Scalable calibration of individual-based epidemic models through categorical approximations

Lorenzo Rimella
Dipartimento di Scienze Economiche, Università degli studi di Bergamo, Bergamo, IT
Nick Whiteley
School of Mathematics, University of Bristol, Bristol, UK
Chris Jewell
School of Mathematical Sciences, Lancaster University, Lancaster, UK
Paul Fearnhead
School of Mathematical Sciences, Lancaster University, Lancaster, UK
Michael Whitehouse
School of Public Health, Imperial College London, London, UK

May 15, 2026

Abstract

Traditional compartmental models capture population-level dynamics but fail to characterize individual-level risk. The computational cost of exact likelihood evaluation for partially observed individual-based models, however, grows exponentially with the population size, necessitating approximate inference. Existing sampling-based methods usually require multiple simulations of the individuals in the population and rely on bespoke proposal distributions or summary statistics. We propose a deterministic approach to approximating the likelihood using categorical distributions. The approximate likelihood is amenable to automatic differentiation so that parameters can be estimated by maximization or posterior sampling using standard software libraries such as Stan or TensorFlow with little user effort. We prove the consistency of the maximum approximate likelihood estimator. We empirically test our approach on several classes of individual-based models for epidemiology: different sets of disease states, individual-specific transition rates, spatial interactions, under-reporting and misreporting. We demonstrate ground truth recovery and comparable marginal log-likelihood values at substantially reduced cost compared to competitor methods. Finally, we show the scalability and effectiveness of our approach with a real-world application on the 2001 UK Foot-and-Mouth outbreak, where the simplicity of the CAL allows us to include 162775 farms.

1 Introduction

The traditional approach to modeling epidemics involves dividing a population into compartments representing epidemic states through which individuals progress over time. The dynamics of such models are mathematically represented by, for instance, a Markov chain (Lekone and Finkenstädt, 2006), a system of ODEs (Chowell et al., 2004), or a system of SDEs (Ionides et al., 2006). These models treat the population as homogeneous in two senses: individuals are equal in their epidemiological attributes, such as susceptibility and infectivity (*homogeneous attributes*); and individuals contact each other at equal rates (*homogeneous-mixing*). These homogeneity assumptions can be relaxed by stratifying individuals into meta-populations (Whitehouse et al., 2023), where individuals have homogeneous attributes and are homogeneous-mixing within each meta-population but they have heterogeneous attributes and are heterogeneous-mixing across different meta-populations. While homogeneous/heterogeneous-mixing is a well-known terminology in the epidemiology literature (Jewell et al., 2009), homogeneous/heterogeneous attributes is somewhat uncommon but we use it to emphasize the different levels of heterogeneity we are considering.

Individual-based models (IBMs) fully relax the homogeneity assumptions by modeling each individual's disease states explicitly, instead of partitioning and aggregating them across different compartments. The interpretation of "individual" is context-specific and could refer, for example, to an individual person (Bu et al., 2025) or farm (Jewell et al., 2009). Whilst stratified compartmental models may handle covariates taking values in a discrete set, IBMs allow practitioners to work with both continuous and discrete covariates defined at an individual level, such as physical and physiological characteristics (age, health records, etc. (Cocker et al., 2023)), location (geographical and/or community membership (Jewell et al., 2009; Rimella et al., 2023)) or contact networks (Estrada et al., 2010, Chapter 4). The increasing availability of both epidemiological testing data and accompanying covariates at an individual level has motivated the development of many fine-scale IBMs of disease transmission (Cocker et al., 2023). However, the computational cost of performing exact inference for such IBMs with noisy and incomplete data necessarily grows astronomically in the population size N, with exact computation of likelihoods having complexity which is exponential in N (Rimella et al., 2025). Fast, simple, and theoretically justified calibration of IBMs is one of the open challenges in infectious disease modeling, motivating the present work.

We now summarize our novel contributions. **(1)** We propose a new form of approximate and recursive likelihood evaluation in partially observed individual-based epidemic models via categorical distributions, which involve no simulation from the model. **(2)** We prove strong consistency of parameter estimators obtained by maximizing our approximate likelihood when data are generated by the exact model. **(3)** The computational simplicity of our methodology allows it to scale up to large population sizes, even with a simple Python implementation. Moreover, it is particularly suited to automatic differentiation, for example gradient-based optimization or Hamiltonian Monte Carlo. **(4)** We calibrate an individual-based model to the 2001 Aphtovirus (Foot-and-Mouth) outbreak in the United Kingdom (UK), scaling up to include 162775 farms in the study.

The paper is organized as follows. We conclude this section with a motivating example and related works. In Section 2, we introduce our model, show how the motivating example can be formulated in that framework, and discuss closed-form likelihood computation. Section 3 presents

the main algorithm and explains the rationale behind the approximation. In Section 4, we state our consistency result, Theorem 1, and outline its proof. Numerical results on both synthetic and real data are reported in Section 5. Section 6 summarizes the paper and discusses limitations and future work.

1.1 Motivating example

Consider a discrete-time Susceptible-Infective-Susceptible (SIS) model for a population of size N. At each discrete time step $t = 0, 1, \ldots$, each individual assumes one of the disease states $\{S, I\}$. The following example could easily be extended to models with more than two disease states, we focus on SIS for ease of presentation. We present alternative building blocks of the model in separate paragraphs. Similar models incorporating individual-specific covariates have been considered in previous works, for example, by Ju et al. (2021); Bu et al. (2022); Rimella et al. (2023); Bu et al. (2025).

Heterogeneous attributes. IBMs allow individuals' covariates to be reflected in disease state transition probabilities. Let $\beta_{nk} > 0$ denote the rate at which the k-th individual infects the n-th when they come into contact, assuming the former is infective and the latter is susceptible. One may consider a regression model for the logarithm of the pairwise individual-specific transmission rate, e.g. $\log \beta_{nk} = \log \beta + \mathbf{c}_n^\top \mathbf{b}_S + \mathbf{c}_k^\top \mathbf{b}_I$, where β is the background infection rate, $\mathbf{c}_n$ is a vector of observed covariates associated with the n-th individual, and $\mathbf{b}_S, \mathbf{b}_I$ are parameter vectors respectively determining susceptibility to infection and propensity to pass the infection on. Similarly, one may consider a regression model for the recovery rate $\gamma_n > 0$ at which an infective individual recovers and returns to being susceptible: $\log \gamma_n = \log \gamma + \mathbf{c}_n^\top \mathbf{b}_R$, where $\mathbf{b}_R$ is a parameter vector.

Homogeneous- and heterogeneous-mixing dynamics. Under the assumption that the population mixes homogeneously in discrete time, the n-th individual is equally likely to contact any one individual, and the probability of the transition $S \to I$ at time $t > 0$ is:

$$1 - \exp\left(-\frac{h}{N} \sum_{k \in \mathcal{I}_{t-1}} \beta_{nk}\right) = 1 - \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\} \frac{1}{N} \sum_{k \in \mathcal{I}_{t-1}} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\}\right), \tag{1}$$

where $h > 0$ is the time period length (set equal to 1 unless stated otherwise) and $\mathcal{I}_{t-1}$ is the set of individuals who are infective at time $t - 1$. Here $\exp\{\mathbf{c}_n^\top \mathbf{b}_S\}$ has the interpretation of the susceptibility of the n-th individual, while $\exp\{\mathbf{c}_k^\top \mathbf{b}_I\}$ is the infectivity. At each time step, the probability of the n-th individual transitioning $I \to S$ is: $1 - \exp(-h\gamma_n)$.

In some cases, information may be available about the geographic location of individuals, network structure, or other factors that influence transmission rates between pairs of individuals. For example, if $\mathbf{z}_n$ denotes the Euclidean position of the n-th individual, then spatial weighting could

be introduced in the $S \to I$ transition probability (1) in the form:

$$1 - \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\} \frac{1}{N} \sum_{k \in \mathcal{I}_{t-1}} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{\|\mathbf{z}_n - \mathbf{z}_k\|^2}{2\phi^2} + \mathbf{c}_k^\top \mathbf{b}_I\right\}\right),$$

where $\phi > 0$ is a parameter and $\|\cdot\|$ is the Euclidean distance. Here, the term $\frac{1}{N}\frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{\|\mathbf{z}_n - \mathbf{z}_k\|^2}{2\phi^2}\right\}$ can be interpreted as the rate at which the n-th individual contacts the k-th individual, whilst the other terms are as in the homogeneous-mixing case. One could similarly model heterogeneity arising from a known network rather than spatial structure, by choosing the $S \to I$ transition probability for individual n to reflect its connectivity to other members of the population.

Observations. At each time step t, an individual is either reported in their true state, misreported in some other state, or not reported at all: in our SIS model, the observed state of each individual is therefore one of $\{U, S, I\}$, where U represents being unreported (e.g. missing test results). For any individual, let q_S (resp. q_I) be the probability of either correctly reporting or misreporting their state, given that they are susceptible (resp. infected). Conditional that the state of the individual is either reported or misreported, let q_{Se} (resp. q_{Sp}) be the probability of observing I if I (resp. S if S) is the true state. In the context of testing q_{Se} and q_{Sp} are the sensitivity and specificity.

Inference challenges. A typical inference task would be to calibrate the model by estimating the parameters, $\beta, \mathbf{b}_S, \mathbf{b}_I, \gamma, \mathbf{b}_R, \phi, q_S, q_I, q_{Se}, q_{Sp}$ or some subset thereof, allowing us to understand how the covariates $\mathbf{c}_n$ and spatial or network interactions contribute to the dynamics of the disease in question (Rimella et al., 2023; Seymour et al., 2022). Due to the partial observation structure of the above model, for population size and time horizon $N, T \in \mathbb{N}$ exact likelihood evaluation would involve marginalizing over 2^{NT} possible latent states.

1.2 Related work

The literature on IBMs is vast, and a full review would be impossible within the length constraints of the present work, here we present a brief summary. Inference for partially observed stochastic epidemic models is difficult, and, even when homogeneity assumptions are permitted, simplifications (King et al., 2015) or simulation-based procedures (Ionides et al., 2006) are required. Analogously, many studies of partially observed IBMs make simplifications or approximations, e.g. by deterministic modeling (Sharkey, 2008), mean-field approximations (Sherborne et al., 2018), or noiseless observation mechanisms (Deardon et al., 2010). Sophisticated simulation-based techniques have also been developed for inference in IBMs: bespoke proposals for sequential Monte Carlo (Rimella et al., 2023), approximate Bayesian computation procedures (McKinley et al., 2018), composite likelihood methods (Rimella et al., 2025), data augmentation schemes (Bu et al., 2022, 2025), neural posterior estimation (Chatha et al., 2024), and Markov chain Monte Carlo samplers (Touloupou et al., 2020); alongside Bayesian non-parametric approaches (Seymour et al., 2022) and kernel-linearization with

imputation based techniques (Deardon et al., 2010). Typically, when applied to noisy observations, it is hard to effectively scale the above methods to large population sizes.

The approximation techniques in the present work extend ideas for homogeneous population compartmental models (Whiteley and Rimella, 2021; Whitehouse et al., 2023; Whitehouse, 2025) to the case of individual-based models. This approach of making distributional approximations is related to assumed density filtering (Sorenson and Stubberud, 1968) and expectation propagation algorithms (Minka, 2001, Ch.1), but the details are different and specially designed to exploit the structure of the individual-based model. Furthermore, as far as the authors are aware, this is the first work to provide results on the consistency of parameter estimates for IBMs under analysis of the large population regime.

2 Individual-based compartmental model

2.1 Notation

Given $M \in \mathbb{N}$, we use $x_{0:M} \coloneqq x_0, \ldots, x_M$ for indexing sequences, $[M] \coloneqq \{1, \ldots, M\}$ for the set of the first M integers, and $\mathbf{x} \coloneqq \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\right]^\top$ for an M-dimensional vector. Given two M-dimensional vectors $\mathbf{x}_1, \mathbf{x}_2$, or $M \times M$-dimensional matrices, we denote with $\mathbf{x}_1 \odot \mathbf{x}_2$ the element-wise product and with $\mathbf{x}_1 \oslash \mathbf{x}_2$ the element-wise division, and we use $[\mathbf{x}_1, \mathbf{x}_2]$ for the vector stacking together $\mathbf{x}_1, \mathbf{x}_2$, i.e. $[\mathbf{x}_1, \mathbf{x}_2] \coloneqq \left[\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_1^{(M)}, \mathbf{x}_2^{(1)}, \ldots, \mathbf{x}_2^{(M)}\right]^\top$. We write 1_M for the M-dimensional vector of all ones, Δ_M for the M-dimensional probability simplex, i.e. $\Delta_M \coloneqq \left\{\mathbf{x} \in [0,1]^M : \sum_{i=1}^M \mathbf{x}^{(i)} = 1\right\}$, and $\mathbb{O}_M$ for the set of one-hot encoding vectors with dimension M, i.e. $\mathbb{O}_M \coloneqq \{\mathbf{x} \in \{0,1\}^M : \exists j \in [M] : \mathbf{x}^{(j)} = 1 \text{ and } \mathbf{x}^{(i)} = 0 \text{ if } i \neq j\}$, with $\mathbb{O}_M \subset \Delta_M$. Given $\boldsymbol{\pi} \in \Delta_M$ we denote with $\mathrm{Cat}(\cdot|\boldsymbol{\pi})$ the categorical distribution over $\mathbb{O}_M$ which assigns probability $\boldsymbol{\pi}^{(i)}$ to the vector $\mathbf{x} \in \mathbb{O}_M$ with $\mathbf{x}^{(i)} = 1$ and $\mathbf{x}^{(j)} = 0$ for $j \neq i$.

2.2 Model

We now introduce a generic form of individual-based model. We consider a population of $N \in \mathbb{N}$ individuals and assume that a vector of known covariates $\mathbf{w}_n \in \mathbb{W}$ is associated with individual $n \in [N]$, where $\mathbb{W}$ is a subset of Euclidean space. We denote by $\mathbf{W}$ the collection of covariates of the entire population, $\mathbf{W} \coloneqq (\mathbf{w}_1, \ldots, \mathbf{w}_N)$. These covariates allow us to express heterogeneity in how the disease propagates through the population. Each individual assumes any one of $M \in \mathbb{N}$ latent disease states at any one discrete time step $t \geq 0$. We use one-hot encoding vectors to represent the states of the individuals; this is a little non-standard but it will simplify mathematical expressions. The state of the n-th individual at time t is denoted $\mathbf{x}_{n,t} \in \mathbb{O}_M$, meaning that the i-th component of $\mathbf{x}_{n,t}$ is 1 if and only if the n individual is in state i at time t. The state of the entire population is written $\mathbf{X}_t \coloneqq (\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{N,t})$. With the covariates $\mathbf{W}$ fixed, the process $(\mathbf{X}_t)_{t \geq 0}$ is a Markov chain and the individual disease states $\mathbf{x}_{n,t}$ are distributed as follows.

Latent dynamics At time step $t = 0$, the state of each individual is drawn independently from an initial distribution which is a function of the individual-specific covariates $\mathbf{x}_{n,0}|\mathbf{w}_n \sim \text{Cat}\,(\,\cdot\,|p_0(\mathbf{w}_n))$, for some probability vector $p_0(\mathbf{w}_n)$. At time steps $t \geq 1$, conditional on the population state $\mathbf{X}_{t-1}$, the n-th individual evolves according to an $M \times M$ row-stochastic transition matrix $K_{\boldsymbol{\eta}_{n,t}}(\mathbf{w}_n)$:

$$\mathbf{x}_{n,t}|\mathbf{X}_{t-1}, \mathbf{W} \sim \text{Cat}\left(\,\cdot\,\middle|\left[\mathbf{x}_{n,t-1}^\top K_{\boldsymbol{\eta}_{n,t-1}}(\mathbf{w}_n)\right]^\top\right).$$

Here $\boldsymbol{\eta}_{n,t} \coloneqq \eta\,(\mathbf{w}_n, \mathbf{W}, \mathbf{X}_t)$ where $\eta : \mathbb{W} \times \mathbb{W}^N \times \Delta_M^N \to [0, C]$ with $C \in \mathbb{R}_+$ is a function which will allow us to express how individuals interact with the population in terms of their respective covariates and disease states. In this formulation, the transition matrix $K_{\boldsymbol{\eta}_{n,t}}(\mathbf{w}_n)$ depends on t only via $\mathbf{X}_t$. We consider this case for ease of presentation; our model, algorithm, and theory can be extended to transition matrices that evolve over time.

Observations At time $t \geq 1$, we observe a collection of vectors $\mathbf{Y}_t \coloneqq (\mathbf{y}_{1,t}, \ldots, \mathbf{y}_{N,t})$, where each $\mathbf{y}_{n,t}$ is a $\mathbb{O}_{M+1}$-valued random measurement associated with the n-th individual. Given $\mathbf{X}_t$, $\mathbf{y}_{1,t}, \ldots, \mathbf{y}_{N,t}$ are conditionally independent and distributed:

$$\mathbf{y}_{n,t}|\mathbf{x}_{n,t}, \mathbf{w}_n \sim \text{Cat}\left(\,\cdot\,\middle|\left[\mathbf{x}_{n,t}^\top G(\mathbf{w}_n)\right]^\top\right),$$

where $G(\mathbf{w}_n)$ is a $M \times (M+1)$-dimensional row-stochastic matrix. This matrix allows probabilities to be assigned to the n-th individual being: unreported, representing the extra compartment; correctly reported as in the disease state specified $\mathbf{x}_{n,t}$; erroneously reported as assuming one of the other $M-1$ disease states. The matrix $G(\mathbf{w}_n)$ could be allowed to depend on t with only notational changes to our algorithm and theory needed. We also assume the observations are evenly spaced in time, i.e. a period length of $h = 1$, but, once again, this is just for presentation purposes.

2.3 Motivating example

We now show how the motivating example from Section 1.1 can be cast as an instance of the generic model described in Section 2.2. As it is an SIS model, $M = 2$, $\mathbf{x}_{n,t} \in \{[1,0]^\top, [0,1]^\top\}$ is a 2-dimensional one-hot encoding vector representing disease states $\{S, I\}$. The individual-specific covariates are $\mathbf{w}_n = [\mathbf{c}_n, \mathbf{z}_n]$, where the latter are as in Section 1.1. As initial infection probabilities, we consider: $p_0(\mathbf{w}_n) = [1 - p_0,\ p_0]^\top$, i.e. each individual has the same probability of being infected at the beginning of the epidemic.

Homogeneous- and heterogeneous-mixing dynamics. Following the formulation from Section 2.2, we can express homogeneous- and heterogeneous-mixing dynamics by reformulating the interaction term $\boldsymbol{\eta}_{n,t-1}$. For the homogeneous-mixing case we can write $\boldsymbol{\eta}_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\}\mathbf{x}_{k,t-1}^{(2)}$, while for the heterogeneous-mixing case we have $\boldsymbol{\eta}_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\}\frac{\exp\left\{-\frac{\|\mathbf{z}_n - \mathbf{z}_k\|^2}{2\phi^2}\right\}}{\sqrt{2\pi\phi^2}}\mathbf{x}_{k,t-1}^{(2)}$. Then in either case we can write:

$$K_{\boldsymbol{\eta}_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} \exp\left(-h\beta\exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}_{n,t-1}\right) & 1 - \exp\left(-h\beta\exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}_{n,t-1}\right) \\ 1 - \exp\left(-h\gamma_n\right) & \exp\left(-h\gamma_n\right) \end{bmatrix}. \tag{2}$$

Observation model. The observation for each individual $\mathbf{y}_{n,t}$ is a 3-dimensional one-hot encoding vector representing the states: not reported (e.g. missing test results), reported as S, and reported as I. In this scenario, a stochastic matrix for our observation model is:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1 - q_S & q_S q_{Sp} & q_S(1 - q_{Sp}) \\ 1 - q_I & q_I(1 - q_{Se}) & q_I q_{Se} \end{bmatrix},$$

where $q_S, q_I, q_{Se}, q_{Sp} \in [0, 1]$ represent the reporting probabilities when S and when I, and the sensitivity and specificity of the test. More generally, these probabilities could be a function of covariates $\mathbf{w}_n$ and time-varying.

2.4 Exact likelihood

Under the definitions in Section 2.2, with the covariates $\mathbf{W}$ fixed, the joint process of population states $(\mathbf{X}_t)_{t\geq 0}$ and observations $(\mathbf{Y}_t)_{t\geq 1}$ is a Hidden Markov Model (HMM) (Chopin and Papaspiliopoulos, 2020). Over a time horizon T, the marginal likelihood of $\mathbf{Y}_{1:T}$ is:

$$p(\mathbf{Y}_{1:T}|\mathbf{W}) = \sum_{\mathbf{X}_{0:T} \in \mathbb{O}_M^{NT}} p(\mathbf{X}_0|\mathbf{W}) \prod_{t=1}^{T} p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{W}) p(\mathbf{Y}_t|\mathbf{X}_t, \mathbf{W}),$$

where, using some HMM terminology, the initial distribution is $p(\mathbf{X}_0|\mathbf{W}) \coloneqq \prod_{n\in[N]} \mathbf{x}_{n,0}^\top p_0(\mathbf{w}_n)$, the transition kernel is $p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{W}) \coloneqq \prod_{n\in[N]} \mathbf{x}_{n,t-1}^\top K_{\boldsymbol{\eta}_{n,t-1}}(\mathbf{w}_n)\mathbf{x}_{n,t}$, and the emission distribution is $p(\mathbf{Y}_t|\mathbf{X}_t, \mathbf{W}) \coloneqq \prod_{n\in[N]} \mathbf{x}_{n,t}^\top G(\mathbf{w}_n)\mathbf{y}_{n,t}$. The computation of the marginal likelihood requires a summation over the set $\mathbb{O}_M^{NT}$. The forward algorithm (Chopin and Papaspiliopoulos, 2020) computes the sum at a cost linear in T, by recursively computing the prediction distributions $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, \mathbf{W})$, the filtering distributions $p(\mathbf{X}_t|\mathbf{Y}_{1:t}, \mathbf{W})$, and the marginal likelihood increments $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, \mathbf{W})$, via the so-called "prediction" and "correction" steps of the filtering recursion. Indeed, given $p(\mathbf{X}_0|\mathbf{Y}_{1:0}, \mathbf{W}) \coloneqq p(\mathbf{X}_0|\mathbf{W})$:

$$\textbf{Prediction: } p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, \mathbf{W}) \coloneqq \sum_{\mathbf{X}_{t-1} \in \mathbb{O}_M^N} p(\mathbf{X}_t|\mathbf{X}_{t-1}, \mathbf{W}) p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1}, \mathbf{W});$$

$$\textbf{Correction: } p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, \mathbf{W}) \coloneqq \sum_{\mathbf{X}_t \in \mathbb{O}_M^N} p(\mathbf{Y}_t|\mathbf{X}_t, \mathbf{W}) p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, \mathbf{W}) \text{ and}$$

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}, \mathbf{W}) \coloneqq \frac{p(\mathbf{Y}_t|\mathbf{X}_t, \mathbf{W}) p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, \mathbf{W})}{p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, \mathbf{W})};$$

from which get $p(\mathbf{Y}_{1:T}|\mathbf{W}) = p(\mathbf{Y}_1|\mathbf{W}) \prod_{t=2}^{T} p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, \mathbf{W})$. Whilst the forward algorithm simplifies the likelihood computation with respect to time, it still requires summation over the set $\mathbb{O}_M^N$, making it infeasible for even small population sizes.

3 CAL: Categorical Approximate Likelihood

Given the state at time $t-1$, the transitions and observations of each individual at time t are independent of one another. The challenge arises from uncertainty in $\mathbf{X}_{t-1}$, which introduces dependence. To address this, we substitute $\mathbf{X}_{t-1}$ with its expectation during the prediction step and propagate the resulting approximation of $p(\mathbf{X}_t \mid \mathbf{Y}_{1:t-1}, \mathbf{W})$ through the correction step. This leads to approximations of the predictive distributions, filtering distributions, and marginal likelihood as products of categorical distributions.

3.1 Approximate filtering algorithm

In this section we propose the approximations $p(\mathbf{X}_t|\mathbf{Y}_{1:t-1}, \mathbf{W}) \approx \prod_{n\in[N]} \mathrm{Cat}(\mathbf{x}_{n,t} \mid \boldsymbol{\pi}_{n,t|t-1})$, and $p(\mathbf{Y}_t|\mathbf{Y}_{1:t-1}, \mathbf{W}) \approx \prod_{n\in[N]} \mathrm{Cat}(\mathbf{y}_{n,t} \mid \boldsymbol{\mu}_{n,t})$, and $p(\mathbf{X}_t|\mathbf{Y}_{1:t}, \mathbf{W}) \approx \prod_{n\in[N]} \mathrm{Cat}(\mathbf{x}_{n,t} \mid \boldsymbol{\pi}_{n,t})$. We next explain how the probability vectors $\boldsymbol{\pi}_{n,t|t-1}, \boldsymbol{\pi}_{n,t} \in \Delta_M$ and $\boldsymbol{\mu}_{n,t} \in \Delta_{M+1}$ are computed for $n = 1, \ldots, N$ and $t \geq 1$.

Starting from $p(\mathbf{X}_0 \mid \mathbf{Y}_{1:0}, \mathbf{W})$, we set, for each $n \in [N]$, $\boldsymbol{\pi}_{n,0}$ to be the length-M probability vector associated with the initial distribution $p_0(\mathbf{w}_n)$. Hence, at the initial time, no approximation is required, as the distribution of $\mathbf{X}_0$ already factorizes as a product of categorical distributions. Now consider a general time $t \geq 1$. Suppose we have computed $\boldsymbol{\Pi}_{t-1} \coloneqq (\boldsymbol{\pi}_{1,t-1}, \ldots, \boldsymbol{\pi}_{N,t-1})$, so that $p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1}\mathbf{W}) \approx \prod_{n\in[N]} \mathrm{Cat}(\mathbf{x}_{n,t-1} \mid \boldsymbol{\pi}_{n,t-1})$. In the **prediction** step, we approximate the transition probability $p(\mathbf{X}_t \mid \mathbf{X}_{t-1}, \mathbf{W})$ by substituting $\boldsymbol{\Pi}_{t-1}$ in place of $\mathbf{X}_{t-1}$ in the quantity $\boldsymbol{\eta}_{n,t-1} = \eta(\mathbf{w}_n, \mathbf{W}, \mathbf{X}_{t-1})$. Under the categorical approximation, $\boldsymbol{\Pi}_{t-1}$ coincides with the expectation of $\mathbf{X}_{t-1}$. After this substitution, the **prediction** step admits a closed-form expression. Precisely:

$$\boldsymbol{\pi}_{n,t|t-1} \coloneqq \left[\boldsymbol{\pi}_{n,t-1}^\top K_{\widetilde{\boldsymbol{\eta}}_{n,t-1}}(\mathbf{w}_n)\right]^\top, \quad \text{for } n \in [N],$$

where $\widetilde{\boldsymbol{\eta}}_{n,t-1} \coloneqq \eta(\mathbf{w}_n, \mathbf{W}, \boldsymbol{\Pi}_{t-1})$. The **correction** step is then applied to the categorical approximation $\bigotimes_{n\in[N]} \mathrm{Cat}(\cdot \mid \boldsymbol{\pi}_{n,t|t-1})$ of $p(\mathbf{X}_t \mid \mathbf{Y}_{1:t-1}, \mathbf{W})$, using the exact observation model $p(\mathbf{Y}_t \mid \mathbf{X}_t, \mathbf{W})$. This model factorizes across individuals under the conditional independence structure specified in Section 2.2, giving us:

$$\boldsymbol{\mu}_{n,t} \coloneqq \left[\boldsymbol{\pi}_{n,t|t-1}^\top G(\mathbf{w}_n)\right]^\top \text{ and } \boldsymbol{\pi}_{n,t} \coloneqq \boldsymbol{\pi}_{n,t|t-1} \odot \left\{\left[G(\mathbf{w}_n) \oslash \left(1_M \boldsymbol{\mu}_{n,t}^\top\right)\right] \mathbf{y}_{n,t}\right\} \quad \text{for } n \in [N],$$

where we use the convention $\frac{0}{0} = 0$. By sequentially combining these approximate prediction and correction steps we get Algorithm 1, which computes all the aforementioned quantities. Algorithm 1 could output all the categorical approximations, but for the sake of presentation we make it output only the Categorical Approximate Likelihood (CAL):

$$p(\mathbf{Y}_{1:T}|\mathbf{W}) \approx \prod_{t=1}^{T} \prod_{n\in[N]} \mathrm{Cat}(\mathbf{y}_{n,t} \mid \boldsymbol{\mu}_{n,t}) = \prod_{t=1}^{T} \prod_{n\in[N]} \mathbf{y}_{n,t}^\top \boldsymbol{\mu}_{n,t}.$$

We refer the reader to Section B of the supplementary material for complete derivations.

Algorithm 1 Categorical Approximate Likelihood

Require: $\mathbf{W}, \mathbf{Y}_{1:T}, p_0(\cdot), K_\cdot(\cdot), G(\cdot)$
  Initialize $\boldsymbol{\pi}_{n,0}$ with $p_0(\mathbf{w}_n)$ for all $n \in [N]$
  for $t \in 1, \ldots, T$ **do**
    $\boldsymbol{\Pi}_{t-1} = (\boldsymbol{\pi}_{1,t-1}, \ldots, \boldsymbol{\pi}_{N,t-1})$
    for $n \in [N]$ **do**
      $\widetilde{\boldsymbol{\eta}}_{n,t-1} = \eta(\mathbf{w}_n, \mathbf{W}, \boldsymbol{\Pi}_{t-1})$
      $\boldsymbol{\pi}_{n,t|t-1} = \left[\boldsymbol{\pi}_{n,t-1}^\top K_{\widetilde{\boldsymbol{\eta}}_{n,t-1}}(\mathbf{w}_n)\right]^\top$
      $\boldsymbol{\mu}_{n,t} = \left[\boldsymbol{\pi}_{n,t|t-1}^\top G(\mathbf{w}_n)\right]^\top$
      $\boldsymbol{\pi}_{n,t} = \boldsymbol{\pi}_{n,t|t-1} \odot \left\{\left[G(\mathbf{w}_n) \oslash \left(1_M \boldsymbol{\mu}_{n,t}^\top\right)\right] \mathbf{y}_{n,t}\right\}$
    end for
  end for
  Return the approximate likelihood $\prod_{t=1}^T \prod_{n\in[N]} \mathbf{y}_{n,t}^\top \boldsymbol{\mu}_{n,t}$

3.2 CAL as an exact likelihood in an approximate model

Although the CAL is derived as an approximation to the marginal likelihood for the model in Section 2.2, it can also be interpreted as the exact marginal likelihood for the approximate model where $\widetilde{\boldsymbol{\eta}}_{n,t}$ is used instead of $\boldsymbol{\eta}_{n,t}$. Here, as in Section 3.1, $\widetilde{\boldsymbol{\eta}}_{n,t-1} = \eta(\mathbf{w}_n, \mathbf{W}, \boldsymbol{\Pi}_{t-1})$, and $\boldsymbol{\Pi}_{t-1}$ is computed as in Algorithm 1. Indeed, we can define the state process $\widetilde{\mathbf{X}}_t = (\widetilde{\mathbf{x}}_{1,t}, \ldots, \widetilde{\mathbf{x}}_{N,t})$ and the observation process $\widetilde{\mathbf{Y}}_t = (\widetilde{\mathbf{y}}_{1,t}, \ldots, \widetilde{\mathbf{y}}_{N,t})$ of the approximate model as follows:

$$\widetilde{\mathbf{x}}_{n,0}|\mathbf{w}_n \sim \mathrm{Cat}\left(\cdot|p_0(\mathbf{w}_n)\right), \quad \widetilde{\mathbf{x}}_{n,t}|\widetilde{\mathbf{X}}_{t-1}, \widetilde{\mathbf{Y}}_{1:t-1}, \mathbf{W} \sim \mathrm{Cat}\left(\cdot|\left[\widetilde{\mathbf{x}}_{n,t-1}^\top K_{\widetilde{\boldsymbol{\eta}}_{n,t-1}}(\mathbf{w}_n)\right]^\top\right),$$
$$\widetilde{\mathbf{y}}_{n,t}|\widetilde{\mathbf{x}}_{n,t}, \mathbf{w}_n \sim \mathrm{Cat}\left(\cdot|\left[\widetilde{\mathbf{x}}_{n,t}^\top G(\mathbf{w}_n)\right]^\top\right).$$

Under the approximate model above, the marginal likelihood of $\widetilde{\mathbf{Y}}_{1:T}$ can be computed in closed-form using Algorithm 1. Indeed, we have $p(\widetilde{\mathbf{Y}}_{1:T}|\mathbf{W}) = \prod_{n\in[N]} \prod_{t=1}^T \widetilde{\mathbf{y}}_{n,t}^\top \boldsymbol{\mu}_{n,t}$, which coincides with the CAL when the observations are $\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_T$.

4 Consistency of the maximum CAL estimator

In this section, we state our results about consistency of the maximum CAL estimator of model parameters, when data are generated from the exact model from Section 2.2. Proofs and supporting results are given in Section C of the supplementary materials. This theory has links to mean-field approximations (Sherborne et al., 2018) and propagation of chaos (Sharrock et al., 2023; Le Boudec et al., 2007) as it relies on the construction of what we call a "saturated" system of independently evolving individuals.

4.1 Notation, definitions and assumptions

We denote with Θ the parameter space, with $\mathbb{W}$ the covariate space. We assume all the random variables appearing in our theory to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We augment our notation from Section 2.2 by writing $\mathbf{W}^N, \mathbf{x}_{n,t}^N, \mathbf{X}_t^N, \mathbf{y}_{n,t}^N, \mathbf{Y}_t^N$ for $\mathbf{W}, \mathbf{x}_{n,t}, \mathbf{X}_t, \mathbf{y}_{n,t}, \mathbf{Y}_t$; given a parameter vector $\theta \in \Theta$, the initial distribution, transition matrix, and emission matrix are denoted: $p_0(\mathbf{w}_n, \theta)$, $K_\cdot(\mathbf{w}_n, \theta)$, and $G(\mathbf{w}_n, \theta)$, with $\eta(\mathbf{w}_n, \mathbf{W}, \mathbf{X})$ becoming $\eta^N(\mathbf{w}_n, \theta, \mathbf{W}^N, \mathbf{X}^N)$, and $\boldsymbol{\eta}_{n,t}$ becoming $\boldsymbol{\eta}_t^N(\mathbf{w}_n, \theta)$. Similarly, the CAL quantities $\boldsymbol{\pi}_{n,t|t-1}, \boldsymbol{\mu}_{n,t}, \boldsymbol{\pi}_{n,t}, \widetilde{\boldsymbol{\eta}}_{n,t}$ in Algorithm 1 become $\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta), \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)$, $\boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta), \widetilde{\boldsymbol{\eta}}_t^N(\mathbf{w}_n, \theta)$. We denote by $\theta^\star \in \Theta$ the data-generating parameter (DGP) value, which determines the distributions of $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Y}_t)_{t\geq 1}$ under $\mathbb{P}$ conditional on $\mathbf{W}$. For an M-dimensional vector π, an $M \times M$-dimensional matrix K, and an $\mathbb{R}$-valued random variable $\mathbf{x}$ we define the following norms:

$$\|\pi\|_\infty \coloneqq \max_{i\in[M]} \left|\pi^{(i)}\right|, \quad \|K\|_\infty \coloneqq \max_{i\in[M]} \sum_{j=1}^{M} \left|K^{(i,j)}\right|, \quad \text{and} \quad \|\mathbf{x}\|_4 \coloneqq \left(\mathbb{E}\left[|\mathbf{x}|^4\right]\right)^{\frac{1}{4}}.$$

We refer to the non-zero elements of probability vectors and matrices as the support, which we define as $\mathbf{supp}(\boldsymbol{\pi}) \coloneqq \{i \in [M] : \boldsymbol{\pi}^{(i)} \neq 0\}$ for an M-dimensional vector $\boldsymbol{\pi}$, and $\mathbf{supp}(\boldsymbol{\Pi}) \coloneqq \{(i,j) \in [M]^2 : \boldsymbol{\Pi}^{(i,j)} \neq 0\}$ for a $M \times M$ matrix $\boldsymbol{\Pi}$.

We next state our assumptions, which we first comment on and then list. Assumption 1 collects standard compactness and continuity assumptions. Assumption 2 is a random design assumption on the covariate vectors, which is inspired by classical theoretical analysis of regression models. Assumption 3 represents a technical assumption on the support of initial distribution, transition matrix, and emission matrix, which guarantees the invariance of the support when considering different parameters and covariates. We conclude with Assumption 4 on the structure of η, which ensures a law of large numbers for this interaction term, and Assumption 5 about the Lipschitz continuity of the transition matrix in η.

Assumption 1. *The parameter space Θ and the covariate space $\mathbb{W}$ are compact subsets of Euclidean spaces. Moreover, the initial distribution $p_0(w, \theta)$, the transition matrix $K_\eta(w, \theta)$, and the emission matrix $G(w, \theta)$ are all continuous functions in their arguments w, θ.*

Assumption 2. *The covariates $\mathbf{w}_1, \mathbf{w}_2, \ldots,$ are independent and identically distributed according to a distribution Γ on $\mathbb{W}$.*

Assumption 3. *The following hold: for any $w \in \mathbb{W}$ and $\theta, \theta' \in \Theta$ we have that $\boldsymbol{supp}(p_0(w, \theta)) = \boldsymbol{supp}(p_0(w, \theta'))$; for any $w \in \mathbb{W}$, $\eta, \eta' \in [0, C]$ and $\theta, \theta' \in \Theta$ we have that $\boldsymbol{supp}(K_\eta(w, \theta)) = \boldsymbol{supp}(K_{\eta'}(w, \theta'))$; for any $w \in \mathbb{W}$ and $\theta, \theta' \in \Theta$ we have that $\boldsymbol{supp}(G(w, \theta)) = \boldsymbol{supp}(G(w, \theta'))$.*

Assumption 4. *For any $\theta \in \Theta, w \in \mathbb{W}, N \in \mathbb{N}$, and for any $W^N = (w_1, \ldots, w_N), \Pi^N = (\pi_1, \ldots, \pi_N)$ with $w_n \in \mathbb{W}, \pi_n \in \Delta_M$ for all $n \in [N]$, we have:*

$$\eta^N(w, \theta, W^N, \Pi^N) = \frac{1}{N} \sum_{n\in[N]} d(w, w_n, \theta)^\top \pi_n,$$

where $d : \mathbb{W}\times\mathbb{W}\times\Theta \to [0,C]^M$ *is a bounded function, i.e.* $\|d(w,\widetilde{w},\theta)\|_\infty \leq C < \infty$ *for all* $w,\widetilde{w}\in\mathbb{W}$ *and* $\theta\in\Theta$*. This assumption also implies* $\eta^N(w,\theta,W^N,\Pi^N)\in[0,C]$ *for any* N,w,π,W^N,Π^N*.*

Assumption 5. *For any* $\theta\in\Theta$ *and* $w\in\mathbb{W}$*, the matrix* $K_\eta(w,\theta)$ *is Lipschitz continuous in* η *with Lipschitz constant* L*, that is for any* $\eta,\eta'\in[0,C]$ *we have:*

$$\|K_\eta(w,\theta) - K_{\eta'}(w,\theta)\|_\infty \leq L\,|\eta-\eta'|\,.$$

4.2 Main consistency theorem and outline of the proof

Consider a fixed time horizon $T\geq 1$, and the log-CAL evaluated at $\theta\in\Theta$:

$$\ell^N_{1:T}(\theta) := \sum_{t=1}^{T}\sum_{n\in[N]} \log\left[\left(\mathbf{y}^N_{n,t}\right)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right].$$

This section outlines the proof that the maximum CAL estimator $\hat{\theta}^N := \operatorname{argmax}_{\theta\in\Theta}\ell^N_{1:T}(\theta)$ is consistent in the large population limit. All the details are available in Section C of the supplementary material. The main challenge is to prove that $N^{-1}(\ell^N_{1:T}(\theta)-\ell^N_{1:T}(\theta^\star))$ converges uniformly $\mathbb{P}$-almost surely to a contrast function which is maximised by $\theta^\star$. This then allows standard continuity arguments to be used in proving almost sure convergence of the maximizer $\hat{\theta}^N$ to some equivalence set containing $\theta^\star$. Due to the presence of covariates, the details of the analysis are substantially richer than those of Whitehouse et al. (2023).

