\min\sigma(\tilde{U}^\top U)^2 = 1 - \max\sigma(\tilde{U}_\perp^\top U)^2 \geq 1 - \|\tilde{U}_\perp^\top U\|_{\mathrm{F}}^2 \geq 1 - \frac{2\|E\|_{\mathrm{F}}^2}{\delta^2}.$$

where $\delta = \sigma_m(A)$. □

Next we prove the pseudo-inverse theorem. Recall that for $A \in \mathbb{R}^{[0,1]\times m}$ the SVD is $A = U\Sigma V^\top$ where $U \in \mathbb{R}^{[0,1]\times m}$, $\Sigma \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{m\times m}$ where U, V have orthonormal columns. Denote its pseudo-inverse by $A^\dagger = V\Sigma^{-1}U^\top$.

Proof of Lemma 7. Let $A = U\Sigma V$ be the SVD of A and $\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}$ be the SVD of $\tilde{A}$. Let $\widetilde{P} = \tilde{U}\tilde{U}^\top$, $R = VV^\top$, $\widetilde{R} = \tilde{V}\tilde{V}^\top$, $P_\perp = I_{[0,1]} - UU^\top$, $\widetilde{R}_\perp = I_{[0,1]} - \tilde{V}\tilde{V}^\top$ and $P = UU^\top$. We then have,

$$\begin{aligned}
\tilde{A}^\dagger - A^\dagger &= -\tilde{A}^\dagger \widetilde{P} E R A^\dagger + (\tilde{A}^\top \tilde{A})^\dagger \widetilde{R} E^\top P_\perp + \widetilde{R}_\perp E P (AA^\top)^\dagger \\
\|\tilde{A}^\dagger - A^\dagger\|_2 &\leq \|\tilde{A}^\dagger\|_2 \|E\|_2 \|A^\dagger\|_2 + \|(\tilde{A}^\top \tilde{A})^\dagger\|_2 \|E\|_2 + \|E\|_2 \|(AA^\top)^\dagger\|_2 \\
&= \left(\|\tilde{A}^\dagger\|_2 \|A^\dagger\|_2 + \|\tilde{A}^\dagger\|_2^2 + \|A^\dagger\|_2^2\right) \|E\|_2 \leq 3\max\{\|\tilde{A}\|_2^2, \|A\|_2^2\} \|E\|_2
\end{aligned}$$

The first step is obtained by substitutine for $\widetilde{P}, E, R, \widetilde{R}, P_\perp, \widetilde{R}_\perp$ and P, the second step uses the triangle inequality, and the third step uses $\tilde{A}^\top \tilde{A} = U\Sigma^2 U^\top$, $AA^\top = V\Sigma^2 V^\top$. □

Remark 20. *$P, \widetilde{P}, R, \widetilde{R}$ can be shown to be the projection operators to $\mathcal{R}(A)$, $\mathcal{R}(\tilde{A})$, $\mathcal{R}(A^\top)$ and $\mathcal{R}(\tilde{A}^\top)$. Here, $\mathcal{R}(A) = \{Ax; x \in \mathbb{R}^m\} \subset \mathbb{R}^{[0,1]}$ is the range of A. $\mathcal{R}(\tilde{A}) \subset \mathbb{R}^{[0,1]}$, $\mathcal{R}(A^\top) \subset \mathbb{R}^m$ and $\mathcal{R}(\tilde{A}^\top) \subset \mathbb{R}^m$ are defined similarly. $P_\perp$, $\widetilde{R}_\perp$ are the complementary projectors of $P, \widetilde{R}$.*

Finally, we state an analogue of Weyl's theorem for cmatrices which bounds the difference in the singular values in terms of the operator norm of the perturbation. While Weyl's theorem has been studied for general operators [24], we use the form below from Townsend [6] for cmatrices.

Lemma 21 (Weyl's Theorem for Cmatrices, [6].)**.** *Let $A, E \in \mathbb{R}^{[a,b]\times[c,d]}$ and $\tilde{A} = A + E$. Let the singular values of A be $\sigma_1 \geq \sigma_2, \ldots$ and those of $\tilde{A}$ be $\tilde{\sigma}_1 \geq \tilde{\sigma}_2, \ldots$. Then,*

$$|\sigma_i - \tilde{\sigma}_i| \leq \|E\|_2 \quad \forall i \geq 1.$$

C Concentration of Kernel Density Estimation

We will first define the Hölder class in high dimensions.

