Picking Up the Pieces: Causal States in Noisy Data, and How to Recover Them
Supplementary Material

Gustav Eje Henter[a,*], W. Bastiaan Kleijn[a,b]

[a] Sound and Image Processing Laboratory, School of Electrical Engineering,
KTH – Royal Institute of Technology, SE-100 44 Stockholm, Sweden
[b] School of Engineering and Computer Science, Victoria University of Wellington,
PO Box 600, Wellington 6140, New Zealand

Abstract

This document contains supplementary material for the Pattern Recognition Letters article "Picking Up the Pieces: Causal States in Noisy Data, and How to Recover Them."

A. Proof of Theorem 1

In this appendix we prove that the criterion in Theorem 1 is sufficient for H_{AB} to have an infinite number of causal states. To begin with, we assume that the number of causal states is $M < \infty$. We then show that this leads to a contradiction.

Note that the distribution of future observations of the HMM from $t + 1$ on is completely determined by the hidden state probabilities at time t. The distribution over hidden states at t can be identified from the known HMM parameters and the symbols observed up until t. Consider a sequence of HMM observations $X_{-\infty}^{t} = x_{-\infty}^{t}$ for which the causal state at t is well defined. (The probability that $x_{-\infty}^{t}$ has an undefined causal state is zero.) Let the hidden-state probabilities associated with the current causal state be encoded by a vector $\boldsymbol{p}_t \in \mathbb{R}^n \setminus \boldsymbol{0}$ with elements proportional to the probabilities of each hidden state

[*] Corresponding author. Tel: +46 8 790 7420.
Email addresses: gustav.henter@ee.kth.se (Gustav Eje Henter), bastiaan.kleijn@ecs.vuw.ac.nz (W. Bastiaan Kleijn)

i, i.e., $(\boldsymbol{p}_t)_i \propto \mathbb{P}\left(S_t{=}i \mid X_{-\infty}^t{=}x_{-\infty}^t\right)$, for i from 1 to n. The direction of $\boldsymbol{p}_t$ uniquely gives the hidden-state distribution at t. By ignoring normalization we do not need to consider nonlinear renormalization operations later on.

Knowing $\boldsymbol{p}_t$ determines the causal state, but there may be different directions of $\boldsymbol{p}_t$ that correspond to the same causal state. Decompose $\boldsymbol{p}_t = \left[\boldsymbol{G}_r \boldsymbol{G}_0\right]\left[\boldsymbol{v}^T \boldsymbol{o}^T\right]^T$ where $\boldsymbol{v} \in \mathbb{C}^r \setminus \boldsymbol{0}$ and $\boldsymbol{o} \in \mathbb{C}^{n-r}$. Trivially $\mathbb{P}\left(S_{t+1}{=}i \mid X_{-\infty}^t{=}x_{-\infty}^t\right) \propto \left(\boldsymbol{A}^T \boldsymbol{p}_t\right)_i$. Since $\boldsymbol{G}_0 \boldsymbol{o}$ is in the null space of $\boldsymbol{A}^T$ we have that $\boldsymbol{A}^T \boldsymbol{p}_t = \boldsymbol{A}^T \boldsymbol{G}_r \boldsymbol{v}$. Hence the values in $\boldsymbol{o}$ do not affect future hidden state probabilities or observations, and the information in $\boldsymbol{v}$ suffices to fix the state.

Distinct causal states are distinguished by different probability distributions for the future symbols. We have that $\mathbb{P}\left(X_{t+1} = \sigma \mid X_{-\infty}^t = x_{-\infty}^t\right) \propto \left(\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r \boldsymbol{v}\right)_\sigma$. Therefore any change in direction for $\boldsymbol{v}$ will surely translate to a different next-symbol distribution if $\operatorname{rank}\left(\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r\right) = r$. There is then a one-to-one mapping between the different directions of $\boldsymbol{v} \in \mathbb{C}^r \setminus \boldsymbol{0}$ and all the different beliefs about the future that can be conceived. Not all these beliefs will necessarily be causal states of H_{AB} (not everything that can be believed is right or incorporates the information in $x_{-\infty}^t$), but each causal state is represented by a unique direction of $\boldsymbol{v}$.