Saturated process and saturated CAL. From Section 2.2 it is clear that all the individuals are interacting via the interaction term $\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)$ in the transition matrix. We can prove that under Assumptions 2,4,5 for any $t\geq 0$ there exists a deterministic function $w\mapsto\boldsymbol{\eta}^\infty_t(w,\theta^\star)$ from $\mathbb{W}$ to $[0,C]$ such that for any $n\in[N]$:

$$\left\|\boldsymbol{\eta}^N_t(\mathbf{w}_n,\theta^\star) - \boldsymbol{\eta}^\infty_t(\mathbf{w}_n,\theta^\star)\right\|_4 = \mathcal{O}\left(N^{-\frac{1}{2}}\right), \tag{3}$$

see Section C.2 of the supplementary material for details. From (3) we observe that when the system becomes "saturated" with individuals, i.e. $N\to\infty$, the effect from the population has a deterministic behavior. Substituting $\boldsymbol{\eta}^\infty_t(\cdot,\theta^\star)$ in our latent dynamic defines a saturated process at the individual level. Specifically, for an individual with covariate $\mathbf{w}^\infty\sim\Gamma$ the saturated process is:

$$\begin{aligned}
&\mathbf{x}^\infty_0|\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot\,|\,p_0(\mathbf{w}^\infty,\theta^\star)\right),\\
&\mathbf{x}^\infty_t|\mathbf{x}^\infty_{t-1},\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot\,\middle|\,\left[(\mathbf{x}^\infty_{t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}^\infty,\theta^\star)}(\mathbf{w}^\infty,\theta^\star)\right]^\top\right),\\
&\mathbf{y}^\infty_t|\mathbf{x}^\infty_t,\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot\,\middle|\,\left[(\mathbf{x}^\infty_t)^\top G(\mathbf{w}^\infty,\theta^\star)\right]^\top\right).
\end{aligned} \tag{4}$$

We can observe that in the saturated process the individuals evolve independently, providing an asymptotic justification for the CAL prediction step.

Similarly, we can prove that under Assumptions 2,4,5 for any $t \geq 0$ there exists a deterministic function $w \mapsto \bar{\boldsymbol{\eta}}_t^{\infty}(w, \theta)$ from $\mathbb{W}$ to $[0, C]$, which is such that for any $w \in \mathbb{W}$ we have $\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(w, \theta^\star) = \boldsymbol{\eta}_{t-1}^{\infty}(w, \theta^\star)$ and for any $n \in [N]$:

$$\left\| \widetilde{\boldsymbol{\eta}}_t^N(\mathbf{w}_n, \theta) - \bar{\boldsymbol{\eta}}_t^{\infty}(\mathbf{w}_n, \theta) \right\|_4 = \mathcal{O}\left(N^{-\frac{1}{2}}\right),$$

see Section C.3 of the supplementary material. We can then define the saturated CAL recursion by substituting individual saturated process observations from (4) and $\bar{\boldsymbol{\eta}}_t^{\infty}(\mathbf{w}_n, \theta)$ into Algorithm 1. Precisely, set $\boldsymbol{\pi}_0^{\infty}(\mathbf{w}^{\infty}, \theta) \coloneqq p_0(\mathbf{w}^{\infty}, \theta)$ and then for $t \geq 1$:

$$\begin{aligned}
\boldsymbol{\pi}_{t|t-1}^{\infty}(\mathbf{w}^{\infty}, \theta) &\coloneqq \left[\boldsymbol{\pi}_{t-1}^{\infty}(\mathbf{w}^{\infty}, \theta)^{\top} K_{\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}^{\infty}, \theta)}(\mathbf{w}^{\infty}, \theta)\right]^{\top}, \\
\boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta) &\coloneqq \left[\boldsymbol{\pi}_{t|t-1}^{\infty}(\mathbf{w}^{\infty}, \theta)^{\top} G(\mathbf{w}^{\infty}, \theta)\right]^{\top}, \\
\boldsymbol{\pi}_t^{\infty}(\mathbf{w}^{\infty}, \theta) &\coloneqq \boldsymbol{\pi}_{t|t-1}^{\infty}(\mathbf{w}^{\infty}, \theta) \odot \left\{\left[G(\mathbf{w}^{\infty}, \theta) \oslash \left(1_M \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)^{\top}\right)\right] \mathbf{y}_t^{\infty}\right\}.
\end{aligned} \tag{5}$$

As, conditional on $\mathbf{w}^{\infty}$, the joint process $(\mathbf{x}_t^{\infty})_{t\geq 0}$, $(\mathbf{y}_t^{\infty})_{t\geq 1}$ in (4) is a HMM, Recursion (5) becomes the forward algorithm associated with this HMM when $\theta = \theta^\star$, i.e. at the DGP.

Contrast function and set of maximizers. Under Assumptions 1,2,3,4,5, for any $\theta \in \Theta$ we prove that $N^{-1}(\ell_{1:T}^N(\theta) - \ell_{1:T}^N(\theta^\star))$ converges $\mathbb{P}$-almost surely to a contrast function $\mathcal{C}_T(\theta, \theta^\star)$ as $N \to \infty$, which takes the form of an expected Kullback-Leibler (KL) divergence:

$$\mathcal{C}_T(\theta, \theta^\star) \coloneqq -\sum_{t=1}^{T} \mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta^\star)\right) || \mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right]\right\}.$$

Moreover, we use properties of the KL divergence to show the DGP belongs to the set of maximizers of the contrast function, i.e. $\theta^\star \in \Theta^\star \coloneqq \mathrm{argmax}_{\theta\in\Theta} \mathcal{C}_T(\theta, \theta^\star)$. Full proof is available in Section C.5 and Section C.4 of the supplementary material.

Convergence of the maximum CAL estimator and identifiability. After proving some technical results, we can complete the proof of Theorem 1, which states the consistency of the maximum CAL estimator, see Section C.5 of the supplementary material.

Theorem 1. *Let Assumptions 1,2,3,4,5 hold and let $\hat{\theta}_N$ be a maximizer of $\ell_{1:T}^N(\theta)$. Then $\hat{\theta}_N$ converges to $\Theta^\star$ as $N \to \infty$, $\mathbb{P}$-almost surely.*

The theorem states that the maximum CAL estimator converges to a set of maximizers $\Theta^\star$, which is a set of parameters that define statistically indistinguishable one-individual saturated processes. More formally, denote with $\mathbb{P}_{\infty}^{\theta^\star, w}$ the law of $(\mathbf{y}_t^{\infty})_{t\geq 1}$ conditional on $\mathbf{w}^{\infty} = w$ and with DGP $\theta^\star$. We can show that for any $\theta_1^\star, \theta_2^\star \in \Theta^\star$ we have $\mathbb{P}_{\infty}^{\theta_1^\star, w} = \mathbb{P}_{\infty}^{\theta_2^\star, w}$ for Γ-almost all $w \in \mathbb{W}$, see Section C.5 of the supplementary material.

5 Experiments

We now implement the CAL over a range of IBMs with heterogeneous attributes and different heterogeneous-mixing behaviors. The results can be reproduced following the GitHub repository LorenzoRimella/CAL. All the experiments were run on a 32GB Tesla V100 GPU available on "The High-End Computing" (HEC) facility at Lancaster University.

Computational considerations The CAL is "embarrassingly parallel" in N at each time step, making it well-suited for parallel architectures such as GPUs. Furthermore, the CAL does not rely on simulations from the model or permutations of indices, which makes it particularly amenable to just-in-time (JIT) compilation (Aycock, 2003), enabling efficient execution without complex code design. Because of its simplicity, the gradient of the log-CAL with respect to the model parameters can be computed via automatic differentiation (AD). As a result, popular AD libraries (Abadi et al., 2015) can be used for efficient optimization, or the CAL can be embedded within a Hamiltonian Monte Carlo (HMC) sampler for Bayesian inference via probabilistic programming languages (**?**). Both approaches are used and combined in our experiments, with the former implemented using the Adam optimizer (Kingma and Ba, 2014). More computational considerations can be found in Section D.1 of the supplementary materials.

5.1 CAL-posterior inference using HMC in TensorFlow

We demonstrate Bayesian inference using an HMC sampler in TensorFlow (Abadi et al., 2015) to target a posterior distribution defined in terms of the CAL. The appeal of this approach is that once the model is formulated as in Section 2, evaluating the CAL involves no tuning parameters and can be readily embedded within an "off the shelf" HMC program.

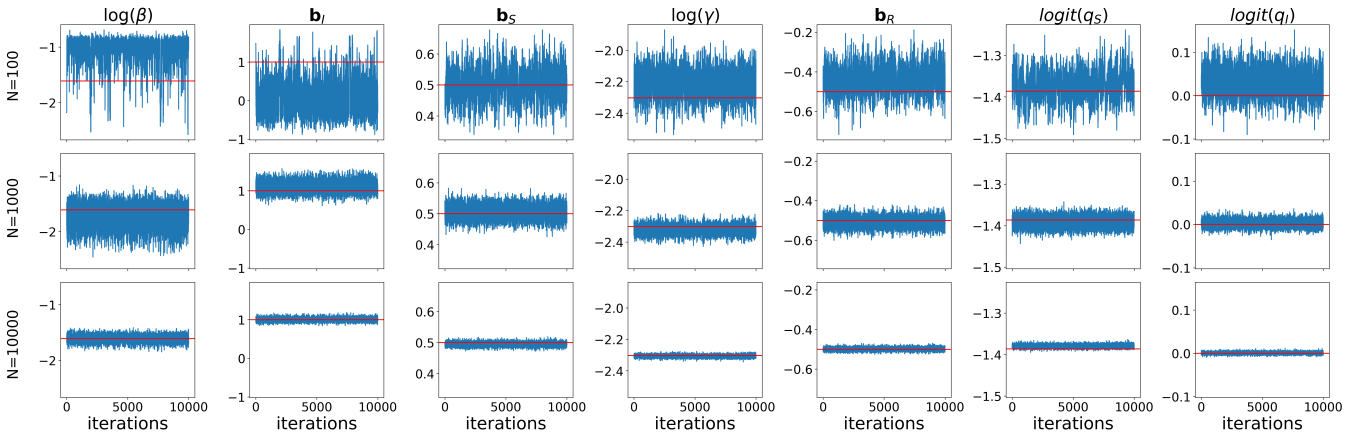

Figure 1: Trace plots for HMC under different population sizes. Solid red lines denote the DGP.

Consider the homogeneous-mixing SIS model from Section 1.1 and Section 2.3 with synthetic covariates $\mathbf{c}_n \sim \mathcal{N}(\cdot|0,1)$. We simulate data from models with increasing population sizes $N = 100, 1000, 10000$ and time horizon $T = 200$. The full parameter settings can be found in Section D.2

of the supplementary materials. The chains show no signs of poor mixing and recovery of the DGP, with a posterior distribution that becomes increasingly concentrated as the population size grows, complementing our consistency theory. The running time for the experiment with $N = 10000$ was around 0.5s for each iteration, see Section D.2 of the supplementary material for full details of the HMC scheme.

We further assess the coverage of the credible intervals by considering $N = 1000$ and rerunning the experiment 100 times. Specifically, we simulate 100 epidemics and replicate the previous HMC procedure for each of them. This results in 95% marginal credible interval coverages of 0.83 for $\log(\beta)$, 0.90 for $\mathbf{b}_S$, 0.80 for $\mathbf{b}_I$, 0.91 for $\mathbf{b}_R$, 0.93 for $logit(q_S)$, and 0.95 for $logit(q_I)$. Overall, the coverages are close to the target value of 0.95, except for $\log(\beta)$ and $\mathbf{b}_I$, for which uncertainty is underestimated.

5.2 Gradient-based calibration for heterogeneous-mixing SIS

We consider two heterogeneous-mixing IBMs: one with a continuous spatial interpretation, and the other with a network interpretation. We demonstrate recovery of the DGP by optimizing the CAL with Adam (Kingma and Ba, 2014) and with an accuracy that increases in N.

Model 1. Consider the heterogeneous-mixing SIS model from Section 1.1 and Section 2.3. Precisely, we have initial infection probabilities and $\boldsymbol{\eta}_{n,t-1}$ as in Section 2.3, and a transition matrix as in (2), where we also include the term ϵ to represent a constant rate of infection from the environment. The covariates $\mathbf{w}_n = [\mathbf{z}_n, \mathbf{c}_n]$ are synthetic, such that $\mathbf{z}_n$ is the location in space of each individual and $\mathbf{c}_n \sim \mathcal{N}(\cdot|0, 1)$. The location $\mathbf{z}_n$ is drawn from a mixture of 10 bivariate Gaussian distributions, each component of which can be interpreted as a geographic hub, e.g. a city. A full mathematical description of the model can also be found in Section D.3 of the supplementary material.

Model 2. We now group individuals into communities. All the quantities are as in Model 1 except the individuals' location $\mathbf{z}_n$ which is now replaced by $\mathbf{m}_n$ the mean of the mixture component the n-th individual was assigned to in Model 1. It is important to note that the computational cost of computing all interaction terms for Model 1 is N^2, while this cost can be reduced to N times the number of communities for Model 2. More details on the model are available in Section D.3 of the supplementary material.

Even though the two models have different heterogeneous-mixing properties, they share the same parameters and we set the DGP to the same values for both models. Precisely, $p_0 = 0.01, \beta = 2.0, \mathbf{b}_I = 1.0, \mathbf{b}_S = 0.5, \gamma = 0.1, \mathbf{b}_R = -0.5, \phi = 1, \epsilon = 0.0001, q_S = 0.2, q_I = 0.5, q_{Se} = 0.9, q_{Sp} = 0.95$. Given the DGP we can simulate from the two models for a fixed time horizon and population, see Figure 2 for a graphical representation of the disease's spread for $t = 5, 10, 20, 50$ and when $N = 1000$. Here we can observe the effect of the spatial component in both models, with the disease spreading faster in regions/communities with a higher number of infected, while isolated regions/communities are difficult to reach, and remain untouched by the disease, see Figure 2.

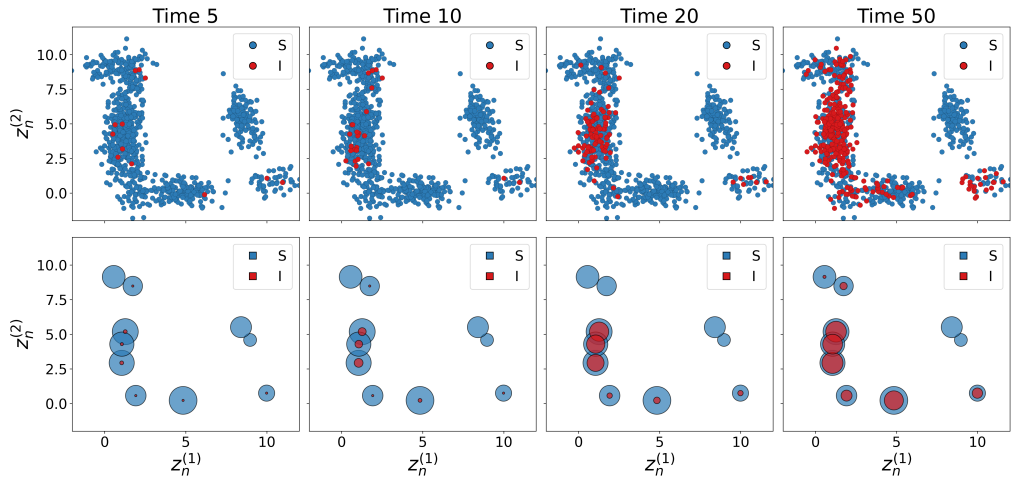

Figure 2: A realization of the latent process from Model 1 (first row) and Model 2 (second row) when $N = 1000$. Different columns are associated with different time steps. For Model 1, blue and red dots refer to susceptible and infected individuals, respectively. For Model 2, the communities are blue circles with a radius that is proportional to their population, while the red circles are proportional to the number of infected inside the communities.

Table 1: The effect of increasing N on the maximum CAL estimator for Model 1 and Model 2. The first column shows the DGP. In brackets the standard deviation of the maximum CAL estimator computed over 100 simulations.

	Model 1			Model 2		
Parameter	$N = 500$	$N = 1000$	$N = 2000$	$N = 500$	$N = 5000$	$N = 50000$
$\log(\beta)$=0.69	0.71(0.21)	0.71(0.17)	0.72(0.085)	0.72(0.2)	0.69(0.08)	0.7(0.03)
$\mathbf{b}_S$=0.5	0.45(0.06)	0.48(0.03)	0.49(0.022)	0.42(0.08)	0.5(0.01)	0.5(0.003)
$\mathbf{b}_I$=1.0	0.95(0.16)	0.97(0.15)	0.98(0.063)	0.91(0.2)	1.0(0.06)	1.0(0.025)
$\log(\gamma)$=-2.3	-2.31(0.04)	-2.29(0.03)	-2.3(0.016)	-2.32(0.05)	-2.3(0.01)	-2.3(0.003)
$\mathbf{b}_R$=-0.5	-0.51(0.05)	-0.5(0.03)	-0.5(0.018)	-0.55(0.07)	-0.5(0.01)	-0.5(0.004)
$\log(\phi)$=0.0	0.01(0.05)	-0.0(0.05)	-0.0(0.026)	-0.02(0.07)	0.0(0.02)	-0.0(0.007)
$logit(q_\cdot)$=-1.39	-1.39(0.01)	-1.39(0.01)	-1.39(0.005)	-1.39(0.01)	-1.39(0.004)	-1.39(0.001)
$logit(q_\cdot)$=0.0	0.0(0.02)	0.0(0.01)	0.0(0.007)	0.01(0.02)	0.0(0.004)	0.0(0.001)

For the experiment, we consider $N = 500, 1000, 2000$ for Model 1 and $N = 500, 5000, 50000$ for Model 2, where we can use a larger population for Model 2 because of the reduced computational cost. We then simulate for each population size and for each model 100 realizations according to the considered dynamics and observation model, with the covariates fixed.

For both models, we treat $p_0, \epsilon, q_{Se}, q_{Sp}$ as known, and infer for each simulated dataset $p_0, \beta, \mathbf{b}_S$, $\mathbf{b}_R, \phi, q_S, q_I$ by running Adam with 10 different initial conditions for 1000 gradients steps. At the end of the optimization, we choose the best out of the 10 in terms of CAL log-likelihood for each dataset. We report the results of this optimization in Table 1. Here, we can observe that, for both

models, the mean of our estimator is close to the true value of the DGP and that the variance of the maximum CAL estimator is shrinking with the population size.

5.3 Calibration and filtering for heterogeneous-mixing SIR

In this section, we present a simple pipeline explaining how the CAL can be used to track individuals' disease states within the population when considering an individual-based susceptible-infected-removed (SIR) model. We analyze both a scenario where the model is well-specified and a scenario where the model is misspecified.

Well-specified model. Consider an individual-based SIR where the individuals have the same covariates as Model 1 from Section 5.2, including the same spatial locations. We consider a $p_0(\mathbf{w}_n)$ such that the individuals in the top-left of the spatial region become infected with probability p_0, while the others are susceptible with probability 1. The transition matrix is now 3×3 with transition probabilities that are governed by the same parameters as in Model 1 from Section 5.2. The observation model now needs three parameters q_S, q_I, q_R which represent the probability of reporting S as S, I as I, R as R, respectively. We do not allow for misreporting and we force half of the population to always remain unreported. Full details are available in Section D.4 of the supplementary material. We consider this as the model that generates the data and we set the DGP to $p_0 = 0.5, \beta = 3.0, \mathbf{b}_I = 1.0, \mathbf{b}_S = 0.5, \mathbf{b}_R = -0.1, \phi = 1.5, \epsilon = 0.0001, q_S = 0.1, q_I = 0.2, q_R = 0.5$.

Misspecified model. We now present a misspecified model, with the same initial distribution as the well-specified model, but a transition kernel with an interaction term $\boldsymbol{\eta}_{n,t}$ that groups individuals into communities as in Model 2 of Section 5.2. Precisely, we use the same formulation as in Model 2 but an interaction term that considers $\bar{\mathbf{z}}_n$, the mean distance between all pairs of individuals within the community of individual n:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N} \sum_{k \in [N]} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{ -\frac{\|\mathbf{m}_n - \mathbf{m}_k\|^2 B_{n,k} + \bar{\mathbf{z}}_n^2 (1 - B_{n,k})}{2\phi^2} \right\} \mathbf{x}_{k,t-1}^{(2)},$$

where $B_{n,k} := \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| \neq 0)$. The covariates of the misspecified model are then $\mathbf{w}_n = [\mathbf{m}_n, \bar{\mathbf{z}}_n, \mathbf{c}_n]$. Similarly to Model 2, the computational cost of computing all the interaction terms is N times the number of communities, making it significantly cheaper to fit compared to the well-specified model. More details are available in Section D.4 of the supplementary material.

We generate an epidemic from the well-specified model, and we then optimize the parameters of both the well-specified and the misspecified model by running Adam with a learning rate of 0.1 for 500 iterations. After optimization, we set θ to the maximum CAL estimator and run Algorithm 1 for both models, where the CAL filter $\boldsymbol{\pi}_{n,t}$ is stored and used as an approximation of the true state. Figure 3 reports for $t = 5, 10, 20, 50$ the observations in the first row, the true states in the second row, the CAL estimate under the well-specified model and the misspecified model in the third and fourth rows. The CAL estimate on the state of the n-th individual at time t is obtained as the argmax of $\boldsymbol{\pi}_{n,t}$. It can be observed that the CAL filter is able to reproduce the spread of

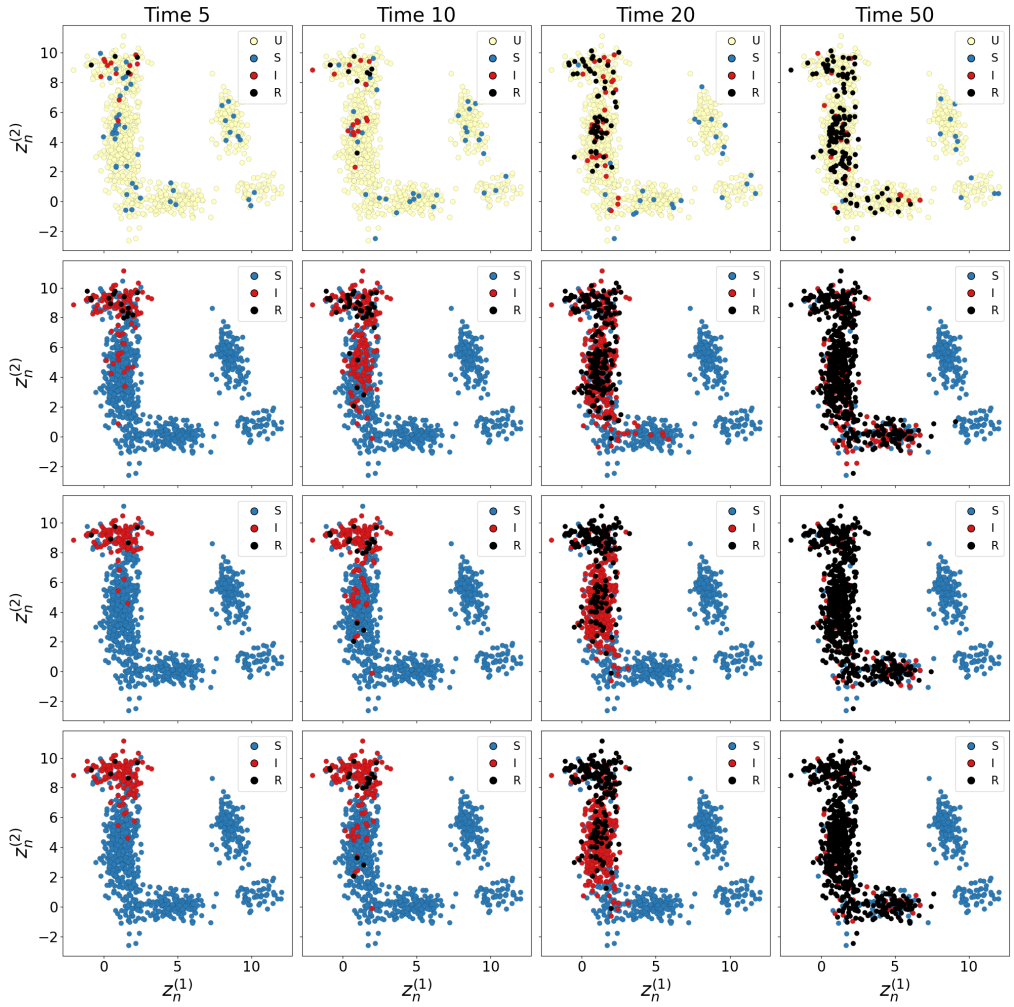

Figure 3: CAL filtering for $t = 5, 10, 20, 50$ under the well-specified and misspecified scenario. Rows from top to bottom: observed data, true latent disease states, and inferred latent disease states from CAL filtering under the well-specified and misspecified models. The yellow dots are used for unreported individuals, while blue, red, black are susceptible, infected, removed.

the epidemic from top to bottom, and even track infected individuals that are unreported for both models.

Table 2: Cross-entropy loss (the lower the better) and accuracy (the higher the better) for the CAL well-specified and misspecified, along with some baselines. The predicted state for accuracy is the argmax of the probability vector.

Metric	Random	Prev. uncertain	Prev. certain	CAL	CAL missp.
Cross-entropy	1.10	1.09	0.67	0.28	0.29
Accuracy	34.85%	65.28%	65.28%	88.48%	88.08%

Graphically, it seems we do not lose much with the misspecified model. To verify this, we consider the prediction performance on $\mathbf{x}_{n,t}$ of $\boldsymbol{\pi}_{n,t}$ for both models. This is measured via two metrics: the cross-entropy loss (De Boer et al., 2005), or equivalently minus the mean categorical log-likelihood; the accuracy, which is the percentage of correct estimates of $\mathbf{x}_{n,t}$. We also consider three baselines: "Random" where we predict individual n randomly unless we report their state; "Prev. uncertain" where we predict individual n with their latest reported state with probability 0.34 (and the other states with $(1-0.34)/2$) unless we report their state; and "Prev. certain" where we predict individual n with their latest reported state with probability 0.99 (and the other with $(1-0.99)/2$) unless we report their state. As in Figure 3, the estimate is obtained as the argmax of the probabilities. Table 2 reports the results and shows that the CAL methods perform better than the baselines in both metrics. We observe that, as expected from the graphical interpretation, little accuracy is lost when switching to the misspecified model. All the mathematical definitions of the metrics and baselines are provided in Section D.4 of the supplementary material, which also discusses how to run the same experiment with different parameter values.

5.4 Comparing CAL with sequential Monte Carlo

In this section, we compare the runtime and marginal likelihood values obtained by the CAL with those produced by two sequential Monte Carlo (SMC) algorithms: the Auxiliary Particle Filter (APF) (Johansen and Doucet, 2008) and the Block Auxiliary Particle Filter (Block APF) (Rebeschini and Van Handel, 2015), as both the number of particles and the population size increase. We also consider a just-in-time (JIT) compiled version of the CAL (Aycock, 2003) and a batched version of the Block APF. The CAL can be JIT-compiled without any additional effort because of the simplicity of its operations. It is worth noting that SMC variants exist that can be both JIT-compiled and automatically differentiated (Corenflos et al., 2021; Tan et al., 2024). For the batched Block APF, both individuals and particles are split into batches to reduce memory consumption; see Section D.6 of the supplementary materials for further details.

We consider the homogeneous-mixing SIS model from Section 1.1 and Section 2.3 with synthetic covariates $\mathbf{c}_n \sim \mathcal{N}(\cdot|0,1)$. We simulate data from models with increasing population sizes $N = 10, 100, 1000, 10000$ and time horizon $T = 100$. The full parameter settings can again be found in Section D.6 of the supplementary materials.

For each population size, we run the considered algorithms on the DGP to estimate the log-marginal likelihood, which we report in Table 3 as the mean and standard deviation over 100 runs (the CAL requires only a single run, as it is a deterministic algorithm). Recall that the APF provides unbiased estimates of the marginal likelihood (Johansen and Doucet, 2008), but it is affected by the curse of dimensionality: typically the number of particles must increase exponentially with the dimension to ensure reliable (i.e. low-variance) estimates. Moreover, although the marginal likelihood estimates are unbiased, the corresponding log-marginal likelihood estimates are negatively biased due to Jensen's inequality.

For $N = 10$ and $N = 100$, the APF exhibits low variance, giving us a suitable proxy for the ground truth. Both the CAL and the Block APF produce results close to those of the APF, with the Block APF being slightly closer. In terms of running times, the CAL (not JIT-compiled) is

Table 3: Log-marginal likelihood means and standard deviations, over 100 runs, for the SIS model from Section 1.1. Running times are reported in seconds for a single run and as averages across the 100 runs.

N	P	CAL	time	CAL compiled	time	APF	time	Block APF	time	Block APF batched	time
10	256	-54.16	1.34	-54.16	0.001	-54.09(0.03)	0.11	-54.14(0.06)	0.13	-54.14(0.06)	2.43
10	512					-54.09(0.03)	0.12	-54.14(0.04)	0.13	-54.14(0.04)	2.43
10	1024					-54.09(0.02)	0.12	-54.14(0.03)	0.13	-54.14(0.03)	2.43
10	2048					-54.09(0.01)	0.12	-54.14(0.02)	0.13	-54.14(0.02)	2.43
10	4096					-54.09(0.01)	0.12	-54.14(0.01)	0.16	-54.14(0.01)	2.43
10	8192					-54.09(0.01)	0.12	-54.14(0.01)	0.3	-54.14(0.01)	2.44
100	256	-5275.99	1.77	-5275.99	0.001	-5275.15(0.12)	1.12	-5275.94(0.2)	1.31	-5275.91(0.25)	24.0
100	512					-5275.13(0.07)	1.13	-5275.95(0.16)	1.31	-5275.96(0.2)	24.1
100	1024					-5275.14(0.05)	1.14	-5275.97(0.14)	1.47	-5275.95(0.14)	24.2
100	2048					-5275.14(0.04)	1.18	-5275.96(0.1)	2.39	-5275.96(0.09)	24.6
100	4096					-5275.14(0.03)	1.2	-5276.0(0.07)	6.01	-5276.0(0.07)	24.5
100	8192					-5275.14(0.02)	1.28	Out of memory		-5275.97(0.05)	36.2
1000	256	-71208.4	1.87	-71208.4	0.001	-74890.3(87.9)	1.23	-71247.7(10.2)	1.47	-71248.4(9.0)	25.8
1000	512					-74736.0(72.0)	1.26	-71228.8(6.3)	2.09	-71229.1(6.5)	25.8
1000	1024					-74563.2(66.8)	1.33	-71218.3(4.0)	4.55	-71217.7(4.8)	26.2
1000	2048					-74432.7(85.9)	1.48	Out of memory		-71213.1(2.9)	31.3
1000	4096					-74322.0(74.5)	1.81	Out of memory		-71210.8(2.1)	67.6
1000	8192					-74178.2(71.1)	2.63	Out of memory		-71209.7(1.7)	211
10000	256	-731878	1.81	-731877	0.001	-786077(294)	1.44	-732295(31)	3.8	-732294(28)	223
10000	512					-785395(327)	1.99	Out of memory		-732085(18)	231
10000	1024					-784872(328)	3.05	Out of memory		-731983(16)	231
10000	2048					-784332(261)	5.16	Out of memory		-731928(10)	287
10000	4096					-783742(313)	9.78	Out of memory		-731904(7)	649
10000	8192					-783216(297)	19.3	Out of memory		-731891(5)	2097

slower than both the APF and the Block APF, this is likely because sampling is more efficient than linear-algebra operations when the population is small. For $N = 1000$ and $N = 10000$, two issues arise for the SMC algorithms. First, the variance of the APF is large and the estimated log-marginal likelihood increases with the number of particles, suggesting more particles are needed. This is also confirmed by the fact that the variance does not decrease with the number of particles. Second, the Block APF requires more memory to run in parallel and must be converted into a sequential version (the batched Block APF). In this large-population regime, the CAL outperforms the SMC algorithms in terms of running time, and, being deterministic, it avoids noisy log-marginal likelihood estimates. It also appears that the (batched) Block APF has a similar asymptotic behavior, as $N, P \to \infty$, to the CAL, when $N \to \infty$, which may be of independent interest, see also Section D.6 of the supplementary materials.

It is important to note that if we consider more compartments, and models in which not all movements across compartments are possible (e.g. in an SIR model individuals cannot move from S to R, and viceversa) the APF and Block APF will become degenerate as N increases. Indeed, the proposal distributions of the APF and Block APF rely only on the observation at the current time step and cannot prevent future mismatches. In such situations, it may be necessary to use lookahead filters (Rimella et al., 2023), see Section D.5 of the supplementary materials for some experiments.

5.5 2001 UK foot-and-mouth disease outbreak

In 2001 a foot-and-mouth disease outbreak infected 2026 out of 188361 farms in the United Kingdom, resulting in damages and costs of about 8 billion pounds. This dataset has been studied in the context of individual-based models (Jewell et al., 2009; Ster et al., 2009; Deardon et al., 2010), where farms are considered as individuals with the location and the number of animals forming the individual-specific covariates. A farm can be susceptible if no animals are infected, infected if at least one animal is infected, and removed if the farm exits the epidemic due to quarantine or culling. For this study, we consider 162775 farms, from England, Wales, and the south of Scotland, see Figure 4. A fully dense spatial kernel model would be costly to run for all farms, hence we adopt the network model strategy of Section 5.3, and assign each farm to a local authority, full details are in Section D.7 of the supplementary materials.

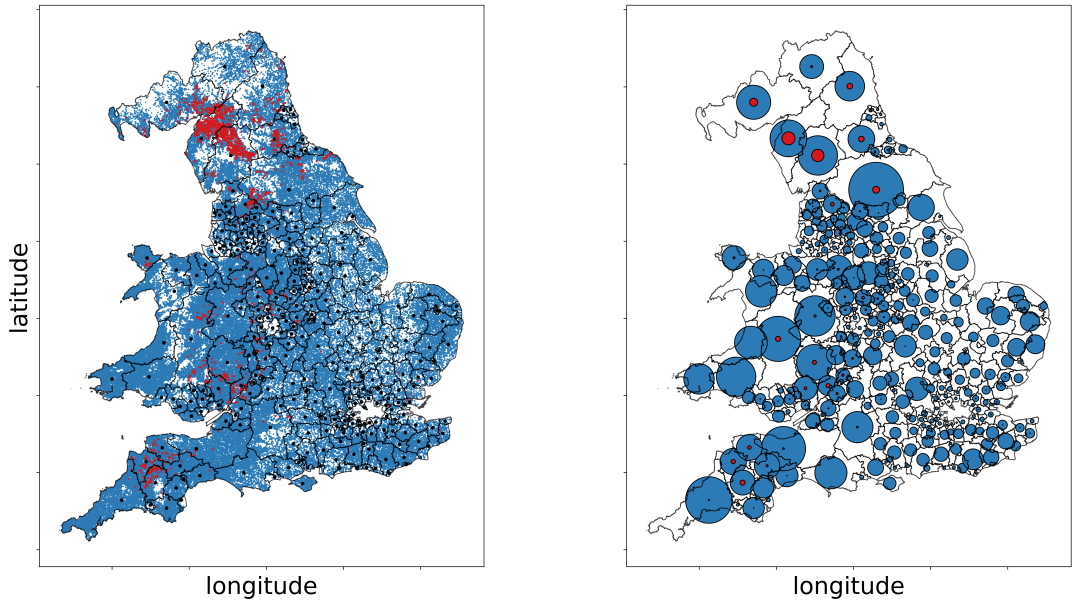

Figure 4: The farms and the local authorities included in the study. On the left, dots represent farms, with red indicating that the farm was reported infected at some point in time. On the right, the local authorities are blue circles with a radius proportional to the number of farms. Red inner circles are proportional to the number of farms within the local authority that were reported infected during the outbreak. Black contours represent the geometries of the local authorities.

We consider a heterogeneous-mixing individual-based SIR model as in Section 5.3, where transitions from S to R are also allowed, representing the culling/quarantine of healthy farms to create a containment zone around infected farms. In this model, we have two interaction terms: one controlling the spread of the disease, the other controlling the intensity of culling/quarantining. For the observation model, we do not allow for misreporting and we assume that susceptible and removed are always unreported.

We optimized the parameters with Adam in two steps. In the first step, we performed a pre-optimization over different initial conditions, with each optimization taking around 70 minutes

(10000 gradient steps). This resulted in a log-CAL of –17947.27 for the best combination of parameters. In the second step, we selected the best log-CAL parameters from the first step and used them as a warm start for a longer optimization with a varying step size, which took about 48 hours (400000 gradient steps). This produced a log-CAL of –17946.734. Note that the improvement is marginal, and we could probably have used fewer gradient steps. After the optimization we further used the resulting parameters as a warm start for an HMC to provide credible bands. Full details on the model, optimization and HMC are provided in Section D.7 of the supplementary materials.