Definition 22. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact space. For any $r = (r_1, \ldots, r_d)$, $r_i \in \mathbb{N}$, let $|r| = \sum_i r_i$ and $D^r = \frac{\partial^{|r|}}{\partial x_1^{r_1} \ldots x_d^{r_d}}$. The Hölder class $\mathcal{H}_d(\beta, L)$ is the set of functions of $L_2(\mathcal{X})$ satisfying*

$$|D^r f(x) - D^r f(y)| \leq L\|x - y\|^{\beta - |r|}, \tag{8}$$

for all r such that $|r| \leq \lfloor \beta \rfloor$ and for all $x, y \in \mathcal{X}$.

The following result establishes concentration of kernel density estimators. At a high level, we follow the standard KDE analysis techniques to decompose the L^2 error into bias and variance terms and bound them separately. A similar result for 2-dimensional densities was given by Liu et al. [25]. Unlike the previous work, here we deal with the general d-dimensional case as well as explicitly delineate the dependencies of the concentration bounds on the deviation, ε.

Lemma 23. *Let $f \in \mathcal{H}_d(\beta, L)$ be a density on $[0,1]^d$ and assume we have N i.i.d samples $\{X_i\}_{i=1}^N \sim f$. Let $\hat{f}$ be the kernel density estimate obtained using a kernel with order at least β and bandwidth $h = \left(\frac{\log N}{N}\right)^{\frac{1}{2\beta+d}}$. Then there exist constants $\kappa_1, \kappa_2, \kappa_3, \kappa_4 > 0$ such that for all $\varepsilon < \kappa_4$ and number of samples satisfying $\frac{N}{\log N} > \frac{\kappa_1}{\varepsilon^{2+\frac{d}{\beta}}}$ we have,*

$$\mathbb{P}\left(\|\hat{f} - f\|_{L^2} > \varepsilon\right) \leq \kappa_2 \exp\left(-\kappa_3 N^{\frac{2\beta}{2\beta+d}} (\log N)^{\frac{d}{2\beta+d}} \varepsilon^2\right) \tag{9}$$

Proof. First note that

$$\mathbb{P}\big(\|\hat{f} - f\|_{L^2} > \varepsilon\big) \leq \mathbb{P}\big(\|\hat{f} - \mathbb{E}\hat{f}\|_{L^2} + \|\mathbb{E}\hat{f} - f\|_{L^2} > \varepsilon\big). \tag{10}$$

Using the Hölderian conditions and assumptions on the kernel, standard techniques for analyzing the KDE [13, 18], give us a bound on the bias, $\|\mathbb{E}\hat{f} - f\|_{L^2} \leq \kappa_5 h^\beta$, where $\kappa_5 = L \int K(u) u^\beta \mathrm{d}u$. When the number of samples, N, satisfies

$$\frac{N}{\log N} > \left(\frac{2\kappa_5'}{\varepsilon}\right)^{2+\frac{d}{\beta}} = \frac{\kappa_5}{\varepsilon^{2+\frac{d}{\beta}}}, \text{ where } \kappa_5 \triangleq (2\kappa_5')^{2+\frac{d}{\beta}} \tag{11}$$

we have $\|\mathbb{E}\hat{f} - f\|_{L^2} \leq \varepsilon/2$, and hence (10) turns into $\mathbb{P}\big(\|\hat{f} - f\|_{L^2} > \varepsilon\big) \leq \mathbb{P}\big(\|\hat{f} - \mathbb{E}\hat{f}\|_{L^2} > \varepsilon/2\big)$.

The main challenge in bounding the first term is that we want the difference to hold in L^2. The standard techniques that bound the pointwise variance would not be sufficient here. To overcome the limitations, we use Corollary 2.2 from Giné and Guillou [26]. Using their notation we have,

$$\begin{aligned}
\sigma^2 &= \sup_{t\in[0,1]^d} \mathbb{V}_{X\sim f}\left[\frac{1}{h^d}K\left(\frac{X-t}{h}\right)\right] \\
&\leq \sup_{t\in[0,1]^d} \frac{1}{h^{2d}}\int K^2\left(\frac{x-t}{h}\right) f(x)\mathrm{d}x \\
&= \sup_{t\in[0,1]^d} \frac{1}{h^d}\int K^2(u) f(t+uh)\mathrm{d}u \;\leq\; \frac{\|f\|_\infty \|K\|_{L^2}}{h^d} \\
U &= \sup_{t\in[0,1]^d} \left\|\frac{1}{h^d}K\left(\frac{X-t}{h}\right)\right\|_\infty \;=\; \frac{\|K\|_{L^\infty}}{h^d}.
\end{aligned}$$

