6. Supplementary Material

We now give a detailed account for the theorems stated in section 4.

6.1. Preliminaries I

In what follows we use the following notation: For an $n \times n$ matrix A, $\text{vec}(A) \in \mathbb{R}^{n^2}$ is the result of stacking its columns vertically into a single long vector. Thus, its Frobenius matrix norm is $\|A\|_F = \|\text{vec}(A)\|_2$.

Recall the definition of g_ψ:

$$g_\psi \equiv \frac{2G}{1-\psi}.$$

One can easily verify that for $2G \geq 1$, we have $1 + \psi g_\psi \leq g_\psi^2$. Also recall that assumption (2b) states that the distributions in $\mathcal{F}_n^\theta$ are bounded by L, which is defined by:

$$\max_{i \in [n]} \sup_{y \in \mathbb{R}} f_{\theta_i}(y) \leq L < \infty.$$

The following concentration result from Kontorovich & Weiss (2012, Theorem 1) is our main tool in proving the error bounds given here.

Lemma 1. *Let $Y = Y_0, \ldots, Y_{T-1} \in \mathcal{Y}^T$ be the output of a Hidden Markov chain with transition matrix A and output distributions $\mathcal{F}_n^\theta$. Assume that A is geometrically ergodic with constants G, ψ as in (1). Let $F : (Y_0, \ldots, Y_{T-1}) \mapsto \mathbb{R}$ be any function that is l-Lipschitz with respect to the Hamming metric on $\mathcal{Y}^T$. Then, for all $\epsilon > 0$,*

$$P(|F(Y) - \mathbf{E}F| > \epsilon T) \leq 2\exp\left(-\frac{T(1-\psi)^2\epsilon^2}{2l^2G^2}\right). \quad (31)$$

We will also need the following Lemma (proved in (Kontorovich & Weiss, 2012) for the discrete output case but easily generalize to continuous outputs) for bounding the variance of our estimators.

Lemma 2. *Let $f(y) : \mathbb{R} \to \mathbb{R}^+$ be a function of the observables of an n states geometrically ergodic HMM with constants (G, ψ) and*

$$\int_{\mathcal{Y}} f(y)dy \leq 1.$$

Assume the HMM is started with the stationary distribution $\boldsymbol{\pi}$. Then

$$\text{Var}\left[\frac{1}{T}\sum_{t=0}^{T-1} f(Y_t)\right] \leq \frac{\text{Var}[f(Y)]}{T} + \frac{\psi g_\psi \mathbf{E}[f(Y)]}{T}.$$

Similarly, let $g(y, y') : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ be a function of **consecutive** *observations (y, y') such that*

$$\iint_{\mathcal{Y}} g(y, y')dydy' \leq 1.$$

Then

$$\text{Var}\left[\frac{1}{T}\sum_{t=1}^{T-1} g(Y_t, Y_{t+1})\right] \leq \frac{\text{Var}[g(Y,Y')]}{T-1} + \frac{(1+\psi g_\psi)\mathbf{E}[g(Y,Y')]}{T-1}.$$

6.2. Accuracy of $\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\eta}}$

Since our estimators $\hat{\boldsymbol{\pi}}$ and $\hat{A}$ are constructed in terms of $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\sigma}}$ in the discrete case, and $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\eta}}$ in the continuous case, let us first examine the accuracy of the later. The following results shows that geometric ergodicity is sufficient to ensure their rapid convergence to the true values.

Lemma 3. Discrete case. *Let $(y_t)_{t=1}^T$ be an observed sequence from a discrete output HMM whose initial state X_0 follows the stationary distribution $\boldsymbol{\pi}$. Let $\boldsymbol{\rho}$ be given by (3) and $\boldsymbol{\sigma}$ by (4) with their empirical estimates given in (5). Then*

$$\mathbf{E}[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2] \leq \sqrt{\frac{1+\psi g_\psi}{T}} \quad (32)$$

$$\mathbf{E}[\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2] \leq \sqrt{\frac{2+\psi g_\psi}{T-1}} \quad (33)$$

Furthermore, for any $\epsilon > 0$,

$$P(\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2 > \sqrt{\tfrac{1+\psi g_\psi}{T}} + \epsilon) \leq 2\exp\left(-\tfrac{2T\epsilon^2}{g_\psi^2}\right) \quad (34)$$

and

$$P\left(\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2 > \sqrt{\frac{2+\psi g_\psi}{T-1}} + \epsilon\right) \leq 2\exp\left(\frac{-2(T-1)\epsilon^2}{g_\psi^2}\right). \quad (35)$$