For each additional symbol the HMM emits our beliefs about the future evolve according to the forward algorithm (Rabiner, 1989), as expressed by formula (5). Since we do not care about normalization, the algorithm can be written $\boldsymbol{p}_{t+1} = \operatorname{diag}\left(\boldsymbol{b}_{\cdot x_{t+1}}\right) \boldsymbol{A}^T \boldsymbol{p}_t$, where the history up until t is $x_{-\infty}^t$ and x_{t+1} is the new observation. If $\boldsymbol{p}_t$ is a causal state such that $\mathbb{P}\left(X_{t+1} = \sigma \mid X_{-\infty}^t = x_{-\infty}^t\right) > 0$ then $\boldsymbol{p}_{t+1}$ must also be a causal state of the process.

Consider the causal state at t given the history $x_{-\infty}^t$. Let $\boldsymbol{v}_t$ be a vector in $\mathbb{C}^r \setminus \boldsymbol{0}$ in the unique direction corresponding to this causal state $\varepsilon\left(x_{-\infty}^t\right)$. From the forward algorithm we see that we always can write $\boldsymbol{p}_{t+1} = \boldsymbol{G}_r \boldsymbol{v}_{t+1}$, where $\boldsymbol{v}_{t+1} = \boldsymbol{H}_r \operatorname{diag}\left(\boldsymbol{b}_{\cdot x_{t+1}}\right) \boldsymbol{A}^T \boldsymbol{G}_r \boldsymbol{v}_t = \boldsymbol{C}_{x_{t+1}} \boldsymbol{v}_t$.

Now assume that conditions 1 through 3 in the theorem statement hold. Consider a sequence of $M' \geq M$ new, all identical observations $X_{t+m} = \sigma$ for $1 \leq m \leq M'$. σ is a symbol from the alphabet subset $\mathcal{A}_{\text{sub}} \subseteq \mathcal{A}$ (which is nonempty) chosen such that $\boldsymbol{q}_t = \boldsymbol{Q}_\sigma^{-1}\boldsymbol{v}_t$ has two components $(\boldsymbol{q}_t)_i \neq 0$ and $(\boldsymbol{q}_t)_j \neq 0$ that satisfy $|\,(\boldsymbol{\Lambda}_\sigma)_{ii}\,| \neq |\,(\boldsymbol{\Lambda}_\sigma)_{jj}\,|$ (this is possible due to condition 3; the decomposition $\boldsymbol{C}_\sigma = \boldsymbol{Q}_\sigma \boldsymbol{\Lambda}_\sigma \boldsymbol{Q}_\sigma^{-1}$ exists by condition 2). Condition 1 ensures that $\mathbb{P}\left(X_{t+1}^{t+m} = [\sigma, \ldots, \sigma] \mid X_{-\infty}^{t} = x_{-\infty}^{t}\right) > 0$.

Define the ratio $u_{t+m} = |\left(\boldsymbol{q}_{t+m}\right)_i / \left(\boldsymbol{q}_{t+m}\right)_j|$, which is finite and nonzero. Applying the forward algorithm iteratively we can establish that $\boldsymbol{q}_{t+m} = \boldsymbol{\Lambda}_\sigma^m \boldsymbol{q}_t$. This implies that $u_{t+m} = |\,(\boldsymbol{\Lambda}_\sigma)_{ii} / (\boldsymbol{\Lambda}_\sigma)_{jj}\,|^m u_t$. Since $\boldsymbol{C}_\sigma$ is nonsingular and $|\,(\boldsymbol{\Lambda}_\sigma)_{ii}\,| \neq |\,(\boldsymbol{\Lambda}_\sigma)_{jj}\,|$ we have that the u-ratios satisfy either $u_t < u_{t+1} < \ldots < u_{t+M'}$ or $u_t > u_{t+1} > \ldots > u_{t+M'}$, and so all $\boldsymbol{q}_{t+m}$-vectors (and all $\boldsymbol{v}_{t+m}$-vectors) have different ratios between these two components and must point in different directions. Since we started from a causal state, all these $\boldsymbol{v}$-directions must represent other causal states of the process.