Then, there exist constants $\kappa_2, \kappa_3, \kappa_4'$ such that for all $\varepsilon \in \left(\kappa_4' \frac{\sigma}{\sqrt{n}}\sqrt{\log\frac{U}{\sigma}},\ \frac{\sigma^2}{U}\kappa_4'\right)$ we have,

$$\mathbb{P}\left(\|\hat{f} - \mathbb{E}\hat{f}\|_{L^2} > \frac{\varepsilon}{2}\right) \leq \kappa_2 \exp\left(-\kappa_3 N h^d \varepsilon^2\right).$$

Substituting for h and then combining this with (10) gives us the probability inequality of the theorem. All that is left to do is to verify the that the conditions on ε hold. The upper bound condition requires $\varepsilon \leq \frac{\kappa_4'\|f\|_\infty\|K\|_{L^2}}{\|K\|_{L^\infty}} \triangleq \kappa_4$. After some algebra, the lower bound on ε reduces to $\frac{N}{\log N} > \frac{\kappa_6}{\varepsilon^{2+\frac{d}{\beta}}}$. Combining this with the condtion (11) and taking $\kappa_1 = \max(\kappa_6, \kappa_5)$ gives the theorem. □

In order to apply the above lemma, we need P_1, P_{21}, P_{321} to satisfy the Hölder condition. The following lemma shows that if all O_k's are Hölderian, so are P_1, P_{21}, P_{321}.

Lemma 24. *Assume that the observation probabilities belong to the one dimensional Hölder class; $\forall \ell \in [m], O_\ell \in \mathcal{H}_1(\beta, L)$. Then for some constants L_1, L_2, L_3, $P_1 \in \mathcal{H}_1(\beta, L_1)$, $P_{21} \in \mathcal{H}_2(\beta, L_2)$, $P_{321} \in \mathcal{H}_3(\beta, L_3)$.*

Proof. We prove the statement for P_{21}. The other two follow via a similar argument. Let $r = (r_1, r_2)$, $r_i \in \mathbb{N}$, $|r| = r_1 + r_2 \leq \beta$, and let $(s,t), (s',t') \in [0,1]^d$. Note that we can write,

$$P_{21}(s,t) = \sum_{k\in[m]}\sum_{\ell\in[m]} p(x_2 = s, x_1 = t, h_2 = k, h_1 = \ell) = \sum_{k\in[m]}\sum_{\ell\in[m]} \alpha_{kl} O_k(s) O_\ell(t),$$

where $\sum_{k,\ell}\alpha_{k\ell} = 1$. Then,

$$\begin{aligned}
&\frac{\partial^{|r|}P_{21}(s,t)}{\partial s^{r_1}\partial t^{r_2}} - \frac{\partial^{|r|}P_{21}(s',t')}{\partial s^{r_1}\partial t^{r_2}} \\
&\quad= \sum_{k,\ell}\alpha_{k\ell}\left(\frac{\partial O_k(s)}{\partial s^{r_1}}\frac{\partial O_\ell(t)}{\partial t^{r_2}} - \frac{\partial O_k(s')}{\partial s^{r_1}}\frac{\partial O_\ell(t')}{\partial t^{r_2}}\right) \\
&\quad\leq \sum_{k,\ell}\alpha_{k\ell}\left(\left|\frac{\partial O_k(s)}{\partial s^{r_1}}\right|\left|\frac{\partial O_\ell(t)}{\partial t^{r_2}} - \frac{\partial O_\ell(t')}{\partial t^{r_2}}\right| + \right. \\
&\qquad\qquad \left.\left|\frac{\partial O_\ell(t')}{\partial t^{r_2}}\right|\left|\frac{\partial O_k(s)}{\partial s^{r_1}} - \frac{\partial O_\ell(s')}{\partial s^{r_1}}\right|\right) \\
&\quad\leq \sum_{k,\ell}\alpha_{kl}\left(L'L|t-t'|^{\beta-r_2} + L'L|s-s'|^{\beta-r_1}\right) && \text{(Hölder condition)} \\
&\quad\leq L'L\left(|t-t'|^{\beta-|r|} + |s-s'|^{\beta-|r|}\right) && \text{(domain of } s,s' \text{ and } t,t') \\
&\quad\leq L_2\sqrt{(t-t')^2 + (s-s')^2}^{\beta-|r|}
\end{aligned}$$