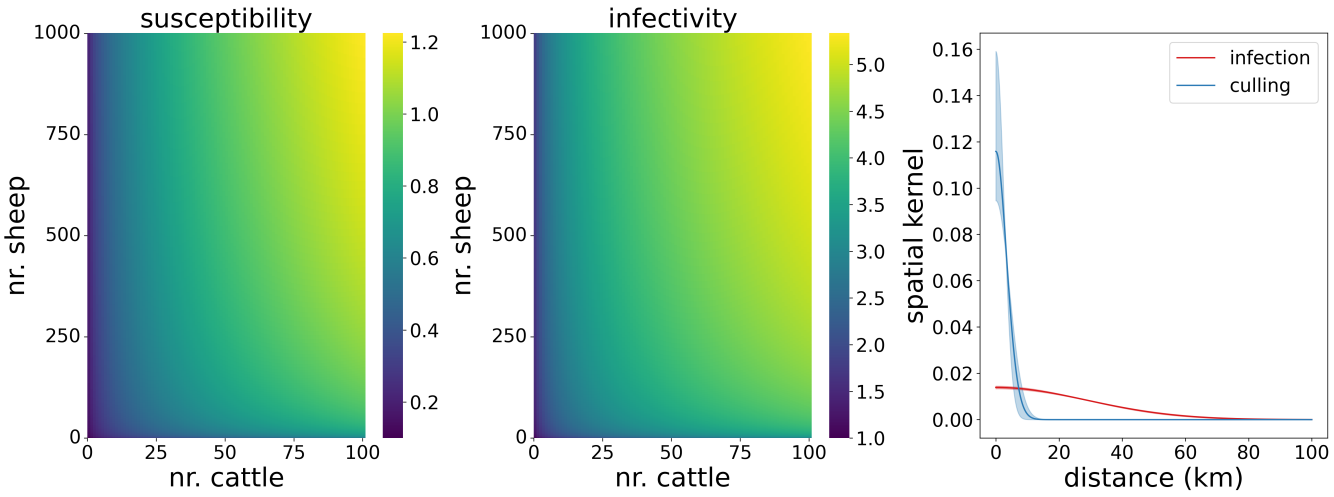

Figure 5: On the left and in the center are the heat maps of the inferred susceptibility and infectivity. On the right is the inferred spatial kernel effect on both infection and culling as a function of the distance in kilometers. The solid lines represent posterior means and the shaded bands represent 95% credible intervals.

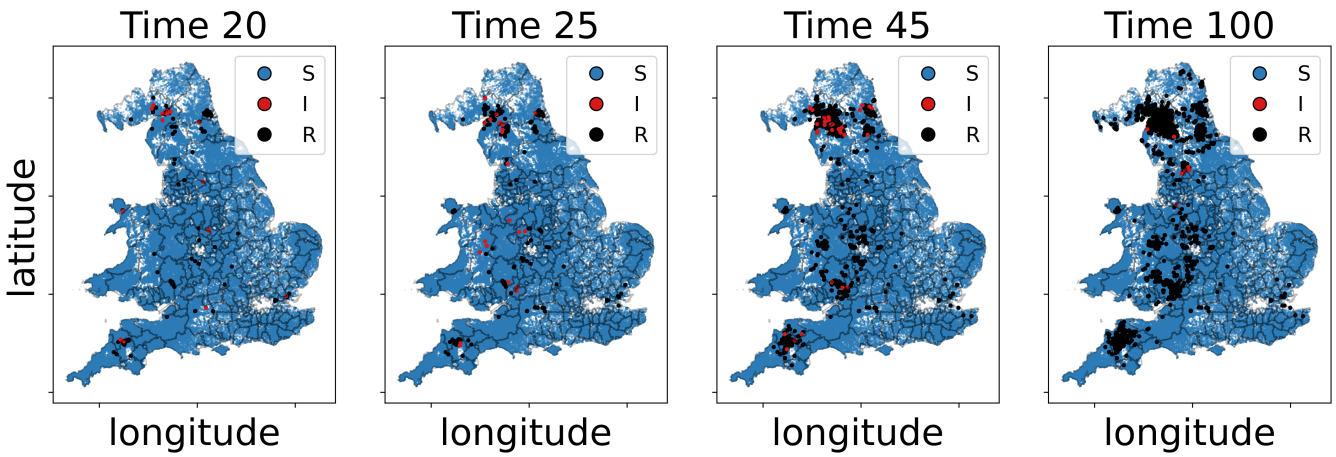

Figure 6: The CAL prediction over time of susceptible farms (blue), infected farms (red), and removed farms (black), over $t = 20, 25, 45, 100$. Parameters are set to the posterior mean of the HMC.

After the HMC run, we can study the inferred mean effects of owning cattle/sheep on susceptibility/infectivity. The plots on the left and in the center of Figure 5 show that owning cattle affects both susceptibility and infectivity more than owning sheep, which is a similar conclusion to Jewell et al. (2009) and Rimella et al. (2025). We can also analyze how distance affects both the infection and the culling/quarantine processes. No data about culled farms are included in the calibration, meaning that the culling/quarantine process is inferred from the infected farms only. The right plot of Figure 5 shows that the infection process travels on the order of 60km, with a culling/quarantine process that is applied up to a radius of 10km, which is also in line with the "surveillance zone" protocol considered by the government (gov.uk, 2025). We further included 95% credible interval as shaded bands. Following the reasoning of Section 5.3, we can plot over time the predicted state via the CAL filter, see Figure 6, when the parameters are set to the posterior mean of the HMC. We consider $t = 20, 25$ to show the spatial spread and culling/quarantine process in the short term, $t = 45$ to show the spread when we are close to the peak of the epidemic, and $t = 100$ to show when we are close to the end. We observe that the epidemic rapidly spread in the north of England, Cumbria in particular, and Cornwall, to then affect Wales on a lower scale. Even though no information about culling/quarantine was fed to the model, the CAL is able to spatially detect them and match the locations of confirmed and suspected premises towards the end of the outbreak (October 2001) as reported graphically on the UK government website (data.gov.uk, 2001).

5.5.1 Benchmarking, overdispersion and extensions

A natural question that arises is whether the considered model is correctly specified. This can be assessed by comparing its log-likelihood values with those of simple benchmarks that are much easier to fit. With this in mind, we consider $\widetilde{\mathbf{y}}_{n,t}$ which takes values $1, 0$ depending on whether the farm is reported as infected or not. We then model $p(\widetilde{\mathbf{y}}_{n,0} = 1), p(\widetilde{\mathbf{y}}_{n,t}|\widetilde{\mathbf{y}}_{n,t-1})$ with an AR logistic regression:

$$p(\widetilde{\mathbf{y}}_{n,1} = 1) = \frac{\gamma}{1 + e^{-\mathbf{b}^\top \mathbf{w}_n}}, \quad p(\widetilde{\mathbf{y}}_{n,t} = 1|\widetilde{\mathbf{y}}_{n,t-1}) = \frac{\gamma}{1 + e^{-\beta \widetilde{\mathbf{y}}_{n,t-1} - \mathbf{b}^\top \mathbf{w}_n}},$$

where $\gamma \in [0, 1]$, $\beta \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^{C+1}$, with C covariates for each individual (plus one for the intercept). We optimize $\gamma, \beta, \mathbf{b}$ using gradient ascent on the log-likelihood with the Adam optimizer, obtaining a log-likelihood of -20858.99. We further consider another AR logistic regression benchmark in which $\gamma, \beta, \mathbf{b}$ are specific to each local authority, and fit the parameters in the same way. This second benchmark yields a log-likelihood of -18866.53. Both benchmarks provide sensible log-likelihood values that are not far from the -17946.73 achieved by the CAL, but they remain lower, suggesting that the individual-based model approach explains the data better. The usefulness of such benchmarks extends beyond the foot-and-mouth application and the CAL: these AR logistic regressions can be easily generalized to multiple states and can be used to check whether an individual-based model is misspecified.

In our study, we focused on models that may underestimate uncertainty and therefore produce overconfident parameter estimates. Evaluating overdispersed models is thus crucial for assessing the quality of the inferential procedure and avoiding model misspecification (Stocks et al., 2020; Whitehouse et al., 2023; Li et al., 2024).

We next consider the same SIR model as in Section 5.5 but replace the transmission parameter β with $\beta\xi_t$, where $\log \xi_t$ is distributed according to a Gaussian distribution with unknown mean μ_o and unknown standard deviation σ_o. We examine two cases: a model with "shared" overdispersion, where ξ_t is common to all farms, and a model with "local authority" overdispersion, where a separate $\xi_{B,t}$ is defined for each local authority B and shared across the farms within that authority.

Following the procedure of Whitehouse et al. (2023), we nest the CAL within an SMC algorithm to estimate the marginal likelihood of the overdispersed models. Here the CAL marginalizes over the latent individuals' states, while the SMC marginalizes over the stochastic parameters. We perform a grid search over (μ_o, σ_o) for the "shared" overdispersion model, where the values on the grid are given by the CAL within SMC algorithm. This yields $\mu_o = 0$, $\sigma_o = 0.25$, and a log-marginal likelihood of $-17926.99(0.46)$ (the standard deviation is estimated over 100 runs of the SMC). We then estimate the log-marginal likelihood for the "local authority" model and obtain a value of $-17907.55(0.84)$ (the standard deviation is estimated over 100 runs of the SMC). As the "local authority" model considers 289 stochastic parameters (one for each local authority) we considered a block SMC approach (Rebeschini and Van Handel, 2015). More details on the models, the SMC algorithms, and the grid search are available in Section D.7 of the supplementary materials.

We observe that both overdispersed models achieve better log-marginal likelihoods than the other model, but this improvement comes at a high computational cost: a single run of CAL within SMC takes approximately 7 minutes for the "shared" overdispersion model with 1024 particles and about 43 minutes for the "local authority" overdispersion model with 512 particles. Moreover, neither JIT compilation nor automatic differentiation is straightforward in this context. It is therefore important to consider alternative SMC methods that preserve these properties (Corenflos et al., 2021; Tan et al., 2024) and/or leverage parallel-in-time implementations (Corenflos et al., 2022).

From a theoretical perspective, we know that the CAL performs well when individuals decouple in the large-population limit. In this regime, neighboring individuals being infected have negligible predictive consequences for a considered individual as long as we know the regional force of infection. In the case of the foot-and-mouth disease, the spreading process is not necessarily determined by adjacent contacts but can also involve animal or insect vectors and the livestock transport network (Keeling et al., 2001; Diggle, 2006; Jewell et al., 2009). These considerations align with the assumptions and modeling framework of the CAL, we leave to future work the development of modeling approaches in which farms do not decouple.

6 Discussion, limitations and future work

We have proposed a computationally and mathematically simple algorithm to enable approximate likelihood-based inference for a broad class of individual-based models of epidemics, supported by both theoretical foundations and practical implementations.

At first glance, the CAL and block particle filters (Rebeschini and Van Handel, 2015) are similar, as they both exploit certain factorization properties of the model. However, there is a key difference in how their approximations are constructed. The CAL replaces the state of the system at time $t-1$ with its expectation under the previous filtering distribution, whereas the block particle filter applies a blocking approximation after prediction, in the form of independent resampling for each

block. Hence, for finite N, the CAL targets the likelihood of the approximate model described in Section 3.2, while the block particle filter targets the likelihood of the approximate model obtained by marginalizing over the blocks at each time step. Their theoretical justifications also differ: the CAL is motivated by a large-population limit (i.e. the dimension goes to infinity), while the block particle filter relies on large-block asymptotics (i.e. the partition on the dimensions becomes coarser and coarser). Table 3 and the graphical illustration in Section D.6 of the supplementary material both suggest that the marginal likelihood approximation from the block particle filter behaves, as $N, P \to \infty$, similarly to the log-CAL, as $N \to \infty$. This indicates that the CAL theory may provide a justification for the use of block particle filters in high-dimensional settings (Li et al., 2024; Wheeler et al., 2024), which is an interesting direction for further investigation.

One limitation of the CAL is that individuals are updated independently, without accounting for correlations between them. This becomes problematic when the individuals do not decouple in the large population limit, and hence when the state of one individual provides information about others. Some clear examples of such models are household models (Rimella et al., 2023), where individuals are assigned to households and as the number of individuals grows so does the number of households. In such a scenario, we expect the CAL to provide a bad approximation of the likelihood as it does not account for the dependencies within the households. One possible solution might be to model the state of each household instead of the individuals, hence considering the households to decouple in the large population limit. We are confident that the CAL can be extended to this modeling framework as the categorical representation can be used in any discrete state-space. It is important to note that this will trade off the quality of the approximation with the performance as the computational cost will deteriorate exponentially in the households sizes.

Another key consideration is the implicit geometric distribution assumed for compartmental waiting periods, e.g. infectious periods. Negative binomial waiting times can be approximated by introducing additional compartments, though this approach is not always satisfactory. A promising avenue for future research would be to extend our model to higher-order Markovian or semi-Markovian dynamics (Jewell et al., 2009; Touloupou et al., 2020).

Throughout our examples, we assume $\mathbf{W}$ to be static over time, though this assumption could be relaxed to incorporate dynamic covariates, and even integrate a time-evolving contact network process (Bu et al., 2022) within the dynamics. Additionally, there are interesting avenues for extending our methodology to continuous observation spaces, with applications in areas such as target tracking (Whiteley et al., 2010) and epidemic modeling using continuous serological data (Hay et al., 2024).

Funding

LR, MW, NW are supported by EPSRC Grant UKRI3641: "Scalable statistical inference for disease contact networks". MW acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (Reference No. MR/X020258/1), funded by the UK Medical Research Council. This UK-funded grant is carried out within the framework of the Global Health EDCTP3 Joint Undertaking. PF acknowledges funding from the EPSRC grant "Prob_AI" (Reference No. EP/Y028783/1). CJ acknowledges funding from the MRC (MR/S004793/1), BBSRC (BB/T004312/1), EPSRC

(EP/V042866/1), and Research England as part of the E3: Expanding Excellence in England programme.

Acknowledgements

The authors thank "The High-End Computing" (HEC) facility at Lancaster University for providing the computational resources to run the experiments. The authors thank Dr. Sam Power for the useful discussion on "exact inference for an approximate model" and Dr. Louis Sharrock for providing helpful references on the propagation of chaos.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from https://www.tensorflow.org/.

Andrews, D. W. (1992). Generic uniform convergence. *Econometric theory 8*(2), 241–257.

Aycock, J. (2003). A brief history of just-in-time. *ACM Computing Surveys (CSUR) 35*(2), 97–113.

Bu, F., A. E. Aiello, A. Volfovsky, and J. Xu (2025). Stochastic EM algorithm for partially observed stochastic epidemics with individual heterogeneity. *Biostatistics 26*(1), kxae018.

Bu, F., A. E. Aiello, J. Xu, and A. Volfovsky (2022). Likelihood-based inference for partially observed epidemics on dynamic networks. *Journal of the American Statistical Association 117*(537), 510–526.

Chatha, P., F. Bu, J. Regier, E. Snitkin, and J. Zelner (2024). Neural posterior estimation for stochastic epidemic modeling. *arXiv:2412.12967*.

Chopin, N. and O. Papaspiliopoulos (2020). *Introduction to Sequential Monte Carlo*. Springer International Publishing.

Chowell, G., N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman (2004). The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *Journal of Theoretical Biology 229*(1), 119–126.

Cocker, D., M. Sammarro, K. Chidziwisano, N. Elviss, S. T. Jacob, H. Kajumbula, L. Mugisha, D. Musoke, P. Musicha, A. P. Roberts, et al. (2023). Drivers of resistance in uganda and malawi (drum): a protocol for the evaluation of one-health drivers of extended spectrum beta lactamase (esbl) resistance in low-middle income countries (lmics). *Wellcome open research 7*, 55.

Corenflos, A., N. Chopin, and S. Särkkä (2022). De-sequentialized monte carlo: a parallel-in-time particle smoother. *Journal of Machine Learning Research 23*(283), 1–39.

Corenflos, A., J. Thornton, G. Deligiannidis, and A. Doucet (2021, 18–24 Jul). Differentiable particle filtering via entropy-regularized optimal transport. In M. Meila and T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning*, Volume 139 of *Proceedings of Machine Learning Research*, pp. 2100–2111. PMLR.

data.gov.uk (2001). Foot and mouth disease outbreak on the 14th October 2001. http://data.defra.gov.uk/Agriculture/APHA0912-FMD_Overview_Map_20011014.jpg (Last access 22-01-2026).

data.gov.uk (2023). Local authority districts (December 2023) boundaries UK BFE. https://www.data.gov.uk/dataset/288458f7-7789-47d0-80d4-ffdf746c6b75/local-authority-districts-december-2023-boundaries-uk-bfe (Last access 22-01-2026).

De Boer, P.-T., D. P. Kroese, S. Mannor, and R. Y. Rubinstein (2005). A tutorial on the cross-entropy method. *Annals of Operations Research 134*, 19–67.

Deardon, R., S. P. Brooks, B. T. Grenfell, M. J. Keeling, M. J. Tildesley, N. J. Savill, D. J. Shaw, and M. E. Woolhouse (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica 20*(1), 239.

Diggle, P. J. (2006). Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical methods in medical research 15*(4), 325–336.

Estrada, E., M. Fox, D. J. Higham, and G.-L. Oppo (2010). *Network science: complexity in nature and technology*. Springer Science & Business Media.

gov.uk (2025). Foot and mouth disease control strategy for Great Britain. https://www.gov.uk/government/publications/foot-and-mouth-disease-control-strategy (Last access 22-01-2026).

Hay, J. A., I. Routledge, and S. Takahashi (2024). Serodynamics: a primer and synthetic review of methods for epidemiological inference using serological data. *Epidemics*, 100806.

Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

Jewell, C. P., T. Kypraios, P. Neal, and G. O. Roberts (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis 4*, 465–496.

Johansen, A. M. and A. Doucet (2008). A note on auxiliary particle filters. *Statistics & Probability Letters 78*(12), 1498–1504.

Ju, N., J. Heng, and P. E. Jacob (2021). Sequential Monte Carlo algorithms for agent-based models of disease transmission. *arXiv:2101.12156*.

Keeling, M. J., M. E. J. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell (2001). Dynamics of the 2001 uk foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science 294*(5543), 813–817.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. *Proceedings of the Royal Society B: Biological Sciences 282*(1806), 20150347.

Kingma, D. P. and J. Ba (2014). Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Le Boudec, J.-Y., D. McDonald, and J. Mundinger (2007). A generic mean field convergence result for systems of interacting objects. In *Fourth International Conference on the Quantitative Evaluation of Systems (QEST 2007)*, pp. 3–18. IEEE.

Lekone, P. E. and B. F. Finkenstädt (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics 62*(4), 1170–1177.

Li, J., E. L. Ionides, A. A. King, M. Pascual, and N. Ning (2024, 07). Inference on spatiotemporal dynamics for coupled biological populations. *Journal of The Royal Society Interface 21*(216), 20240217.

McKinley, T. J., I. Vernon, I. Andrianakis, N. McCreesh, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White (2018). Approximate Bayesian Computation and Simulation-Based Inference for Complex Stochastic Epidemic Models. *Statistical Science 33*(1), 4 – 18.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph. D. thesis, Massachusetts Institute of Technology.

Neal, R. M. (2012). Mcmc using hamiltonian dynamics. *arXiv:1206.1901*.

Rebeschini, P. and R. Van Handel (2015). Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability 25*, 2809–2866.

Rimella, L., S. Alderton, M. Sammarro, B. Rowlingson, D. Cocker, N. Feasey, P. Fearnhead, and C. Jewell (2023, 07). Inference on extended-spectrum beta-lactamase escherichia coli and klebsiella pneumoniae data through smc2. *Journal of the Royal Statistical Society Series C: Applied Statistics 72*(5), 1435–1451.

Rimella, L., C. Jewell, and P. Fearnhead (2023). Approximating Optimal SMC Proposal Distributions in Individual-Based Epidemic Models. *Statistica Sinica*, SS–2022–0198.

Rimella, L., C. Jewell, and P. Fearnhead (2025). Simulation based composite likelihood. *Statistics and Computing 35*(3), 58.

Seymour, R. G., T. Kypraios, and P. D. O'Neill (2022). Bayesian nonparametric inference for heterogeneously mixing infectious disease models. *Proceedings of the National Academy of Sciences 119*(10), e2118425119.

Sharkey, K. J. (2008). Deterministic epidemiological models at the individual level. *Journal of Mathematical Biology 57*, 311–331.

Sharrock, L., N. Kantas, P. Parpas, and G. A. Pavliotis (2023). Online parameter estimation for the McKean–Vlasov stochastic differential equation. *Stochastic Processes and their Applications 162*, 481–546.

Sherborne, N., J. C. Miller, K. B. Blyuss, and I. Z. Kiss (2018). Mean-field models for non-markovian epidemics on networks. *Journal of Mathematical Biology 76*, 755–778.

Sorenson, H. W. and A. R. Stubberud (1968). Non-linear filtering by approximation of the a posteriori density. *International Journal of Control 8*(1), 33–51.

Ster, I. C., B. K. Singh, and N. M. Ferguson (2009). Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in Great Britain. *Epidemics 1*(1), 21–34.

Stocks, T., T. Britton, and M. Höhle (2020). Model selection and parameter estimation for dynamic epidemic models via iterated filtering: application to rotavirus in germany. *Biostatistics 21*(3), 400–416.

Tan, K., G. Hooker, and E. L. Ionides (2024). Accelerated inference for partially observed markov processes using automatic differentiation. *arXiv preprint arXiv:2407.03085*.

Touloupou, P., B. Finkenstädt, and S. E. Spencer (2020). Scalable Bayesian inference for coupled hidden Markov and semi-Markov models. *Journal of Computational and Graphical Statistics 29*(2), 238–249.

Wheeler, J., A. Rosengart, Z. Jiang, K. Tan, N. Treutle, and E. L. Ionides (2024). Informing policy via dynamic models: Cholera in haiti. *PLOS Computational Biology 20*(4), e1012032.

Whitehouse, M. (2025). Accelerated inference for stochastic compartmental models with overdispersed partial observations. *arXiv preprint arXiv:2505.06935*.

Whitehouse, M., N. Whiteley, and L. Rimella (2023). Consistent and fast inference in compartmental models of epidemics using poisson approximate likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology 85*(4), 1173–1203.

Whiteley, N. and L. Rimella (2021). Inference in stochastic epidemic models via multinomial approximations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1297–1305. PMLR.

Whiteley, N., S. Singh, and S. Godsill (2010). Auxiliary particle implementation of probability hypothesis density filter. *IEEE Transactions on Aerospace and Electronic Systems 46*(3), 1437–1454.

A Notation and assumptions

A.1 General notation

Given an integer $M \in \mathbb{N}$, we use $x_{0:M} \coloneqq x_0, \ldots, x_M$ for indexing sequences, $[M] \coloneqq \{1, \ldots, M\}$ for the set of the first M positive integers, and $\mathbf{x} \coloneqq \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}\right)$ for an M-dimensional vector. Given two M-dimensional vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ we denote with $\mathbf{x}_1 \odot \mathbf{x}_2$ the element-wise product and with $\mathbf{x}_1 \oslash \mathbf{x}_2$ the element-wise division. We write 1_M for the M-dimensional vector of all ones, Δ_M for the M-dimensional probability simplex, i.e. $\Delta_M \coloneqq \left\{\mathbf{x} \in [0,1]^M : \sum_{i=1}^M \mathbf{x}^{(i)} = 1\right\}$, and $\mathbb{O}_M$ for the set of one-hot encoding vectors with dimension M, i.e. the set $\mathbb{O}_M \coloneqq \left\{\mathbf{x} \in \{0,1\}^M : \exists j \in [M] : \mathbf{x}^{(j)} = 1 \text{ and } \mathbf{x}^{(i)} =$ Note that $\mathbb{O}_M \subset \Delta_M$. Given $\boldsymbol{\pi} \in \Delta_M$ we denote with $\mathrm{Cat}(\cdot|\boldsymbol{\pi})$ the categorical distribution over $\mathbb{O}_M$ which assigns probability $\boldsymbol{\pi}^{(i)}$ to the vector $\mathbf{x} \in \mathbb{O}_M$ with $\mathbf{x}^{(i)} = 1$ and $\mathbf{x}^{(j)} = 0$ for $j \neq i$.

We shall work with the following norms. For an M-dimensional vector π and an $M \times M$-matrix K we define

$$\|\pi\|_\infty \coloneqq \max_{i \in [M]} \left|\pi^{(i)}\right|, \quad \|K\|_\infty \coloneqq \max_{i \in [M]} \sum_{j=1}^M \left|K^{(i,j)}\right|.$$

For a vector-valued function $f : \mathbb{S} \to \mathbb{R}^M$, we write

$$\|f\|_\infty \coloneqq \sup_{s \in \mathbb{S}} \max_{i \in [M]} \left|f(s)^{(i)}\right|,$$

For an $\mathbb{R}$-valued random variable $\mathbf{x}$ the L^4 norm is written:

$$\|\mathbf{x}\|_4 \coloneqq \left(\mathbb{E}\left[|\mathbf{x}|^4\right]\right)^{\frac{1}{4}}.$$

We need a notation for the support of probability vectors and matrices. For an M-dimensional vector $\boldsymbol{\pi}$ we define $\mathbf{supp}(\boldsymbol{\pi}) \coloneqq \{i \in [M] : \boldsymbol{\pi}^{(i)} \neq 0\}$. Similarly for a $M \times M$ matrix $\boldsymbol{\Pi}$ we define $\mathbf{supp}(\boldsymbol{\Pi}) \coloneqq \{(i,j) \in [M]^2 : \boldsymbol{\Pi}^{(i,j)} \neq 0\}$.

A.2 Model and CAL notation

We consider the individual-based model and the CAL algorithm described in the main paper. To formulate and prove our theoretical results, we need to make explicit the dependence of various quantities on the parameter vector $\theta \in \Theta$, covariates $\mathbf{w}_n$, and/or the population size N.

We write $\mathbf{W}^N = (\mathbf{w}_1, ..., \mathbf{w}_N)$ for the first N population covariate vectors, $\mathbf{x}_{n,t}^N$ for the state at time t of the n-th individual, $\mathbf{X}_t^N = (\mathbf{x}_{1,t}^N, \ldots, \mathbf{x}_{N,t}^N)$ for the state of the population at time t, $\mathbf{y}_{n,t}^N$ and $\mathbf{Y}_t^N$ for the observations at an individual and population level.

We use for the initial distribution, transition matrix, and emission matrix the notation: $p_0(\mathbf{w}_n, \theta)$, $K_\cdot(\mathbf{w}_n, \theta)$, and $G(\mathbf{w}_n, \theta)$, with $\eta(\mathbf{w}_n, \mathbf{W}, \mathbf{X})$ becoming $\eta^N(\mathbf{w}_n, \theta, \mathbf{W}^N, \mathbf{X}^N)$, and $\boldsymbol{\eta}_{n,t}$ becoming $\boldsymbol{\eta}_t^N(\mathbf{w}_n, \theta)$. For the quantities computed in the CAL algorithm, we highlight the functional dependence on the covariates and the considered parameter value, with $\boldsymbol{\pi}_{n,t|t-1}, \boldsymbol{\mu}_{n,t}, \boldsymbol{\pi}_{n,t}, \widetilde{\boldsymbol{\eta}}_{n,t}$ becoming $\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta)$, $\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)$, $\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta), \widetilde{\boldsymbol{\eta}}^N_t(\mathbf{w}_n, \theta)$. Note that for $t \geq 1$ the quantities $\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta), \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)$ also depend on the population covariates $\mathbf{W}^N$ and the observations $\mathbf{Y}^N_{1:t-1}$, and similarly the quantities $\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta), \widetilde{\boldsymbol{\eta}}^N_t(\mathbf{w}_n, \theta)$ depend on $\mathbf{W}^N$ and $\mathbf{Y}^N_{1:t}$, but these dependencies are not shown in the notation.

For completeness we write the CAL recursion as in Algorithm 1 with the dependence on $\mathbf{w}_n$ and θ explicit in the notation, for any $n \in [N]$:

$$
\begin{aligned}
\boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n, \theta) &\coloneqq p_0(\mathbf{w}_n, \theta), \\
\boldsymbol{\Pi}^N_{t-1} &\coloneqq \left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_1, \theta), \ldots, \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_N, \theta)\right), \\
\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n, \theta) &\coloneqq \eta^N(\mathbf{w}_n, \theta, \mathbf{W}^N, \boldsymbol{\Pi}^N_{t-1}), \\
\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) &\coloneqq \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta)^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n, \theta)}(\mathbf{w}_n, \theta)\right]^\top, \\
\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) &\coloneqq \left[\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta)\right]^\top, \\
\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta) &\coloneqq \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) \odot \left\{\left[G(\mathbf{w}_n, \theta) \oslash \left(1_M \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)^\top\right)\right] \mathbf{y}^N_{n,t}\right\}.
\end{aligned}
\tag{6}
$$

A.3 Assumptions

The compactness and continuity in the following assumption are standard conditions in the consistency theory of maximum likelihood estimators.

Assumption 6. *The parameter space Θ and the covariate space $\mathbb{W}$ are compact subsets of Euclidean spaces. Moreover, the initial distribution $p_0(w, \theta)$, the transition matrix $K_\eta(w, \theta)$, and the emission matrix $G(w, \theta)$ are all continuous functions in their arguments w, θ.*

For the purposes of our theory, we shall treat the individual-specific covariate vectors as random, independent, and identically distributed across the population. This can be interpreted as a random design assumption of a sort, which is commonly adopted in the asymptotic studies of regression and classification methods.

Assumption 7. *The covariates $\mathbf{w}_1, \mathbf{w}_2, \ldots$ are independent and identically distributed according to a distribution Γ on $\mathbb{W}$.*

The following assumption will be used to establish that the logarithm of the CAL is well-defined across all possible values of parameters, covariates, and almost all realizations of the data.

Assumption 8. *The following hold:*

- *for any $w \in \mathbb{W}$ and $\theta, \theta' \in \Theta$ we have that $\boldsymbol{supp}(p_0(w, \theta)) = \boldsymbol{supp}(p_0(w, \theta'))$;*
- *for any $w \in \mathbb{W}$, $\eta, \eta' \in [0, C]$ and $\theta, \theta' \in \Theta$ we have that $\boldsymbol{supp}(K_\eta(w, \theta)) = \boldsymbol{supp}(K_{\eta'}(w, \theta'))$;*

- *for any* $w \in \mathbb{W}$ *and* $\theta, \theta' \in \Theta$ *we have that* $\boldsymbol{supp}(G(w,\theta)) = \boldsymbol{supp}(G(w,\theta'))$.

The following assumption constrains the form of the function η which determines the mechanism of interaction amongst the population.

Assumption 9. *For any* $\theta \in \Theta, w \in \mathbb{W}, N \in \mathbb{N}$, *and for any* $W^N = (w_1, \ldots, w_N), \Pi^N = (\pi_1, \ldots, \pi_N)$ *with* $w_n \in \mathbb{W}, \pi_n \in \Delta_M$ *for all* $n \in [N]$, *we have:*

$$\eta^N(w, \theta, W^N, \Pi^N) = \frac{1}{N} \sum_{n \in [N]} d(w, w_n, \theta)^\top \pi_n,$$

where $d : \mathbb{W} \times \mathbb{W} \times \Theta \to [0, C]^M$ *is a bounded function, i.e.* $\|d\|_\infty \leq C < \infty$, *from which we also obtain* $\eta^N(w, \theta, W^N, \Pi^N) \in [0, C]$ *for any* N, w, π, W^N, Π^N.

Assumption 10. *For any* $\theta \in \Theta$ *and* $w \in \mathbb{W}$, *the matrix* $K_\eta(w, \theta)$ *is Lipschitz continuous in* η *with Lipschitz constant* L, *that is for any* $\eta, \eta' \in [0, C]$ *we have:*

$$\|K_\eta(w, \theta) - K_{\eta'}(w, \theta)\|_\infty \leq L \, |\eta - \eta'| \, .$$

A.4 The data-generating process

All the random variables appearing in our theory are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Thus in accordance with Assumption 7, under $\mathbb{P}$, $\mathbf{w}_1, \mathbf{w}_2, \ldots$ are i.i.d.

Let $\theta^\star$ be an arbitrarily chosen but a fixed member of Θ, which will be referred to as the data-generating parameter (DGP). For each $N \geq 1$, the DGP determines the distributions of $(\mathbf{X}_t^N)_{t \geq 0}$ and $(\mathbf{Y}_t^N)_{t \geq 1}$ under $\mathbb{P}$ conditional on $\mathbf{W}^N$, in that for each $n \in [N]$,

$$\begin{aligned}
\mathbf{x}_{n,0}^N | \mathbf{w}_n &\sim \mathrm{Cat}\left(\cdot \,|\, p_0(\mathbf{w}_n, \theta^\star) \right), \\
\mathbf{x}_{n,t}^N | \mathbf{X}_{t-1}^N, \mathbf{W}^N &\sim \mathrm{Cat}\left(\cdot \,\middle|\, \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) \right]^\top \right), \\
\mathbf{y}_{n,t}^N | \mathbf{x}_{n,t}^N, \mathbf{w}_n &\sim \mathrm{Cat}\left(\cdot \,\middle|\, \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n, \theta^\star) \right]^\top \right).
\end{aligned}$$

B Closed-forms under categorical approximations

Proposition 2 and Proposition 3 below show how the formulae appearing in the CAL algorithm in Section 3.1 of the main paper are derived and can be interpreted as prediction and correction operations associated with the approximate model specified in Section 3.2 of the main paper.

Proposition 2. *If* $\widetilde{\mathbf{X}}_{t-1}^N | \mathbf{W}^N \sim \bigotimes_{n \in [N]} Cat(\cdot | \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta))$ *and:*

$$\widetilde{\mathbf{x}}_{n,t}^N | \widetilde{\mathbf{X}}_{t-1}^N, \mathbf{W}^N \sim Cat\left(\cdot | \left[(\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n, \theta)}(\mathbf{w}_n, \theta) \right]^\top \right) \quad \text{for } n \in [N],$$

then $\widetilde{\mathbf{X}}_t^N | \mathbf{W}^N \sim \bigotimes_{n \in [N]} Cat(\cdot | \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta))$ *with:*

$$\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta) = \left(\boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n, \theta)}(\mathbf{w}_n, \theta) \right)^\top \quad \text{for } n \in [N].$$

Proof. We just want to compute the marginal at time t after applying the approximate transition kernel:

$$\begin{aligned}
&\mathbb{P}(\widetilde{\mathbf{X}}_t^N = \widetilde{X}_t^N | \mathbf{W}^N) \\
&= \sum_{X^N=(x_1,\dots,x_N)} \mathbb{P}(\widetilde{\mathbf{X}}_{t-1}^N = X^N | \mathbf{W}^N) \mathbb{P}(\widetilde{\mathbf{X}}_t^N = \widetilde{X}_t^N | \mathbf{W}^N, \widetilde{\mathbf{X}}_{t-1}^N = X^N) \\
&= \sum_{X^N=(x_1,\dots,x_N)} \prod_{n\in[N]} \left[\left(x_n^\top \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta) \right) \left(x_n^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \widetilde{x}_{n,t} \right) \right] \\
&= \prod_{n\in[N]} \sum_{x_n} \left[\left(x_n^\top \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta) \right) \left(x_n^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \widetilde{x}_{n,t}^N \right) \right] \\
&= \prod_{n\in[N]} \sum_{i=1}^{M} \left[\boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta)^{(i)} K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)^{(i,\cdot)} \widetilde{x}_{n,t}^N \right] \\
&= \prod_{n\in[N]} \left[\boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top \widetilde{x}_{n,t}^N.
\end{aligned}$$

□

The next proposition considers the correction step, where we perform a Bayes update with a factorized categorical distribution prior and a likelihood that is given by the emission distribution of the individual-based model. As the posterior distribution is in the family of the prior distribution, we say that the factorized categorical distribution is a conjugate prior for the individual-based model observation model.

Proposition 3. *If* $\widetilde{\mathbf{X}}_t^N | \mathbf{W}^N \sim \bigotimes_{n\in[N]} Cat(\cdot | \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta))$ *and:*

$$\widetilde{\mathbf{y}}_{n,t}^N | \widetilde{\mathbf{x}}_{n,t}^N, \mathbf{w}_n \sim Cat\left(\cdot | \left[(\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n, \theta) \right]^\top \right) \quad \text{for } n \in [N],$$

then $\widetilde{\mathbf{Y}}_t^N | \mathbf{W}^N \sim \bigotimes_{n\in[N]} Cat(\cdot | \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta))$ *with:*

$$\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta) = \left[\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta) \right]^\top \quad \text{for } n \in [N],$$

and $\widetilde{\mathbf{X}}_t^N | \widetilde{\mathbf{Y}}_t^N, \mathbf{W}^N \sim \bigotimes_{n\in[N]} Cat(\cdot | \boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta))$ *with:*

$$\boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta) = \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta) \odot \left\{ \left[G(\mathbf{w}_n, \theta) \oslash \left(1_M \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^\top \right) \right] \widetilde{\mathbf{y}}_{n,t}^N \right\} \quad \text{for } n \in [N].$$

Proof. We start by computing the marginal likelihood:

$$\begin{aligned}
\mathbb{P}(\widetilde{\mathbf{Y}}_t^N = \widetilde{Y}_t^N | \mathbf{W}^N) &= \sum_{X^N=(x_1,\dots,x_N)} \mathbb{P}(\widetilde{\mathbf{Y}}_t^N = \widetilde{Y}_t^N | \mathbf{W}^N, \widetilde{\mathbf{X}}_t^N = X^N) \mathbb{P}(\widetilde{\mathbf{X}}_t^N = X^N | \mathbf{W}^N) \\
&= \sum_{X^N=(x_1,\dots,x_N)} \prod_{n\in[N]} \left(x_n^\top G(\mathbf{w}_n,\theta)\widetilde{y}_{n,t}^N\right)\left(x_n^\top \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)\right) \\
&= \prod_{n\in[N]} \sum_{x_n} \left(x_n^\top G(\mathbf{w}_n,\theta)\widetilde{y}_{n,t}^N\right)\left(x_n^\top \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)\right) \\
&= \prod_{n\in[N]} \sum_{i=1}^{M} \left(G(\mathbf{w}_n,\theta)^{(i,\cdot)}\widetilde{y}_{n,t}^N\right) \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)^{(i)} \\
&= \prod_{n\in[N]} \left[\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)^\top G(\mathbf{w}_n,\theta)\right] \widetilde{y}_{n,t}^N,
\end{aligned}$$

showing that the observations are marginally distributed as categorical distributions with the desired parameters. We can now compute the posterior distribution using Bayes theorem:

$$\begin{aligned}
\mathbb{P}(\widetilde{\mathbf{X}}_t^N = \widetilde{X}_t^N | \mathbf{W}^N, \widetilde{\mathbf{Y}}_t^N = \widetilde{Y}_t^N) &= \frac{\mathbb{P}(\widetilde{\mathbf{Y}}_t^N = \widetilde{Y}_t^N | \mathbf{W}^N, \widetilde{\mathbf{X}}_t^N = \widetilde{X}_t^N)\mathbb{P}(\widetilde{\mathbf{X}}_t^N = \widetilde{X}_t^N | \mathbf{W}^N)}{\mathbb{P}(\widetilde{\mathbf{Y}}_t^N = \widetilde{Y}_t^N | \mathbf{W}^N)} \\
&= \prod_{n\in[N]} \frac{\left((\widetilde{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta)\widetilde{y}_{n,t}^N\right)\left((\widetilde{x}_{n,t}^N)^\top \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)\right)}{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^\top \widetilde{y}_{n,t}^N} \\
&= \prod_{n\in[N]} \frac{(\widetilde{x}_{n,t}^N)^\top \left(G(\mathbf{w}_n,\theta)\widetilde{y}_{n,t}^N \odot \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)\right)}{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^\top \widetilde{y}_{n,t}^N} \\
&= \prod_{n\in[N]} (\widetilde{x}_{n,t}^N)^\top \left[\left(G(\mathbf{w}_n,\theta)\widetilde{y}_{n,t}^N \odot \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)\right) \oslash \left(\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^\top \widetilde{y}_{n,t}^N\right)\right] \\
&= \prod_{n\in[N]} \sum_{i=1}^{M}\sum_{j=1}^{M} (\widetilde{x}_{n,t}^N)^{(i)} \frac{G(\mathbf{w}_n,\theta)^{(i,j)}(\widetilde{y}_{n,t}^N)^{(j)}\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)^{(i)}}{\sum_{k\in[M]} \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(k)}(\widetilde{y}_{n,t}^N)^{(k)}} \\
&= \prod_{n\in[N]} \sum_{i=1}^{M}\sum_{j=1}^{M} (\widetilde{x}_{n,t}^N)^{(i)} \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta)^{(i)} \frac{G(\mathbf{w}_n,\theta)^{(i,j)}}{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(j)}} (\widetilde{y}_{n,t}^N)^{(j)} \\
&= \prod_{n\in[N]} \left\{\boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n,\theta) \odot \left\{\left[G(\mathbf{w}_n,\theta) \oslash \left(1_M \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^\top\right)\right]\widetilde{y}_{n,t}^N\right\}\right\}^\top \widetilde{x}_{n,t}^N.
\end{aligned}$$

□

As a consequence of Proposition 2 and 3, we can see the CAL as an exact marginal likelihood

for the approximate model:

$$\begin{aligned}
&\widetilde{\mathbf{x}}_{n,0}^N|\mathbf{w}_n \sim \mathrm{Cat}\left(\cdot|p_0(\mathbf{w}_n,\theta)\right),\\
&\widetilde{\mathbf{x}}_{n,t}^N|\widetilde{\mathbf{X}}_{t-1}^N, \widetilde{\mathbf{Y}}_{1:t-1}^N, \mathbf{W}^N \sim \mathrm{Cat}\left(\cdot|\left[(\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\right]^\top\right),\\
&\widetilde{\mathbf{y}}_{n,t}^N|\widetilde{\mathbf{x}}_{n,t}^N, \mathbf{w}_n \sim \mathrm{Cat}\left(\cdot|\left[(\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n,\theta)\right]^\top\right),
\end{aligned}$$

and we can compute both its joint likelihood and marginal likelihood, see the next proposition.