We have used an observation series with strictly positive probability to generate a sequence of M' new causal states that are both mutually distinct and distinct from the (arbitrary) starting state, meaning that the process has at least $M' + 1 \geq M + 1$ causal states. This contradicts the original assumption that the process had only M causal states. Since our reasoning applies for any finite M the number of causal states must be infinite.

B. Proof of Theorem 2

We wish to recover the causal structure of a CSSR-learnable process X_t using data from a distinguishable corruption Y_t, and show that this is possible if disturbances are not too strong, so that the precausal states of X_t are still discernible as well-separated clusters in the next-symbol probability space of Y_t.

The central argument of the theorem is to establish that the probability that any limited-resolution statistical test performed during robust homogenization makes an error—that is, makes a decision inconsistent with the suffix assignment corresponding to the precausal states of X_t—goes to zero as $N \to \infty$. This result establishes that RCS produces a suffix clustering identical to the precausal states of X_t with probability one in the limit. It is then trivial that the same suffix grouping as the causal states of X_t results after determinization.

To begin with, we must establish that all nonzero-probability suffixes will have appeared $n_{\max}$ times or more in the data already after a finite (if stochastic) time, so that the limited resolution criterion in (10) is used for every test when RCS is applied from that point on. Noting that Σ_X^L is finite, it is a straightforward consequence of the law of large numbers that such a sample size exists with probability one. Once this point is reached, all robust homogenization suffix assignments when applying RCS will be based on whether $d_m\left(\boldsymbol{p}, \boldsymbol{q}\right) > F_{\text{sig}}\left(\alpha\right)$ holds or not (so they are based on the metric d_m), and it suffices to show that, as $N \to \infty$, the (pre)causal states of X_t can be extracted reliably using this criterion applied to data from the corruption Y_t.

B.1. Distances in robust homogenization

Consider an arbitrary test performed during homogenization, where the estimated next-symbol distribution of a suffix $u \in \Sigma_X^L$ is compared against the estimated next-symbol distribution of a nonempty collection of other suffixes, the working state $V \subset \Sigma_X^L$. All suffixes $v \in V$ are taken to be from the same precausal state of X_t, and thus have the same unperturbed next-symbol distribution $\boldsymbol{q}_V$. (Unless any previous test makes an error, all tests performed during homogenization of X_t and robust homogenization of Y_t are of this type.) For RCS, the distribution estimates are based on an N-symbol string sampled from Y_t. We let $\widehat{\boldsymbol{p}}_u$ denote the next-symbol probability given history suffix u, as estimated from the available noisy data, while $\widetilde{\boldsymbol{p}}_u$ represents the actual next-symbol probability given u for the noisy process Y_t (on which $\widehat{\boldsymbol{p}}_u$ converges as N grows large).

Since d_m is a metric, the triangle inequality gives

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u, \widehat{\boldsymbol{q}}_V\right) \le\ & d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{p}}_u, \boldsymbol{p}_u\right) + d_m\left(\boldsymbol{p}_u, \boldsymbol{q}_V\right) \\ & + d_m\left(\boldsymbol{q}_V, \widetilde{\boldsymbol{q}}_V\right) + d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right) \qquad \text{(B.1)} \\ =\ & d_m\left(\widehat{\boldsymbol{p}}_u, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{p}}_u, \boldsymbol{p}_u\right) \\ & + d_m\left(\boldsymbol{q}_V, \widetilde{\boldsymbol{q}}_V\right) + d_m\left(\widetilde{\boldsymbol{q}}_V, \widehat{\boldsymbol{q}}_V\right) \qquad \text{(B.2)} \end{aligned}$$