Finally, we have for any fixed $\mathbf{v} \in \mathbb{R}^m$ with $\|\mathbf{v}\|_2 = 1$,

$$P(|\langle\hat{\boldsymbol{\rho}}, \mathbf{v}\rangle - \langle\boldsymbol{\rho}, \mathbf{v}\rangle| > \epsilon) \leq 2\exp\left(-\frac{2T\epsilon^2}{g_\psi^2}\right). \quad (36)$$

Proof. First note that w.r.t the Hamming metric, $T\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2$ and $|\langle\hat{\boldsymbol{\rho}}, \mathbf{v}\rangle - \langle\boldsymbol{\rho}, \mathbf{v}\rangle|$ are 1-Lipschitz and $T\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2$ is 2-Lipschitz. Thus the claims in (34, 35, 36) all follows directly from Lemma 1 where for

(34, 35) we also take into account (32) and (33) respectively. In order to prove (32) note that

$$\mathbf{E}[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2] = \sum_{k\in[n]} \mathbf{E}(\hat{\rho}_k - \rho_k)^2 = \sum_{k\in[n]} Var(\hat{\rho_k}).$$

So by taking in Lemma 2, $f(y) = \mathbb{1}_{y=k}$, we have $\mathbf{E}[\mathbb{1}_{y=k}] = \rho_k$ and $Var(\mathbb{1}_{y=k}) = \rho_k(1-\rho_k) \le \rho_k$. Since $\sum_{k=1}^m \rho_k = 1$ we get the desired bound.

The bound in (33) is obtained similarly by taking $g(y,y') = \mathbb{1}_{y=k}\mathbb{1}_{y'=k'}$ in Lemma 2 with the fact that $\sum_{kk'} \sigma_{kk'} = 1$. □

Lemma 4. Continuous case. *Let $(Y_t)_{t=1}^T$ be an observed sequence from a continuous observations HMM whose initial state X_0 follows the stationary distribution $\boldsymbol{\pi}$. Let $\boldsymbol{\xi}$ be given by (13) , $\boldsymbol{\eta}$ by (18) and $\hat{\boldsymbol{\xi}}$ and $\hat{\boldsymbol{\eta}}$ be their empirical estimates, given by (14) and (19) respectively. Then for any $\epsilon > 0$,*

$$P\left(\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2 > \epsilon\right) \le 2n \exp\left(-\frac{2T\epsilon^2}{g_\psi^2 n L^2}\right), \qquad (37)$$

and

$$P\left(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 > \epsilon\right) \le \qquad (38)$$
$$2n^2 \exp\left(-\frac{2(T-1)\epsilon^2}{g_\psi^2 n^2}\right).$$

Proof. Note that $\mathbf{E}\hat{\xi}_k = \xi_k$ and $T\hat{\xi}_k$ is L-Lipschitz for all $k \in [n]$. Thus by Lemma 1 and the union bound we have

$$P\left(\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_\infty > \epsilon'\right) \le 2n \exp\left(-\frac{2T\epsilon'^2}{g_\psi^2 L^2}\right). \qquad (39)$$

Since

$$\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2^2 = \sum_{k\in[n]} (\hat{\xi}_k - \xi_k)^2 \le n \left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_\infty^2,$$

we have

$$P\left(\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2 > \epsilon\right) \le P\left(\sqrt{n}\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_\infty > \epsilon\right).$$

putting $\epsilon' = \epsilon/\sqrt{n}$ in (39), the claim in (37) follows.

The proof of (38) follows the same paradigm as the proof for (39). Indeed $\mathbf{E}[\hat{\eta}_{kk'}] = \eta_{kk'}$ and $T\hat{\eta_{kk'}}$ is 1-Lipschitz so by Lemma 1 and the union bound we have

$$P\left(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_\infty > \epsilon'\right) \le 2n^2 \exp\left(-\frac{2T\epsilon'^2}{g_\psi^2 L^2}\right). \qquad (40)$$

Since

$$\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2^2 = \sum_{k,k'\in[n]\times[n]} (\hat{\eta}_{kk'} - \eta_{kk'})^2 \le n^2 \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_\infty^2,$$

we have

$$P\left(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 > \epsilon\right) \le P\left(n\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_\infty > \epsilon\right).$$

putting $\epsilon' = \epsilon/n$ in (40), the claim in (38) follows. □

6.3. Proof of theorem 1 - Strong consistency

We now prove the strong consistency of our estimators stated in Theorem 1.