Here, the third step uses the Hölder conditions on O_k and O_ℓ and the fact that the partial fractions are bounded in a bounded domain by a constant, which we denoted L', due to the Hölder condition. Since $r_1 + r_2 = |r| \leq \beta$ and r_1, r_2 are positive integers, we have $x^{\beta - r_i} \leq x^{\beta - r}, i = 1, 2$ for any $x \in [0, 1]$, which implies the fourth step. The last step uses Jensen's inequality and sets $L_2 \equiv L'L$. □

The corollary belows follws as a direct consequence of Lemmas 23 and 24. We have absorbed the constants L_1, L_2, L_3 into $\kappa_1, \kappa_2, \kappa_3, \kappa_4$.

Corollary 25. *Assume the HMM satisfies the conditions given in Section 3. Let $\epsilon_1, \epsilon_{21}, \epsilon_{321} \in (0, \kappa_4)$ and $\eta \in (0, 1)$. If the number of samples N is large enough such that the following are true,*

$$\frac{N}{\log N} > \frac{\kappa_1}{\epsilon_1^{2+\frac{1}{\beta}}}, \qquad \frac{N}{\log N} > \frac{\kappa_1}{\epsilon_{21}^{2+\frac{2}{\beta}}}, \qquad \frac{N}{\log N} > \frac{\kappa_1}{\epsilon_{321}^{2+\frac{3}{\beta}}},$$

$$N(\log N)^{\frac{1}{2\beta}} > \frac{1}{\epsilon_1^{2+\frac{1}{\beta}}} \left(\frac{1}{\kappa_3} \log \left(\frac{3\kappa_2}{\eta} \right) \right)^{1+\frac{1}{2\beta}}$$

$$N(\log N)^{\frac{2}{2\beta}} > \frac{1}{\epsilon_{21}^{2+\frac{2}{\beta}}} \left(\frac{1}{\kappa_3} \log \left(\frac{3\kappa_2}{\eta} \right) \right)^{1+\frac{2}{2\beta}}$$

$$N(\log N)^{\frac{3}{2\beta}} > \frac{1}{\epsilon_{321}^{2+\frac{3}{\beta}}} \left(\frac{1}{\kappa_3} \log \left(\frac{3\kappa_2}{\eta} \right) \right)^{1+\frac{3}{2\beta}}$$

then with at least $1 - \eta$ probability the L^2 errors between P_1, P_{21}, P_{321} and the KDE estimates $\widehat{P}_1, \widehat{P}_{21}, \widehat{P}_{321}$ satisfy,

$$\|P_1 - \widehat{P}_1\|_{L^2} \leq \epsilon_1, \qquad \|P_{21} - \widehat{P}_{21}\|_{L^2} \leq \epsilon_{21}, \qquad \|P_{321} - \widehat{P}_{321}\|_{L^2} \leq \epsilon_{321}.$$

D Analysis of the Spectral Algorithm

Our proof is a brute force generalization of the analysis in Hsu et al. [2]. Following their template, we use establish a few technical lemmas. We mainly focus on the cases where our analysis is different.

Throughout this section $\epsilon_1, \epsilon_{21}, \epsilon_{321}$ will refer to L^2 errors. Using our notation for c/q-matrices the errors can be written as,

$$\begin{aligned} \epsilon_1 &= \|P_1 - \widehat{P}_1\|_{L^2} = \|P_1 - \widehat{P}_1\|_F, \\ \epsilon_{21} &= \|P_{21} - \widehat{P}_{21}\|_{L^2} = \|P_{21} - \widehat{P}_{21}\|_F, \\ \epsilon_{321} &= \|P_{321} - \widehat{P}_{321}\|_{L^2}. \end{aligned}$$

We begin with a series of Lemmas.