in case u belongs in the same precausal state of the unperturbed process as V, and

$$\begin{aligned} d_m\left(\boldsymbol{p}_u,\, \boldsymbol{q}_V\right) \leq\; & d_m\left(\boldsymbol{p}_u,\, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{p}}_u\right) + d_m\left(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V\right) \\ & + d_m\left(\widehat{\boldsymbol{q}}_V,\, \widetilde{\boldsymbol{q}}_V\right) + d_m\left(\widetilde{\boldsymbol{q}}_V,\, \boldsymbol{q}_V\right), \end{aligned} \tag{B.3}$$

which we rearrange to establish

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V\right) \geq\; & d_m\left(\boldsymbol{p}_u,\, \boldsymbol{q}_V\right) - d_m\left(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u\right) - d_m\left(\widetilde{\boldsymbol{p}}_u,\, \boldsymbol{p}_u\right) \\ & - d_m\left(\boldsymbol{q}_V,\, \widetilde{\boldsymbol{q}}_V\right) - d_m\left(\widetilde{\boldsymbol{q}}_V,\, \widehat{\boldsymbol{q}}_V\right), \end{aligned} \tag{B.4}$$

if u does not belong in that precausal state. Note that the function d_m is nonnegative.

We now invoke the disturbance bound $\widetilde{d}$ from (12). Since the theorem assumes d_m to be convex and symmetric in the arguments, we have

$$d_m\left(\boldsymbol{q}_V,\, \widetilde{\boldsymbol{q}}_V\right) \leq \max_{v \in V} d_m\left(\boldsymbol{q}_V,\, \widetilde{\boldsymbol{p}}_v\right) \leq \widetilde{d}. \tag{B.5}$$

Thus the (ML estimated) expected perturbed precausal-state next-symbol distributions show limited differences from the unperturbed, original distributions, just like individual suffix distributions do. This bound—together with the earlier triangle inequalities (B.2) and (B.4), and the distinguishability $d_{\min}$ of next-step distributions defined in (11)—can be used to establish

$$d_m\left(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V\right) \leq 2\widetilde{d} + d_m\left(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u\right) + d_m\left(\widetilde{\boldsymbol{q}}_V,\, \widehat{\boldsymbol{q}}_V\right), \tag{B.6}$$

in case u belongs in the precausal state V, or

$$\begin{aligned} d_m\left(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V\right) \geq\; & d_{\min} - 2\widetilde{d} \\ & - d_m\left(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u\right) - d_m\left(\widetilde{\boldsymbol{q}}_V,\, \widehat{\boldsymbol{q}}_V\right), \end{aligned} \tag{B.7}$$

if u does not belong in V.

To discriminate between perturbed distributions from the same unperturbed precausal state, and those from different states, we want to choose the significance parameter α such that $F_{\text{sig}}\left(\alpha\right)$ falls between the upper and lower bounds above.

Since F_{sig} is assumed to be monotonic and continuous on $\alpha \in [0,\, 1]$, and extends over the entire range of d, there exists a nonempty interval I_{sig} such that $\alpha \in I_{\mathrm{sig}} \Rightarrow F_{\mathrm{sig}}(\alpha) \in (2\widetilde{d},\, d_{\mathrm{min}} - 2\widetilde{d})$, which is also nonempty since $2\widetilde{d} < \frac{1}{2} d_{\mathrm{min}}$ by (12). Under the chosen n_{max}, I_{sig} is the significance interval for which robust homogenization asymptotically will produce the desired suffix partitioning. It is centered on $\frac{1}{2} d_{\mathrm{min}}$. From now on, we assume $\alpha \in I_{\mathrm{sig}}$.