Proposition 4. *Over a time horizon T, the marginal likelihood of $\widetilde{\mathbf{Y}}_{1:T}^N$ is given by:*

$$p(\widetilde{\mathbf{Y}}_{1:T}^N|\mathbf{W}^N,\theta) \coloneqq \prod_{t=1}^T \prod_{n\in[N]} Cat\left(\widetilde{\mathbf{y}}_{n,t}^N|\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)\right),$$

which is the marginal of:

$$\begin{aligned}
p(\widetilde{\mathbf{X}}_{0:T}^N, \widetilde{\mathbf{Y}}_{1:T}^N|\mathbf{W}^N,\theta) \coloneqq &\prod_{n\in[N]} (\widetilde{\mathbf{x}}_{n,t}^N)^\top p_0(\mathbf{w}_n,\theta) \prod_{t=1}^T \left[(\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\right]\\
&\cdot \prod_{t=1}^T \left[(\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}_{n,t}^N\right],
\end{aligned}$$

i.e. $\sum_{\widetilde{\mathbf{x}}_{0:T}^N} p(\widetilde{\mathbf{X}}_{0:T}^N, \widetilde{\mathbf{Y}}_{1:T}^N|\mathbf{W}^N,\theta) = p(\widetilde{\mathbf{Y}}_{1:T}^N|\mathbf{W}^N,\theta)$.

Proof. For ease of presentation throughout the proof we omit conditioning on $\mathbf{W}^N$ from the notation. We prove the statement of the proposition by induction on T and start by showing that $p(\widetilde{\mathbf{Y}}_1^N|\theta)$ satisfies the statement:

$$\begin{aligned}
p(\widetilde{\mathbf{Y}}_1^N|\theta) &= \prod_{n\in[N]} \boldsymbol{\mu}_{n,1}^N(\mathbf{w}_n,\theta)^\top \widetilde{\mathbf{y}}_{n,1}^N = \prod_{n\in[N]} \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta)^\top K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}_{n,1}^N\\
&= \prod_{n\in[N]} \sum_{i,j,k} \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta)^{(i)} K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(i,j)} G(\mathbf{w}_n,\theta)^{(j,k)} (\widetilde{\mathbf{y}}_{n,1}^N)^{(k)}\\
&= \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}_{n,0}^N, \widetilde{\mathbf{x}}_{n,1}^N} \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta)\right) \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}_{n,1}^N\right)\\
&\qquad \cdot \left((\widetilde{\mathbf{x}}_{n,1}^N)^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}_{n,1}^N\right)\\
&= \sum_{\widetilde{\mathbf{x}}_0^N, \widetilde{\mathbf{x}}_1^N} \prod_{n\in[N]} \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta)\right) \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}_{n,1}^N\right)\\
&\qquad \cdot \left((\widetilde{\mathbf{x}}_{n,1}^N)^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}_{n,1}^N\right)\\
&= \sum_{\widetilde{\mathbf{x}}_0^N, \widetilde{\mathbf{x}}_1^N} p(\widetilde{\mathbf{X}}_{0:1}^N, \widetilde{\mathbf{Y}}_1^N|\theta)
\end{aligned}$$

for $\widetilde{\mathbf{X}}_t^N = (\widetilde{\mathbf{x}}_{1,t}^N, \ldots, \widetilde{\mathbf{x}}_{N,t}^N)$ with $t = 0, 1$, which complete the proof for the first time step. From the above we also get:

$$\begin{aligned}\boldsymbol{\pi}_{n,1}^N(\mathbf{w}_n, \theta) = &\left[\sum_{\widetilde{\mathbf{x}}_{n,0}^N} \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta)\right) \odot \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\right)^\top\right] \\ &\oslash \left[\sum_{\widetilde{\mathbf{x}}_{n,0}^N, \widetilde{\mathbf{x}}_{n,1}^N} \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta)\right) \odot \left((\widetilde{\mathbf{x}}_{n,0}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_0^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\right)^\top\right] \\ &\odot \left(G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,1}^N\right).\end{aligned}$$

Now assume that the statement is valid for $T-1$:

$$\begin{aligned}p(\widetilde{\mathbf{Y}}_{1:T-1}^N|\theta) &= \sum_{\widetilde{\mathbf{X}}_{0:T-1}^N} \prod_{n\in[N]} \left[(\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta) \prod_{t=1}^{T-1} \left((\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\widetilde{\mathbf{x}}_{n,t}^N\right)\right. \\ &\qquad \left.\cdot \left((\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,t}^N\right)\right] \\ &= \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}_{n,0:T-1}^N} (\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta) \prod_{t=1}^{T-1} \left((\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\widetilde{\mathbf{x}}_{n,t}^N\right) \\ &\qquad \cdot \left((\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,t}^N\right),\end{aligned}$$

from which we also get:

$$\begin{aligned}\boldsymbol{\pi}_{n,T-1}^N(\mathbf{w}_n, \theta) = &\left\{\sum_{\widetilde{\mathbf{x}}_{n,0:T-2}^N} \left[(\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta) \prod_{t=1}^{T-2} \left((\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\widetilde{\mathbf{x}}_{n,t}^N\right)\right. \right. \\ &\left. \left.\cdot \left((\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,t}^N\right)\right] \left((\widetilde{\mathbf{x}}_{n,T-2}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{T-2}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\right)^\top\right\} \\ &\oslash \left[\sum_{\widetilde{\mathbf{x}}_{n,0:T-1}^N} (\widetilde{\mathbf{x}}_{n,0}^N)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta) \prod_{t=1}^{T-1} \left((\widetilde{\mathbf{x}}_{n,t-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\widetilde{\mathbf{x}}_{n,t}^N\right)\right. \\ &\left.\cdot \left((\widetilde{\mathbf{x}}_{n,t}^N)^\top G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,t}^N\right)\right] \odot G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,T-1}^N.\end{aligned}$$

Consider now time T:

$$\begin{aligned}p(\widetilde{\mathbf{Y}}_{1:T}^N|\theta) &= p(\widetilde{\mathbf{Y}}_{1:T-1}^N|\theta) \prod_{n\in[N]} \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^\top \widetilde{\mathbf{y}}_{n,T}^N \\ &= p(\widetilde{\mathbf{Y}}_{1:T-1}^N|\theta) \prod_{n\in[N]} \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,T}^N \\ &= p(\widetilde{\mathbf{Y}}_{1:T-1}^N|\theta) \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}_{n,T-1}^N, \widetilde{\mathbf{x}}_{n,T}^N} \left((\widetilde{\mathbf{x}}_{n,T-1}^N)^\top \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta)\right) \\ &\qquad \cdot \left((\widetilde{\mathbf{x}}_{n,T-1}^N)^\top K_{\widetilde{\boldsymbol{\eta}}_{t-1}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\widetilde{\mathbf{x}}_{n,T}^N\right) \left((\widetilde{\mathbf{x}}_{n,T}^N)^\top G(\mathbf{w}_n, \theta)\widetilde{\mathbf{y}}_{n,T}^N\right),\end{aligned}$$

from which we remark that:

$$
\begin{aligned}
&(\widetilde{\mathbf{x}}^N_{n,T-1})^\top \boldsymbol{\pi}^N_{n,T-1}(\mathbf{w}_n,\theta) \\
&= \Bigg\{ \sum_{\widetilde{\mathbf{x}}^N_{n,0:T-2}} \Bigg[(\widetilde{\mathbf{x}}^N_{n,0})^\top \boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta) \prod_{t=1}^{T-2} \left((\widetilde{\mathbf{x}}^N_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \widetilde{\mathbf{x}}^N_{n,t} \right) \\
&\qquad \cdot \left((\widetilde{\mathbf{x}}^N_{n,t})^\top G(\mathbf{w}_n,\theta) \widetilde{\mathbf{y}}^N_{n,t} \right) \Bigg] \left((\widetilde{\mathbf{x}}^N_{n,T-2})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{T-2}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \widetilde{\mathbf{x}}^N_{n,T-1} \right) \Bigg\} \\
&\quad / \Bigg[\sum_{\widetilde{\mathbf{x}}^N_{n,0:T-1}} (\widetilde{\mathbf{x}}^N_{n,0})^\top \boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta) \prod_{t=1}^{T-1} \left((\widetilde{\mathbf{x}}^N_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \widetilde{\mathbf{x}}^N_{n,t} \right) \\
&\qquad \cdot \left((\widetilde{\mathbf{x}}^N_{n,t})^\top G(\mathbf{w}_n,\theta) \widetilde{\mathbf{y}}^N_{n,t} \right) \Bigg] \\
&\quad \cdot (\widetilde{\mathbf{x}}^N_{n,T-1})^\top G(\mathbf{w}_n,\theta) \widetilde{\mathbf{y}}^N_{n,T-1}.
\end{aligned}
$$

As from our inductive hypothesis we have:

$$
\begin{aligned}
p(\widetilde{\mathbf{Y}}^N_{1:T-1}|\theta) = \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}^N_{n,0:T-1}} &(\widetilde{\mathbf{x}}^N_{n,0})^\top \boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta) \prod_{t=1}^{T-1} \left((\widetilde{\mathbf{x}}^N_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \widetilde{\mathbf{x}}^N_{n,t} \right) \\
&\cdot \left((\widetilde{\mathbf{x}}^N_{n,t})^\top G(\mathbf{w}_n,\theta) \widetilde{\mathbf{y}}^N_{n,t} \right),
\end{aligned}
$$

which is the denominator in $(\widetilde{\mathbf{x}}^N_{n,T-1})^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)$, we can conclude:

$$
\begin{aligned}
p(\widetilde{\mathbf{Y}}^N_{1:t}|\theta) &= p(\widetilde{\mathbf{Y}}^N_{1:T-1}|\theta) \\
&\quad \cdot \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}^N_{n,T-1},\widetilde{\mathbf{x}}^N_{n,T}} \left((\widetilde{\mathbf{x}}^N_{n,T-1})^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right) \left((\widetilde{\mathbf{x}}^N_{n,T-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}^N_{n,T}\right) \\
&\qquad \cdot \left((\widetilde{\mathbf{x}}^N_{n,T})^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}^N_{n,T}\right) \\
&= \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}^N_{n,T-1},\widetilde{\mathbf{x}}^N_{n,T}} \sum_{\widetilde{\mathbf{x}}^N_{n,0:T-2}} (\widetilde{\mathbf{x}}^N_{n,0})^\top \boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta) \\
&\qquad \cdot \prod_{t=1}^{T-2} \left((\widetilde{\mathbf{x}}^N_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}^N_{n,t}\right) \left((\widetilde{\mathbf{x}}^N_{n,t})^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}^N_{n,t}\right) \\
&\qquad \cdot \left((\widetilde{\mathbf{x}}^N_{n,T-2})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{T-2}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}^N_{n,T-1}\right) \left((\widetilde{\mathbf{x}}^N_{n,T-1})^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}^N_{n,T-1}\right) \\
&\qquad \cdot \left((\widetilde{\mathbf{x}}^N_{n,T-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}^N_{n,T}\right) \left((\widetilde{\mathbf{x}}^N_{n,T})^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}^N_{n,T}\right) \\
&= \prod_{n\in[N]} \sum_{\widetilde{\mathbf{x}}^N_{n,0:T}} (\widetilde{\mathbf{x}}^N_{n,0})^\top \boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta) \\
&\qquad \cdot \prod_{t=1}^{T} \left((\widetilde{\mathbf{x}}^N_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\widetilde{\mathbf{x}}^N_{n,t}\right) \left((\widetilde{\mathbf{x}}^N_{n,t})^\top G(\mathbf{w}_n,\theta)\widetilde{\mathbf{y}}^N_{n,t}\right) \\
&= \sum_{\widetilde{\mathbf{x}}^N_{0:T}} p(\widetilde{\mathbf{X}}^N_{0:T}, \widetilde{\mathbf{Y}}^N_{1:T}|\theta),
\end{aligned}
$$

which completes the proof. □

C Consistency of the maximum CAL estimator

With the definition:

$$\ell^N_t(\theta) \coloneqq \sum_{n\in[N]} \log\left[\left(\mathbf{y}^N_{n,t}\right)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right], \tag{7}$$

our ultimate goal is to prove Theorem 33, which establishes consistency of the maximum CAL estimator:

$$\hat{\theta}^N \coloneqq \operatorname*{argmax}_{\theta\in\Theta} \sum_{t=1}^{T} \frac{1}{N} \ell^N_t(\theta),$$

in the sense that $\hat{\theta}^N$ converges to $\Theta^\star$ as $N\to\infty$, $\mathbb{P}$-almost surely, where $\Theta^\star \subset \Theta$ is a set of parameter values which are, in a sense to be made precise, equivalent to the DGP $\theta^\star$.

The main steps are:

1. prove almost sure pointwise in θ convergence of $\sum_{t=1}^{T} \frac{1}{N}\ell_t^N(\theta) - \frac{1}{N}\ell_t^N(\theta^\star)$ to a contrast function $\mathcal{C}(\theta, \theta^\star)$, this is the subject of Theorem 29;

2. Use stochastic equi-continuity Andrews (1992) to prove that the almost sure convergence $\sum_{t=1}^{T} \frac{1}{N}\ell_t^N(\theta) - \frac{1}{N}\ell_t^N(\theta^\star) \to \mathcal{C}(\theta, \theta^\star)$ is uniform in θ, this is the subject of Lemma 32;

3. prove Theorem 33, and so the convergence of the maximum CAL estimator to a set of maximizers $\Theta^\star$;

4. Characterize the set of maximizers of the contrast function $\Theta^\star$ and prove that $\Theta^\star$ contains $\theta^\star$; this is the subject of Theorem 29 and Lemma 35.

Step 1. is by far the most complicated of the three. It involves proving L^4 bounds for averages across the population in the data-generating process (Section C.2), averages of various quantities computed in the CAL algorithm (Section C.3), and comparison to averages across what we call the *saturated processes* and *saturated CAL algorithm* (Section C.4), which are processes we construct for purposes of our proofs in which members of the population are statistically decoupled. All of these ingredients are then combined to establish convergence to the contrast function (Theorem 29).

C.1 Preliminaries

C.1.1 L^4 bound for conditionally independent random variables

We state and prove a result from Whitehouse et al. (2023), which is useful to find L^4 bounds of averages of random variables that are conditionally independent, bounded, and mean zero. This is going to be one of the main building blocks in our proof strategy.

Lemma 5. *Consider a collection of random variable $\boldsymbol{\delta}_n$ with $n \in [N]$. Assume that given a filtration $\mathcal{F}$ the random variables $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_N$ are conditionally independent, bounded by a constant $B < \infty$, i.e. $|\boldsymbol{\delta}_n| \leq B$ almost surely, and satisfy $\mathbb{E}[\boldsymbol{\delta}_n|\mathcal{F}] = 0$, then:*

$$\left\| \frac{1}{N} \sum_{n\in[N]} \boldsymbol{\delta}_n \right\|_4 \leq B\sqrt[4]{6} N^{-\frac{1}{2}}.$$

Proof. From the Multinomial theorem we can see that:

$$\left(\sum_{n\in[N]} \boldsymbol{\delta}_n \right)^4 = \sum_{k_1,\dots,k_N \in \mathbb{N}: k_1+\cdots+k_N=4} \binom{4}{k_1, \dots, k_N} \prod_{n\in[N]} (\boldsymbol{\delta}_n)^{k_n},$$

hence if we compute expectations with respect to the filtration we have:

$$\mathbb{E}\left[\left(\sum_{n\in[N]} \boldsymbol{\delta}_n \right)^4 \middle| \mathcal{F} \right] = \sum_{k_1,\dots,k_N \in \mathbb{N}: k_1+\cdots+k_N=4} \binom{4}{k_1, \dots, k_N} \prod_{n\in[N]} \mathbb{E}\left[(\boldsymbol{\delta}_n)^{k_n} \middle| \mathcal{F} \right],$$

because of the conditional independence assumption. Remark that we can combine a maximum of four terms as we are considering a power of 4 and we are constrained to $\sum_{n\in[N]} k_n = 4$, meaning that we have only these possible combinations:

- $k_{i_1} + k_{i_2} + k_{i_3} + k_{i_4} = 4$;
- $k_{i_1} + k_{i_2} + k_{i_3} = 4$;
- $k_{i_1} + k_{i_2} = 4$;
- $k_{i_1} = 4$;

with all the other k's being zero. Given that $\mathbb{E}[\boldsymbol{\delta}_n|\mathcal{F}] = 0$ then all the combinations which involve a $k_i = 1$ can be safely removed, we then end up with:

$$\begin{aligned}
\mathbb{E}\left[\left(\sum_{n\in[N]} \boldsymbol{\delta}_n\right)^4 \middle| \mathcal{F}\right] &= \sum_{n\in[N]} \mathbb{E}\left[(\boldsymbol{\delta}_n)^4 \middle| \mathcal{F}\right] + \binom{4}{2,2} \sum_{n,n'\in[N],n\neq n'} \mathbb{E}\left[(\boldsymbol{\delta}_n)^2|\mathcal{F}\right] \mathbb{E}\left[(\boldsymbol{\delta}_{n'})^2|\mathcal{F}\right] \\
&\leq \sum_{n\in[N]} (B)^4 + 6 \sum_{n,n'\in[N],n\neq n'} (B)^2(B)^2 \\
&= NB^4 + 6N(N-1)B^4 \\
&= B^4(N + 6N^2 - 6N) \leq 6B^4N^2,
\end{aligned}$$

from which the statement of the Lemma follows by the tower rule, dividing by N^4 and exponentiating by $\frac{1}{4}$. □

C.1.2 Checking the CAL is almost surely well-defined

If $(\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) = 0$, then (7) would evaluate to $\log(0)$. The aim of this section is to prove that the CAL is almost surely well-defined, in the sense that $(\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) = 0$ happens with zero probability no matter the values of N and θ.

The main result is Theorem 10, and its proof builds upon propositions 6 - 9 below, which exploit the recursive nature of the CAL algorithm and the data-generating process.

Proposition 6. *Under Assumption 8, for any $n \in [N]$ if there exists $\theta \in \Theta, i \in [M]$ such that $p_0(w_n, \theta)^{(i)} = 0$ then $(\mathbf{x}^N_{n,0})^{(i)} = 0$, $\mathbb{P}(\cdot|\mathbf{W}^N = W^N)$-almost surely.*

Proof. Under Assumption 8 we have that $p_0(w_n, \theta)^{(i)} = 0$ implies $p_0(w_n, \theta^\star)^{(i)} = 0$, hence:

$$\mathbb{P}\left((\mathbf{x}^N_{n,0})^{(i)} = 0|\mathbf{W}^N = W^N\right) = p_0(w_n, \theta^\star)^{(i)} = 0,$$

which concludes the proof. □

Proposition 7. *Under Assumption 8, for any $t \geq 1$ and $n \in [N]$ if there exists $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^N_{n,t-1}(w_n, \theta)^{(i)} = 0$ implies $(\mathbf{x}^N_{n,t-1})^{(i)} = 0$, $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$-almost surely, then, there exists $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^N_{n,t|t-1}(w_n, \theta)^{(i)} = 0$ implies $(\mathbf{x}^N_{n,t})^{(i)} = 0$, $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$-almost surely.*

Proof. Note that for $\theta \in \Theta, i \in [M]$ the following are equivalent:

$$\begin{aligned}\boldsymbol{\pi}^N_{n,t|t-1}(w_n,\theta)^{(i)} = 0 &\iff \sum_j \boldsymbol{\pi}^N_{n,t-1}(w_n,\theta)^{(j)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(w_n,\theta)}(w_n,\theta)^{(j,i)} = 0 \\ &\iff \forall j \in [M] \quad \boldsymbol{\pi}^N_{n,t-1}(w_n,\theta)^{(j)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(w_n,\theta)}(w_n,\theta)^{(j,i)} = 0.\end{aligned}$$

We then have that for any j either:

1. $\boldsymbol{\pi}^N_{n,t-1}(w_n,\theta)^{(j)} = 0$ which implies $(\mathbf{x}^N_{n,t-1})^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$ by assumption, or
2. $K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(w_n,\theta)}(w_n,\theta)^{(j,i)} = 0$, which implies that there exists $\eta = \widetilde{\boldsymbol{\eta}}^N_{t-1}(w_n,\theta)$ such that $K_\eta(w_n,\theta)^{(j,i)} = 0$. Then by Assumption 8 we have that $K_\eta(w_n,\theta)^{(j,i)} = 0$ for all $\eta \in [0,C]$ and $\theta \in \Theta$ meaning $K_{\boldsymbol{\eta}^N_{t-1}(w_n,\theta^\star)}(w_n,\theta^\star)^{(j,i)} = 0$, $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$-almost surely as it holds for almost any realization of $\mathbf{X}^N_{t-1} \in \mathbb{O}^N_M \subset \Delta^N_M$, where we remark that $\mathbf{X}^N_{t-1}$ appears in $\boldsymbol{\eta}^N_{t-1}(w_n,\theta^\star) = \eta^N(w_n,\theta^\star,W^N,\mathbf{X}^N_{t-1})$.

Now given that:

$$\mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 0|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 1 - \mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right),$$

and by defining $\mathbf{X}^{N\backslash n}_{t-1}$ as $\mathbf{X}^N_{t-1}$ with the n-th individual removed, we can notice that:

$$\begin{aligned}&\mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &= \sum_{X^{N\backslash n}_{t-1}} \mathbb{P}\left(\mathbf{X}^{N\backslash n}_{t-1} = X^{N\backslash n}_{t-1}|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &\quad \sum_{x^N_{n,t-1}} \mathbb{P}\left(\mathbf{x}^N_{n,t-1} = x^N_{n,t-1}|\mathbf{W}^N = W^N, \mathbf{X}^{N\backslash n}_{t-1} = X^{N\backslash n}_{t-1}, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &\quad \cdot \mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{X}^N_{t-1} = X^N_{t-1}, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &= \sum_{X^{N\backslash n}_{t-1}} \mathbb{P}\left(\mathbf{X}^{N\backslash n}_{t-1} = X^{N\backslash n}_{t-1}|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &\quad \sum_j \mathbb{P}\left((\mathbf{x}^N_{n,t-1})^{(j)} = 1|\mathbf{W}^N = W^N, \mathbf{X}^{N\backslash n}_{t-1} = X^{N\backslash n}_{t-1}, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) K_{\boldsymbol{\eta}^N_{t-1}(w_n,\theta^\star)}(w_n,\theta^\star)^{(j,i)} \\ &= 0,\end{aligned}$$

where the last step follows from the fact that for all j we either have $(\mathbf{x}^N_{n,t-1})^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$, or $K_\eta(w_n,\theta^\star)^{(j,i)} = 0$ for any $\eta \in [0,C]$ and so $K_{\boldsymbol{\eta}^N_{t-1}(w_n,\theta^\star)}(w_n,\theta^\star)^{(j,i)} = 0$ for any X^N_{t-1}. As $\mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 0$ we can conclude $\mathbb{P}\left((\mathbf{x}^N_{n,t})^{(i)} = 0|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 1$, which concludes the proof. □

Proposition 8. *Under Assumption 8, for any $t \geq 1$ and $n \in [N]$ if there exists $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^N_{n,t|t-1}(w_n, \theta)^{(i)} = 0$ implies $(\mathbf{x}^N_{n,t})^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$, then:*

- *there exists $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\mu}^N_{n,t}(w_n, \theta)^{(i)} = 0$ implies $(\mathbf{y}^N_{n,t})^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$;*
- *there exists $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^N_{n,t}(w_n, \theta)^{(i)} = 0$ implies $(\mathbf{x}^N_{n,t})^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t} = Y^N_{1:t})$.*

Proof. Let us start with the results regarding the observations. Note that for $\theta \in \Theta, i \in [M]$ the following are equivalent:

$$\begin{aligned}\boldsymbol{\mu}^N_{n,t}(w_n, \theta)^{(i)} = 0 &\iff \sum_j \boldsymbol{\pi}^N_{n,t|t-1}(w_n, \theta)^{(j)} G(w_n, \theta)^{(j,i)} = 0 \\ &\iff \forall j \in [M] \quad \boldsymbol{\pi}^N_{n,t|t-1}(w_n, \theta)^{(j)} G(w_n, \theta)^{(j,i)} = 0.\end{aligned}$$

We then have that for all j either:

1. $\boldsymbol{\pi}^N_{n,t|t-1}(w_n, \theta)^{(j)} = 0$ which implies $(\mathbf{x}^N_{n,t})^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$ by assumption, or
2. $G(w_n, \theta)^{(j,i)} = 0$ which implies $G(w_n, \theta^\star)^{(j,i)} = 0$ because of Assumption 8.

Now given that:

$$\mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 0|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 1 - \mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right),$$

and:

$$\begin{aligned}&\mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &= \sum_{x^N_{n,t}} \mathbb{P}\left(\mathbf{x}^N_{n,t} = x^N_{n,t}|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &\quad \cdot \mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{x}^N_{n,t} = x^N_{n,t}, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) \\ &= \sum_j \mathbb{P}\left((\mathbf{x}^N_{n,t})^{(j)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) G(w_n, \theta^\star)^{(j,i)} = 0,\end{aligned}$$

where the last step follows from the fact that for all j we either have $(\mathbf{x}^N_{n,t-1})^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1})$ or $G(w_n, \theta^\star)^{(j,i)} = 0$. As

$$\mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 1|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 0,$$

we can conclude

$$\mathbb{P}\left((\mathbf{y}^N_{n,t})^{(i)} = 0|\mathbf{W}^N = W^N, \mathbf{Y}^N_{1:t-1} = Y^N_{1:t-1}\right) = 1,$$

which concludes the proof of the first part. As a consequence if there exists $j \in [M]$ such that $\mathbb{P}((\mathbf{y}_{n,t}^N)^{(j)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N) > 0$ then $\boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)} \neq 0$, meaning that

$$\mathbb{P}\left(\sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)} \neq 0 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N\right) = 1. \tag{8}$$

Consider now $\boldsymbol{\pi}_{n,t}^N(w_n, \theta)^{(i)}$, and observe that:

$$\begin{aligned}
&\boldsymbol{\pi}_{n,t}^N(w_n, \theta)^{(i)} = 0 \iff \boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} \frac{\sum_j G(w_n, \theta)^{(i,j)} (\mathbf{y}_{n,t}^N)^{(j)}}{\sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)}} = 0 \\
&\iff \boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} \sum_j G(w_n, \theta)^{(i,j)} (\mathbf{y}_{n,t}^N)^{(j)} = 0 \text{ and } \sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)} \neq 0.
\end{aligned}$$

From the (8), we know that under $\mathbb{P}(\cdot | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N)$ we have that the denominator $\sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)}$ is almost surely different from 0 hence:

$$\begin{aligned}
&\mathbb{P}\Bigg(\boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} \sum_j G(w_n, \theta)^{(i,j)} (\mathbf{y}_{n,t}^N)^{(j)} = 0 \\
&\qquad \text{and } \sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)} \neq 0 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N\Bigg) \\
&= \mathbb{P}\left(\sum_j (\mathbf{y}_{n,t}^N)^{(j)} \boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(j)} = 0 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N\right),
\end{aligned}$$

as we are considering an intersection with an almost sure event.

We then just need to prove $\boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} \sum_j G(w_n, \theta)^{(i,j)} (\mathbf{y}_{n,t}^N)^{(j)} = 0$ almost surely. We have either:

- $\boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} = 0$, implying $(\mathbf{x}_{n,t}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N)$ because we consider this statement to be true, or
- $\sum_j G(w_n, \theta)^{(i,j)} (\mathbf{y}_{n,t}^N)^{(j)} = 0$, which tells us that there exists $k \in [M]$ such that $\mathbb{P}\left((\mathbf{y}_{n,t}^N)^{(k)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N\right) > 0$ and $G(w_n, \theta)^{(i,k)} = 0$, note that under $\mathbb{P}(\cdot | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t}^N = Y_{1:t}^N)$ we know k as we are conditioning on $\mathbf{y}_{n,t}^N$.

Now given that:

$$\mathbb{P}((\mathbf{x}_{n,t}^N)^{(i)} = 0 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t}^N = Y_{1:t}^N) = 1 - \mathbb{P}((\mathbf{x}_{n,t}^N)^{(i)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t}^N = Y_{1:t}^N),$$

we can notice that:

$$\begin{aligned}
&\mathbb{P}((\mathbf{x}_{n,t}^N)^{(i)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t}^N = Y_{1:t}^N) \\
&\propto \mathbb{P}((\mathbf{x}_{n,t}^N)^{(i)} = 1, (\mathbf{y}_{n,t}^N)^{(k)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N) \\
&= G(w_n, \theta)^{(i,k)} \mathbb{P}((\mathbf{x}_{n,t}^N)^{(i)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N) = 0,
\end{aligned}$$

because we have either $(\mathbf{x}_{n,t}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N)$ or $G(w_n, \theta)^{(i,k)} = 0$, which concludes the proof. □

We can now combine Proposition 6, Proposition7, and Proposition 8 in the following proposition.

Proposition 9. *Under Assumption 8, for any $t \geq 1$, $n \in [N]$, the following hold:*

- *if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}_{n,t|t-1}^N(w_n, \theta)^{(i)} = 0$, then $(\mathbf{x}_{n,t}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N)$;*
- *if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\mu}_{n,t}^N(w_n, \theta)^{(i)} = 0$, then $(\mathbf{y}_{n,t}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N)$;*
- *if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}_{n,t}^N(w_n, \theta)^{(i)} = 0$, then $(\mathbf{x}_{n,t}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N, \mathbf{Y}_{1:t}^N = Y_{1:t}^N)$.*

Proof. Suppose that the third statement is true at time $t-1$, which is valid at $t-1=0$ because of Proposition 6, which guarantees that if $\exists \theta \in \Theta, i \in [M] : \boldsymbol{\pi}_{n,0}^N(w_n, \theta)^{(i)} = 0 \implies (\mathbf{x}_{n,0}^N)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{W}^N = W^N)$. We can in turn apply Proposition 7 to prove the first statement and Proposition 8 to prove the second and the third for t. As the third statement is our inductive hypothesis at the next time step t we can then close the induction and conclude that the three statements are valid for an arbitrary t. □

We finally prove the main result.

Theorem 10. *Under assumptions 7,8, for any $N \in \mathbb{N}, t \geq 1, n \in [N]$ and $\theta \in \Theta$ we have that $\sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)} \neq 0$ almost surely in $\mathbb{P}$.*

Proof. We can see that:

$$\begin{aligned}
&\mathbb{P}\left(\exists \theta \in \Theta, t \geq 1, n \in [N] : \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)} = 0 | \mathbf{W}^N = W^N\right) \\
&= \mathbb{P}\left(\exists \theta \in \Theta, t \geq 1, n \in [N] : \quad \forall i \in [M] \quad \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)} = 0 | \mathbf{W}^N = W^N\right) \\
&= \mathbb{P}\left(\exists \theta \in \Theta, t \geq 1, n \in [N], k \in [M] : \quad \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(k)} = 0 \text{ and } (\mathbf{y}_{n,t}^N)^{(k)} = 1 | \mathbf{W}^N = W^N\right),
\end{aligned}$$

where the second equality uses the fact that $\mathbf{y}_{n,t}^N$ is a one-hot encoding vector, i.e. the kth component is 1 while the others are 0. Moreover:

$$\begin{aligned}
&\mathbb{P}\left(\exists \theta \in \Theta, t \geq 1, n \in [N] : \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)} = 0 \middle| \mathbf{W}^N = W^N\right) \\
&= \sum_{Y_{1:t-1}^N} \mathbb{P}\left(\exists \theta \in \Theta, t \geq 1, n \in [N], k \in [M] : \right. \\
&\qquad \left. \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)^{(k)} = 0 \text{ and } (\mathbf{y}_{n,t}^N)^{(k)} = 1 | \mathbf{W}^N = W^N, \mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N\right) \\
&\qquad \cdot \mathbb{P}\left(\mathbf{Y}_{1:t-1}^N = Y_{1:t-1}^N | \mathbf{W}^N = W^N\right) = 0,
\end{aligned}$$

as from the second statement of Proposition 9 we know that event $\left\{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(k)}=0\right\}\cap\left\{(\mathbf{y}_{n,t}^N)^{(k)}=1\right\}$ is probability zero conditionally on $\mathbf{W}^N=W^N, \mathbf{Y}_{1:t-1}^N=Y_{1:t-1}^N$, indeed:

$$\begin{aligned}
&\mathbb{P}\left(\exists\theta\in\Theta, t\geq 1, n\in[N], k\in[M]:\right.\\
&\qquad\left.\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(k)}=0 \text{ and } (\mathbf{y}_{n,t}^N)^{(k)}=1 | \mathbf{W}^N=W^N, \mathbf{Y}_{1:t-1}^N=Y_{1:t-1}^N\right)\\
&=1-\mathbb{P}\left(\forall\theta\in\Theta, t\geq 1, n\in[N], k\in[M]:\right.\\
&\qquad\left.\text{if } \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(k)}=0 \text{ then } (\mathbf{y}_{n,t}^N)^{(k)}=0 | \mathbf{W}^N=W^N, \mathbf{Y}_{1:t-1}^N=Y_{1:t-1}^N\right)\\
&=0.
\end{aligned}$$

We can then conclude the proof of the first statement as:

$$\begin{aligned}
&\mathbb{P}\left(\forall\theta\in\Theta, t\geq 1, n\in[N] \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)}\neq 0 \,\middle|\, \mathbf{W}^N=W^N\right)\\
&=1-\mathbb{P}\left(\exists\theta\in\Theta, t\geq 1, n\in[N]: \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)}=0 \,\middle|\, \mathbf{W}^N=W^N\right)=1.
\end{aligned}$$

To conclude the proof we need:

$$\mathbb{P}\left(\forall\theta\in\Theta, t\geq 1, n\in[N] \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)}\neq 0\right)=1,$$

which can be proven by observing that:

$$\begin{aligned}
&\mathbb{P}\left(\forall\theta\in\Theta, t\geq 1, n\in[N] \quad \sum_i \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)}\neq 0\right)\\
&=\int \mathbb{P}\left(\forall\theta\in\Theta, t\geq 1, n\in[N] \quad \sum_i \boldsymbol{\mu}_{n,t}^N(w_n,\theta)^{(i)}(\mathbf{y}_{n,t}^N)^{(i)}\neq 0 \,\middle|\, \mathbf{W}^N=W^N\right)\\
&\qquad\Gamma(dw_1)\ldots\Gamma(dw_N)\\
&=1,
\end{aligned}$$

where we applied Assumption 7 and where the last step follows from what we have just proven. □

C.1.3 Checking the CAL is almost surely bounded

In this section, we want to prove that all the non-zero elements of $\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)$ are almost surely bounded below by a quantity $m_t>0$ that does not depend on N, this will be put to use in establishing L^4 bounds in Section C.3.

