Appendix: Kernel Spectral Algorithm for HMM

Le Song LESONG@CS.CMU.EDU
Byron Boots BEB@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Sajid M. Siddiqi SIDDIQI@GOOGLE.COM
Google, Pittsburgh, PA 15213, USA

Geoffrey Gordon GGORDON@CS.CMU.EDU
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Alex Smola ALEX@SMOLA.ORG
Yahoo! Research, Santa Clara, CA 95051, USA

1. Sample Complexity

Let $k(x,x') = \langle\varphi(x),\varphi(x')\rangle_{\mathcal{F}}$ and $l(h,h') = \langle\phi(h),\phi(h')\rangle_{\mathcal{G}}$ be the kernels for the observation variables X_t and hidden states H_t respectively. We assume that the feature map is bounded, *i.e.* $\|\varphi(x)\|_{\mathcal{F}} \le 1$ and $\|\phi(h)\|_{\mathcal{G}} \le 1$. In this case, the norms of the conditional embeddings of H_t are bounded by 1, *i.e.* by convexity

$$\begin{aligned}&\left\|\mu_{H_t|x_{t-1:1}}\right\|_{\mathcal{G}} = \left\|\mathbb{E}_{H_t|x_{t-1:1}}[\phi(H_t)]\right\|_{\mathcal{G}}\\ &\le \mathbb{E}_{H_t|x_{t-1:1}}[\|\phi(H_t)\|_{\mathcal{G}}] \le 1 \end{aligned} \tag{1}$$

We will use $\|\cdot\|_2$ to denote the spectral norm of an operator. We assume that the spectral norms of $\mathcal{O}$ and $\mathcal{T}$ are finite, *i.e.* $\|\mathcal{O}\|_2 \le C_{\mathcal{O}}$ and $\|\mathcal{T}\|_2 \le C_{\mathcal{T}}$. Let $\hat{\mathcal{U}}$ be the column concatenation of the top N left singular vectors of $\hat{\mathcal{C}}_{2,1}$. We define

$$\tilde{\beta}_1 := \hat{\mathcal{U}}^\top \mu_1 \tag{2}$$
$$\tilde{\beta}_\infty := \mathcal{C}_{2,1}(\hat{\mathcal{U}}^\top \mathcal{C}_{2,1})^\dagger \tag{3}$$
$$\tilde{\mathcal{B}}_x := \mathbb{P}(x)\left(\hat{\mathcal{U}}^\top(\mathcal{C}_{3,1|2}\varphi(x))\right)\left(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1}\right)^\dagger \tag{4}$$

We further define

$$\delta_1 := \left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\tilde{\beta}_1 - \hat{\beta}_1)\right\|_{\mathcal{G}} \tag{5}$$
$$\delta_\infty := \left\|\tilde{\beta}_\infty - \hat{\beta}_\infty\right\|_2 \tag{6}$$
$$\Delta := \max_{x_t}\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2 \tag{7}$$
$$\gamma := \max_{x_t}\|\mathcal{A}_{x_t}\|_2 \quad \text{(we assume } \gamma \le 1\text{)}. \tag{8}$$

Lemma 1 $\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1 - \tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{G}} \le (1+\Delta)^t\delta_1 + (1+\Delta)^t - 1$

Proof By induction on t. $\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_1 - \tilde{\beta}_1)\right\|_{\mathcal{G}} = \delta_1 = (1+\Delta)^0\delta_1 + (1+\Delta)^0 - 1$ when $t = 0$. For the induction step define $\hat{\beta}_t := \hat{\mathcal{B}}_{x_{t-1:1}}\hat{\beta}_1$ and $\tilde{\beta}_t := \tilde{\mathcal{B}}_{x_{t-1:1}}\tilde{\beta}_1$. Assume for $t > 1$:

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t - \tilde{\beta}_t)\right\|_{\mathcal{G}}\\ &\le (1+\Delta)^{t-1}\delta_1 + (1+\Delta)^{t-1} - 1\end{aligned} \tag{9}$$