Proof. For the discrete case, by Lemma 3, the expectation $\mathbf{E}[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2]$ goes to zero as $T \to \infty$. Furthermore, using the Borel-Cantelli lemma, $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2$ converge to its expectation a.s. concluding that $\hat{\boldsymbol{\rho}}$ converges a.s. to $\boldsymbol{\rho}$. The same argument goes for $\hat{\boldsymbol{\sigma}}$, $\hat{\boldsymbol{\xi}}$, $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\sigma}$, $\boldsymbol{\xi}$, $\boldsymbol{\eta}$ respectively.

Now, the function $f : \mathbb{R}^m \to \mathbb{R}^n$ given by $f(x) = (B^\intercal \operatorname{diag}(1/x)B)^{-1}\mathbf{1}$ is continuous on $\mathbb{R}_+^m$. Moreover, $f(\boldsymbol{\rho}) = \boldsymbol{\pi}$ since the optimization problem (7) has a unique minimizer x^* for all $\hat{\boldsymbol{\rho}}$, which in particular is given by $x^* = \boldsymbol{\pi}$ when $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}$. Since $\boldsymbol{\rho} \in \mathbb{R}_+^m$ by assumption, the argument above shows that almost surely, $\hat{\boldsymbol{\rho}} \in \mathbb{R}_+^m$ for all sufficiently large T. Therefore, $\lim_{T\to\infty} f(\hat{\boldsymbol{\rho}}) = f(\boldsymbol{\rho}) = \boldsymbol{\pi}$ almost surely, and the asymptotic strong consistency of $\hat{\boldsymbol{\pi}}$ is established.

To prove the asymptotic strong consistency of $\hat{A}$ in the discrete case, recall that the minimizer of the quadratic program $x^\intercal Kx - h^\intercal x$ subject to $Gx \le g$, $Dx = d$, is continuous under small perturbations of K, h, G, D, d (Dantzig et al., 1967). In particular, if $\hat{\boldsymbol{\pi}}$ is sufficiently close to $\boldsymbol{\pi}$ then $\hat{A}$ is close to A. Since $\hat{\boldsymbol{\pi}} \to \boldsymbol{\pi}$ and $\hat{\boldsymbol{\sigma}} \to \boldsymbol{\sigma}$ almost surely, we also have $\hat{A} \xrightarrow{\text{a.s.}} A$.

For the continuous observations case, note that $\hat{\pi}$ and $\hat{A}$ are also solutions of quadratic programs. Also note that $\hat{\boldsymbol{\xi}} \to \boldsymbol{\xi}$ and $\hat{\boldsymbol{\eta}} \to \boldsymbol{\eta}$ almost surely. Thus we have that $\hat{A} \xrightarrow{\text{a.s.}} A$ and $\hat{\boldsymbol{\pi}} \xrightarrow{\text{a.s.}} \boldsymbol{\pi}$ as above. □

6.4. Proof of Theorem 2: Bounding the error for $\hat{\pi}$ in the discrete observations case

Proof. Lemma 3 and the fact that $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty \le \|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2$ implies that $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty = O_P(1/\sqrt{T})$. Hence we make a change of variables,

$$\hat{\boldsymbol{\rho}} = \boldsymbol{\rho} + \frac{1}{\sqrt{T}}\zeta. \qquad (41)$$

To establish the (eventual) positivity of the entries of $\hat{\boldsymbol{\pi}}$, we consider the solution x^* of (8) with $\lambda = 0$, e.g.

without the normalization $\sum x_i = 1$, and write it as $x^* = \boldsymbol{\pi} + \delta$. Our goal is to understand the relation between δ and ζ.

Observe that δ satisfies the system of linear equations

$$\sum_j \Big(\sum_k \frac{B_{kj} B_{ki}}{\rho_k \left(1 + \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k}\right)} \Big) (\pi_j + \delta_j) = 1.$$

We need T sufficiently large so that, with high probability, $\max_k \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k} \ll 1$, or equivalently, $|\hat{\rho}_k - \rho_k| \ll \rho_k$.