B.2. Limiting behavior

For the terms representing the effects of stochastic variation in the finite samples used, we note that $\|\widehat{\boldsymbol{p}}_u - \widetilde{\boldsymbol{p}}_u\|_\infty \to 0$ as $N \to \infty$—specifically

$$\lim_{N\to\infty} \mathbb{P}\left(\|\widehat{\boldsymbol{p}}_u - \widetilde{\boldsymbol{p}}_u\|_\infty > \mu\right) = 0 \quad \forall u \in \Sigma_X^L,\, \mu > 0 \tag{B.8}$$

due to the weak law of large numbers. Because d_m is continuous, it follows that $d_m(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u)$ must converge on $d_m(\widetilde{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u)$, which is zero—in other words,

$$|d_m(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u) - d_m(\widetilde{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u)| = d_m(\widehat{\boldsymbol{p}}_u,\, \widetilde{\boldsymbol{p}}_u) \to 0 \tag{B.9}$$

with probability one as $N \to \infty$. A similar argument can be applied to show that $d_m(\widetilde{\boldsymbol{q}}_V,\, \widehat{\boldsymbol{q}}_V)$ goes to zero in the limit as well. (Unlike $\widehat{\boldsymbol{p}}_u$, which is based on statistics from a single suffix, the working state V may contain many component suffixes, all of which influence $\widehat{\boldsymbol{q}}_V$. However, the probability of $\widehat{\boldsymbol{q}}_V$ failing to converge on $\widetilde{\boldsymbol{q}}_V$ is at most the sum of the probabilities of any component suffix failing to converge, which is a finite sum of zeros, and thus also evaluates to zero.)

Since $F_{\mathrm{sig}}(\alpha) > 2\widetilde{d}$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(d_m(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V) > F_{\mathrm{sig}}(\alpha)\right) = 0 \tag{B.10}$$

if u belongs in the precausal working state V of the unperturbed X_t process. Similarly, $F_{\mathrm{sig}}(\alpha) < d_{\mathrm{min}} - 2\widetilde{d}$ ensures

$$\lim_{N\to\infty} \mathbb{P}\left(d_m(\widehat{\boldsymbol{p}}_u,\, \widehat{\boldsymbol{q}}_V) \leq F_{\mathrm{sig}}(\alpha)\right) = 0 \tag{B.11}$$

in case u does not belong in V. As RCS and CSSR only perform a finite number of tests, each of which (assuming no earlier test made an error) has an error probability that approaches zero as above, and the error probability of the entire procedure is limited by the sum of all the individual test error-probabilities, the probability of any error in RCS also goes to zero in the limit $N \to \infty$. Therefore RCS with the current L, $n_{\max}$, and $\alpha \in I_{\text{sig}}$ applied to data from the distinguishable corruption Y_t converges in probability on the string clustering representing the precausal states of X_t.

Finally, for determinization we note that our assumptions ensure that only the suffixes in Σ_X^{L+1} ever occur in the data from Y_t. This implies that, not only are the precausal states identical to those of X_t, the nonzero-probability next-step symbols for each precausal state are the same, too, even when estimated from data (after some finite amount of samples has been amassed, with probability one). Since these are the only quantities relevant to determinization, and the determinization procedure is deterministic, the same suffix clustering as the causal states of X_t must result after determinization. This completes the argument.

C. Causal states of the flip process

In this appendix, we show that the noisy flip process has an infinite number of causal states, rendering it non-learnable using CSSR, by verifying that all parts of the criterion in Theorem 1 apply. The steps of the proof mirror the checks performed in Algorithm 1.

C.1. First parts of the theorem

The flip process can be described as a four-state stationary and ergodic HMM with parameter matrices

$$\boldsymbol{A} = \begin{bmatrix} 1-p_f & p_f & 0 & 0 \\ 0 & 0 & p_f & 1-p_f \\ 1-p_f & p_f & 0 & 0 \\ 0 & 0 & p_f & 1-p_f \end{bmatrix} \tag{C.1}$$

and

$$\boldsymbol{B} = \begin{bmatrix} 1 - {}^{\epsilon}\!/2 & {}^{\epsilon}\!/2 \\ {}^{\epsilon}\!/2 & 1 - {}^{\epsilon}\!/2 \\ 1 - {}^{\epsilon}\!/2 & {}^{\epsilon}\!/2 \\ {}^{\epsilon}\!/2 & 1 - {}^{\epsilon}\!/2 \end{bmatrix} . \tag{C.2}$$