Proposition 11. *Under assumptions 6,8,10, for $t\geq 1$ there exists $m_t>0$ such that for any $N\in\mathbb{N}$ and $n\in[N]$ we have:*

$$\mathbb{P}\left(\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)^{(i)}\geq m_t \quad \forall i\in \boldsymbol{supp}\left(\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)\right)\right)=1 \quad \forall\theta\in\Theta.$$

Proof. For a fixed $N \in \mathbb{N}$, consider the following inductive hypothesis. There exists $\bar{m}_{t-1} > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_{t-1} \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right)\right) = 1 \quad \forall n \in [N], \theta \in \Theta.$$

We start by proving that the inductive hypothesis is true when $t-1=0$. From Assumption 6 we have that $p_0(w,\theta)$ is continuous in w,θ and both $\mathbb{W}$ and Θ are compact, we then get from Weierstrass theorem that there exists a minimum m_0 such that for any realization of $\mathbf{W}^N$ and for any $i \in \mathbf{supp}\left(p_0(\mathbf{w}_n,\theta)^{(i)}\right)$:

$$\boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta)^{(i)} = p_0(\mathbf{w}_n,\theta)^{(i)} \geq \min_{w\in\mathbb{W},\theta\in\Theta} \min_{j\in\mathbf{supp}\left(p_0(w,\theta)^{(j)}\right)} p_0(w,\theta)^{(j)} =: m_0,$$

with $m_0 > 0$ as we are considering a minimum over $j \in \mathbf{supp}\left(p_0(w,\theta)^{(j)}\right)$ which excludes all the zeros. As m_0 does not depend on $\mathbf{W}^N$ we conclude that there exists $m_0 > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta)^{(i)} \geq m_0 \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,0}(\mathbf{w}_n,\theta)\right)\right) = 1 \quad \forall n \in [N], \theta \in \Theta.$$

Let us now work on a general time step t. For $i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right)$,

$$\begin{aligned}
\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(i)} &= \sum_j \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^{(j)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} \\
&= \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right)} \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^{(j)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} \\
&\geq \bar{m}_{t-1} \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)},
\end{aligned}$$

where the inequality holds $\mathbb{P}$-almost surely by the inductive hypothesis. Several other inequalities in the remainder of the proof hold $\mathbb{P}$-almost surely, but to avoid repetition we do not state this explicitly. As:

$$i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right) \iff \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} \neq 0,$$

we can conclude that there exists at least one component of $K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(\cdot,i)}$ which is different from zero, hence:

$$\sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)\right)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} \geq \min_{j\in\mathbf{supp}\left(K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(\cdot,i)}\right)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} > 0. \quad (9)$$

Because of Assumption 8 we have:

$$K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} = 0 \iff K_\eta(\mathbf{w}_n,\theta)^{(j,i)} = 0 \quad \forall \eta \in [0,C],$$

meaning that

$$\min_{j\in\mathbf{supp}\left(K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(\cdot,i)}\right)} K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(j,i)} \geq \min_{\eta\in[0,C]}\min_{j\in\mathbf{supp}\left(K_\eta(\mathbf{w}_n,\theta)^{(\cdot,i)}\right)} K_\eta(\mathbf{w}_n,\theta)^{(j,i)}.$$

We can then conclude:

$$\begin{aligned}\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(i)} &\geq \bar{m}_{t-1}\min_{\theta\in\Theta,w\in\mathbb{W},\eta\in[0,C]}\min_{j\in\mathbf{supp}\left(K_\eta(w,\theta)^{(\cdot,i)}\right)} K_\eta(w,\theta)^{(j,i)}\\ &\geq \bar{m}_{t-1}\min_{\theta\in\Theta,w\in\mathbb{W},\eta\in[0,C]}\min_{(i,j)\in\mathbf{supp}(K_\eta(w,\theta))} K_\eta(w,\theta)^{(j,i)}.\end{aligned}$$

Note that $K_\eta(w,\theta)$ is continuous in η because of Assumption 10 and also in w,θ because of Assumption 6. Furthermore, the domain of η is $[0,C]$, which does not depend on N and it is compact. Additionally, $\mathbb{W},\Theta$ are compact by Assumption 6. Hence we can conclude by the Weirstrass theorem that there exist $\eta_{min}\in[0,C]$, $w_{min}\in\mathbb{W}$, $(i,j)\in\mathbf{supp}\left(K_\eta(w,\theta)\right)$, and $\theta_{min}\in\Theta$ such that:

$$m_K \coloneqq K_{\eta_{min}}(w_{min},\theta_{min})^{(j,i)} = \min_{\theta\in\Theta,w\in\mathbb{W},\eta\in[0,C]}\min_{(i,j)\in\mathbf{supp}(K_\eta(w,\theta))} K_\eta(w,\theta)^{(j,i)} > 0,$$

where strict positivity follows from the observation made in Equation (9). Hence:

$$\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_{t-1}m_K > 0,$$

where the lower bounding constants do not depend on $\mathbf{W}^N,\mathbf{Y}^N_{1:t-1}$, θ or N. Therefore we can conclude that there exist $\bar{m}_{t-1},m_K>0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(i)}\geq\bar{m}_{t-1}m_K \quad \forall i\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right)\right)=1 \quad \forall\theta\in\Theta.$$

Similarly, for $i\in\mathbf{supp}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right)$,

$$\begin{aligned}\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} &= \sum_j \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(j)}G(\mathbf{w}_n,\theta)^{(j,i)}\\ &\geq \bar{m}_{t-1}m_K \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right)} G(\mathbf{w}_n,\theta)^{(j,i)},\end{aligned}$$

where the inequality follows from what we have proven above. Moreover:

$$i\in\mathbf{supp}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right) \iff \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right)} G(\mathbf{w}_n,\theta)^{(j,i)}\neq 0,$$

meaning that there exists at least one component of $G(\mathbf{w}_n,\theta)^{(\cdot,i)}$ which is different from zero. Hence following the same reasoning as above:

$$\begin{aligned}\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} &\geq \bar{m}_{t-1}m_K\min_{w\in\mathbb{W},\theta\in\Theta}\min_{j\in\mathbf{supp}\left(G(w,\theta)^{(\cdot,i)}\right)} G(w,\theta)^{(j,i)}\\ &\geq \bar{m}_{t-1}m_K\min_{w\in\mathbb{W},\theta\in\Theta}\min_{(i,j)\in\mathbf{supp}(G(w,\theta))} G(w,\theta)^{(j,i)},\end{aligned}$$

and as $G(w,\theta)$ is continuous in w,θ because of Assumption 6 and $\mathbb{W},\Theta$ are compact because of Assumption 6 we can conclude by Weirstrass theorem that there exists a minimum m_G such that:

$$\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G > 0,$$

where the strict inequality follows from considering a minimum on the support of matrix G.

We conclude that there exist $\bar{m}_{t-1}, m_K, m_G > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right)\right) = 1 \quad \forall \theta \in \Theta.$$

Consider $i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)\right)$ then:

$$\begin{aligned}
\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} &= \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^{(i)} \frac{\sum_j G(\mathbf{w}_n,\theta)^{(i,j)} (\mathbf{y}^N_{n,t})^{(j)}}{\sum_j (\mathbf{y}^N_{n,t})^{(j)} \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(j)}} \\
&\geq \bar{m}_{t-1} m_K \sum_j G(\mathbf{w}_n,\theta)^{(i,j)} (\mathbf{y}^N_{n,t})^{(j)},
\end{aligned}$$

where the inequality follows from what we have proven above, from $\sum_j (\mathbf{y}^N_{n,t})^{(j)} \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(j)} \leq 1$ by definition and $\sum_j (\mathbf{y}^N_{n,t})^{(j)} \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(j)} \neq 0$ $\mathbb{P}$-almost surely because of Theorem 10. We now observe that:

$$i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)\right) \iff \sum_j G(\mathbf{w}_n,\theta)^{(i,j)} (\mathbf{y}^N_{n,t})^{(j)} \neq 0,$$

meaning that there is at least one element of the $G(\mathbf{w}_n,\theta)^{(i,\cdot)}$ which is different from zero. Following the same reasoning of $\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)}$ we can conclude:

$$\begin{aligned}
\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} &\geq \bar{m}_{t-1} m_K \min_{j \in \mathbf{supp}\left(G(\mathbf{w}_n,\theta)^{(i,\cdot)}\right)} G(\mathbf{w}_n,\theta)^{(i,j)} \\
&\geq \bar{m}_{t-1} m_K \min_{w \in \mathbb{W}, \theta \in \Theta} \min_{(i,j) \in \mathbf{supp}(G(w,\theta))} G(w,\theta)^{(i,j)} \\
&\geq \bar{m}_{t-1} m_K m_G > 0.
\end{aligned}$$

We can conclude that there exist constants $\bar{m}_{t-1}, m_K, m_G > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)\right)\right) = 1, \quad \forall \theta \in \Theta.$$

We can then set $\bar{m}_t \coloneqq \bar{m}_{t-1} m_K m_G$ and conclude that there exists $\bar{m}_t > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} \geq \bar{m}_t \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta)\right)\right) = 1, \quad \forall \theta \in \Theta,$$

which closes the induction, meaning that the above equality holds for an arbitrary t. As a consequence, we also have that for any $t \geq 1$ there exists $m_t > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)^{(i)} \geq m_t \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right)\right) = 1, \quad \forall \theta \in \Theta,$$

with $m_t \coloneqq \bar{m}_{t-1} m_K m_G$, which concludes the proof.

□

C.2 L^4 bounds for averages of the data-generating process

In Section C.2 we establish L^4 bounds for averages across the population of disease states and observations.

Initial condition. We start by considering the population at $t = 0$.

Proposition 12. *Under Assumption 7, there exists $\alpha_0 \geq 0$ such that for any bounded function $f : \mathbb{W} \to [0, B]^M$, i.e. $\|f\|_\infty \leq B < \infty$,*

$$\left\| \frac{1}{N} \sum_{n \in [N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N - \int f(w)^\top p_0(w, \theta^\star) \, \Gamma(dw) \right\|_4 \leq 2B \sqrt[4]{6} N^{-\frac{1}{2}} \alpha_0.$$

Moreover, there exists $\widetilde{\alpha}_0 \geq 0$ such that for any bounded function $f : \mathbb{W} \to [0, B]^M$,

$$\left\| \frac{1}{N} \sum_{n \in [N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) \right\|_4 \leq 2B \sqrt[4]{6} N^{-\frac{1}{2}} \widetilde{\alpha}_0.$$

Proof. For the first statement note that by Minkowski inequality:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n \in [N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N - \int f(w)^\top p_0(w, \theta^\star) \, \Gamma(dw) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n \in [N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) \right\|_4 \qquad (10) \\
&\quad + \left\| \frac{1}{N} \sum_{n \in [N]} f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) - \int f(w)^\top p_0(w, \theta^\star) \, \Gamma(dw) \right\|_4 . \qquad (11)
\end{aligned}$$

Starting from (10), which is also the second statement, we observe that:

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N | \mathbf{W}^N \right] &= \sum_{i=1}^M \mathbb{E}\left[f(\mathbf{w}_n)^{(i)} (\mathbf{x}_{n,0}^N)^{(i)} | \mathbf{W}^N \right] \\
&= \sum_{i=1}^M f(\mathbf{w}_n)^{(i)} p_0(\mathbf{w}_n, \theta^\star)^{(i)} = f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star),
\end{aligned}$$

moreover:

$$\begin{aligned}
&\left| f(\mathbf{w}_n)^\top \mathbf{x}_{n,0}^N - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) \right| \\
&\leq \sum_{i=1}^M \left[\left| f(\mathbf{w}_n)^{(i)} (\mathbf{x}_{n,0}^N)^{(i)} \right| + \left| f(\mathbf{w}_n)^{(i)} p_0(\mathbf{w}_n, \theta^\star)^{(i)} \right| \right] \\
&\leq B \sum_{i=1}^M \left[(\mathbf{x}_{n,0}^N)^{(i)} + p_0(\mathbf{w}_n, \theta^\star)^{(i)} \right] = 2B, \quad \mathbb{P}\text{-almost surely},
\end{aligned}$$

as $\mathbf{x}^N_{n,0}$ is a one-hot encoding vector and $p_0(\mathbf{w}_n, \theta^\star)$ is a probability distribution for any $\mathbf{w}_n$. As $f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,0} - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star)$ are also conditionally independent across n given the population covariates $\mathbf{W}^N$ because of the factorization of the initial distribution, we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,0} - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}},$$

which also proves the second statement for $\widetilde{\alpha}_0 = 1$.

Similarly for (11) we have:

$$\mathbb{E}\left[f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star)\right] = \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw),$$

and also:

$$\left| f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) - \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw) \right| \leq 2B, \quad \mathbb{P}\text{-almost surely.}$$

As $f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) - \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw)$ are independent because functions of independent random variables, indeed $\mathbf{w}_n$ are i.i.d. samples from Γ, we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) - \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}.$$

By putting everything together we conclude the proof by setting $\alpha_0 = 2$, indeed:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,0} - \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,0} - f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) \right\|_4 \\
&\quad + \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top p_0(\mathbf{w}_n, \theta^\star) - \int f(w)^\top p_0(w, \theta^\star)\, \Gamma(dw) \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} + 2B\sqrt[4]{6}N^{-\frac{1}{2}} = 2B\sqrt[4]{6}N^{-\frac{1}{2}}2.
\end{aligned}$$

□

Dynamics. We now turn to the behavior of $\frac{1}{N}\sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,t}$, for $t \geq 1$. We need the following definitions, for $w \in \mathbb{W}$ and $t \geq 1$:

$$\begin{aligned}
\boldsymbol{\lambda}^\infty_0(w, \theta^\star) &\coloneqq p_0(w, \theta^\star) \\
\boldsymbol{\eta}^\infty_{t-1}(w, \theta^\star) &\coloneqq \int d(w, \widetilde{w}, \theta^\star)^\top \boldsymbol{\lambda}^\infty_{t-1}(\widetilde{w}, \theta^\star) \Gamma(d\widetilde{w}) \\
\boldsymbol{\lambda}^\infty_t(w, \theta^\star) &\coloneqq \left[\boldsymbol{\lambda}^\infty_{t-1}(w, \theta^\star)^\top K_{\boldsymbol{\eta}^\infty_{t-1}(w,\theta^\star)}(w, \theta^\star)\right]^\top.
\end{aligned} \tag{12}$$

The quantity $\boldsymbol{\eta}_{t-1}^{\infty}$ can be loosely interpreted as being $\eta^N(\cdot,\cdot,\mathbf{W}^N,\mathbf{X}_{t-1}^N)$ but with $\mathbf{W}^N$ and $\mathbf{X}_{t-1}^N$ integrated out in the $N\to\infty$ limit.

Proposition 13. *Under assumptions 7,9,10, for any $t\geq 1$ there exists $\alpha_t>0$ such that for any bounded function $f:\mathbb{W}\to[0,B]^M$*

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^\top\mathbf{x}_{n,t}^N-\int f(w)^\top\boldsymbol{\lambda}_t^\infty(w,\theta^\star)\Gamma(dw)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_t.$$

Moreover, under assumptions 9,10, for any $t\geq 1$ there exists $\widetilde{\alpha}_t>0$ such that for any bounded function $f:\mathbb{W}\to[0,B]^M$,

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^\top\mathbf{x}_{n,t}^N-f(\mathbf{w}_n)^\top\boldsymbol{\lambda}_t^\infty(\mathbf{w}_n,\theta^\star)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\widetilde{\alpha}_t.$$

Proof. We prove the first statement by induction on t and start by assuming that there exists $\alpha_{t-1}\in\mathbb{R}_+$ such that for any bounded function $f:\mathbb{W}\to[0,B]^M$ we have:

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^\top\mathbf{x}_{n,t-1}^N-\int f(w)^\top\boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)\Gamma(dw)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1},$$

which is true at the initial time step under $\alpha_0=2$ because of the first statement of Proposition 12.

We then look at time step t and decompose the problem into three sub-problems:

$$\begin{aligned}
&f(\mathbf{w}_n)^\top\mathbf{x}_{n,t}^N-\int f(w)^\top\boldsymbol{\lambda}_t^\infty(w,\theta^\star)\Gamma(dw)\\
&=f(\mathbf{w}_n)^\top\mathbf{x}_{n,t}^N-f(\mathbf{w}_n)^\top\left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right]^\top\\
&+f(\mathbf{w}_n)^\top\left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right]^\top-f(\mathbf{w}_n)^\top\left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right]^\top\\
&+f(\mathbf{w}_n)^\top\left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right]^\top-\int f(w)^\top\left[\boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(w,\theta^\star)}(w,\theta^\star)\right]^\top\Gamma(dw).
\end{aligned}$$

By the Minkowski inequality, we can conclude that

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N - \int f(w)^\top \boldsymbol{\lambda}_t^\infty(w,\theta^\star)\Gamma(dw) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N - f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \right\|_4 \qquad (13) \\
&+ \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \mathbf{x}_{n,t-1}^N \right\|_4 \qquad (14) \\
&+ \left\| \frac{1}{N} \sum_{n\in[N]} \left[K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) f(\mathbf{w}_n) \right]^\top \mathbf{x}_{n,t-1}^N \right. \\
&\qquad \left. - \int f(w)^\top \left[\boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(w,\theta^\star)}(w,\theta^\star) \right]^\top \Gamma(dw) \right\|_4 . \qquad (15)
\end{aligned}$$

Consider (13), we can notice that:

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N | \mathbf{X}_{t-1}^N, \mathbf{W}^N \right] &= \sum_{j=1}^M \mathbb{E}\left[f(\mathbf{w}_n)^{(j)} (\mathbf{x}_{n,t}^N)^{(j)} | \mathbf{X}_{t-1}^N, \mathbf{W}^N \right] \\
&= \sum_{j=1}^M \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^{(j)} f(\mathbf{w}_n)^{(j)} \\
&= f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top .
\end{aligned}$$

Given that we are considering differences of scalar products of $f(\mathbf{w}_n)$ with probability/one-hot encoding vectors, we have:

$$\left| f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N - f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \right| \leq 2B \quad \mathbb{P}-\text{almost surely.}$$

As $\mathbf{x}_{n,t}^N$ are conditionally independent across n given the population covariates and the state of the population at the previous time step we can apply Lemma 5 and get:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N - f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \right\|_4 \leq 2B\sqrt[4]{6} N^{-\frac{1}{2}}.$$

We now consider (14) and note that:

$$\begin{aligned}
&\left| f(\mathbf{w}_n)^\top \left[(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top - f(\mathbf{w}_n)^\top \left[(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \right| \\
&\leq \sum_{i=1}^M \sum_{j=1}^M \left| (\mathbf{x}^N_{n,t-1})^{(i)} \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^{(i,j)} f(\mathbf{w}_n)^{(j)} \right| \\
&\leq B \left\| K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right\|_\infty \\
&\leq BL \left| \boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star) - \boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star) \right|,
\end{aligned}$$

where the final steps follow from the boundedness of f, the definition of one-hot encoding vectors, and the Lipschitz-continuity assumption on the transition matrix stated in Assumption 10. We now use the structure of $\boldsymbol{\eta}^N_{n,t-1}$ given in Assumption 9:

$$\begin{aligned}
&\left| \boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star) - \boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star) \right| \\
&= \left| \frac{1}{N} \sum_{k\in[N]} d(\mathbf{w}_n,\mathbf{w}_k,\theta^\star)^\top \mathbf{x}^N_{k,t-1} - \int d(\mathbf{w}_n,\mathbf{w},\theta^\star)^\top \boldsymbol{\lambda}^\infty_{t-1}(\mathbf{w},\theta^\star)\Gamma(d\mathbf{w}) \right|,
\end{aligned} \tag{16}$$

hence we get the following bound for (14):

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \mathbf{x}^N_{n,t-1} \right\|_4 \\
&\leq \frac{1}{N} \sum_{n\in[N]} \left\| f(\mathbf{w}_n)^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \mathbf{x}^N_{n,t-1} \right\|_4 \\
&\leq \frac{BL}{N} \sum_{n\in[N]} \left\| \frac{1}{N} \sum_{k\in[N]} d(\mathbf{w}_n,\mathbf{w}_k,\theta^\star)^\top \mathbf{x}^N_{k,t-1} - \int d(\mathbf{w}_n,w,\theta^\star) \boldsymbol{\lambda}^\infty_{t-1}(w,\theta^\star)\Gamma(dw) \right\|_4.
\end{aligned}$$

As for almost any realization $w_n \in \mathbb{W}$ of $\mathbf{w}_n$ the function $d(w_n,\cdot)$ is a bounded function because of Assumption 9 we can apply our inductive hypothesis and get:

$$\left\| \frac{1}{N} \sum_{k\in[N]} d(w_n,\mathbf{w}_k,\theta^\star)^\top \mathbf{x}^N_{k,t-1} - \int d(w_n,w,\theta^\star) \boldsymbol{\lambda}^\infty_{t-1}(w,\theta^\star)\Gamma(dw) \right\|_4 \leq 2C\sqrt[4]{6} N^{-\frac{1}{2}} \alpha_{t-1}.$$

Since the above bound holds for any $w_n \in \mathbb{W}$, the same inequality holds $\mathbb{P}$-almost surely when the

random covariate $\mathbf{w}_n$ is substituted in place of w_n. We then conclude:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) \right]^\top \mathbf{x}^N_{n,t-1} \right\|_4 \\
&\leq \frac{BL}{N} \sum_{n\in[N]} \left\| \frac{1}{N} \sum_{k\in[N]} d(\mathbf{w}_n,\mathbf{w}_k,\theta^\star)^\top \mathbf{x}^N_{k,t-1} - \int d(\mathbf{w}_n,w,\theta^\star) \boldsymbol{\lambda}^\infty_{t-1}(w,\theta^\star)\Gamma(dw) \right\|_4 \\
&\leq \frac{BL}{N} \sum_{n\in[N]} 2C\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1} = 2BLC\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1}.
\end{aligned}$$

where the first step follows from Minkowski inequality, the second one from our previous calculations, and the final one from the inductive assumption and the fact that $\|d(\mathbf{w}_n,\cdot)\|_\infty \leq C$, $\mathbb{P}$-almost surely from Assumption 9.

The last term that is left to bound is (15). Given that it can be easily proven that for any row-stochastic matrix K the function $Kf(\cdot)$ given by $\mathbf{w} \mapsto Kf(\mathbf{w})$ is such that $\|Kf(\cdot)\|_\infty \leq \|f\|_\infty$, we can conclude that $\left\| K_{\boldsymbol{\eta}^\infty_{t-1}(\cdot,\theta^\star)}(\cdot,\theta^\star) f(\cdot) \right\|_\infty \leq B$. We can then apply our inductive assumption and conclude:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} \left[K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) f(\mathbf{w}_n) \right]^\top \mathbf{x}^N_{n,t-1} \right. \\
&\qquad \left. - \int \left[K_{\boldsymbol{\eta}^\infty_{t-1}(w,\theta^\star)}(w,\theta^\star) f(w) \right]^\top \boldsymbol{\lambda}^\infty_{t-1}(w,\theta^\star)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1}.
\end{aligned}$$

By putting everything together we have:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)\mathbf{x}^N_{n,t} - \int f(w)^\top \boldsymbol{\lambda}^\infty_t(w,\theta^\star)\Gamma(dw) \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} + 2BLC\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1} + 2B\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1} \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} \left[1 + (1+LC)\alpha_{t-1}\right],
\end{aligned}$$

from which we can conclude the proof of the first statement by setting $\alpha_t = [1 + (1+LC)\alpha_{t-1}]$.

The proof of the second statement follows similarly by induction on t. Start by assuming that there exists $\widetilde{\alpha}_{t-1} \in \mathbb{R}_+$ for any $f : \mathbb{W} \to [0,B]^M$ bounded function such that for time $t-1$ we have:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}^N_{n,t-1} - f(\mathbf{w}_n)^\top \boldsymbol{\lambda}^\infty_{t-1}(\mathbf{w}_n,\theta^\star) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\widetilde{\alpha}_{t-1},$$

which is true at the initial time step under $\widetilde{\alpha}_0 = 1$ because of the second statement of Proposition 12. We then follow the same steps as the previous proof, but we substitute (15) with:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \left[K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) f(\mathbf{w}_n) \right]^\top \left[\mathbf{x}^N_{n,t-1} - \boldsymbol{\lambda}^\infty_{t-1}(\mathbf{w}_n,\theta^\star) \right] \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\widetilde{\alpha}_{t-1},$$

which is bounded because of the induction hypothesis. We can then conclude:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{x}_{n,t}^N - f(\mathbf{w}_n)^\top \boldsymbol{\lambda}_t^\infty(\mathbf{w}_n, \theta^\star)^\top \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} + 2BC\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1} + 2B\sqrt[4]{6}N^{-\frac{1}{2}}\widetilde{\alpha}_{t-1} \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} \left[1 + C\alpha_{t-1} + \widetilde{\alpha}_{t-1}\right],
\end{aligned}$$

from which we can conclude the proof of the second statement by setting $\widetilde{\alpha}_t = [1 + C\alpha_{t-1} + \widetilde{\alpha}_{t-1}]$. □

The following corollary tells us that in the large population limit, the interaction term $\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n, \theta^\star)$ converges to the quantity $\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)$ defined in (12) which depends on individual-specific covariate $\mathbf{w}_n$, but not on the covariates or diseases states of the rest of the population. This can be interpreted as meaning that individuals become statistically decoupled as the population size grows.

Corollary 14. *Under assumptions 7,9,10, for any $t \geq 1$ there exists $\alpha_{t-1} > 0$ such that:*

$$\left\| \boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n, \theta^\star) - \boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star) \right\|_4 \leq 2C\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1}.$$

Proof. The result is a byproduct of the proof of Proposition 13. In particular, the passage proving the bound on Equation (16) proves exactly the statement of the present corollary. □

Observations. Using Proposition 13, we next establish an L^4 bound for averages across the population of observations. We need the following definition, for $w \in \mathbb{W}$ and $t \geq 1$:

$$\boldsymbol{\nu}_t^\infty(w, \theta^\star) \coloneqq \left[\boldsymbol{\lambda}_t^\infty(w, \theta^\star)^\top G(w, \theta^\star)\right]^\top. \tag{17}$$

Proposition 15. *Under assumptions 7,9,10, for any $t \geq 1$ there exists $\alpha_t > 0$ such that for any bounded $f : \mathbb{W} \to [0, B]^{M+1}$*

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - \int f(w)^\top \boldsymbol{\nu}_t^\infty(w, \theta^\star)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} (1 + \alpha_t).$$

Moreover, under assumptions 9,10, for any $t \geq 1$ there exists $\widetilde{\alpha}_t > 0$ such that for bounded function $f : \mathbb{W} \to [0, B]^{M+1}$:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - f(\mathbf{w}_n)^\top \boldsymbol{\nu}_t^\infty(\mathbf{w}_n, \theta^\star) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} (1 + \widetilde{\alpha}_t).$$

Proof. To prove the first statement, the first step is to note that by Minkowski inequality:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - \int f(w)^\top \boldsymbol{\nu}_t^\infty(w,\theta^\star)\Gamma(dw) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right]^\top \right\|_4 \\
&\quad + \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right]^\top - \int f(w)^\top \left[\boldsymbol{\lambda}_t^\infty(w,\theta^\star)^\top G(w,\theta^\star)\right]^\top \Gamma(dw) \right\|_4,
\end{aligned}$$

as by definition $\int f(w)^\top \boldsymbol{\nu}_t^\infty(w,\theta^\star)\Gamma(dw) = \int f(w)^\top \left[\boldsymbol{\lambda}_t^\infty(w,\theta^\star)^\top G(w,\theta^\star)\right]^\top \Gamma(dw)$. For the first term, we notice that:

$$\mathbb{E}\left[f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N \middle| \mathbf{X}_t^N, \mathbf{W}^N \right] = f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right] = \left[G(\mathbf{w}_n,\theta^\star) f(\mathbf{w}_n)\right]^\top \mathbf{x}_{n,t}^N,$$

meaning that the arguments of the sum are all conditionally independent and with mean zero. Moreover, given that $[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)]^\top$ is a probability vector and $\mathbf{y}_{n,t}^N$ is a one-hot encoding vector, we can also conclude that the arguments of the sums are all bounded by $2B$, hence we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right]^\top \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}.$$

For the second term we just need some refactoring:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right]^\top - \int f(w)^\top \left[\boldsymbol{\lambda}_t^\infty(w,\theta^\star)^\top G(w,\theta^\star)\right]^\top \Gamma(dw) \right\|_4 \\
&= \left\| \frac{1}{N} \sum_{n\in[N]} \left[G(\mathbf{w}_n,\theta^\star) f(\mathbf{w}_n)\right]^\top \mathbf{x}_{n,t}^N - \int \left[G(w,\theta^\star) f(w)\right]^\top \boldsymbol{\lambda}_t^\infty(w,\theta^\star)\Gamma(dw) \right\|_4.
\end{aligned}$$

As $\|G(\cdot,\theta^\star)f(\cdot)\|_\infty \leq \|f\|_\infty$ because $G(\cdot,\theta^\star)$ is a row-stochastic matrix, we can just apply the first statement of Proposition 13 and conclude:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n,\theta^\star)\right]^\top - \int f(w)^\top \left[\boldsymbol{\lambda}_t^\infty(w,\theta^\star)^\top G(w,\theta^\star)\right]^\top \Gamma(dw) \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_t.
\end{aligned}$$

We then conclude the proof by putting everything together:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \mathbf{y}_{n,t}^N - \int f(w)^\top \boldsymbol{\nu}_t^\infty(w,\theta^\star)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\left(1+\alpha_t\right).$$

Given the above proof strategy, it is trivial to establish the second statement by simply substituting the last term of the Minkowski inequality with:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[(\mathbf{x}_{n,t}^N)^\top G(\mathbf{w}_n, \theta^\star) \right]^\top - f(\mathbf{w}_n)^\top \boldsymbol{\lambda}_t^\infty(\mathbf{w}_n, \theta^\star) \right\|_4,$$

and by applying the second statement of Proposition 13. □

C.3 L^4 bounds for the CAL filtering algorithm

The aim of this section is to establish L^4 bounds for averages across the population of the various probability vectors computed in the CAL filtering recursion, (6). In Proposition 16 and Proposition 18 below, we shall see that the large-population behavior of these averages is determined by the following quantities. Specifically, for $w \in \mathbb{W}$:

$$\begin{aligned}
\bar{\boldsymbol{\pi}}_0^\infty(w,\theta) &\coloneqq p_0(w,\theta), \\
\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta) &\coloneqq \int d(w,\widetilde{w},\theta)^\top \bar{\boldsymbol{\pi}}_{t-1}^\infty(\widetilde{w},\theta)\Gamma(d\widetilde{w}), \\
\bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta) &\coloneqq \left[\bar{\boldsymbol{\pi}}_{t-1}^\infty(w,\theta)^\top K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w,\theta) \right]^\top, \\
\bar{\boldsymbol{\mu}}_t^\infty(w,\theta) &\coloneqq \left[\bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta)^\top G(w,\theta) \right]^\top, \\
\bar{\boldsymbol{\pi}}_t^\infty(w,\theta) &\coloneqq \bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta) \odot \left\{ \left[G(w,\theta) \oslash \left(1_M \bar{\boldsymbol{\mu}}_t^\infty(w,\theta)^\top \right) \right] \boldsymbol{\nu}_t^\infty(w,\theta^\star) \right\},
\end{aligned} \tag{18}$$

where $\boldsymbol{\nu}_t^\infty(w,\theta^\star)$ is defined in (17).

Proposition 16. *Under Assumption 7, for any bounded function $f : \mathbb{W} \to [0,B]^M$ it holds that:*

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta) - \int f(\mathbf{w})^\top \bar{\boldsymbol{\pi}}_0^\infty(\mathbf{w},\theta)\Gamma(d\mathbf{w}) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}.$$

Moreover, for any bounded function $f : \mathbb{W} \to [0,B]^M$:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta) - f(\mathbf{w}_n)^\top \bar{\boldsymbol{\pi}}_0^\infty(\mathbf{w}_n,\theta) \right\|_4 = 0.$$

Proof. The first statement follows the same reasoning of Proposition 12, but we include it here for completeness. As:

$$\mathbb{E}\left[f(\mathbf{w}_n)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta) \right] = \int f(w)^\top \bar{\boldsymbol{\pi}}_0^\infty(w,\theta)\Gamma(dw),$$

and we have a sequence of independent and bounded random variables:

$$\left| f(\mathbf{w}_n)^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}_0^\infty(w,\theta)\Gamma(dw) \right| \leq 2B,$$

we can conclude the proof by applying Lemma 5. The second statement follows trivially from the definition of the limiting process, indeed $\boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n,\theta) = \bar{\boldsymbol{\pi}}_0^\infty(\mathbf{w}_n,\theta)$. □

Proposition 17. *If there exists a constant $\gamma_{t-1} > 0$ such that for any $f : \mathbb{W} \to [0, B]^M$*

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta) - \bar{\boldsymbol{\pi}}^\infty_{t-1}(\mathbf{w}_n, \theta) \right] \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t-1},$$

then under assumptions 9,10 there exists $\gamma_{t|t-1} > 0$ such that for any bounded function $f : \mathbb{W} \to [0, B]^M$,

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) - \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n, \theta) \right] \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1}.$$

Moreover, under assumptions 7,9,10, for any bounded function $f : \mathbb{W} \to [0, B]^M$ it also holds:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(w, \theta)\Gamma(dw)) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}} \left(\gamma_{t|t-1} + 1\right).$$

Proof. As a preliminary notice that:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(w, \theta)\Gamma(dw) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta) - \bar{\boldsymbol{\pi}}^\infty_{t-1}(\mathbf{w}_n, \theta) \right] \right\|_4 \\
&+ \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(w, \theta)\Gamma(dw) \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1} + 1),
\end{aligned} \tag{19}$$

as we can apply Minkowski inequality and then bound the first quantity with our assumption at time $t-1$ and the second quantity with Lemma 5, as we are dealing with an average of independent random variables that are mean zero and bounded (this follows the same steps as the proof of Proposition 16).

As a consequence, we have the following implication of our assumption:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(w, \theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1} + 1),$$

and we know that if we prove the first statement it is enough to follow the above reasoning to prove the second one.

We now start the proof of the first statement by noting:

$$\begin{aligned}
&f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) - f(\mathbf{w}_n)^\top \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n, \theta) \\
&= f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta)^\top K_{\tilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top - f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top \\
&+ f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n, \theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top - f(\mathbf{w}_n)^\top \left[\bar{\boldsymbol{\pi}}^\infty_{t-1}(\mathbf{w}_n, \theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top.
\end{aligned}$$

Hence we can apply Minkowski inequality:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - f(\mathbf{w}_n)^\top \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta) \right\|_4 \\
&\leq \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) - K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \right]^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta) \right\|_4 \qquad (20) \\
&\quad + \left\| \frac{1}{N} \sum_{n\in[N]} \left[K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) f(\mathbf{w}_n) \right]^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\pi}}^\infty_{t-1}(\mathbf{w}_n,\theta) \right] \right\|_4 . \qquad (21)
\end{aligned}$$

On (20) we can apply Assumption 10 and get:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) - K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \right]^\top \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta) \right\|_4 \\
&\leq \frac{B}{N} \sum_{n\in[N]} \left\| \widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta) \right\|_4 ,
\end{aligned}$$

which follows the same steps as the proof of Proposition 13. Again, similarly to Proposition 13, we can apply Assumption 10 on the Lipschitz continuity of the transition matrix, Assumption 9 on the structure of η^N and, given the definition of $\widetilde{\boldsymbol{\eta}}^N_t$, $\bar{\boldsymbol{\eta}}^\infty_t$, we get:

$$\begin{aligned}
&\left\| \widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta) \right\|_4 \\
&= \left\| \frac{1}{N} \sum_{k\in[N]} d(\mathbf{w}_n,\mathbf{w}_k,\theta)^\top \boldsymbol{\pi}^N_{k,t-1}(\mathbf{w}_k,\theta) - \int d(\mathbf{w}_n,w,\theta)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(w,\theta)\Gamma(dw) \right\|_4 \qquad (22) \\
&\leq 2LC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1),
\end{aligned}$$

because of the inductive assumption, and which holds $\mathbb{P}$-almost surely following the same reasoning of Proposition 13. Hence:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) - \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta) \right]^\top \right\|_4 \\
&\leq \frac{B}{N} \sum_{n\in[N]} \left\| \frac{1}{N} \sum_{k\in[N]} d(\mathbf{w}_n,\mathbf{w}_k,\theta)^\top \boldsymbol{\pi}^N_{k,t-1}(\mathbf{w}_k,\theta) - \int d(\mathbf{w}_n,w,\theta)^\top \bar{\boldsymbol{\pi}}^\infty_{t-1}(w,\theta)\Gamma(dw) \right\|_4 \qquad (23) \\
&\leq \frac{B}{N} \sum_{n\in[N]} 2LC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1) = 2BLC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1),
\end{aligned}$$

where the last step follows from (19) and from the same reasoning of Proposition 13, i.e. we prove the inequality $\mathbb{P}$-almost surely.

On (21), we can observe that the vectorial function $\mathbf{w} \mapsto K_{\bar{\boldsymbol{\eta}}^{\infty}_{t-1}(\mathbf{w},\theta)}(\mathbf{w},\theta)f(\mathbf{w})$ is bounded, i.e. $\left\|K_{\bar{\boldsymbol{\eta}}^{\infty}_{t-1}(\cdot,\theta)}(\cdot,\theta)f\right\|_{\infty} \leq B$, as $K_{\bar{\boldsymbol{\eta}}^{\infty}_{t-1}(\mathbf{w},\theta)}$ is a row-stochastic matrix, hence we can apply our assumption on the time step $t-1$ and conclude:

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{n\in[N]}\left[K_{\bar{\boldsymbol{\eta}}^{\infty}_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)f(\mathbf{w}_n)\right]^{\top}\left[\boldsymbol{\pi}^{N}_{n,t-1}(\mathbf{w}_n,\theta)-\bar{\boldsymbol{\pi}}^{\infty}_{t-1}(\mathbf{w}_n,\theta)\right]\right\|_4\\&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t-1}.\end{aligned}$$

By putting everything together we get:

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^{\top}\boldsymbol{\pi}^{N}_{n,t|t-1}(\mathbf{w}_n,\theta)-f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right\|_4\\&\leq 2BLC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1)+2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t-1}=2B\sqrt[4]{6}N^{-\frac{1}{2}}\left[LC(\gamma_{t-1}+1)+\gamma_{t-1}\right],\end{aligned}$$

by setting $\gamma_{t|t-1}=[C(\gamma_{t-1}+1)+\gamma_{t-1}]$ we conclude the proof of the first statement.