Lemma 26. *Let $\epsilon_{21} \leq \varepsilon \sigma_m(P_{21})$ where $\varepsilon < \frac{1}{1+\sqrt{2}}$. Denote $\varepsilon_0 = \frac{\epsilon_{21}^2}{((1-\varepsilon)\sigma_m(P_{21}))^2} < 1$. Then the following hold,*

1. $\sigma_m(\widehat{U}^\top \widehat{P}_{21}) \geq (1 - \varepsilon)\sigma_m(P_{21})$.
2. $\sigma_m(\widehat{U}^\top P_{21}) \geq \sqrt{1 - \varepsilon_0}\sigma_m(P_{21})$.
3. $\sigma_m(\widehat{U}^\top P_{21}) \geq \sqrt{1 - \varepsilon_0}\sigma_m(P_{21})$.

Proof. The proof follows Hsu et al. [2] after an application of Weyl's theorem (Lemma 21) and Wedin's sine theorem (Lemma 6) for cmatrices. □

We define an alternative observable representation for the true HMM given by, $\widetilde{b}_\infty, \widetilde{b}_1 \in \mathbb{R}^m$ and $\widetilde{B} : [0,1] \to \mathbb{R}^{m \times m}$.

$$\begin{aligned}
\widetilde{b}_1 &= \widehat{U}^\top P_1 = (\widehat{U}^\top O)\pi \\
\widetilde{b}_\infty &= (P_{21}^\top \widehat{U}) P_1 = (\widehat{U}^\top O)^{-1} \mathbf{1}_m \\
\widetilde{B}(x) &= (\widehat{U}^\top P_{3x1})(\widehat{U}^\top P_{21})^\dagger = (\widehat{U}^\top O) A(x) (\widehat{U}^\top O)^{-1}.
\end{aligned}$$

As long as $\widehat{U}^\top O$ is invertible, the above parameters constitute a valid observable representation. This is guaranteed if $\widehat{U}$ is sufficiently close to U. We now define the following error terms,

$$\begin{aligned}
\delta_\infty &= \|(\widehat{U}^\top O)^\top (\widehat{b}_\infty - \widetilde{b}_\infty)\|_\infty = \|(\widehat{U}^\top O)^\top \widehat{b}_\infty - \mathbf{1}_m\|_\infty \\
\delta_1 &= \|(\widehat{U}^\top O)^{-1} (\widehat{B}(x) - \widetilde{B}(x))(\widehat{U}^\top O)\|_1 = \|(\widehat{U}^\top O)^{-1} \widehat{B}(x)(\widehat{U}^\top O) - A(x)\|_1 \\
\Delta(x) &= \|(\widehat{U}^\top O)^{-1} (\widehat{B}(x) - \widetilde{B}(x)) \widehat{U}^\top O\|_1 = \|(\widehat{U}^\top O)^{-1} \widehat{B}(x) - A(x)\|_1 \\
\Delta &= \int_{x \in [0,1]} \Delta(x) \mathrm{d}x
\end{aligned}$$

The next lemma bounds the above quantities in terms of $\epsilon_1, \epsilon_{21}, \epsilon_{321}$.

Lemma 27. *Assume $\epsilon_{21} < \sigma_m(P_{21})/3$. Then, there exists constants c_1, c_2, c_3, c_4 such that,*

$$\begin{aligned}
\delta_\infty &\leq c_1\, \sigma_1(O) \left(\frac{\epsilon_{21}}{\sigma_m(P_{21})^2} + \frac{\epsilon_1}{\sigma_m(P_{21})} \right) \\
\delta_1 &\leq c_2\, \frac{\epsilon_1}{\sigma_m(O)} \\
\Delta(x) &\leq c_3\, \sqrt{m}\, \kappa(O) \left(\frac{\epsilon_{21}}{\sigma_m(P_{21})^2} \|P_{3x1}\|_2 + \frac{\|P_{3x1} - \widehat{P}_{3x1}\|_2}{\sigma_m(P_{21})^2} \right) \\
\Delta &\leq c_4\, \sqrt{m}\, \kappa(O) \left(\frac{\epsilon_{21}}{\sigma_m(P_{21})^2} + \frac{\epsilon_{321}}{\sigma_m(P_{21})^2} \right)
\end{aligned}$$