Then, we have

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1 - \tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{G}}\\ &\le \left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\tilde{\beta}_t\right\|_{\mathcal{G}} &(10)\\ &+ \left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t - \tilde{\beta}_t)\right\|_{\mathcal{G}} &(11)\\ &+ \left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\hat{\mathcal{B}}_{x_t}(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t - \tilde{\beta}_t)\right\|_{\mathcal{G}} &(12)\end{aligned}$$

For (10), we have

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\tilde{\beta}_t\right\|_{\mathcal{G}}\\ &\le \Delta\left\|\mu_{H_t|x_{t-1:1}}\right\|_{\mathcal{G}} \le \Delta\end{aligned} \tag{13}$$

For (11), we have

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t - \tilde{\beta}_t)\right\|_{\mathcal{G}}\\ &\le \Delta((1+\Delta)^{t-1}\delta_1 + (1+\Delta)^{t-1} - 1)\end{aligned} \tag{14}$$

For (12), we have

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\hat{\mathcal{B}}_{x_t}(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t-\tilde{\beta}_t)\right\|_{\mathcal{G}}\\ \leq&\left\|\mathcal{A}_{x_t}\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\beta}_t-\tilde{\beta}_t)\right\|_{\mathcal{G}}\\ \leq&(1+\Delta)^{t-1}\delta_1+(1+\Delta)^{t-1}-1 \qquad (15)\end{aligned}$$

Summing the above three bounds, we have

$$\begin{aligned}&(1+\Delta)^t\delta_1+(1+\Delta)^t-(1+\Delta)+\Delta\\ =&(1+\Delta)^t\delta_1+(1+\Delta)^t-1 \qquad (16)\end{aligned}$$

■

Lemma 2 $\left\|\mu_{X_{t+1}|x_{t:1}}-\hat{\mu}_{X_{t+1}|x_{t:1}}\right\|_{\mathcal{F}} \leq C_{\mathcal{O}}\delta_\infty + (C_{\mathcal{O}}C_{\mathcal{T}}+C_{\mathcal{O}}\delta_\infty)\left((1+\Delta)^t\delta_1+(1+\Delta)^t-1\right)$

Proof Using triangle inequality

$$\begin{aligned}&\left\|\mu_{X_{t+1}|x_{t:1}}-\hat{\mu}_{X_{t+1}|x_{t:1}}\right\|_{\mathcal{F}}\\ =&\left\|\tilde{\beta}_\infty\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1-\hat{\beta}_\infty\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1\right\|_{\mathcal{F}}\\ \leq&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1\right\|_{\mathcal{F}} \qquad (17)\\ +&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{F}} \qquad (18)\\ +&\left\|\tilde{\beta}_\infty(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{F}} \qquad (19)\end{aligned}$$

For (17), we have

$$\begin{aligned}&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1\right\|_{\mathcal{F}}\\ \leq&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1\right\|_{\mathcal{G}}\\ \leq&\left\|\tilde{\beta}_\infty-\hat{\beta}_\infty\right\|_2\left\|\mathcal{O}\right\|_2\left\|\mu_{H_{t+1}|x_{t:1}}\right\|_{\mathcal{G}}\\ \leq&C_{\mathcal{O}}\left\|\tilde{\beta}_\infty-\hat{\beta}_\infty\right\|_2\leq C_{\mathcal{O}}\delta_\infty \qquad (20)\end{aligned}$$

For (18), we have

$$\begin{aligned}&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{F}}\\ \leq&\left\|(\tilde{\beta}_\infty-\hat{\beta}_\infty)(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{G}}\\ \leq&\left\|\tilde{\beta}_\infty-\hat{\beta}_\infty\right\|_2\left\|\mathcal{O}\right\|_2\left((1+\Delta)^t\delta_1+(1+\Delta)^t-\gamma\right)\\ \leq&C_{\mathcal{O}}\delta_\infty\left((1+\Delta)^t\delta_1+(1+\Delta)^t-1\right) \qquad (21)\end{aligned}$$