By taking $T \gtrsim 4 g_\psi / a_1^2$ we have

$$\mathbf{E}[\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty] \le a_1/2.$$

So choosing $\epsilon = \min \rho_k/2 \ge a_1/2$ in (34), this condition is satisfied for $T \gtrsim g_\psi^2/a_1^2$. Then, approximating $1/(1+\epsilon) = 1 - \epsilon + O(\epsilon^2)$ gives

$$\sum_j \left[\sum_k \frac{B_{kj} B_{ki}}{\rho_k} \left(1 - \frac{1}{\sqrt{T}} \frac{\zeta_k}{\rho_k}\right) \right] (\pi_j + \delta_j) = 1 + O_P\left(\frac{1}{T}\right).$$

Note that since $B\boldsymbol{\pi} = \boldsymbol{\rho}$, the leading order correction for δ is simply

$$\delta = \frac{1}{\sqrt{T}} (\tilde{B}^\intercal \tilde{B})^{-1} \tilde{B}^\intercal \left(\frac{\zeta}{\boldsymbol{\rho}}\right) + O_P\left(\frac{1}{T}\right),$$

where the matrix $\tilde{B} = \mathrm{diag}(1/\sqrt{\boldsymbol{\rho}}) B$.

Let $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ be the right and left singular vectors of $\tilde{B}$ with non-zero singular values $\sigma_i(\tilde{B})$, where $\sigma_1 \le \sigma_2 \ldots \le \sigma_n$; thus, $\tilde{B}\mathbf{u}_i = \sigma_i \mathbf{v}_i$. The fact that $\tilde{B}$ also has n non-zero singular values follows from its definition combined with our Assumption 2d that B has rank n. Then

$$\tilde{B}^\intercal \tilde{B} = \sum_i \sigma_i^2 \mathbf{u}_i \mathbf{u}_i^\intercal \tag{42}$$

and hence,

$$\delta = \frac{1}{\sqrt{T}} \sum_i \frac{1}{\sigma_i} \langle \frac{\zeta}{\boldsymbol{\rho}}, \mathbf{v}_i \rangle \mathbf{u}_i + O_P\left(\frac{1}{T}\right) \tag{43}$$

For the solution x to have strictly positive coordinates we need that $|\delta_j| < \pi_j$ for each of $j = 1, \ldots, n$. Without loss of generality, assume that $\pi_1 = \min_j \pi_j$ and analyze the worst-case setting. This occurs when the singular vector $\mathbf{u}_1$ with smallest singular value coincides with the standard basis vector $\mathbf{e}_1$. Then,

$$|\delta_1| \le \frac{1}{\sqrt{T}} \frac{1}{\sigma_1(\tilde{B}) \min_j \rho_j} |\langle \zeta, \mathbf{v}_1 \rangle| + O_P\left(\frac{1}{T}\right). \tag{44}$$

It follows from (36) that $|\delta_1|$ will be dominated by $\min \pi_j \ge a_0$ provided that

$$T \gtrsim \frac{g_\psi}{a_0 a_1 \sigma_1(\tilde{B})}. \tag{45}$$

In the unlikely event that (i) the vector $\boldsymbol{\pi}$ is uniform ($\pi_j = 1/n$ for all j), (ii) the matrix $\tilde{B}$ has n identical singular values, we need the equation analogous to (44) to hold for all n coordinates. By a union bound argument, an additional factor of $\log n$ in the number of samples suffices to ensure, with high probability, the non-negativity of the solution x.

Next we proceed to bound $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2^2$. To this end, we write

$$x^* - \boldsymbol{\pi} = \delta = \sum_i \frac{1}{\sigma_i(\tilde{B})} \langle \frac{\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}}{\boldsymbol{\rho}}, \mathbf{v}_i \rangle \mathbf{u}_i + O_P\left(\frac{1}{T}\right).$$

Since both the $\{\mathbf{u}_i\}$ and the $\{\mathbf{v}_i\}$ are orthonormal,

$$\begin{aligned} \|\delta\|_2^2 &= \sum_i \frac{1}{\sigma_i^2(\tilde{B})} \langle \frac{\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}}{\boldsymbol{\rho}}, \mathbf{v}_i \rangle^2 \\ &\le \frac{1}{\sigma_1^2(\tilde{B}) (\min \rho_k)^2} \sum_i \langle \hat{\boldsymbol{\rho}} - \boldsymbol{\rho}, \mathbf{v}_i \rangle^2 \\ &\le \frac{\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2}{\sigma_1^2(\tilde{B}) a_1^2}. \end{aligned}$$

Bounding $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_2^2$ via Lemma 3 and noting that

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 = \left\| \tfrac{x^*}{\|x^*\|_1} - \boldsymbol{\pi} \right\|_2 \le 2 \|x^* - \boldsymbol{\pi}\|_2 = 2 \|\delta\|_2,$$

the result in (22) follows. □

6.5. Preliminaries II

The remaining estimators ($\hat{\pi}$ for the continuous observations case, and $\hat{A}$ for both the discrete and continuous observations cases) are obtained as solutions for quadratic programs. Let us take for example the QP for calculating $\hat{\boldsymbol{\pi}}$ with continuous observations HMM, given in (23). For this case, the QP is equivalent to

$$\hat{\boldsymbol{\pi}} = \operatorname*{argmin}_x \frac{1}{2} x^\intercal K^\intercal K x - x^\intercal K^\intercal \hat{\boldsymbol{\xi}}$$

subject to $x \ge 0$ and $\sum_i x_i = 1$.