Proof. We will use $\lesssim, \gtrsim$ to denote inequalities ignoring constants. First we bound $\delta_\infty \leq \|(\widehat{U}^\top O)^\top (\widehat{b}_\infty - \widetilde{b}_\infty)\|_2 \leq \sigma_1(O) \|\widehat{b}_\infty - \widetilde{b}_\infty\|_2$. Then we note,

$$\begin{aligned}
\|\widehat{b}_\infty - \widetilde{b}_\infty\|_2 &\leq \|(\widehat{P}_{21}^\top \widehat{U})^\dagger \widehat{P}_1 - (P_{21}\widehat{U})^\dagger P_1\|_2 \\
&\leq \|(\widehat{P}_{21}^\top \widehat{U})^\dagger - (P_{21}^\top \widehat{U})^\dagger\|_2 \|\widehat{P}_1\|_2 + \|(P_{21}^\top \widehat{U})^\dagger\|_2 \|\widehat{P}_{21} - P_1\|_2 \\
&\lesssim \frac{\epsilon_{21}}{\min\{\sigma_m(\widehat{P}_{21}^\top), \sigma_m(P_{21}^\top \widehat{U})\}^2} + \frac{\epsilon_1}{\sigma_m(P_{21}^\top \widehat{U})} \\
&\lesssim \frac{\epsilon_{21}}{\sigma_m(P_{21})^2} + \frac{\epsilon_1}{\sigma_m(P_{21})},
\end{aligned}$$

where the third and fourth steps use Lemma 26 and Lemma 7 (the pseudoinverse theorem for qmatrices). This establishes the first result. The second result is straightforward from Lemma 26.

$$\delta_1 \leq \sqrt{m} \|(\widehat{U}^\top O)^{-1}\|_2 \|\widehat{b}_1 - \widetilde{b}_1\|_2 \leq \sqrt{m} \frac{\|\widehat{b}_1 - \widetilde{b}_1\|_2}{\sigma_m(\widehat{U}^\top O)} \lesssim \sqrt{m} \frac{\|\widehat{U}^\top (\widehat{P}_1 - P_1)\|_2}{\sigma_m(O)} \lesssim \frac{\sqrt{m}\epsilon_1}{\sigma_m(O)}.$$

For the third result, we first note

$$\begin{aligned}
\Delta(x) &\leq \sqrt{m} \|(\widehat{U}^\top O)^{-1}\|_2 \|\widehat{B}(x) - \widetilde{B}(x)\|_2 \|\widehat{U}^\top O\|_2 \leq \sqrt{m} \frac{\sigma_1(O)}{\sigma_m(\widehat{U}^\top O)} \|\widehat{B}(x) - \widetilde{B}(x)\|_2 \\
&\lesssim \sqrt{m}\, \kappa(O) \|\widehat{B}(x) - \widetilde{B}(x)\|_2
\end{aligned}$$

To bound the last term we decompose it as follows.

$$\begin{aligned}\|\widehat{B}(x) - \widetilde{B}(x)\|_2 &= \|(\widehat{U}^\top P_{3x1})(\widehat{U}^\top P_{21})^\dagger - (\widehat{U}^\top \widehat{P}_{3x1})(\widehat{U}^\top \widehat{P}_{21})^\dagger\|_2 \\ &\leq \|(\widehat{U}^\top P_{3x1})((\widehat{U}^\top P_{21})^\dagger - (\widehat{U}^\top \widehat{P}_{21})^\dagger)\|_2 \;+\; \|\widehat{U}^\top (P_{3x1} - \widehat{P}_{3x1})(\widehat{U}^\top \widehat{P}_{21})^\dagger\|_2 \\ &\leq \|P_{3x1}\|_2 \|(\widehat{U}^\top P_{21})^\dagger - (\widehat{U}^\top \widehat{P}_{21})^\dagger\|_2 \;+\; \|P_{3x1} - \widehat{P}_{3x1}\|_2 \|(\widehat{U}^\top \widehat{P}_{21})^\dagger\|_2 \\ &\lesssim \|P_{3x1}\|_2 \frac{\epsilon_{21}}{\sigma_m(P_{21})^2} + \frac{\|P_{3x1} - \widehat{P}_{3x1}\|_2}{\sigma_m(P_{21})}.\end{aligned}$$

This proves the third claim. For the last claim, we make use of the proven statements. Observe,

$$\int \|P_{3x1}\|_2 \mathrm{d}x \;\leq\; \left(\int \|P_{3x1}\|_2^2 \mathrm{d}x\right)^{1/2} \;\leq\; \left(\int\int\int P_{321}(s,x,t)^2 \mathrm{d}s\mathrm{d}t\mathrm{d}x\right)^{1/2} = \|P_{321}\|_{L^2},$$

where the first step uses inclusion of the L^p norms in $[0,1]$. The second step uses $\|\cdot\|_2 \leq \|\cdot\|_{\mathrm{F}}$ for cmatrices. A similar argument shows $\int_x \|P_{3x1} - \widehat{P}_{3x1}\|_2 \leq \epsilon_{321}$. Combining these results gives the fourth claim. □

Finally, we need the following Lemma. The proof almost exactly replicates the proof of Lemma 12 in Hsu et al. [2], as all operations can be done with just matrices.