For (19), we have

$$\begin{aligned}&\left\|\tilde{\beta}_\infty(\hat{\mathcal{U}}^\top\mathcal{O})(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{F}}\\ \leq&\left\|\tilde{\beta}_\infty(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_{t:1}}\hat{\beta}_1-\tilde{\mathcal{B}}_{x_{t:1}}\tilde{\beta}_1)\right\|_{\mathcal{G}}\\ \leq&\left\|\mathcal{O}\mathcal{T}\right\|_2\left((1+\Delta)^t\delta_1+(1+\Delta)^t-1\right)\\ \leq&C_{\mathcal{O}}C_{\mathcal{T}}\left((1+\Delta)^t\delta_1+(1+\Delta)^t-1\right) \qquad (22)\end{aligned}$$

■

Define:

$$\epsilon_1:=\left\|\mu_1-\hat{\mu}_1\right\|_{\mathcal{F}} \qquad (23)$$

$$\epsilon_2:=\left\|\mathcal{C}_{2,1}-\hat{\mathcal{C}}_{2,1}\right\|_{\mathcal{F}\otimes\mathcal{F}} \qquad (24)$$

$$\epsilon_3:=\max_x\left\|\mu_{3,1|x}-\hat{\mu}_{3,1|x}\right\|_{\mathcal{F}\otimes\mathcal{F}} \qquad (25)$$

Lemma 3 *Suppose $\epsilon_2\leq\varepsilon\sigma_N(\mathcal{C}_{2,1})$ for some $\varepsilon\leq 1/2$. Let $\varepsilon_0=\epsilon_2^2/((1-\varepsilon)\sigma_N(\mathcal{C}_{2,1}))^2$. Then:*

1. $\varepsilon_0<1$,
2. $\sigma_n(\hat{\mathcal{U}}^\top\hat{\mathcal{C}}_{2,1})\geq(1-\varepsilon)\sigma_N(\mathcal{C}_{2,1})$,
3. $\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})\geq\sqrt{1-\varepsilon_0}\sigma_N(\mathcal{C}_{2,1})$,
4. $\sigma_N(\hat{\mathcal{U}}^\top\mathcal{O})\geq\sqrt{1-\varepsilon_0}\sigma_N(\mathcal{O})$.

Proof This lemma can be proved by an extension of Lemma 9 in **?**. ■

Lemma 4 $\delta_1:=\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\tilde{\beta}_1-\hat{\beta}_1)\right\|_{\mathcal{G}}\leq\frac{\epsilon_1}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{O})}$

Proof

$$\begin{aligned}&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\tilde{\beta}_1-\hat{\beta}_1)\right\|_{\mathcal{G}}\\ =&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{U}}^\top\mu_1-\hat{\mathcal{U}}^\top\hat{\mu}_1)\right\|_{\mathcal{G}}\\ \leq&\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}\hat{\mathcal{U}}^\top\right\|_2\left\|\mu_1-\hat{\mu}_1\right\|_{\mathcal{F}} \qquad (26)\\ \leq&\frac{\epsilon_1}{\sigma_n(\hat{\mathcal{U}}^\top\mathcal{O})} \qquad (27)\end{aligned}$$

■

Lemma 5 $\delta_\infty:=\left\|\tilde{\beta}_\infty-\beta_\infty\right\|_2\leq\frac{1+\sqrt{5}}{2}\frac{\epsilon_2}{\min\{\sigma_N(\hat{\mathcal{C}}_{2,1}),\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})\}^2}+\frac{\epsilon_2}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})}$