Note that if $\hat{\boldsymbol{\xi}}$ was equal to its true values $\boldsymbol{\xi}$, the solution of the above QP would simply be the true $\boldsymbol{\pi}$. In reality, we only have the estimate $\hat{\boldsymbol{\xi}}$. In order to analyze the error $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2$, we will need to consider how the solutions of such a quadratic program are affected by errors in $\boldsymbol{\xi}$.

More generally, we are concerned with two QPs

$$\min Q(x) \quad = \quad \min \frac{1}{2} x^\intercal M x - x^\intercal h, \tag{46}$$

$$\min \hat{Q}(x) \quad = \quad \min \frac{1}{2} x^\intercal \hat{M} x - x^\intercal \hat{h}, \tag{47}$$

both subject to $Gx \leq g$, $Dx = d$. We assume that the solution to the first QP is the "true" value while the solution to the second is our estimate. So bounding the estimate error is equivalent to bounding the error between the solutions obtained by the above two QPs, where $\hat{M}$ and $\hat{h}$ are perturbed versions of M and h.

Given that, note that only the objective function has been perturbed, while the linear constraints remained unaffected. We may thus apply the following classical result on the solution stability of definite quadratic programs.

Theorem 7. *(Daniel, 1973) Let $\lambda = \lambda_{\min}(M)$ be the smallest eigenvalue of M, and let $\epsilon = \max\{\|\hat{M} - M\|_2, \|\hat{h} - h\|_2\}$. Let x and $\hat{x}$ be the minimizers of Eqs.(46) and (47), respectively. Then, for $\epsilon < \lambda$,*

$$\|x - \hat{x}\|_2 \leq \frac{\epsilon}{\lambda - \epsilon}(1 + \|x\|_2).$$

In the following we will obtain bounds on ϵ and λ for the different estimators and invoke the above theorem.

6.6. Proof of Theorem 3: Bounding the error for $\hat{\pi}$ in the continuous observations case

Proof. Note that in the notation given in Theorem 7, we have $h = \boldsymbol{\xi}^\intercal K$ and $\hat{h} = \hat{\boldsymbol{\xi}}^\intercal K$. Since we assumed that the output density parameters are known exactly we have no error in $M = K^\intercal K$.

It is immediate that

$$\lambda_{min}(K^\intercal K) = \sigma_1^2(K),$$

and

$$\epsilon \leq \left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2 \|K\|_2 \leq nL \left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2.$$

From Lemma 4 we have

$$\left\|\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right\|_2 \lesssim_P \sqrt{\frac{(n \ln n) g_\psi^2 L^2}{T}},$$

while by Theorem 7 we have

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 \lesssim \frac{\epsilon}{\lambda_{min}(K^\intercal K)}(1 + \|\boldsymbol{\pi}\|_2).$$

Since $\|\boldsymbol{\pi}\|_2 \leq 1$, the claim follows. □

As a side remark we note that the form of (24) is somewhat counter-intuitive, as it suggests a worse behavior for larger L. Intuitively, however, larger L corresponds to a more peaked — and hence lower-variance — density, which ought to imply sharper estimates. Note however that as numerical simulations suggest we typically have

$$\frac{\sigma_1^2(\tilde{F}) L^2}{\sigma_1^2(\tilde{K})} = O(1).$$

Thus, whenever $\sigma_1^2(\tilde{F})$ is well behaved so is the estimate in (24) and the bound is reasonable after all. Finally note that F is stochastic so it behaves very much like the matrix B in the discrete outputs case.