Lemma 28. *Assume* $\epsilon_{321} \leq \sigma_m(P_{21})/3$. *Then* $\forall t \geq 0$,

$$\int |p(x_{1:t}) - \widehat{p}(x_{1:t})| \mathrm{d}x_{1:t} \;\leq\; \delta_\infty + (1+\delta_\infty)\left((1+\Delta)^t \delta_1 + (1+\Delta)^t - 1\right), \tag{12}$$

where the integral is over $[0,1]^t$.

We are now ready to prove Theorem 5.

Proof of Theorem 5. If $\epsilon_1, \epsilon_{21}, \epsilon_{321}$ satisfy the following for appropriate choices of c_5, c_6, c_7,

$$\epsilon_1 \;\leq\; c_5 \min(\sigma_m(P_{21}), \frac{\kappa(O)}{\sqrt{m}})\epsilon, \quad \epsilon_{21} \;\leq\; c_6 \frac{\sigma_m(P_{21})^2}{\kappa(O)}\epsilon, \quad \epsilon_{321} \;\leq\; c_7 \frac{\sigma_m(P_{21})}{\sigma_1(O)} \frac{1}{t\sqrt{m}}\epsilon, \tag{13}$$

we then have $\delta_1 \leq \epsilon/20$, $\delta_\infty \leq \epsilon/20$ and $\Delta \leq 0.4\epsilon/t$. Plugging these expressions into Lemma 28 gives $\int |p(x_{1:t}) - \widehat{p}(x_{1:t})| \mathrm{d}x_{1:t} \leq \epsilon$. When we plug the expresssions for $\epsilon_1, \epsilon_{21}, \epsilon_{321}$ in (13) into Corollary 25 we get the required sample complexity. □

E Addendum to Experiments

Details on Synthetic Experiments: Figure 3 shows the emission probabilities used in our synthetic experiments. For the transition matrices, we sampled the entries of the matrix from a $U(0,1)$ distribution and then renormalised the columns to sum to 1.

In our implementation, we use a Gaussian kernel for the KDE which is of order $\beta = 2$. While higher order kernels can be constructed using Legendre polynomials [18], the Gaussian kernel was more robust in practice. The bandwidth for the kernel was chosen via cross validation on density estimation.

Details on Real Datasets: Here, we first estimate the model parameters using the training sequence. Given a test sequence $x_{1:n}$, we predict x_{t+1} conditioned on the previous $x_{1:t}$ for $t = 1:n$.

1. Internet Traffic. Training sequence length: $10,000$. Test sequence length: 10.
2. Laser Generation. Training sequence length: $10,000$. Test sequence length: 100.
3. Physiological data. Training sequence length: $15,000$. Test sequence length: 100.

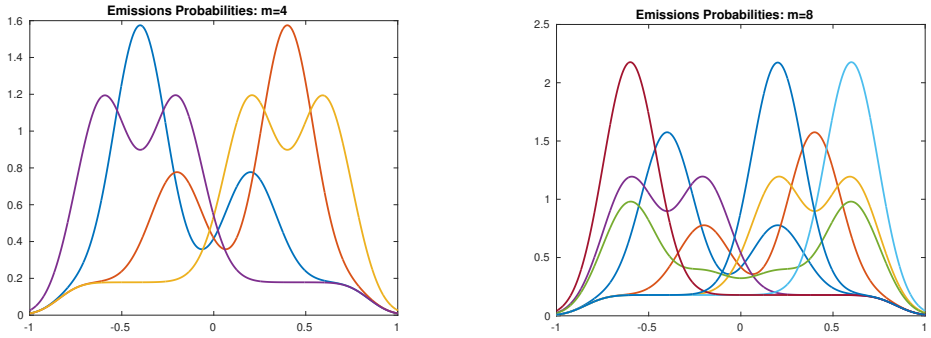

Figure 3: An illustration of the nonparametric emission probabilities used in our experiments.