Remark that as we already mentioned at the beginning of the proof we have:

$$\mathbb{E}\left[f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]=\int f(w)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(w,\theta)\Gamma(dw),$$

and the random variables $f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)-\int f(w)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(w,\theta)\Gamma(dw)$ are all bounded by $2B$ and independent, hence by Lemma 5:

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)-\int f(w)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(w,\theta)\Gamma(dw)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}},$$

which proves the second statement. □

Proposition 18. *There exists $\gamma_{t|t-1}>0$ for any $f:\mathbb{W}\to[0,B]^M$ bounded function such that*

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^{\top}\boldsymbol{\pi}^{N}_{n,t|t-1}(\mathbf{w}_n,\theta)-f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1},$$

then under assumptions 6,8,9,10 for any $f:\mathbb{W}\to[0,B]^M$ bounded function we have:

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^{\top}\boldsymbol{\mu}^{N}_{n,t}(\mathbf{w}_n,\theta)-f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\mu}}^{\infty}_{t}(\mathbf{w}_n,\theta)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1},$$

and there exists $\gamma_t>0$ such that for any bounded function $f:\mathbb{W}\to[0,B]^M$,

$$\left\|\frac{1}{N}\sum_{n\in[N]}f(\mathbf{w}_n)^{\top}\boldsymbol{\pi}^{N}_{n,t}(\mathbf{w}_n,\theta)-f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}^{\infty}_{t}(\mathbf{w}_n,\theta)\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t}.$$

Moreover, under assumptions 6,7,8,9,10, for any $f : \mathbb{W} \to [0, B]^M$ bounded function

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\mu}}^\infty_t(w, \theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}(1 + \gamma_{t|t-1}),$$

and:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta) - \int f(w)^\top \bar{\boldsymbol{\pi}}^\infty_t(w, \theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}(1 + \gamma_t).$$

Proof. The first statement is straightforward, indeed:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) - f(\mathbf{w}_n)^\top \bar{\boldsymbol{\mu}}^\infty_t(\mathbf{w}_n, \theta) \right\|_4 \\
&= \left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \left[\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta)\right]^\top - f(\mathbf{w}_n)^\top \left[\bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta)\right]^\top \right\|_4 \\
&= \left\| \frac{1}{N} \sum_{n\in[N]} \left[G(\mathbf{w}_n, \theta) f(\mathbf{w}_n)\right]^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta) - \left[G(\mathbf{w}_n, \theta) f(\mathbf{w}_n)\right]^\top \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n, \theta) \right\|_4 .
\end{aligned}$$

As $G(\mathbf{w}, \theta)$ is a row-stochastic matrix the function $\mathbf{w} \mapsto G(\mathbf{w}, \theta)f(\mathbf{w})$ is bounded, and precisely $\|G(\cdot, \theta)f(\cdot)\|_\infty \leq B$, meaning that the proof of the statement follows simply by the assumption on step $t - 1$:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta) - f(\mathbf{w}_n)^\top \bar{\boldsymbol{\mu}}^\infty_t(\mathbf{w}_n, \theta) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1}. \tag{24}$$

The proof of the second statement is more involved and we need to start by reformulating $f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta)$:

$$\begin{aligned}
f(\mathbf{w}_n)^\top \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta) &= \sum_{i=1}^M f(\mathbf{w}_n)^{(i)} \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n, \theta)^{(i)} \\
&= \sum_{i=1}^M f(\mathbf{w}_n)^{(i)} \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta)^{(i)} \sum_{j=1}^M \frac{G(\mathbf{w}_n, \theta)^{(i,j)}}{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)^{(j)}} (\mathbf{y}^N_{n,t})^{(j)} \\
&= \left[f(\mathbf{w}_n) \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n, \theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}^N_{n,t}\right],
\end{aligned} \tag{25}$$

where we define $G_{\boldsymbol{\mu}}(\mathbf{w}, \theta)$ as the matrix with elements $G_{\boldsymbol{\mu}}(\mathbf{w}, \theta)^{(i,j)} = \frac{G(\mathbf{w},\theta)^{(i,j)}}{\boldsymbol{\mu}^{(j)}}$ where $\frac{0}{0} = 0$ by convention. Remark that, because of Theorem 10, the CAL is well-defined $\mathbb{P}$-almost surely. Also

a similar reformulation to the one in (25) can be done on $\bar{\boldsymbol{\pi}}_t^{\infty}(\mathbf{w},\theta)$ meaning: $f(\mathbf{w})^{\top}\bar{\boldsymbol{\pi}}_t^{\infty}(\mathbf{w},\theta) = \left[f(\mathbf{w}) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w},\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w},\theta)}(\mathbf{w},\theta)\boldsymbol{\nu}_t^{\infty}(\mathbf{w},\theta^{\star})\right]$. Then we note that:

$$
\begin{aligned}
& f(\mathbf{w}_n)^{\top}\boldsymbol{\pi}_{n,t}^{N}(\mathbf{w}_n,\theta) - f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}_t^{\infty}(\mathbf{w}_n,\theta) \\
&= \left[f(\mathbf{w}_n) \odot \boldsymbol{\pi}_{n,t|t-1}^{N}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\quad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\quad + \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\quad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\quad + \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\quad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\boldsymbol{\nu}_t^{\infty}(\mathbf{w}_n,\theta^{\star})\right].
\end{aligned}
$$

With the above decomposition we can apply Minkowski inequality and conclude:

$$
\begin{aligned}
& \left\| \frac{1}{N}\sum_{n\in[N]} f(\mathbf{w}_n)^{\top}\boldsymbol{\pi}_{n,t}^{N}(\mathbf{w}_n,\theta) - f(\mathbf{w}_n)^{\top}\bar{\boldsymbol{\pi}}_t^{\infty}(\mathbf{w}_n,\theta) \right\|_4 \\
&\leq \Bigg\| \frac{1}{N}\sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \boldsymbol{\pi}_{n,t|t-1}^{N}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \Bigg\|_4 \qquad (26) \\
&\qquad + \Bigg\| \frac{1}{N}\sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \Bigg\|_4 \qquad (27) \\
&\qquad + \Bigg\| \frac{1}{N}\sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}_{n,t}^{N}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}_{t|t-1}^{\infty}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}_t^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\boldsymbol{\nu}_t^{\infty}(\mathbf{w}_n,\theta^{\star})\right] \Bigg\|_4 . \qquad (28)
\end{aligned}
$$

Starting from (26), we notice that:

$$\begin{aligned}
&\Bigg\| \frac{1}{N} \sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Bigg\|_4 \\
&= \Bigg\| \frac{1}{N} \sum_{n\in[N]} \left[\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[f(\mathbf{w}_n) \odot G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Bigg\|_4 \\
&= \Bigg\| \frac{1}{N} \sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right]^\top \left[\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right] \Bigg\|_4,
\end{aligned}$$

as $G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}$ is a vector of probabilities, i.e. elements that are less or equal than 1, we can conclude that:

$$\left\| f(\cdot) \odot G_{\boldsymbol{\mu}^N_{n,t}(\cdot,\theta)}(\cdot,\theta)\mathbf{y}^N_{n,t} \right\|_\infty \leq B, \quad \mathbb{P}\text{-almost surely},$$

hence, similarly to what we do with $d(\mathbf{w}_n,\cdot)$ in the proof of Proposition 13, we can apply our assumption on time step $t-1$ for almost any realization of $\mathbf{y}^N_{n,t}$, and conclude:

$$\begin{aligned}
&\Bigg\| \frac{1}{N} \sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Bigg\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1}.
\end{aligned}$$

As we define G_μ as the matrix with elements $G^{(i,j)}/\mu^{(j)}$, we can notice that given two vectors a, b with the same dimensions of G_μ we have:

$$x^\top G_\mu b - x^\top G_{\widetilde{\mu}} b = \sum_{i,j} x^{(i)} y^{(j)} \frac{G^{(i,j)}\widetilde{\mu}^{(j)} - G^{(i,j)}\mu^{(j)}}{\widetilde{\mu}^{(j)}\mu^{(j)}} = \sum_{i,j} x^{(i)} \frac{y^{(j)}}{\widetilde{\mu}^{(j)}\mu^{(j)}} G^{(i,j)} \left(\widetilde{\mu}^{(j)} - \mu^{(j)}\right)$$

Hence we can reformulate (27):

$$\begin{aligned}
&\Bigg\| \frac{1}{N} \sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\bar{\boldsymbol{\mu}}^\infty_t(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Bigg\|_4 \\
&= \Bigg\| \frac{1}{N} \sum_{n\in[N]} \left\{ \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta) \right\}^\top \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \bar{\boldsymbol{\mu}}^\infty_t(\mathbf{w}_n,\theta)\right]^\top \\
&\qquad \left[\bar{\boldsymbol{\mu}}^\infty_t(\mathbf{w}_n,\theta) - \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right] \Bigg\|_4,
\end{aligned}$$

from which we can notice that for any $\mathbf{w}_n$:

$$\begin{aligned}&\left\|\left\{\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}G(\mathbf{w}_n,\theta)\right\}\odot\left[\mathbf{y}^N_{n,t}\oslash\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\oslash\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)\right]\right\|_{\infty}\\&\leq\|f\|_{\infty}\left\|\mathbf{y}^N_{n,t}\oslash\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right\|_{\infty},\quad\mathbb{P}\text{-almost surely,}\end{aligned}$$

where the first step follows from $\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w},\theta)=\left[\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w},\theta)^{\top}G(\mathbf{w},\theta)\right]^{\top}$ and the elementwise ratio $\mathbf{y}^N_{n,t}\oslash\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)$ is well-defined because of Theorem 10.

Because of Proposition 11 we know that there exists $m_t>0$ such that $\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\geq m_t$ almost surely. We can then conclude that for any $n\in[N]$:

$$\begin{aligned}&\left\|\left\{\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}G(\mathbf{w}_n,\theta)\right\}\odot\left[\mathbf{y}^N_{n,t}\oslash\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\oslash\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)\right]\right\|_{\infty}\\&\leq\frac{\|f\|_{\infty}}{m_t}\leq\frac{B}{m_t},\quad\mathbb{P}\text{-almost surely.}\end{aligned}$$

Hence we can apply to (27) the same reasoning that we do with $d(\mathbf{w}_n,\cdot)$ in the proof of Proposition 13 and apply (24) for almost any realization of $\mathbf{Y}^N_{1:t-1},\mathbf{y}^N_{n,t}$, i.e. $\mathbb{P}$-almost surely, and conclude:

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{n\in[N]}\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right]\right.\\&\qquad\left.-\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right]\right\|_4\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\frac{\gamma_{t|t-1}}{m_t}.\end{aligned}$$

Finally, we notice that for (28):

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{n\in[N]}\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right]\right.\\&\qquad\left.-\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}\left[G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\boldsymbol{\nu}^{\infty}_t(\mathbf{w}_n,\theta^{\star})\right]\right\|_4\\&=\left\|\frac{1}{N}\sum_{n\in[N]}\left\{\left[f(\mathbf{w}_n)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top}G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\right\}\left[\mathbf{y}^N_{n,t}-\boldsymbol{\nu}^{\infty}_t(\mathbf{w}_n,\theta^{\star})\right]\right\|_4,\end{aligned}$$

from which we can see that:

$$\begin{aligned}\left\|\left[f(\cdot)\odot\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\cdot,\theta)\right]^{\top}G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\cdot,\theta)}(\cdot,\theta)\right\|_{\infty}&\leq\|f\|_{\infty}\left\|\left[\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\cdot,\theta)\right]^{\top}G_{\bar{\boldsymbol{\mu}}^{\infty}_t(\cdot,\theta)}(\cdot,\theta)\right\|_{\infty}\\&\leq\|f\|_{\infty}\left\|\sum_i\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\cdot,\theta)^{(i)}\frac{G(\cdot,\theta)^{(i,j)}}{\bar{\boldsymbol{\mu}}^{\infty}_t(\cdot,\theta)^{(j)}}\right\|_{\infty}\\&=\|f\|_{\infty},\end{aligned}$$

where the last step follows from $\bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\cdot,\theta), G(\cdot,\theta), \bar{\boldsymbol{\mu}}^{\infty}_{t}(\cdot,\theta)$ being almost surely positive and from $[\bar{\boldsymbol{\mu}}^{\infty}_{t}(\cdot,\theta) = \bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\cdot,\theta)^{\top} G(\cdot,\theta)]^{\top}$. We can then apply Proposition 15 and conclude

$$\left\| \frac{1}{N} \sum_{n\in[N]} \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top} \left[G_{\bar{\boldsymbol{\mu}}^{\infty}_{t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^{N}_{n,t}\right] \right.$$
$$\left. - \left[f(\mathbf{w}_n) \odot \bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(\mathbf{w}_n,\theta)\right]^{\top} \left[G_{\bar{\boldsymbol{\mu}}^{\infty}_{t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\boldsymbol{\nu}^{\infty}_{t}(\mathbf{w}_n,\theta^{\star})\right] \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\left(1+\alpha_t\right).$$

By putting everything together we then have:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^{\top} \boldsymbol{\pi}^{N}_{n,t}(\mathbf{w}_n,\theta) - f(\mathbf{w}_n)^{\top} \bar{\boldsymbol{\pi}}^{\infty}_{t}(\mathbf{w}_n,\theta) \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1} + 2B\sqrt[4]{6}N^{-\frac{1}{2}}\frac{\gamma_{t|t-1}}{m_t} + 2B\sqrt[4]{6}N^{-\frac{1}{2}}\left(1+\alpha_t\right) \\
&= 2B\sqrt[4]{6}N^{-\frac{1}{2}}\left[\gamma_{t|t-1} + \frac{\gamma_{t|t-1}}{m_t} + (1+\alpha_t)\right],
\end{aligned}$$

which concludes the proof for $\gamma_t = \left[\gamma_{t|t-1} + \frac{\gamma_{t|t-1}}{m_t} + (1+\alpha_t)\right]$.

The third and the fourth statements are simple consequences of the first and the second statement, Minkowski inequality and Lemma 5, see the reasoning to prove the second statement of Proposition 17. □

We now combine propositions 16 - 18.

Proposition 19. *Under assumptions 6,7,8,9,10, for any $t \geq 1$ there exist $\gamma_{t|t-1} > 0$ and $\gamma_t > 0$ such that for any bounded function $f : \mathbb{W} \to [0,B]^M$*

$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^{\top} \boldsymbol{\pi}^{N}_{n,t|t-1}(\mathbf{w}_n,\theta) - \int f(w)^{\top} \bar{\boldsymbol{\pi}}^{\infty}_{t|t-1}(w,\theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1},$$
$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^{\top} \boldsymbol{\mu}^{N}_{n,t}(\mathbf{w}_n,\theta) - \int f(w)^{\top} \bar{\boldsymbol{\mu}}^{\infty}_{t}(w,\theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t|t-1},$$
$$\left\| \frac{1}{N} \sum_{n\in[N]} f(\mathbf{w}_n)^{\top} \boldsymbol{\pi}^{N}_{n,t}(\mathbf{w}_n,\theta) - \int f(w)^{\top} \bar{\boldsymbol{\pi}}^{\infty}_{t}(w,\theta)\Gamma(dw) \right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\gamma_{t}.$$

Proof. Assume that the third statement is true at time $t-1$, which is verified for $t-1=0$ because of Proposition 16. Then we can in turn apply Proposition 17 to prove the first statement, and Proposition 18 to prove the second and the third. As the third statement is also our inductive hypothesis and we have proven it to hold for the next time step t we can state that the three statements hold for an arbitrary time step t. □

The following corollary is analogous to Corollary 14, but addresses the quantity $\widetilde{\boldsymbol{\eta}}_{t-1}^N(\cdot,\cdot)$ computed in the CAL algorithm (recall 6). Corollary 20 can be interpreted as meaning that the probability vectors computed in the CAL algorithm become decoupled across the population as $N \to \infty$. Recall from (12) and (18) the definitions of $\boldsymbol{\eta}_{t-1}^\infty(\cdot,\cdot)$ and $\bar{\boldsymbol{\eta}}_{t-1}^\infty(\cdot,\cdot)$.

Corollary 20. *Under assumptions 7,9,10, for any $t \geq 1$ there exists $\gamma_{t-1} > 0$ such that for any $w \in \mathbb{W}$ and $\theta \in \Theta$,*

$$\left\| \widetilde{\boldsymbol{\eta}}_{t-1}^N(w,\theta) - \bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta) \right\|_4 \leq 2LC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1).$$

Moreover, $\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta^\star) = \boldsymbol{\eta}_{t-1}^\infty(w,\theta^\star)$.

Proof. The first statement is a byproduct of Proposition 19 and Equation (22) in the proof of Proposition 17. Indeed, Proposition 19 ensures that the assumptions of Proposition 17 hold.

The second statement follows by induction, indeed $\bar{\boldsymbol{\pi}}_0^\infty(w,\theta^\star) = \boldsymbol{\lambda}_0^\infty(w,\theta^\star)$ by definition hence $\bar{\boldsymbol{\pi}}_{t-1}^\infty(w,\theta^\star) = \boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)$ holds for $t=1$. Suppose now $\bar{\boldsymbol{\pi}}_{t-1}^\infty(w,\theta^\star) = \boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)$ is true, then we get that for any $i \in [M]$:

$$\begin{aligned}
\bar{\boldsymbol{\pi}}_t^\infty(w,\theta^\star)^{(i)} &= \bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta^\star)^{(i)} \left\{ \left[G(w,\theta^\star) \oslash \left(1_M \bar{\boldsymbol{\mu}}_t^\infty(w,\theta^\star)^\top \right) \right] \boldsymbol{\nu}_t^\infty(w,\theta^\star) \right\}^{(i)} \\
&= \bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta^\star)^{(i)} \left\{ \left[G(w,\theta^\star) \oslash \left(1_M \nu_t^\infty(w,\theta^\star)^\top \right) \right] \boldsymbol{\nu}_t^\infty(w,\theta^\star) \right\}^{(i)} \\
&= \bar{\boldsymbol{\pi}}_{t|t-1}^\infty(w,\theta^\star)^{(i)} \sum_{j=1}^M \frac{G(w,\theta^\star)^{(i,j)}}{\boldsymbol{\nu}_t^\infty(w,\theta^\star)^{(j)}} \boldsymbol{\nu}_t^\infty(w,\theta^\star)^{(j)} \\
&= \sum_{j=1}^M \bar{\boldsymbol{\pi}}_{t-1}^\infty(w,\theta^\star)^{(j)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta^\star)}(w,\theta^\star)^{(j,i)} \\
&= \sum_{j=1}^M \boldsymbol{\lambda}_{t-1}^\infty(w,\theta^\star)^{(j)} K_{\boldsymbol{\eta}_{t-1}^\infty(w,\theta^\star)}(w,\theta^\star)^{(j,i)} = \boldsymbol{\lambda}_t^\infty(w,\theta^\star)^{(i)},
\end{aligned}$$

which concludes the proof. □

C.4 The saturated processes and saturated CAL algorithm

In Section C.4 we introduce key objects which will allow us to understand the large-population behavior of the data-generating process and the CAL algorithm. The first is what we call the *one-individual saturated process*. This process can be interpreted as describing the evolution of a single individual in an infinite population, where the covariate vector associated with this one individual is drawn from the distribution Γ appearing in Assumption 7. The term "saturated" refers to the fact that the law of the process is defined in terms of the limiting interaction terms $\boldsymbol{\eta}_t^\infty(\cdot,\cdot)$, $t \geq 0$, cf. Corollary 14.

We then introduce the *population saturated process*, which consists of independent individuals, each of which is distributed in the same way as the one-individual saturated process, but with

covariate vectors $(\mathbf{w}_n)_{n\in[N]}$ which are shared with those in the data-generating process specified in Section A.4. In Proposition 24 we shall show that averages across the population in the data-generating process approximate averages across the population saturated process.

Lastly, we introduce the *saturated CAL algorithm*, which is very similar to the CAL algorithm, but uses a model that is decoupled across individuals. Proposition 27 below will tell us that the logarithm of the CAL approximates the logarithm of the corresponding quantity from the saturated CAL algorithm.

One-individual saturated process. We define the one-individual saturated process as:

$$\begin{aligned}
&\mathbf{w}^\infty \sim \Gamma,\\
&\mathbf{x}_0^\infty|\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot \,|\, p_0(\mathbf{w}^\infty,\theta^\star)\right),\\
&\mathbf{x}_t^\infty|\mathbf{x}_{t-1}^\infty,\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot \,\middle|\, \left[(\mathbf{x}_{t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}^\infty,\theta^\star)}(\mathbf{w}^\infty,\theta^\star)\right]^\top\right),\\
&\mathbf{y}_t^\infty|\mathbf{x}_t^\infty,\mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot \,\middle|\, \left[(\mathbf{x}_t^\infty)^\top G(\mathbf{w}^\infty,\theta^\star)\right]^\top\right).
\end{aligned} \tag{29}$$

Conditional on $\mathbf{w}^\infty$, the joint process $(\mathbf{x}_t^\infty)_{t\geq 0}$, $(\mathbf{y}_t^\infty)_{t\geq 1}$ described in (29) is a HMM. If in Recursion (30) below $\mathbf{w}^\infty$ and $\theta^\star$ are substituted in place of w and θ, then (30) becomes the forward algorithm associated with this HMM.

Recall from (18) the definition of $\bar{\boldsymbol{\eta}}_{t-1}^\infty(\cdot,\cdot)$, and define for $w\in\mathbb{W}$ and $t\geq 1$:

$$\begin{aligned}
&\boldsymbol{\pi}_0^\infty(w,\theta) \coloneqq p_0(w,\theta),\\
&\boldsymbol{\pi}_{t|t-1}^\infty(w,\theta) \coloneqq \left[\boldsymbol{\pi}_{t-1}^\infty(w,\theta)^\top K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w,\theta)\right]^\top,\\
&\boldsymbol{\mu}_t^\infty(w,\theta) \coloneqq \left[\boldsymbol{\pi}_{t|t-1}^\infty(w,\theta)^\top G(w,\theta)\right]^\top,\\
&\boldsymbol{\pi}_t^\infty(w,\theta) \coloneqq \boldsymbol{\pi}_{t|t-1}^\infty(w,\theta) \odot \left\{\left[G(w,\theta) \oslash \left(1_M \boldsymbol{\mu}_t^\infty(w,\theta)^\top\right)\right] \mathbf{y}_t^\infty\right\}.
\end{aligned} \tag{30}$$

Recall from (30) the definition of $\boldsymbol{\mu}_t^\infty(\cdot,\cdot)$.

Proposition 21. *Over a time horizon T, let $w\in\mathbb{W}$ and $y_{1:T}\in\mathbb{O}_{M+1}$:*

$$p(\mathbf{y}_{1:T}^\infty|\mathbf{w}^\infty,\theta^\star) \coloneqq \prod_{t=1}^T Cat(\mathbf{y}_t^\infty|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)) = \prod_{t=1}^T (\mathbf{y}_t^\infty)^\top \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star),$$

and it is obtained as a marginal distribution over $x_{0:T}\in\mathbb{O}_M$ of:

$$\begin{aligned}
&p(\mathbf{x}_{0:T}^\infty,\mathbf{y}_{1:T}^\infty|\mathbf{w}^\infty,\theta^\star) \coloneqq\\
&Cat(\mathbf{x}_0^\infty|p_0(\mathbf{w}^\infty,\theta^\star)) \prod_{t=1}^T Cat\left(\mathbf{x}_t^\infty|\left[(\mathbf{x}_{t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}^\infty,\theta^\star)}(\mathbf{w}^\infty,\theta^\star)\right]^\top\right)\\
&\qquad\qquad \cdot \prod_{t=1}^T Cat\left(\mathbf{y}_t^\infty|\left[(\mathbf{x}_t^\infty)^\top G(\mathbf{w}^\infty,\theta^\star)\right]^\top\right)\\
&= \left((\mathbf{x}_0^\infty)^\top p_0(\mathbf{w}^\infty,\theta^\star)\right) \prod_{t=1}^T \left[\left((\mathbf{x}_{t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}^\infty,\theta^\star)}(\mathbf{w}^\infty,\theta^\star)\mathbf{x}_t^\infty\right)\left((\mathbf{x}_t^\infty)^\top G(\mathbf{w}^\infty,\theta^\star)\mathbf{y}_t^\infty\right)\right].
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\mathbf{x}_t^\infty|\mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty &\sim Cat\left(\cdot|\boldsymbol{\pi}_{t|t-1}^\infty(\mathbf{w}^\infty, \theta^\star)\right), \\
\mathbf{x}_t^\infty|\mathbf{y}_{1:t}^\infty, \mathbf{w}^\infty &\sim Cat\left(\cdot|\boldsymbol{\pi}_{t}^\infty(\mathbf{w}^\infty, \theta^\star)\right), \\
\mathbf{y}_t^\infty|\mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty &\sim Cat\left(\cdot|\boldsymbol{\mu}_{t}^\infty(\mathbf{w}^\infty, \theta^\star)\right).
\end{aligned}$$

Proof. Note that because of Corollary 20 we have $\boldsymbol{\eta}_t^\infty(\mathbf{w}^\infty, \theta^\star) = \bar{\boldsymbol{\eta}}_t^\infty(\mathbf{w}^\infty, \theta^\star)$. The proof of the first statement follows then the same calculation of the proof of Proposition 4, with the infinite transition matrix $K_{\boldsymbol{\eta}_t^\infty(\mathbf{w}^\infty, \theta^\star)}(\mathbf{w}^\infty, \theta^\star)$. The proof of the remaining statements is trivial by noting that $(\mathbf{x}_t^\infty)_{t\geq 0}, (\mathbf{y}_t^\infty)_{t\geq 1}$ given $\mathbf{w}^\infty$ is a HMM. Indeed, Recursion 30 can be interpreted as the forward algorithm for the HMM $(\mathbf{x}_t^\infty)_{t\geq 0}, (\mathbf{y}_t^\infty)_{t\geq 1}$ given $\mathbf{w}^\infty$. □

Proposition 22. *Under assumptions 7,8, for any $t \geq 1$ and $\theta \in \Theta$ we have that* $\sum_i \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)^{(i)}(\mathbf{y}_t^\infty)^{(i)} \neq 0 \quad \mathbb{P} - \text{almost surely.}$

Proof. The proof follows the same steps as the proof of Theorem 10, but we work with the one-individual saturated process.

We can note that because of Assumption 8 if there exists $\theta \in \Theta$, $i \in [M]$ such that $\boldsymbol{\pi}_0^\infty(w, \theta)^{(i)} = p_0(w, \theta)^{(i)} = 0$, then $p_0(w, \theta^\star)^{(i)} = 0$ and so:

$$\mathbb{P}((\mathbf{x}_0^\infty)^{(i)} = 0|\mathbf{w}^\infty = w) = p_0(w, \theta) = \boldsymbol{\pi}_0^\infty(w, \theta)^{(i)} = 0,$$

meaning that $(\mathbf{x}_0^\infty)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^\infty = w)$.

Assume now by induction that if there exists $\theta \in \Theta$, $i \in [M]$ such that $\boldsymbol{\pi}_{t-1}^\infty(w, \theta)^{(i)} = 0$, then $(\mathbf{x}_{t-1}^\infty)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^\infty = w, \mathbf{y}_{1:t-1}^\infty = y_{1:t-1})$, which is satisfied at time $t = 1$ because of our previous reasoning.

We can note that:

$$\begin{aligned}
\boldsymbol{\pi}_{t|t-1}^\infty(w, \theta)^{(i)} = 0 &\iff \sum_j \boldsymbol{\pi}_{t-1}^\infty(w, \theta)^{(j)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w_n, \theta)^{(j,i)} = 0 \\
&\iff \forall j \in [M] \quad \boldsymbol{\pi}_{t-1}^\infty(w, \theta)^{(j)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w_n, \theta)^{(j,i)} = 0,
\end{aligned}$$

hence for all $j \in [M]$ we either have:

1. $\boldsymbol{\pi}_{t-1}^\infty(w, \theta)^{(j)} = 0$ which implies $(\mathbf{x}_{t-1}^\infty)^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^\infty = w, \mathbf{y}_{1:t-1}^\infty = y_{1:t-1})$ by our inductive assumption, or

2. $K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w, \theta)^{(j,i)} = 0$, which implies that there exists η such that $K_\eta(w, \theta)^{(j,i)} = 0$ and so, by Assumption 8, $K_\eta(w, \theta)^{(j,i)} = 0$ for all $\eta \in [0, C]$ and $\theta \in \Theta$ meaning $K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta^\star)}(w, \theta^\star)^{(j,i)} = 0$.

from which we conclude:

$$\begin{aligned}
&\mathbb{P}((\mathbf{x}_t^\infty)^{(i)} = 0|\mathbf{w}^\infty = w, \mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) \\
&= 1 - \sum_{x_{t-1}} \mathbb{P}((\mathbf{x}_t^\infty)^{(i)} = 1|\mathbf{w}^\infty = w, \mathbf{x}_{t-1}^\infty = x_{t-1}, \mathbf{y}_{1:t}^\infty = y_{1:t}) \\
&\qquad \cdot \mathbb{P}(\mathbf{x}_{t-1}^\infty = x_{t-1}|\mathbf{w}^\infty = w, \mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) \\
&= 1 - \sum_j K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(w,\theta)}(w, \theta)^{(j,i)} \mathbb{P}((\mathbf{x}_{t-1}^\infty)^{(j)} = 1|\mathbf{w}^\infty = w, \mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) = 1,
\end{aligned}$$

hence if $\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} = 0$ then $(\mathbf{y}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$.

Moving to:

$$\begin{aligned}\boldsymbol{\mu}^{\infty}_t(w,\theta)^{(i)} = 0 &\iff \sum_j \boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(j)} G(w,\theta)^{(j,i)} = 0 \\ &\iff \forall j \in [M] \quad \boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(j)} G(w,\theta)^{(j,i)} = 0,\end{aligned}$$

we then have that for all j either:

1. $\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(j)} = 0$ which implies $(\mathbf{x}^{\infty}_t)^{(j)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$ because of what we have proven above, or
2. $G(w,\theta)^{(j,i)} = 0$ which implies $G(w,\theta^{\star})^{(j,i)} = 0$ because of Assumption 8,

meaning that:

$$\begin{aligned}&\mathbb{P}\left((\mathbf{y}^{\infty}_t)^{(i)} = 0|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) = 1 - \mathbb{P}\left((\mathbf{y}^{\infty}_t)^{(i)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) \\ &= 1 - \sum_{x_t} \mathbb{P}\left(\mathbf{x}^{\infty}_t = x_t|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) \\ &\qquad \cdot \mathbb{P}\left((\mathbf{y}^{\infty}_t)^{(i)} = 1|\mathbf{w}^{\infty} = w, \mathbf{x}^{\infty}_t = x_t, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) \\ &= 1 - \sum_j \mathbb{P}\left((\mathbf{x}^{\infty}_t)^{(j)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) G(w,\theta^{\star})^{(j,i)} = 1,\end{aligned}$$

hence if $\boldsymbol{\mu}^{\infty}_t(w,\theta)^{(i)} = 0$ then $(\mathbf{y}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$.

Consider now:

$$\begin{aligned}&\boldsymbol{\pi}^{\infty}_t(w,\theta)^{(i)} = 0 \iff \boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} \frac{\sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)}}{\sum_j (\mathbf{y}^{\infty}_t)^{(j)} \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(j)}} = 0 \\ &\iff \boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} \sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)} = 0 \text{ and } \sum_j (\mathbf{y}^{\infty}_t)^{(j)} \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(j)} \neq 0.\end{aligned}$$

Similarly to Theorem 10 we have $\sum_j (\mathbf{y}^{\infty}_t)^{(j)} \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(j)} \neq 0$ being an almost sure event under $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$, hence:

$$\begin{aligned}&\mathbb{P}\left(\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} \sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)} = 0 \right. \\ &\qquad \left. \text{and } \sum_j (\mathbf{y}^{\infty}_t)^{(j)} \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(j)} \neq 0 | \mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) \\ &= \mathbb{P}\left(\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} \sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)} = 0 | \mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right).\end{aligned}$$

We then just need to work on the event $\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} \sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)} = 0$.

We then have either:

- $\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} = 0$, implying $(\mathbf{x}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$ because of what we have proven above, or
- $\sum_j G(w,\theta)^{(i,j)}(\mathbf{y}^{\infty}_t)^{(j)} = 0$, which tells us that there exists $k \in [M]$ such that $\mathbb{P}\left((\mathbf{y}^{\infty}_t)^{(k)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) > 0$ and $G(w,\theta)^{(i,k)} = 0$, note that under $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t} = y_{1:t})$ we know k as we are conditioning on $\mathbf{y}^{\infty}_t$;

as:

$$\begin{aligned}
&\mathbb{P}((\mathbf{x}^{\infty}_t)^{(i)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t} = y_{1:t}) \\
&\propto \mathbb{P}((\mathbf{x}^{\infty}_t)^{(i)} = 1, (\mathbf{y}^{\infty}_t)^{(k)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}) \\
&= G(w,\theta)^{(i,k)}\mathbb{P}((\mathbf{x}^{\infty}_t)^{(i)} = 1|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}) = 0,
\end{aligned}$$

we can then conclude $(\mathbf{x}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t} = y_{1:t})$.

With this final result we have shown that our inductive assumption is true at time t, hence by induction we can conclude that for any $t \geq 1$:

- if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^{\infty}_{t|t-1}(w,\theta)^{(i)} = 0$, then $(\mathbf{x}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$;
- if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\mu}^{\infty}_t(w,\theta)^{(i)} = 0$, then $(\mathbf{y}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1})$;
- if there exist $\theta \in \Theta, i \in [M]$ such that $\boldsymbol{\pi}^{\infty}_t(w,\theta)^{(i)} = 0$, then $(\mathbf{x}^{\infty}_t)^{(i)} = 0$ almost surely in $\mathbb{P}(\cdot|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t} = y_{1:t})$.

To prove that the random asymptotic CAL is a well-defined algorithm in $\mathbb{P}$ we need to prove that:

$$\mathbb{P}\left(\forall\theta \in \Theta, t \geq 1 \quad \sum_i \boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)}(\mathbf{y}^{\infty}_t)^{(i)} \neq 0\right) = 1,$$

which can be proven by observing that:

$$\begin{aligned}
&\mathbb{P}\left(\forall\theta \in \Theta, t \geq 1, \quad \sum_i \boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)}(\mathbf{y}^{\infty}_t)^{(i)} \neq 0\right) \\
&= \mathbb{P}\left(\forall\theta \in \Theta, t \geq 1, \quad \sum_i \boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)}(\mathbf{y}^{\infty}_t)^{(i)} \neq 0\right) \\
&= \int \mathbb{P}\left(\forall\theta \in \Theta, t \geq 1 \quad \sum_i \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(i)}(\mathbf{y}^{\infty}_t)^{(i)} \neq 0|\mathbf{w}^{\infty} = w\right)\Gamma(dw) \\
&= \int \sum_{y_{1:t-1}} \mathbb{P}\left(\forall\theta \in \Theta, t \geq 1 \quad \sum_i \boldsymbol{\mu}^{\infty}_t(w,\theta)^{(i)}(\mathbf{y}^{\infty}_t)^{(i)} \neq 0|\mathbf{w}^{\infty} = w, \mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}\right) \\
&\qquad \cdot \mathbb{P}(\mathbf{y}^{\infty}_{1:t-1} = y_{1:t-1}|\mathbf{w}^{\infty} = w)\Gamma(dw) \\
&= 1.
\end{aligned}$$

□

Proposition 23. *Under assumptions 6,8,10, there exists $m_t > 0$ as in Proposition 11 for any $t \geq 1$such that:*

$$\mathbb{P}\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)^{(i)} \geq m_t \quad \forall i \in \boldsymbol{supp}\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)\right)\right) = 1 \quad \forall \theta \in \Theta.$$

Proof. Consider the following inductive hypothesis. There exists $\bar{m}_{t-1} > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}_{t-1}^\infty(\mathbf{w}^\infty,\theta)^{(i)} \geq \bar{m}_{t-1} \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}_{t-1}^\infty(\mathbf{w}^\infty,\theta)\right)\right) = 1 \quad \forall n \in [N], \theta \in \Theta.$$

As for Proposition 11, from Assumption 6 we have that $p_0(w,\theta)$ is continuous in w,θ and both $\mathbb{W}$ and Θ are compact, we then get from Weierstrass theorem that there exists a minimum m_0 such that for any realization of $\mathbf{w}^\infty$ and for any $i \in \mathbf{supp}\left(p_0(\mathbf{w}^\infty,\theta)^{(i)}\right)$:

$$\boldsymbol{\pi}_0^\infty(\mathbf{w}^\infty,\theta)^{(i)} = p_0(\mathbf{w}^\infty,\theta)^{(i)} \geq \min_{w\in\mathbb{W},\theta\in\Theta} \min_{j\in\mathbf{supp}\left(p_0(w,\theta)^{(j)}\right)} p_0(w,\theta)^{(j)} =: m_0,$$

with $m_0 > 0$ as we are considering a minimum over $j \in \mathbf{supp}\left(p_0(w,\theta)^{(j)}\right)$ which excludes all the zeros. As m_0 does not depend on $\mathbf{w}^\infty$ we conclude that the inductive hypothesis holds for $t-1=0$.

We now move to $\boldsymbol{\pi}_{t|t-1}^\infty(\mathbf{w}^\infty,\theta)$, and let $i \in \mathbf{supp}\left(\boldsymbol{\pi}_{t|t-1}^\infty(\mathbf{w}^\infty,\theta)\right)$ then:

$$\boldsymbol{\pi}_{t|t-1}^\infty(\mathbf{w}^\infty,\theta)^{(i)} \geq \bar{m}_{t-1} \sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}_{t-1}^\infty(\mathbf{w}^\infty,\theta)\right)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(j,i)},$$

where the inequality holds with probability 1 under the inductive hypothesis. Several of the remaining inequalities in the proof also hold with probability 1, although to avoid repetition we do not state this explicitly. Similarly to Proposition 11:

$$\sum_{j\in\mathbf{supp}\left(\boldsymbol{\pi}_{t-1}^\infty(\mathbf{w}^\infty,\theta)\right)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(j,i)} \geq \min_{j\in\mathbf{supp}\left(K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(\cdot,i)}\right)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(j,i)}.$$

Because of Assumption 8 we have:

$$K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(j,i)} = 0 \iff K_\eta(\mathbf{w}^\infty,\theta)^{(j,i)} = 0 \quad \forall \eta \in [0,C],$$

meaning that

$$\min_{j\in\mathbf{supp}\left(K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(\cdot,i)}\right)} K_{\bar{\boldsymbol{\eta}}_{t-1}^\infty(\mathbf{w}^\infty,\theta)}(\mathbf{w}^\infty,\theta)^{(j,i)} \geq \min_{\eta\in[0,C]} \min_{j\in\mathbf{supp}\left(K_\eta(\mathbf{w}^\infty,\theta)^{(\cdot,i)}\right)} K_\eta(\mathbf{w}^\infty,\theta)^{(j,i)}.$$

We can then conclude:

$$\begin{aligned}
\boldsymbol{\pi}_{t|t-1}^\infty(\mathbf{w}^\infty,\theta)^{(i)} &\geq \bar{m}_{t-1} \min_{\theta\in\Theta,w\in\mathbb{W},\eta\in[0,C]} \min_{j\in\mathbf{supp}\left(K_\eta(w,\theta)^{(\cdot,i)}\right)} K_\eta(w,\theta)^{(j,i)} \\
&\geq \bar{m}_{t-1} \min_{\theta\in\Theta,w\in\mathbb{W},\eta\in[0,C]} \min_{(i,j)\in\mathbf{supp}(K_\eta(w,\theta))} K_\eta(w,\theta)^{(j,i)}.
\end{aligned}$$

As $K_\eta(w,\theta)$ is continuous in η because of Assumption 10 and also in w,θ because of Assumption 6, and both $[0,C]$ is compact by definition and $\mathbb{W},\Theta$ are compact because of Assumption 6, we can conclude by Weirstrass theorem that there exists a minimum m_K, hence:

$$\boldsymbol{\pi}^{\infty}_{t|t-1}(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K > 0,$$

where the strictly greater than zero follows from considering the minimum on the support of the transition matrix. As there is no dependence on $\mathbf{w}^{\infty}, \mathbf{y}^{\infty}_{1:t-1}$ we conclude that there exist $\bar{m}_{t-1}, m_K > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}^{\infty}_{t|t-1}(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}^{\infty}_{t|t-1}(\mathbf{w}^{\infty},\theta)\right)\right) = 1 \quad \forall \theta \in \Theta.$$

Following again the same steps as in Proposition 11, we have that for an arbitrary realization of $\mathbf{w}^{\infty}, \mathbf{y}^{\infty}_{1:t-1}$, and $i \in \mathbf{supp}\left(\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)\right)$:

$$\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K \sum_{j \in \mathbf{supp}\left(\boldsymbol{\pi}^{\infty}_{t|t-1}(\mathbf{w}^{\infty},\theta)\right)} G(\mathbf{w}^{\infty},\theta)^{(j,i)},$$

where the inequality follows from what we have proven above. Moreover:

$$\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K \min_{w \in \mathbb{W}, \theta \in \Theta} \min_{(i,j) \in \mathbf{supp}(G(w,\theta))} G(w,\theta)^{(j,i)},$$

and as $G(w,\theta)$ is continuous in w,θ because of Assumption 6 and $\mathbb{W},\Theta$ are compact because of Assumption 6 we can conclude by Weirstrass theorem that there exists a minimum m_G:

$$\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G > 0,$$

where the strictly greater than zero follows from considering a minimum on the support of the emission matrix. As there is no dependence on $\mathbf{w}^{\infty}, \mathbf{y}^{\infty}_{1:t-1}$ we can then conclude that there exist $\bar{m}_{t-1}, m_K, m_G > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\mu}^{\infty}_t(\mathbf{w}^{\infty},\theta)\right)\right) = 1 \quad \forall \theta \in \Theta.$$

Finally, consider a realization of $\mathbf{w}^{\infty}, \mathbf{y}^{\infty}_{1:t}$ and $i \in \mathbf{supp}\left(\boldsymbol{\pi}^{\infty}_t(\mathbf{w}^{\infty},\theta)\right)$ then:

$$\boldsymbol{\pi}^{\infty}_t(\mathbf{w}^{\infty},\theta)^{(i)} \geq \bar{m}_{t-1} m_K \sum_j G(\mathbf{w}^{\infty},\theta)^{(i,j)} (\mathbf{y}^{\infty}_t)^{(j)},$$

where the inequality follows from what we have proven above and we know that by definition and by Proposition 22:

$$\sum_j G(\mathbf{w}^{\infty},\theta)^{(i,j)} (\mathbf{y}^{\infty}_t)^{(j)} \leq 1 \text{ and } \sum_j G(\mathbf{w}^{\infty},\theta)^{(i,j)} (\mathbf{y}^{\infty}_t)^{(j)} \neq 0,$$

$\mathbb{P}$-almost surely. Following the same steps as in Proposition 11 we can conclude that there exist $\bar{m}_{t-1}, m_K, m_G > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}_t^\infty(\mathbf{w}^\infty, \theta)^{(i)} \geq \bar{m}_{t-1} m_K m_G \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right) = 1 \quad \forall \theta \in \Theta.$$

We can then set $\bar{m}_t \coloneqq \bar{m}_{t-1} m_K m_G$ and conclude that there exists $\bar{m}_t > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\pi}_t^\infty(\mathbf{w}^\infty, \theta)^{(i)} \geq \bar{m}_t \quad \forall i \in \mathbf{supp}\left(\boldsymbol{\pi}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right) = 1 \quad \forall \theta \in \Theta,$$

which closes the induction, meaning that the above statement holds for an arbitrary t. As a consequence, we also have that for any $t \geq 1$ there exists $m_t > 0$ such that:

$$\mathbb{P}\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)^{(i)} \geq m_t \quad \forall i \in \mathbf{supp}\left(\mu_t^\infty(\mathbf{w}^\infty, \theta)\right)\right) = 1 \quad \forall \theta \in \Theta,$$

with $m_t \coloneqq \bar{m}_{t-1} m_K m_G$, which concludes the proof. □

Population saturated process. The population saturated process consists of $\mathbb{O}_M$-valued disease states $(\mathbf{x}_{n,t}^\infty)_{t\geq 0}$ and $\mathbb{O}_{M+1}$-valued observations $(\mathbf{y}_{n,t}^\infty)_{t\geq 1}$, for each $n \in \mathbb{N}$.