6.7. Proof of Theorem 4: Bounding the error of $\hat{A}$ in the discrete observations case

Let $\hat{A}$ be the solution of

$$\min_{A_{ij} \geq 0, \sum_i A_{ij} = 1} \|\hat{\boldsymbol{\sigma}} - \hat{C} A\|_2^2, \tag{25}$$

where $\hat{\boldsymbol{\sigma}}$ is given in (5). Recall that $C_{ij}^{kk'} = \pi_j B_{kj} B_{k'i}$ and $\hat{C}_{ij}^{kk'} = \hat{\pi}_j B_{kj} B_{k'i}$. First note that if $\boldsymbol{\pi}$ and $\boldsymbol{\sigma}$ were known exactly, the above QP could be written as

$$\min Q(A) = \min \frac{1}{2} \operatorname{vec}(A)^\intercal M \operatorname{vec}(A) - \operatorname{vec}(A)^\intercal h \tag{48}$$

where $M = C^\intercal C$ and $h = C^\intercal \operatorname{vec}(\boldsymbol{\sigma})$. Its solution is precisely the transition probability matrix A. In reality, as we only have estimates $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\sigma}}$, the optimization problem is perturbed to

$$\min \hat{Q}(A) = \min \frac{1}{2} \operatorname{vec}(A)^\intercal \hat{M} \operatorname{vec}(A) - \operatorname{vec}(A)^\intercal \hat{h} \tag{49}$$

where $\hat{M} = \hat{C}^\intercal \hat{C}$, and $\hat{h} = \hat{C}^\intercal \operatorname{vec}(\hat{\boldsymbol{\sigma}})$.

To analyze how errors in $\hat{\boldsymbol{\sigma}}$ and $\hat{C}$ affect the optimization problem we follow the same route as above. Thus we need to bound $\|\hat{h} - h\|_2$, $\|\hat{M} - M\|_2$, and the smallest eigenvalue of M. Regarding the latter, by definition, $\lambda_{\min}(M) = \sigma_1^2(C)$, where $\sigma_1(C)$ is the smallest singular value of C. A simple exercise in linear algebra yields

$$\sigma_1(C) \geq a_0 \sigma_1^2(B). \tag{50}$$

The following lemma provides bounds on $\|\hat{M} - M\|_2$ and on $\|\hat{h} - h\|_2$.

Lemma 5. *Asymptotically, as $T \to \infty$,*

$$\|\hat{h} - h\|_2 \lesssim_P \sqrt{n} \left(\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 + \|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_2\right) \tag{51}$$

and

$$\|\hat{M} - M\|_2 \lesssim_P 2n \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2. \tag{52}$$

Proof. By definition, $h_{ij} = \sum_{k,k'} C_{ij}^{kk'} \sigma_{kk'}$, and $\hat{h}_{ij} = \sum_{k,k'} \hat{C}_{ij}^{kk'} \hat{\sigma}_{kk'}$. Using the definitions of C and $\hat{C}$, up to mixed terms $O(\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty \|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_\infty)$, we obtain

$$\begin{aligned} \hat{h}_{ij} - h_{ij} &= (\hat{\pi}_j - \pi_j) \sum_{kk'} B_{kj} B_{k'i} \sigma_{kk'} \\ &\quad + \pi_j \sum_{kk'} B_{kj} B_{k'i} (\hat{\sigma}_{kk'} - \sigma_{kk'}) \end{aligned}$$

Since each of $\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty$ and $\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_\infty$ are $O_P(1/\sqrt{T})$, the neglected mixed terms are asymptotically negligible as compared to each of the first two ones. Next, we use the fact that $\sigma_{kk'} \leq 1, \pi_j \leq 1$ and $\sum_{kk'} B_{kj} B_{k'i} \leq 1$ to obtain that

$$\left\| \hat{h} - h \right\|_2 \lesssim_P \sqrt{n} \, \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 + \sqrt{n} \, \|\mathrm{vec}(\hat{\boldsymbol{\sigma}}) - \mathrm{vec}(\boldsymbol{\sigma})\|_2$$

Similarly, we have that for the $n^2 \times n^2$ matrix M, and not including higher order mixed terms $(\hat{\pi}_j - \pi_j)(\hat{\pi}_\beta - \pi_\beta)$, which are asymptotically negligible,

$$\begin{aligned} (\hat{M} - M)_{ij,\alpha\beta} &= (\hat{\pi}_j - \pi_j)\pi_\beta \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &\quad + (\hat{\pi}_\beta - \pi_\beta)\pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \end{aligned}$$

Note that $\sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} = (\sum_k B_{kj} B_{k\beta})(\sum_{k'} B_{k'i} B_{k'\alpha}) \leq 1$. Hence, by similar arguments as for h, (52) follows. □

We can now prove Theorem 4:

Proof. (**of Theorem 4**) Lemma 3, together with (22), implies that with high probability,

$$\|\hat{\boldsymbol{\sigma}} - \boldsymbol{\sigma}\|_F \lesssim_P \sqrt{\frac{g_\psi^2}{T-1}},$$

and

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 \lesssim_P \sqrt{\frac{g_\psi^2}{T a_1^2 \sigma_1^2(\tilde{B})}}.$$

Inserting these into (51) and (52) yields, w.h.p.,

$$\begin{aligned} \epsilon &= \max\left\{ \left\| \hat{h} - h \right\|_2, \left\| \hat{M} - M \right\|_2 \right\} \\ &\lesssim \sqrt{\frac{n^2 g_\psi^2}{T a_1^2 \sigma_1^2(\tilde{B})}}. \end{aligned} \tag{53}$$

By Theorem 7, we have that

$$\left\| \hat{A} - A \right\|_F \lesssim \frac{\epsilon}{\lambda_1(M)} (1 + \|A\|_F), \tag{54}$$

where $\|A\|_F \leq \sqrt{n}$ since A is column-stochastic. The claim follows by substituting the bounds on ϵ in (53) and on $\lambda_1(M) = \sigma_1^2(C) \geq a_0^2 \sigma_1^4(B)$ in (50) into (54) and noting that $\sigma_1^2(\tilde{B}) \geq \sigma_1^2(B)$. □

6.8. Proof of Theorem 5: Bounding the error of $\hat{A}$ in the continuous observations case

Let $\hat{A}$ be the solution of

$$\min_{A_{ij} \geq 0, \sum_i A_{ij} = 1} \|\hat{\boldsymbol{\eta}} - \hat{C} A\|_2^2, \tag{25}$$

where $\hat{\boldsymbol{\eta}}$ is given in (19) and $C_{ij}^{kk'} = \pi_j F_{kj} F_{k'i}$ and $\hat{C}_{ij}^{kk'} = \hat{\pi}_j F_{kj} F_{k'i}$. The above QP can be written as

$$\min \hat{Q}(A) = \min \frac{1}{2} \mathrm{vec}(A)^\intercal \hat{M} \, \mathrm{vec}(A) - \mathrm{vec}(A)^\intercal \hat{h} \tag{55}$$

where $\hat{M} = \hat{C}^\intercal \hat{C}$, and $\hat{h} = \hat{C}^\intercal \, \mathrm{vec}(\hat{\boldsymbol{\sigma}})$.

Exactly as in the previous subsection, we want to bound the difference between the solutions for the above QP and the unperturbed one.

First note that

$$\sigma_1(C) \geq a_0 \sigma_1^2(F). \tag{56}$$

Next we give the analogue of lemma 5.

Lemma 6. *Asymptotically, as $T \to \infty$,*

$$\|\hat{h} - h\|_2 \lesssim_P \sqrt{n} \left(\frac{1}{a_0} \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 + \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_2 \right) \tag{57}$$

and

$$\|\hat{M} - M\|_2 \lesssim_P 2n \frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2}{a_0}. \tag{58}$$

Proof. In contrast to Lemma 5, here F is also perturbed due to errors in $\hat{\boldsymbol{\pi}}$ with

$$\hat{F}_{ij} = \int_{\mathcal{Y}} \frac{\hat{\pi}_i f_i(y) f_j(y)}{\sum_k \hat{\pi}_k f_k(y)} dy.$$

Expending the difference $\Delta F_{ij} \equiv \left| \hat{F}_{ij} - F_{ij} \right|$ up to first order in $\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}$ we find that

$$\|\Delta F\|_F \leq \frac{\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty}{a_0} \|F\|_F \leq \frac{\sqrt{n} \, \|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_\infty}{a_0},$$

where in the last inequality we used the fact that F is stochastic. Repeating the arguments in the proof for Lemma 5 and noting that $a_0 \ll 1$ we get (57) and (58). □

We now come to the proof of Theorem 5.