We require that the flip probability satisfies $p_f \in (0,\, {}^1\!/2]$ ($p_f = 0$ is nonergodic), and that the substitution probability satisfies $\epsilon \in [0,\, 1]$. We shall see that for the interior of the parameter interval, corresponding to noisy flip processes, the number of causal states of the observed process is infinite.

It is easy to see that $\boldsymbol{A}^T$ has rank two, with right eigenvectors $\boldsymbol{g}_1 = [1 - p_f,\, p_f,\, p_f,\, 1 - p_f]^T$ (eigenvalue $\lambda_1 = 1$) and $\boldsymbol{g}_2 = [1 - p_f,\, p_f,\, -p_f,\, p_f - 1]^T$ (eigenvalue $\lambda_2 = 1 - 2p_f$). This gives

$$\boldsymbol{G}_r = \gamma \begin{bmatrix} 1 - p_f & 1 - p_f \\ p_f & p_f \\ p_f & -p_f \\ 1 - p_f & p_f - 1 \end{bmatrix} , \tag{C.3}$$

where $\gamma = \left(2 - 4p_f + 4p_f^2\right)^{-\frac{1}{2}}$, which is always greater than zero. A corresponding $\boldsymbol{H}_r$-matrix can be constructed from the left eigenvectors, as

$$\boldsymbol{H}_r = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} . \tag{C.4}$$

Some straightforward computations show that

$$\boldsymbol{B}^T \boldsymbol{A}^T \boldsymbol{G}_r = \gamma \begin{bmatrix} 1 & (1 - \epsilon)\,(1 - 2p_f)^2 \\ 1 & -(1 - \epsilon)\,(1 - 2p_f)^2 \end{bmatrix} . \tag{C.5}$$

This matrix has rank $r = 2$ whenever $\epsilon \in (0,\, 1)$ and $p_f \in (0,\, {}^1\!/2)$, that is, in the interior of the interval of parameter values considered. At the edges of the interval, where $(1 - \epsilon)\,(1 - 2p_f)^2$ is zero, the rank is one, and the theorem cannot be applied. These cases correspond either to an observation process Y_t which is i.i.d. ($p_f = {}^1\!/2$), and so has only one state, or to noise-free observations ($\epsilon = 0$), where we know there are exactly two causal states.

We henceforth consider only the interior of the interval of process parameters. Take $\mathcal{A}_{\text{sub}} = \mathcal{A}$, which is the only sensible choice for binary alphabets. Because all elements of $\boldsymbol{B}$ are strictly positive, point 1 of the criterion is always satisfied. Furthermore, it is easily verified that the forward matrices become

$$\boldsymbol{C}_1 = \frac{\gamma}{2} \begin{bmatrix} 1 & (1-\epsilon)(1-2p_f)^2 \\ 1-\epsilon & (1-2p_f)^2 \end{bmatrix} \tag{C.6}$$

and

$$\boldsymbol{C}_2 = \frac{\gamma}{2} \begin{bmatrix} 1 & (\epsilon-1)(1-2p_f)^2 \\ \epsilon-1 & (1-2p_f)^2 \end{bmatrix}. \tag{C.7}$$

These matrices have determinant $|\boldsymbol{C}_1| = |\boldsymbol{C}_2| = \frac{\gamma^2}{4}\epsilon(2-\epsilon)(1-2p_f)^2$, which is greater than zero for the parameter interval considered. Point 2 of the criterion in Theorem 1 is thus satisfied.