Given $\mathbf{w}_1, \mathbf{w}_2, \ldots,$ (which are the same covariate vectors as in the data-generating process), the individuals and observations $(\mathbf{x}_{n,t}^\infty)_{t\geq 0}$, $(\mathbf{y}_{n,t}^\infty)_{t\geq 1}$, are defined to be conditionally independent across n, and distributed as follows:

$$\begin{aligned}
\mathbf{x}_{n,0}^\infty | \mathbf{w}_n &\sim \mathrm{Cat}\left(\cdot | p_0(\mathbf{w}_n, \theta^\star)\right), \\
\mathbf{x}_{n,t}^\infty | \mathbf{x}_{n,t-1}^\infty, \mathbf{w}_n &\sim \mathrm{Cat}\left(\cdot | \left[(\mathbf{x}_{t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star)\right]^\top\right), \\
\mathbf{y}_{n,t}^\infty | \mathbf{x}_{n,t}^\infty, \mathbf{w}_n &\sim \mathrm{Cat}\left(\cdot | \left[(\mathbf{x}_t^\infty)^\top G(\mathbf{w}_n, \theta^\star)\right]^\top\right).
\end{aligned} \tag{31}$$

where $\boldsymbol{\eta}_{t-1}^\infty$ is as in Corollary 14.

Proposition 24. *Under Assumption 7,9,10, for any $t \geq 1$ there exists constants $e_t > 0$ and $B_t > 0$ such that for any function f_t with $f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^N) \in [-B_t, B_t]$ and $f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^\infty) \in [-B_t, B_t]$ almost surely we have:*

$$\left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^N) - f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^\infty) \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} e_t.$$

Proof. Remark that both $\mathbf{x}_{n,t}^N$ and $\mathbf{x}_{n,t}^\infty$ are random variables that take values on $\mathbb{O}_M$, and similarly $\mathbf{y}_{n,t}^N$ and $\mathbf{y}_{n,t}^\infty$ take values on $\mathbb{O}_{M+1}$. We often write expectations over $\mathbf{x}_{n,t}^N$ or $\mathbf{x}_{n,t}^\infty$ (resp. $\mathbf{y}_{n,t}^N$ or $\mathbf{y}_{n,t}^\infty$) as summations over "x" (resp. "y"), implicitly it should be understood that we are marginalizing over $x \in \mathbb{O}_M$ (resp. $y \in \mathbb{O}_{M+1}$).

Consider the following inductive hypothesis. There exists $e_{t-1} > 0$ such that for any function f_{t-1} with $f_{t-1}(\mathbf{w}_n, \mathbf{x}_{n,1:t-1}^N) \in [-B_{t-1}, B_{t-1}]$ and $f_{t-1}(\mathbf{w}_n, \mathbf{x}_{n,1:t-1}^\infty) \in [-B_{t-1}, B_{t-1}]$ we have:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f_{t-1}(\mathbf{w}_n, \mathbf{x}_{n,0:t-1}^N) - f_{t-1}(\mathbf{w}_n, \mathbf{x}_{n,0:t-1}^\infty) \right\|_4 \leq 2B_{t-1} \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_{t-1}.$$

We start by proving it at $t-1=0$, and specifically we want:

$$\left\| \frac{1}{N} \sum_{n \in [N]} f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) - f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty) \right\|_4 \leq 2B_0 \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_0.$$

Observe that $f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) - f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty)$ are conditionally independent given $\mathbf{W}^N$ and mean zero because:

$$\mathbb{E}\left[f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) | \mathbf{W}^N \right] = \sum_{x_0} f_0(\mathbf{w}_n, x_0) p_0(\mathbf{w}_n, \theta^\star) = \mathbb{E}\left[f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty) | \mathbf{W}^N \right].$$

Moreover, they are almost surely bounded because we are assuming $f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) \in [-B_0, B_0]$ and $f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty) \in [-B_0, B_0]$ almost surely, hence:

$$\left\| f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) - f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty) \right\|_\infty \leq 2B_0.$$

meaning that by Lemma 5:

$$\left\| \frac{1}{N} \sum_{n \in [N]} f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^N) - f_0(\mathbf{w}_n, \mathbf{x}_{n,0}^\infty) \right\|_4 \leq 2B_0 \sqrt[4]{6} N^{-\frac{1}{2}},$$

so our inductive hypothesis is true at $t-1=0$ for $\bar{e}_0 = 1$.

Consider now a general time t, we can rewrite:

$$\begin{aligned}
&f_t(\mathbf{w}_n, \mathbf{x}_{n,0:t}^N) - f_t(\mathbf{w}_n, \mathbf{x}_{n,0:t}^\infty) \\
&= f_t(\mathbf{w}_n, \mathbf{x}_{n,0:t}^N) - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}_{n,0:t-1}^N, x))(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) x \\
&\quad + \sum_x f_t(\mathbf{w}_n, (\mathbf{x}_{n,0:t-1}^N, x))(\mathbf{x}_{n,t-1}^N)^\top \left[K_{\boldsymbol{\eta}_{t-1}^N(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) - K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) \right] x \\
&\quad + \sum_x f_t(\mathbf{w}_n, (\mathbf{x}_{n,0:t-1}^N, x))(\mathbf{x}_{n,t-1}^N)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) x \\
&\quad - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}_{n,0:t-1}^\infty, x))(\mathbf{x}_{n,t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) \\
&\quad + \sum_x f_t(\mathbf{w}_n, (\mathbf{x}_{n,0:t-1}^\infty, x))(\mathbf{x}_{n,t-1}^\infty)^\top K_{\boldsymbol{\eta}_{t-1}^\infty(\mathbf{w}_n, \theta^\star)}(\mathbf{w}_n, \theta^\star) x - f_t(\mathbf{w}_n, \mathbf{x}_{n,t}^\infty),
\end{aligned}$$

meaning that by Minkowski inequality:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t}) \right\|_4$$

$$= \left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x \right\|_4 \tag{32}$$

$$+ \left\| \frac{1}{N} \sum_{n\in[N]} \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) \right] x \right\|_4 \tag{33}$$

$$+ \left\| \frac{1}{N} \sum_{n\in[N]} \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x \right\|_4 \tag{34}$$

$$+ \left\| \frac{1}{N} \sum_{n\in[N]} \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,t}) \right\|_4. \tag{35}$$

Starting from (32) we can notice that:

$$\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) | \mathbf{W}^N, \mathbf{x}^N_{0:t-1} \right] = \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x,$$

moreover the random variables $f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x$ are conditionally independent across n given $\mathbf{W}^N, \mathbf{X}^N_{t-1}$ and almost surely bounded by $2B_t$, we can then conclude by Lemma 5:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) x \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}}.$$

Moving to (33) we note that by Minkowski inequality:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) \right] x \right\|_4$$

$$\leq \frac{1}{N} \sum_{n\in[N]} \left\| \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star) \right] x \right\|_4.$$

Remark that:

$$\begin{aligned}
&\left|\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right] x\right| \\
&\leq B_t \sum_x \left|(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right] x\right| \\
&\leq B_t \left\|K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right\|_\infty,
\end{aligned}$$

hence by Assumption 10 we have:

$$\begin{aligned}
&\left|\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star) - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right] x\right| \\
&\leq B_t L \left|\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star) - \boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)\right|.
\end{aligned}$$

We can then rewrite:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{n\in[N]}\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right.\right. \\
&\qquad\qquad\left.\left. - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right] x\right\|_4 \\
&\leq \frac{B_t L}{N}\sum_{n\in[N]}\left\|\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star) - \boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)\right\|_4,
\end{aligned}$$

and because of Corollary 14 and Assumption 10, we can conclude:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{n\in[N]}\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top \left[K_{\boldsymbol{\eta}^N_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right.\right. \\
&\qquad\qquad\left.\left. - K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)\right] x\right\|_4 \\
&\leq 2B_t L C\sqrt[4]{6}N^{-\frac{1}{2}}\alpha_{t-1}.
\end{aligned}$$

Consider (34), it is just enough to apply our inductive hypothesis with test function f_{t-1} given by:

$$f_{t-1}(\mathbf{w}_n, x_{0:t-1}) = \sum_x f_t(\mathbf{w}_n, (x_{0:t-1}, x))x^\top_{t-1} K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n,\theta^\star)x,$$

indeed whether we have $x_{0:t-1} = \mathbf{x}^N_{n,0:t-1}$ or $x_{0:t-1} = \mathbf{x}^\infty_{n,0:t-1}$ in both cases:

$$\begin{aligned}
&\left|\sum_x f_t(\mathbf{w}_n, (x_{0:t-1}, x))x_{t-1}^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x\right| \\
&\leq \sum_x \left|f_t(\mathbf{w}_n, (x_{0:t-1}, x))x_{t-1}^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x\right| \\
&\leq B_t \sum_x \left|x_{t-1}^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x\right| = B_t.
\end{aligned}$$

Hence:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{n\in[N]}\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^N_{n,0:t-1}, x))(\mathbf{x}^N_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x\right. \\
&\qquad\left. - \sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x\right\|_4 \\
&\leq 2B_t\sqrt[4]{6}N^{-\frac{1}{2}}\bar{e}_{t-1}.
\end{aligned}$$

The final term to work on is (35), but:

$$\begin{aligned}
\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t})|\mathbf{W}^N\right] &= \mathbb{E}\left\{\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t})|\mathbf{x}^\infty_{n,0:t-1}, \mathbf{W}^N\right]|\mathbf{W}^N\right\} \\
&= \mathbb{E}\left[\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x|\mathbf{W}^N\right],
\end{aligned}$$

and the random variables:

$$\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t}),$$

are defined to be conditionally independent given $\mathbf{W}^N$ and bounded by $2B_t$. We can then apply Lemma 5 and conclude:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{n\in[N]}\sum_x f_t(\mathbf{w}_n, (\mathbf{x}^\infty_{n,0:t-1}, x))(\mathbf{x}^\infty_{n,t-1})^\top K_{\boldsymbol{\eta}^\infty_{t-1}(\mathbf{w}_n,\theta^\star)}(\mathbf{w}_n, \theta^\star)x - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t})\right\|_4 \\
&\leq 2B_t\sqrt[4]{6}N^{-\frac{1}{2}}.
\end{aligned}$$

By putting everything together we can conclude:

$$\begin{aligned}
&\left\|\frac{1}{N}\sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t})\right\|_4 \\
&\leq 2B_t\sqrt[4]{6}N^{-\frac{1}{2}} + 2B_tLC\sqrt[4]{6}N^{-\frac{1}{2}}\bar{e}_{t-1} + 2B_t\sqrt[4]{6}N^{-\frac{1}{2}}\bar{e}_{t-1} + 2B_t\sqrt[4]{6}N^{-\frac{1}{2}} \\
&= 2B_t\sqrt[4]{6}N^{-\frac{1}{2}}\left[2 + (LC+1)\bar{e}_{t-1}\right].
\end{aligned}$$

Hence we have proven that our inductive hypothesis is valid at time t with the constant $\bar{e}_t \coloneqq [2 + (LMC + 1)\bar{e}_{t-1}]$, which tells us that for any $t \geq 1$ we have:

$$\left\| \frac{1}{N} \sum_{n \in [N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,0:t}) - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,0:t}) \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_t,$$

which also implies:

$$\left\| \frac{1}{N} \sum_{n \in [N]} f_t(\mathbf{w}_n, \mathbf{x}^N_{n,1:t}) - f_t(\mathbf{w}_n, \mathbf{x}^\infty_{n,1:t}) \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_t, \tag{36}$$

as we can rewrite $f_t(\mathbf{w}_n, \mathbf{x}^N_{n,1:t})$ as $f_t(\mathbf{w}_n, \mathbf{x}^N_{n,1:t}) \mathbb{I}(\mathbf{x}_{n,0} \in \mathbb{O}_M)$, where the indicator condition is always satisfied.

We can now move to prove the statement of the proposition, with f_t as therein:

$$\left\| \frac{1}{N} \sum_{n \in [N]} f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t}) \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} e_t.$$

Observe that:

$$\begin{aligned}
&f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t}) \\
&= f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \\
&\quad + \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \left[\prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \right] \\
&\quad + \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,t}),
\end{aligned}$$

then by Minkowski inequality:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n \in [N]} f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t}) \right\|_4 \\
&= \left\| \frac{1}{N} \sum_{n \in [N]} f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \right\|_4 \qquad (37) \\
&+ \left\| \frac{1}{N} \sum_{n \in [N]} \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \left[\prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \right] \right\|_4 \qquad (38) \\
&+ \left\| \frac{1}{N} \sum_{n \in [N]} \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t}) \right\|_4 . \qquad (39)
\end{aligned}$$

Starting from (37) we can notice that:

$$\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t})|\mathbf{W}^N, \mathbf{X}^N_{1:t}\right] = \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s,$$

moreover the random variables:

$$f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s,$$

are conditionally independent given $\mathbf{W}^N, \mathbf{X}^N_{1:t}$ and bounded by $2B_t$, hence we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{y}^N_{n,1:t}) - \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}}.$$

Consider now (38) and notice that if we consider the test function:

$$h_t(\mathbf{w}_n, x_{0:t}) = \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (x_s)^\top G(\mathbf{w}_n, \theta^\star) y_s,$$

whether we have $x_{0:t} = \mathbf{x}^N_{n,0:t}$ or $x_{0:t} = \mathbf{x}^\infty_{n,0:t}$ in both cases:

$$\left| \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (x_s)^\top G(\mathbf{w}_n, \theta^\star) y_s \right| \leq B_t \sum_{y_{1:t}} \left| \prod_{s=1}^{t} (x_s)^\top G(\mathbf{w}_n, \theta^\star) y_s \right| = B_t.$$

The term (38) is then an application of (36):

$$\left\| \frac{1}{N} \sum_{n\in[N]} \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \left[\prod_{s=1}^{t} (\mathbf{x}^N_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s \right] \right\|_4$$
$$\leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_t.$$

The last term we need to work on is (39), but:

$$\begin{aligned}
\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t})|\mathbf{W}^N\right] &= \mathbb{E}\left\{\mathbb{E}\left[f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t})|\mathbf{W}^N, \mathbf{x}^\infty_{n,0:t}\right]|\mathbf{W}^N\right\} \\
&= \sum_{y_{1:t}} \mathbb{E}\left[f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s | \mathbf{W}^N \right],
\end{aligned}$$

and the random variables:

$$\sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}^\infty_{n,s})^\top G(\mathbf{w}_n, \theta^\star) y_s - f_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,1:t}),$$

are defined to be conditionally independent given $\mathbf{W}^N$ and bounded by $2B_t$, hence we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \sum_{y_{1:t}} f_t(\mathbf{w}_n, y_{1:t}) \prod_{s=1}^{t} (\mathbf{x}_{n,s}^{\infty})^\top G(\mathbf{w}_n, \theta^\star) y_s - f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^{\infty}) \right\|_4 \leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}}.$$

By putting everything together we get:

$$\begin{aligned}
&\left\| \frac{1}{N} \sum_{n\in[N]} f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^{N}) - f_t(\mathbf{w}_n, \mathbf{y}_{n,1:t}^{\infty}) \right\|_4 \\
&\leq 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} + 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} \bar{e}_t + 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} \\
&= 2B_t \sqrt[4]{6} N^{-\frac{1}{2}} (2 + \bar{e}_t),
\end{aligned}$$

which conclude the proof under $e_t \coloneqq (2 + \bar{e}_t)$. □

Saturated CAL algorithm. We refer to the following recursion as the "saturated CAL algorithm".

$$\begin{aligned}
\hat{\boldsymbol{\pi}}_{n,0}^{\infty}(\mathbf{w}_n, \theta) &\coloneqq p_0(\mathbf{w}_n, \theta), \\
\hat{\boldsymbol{\pi}}_{n,t|t-1}^{\infty}(\mathbf{w}_n, \theta) &\coloneqq \left[\hat{\boldsymbol{\pi}}_{n,t-1}^{\infty}(\mathbf{w}_n, \theta)^\top K_{\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta) \right]^\top, \\
\hat{\boldsymbol{\mu}}_{n,t}^{\infty}(\mathbf{w}_n, \theta) &\coloneqq \left[\hat{\boldsymbol{\pi}}_{n,t|t-1}^{\infty}(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta) \right]^\top, \\
\hat{\boldsymbol{\pi}}_{n,t}^{\infty}(\mathbf{w}_n, \theta) &\coloneqq \hat{\boldsymbol{\pi}}_{n,t|t-1}^{\infty}(\mathbf{w}_n, \theta) \odot \left\{ \left[G(\mathbf{w}_n, \theta) \oslash \left(1_M \hat{\boldsymbol{\mu}}_{n,t}^{\infty}(\mathbf{w}_n, \theta)^\top \right) \right] \mathbf{y}_{n,t}^{N} \right\},
\end{aligned} \tag{40}$$

where $\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\cdot,\cdot)$ is defined in (18).

Proposition 25. *Under assumptions 7,8, for any $N \in \mathbb{N}, t \geq 1, n \in [N]$ and $\theta \in \Theta$ we have that $\sum_i \hat{\boldsymbol{\mu}}_{n,t}^{\infty}(\mathbf{w}_n, \theta)^{(i)} (\mathbf{y}_{n,t}^{N})^{(i)} \neq 0 \quad \mathbb{P}$-almost surely.*

Proof. The proof follows the same steps as Theorem 10, where we can replicate the same of Proposition 6, Proposition 7, Proposition 8, and Proposition 9 for the saturated CAL. The only difference can be found in the prediction step where instead of $K_{\tilde{\boldsymbol{\eta}}_{t-1}^{N}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)^{(j,i)} = 0$ we have $K_{\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)^{(j,i)} = 0$ which similarly implies that there exists $\eta = \bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}_n, \theta)$ such that $K_{\eta}(\mathbf{w}_n, \theta)^{(j,i)} = 0$, which allows us to follow the same argument as in Proposition 7. □

Proposition 26. *Under assumptions 6,8,10, for any $t \geq 1$ there exists $m_t > 0$ as in Proposition 11 for any $N \in \mathbb{N}$ and $n \in [N]$ such that:*

$$\mathbb{P}\left(\hat{\boldsymbol{\mu}}_{n,t}^{N}(\mathbf{w}_n, \theta)^{(i)} \geq m_t \quad \forall i \in \boldsymbol{supp}\left(\boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta) \right) \right) = 1 \quad \forall \theta \in \Theta.$$

Proof. The proof follows the same steps as the proof of Proposition 11, indeed the only difference between the CAL and the saturated CAL is the use of the saturated dynamic from which we still get:

$$\min_{j \in \mathbf{supp}\left(K_{\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)^{(\cdot,i)} \right)} K_{\bar{\boldsymbol{\eta}}_{t-1}^{\infty}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)^{(j,i)} \geq \min_{\eta \in [0,C]} \min_{j \in \mathbf{supp}\left(K_{\eta}(\mathbf{w}_n,\theta)^{(\cdot,i)} \right)} K_{\eta}(\mathbf{w}_n, \theta)^{(j,i)},$$

as in the proof of Proposition 11. □

To conclude the section we establish that the incremental terms in the logarithm of the CAL approximated the corresponding quantities from the saturated CAL algorithm.

Proposition 27. *Under assumptions 6,7,8,9,10, there exists* $\chi_t > 0$ *such that for any* $t \geq 1$, $n \in [N]$*:*

$$\left\| \log\left((\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)\right) - \log\left((\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta)\right) \right\|_4 \leq 2\sqrt[4]{6} N^{-\frac{1}{2}} \chi_t.$$

Proof. Using Proposition 11 and Proposition 26 for the lower bound, together with the fact that $\mathbf{y}_{n,t}^N$ is a one hot encoding vector and the $\boldsymbol{\mu}_{\cdot}^{\cdot}$ are probability vectors for the upper bound, we can conclude that both $(\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)$ and $(\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta)$ are such that:

$$0 < m_t \leq (\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta) \leq 1, \quad 0 < m_t \leq (\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta) \leq 1,$$

$\mathbb{P}$-almost surely.

As the the function $u \mapsto \log(u)$ is Lipschitz on the compact interval $[m_t, 1]$ we have:

$$\begin{aligned} &\left| \log\left((\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)\right) - \log\left((\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta)\right) \right| \\ &\leq \frac{1}{m_t} \left| (\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta) - (\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right|, \end{aligned}$$

$\mathbb{P}$-almost surely. From the above we can conclude:

$$\begin{aligned} &\left\| \log\left((\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)\right) - \log\left((\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta)\right) \right\|_4 \\ &\leq \frac{1}{m_t} \left\| (\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta) - (\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right\|_4, \end{aligned}$$

meaning that it remains to bound:

$$\left\| (\mathbf{y}_{n,t}^N)^\top \boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta) - (\mathbf{y}_{n,t}^N)^\top \hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right\|_4.$$

Consider as an inductive hypothesis that there exists a constant $\bar{\chi}_{t-1} > 0$ such that for any random vector $\mathbf{f}_n$ which satisfies $\|\mathbf{f}_n\|_\infty \leq B$, $\mathbb{P}$-almost surely:

$$\left\| \mathbf{f}_n^\top \boldsymbol{\pi}_{n,t-1}^N(\mathbf{w}_n, \theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}_{n,t-1}^\infty(\mathbf{w}_n, \theta) \right\|_4 \leq 2B\sqrt[4]{6} N^{-\frac{1}{2}} \bar{\chi}_{t-1}.$$

Observe that this is valid at $t - 1 = 0$ since $\boldsymbol{\pi}_{n,0}^\infty(\mathbf{w}_n, \theta) = \hat{\boldsymbol{\pi}}_{n,0}^\infty(\mathbf{w}_n, \theta)$, therefore we have:

$$\left\| \mathbf{f}_n^\top \boldsymbol{\pi}_{n,0}^N(\mathbf{w}_n, \theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}_{n,0}^\infty(\mathbf{w}_n, \theta) \right\|_4 = 0.$$

In order to show that the inductive hypothesis holds at time t, let us first consider the prediction step, and observe that:

$$\begin{aligned}
&\left\|\mathbf{f}_n^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}^\infty_{n,t|t-1}(\mathbf{w}_n,\theta)\right\|_4 \\
&\leq \left\|\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n - \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_4 \qquad (41)\\
&+ \left\|\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n - \hat{\boldsymbol{\pi}}^\infty_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_4. \qquad (42)
\end{aligned}$$

Starting from (41), we have:

$$\begin{aligned}
&\left\|\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n - \boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_4 \\
&\leq \|\mathbf{f}_n\|_\infty \left\|\widetilde{\boldsymbol{\eta}}^N_{t-1}(\mathbf{w}_n,\theta) - \bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)\right\|_4 \leq 2BC\sqrt[4]{6}N^{-\frac{1}{2}}(\gamma_{t-1}+1),
\end{aligned}$$

which follows in the same way as the proof of Proposition 17, see (23).

Moving to (42), we can apply our inductive hypothesis on the random vector: $K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n$ since $\left\|K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_\infty \leq B$, $\mathbb{P}$-almost surely, hence:

$$\begin{aligned}
&\left\|\boldsymbol{\pi}^N_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n - \hat{\boldsymbol{\pi}}^\infty_{n,t-1}(\mathbf{w}_n,\theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t-1}.
\end{aligned}$$

By putting everything together we can conclude:

$$\begin{aligned}
\left\|\mathbf{f}_n^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}^\infty_{n,t|t-1}(\mathbf{w}_n,\theta)\right\|_4 &\leq 2B[C(\gamma_{t-1}+1)+\bar{\chi}_{t-1}]\sqrt[4]{6}N^{-\frac{1}{2}} \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1},
\end{aligned}$$

where $\bar{\chi}_{t|t-1} \coloneqq C(\gamma_{t-1}+1)+\bar{\chi}_{t-1}$. Given the above we also have:

$$\begin{aligned}
&\left\|\mathbf{f}_n^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \\
&\leq \left\|\boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)^\top G(\mathbf{w}_n,\theta)\mathbf{f}_n - \hat{\boldsymbol{\pi}}^\infty_{n,t|t-1}(\mathbf{w}_n,\theta)^\top G(\mathbf{w}_n,\theta)\mathbf{f}_n\right\|_4 \qquad (43)\\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1},
\end{aligned}$$

which is just an application of the previous result on the bounded random variable $G(\mathbf{w}_n, \theta)\mathbf{f}_n$, as we know $\|G(\mathbf{w}_n, \theta)f\|_\infty \leq B$ is bounded $\mathbb{P}$-almost surely.

Now we need to work on:

$$\left\| \mathbf{f}_n^\top \boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right\|_4 .$$

Note that by using $G_{\boldsymbol{\mu}}(\mathbf{w}, \theta)$ for the matrix with elements $G_{\boldsymbol{\mu}}(\mathbf{w}, \theta)^{(i,j)} = \frac{G(\mathbf{w},\theta)^{(i,j)}}{\boldsymbol{\mu}^{(j)}}$ where $\frac{0}{0} = 0$ by convention, we can rewrite everything in a more compact way:

$$\begin{aligned}
&\left\| \mathbf{f}_n^\top \boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right\|_4 \\
&= \Bigg\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta)\right]^\top \left[G_{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \\
&\qquad - \left[\mathbf{f}_n \odot \hat{\boldsymbol{\pi}}_{t|t-1}^\infty(\mathbf{w}_n, \theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \Bigg\|_4 .
\end{aligned}$$

By Minkowski inequality we can then conclude:

$$\begin{aligned}
&\left\| \mathbf{f}_n^\top \boldsymbol{\pi}_{n,t}^N(\mathbf{w}_n, \theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}_{n,t}^\infty(\mathbf{w}_n, \theta) \right\|_4 \\
&\leq \Bigg\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta)\right]^\top \left[G_{\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \\
&\qquad - \left[\mathbf{f}_n \odot \boldsymbol{\pi}_{t|t-1}^N(\mathbf{w}_n, \theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \Bigg\|_4 \qquad (44) \\
&+ \Bigg\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}_{n,t|t-1}^N(\mathbf{w}_n, \theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \\
&\qquad - \left[\mathbf{f}_n \odot \hat{\boldsymbol{\pi}}_{t|t-1}^\infty(\mathbf{w}_n, \theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}_{n,t}^\infty(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\mathbf{y}_{n,t}^N\right] \Bigg\|_4 . \qquad (45)
\end{aligned}$$

Starting from (44), we remark that:

$$x^\top G_\mu b - x^\top G_{\widetilde{\mu}} b = \sum_{i,j} x^{(i)} y^{(j)} \frac{G^{(i,j)}\widetilde{\mu}^{(j)} - G^{(i,j)}\mu^{(j)}}{\widetilde{\mu}^{(j)}\mu^{(j)}} = \sum_{i,j} x^{(i)} \frac{y^{(j)}}{\widetilde{\mu}^{(j)}\mu^{(j)}} G^{(i,j)} \left(\widetilde{\mu}^{(j)} - \mu^{(j)}\right) .$$

Hence we can reformulate (44):

$$\begin{aligned}
&\Big\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Big\|_4 \\
&= \Bigg\| \left\{\left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta)\right\}^\top \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right]^\top \\
&\qquad \left[\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta) - \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right] \Bigg\|_4,
\end{aligned}$$

from which we can notice that for any $\mathbf{w}_n$:

$$\begin{aligned}
&\left\|\left\{\left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta)\right\} \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right]\right\|_\infty \\
&\leq \|\mathbf{f}_n\|_\infty \left\|\mathbf{y}^N_{n,t} \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_\infty,
\end{aligned}$$

where the first step follows from $\boldsymbol{\mu}^N_{n,t}(\mathbf{w},\theta) = \left[\boldsymbol{\pi}^N_{t|t-1}(\mathbf{w},\theta)^\top G(\mathbf{w},\theta)\right]^\top$ and the elementwise ratio $\mathbf{y}^N_{n,t} \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)$ is well-defined because of Proposition 25.

As from Proposition 26 we know that the saturated CAL is almost surely bounded we have:

$$\begin{aligned}
&\left\|\left\{\left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta)\right\} \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right]\right\|_\infty \\
&\leq \frac{\|\mathbf{f}_n\|_\infty}{m_t} \leq \frac{B}{m_t},
\end{aligned}$$

$\mathbb{P}$-almost surely. Hence we can apply (43) as we are considering an almost surely bounded random vector:

$$\left\{\left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta)\right\} \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right]$$

and conclude:

$$\begin{aligned}
&\Big\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Big\|_4 \\
&= \Bigg\| \left\{\left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top G(\mathbf{w}_n,\theta)\right\}^\top \odot \left[\mathbf{y}^N_{n,t} \oslash \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) \oslash \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right]^\top \\
&\qquad \left[\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta) - \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right] \Bigg\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\frac{\chi_t}{m_t}.
\end{aligned}$$

Moving to (45), we can observe that:

$$\begin{aligned}
&\Big\| \left[\mathbf{f}_n \odot \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \\
&\qquad - \left[\mathbf{f}_n \odot \hat{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right]^\top \left[G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right] \Big\|_4 \\
&= \left\| \left[\mathbf{f}_n \odot G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}\right]^\top \left[\boldsymbol{\pi}^N_{t|t-1}(\mathbf{w}_n,\theta) - \hat{\boldsymbol{\pi}}^\infty_{t|t-1}(\mathbf{w}_n,\theta)\right] \right\|_4 \\
&\leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1},
\end{aligned}$$

where we can apply (43) to the test vector $\mathbf{f}_n \odot G_{\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)}(\mathbf{w}_n,\theta)\mathbf{y}^N_{n,t}$, as it is almost surely bounded.

By putting everything together we can conclude:

$$\left\|\mathbf{f}_n^\top \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\frac{\chi_t}{m_t} + 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1},$$

which closes our inductive hypothesis by setting $\bar{\chi}_t = \frac{\chi_t}{m_t} + \bar{\chi}_{t|t-1}$, and so we can conclude that for any $t \geq 1$ there exists $\bar{\chi}_{t|t-1}, \bar{\chi}_t$ such that for any test random vector $\mathbf{f}_n$, with $\|\mathbf{f}_n\|_\infty \leq B$, $\mathbb{P}$-almost surely:

- $\left\|\mathbf{f}_n^\top \boldsymbol{\pi}^N_{n,t|t-1}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}^\infty_{n,t|t-1}(\mathbf{w}_n,\theta)\right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1}$;
- $\left\|\mathbf{f}_n^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1}$;
- $\left\|\mathbf{f}_n^\top \boldsymbol{\pi}^N_{n,t}(\mathbf{w}_n,\theta) - \mathbf{f}_n^\top \hat{\boldsymbol{\pi}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \leq 2B\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_t$.

As $\mathbf{y}^N_{n,t}$ is almost surely bounded $\left\|\mathbf{y}^N_{n,t}\right\|_\infty \leq 1$ we can conclude:

$$\left\|(\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) - (\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \leq 2\sqrt[4]{6}N^{-\frac{1}{2}}\bar{\chi}_{t|t-1},$$

so we can conclude our proof by setting $\chi_t \coloneqq \frac{\bar{\chi}_{t|t-1}}{m_t}$ as:

$$\begin{aligned}
&\Big\|\log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right) - \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right)\Big\|_4 \\
&\leq \frac{1}{m_t}\left\|(\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta) - (\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right\|_4 \leq 2\sqrt[4]{6}N^{-\frac{1}{2}}\frac{\bar{\chi}_{t|t-1}}{m_t}.
\end{aligned}$$

□

C.5 Strong consistency

Contrast function. In Proposition 28 below we establish the convergence of the rescaled logarithm of the CAL, $\frac{\ell^N_{1:T}(\theta)}{N}$ in the large population limit, and in Theorem 29 show how $\frac{\ell^N_{1:T}(\theta)}{N} - \frac{\ell^N_{1:T}(\theta^\star)}{N}$ converges to a contrast function.

Proposition 28. *Under assumptions 6,7,8,9,10 and for any $\theta \in \Theta$ let:*

$$\ell^N_t(\theta) \coloneqq \sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)\right),$$

then $N^{-1}\ell^N_t(\theta)$ converges to $\mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta)\right)\right]$ as $N \to \infty$, $\mathbb{P}$-almost surely. Moreover:

$$\mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta)\right)\right] = \mathbb{E}\left[\boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta)\right)\right].$$

Proof. Consider the definition of $\boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)$ from (6), the definition of $\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n, \theta)$ from (40), the definition of $\boldsymbol{\mu}^\infty_t(w, \theta)$ from (30), and define:

$$\begin{aligned}
&\boldsymbol{\pi}^\infty_{n,0}(\mathbf{w}_n, \theta) \coloneqq p_0(\mathbf{w}_n, \theta),\\
&\boldsymbol{\pi}^\infty_{n,t|t-1}(\mathbf{w}_n, \theta) \coloneqq \left[\boldsymbol{\pi}^\infty_{n,t-1}(\mathbf{w}_n, \theta)^\top K_{\bar{\boldsymbol{\eta}}^\infty_{t-1}(\mathbf{w}_n,\theta)}(\mathbf{w}_n, \theta)\right]^\top,\\
&\boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n, \theta) \coloneqq \left[\boldsymbol{\pi}^\infty_{n,t|t-1}(\mathbf{w}_n, \theta)^\top G(\mathbf{w}_n, \theta)\right]^\top,\\
&\boldsymbol{\pi}^\infty_{n,t}(\mathbf{w}_n, \theta) \coloneqq \boldsymbol{\pi}^\infty_{n,t|t-1}(\mathbf{w}_n, \theta) \odot \left\{\left[G(\mathbf{w}_n, \theta) \oslash \left(1_M \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n, \theta)^\top\right)\right] \mathbf{y}^\infty_{n,t}\right\},
\end{aligned} \tag{46}$$

where $\bar{\boldsymbol{\eta}}^\infty_{t-1}(\cdot,\cdot)$ is defined in (18) and $\mathbf{y}^\infty_{n,t}$ is defined in (31).