Proof

$$
\begin{aligned}
&\left\|\tilde{\beta}_\infty - \beta_\infty\right\|_2 \\
=&\left\|\mathcal{C}_{2,1}(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger - \hat{\mathcal{C}}_{2,1}(\hat{\mathcal{U}}^\top\hat{\mathcal{C}}_{2,1})^\dagger\right\|_2 \\
\leq&\left\|\hat{\mathcal{C}}_{2,1}((\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger - (\hat{\mathcal{U}}^\top\hat{\mathcal{C}}_{2,1})^\dagger)\right\|_2 \\
+&\left\|(\mathcal{C}_{2,1} - \hat{\mathcal{C}}_{2,1})(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger\right\|_2 \\
\leq&\left\|\hat{\mathcal{C}}_{2,1}\right\|_2\left\|((\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger - (\hat{\mathcal{U}}^\top\hat{\mathcal{C}}_{2,1})^\dagger)\right\|_2 \\
+&\left\|\mathcal{C}_{2,1} - \hat{\mathcal{C}}_{2,1}\right\|_2\left\|(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger\right\|_2 \\
\leq&\frac{1+\sqrt{5}}{2}\frac{\epsilon_2}{\min\{\sigma_N(\hat{\mathcal{C}}_{2,1}),\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})\}^2} + \frac{\epsilon_2}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})}
\end{aligned} \tag{28}
$$

■

Lemma 6 $\Delta := \max_{x_t}\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2 \leq \frac{1}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{O})}\left(\frac{\epsilon_2}{\min\{\sigma_N(\hat{\mathcal{C}}_{2,1}),\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})\}^2} + \frac{\epsilon_3}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})}\right)$

Proof

$$
\begin{aligned}
&\max_{x_t}\left\|(\hat{\mathcal{U}}^\top\mathcal{O})^{-1}(\hat{\mathcal{B}}_{x_t} - \tilde{\mathcal{B}}_{x_t})(\hat{\mathcal{U}}^\top\mathcal{O})\right\|_2 \\
\leq&\max_{x_t}\frac{1}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{O})}\Big(\left\|\hat{\mu}_{3,1|x_t}((\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger - (\hat{\mathcal{U}}^\top\hat{\mathcal{C}}_{2,1})^\dagger)\right\|_2 \\
+&\left\|(\mu_{3,1|x_t} - \hat{\mu}_{3,1|x_t})(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})^\dagger\right\|_2\Big) \\
\leq&\frac{1}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{O})}\left(\frac{\epsilon_2}{\min\{\sigma_N(\hat{\mathcal{C}}_{2,1}),\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})\}^2} + \frac{\epsilon_3}{\sigma_N(\hat{\mathcal{U}}^\top\mathcal{C}_{2,1})}\right)
\end{aligned} \tag{29}
$$

■

Based on the properties of Hilbert space embeddings explained in the main text, we have

Lemma 7 *With probability* $1-\delta$,

$$\epsilon_1 = \|\mu_1 - \hat{\mu}_1\|_{\mathcal{F}} = O_p(m^{-1/2}(\log(1/\delta))^{1/2}) \tag{30}$$

$$\epsilon_2 = \left\|\mathcal{C}_{2,1} - \hat{\mathcal{C}}_{2,1}\right\|_{\mathcal{F}\otimes\mathcal{F}} = O_p(m^{-1/2}(\log(1/\delta))^{1/2}) \tag{31}$$

$$\epsilon_3 = \max_{x_t}\left\|\mu_{3,1|x_t} - \hat{\mu}_{3,1|x_t}\right\|_{\mathcal{F}\otimes\mathcal{F}} = O_p(\lambda^{1/2} + (\lambda m)^{-1/2}(\log(1/\delta))^{1/2}). \tag{32}$$

Plugging these results, and we finally have

Theorem 8 *Assume* $k(x,x')$ *and* $l(h,h')$ *bounded, and* $\max_x \|\mathcal{A}_x\|_2 \leq 1$, *then with probability* $1-\delta$, $\left\|\mu_{X_{t+1}|x_{t:1}} - \hat{\mu}_{X_{t+1}|x_{t:1}}\right\|_{\mathcal{F}} = O_p(\lambda^{1/2} + (\lambda m)^{-1/2}(\log(1/\delta))^{1/2})$.