Proof. (**of Theorem 5**) Lemma 4, together with (24), implies that with high probability,

$$\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\|_F \lesssim_P \sqrt{\frac{(n^2 \ln n) g_\psi^2}{T-1}},$$

and

$$\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_2 \lesssim_P \sqrt{\frac{(n^3 \ln n) g_\psi^2 L^4}{T\sigma_1^4(\tilde{K})}}$$

Inserting these into (57) and (58) yields, w.h.p.,

$$\begin{aligned} \epsilon &= \max\left\{ \left\|\hat{h} - h\right\|_2, \left\|\hat{M} - M\right\|_2 \right\} \\ &\lesssim \sqrt{\frac{(n^5 \ln n) g_\psi^2 L^4}{T\sigma_1^4(\tilde{K})}}. \end{aligned} \tag{59}$$

By Theorem 7, we have that

$$\left\|\hat{A} - A\right\|_F \lesssim \frac{\epsilon}{\lambda_1(M)}(1 + \|A\|_F), \tag{60}$$

where $\|A\|_F \leq \sqrt{n}$ since A is column-stochastic. The claim follows by substituting the bounds on ϵ in (59) and on $\lambda_1(M) = \sigma_1^2(C) \geq a_0^2 \sigma_1^4(F)$ in (50) into (60) and noting that $\sigma_1^2(\tilde{F}) \geq \sigma_1^2(F)$. □

As for remark 1, we point out that estimating $\boldsymbol{\eta}'$ with the help of the matrix K (instead of $\boldsymbol{\eta}$ with F) results in an estimator that is not $O(1/T)$-Lipschitz any more but $O(L^2/T)$-Lipschitz with $L = \max_{i\in[n]} \sup_{y\in\mathbb{R}} f_{\theta_i}(y)$. This means that in principle we will need many more samples to accurately estimate $\boldsymbol{\eta}'$ compared to $\boldsymbol{\eta}$, see Lemma 4. Thus, since in high dimensions calculating F via numerical integration may be computational intensive, choosing between the two estimators is in some sense choosing between working with limited number of samples and computational efficiency.

6.9. Proof of Theorem 6: Perturbations in the output parameters

We give here the proof for the perturbation in the matrix F. The proof for perturbations in the matrix K is similar.

Proof. By definition, $b_{ij} = \sum_{k,k'} C_{ij}^{kk'} \sigma_{kk'}$, and $\hat{b}_{ij} = \sum_{k,k'} \hat{C}_{ij}^{kk'} \hat{\sigma}_{kk'}$. Using the definitions of C and $\hat{C}$, up to first order in $\{\|\hat{\pi} - \pi\|_\infty, \|\hat{\sigma} - \sigma\|_\infty, \epsilon_F\}$ we obtain

$$\begin{aligned} \hat{b}_{ij} - b_{ij} = (\hat{\pi}_j - \pi_j) \sum_{kk'} B_{kj} B_{k'i} \sigma_{kk'} \\ + \pi_j \sum_{kk'} B_{kj} B_{k'i} (\hat{\sigma}_{kk'} - \sigma_{kk'}) \\ + \epsilon_F \pi_j \sum_{kk'} (P_{kj} B_{k'i} + B_{kj} P_{k'i}) \sigma_{kk'}. \end{aligned}$$

As the two first terms already considered we focus on the last term. It can be shown that:

$$\sum_{ij} \left(\pi_j \sum_{kk'} P_{kj} B_{k'i} \sigma_{kk'} \right)^2 \leq n \|P\|_F^2 .$$

Thus

$$\begin{aligned} \left\|\hat{b} - b\right\|_2 \leq \sqrt{n} \left(\|\hat{\pi} - \pi\|_2 + \|vec(\hat{\sigma}) - vec(\sigma)\|_2 + \right. \\ \left. + 2\epsilon_F \|P\|_F \right) (1 + o(1)). \end{aligned} \tag{61}$$

Similarly, for the matrix K up to first order in $\{\|\hat{\pi} - \pi\|_\infty, \epsilon_F\}$ we have

$$\begin{aligned} (\hat{K} - K)_{ij,\alpha\beta} &= (\hat{\pi}_j - \pi_j)\pi_\beta \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &+ (\hat{\pi}_\beta - \pi_\beta)\pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} \\ &+ \epsilon_F \pi_j \pi_\beta \sum_{kk'} P_{kj} B_{k\beta} B_{k'i} B_{k'\alpha} + \ldots \\ &+ \epsilon_F \pi_\beta \pi_j \sum_{kk'} B_{kj} B_{k\beta} B_{k'i} P_{k'\alpha}. \end{aligned}$$

Again considering only the terms including P and using the facts that $\sum_k B_{kj} B_{k\beta} \leq 1$ and $\sum_{kk'} (P_{kj} B_{k'i})^2 \leq \sum_k P_{kj}^2$ we similarly find that

$$\left\|\hat{K} - K\right\|_2 \leq (1 + o_p(1)) 2n \left(\|\hat{\pi} - \pi\|_2 + 4\epsilon_F \|P\|_F \right).$$

Repeating the analysis in the proofs for Theorems 3, 4 and 5 give the desired result. □