C.2. The final point of the theorem

To verify that the third and final point of Theorem 1 applies, we will look at the eigenvalues and eigenvectors of the forward matrices. First, we show that the eigenvalues of the forward matrices have distinct absolute values. The characteristic equation $|\boldsymbol{C}_1 - \lambda'\boldsymbol{I}| = 0$ yields a quadratic equation with a solution of the form

$$\lambda' = -\frac{a}{2} \pm \sqrt{\frac{a^2}{4} - b}. \tag{C.8}$$

In the present case,

$$a = \frac{\gamma}{2}\left(1 + (1-2p_f)^2\right), \tag{C.9}$$

while $\frac{a^2}{4} - b$ evaluates to

$$\frac{\gamma^2}{4}\left(\frac{1}{2} + \frac{1}{2}(1-2p_f)^4 + (1-\epsilon)^2(1-2p_f)^2\right) > 0. \tag{C.10}$$

The second formula shows that the eigenvalues of $\boldsymbol{C}_1$ are real, and (since the determinant additionally is positive) have distinct absolute values and the same sign. The eigenvalues of $\boldsymbol{C}_2$ are identical to those of $\boldsymbol{C}_1$, since the characteristic equation is the same.

As a second result, we establish that the forward matrices have no common eigenvectors. These eigenvectors are simply identified by solving the (singular) system $(\boldsymbol{C}_\sigma - \lambda' \boldsymbol{I})\, \boldsymbol{v} = \boldsymbol{0}$. We only need to consider the first row of the linear system in order to identify the ratio between the eigenvector elements v_1 and v_2, which uniquely determines the direction of the eigenvectors. This gives

$$\frac{v_1}{v_2} = -\gamma \frac{1-\epsilon}{\gamma - 2\lambda'} \left(1 - 2p_f\right)^2 \tag{C.11}$$

for eigenvectors of $\boldsymbol{C}_1$, and

$$\frac{v_1}{v_2} = \gamma \frac{1-\epsilon}{\gamma - 2\lambda'} \left(1 - 2p_f\right)^2 \tag{C.12}$$

for eigenvectors of $\boldsymbol{C}_2$. Trivially, then, eigenvectors of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ corresponding to the same eigenvalue cannot be collinear, since their v_1/v_2-ratios have opposite signs. For eigenvectors corresponding to different eigenvalues, these can only line up if

$$\gamma - 2\lambda_1' = -\left(\gamma - 2\lambda_2'\right). \tag{C.13}$$

Using (C.8), we see that this is equivalent to $\gamma + a = 0$. However,

$$\gamma + a = \frac{\gamma}{2}\left(3 + \left(1 - 2p_f\right)^2\right) > 0, \tag{C.14}$$

so eigenvectors from $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ corresponding to different eigenvalues cannot be collinear (have the same v_1/v_2-ratios) either.

The results above are sufficient to know that point three of Theorem 1 is satisfied. For any nonzero $\boldsymbol{v} \in \mathbb{C}^r$, we can choose some $\sigma \in \{1, 2\}$ such that this $\boldsymbol{v}$ is not an eigenvector of $\boldsymbol{C}_\sigma$, since there are no simultaneous eigenvectors. As $\boldsymbol{v}$ does not line up with any vector in the eigenbasis $\boldsymbol{Q}_\sigma$, $\boldsymbol{q} = \boldsymbol{Q}_\sigma^{-1}\boldsymbol{v}$ must have two nonzero elements. The associated eigenvalues λ_1' and λ_2' always satisfy $|\lambda_1'| \neq |\lambda_2'|$, since the eigenvalues of any $\boldsymbol{C}_\sigma$-matrix all have distinct absolute values.

In summary, we have established that all points of Theorem 1 are satisfied, meaning that the noisy flip process has an infinite number of causal states for $\epsilon \in (0,\, 1) \cap p_f \in (0,\, 1/2)$. We also note that the same computations can be used to show that the number of causal states is infinite for the parameter interval $\epsilon \in (0, 1) \cap p_f \in (1/2, 1)$ as well. The noisy flip process is thus not CSSR-learnable for these parameter values.

References

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. P. IEEE 77, 257–286.