We can then consider the following decomposition:

$$\begin{aligned}
&\frac{1}{N}\sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta)\right)\right]\\
&= \frac{1}{N}\sum_{n\in[N]} \left(\log\left(\mathbf{y}^N_{n,t}\right)^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n, \theta)\right) - \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n, \theta)\right)\\
&+ \frac{1}{N}\sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n, \theta)\right) - \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n, \theta)\right)\\
&+ \frac{1}{N}\sum_{n\in[N]} \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty, \theta)\right)\right],
\end{aligned}$$

by Minkowski inequality we conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty,\theta)\right)\right] \right\|_4$$

$$\leq \left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right) - \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) \right\|_4 \tag{47}$$

$$+ \left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) - \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) \right\|_4 \tag{48}$$

$$+ \left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty,\theta)\right)\right] \right\|_4 . \tag{49}$$

Starting from (47), we can apply Proposition 27 to obtain:

$$\left\| \log\left((\mathbf{y}^N_{n,t})^\top \boldsymbol{\mu}^N_{n,t}(\mathbf{w}_n,\theta)\right) - \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) \right\|_4 \leq 2\sqrt[4]{6} N^{-\frac{1}{2}} \chi_t.$$

Consider now (48), it is important to observe that when we compare (40) with (46), the only difference between the two is that (40) uses $\mathbf{y}^N_{n,t}$ while (46) uses the population saturated process $\mathbf{y}^\infty_{n,t}$. Hence, if we look at $(\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)$ as a function of $\mathbf{y}^N_{n,1:t}$ and $(\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)$ as a function of $\mathbf{y}^\infty_{n,1:t}$ we are considering the same function evaluated in different arguments; we can define for a fixed θ the function $h^\theta_t(w, y_{1:t})$ which is such that $h^\theta_t(\mathbf{w}_n, \mathbf{y}^N_{n,t}) = (\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)$ and $h^\theta_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,t}) = (\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)$. Because of Proposition 26, for all $i \in [M]$ we have that $\hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)^{(i)} > m_t$, meaning that $h^\theta_t(\mathbf{w}_n, \mathbf{y}^N_{n,t}) \in [m_t, 1]$ almost surely. Similarly, because of Proposition 23, for all $i \in [M]$ we have that $\boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)^{(i)} > m_t$, meaning that $h^\theta_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,t}) \in [m_t, 1]$ almost surely. This follows as (31) consists of repeating (29) N times. We can then consider $\log(h^\theta_t(\mathbf{w}_n, \mathbf{y}^N_{n,t})) \in [\log(m_t), 0]$ and $\log(h^\theta_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,t})) \in [\log(m_t), 0]$ almost surely and apply Proposition 24 as both $\log(h^\theta_t(\mathbf{w}_n, \mathbf{y}^N_{n,t}))$ and $\log(h^\theta_t(\mathbf{w}_n, \mathbf{y}^\infty_{n,t}))$ are almost surely in $[\log(m_t), |\log(m_t)|]$:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}^N_{n,t})^\top \hat{\boldsymbol{\mu}}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) - \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}_n,\theta)\right) \right\|_4 \leq 2\,|\log(m_t)|\,\sqrt[4]{6} N^{-\frac{1}{2}} e_t.$$

For (49), note that:

$$\mathbb{E}\left[\log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)\right)\right] = \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty,\theta)\right)\right],$$

see the definitions of $\mathbf{y}^\infty_{n,t}, \mathbf{y}^\infty_t, \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty,\theta)$ in Section C.4 and the definition of $\boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)$ in (46) at the beginning of the proof. Moreover,

$$\left\| \log\left((\mathbf{y}^\infty_{n,t})^\top \boldsymbol{\mu}^\infty_{n,t}(\mathbf{w}_n,\theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}^\infty_t)^\top \boldsymbol{\mu}^\infty_t(\mathbf{w}^\infty,\theta)\right)\right] \right\| \leq 2\,|\log(m_t)|\,,$$

$\mathbb{P}$-almost surely, because of the previous reasoning. As we are considering averages of random variables that are mean zero, bounded, and independent, we can apply Lemma 5 and conclude:

$$\left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{\infty})^{\top} \boldsymbol{\mu}_{n,t}^{\infty}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right\|_4 \leq 2\sqrt[4]{6} N^{-\frac{1}{2}} \left|\log(m_t)\right|.$$

By putting everything together we obtain:

$$\begin{aligned} &\left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right\|_4 \\ &\leq 2\sqrt[4]{6} N^{-\frac{1}{2}} \chi_t + 2\left|\log(m_t)\right| \sqrt[4]{6} N^{-\frac{1}{2}} e_t + 2\sqrt[4]{6} N^{-\frac{1}{2}} \left|\log(m_t)\right| \\ &= 2\sqrt[4]{6} N^{-\frac{1}{2}} \left[\chi_t + \left|\log(m_t)\right| (e_t + 1)\right]. \end{aligned} \tag{50}$$

By applying Markov's inequality, for any $\iota > 0$,

$$\begin{aligned} &\mathbb{P}\left(\left| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right| > \iota \right) \\ &= \mathbb{P}\left(\left| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right|^4 > \iota^4 \right) \\ &\leq \frac{1}{\iota^4} \mathbb{E}\left[\left| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right|^4 \right] \\ &= \iota^{-4} \left\| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right\|_4^4 \\ &\leq 6\frac{2^4}{\iota^4} N^{-2} \left[\chi_t + \left|\log(m_t)\right| (e_t + 1)\right]^4, \end{aligned}$$

where the last bound follows from (50). We then conclude that:

$$\sum_{N=1}^{\infty} \mathbb{P}\left(\left| \frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \right| > \iota \right) < \infty,$$

hence by the Borel-Cantelli lemma:

$$\frac{1}{N} \sum_{n\in[N]} \log\left((\mathbf{y}_{n,t}^{N})^{\top} \boldsymbol{\mu}_{n,t}^{N}(\mathbf{w}_n, \theta)\right) - \mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right] \to 0,$$

as $N \to \infty$, $\mathbb{P}$-almost surely, or equivalently $\frac{1}{N}\ell_t^N(\theta)$ converges to $\mathbb{E}\left[\log\left((\mathbf{y}_t^{\infty})^{\top} \boldsymbol{\mu}_t^{\infty}(\mathbf{w}^{\infty}, \theta)\right)\right]$ as $N \to \infty$, $\mathbb{P}$-almost surely.

For the final statement of the proposition it is enough to observe that as $\mathbf{y}_t^\infty$ is a one-hot encoding vector:

$$\log\left((\mathbf{y}_t^\infty)^\top \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right) = (\mathbf{y}_t^\infty)^\top \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)$$

under the convention $0 \log 0 = 0$, and from Proposition 21 we have:

$$\mathbf{y}_t^\infty | \mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot | \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right).$$

Hence, by the tower rule:

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{y}_t^\infty)^\top \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right] &= \mathbb{E}\left\{\mathbb{E}\left[(\mathbf{y}_t^\infty)^\top \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right) | \mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty\right]\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[(\mathbf{y}_t^\infty)^\top | \mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty\right] \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right\} \\
&= \mathbb{E}\left[\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right].
\end{aligned}$$

□

Because of Proposition 28, we can conclude that the CAL has a contrast function which is an expected Kullback-Leibler divergence, as in the following theorem. Recall from (30) the definition $\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)$.

Theorem 29. *Under assumptions 6,7,8,9,10, for any $T \geq 1$ and $\theta \in \Theta$ let:*

$$\ell_{1:T}^N(\theta) := \sum_{t=1}^T \ell_t^N(\theta)$$

then:

$$\frac{\ell_{1:T}^N(\theta)}{N} - \frac{\ell_{1:T}^N(\theta^\star)}{N} \to \mathcal{C}_T(\theta, \theta^\star),$$

as $N \to \infty$, $\mathbb{P}$-almost surely, where:

$$\mathcal{C}_T(\theta, \theta^\star) := -\sum_{t=1}^T \mathbb{E}\left\{\mathbf{KL}\left[Cat\left(\cdot | \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right) || Cat\left(\cdot | \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right]\right\}. \tag{51}$$

Moreover:

$$\theta^\star \in \Theta^\star := \operatorname*{argmax}_{\theta \in \Theta} \mathcal{C}_T(\theta, \theta^\star). \tag{52}$$

Proof. Because of Proposition 28:

$$\frac{\ell_{1:T}^N(\theta)}{N} - \frac{\ell_{1:T}^N(\theta^\star)}{N} \to \sum_{t=1}^T \int \mathbb{E}\left[\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\frac{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)}{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)}\right) \middle| \mathbf{w}^\infty = w\right] \Gamma(dw),$$

as $N \to \infty$, $\mathbb{P}$-almost surely, where we notice that:

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\frac{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)}{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)}\right) \middle| \mathbf{w}^\infty = w\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left[\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\frac{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)}{\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)}\right) \middle| \mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty = w\right] \middle| \mathbf{w}^\infty = w\right\} \\
&= -\mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot | \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right) || \mathrm{Cat}\left(\cdot | \boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right] | \mathbf{w}^\infty = w\right\},
\end{aligned}$$

hence:

$$\frac{\ell_{1:T}^N(\theta)}{N} - \frac{\ell_{1:T}^N(\theta^\star)}{N}$$
$$\to -\sum_{t=1}^{T} \int \mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)\right)\right]|\mathbf{w}^\infty = w\right\}\Gamma(dw),$$

as $N \to \infty$, $\mathbb{P}$-almost surely, from which we conclude the first part of the proof as:

$$\mathcal{C}_T(\theta,\theta^\star) = -\sum_{t=1}^{T} \int \mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)\right)\right]|\mathbf{w}^\infty = w\right\}\Gamma(dw).$$

The KL-divergence is always greater than or equal to zero, and equal to zero if and only if the two distributions are equal. Hence the maximal value of the negative KL-divergence is zero, and we have:

$$\theta^\star \in \Theta^\star = \operatorname*{argmax}_{\theta\in\Theta} \mathcal{C}_T(\theta,\theta^\star),$$

which concludes the proof. □

Uniform almost sure convergence. Theorem 29 proves the convergence $\frac{\ell_{1:T}^N(\theta)}{N} - \frac{\ell_{1:T}^N(\theta^\star)}{N} \to \mathcal{C}_T(\theta,\theta^\star)$ *pointwise* in θ. In order to make statements about the convergence of the maximizer of $\ell_{1:T}^N(\theta)$ in the same vein as Whitehouse et al. (2023), we must show this convergence is uniform. To proceed we will use the following results.

Definition 30. *Let $(\mathcal{H}_N)_{N\geq 1}$ be a sequence of random functions $\mathcal{H}_N : \theta \in \Theta \mapsto \mathcal{H}_N(\theta) \in \mathbb{R}$ where Θ is a metric space. We say that $(\mathcal{H}_N)_{N\geq 1}$ are stochastically equicontinuous if there exists an event E of probability 1, such that for all $\iota > 0$ and $\omega \in E$, there exists $N(\omega)$ and $\delta > 0$ such that $N > N(\omega)$ implies:*

$$\sup_{|\theta_1-\theta_2|<\delta} |\mathcal{H}_N(\omega,\theta_1) - \mathcal{H}_N(\omega,\theta_2)| < \iota.$$

Lemma 31. *Assume Θ is a compact metric space and let $(\mathcal{H}_N)_{N\geq 1}$ be a sequence of random functions $\mathcal{H}_N : \theta \in \Theta \to \mathcal{H}_N(\theta) \in \mathbb{R}$. If there exists a continuous function $\mathcal{H}$ such that for all $\theta \in \Theta$ we have $|\mathcal{H}_N(\theta) - \mathcal{H}(\theta)| \overset{a.s.}{\to} 0$, and $(\mathcal{H}_N)_{N\geq 1}$ are stochastically equicontinuous, then:*

$$\sup_{\theta\in\Theta} |\mathcal{H}_N(\theta) - \mathcal{H}(\theta)| \overset{a.s.}{\to} 0.$$

That is $\mathcal{H}_N(\theta)$ converges to $\mathcal{H}(\theta)$ almost surely as $N \to \infty$, uniformly in θ.

Proof. See Andrews (1992). □

Lemma 32. *Under Assumptions 6,7,8,9,10,*

$$\sup_{\theta\in\Theta} \left| \frac{\ell_{1:T}^N(\theta)}{N} - \frac{\ell_{1:T}^N(\theta^\star)}{N} - \mathcal{C}_T(\theta,\theta^\star) \right| \to 0,$$

$\mathbb{P}$-almost surely.

Proof. By Theorem 29 we have pointwise convergence. Hence, by Lemma 31 it is enough to show that

$$\mathcal{C}_t^N(\theta) \coloneqq \frac{1}{N} \sum_{n\in[N]} (\mathbf{y}_{n,t}^N)^\top \log\left(\boldsymbol{\mu}_{n,t}^N(\mathbf{w}_n, \theta)\right), \tag{53}$$

is stochastically equicontinuous. Define

$$\mathcal{C}_t^\infty(\theta) \coloneqq \int \mathbb{E}\left[\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)^\top \log\left(\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right) | \mathbf{w}^\infty = \mathbf{w}\right] \Gamma(d\mathbf{w}).$$

Let $\theta_1, \theta_2 \in \Theta$ and consider the decomposition

$$\begin{aligned} |\mathcal{C}_t^N(\theta_1) - \mathcal{C}_t^N(\theta_2)| \leq & |\mathcal{C}_t^N(\theta_1) - \mathcal{C}_t^\infty(\theta_1)| && (54)\\ & + |\mathcal{C}_t^N(\theta_2) - \mathcal{C}_t^\infty(\theta_2)| && (55)\\ & + |\mathcal{C}_t^\infty(\theta_1) - \mathcal{C}_t^\infty(\theta_2)|. && (56) \end{aligned}$$

Note that all of these quantities are well defined by Theorem 10 and by the fact that $\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta) \neq 0$ because of Proposition 22.

Let $E \subset \Omega$ such that $\mathbb{P}(E) = 1$. Let $\theta, \theta^\star \in \Theta$, $\omega \in E$, and $\iota > 0$. By Theorem 29 we have almost sure convergence of (54) and (55) to 0, hence there exists an $N(\omega)$ such that for all $N > N(\omega)$ these terms are bounded by $\iota/3$.

It remains to show that there exists a δ such that for $|\theta_1 - \theta_2| < \delta$ implies that (56) is bounded by $\iota/3$. This follows directly from noticing that $\mathcal{C}_t^\infty(\theta)$ comprises a composition of continuous functions of our model quantities p_0, K, and G, which are themselves continuous functions of θ because of Assumption 6. □

Theorem 33. *Let 6,7,8,9,10 hold and let $\hat{\theta}_N$ be a maximizer of $\ell_{1:T}^N(\theta)$. Then $\hat{\theta}_N$ converges to $\Theta^\star$ as $N \to \infty$, $\mathbb{P}$-almost surely.*

Proof. The proof follows in the same manner as that of Theorem 1 in Whitehouse et al. (2023), we include it for completeness.

Let $\mathcal{C}_T(\theta^\star, \theta)$ be as defined by (51) and let $\mathcal{C}_T^N(\theta) = \sum_{t=1}^T \mathcal{C}_t^N(\theta)$ be as in Equation (53). We have that $\mathcal{C}_T^N(\hat{\theta}_N) \geq \mathcal{C}_T^N(\theta)$ for all $\theta \in \Theta^\star$. Furthermore $\mathcal{C}_T(\theta^\star, \theta^\star) - \mathcal{C}_T(\theta^\star, \theta) \geq 0$ for all $\theta \in \Theta$. We can combine these inequalities to obtain:

$$\begin{aligned} 0 &\leq \mathcal{C}_T(\theta^\star, \theta^\star) - \mathcal{C}_T(\theta^\star, \hat{\theta}_n) \\ &= \mathcal{C}_T(\theta^\star, \theta^\star) - \mathcal{C}_T^N(\theta^\star) + \mathcal{C}_T^N(\theta^\star) - \mathcal{C}_T^N(\hat{\theta}_n) + \mathcal{C}_T^N(\hat{\theta}_n) - \mathcal{C}_T(\theta^\star, \hat{\theta}_n) \\ &\leq 2 \sup_{\theta\in\Theta} \left|\mathcal{C}_T(\theta^\star, \theta) - \mathcal{C}_T^N(\theta)\right| \to 0 \quad \mathbb{P}\text{-almost surely,} \end{aligned} \tag{57}$$

by Lemma 32. Hence $\mathcal{C}_T(\theta^\star, \hat{\theta}_n) \to \mathcal{C}_T(\theta^\star, \theta^\star)$ $\mathbb{P}$-almost surely.

Now assume for purposes of contradiction that there is some positive probability that $\hat{\theta}_n$ does not converge to the set $\Theta^\star$, i.e. assume that there is an event $E \subset \Omega$ with $\mathbb{P}(E) > 0$ such that for all $\omega \in E$ there exists a $\delta > 0$ such that for infinitely many $n \in \mathbb{N}$ we have $\hat{\theta}_n(\omega)$ is not in

the open neighborhood $B_\delta(\Theta^\star) = \{\theta \in \Theta : \exists \theta' \in \Theta^\star : \|\theta - \theta'\| < \delta\}$. Since Θ is compact, the set $B_\delta(\Theta^\star)^c = \Theta \setminus B_\delta(\Theta^\star)$ is closed, bounded, and therefore compact. Furthermore, $\mathcal{C}_T(\theta^\star, \theta)$ is continuous in θ. By the extreme value theorem this means that there exists a $\theta' \in B_\delta(\Theta^\star)^c$ such that for all $\theta \in B_\delta(\Theta^\star)^c$:

$$\mathcal{C}_T(\theta^\star, \theta) \leq \mathcal{C}_T(\theta^\star, \theta')$$

Furthermore, since $\theta' \notin \Theta^\star$ there exists $\iota > 0$ such that:

$$\mathcal{C}_T(\theta^\star, \theta') < \mathcal{C}_T(\theta^\star, \theta^\star) - \iota.$$

By our assumption we have for each $\omega \in E$ there are infinitely many $n \in \mathbb{N}$ such that $\hat{\theta}_n(\omega) \in B_\delta(\Theta^\star)^c$. But this implies that for each $\omega \in E$ there are infinitely many $n \in \mathbb{N}$ such that:

$$\mathcal{C}_T(\theta^\star, \hat{\theta}_n(\omega)) \leq \mathcal{C}_T(\theta^\star, \theta') < \mathcal{C}_T(\theta^\star, \theta^\star) - \iota,$$

$$\implies |\mathcal{C}_T(\theta^\star, \theta^\star) - \mathcal{C}_T(\theta^\star, \hat{\theta}_n(\omega))| > \iota,$$

which contradicts (57). Hence we must have that $\hat{\theta}_n$ converges to the set $\Theta^\star$ $\mathbb{P}$-almost surely. □

C.5.1 Identifiability

Definition 34. *Let $\{\mathbf{y}_t^\infty\}_{t\geq 1}$ be generated according to the process defined by equations (29) with data-generating parameter $\theta^\star \in \Theta$. Denote the law of $\{\mathbf{y}_t^\infty\}_{t\geq 1}$ conditional on $\mathbf{w}^\infty = w$ with $\mathbb{P}_\infty^{\theta^\star, w}$.*

Lemma 35. *Let $\theta^\star \in \Theta$. For any $\theta_1, \theta_2 \in \Theta^\star := \operatorname{argmax}_{\theta\in\Theta} \mathcal{C}_T(\theta, \theta^\star)$ we have that $\mathbb{P}_\infty^{\theta_1, w} = \mathbb{P}_\infty^{\theta_2, w}$ for Γ-almost all $w \in \mathbb{W}$.*

Proof. Recall from (51) the definition of the contrast function:

$$\mathcal{C}_T(\theta, \theta^\star) := -\sum_{t=1}^{T} \mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right) || \mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right]\right\}.$$

where $\theta^\star$ is the DGP and θ is a candidate parameter. With $\theta^\star$ fixed, we want to characterize the set of maximizers of $\mathcal{C}_T(\theta, \theta^\star)$ in the first argument, i.e. the set $\Theta^\star := \operatorname{argmax}_{\theta\in\Theta} \mathcal{C}_T(\theta, \theta^\star)$. Let $w \in \mathbb{W}$ and consider the conditional expectation

$$\mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right) || \mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right)\right] | \mathbf{w}^\infty = w\right\},$$

this is the expectation of a KL-divergence between categorical distributions parameterized by the random vectors $\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)$ and $\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)$. Recall that according to the recursive definition of these vectors given in (30), conditional on $\mathbf{w}^\infty = w$ these vectors are random solely as functions of $\mathbf{y}_{1:t-1}^\infty \sim \mathbb{P}_\infty^{\theta^\star, w}$, with no other sources of stochasticity. Hence this conditional expectation is a summation over $y_{1:t-1} \in \mathbb{O}_{M+1}^{t-1}$:

$$\begin{aligned} &\mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta^\star)\right) || \mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty, \theta)\right) | \mathbf{w}^\infty = w\right]\right\} \\ &= \sum_{y_{1:t-1}} \mathbb{P}_\infty^{\theta^\star, w}(\mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) \mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\mu_t^\infty(w, \theta^\star)\right) || \mathrm{Cat}\left(\cdot|\mu_t^\infty(w, \theta)\right)\right] \end{aligned} \tag{58}$$

where the unbolded $\mu_t^\infty(w,\theta^\star)$ and $\mu_t^\infty(w,\theta)$ (similarly to the process defined by Equation (29)) are calculated with the recursions:

$$\begin{aligned}
&\pi_0^\infty(w,\theta) \coloneqq p_0(w,\theta),\\
&\bar{\eta}_{t-1}^\infty(w,\theta) = \int d(w,\widetilde{w},\theta)^\top \bar{\pi}_{t-1}^\infty(\widetilde{w},\theta)\Gamma(d\widetilde{w}),\\
&\pi_{t|t-1}^\infty(w,\theta) \coloneqq \left[\pi_{t-1}^\infty(w,\theta)^\top K_{\bar{\eta}_{t-1}^\infty(w,\theta)}(w,\theta)\right]^\top,\\
&\mu_t^\infty(w,\theta) \coloneqq \left[\pi_{t|t-1}^\infty(w,\theta)^\top G(w,\theta)\right]^\top,\\
&\pi_t^\infty(w,\theta) \coloneqq \pi_{t|t-1}^\infty(w,\theta) \odot \left\{\left[G(w,\theta) \oslash \left(1_M \mu_t^\infty(w,\theta)^\top\right)\right] y_t\right\},
\end{aligned}$$

where the dependence of various quantities on $y_{1:t}$ is not shown in the notation.

By properties of the KL-divergence we have that

$$\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\mu_t^\infty(w,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\mu_t^\infty(w,\theta)\right)\right] = 0 \iff \mu_t^\infty(w,\theta) = \mu_t^\infty(w,\theta^\star),$$

and

$$\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\mu_t^\infty(w,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\mu_t^\infty(w,\theta)\right)\right] > 0 \iff \mu_t^\infty(w,\theta) \neq \mu_t^\infty(w,\theta^\star),$$

which makes it clear that $\theta^\star \in \Theta^\star$, as already mentioned in (52). It then follows from Equation (58) that if $\theta \in \Theta^\star$ then for $\mathbb{P}_\infty^{\theta^\star,w}$-almost all paths $y_{1:t-1} \in \mathbb{O}_{M+1}^{t-1}$ we must have $\mu_t^\infty(w,\theta) = \mu_t^\infty(w,\theta^\star)$. Now, considering the full contrast function we have by the tower rule:

$$\begin{aligned}
\mathcal{C}_T(\theta,\theta^\star) &\coloneqq -\sum_{t=1}^T \mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)\right)\right]\right\}\\
&= -\sum_{t=1}^T \mathbb{E}\left\{\mathbb{E}\left\{\mathbf{KL}\left[\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)\right)||\mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta)\right)\right]|\mathbf{w}^\infty\right\}\right\}.
\end{aligned}$$

Here we are summing over $t \in [T]$ and taking the expectation of (58) over $\mathbf{w}^\infty \sim \Gamma$. It therefore follows that if $\theta \in \Theta^\star$ then for all $t \in [T]$, $\Gamma-$ almost all $w \in \mathbb{W}$, and $\mathbb{P}_\infty^{\theta^\star,w}$-almost all paths $y_{1:t-1} \in \mathbb{O}_{M+1}^{t-1}$, we must have that $\mu_t^\infty(w,\theta) = \mu_t^\infty(w,\theta^\star)$.

To complete the proof, recall that by Proposition 21 if $\mathbf{y}_{1:T} \sim \mathbb{P}_\infty^{\theta^\star,w}$ we have the conditional distributions $\mathbf{y}_t^\infty|\mathbf{y}_{1:t-1}^\infty, \mathbf{w}^\infty \sim \mathrm{Cat}\left(\cdot|\boldsymbol{\mu}_t^\infty(\mathbf{w}^\infty,\theta^\star)\right)$. Hence for $\Gamma-$almost all $w \in \mathbb{W}$, $\mathbb{P}_\infty^{\theta^\star,w}$-almost all

paths $y_{1:T} \in \mathbb{O}_{M+1}^T$ and any $\theta_1, \theta_2 \in \Theta^\star$ we have

$$\begin{aligned}
\mathbb{P}_\infty^{\theta_1,w}(\mathbf{y}_{1:T}^\infty = y_{1:T}) &= \prod_{t=1}^T \mathbb{P}_\infty^{\theta_1,w}(\mathbf{y}_t^\infty = y_t | \mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) \\
&= \prod_{t=1}^T \mathrm{Cat}\left(y_t | \mu_t^\infty(w, \theta_1)\right) \\
&= \prod_{t=1}^T \mathrm{Cat}\left(y_t | \mu_t^\infty(w, \theta^\star)\right) \\
&= \prod_{t=1}^T \mathrm{Cat}\left(y_t | \mu_t^\infty(w, \theta_2)\right) \\
&= \prod_{t=1}^T \mathbb{P}_\infty^{\theta_2,w}(\mathbf{y}_t^\infty = y_t | \mathbf{y}_{1:t-1}^\infty = y_{1:t-1}) \\
&= \mathbb{P}_\infty^{\theta_2,w}(\mathbf{y}_{1:T}^\infty = y_{1:T}),
\end{aligned}$$

where we used that $\theta \in \Theta^\star$ implies $\mu_t^\infty(w, \theta) = \mu_t^\infty(w, \theta^\star)$. Assumption 8 ensures that the support of $\mathbb{P}_\infty^{\theta,w}$ does not depend on θ, and hence $\mathbb{P}_\infty^{\theta_1,w} = \mathbb{P}_\infty^{\theta_2,w}$ for $\Gamma-$almost all $w \in \mathbb{W}$. □

D Experiments

D.1 Computational considerations

Algorithm 1 requires only N repetitions of simple linear algebra operations on M-dimensional vectors and $M \times M$ matrices at each time step. The resulting computational cost is $\mathcal{O}(TNM^2 + TN\mathcal{C}_\eta(N))$, where T and N arise from recursive operations over time and individuals, M^2 accounts for vector-matrix operations, and $\mathcal{C}_\eta(N)$ is the cost of evaluating the η function for a fixed population size N.

Table 4: Memory and running time summary. $\mathcal{C}_\eta(N)$ is the model specific cost of evaluating η for a fixed population size N. For the SMC, P is the number of particles, while $\mathcal{C}_q(M, 1)$ is the cost of computing the parameters of the categorical distribution q which is used as a proposal (see Rimella et al. (2023) for an example). The memory requirement assumes that everything is stored over time steps.

	Memory	Running time
Data simulation	$\mathcal{O}(TN)$	$\mathcal{O}(TN\mathcal{C}_{Cat}(M, 1) + TN\mathcal{C}_\eta(N))$
SMC	$\mathcal{O}(TNP)$	$\mathcal{O}(TN\mathcal{C}_{Cat}(M, P) + TNP\mathcal{C}_\eta(N) + TN\mathcal{C}_q(M, P))$
CAL	$\mathcal{O}(TNM)$	$\mathcal{O}(TNM^2 + TN\mathcal{C}_\eta(N))$

The cost $\mathcal{C}_\eta(N)$ is model-specific. For instance, for the motivating example in sections 1.1 and 2.3, we have a cost of $\mathcal{C}_\eta(N) = \mathcal{O}(N)$ for the homogeneous-mixing case and a cost of $\mathcal{C}_\eta(N) =$

$\mathcal{O}(N^2)$ for the heterogeneous-mixing scenario or other dense spatial interactions (Jewell et al., 2009; Rimella et al., 2023). Remark that the cost $\mathcal{C}_\eta(N)$ is inherent to the model, rather than a feature of the CAL, as even simulating data from the model would incur in a cost of the same order: $\mathcal{O}(TN\mathcal{C}_{Cat}(M,1)+TN\mathcal{C}_\eta(N))$, where $\mathcal{C}_{Cat}(M,1)$ is the cost of simulating a single draw from a categorical distribution with M categories. Algorithm 1 requires computing probability vectors $\boldsymbol{\pi}_{n,t|t-1}$, $\boldsymbol{\mu}_{n,t}$, and $\boldsymbol{\pi}_{n,t}$ for each individual n at every time step. If all these vectors are stored over T time steps, the memory requirement is $\mathcal{O}(TNM)$. The memory requirement of storing the data is $\mathcal{O}(TN)$. The resulting computational cost is significantly cheaper than SMC, which would require many simulations of each individual. Table 4 provides an SMC-CAL complexity comparison summary. Remark that if the user is only interested in the likelihood computation, it is possible to reduce the memory consumption of the SMC to $\mathcal{O}(NP)$, and similarly to $\mathcal{O}(NM)$ for the CAL.

D.2 Hamiltonian Monte Carlo for homogeneous-mixing SIS

We use the parameter settings $p_0 = 0.01, \beta = 0.2, \mathbf{b}_I = 0.5, \mathbf{b}_S = 1.0, \gamma = 0.1, \mathbf{b}_R = -0.5, q_S = 0.2, q_I = 0.5, q_{Se} = 0.9, q_{Sp} = 0.95$. We treat p_0, q_{Se}, q_{Sp} as known. To create a warm start for our HMC sampler we consider 100 random initializations, where each parameter is drawn from a standard Gaussian, and perform 1000 steps of Adam optimization with a learning rate of 0.1. Across the 100 initializations, we choose the resulting parameter values with the highest CAL log-likelihood as a warm-start for our HMC sampler which we then run for 200000 iterations. We use uninformative independent Gaussian priors each with mean 0 and standard deviation 10.

The learning rate is adapted every 1000 iteration to fall within the acceptance range $[0.55, 0.75]$. Specifically, if the acceptance range is lower than 0.55 then for the next 1000 iterations it is decreased by 35%, while if it is higher than 0.75 it is increased by 35%, otherwise if it falls within the range is kept stable. We run the 201000 HMC iterations and then we choose only the last iterations in which the acceptance rate was stable. We get acceptance rates within the range $0.60 - 0.70$, which are close to the optimal acceptance rate of 0.65 (Neal, 2012). We further apply a thinning procedure that select 10000 samples equally spaced in time (number of iterations).

D.3 Gradient-based calibration for heterogeneously-mixing SIS models

Model 1 Here we provide all the details on Model 1. As initial infection probabilities, we consider:

$$p_0(\mathbf{w}_n) = \begin{bmatrix} 1-p_0 \\ p_0 \end{bmatrix},$$

where $p_0 \in [0,1]$. We consider an interaction term:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \begin{bmatrix} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} & 0 \\ 0 & \frac{1}{\sqrt{2\pi\phi^2}\exp\left\{-\frac{\|\mathbf{z}_n-\mathbf{z}_k\|^2}{2\phi^2}\right\}} \end{bmatrix}^\top \mathbf{x}_{k,t-1},$$

where $\phi > 0$ and $\mathbf{b}_I \in \mathbb{R}$ as we are considering a single covariate. The $\boldsymbol{\eta}_{n,t-1}$ is then used in the transition matrix:

$$K_{\boldsymbol{\eta}_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}_{n,t-1} - h\epsilon\right) & 1 - \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}_{n,t-1} - h\epsilon\right) \\ 1 - \exp\left(-h\gamma_n\right) & \exp\left(-h\gamma_n\right) \end{bmatrix},$$

with $\log \gamma_n = \log \gamma + \mathbf{c}_n^\top \mathbf{b}_R$ and where $\beta, \gamma > 0$ and $\mathbf{b}_S, \mathbf{b}_R \in \mathbb{R}$.

The observation model is given by the matrix:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1 - q_S & q_S q_{Sp} & q_S(1 - q_{Sp}) \\ 1 - q_I & q_I(1 - q_{Se}) & q_I q_{Se} \end{bmatrix},$$

where $q_S, q_{Se}, q_I, q_{Sp} \in [0, 1]$.

Figure 7 provides a graphical representation of the optimization of Model 1.

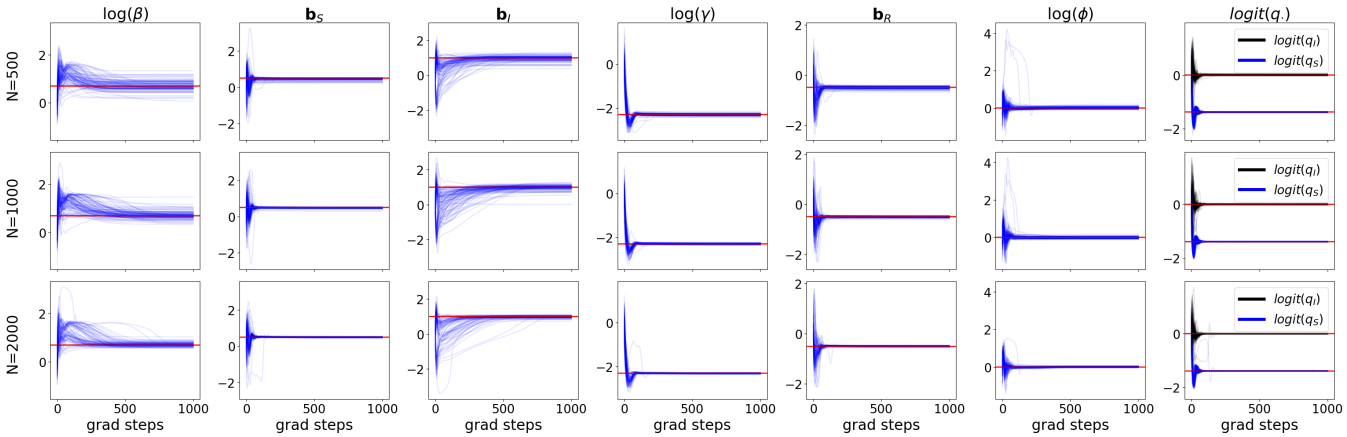

Figure 7: CAL parameters values during optimization for Model 1 from Section 5.2 over the gradient steps of Adam and across different population sizes (from top to bottom). Lines refer to the best out of 10 optimizations on different datasets. Blue and black colors are used in the same plot to distinguish across parameter components.

Model 2 Here we provide all the details on Model 2. Everything is as in Model 1 but the spatial location is now set to the centroid of the community the individual is in. As a consequence, there is also a computationally cheaper representation of $\boldsymbol{\eta}_{n,t-1}$, which exploits the fact that some individuals belong to the same community. Denote with $\mathbf{m}_i^c$ the centroid of community i and with $\mathbf{a}_i$ the set of individuals within the community, i.e. $n \in \mathbf{a}_i$ if n is in the community i. We can then

observe that for $n \in \mathbf{a}_j$:

$$\begin{aligned}
\boldsymbol{\eta}_{n,t-1} &= \frac{1}{N} \sum_{k \in [N]} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{ -\frac{\|\mathbf{m}_n - \mathbf{m}_k\|^2}{2\phi^2} \right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1} \\
&= \frac{1}{N} \sum_i \sum_{k \in \mathbf{a}_i} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{ -\frac{\|\mathbf{m}_n - \mathbf{m}_k\|^2}{2\phi^2} \right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1} \\
&= \frac{1}{N} \sum_i \sum_{k \in \mathbf{a}_i} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{ -\frac{\|\mathbf{m}_j^c - \mathbf{m}_i^c\|^2}{2\phi^2} \right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1},
\end{aligned}$$

from which we observe that for any $n, k \in \mathbf{a}_j$ we have $\boldsymbol{\eta}_{n,t-1} = \boldsymbol{\eta}_{k,t-1}$, hence we just need to compute the interaction term for each community. This reduces the computational cost of computing all $\boldsymbol{\eta}_{n,t-1}$ from N^2 to N times the number of communities.

Figure 8 provides a graphical representation of the optimization of Model 2.

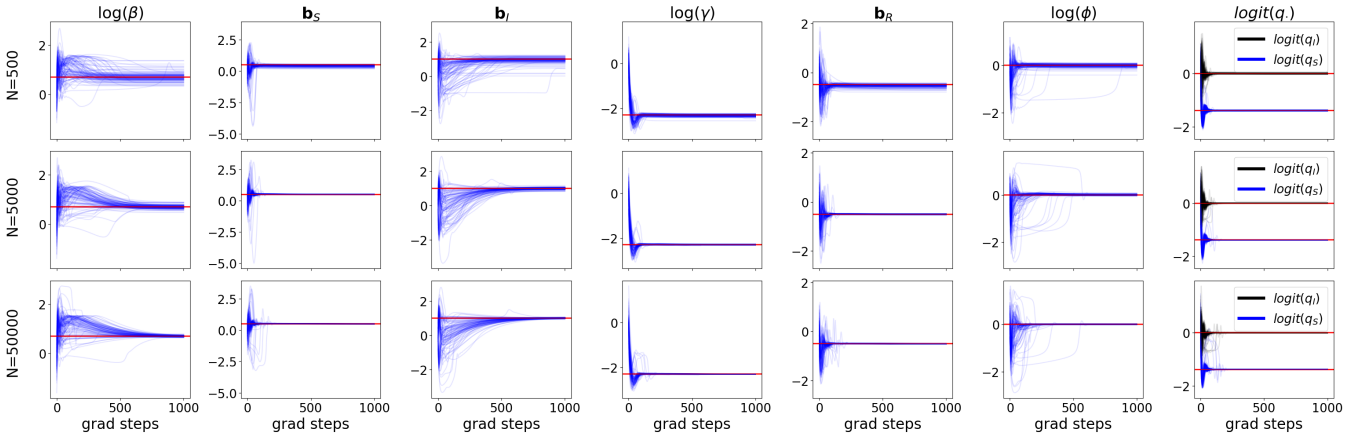

Figure 8: CAL parameters values during optimization for Model 2 from Section 5.2 over the gradient steps of Adam and across different population sizes (from top to bottom). Lines refer to the best out of 10 optimizations on different datasets. Blue and black colors are used in the same plot to distinguish across parameter components.

D.4 Calibration and filtering for heterogeneously-mixing SIR

Well-specified model Here we provide all the details on the correctly specified model from Section 5.3. As initial infection probabilities, we consider:

$$p_0(\mathbf{w}_n) = \begin{bmatrix} 1 - p_0 \mathbb{I}\left(\mathbf{z}_n \in [-\infty, 5] \times [8, +\infty]\right) \\ p_0 \mathbb{I}\left(\mathbf{z}_n \in [-\infty, 5] \times [8, +\infty]\right) \\ 0 \end{bmatrix},$$

where $p_0 \in [0,1]$. We consider an interaction term:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N} \sum_{k \in [N]} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{\exp\left\{-\frac{\|\mathbf{z}_n - \mathbf{z}_k\|^2}{2\phi}\right\}}{\sqrt{2\pi\phi}} \end{bmatrix}^\top \mathbf{x}_{k,t-1},$$

where $\phi > 0$ and $\mathbf{b}_I \in \mathbb{R}$ as we are considering a single covariate. The $\boldsymbol{\eta}_{n,t-1}$ is then used in the transition matrix:

$$K_{\boldsymbol{\eta}_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\} \boldsymbol{\eta}_{n,t-1} - h\epsilon\right) & 1 - \exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\} \boldsymbol{\eta}_{n,t-1} - h\epsilon\right) & 0 \\ 0 & \exp\left(-h\gamma_n\right) & 1 - \exp\left(-h\gamma_n\right) \\ 0 & 0 & 1 \end{bmatrix},$$

with $\log \gamma_n = \log \gamma + \mathbf{c}_n^\top \mathbf{b}_R$ and where $\beta, \gamma > 0$ and $\mathbf{b}_S, \mathbf{b}_R \in \mathbb{R}$.

Half of the population is forced to be unobserved and misreporting is not allowed. We call U the set of individuals that are always unobserved, the observation model is then given by:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1 - q_S \mathbb{I}(n \notin U) & q_S \mathbb{I}(n \notin U) & 0 & 0 \\ 1 - q_I \mathbb{I}(n \notin U) & 0 & q_I \mathbb{I}(n \notin U) & 0 \\ 1 - q_R \mathbb{I}(n \notin U) & 0 & 0 & q_R \mathbb{I}(n \notin U) \end{bmatrix},$$

where $q_S, q_I, q_R \in [0,1]$.

Misspecified model The initial infection probabilities, the transition matrix, and the observation model are the same as for the well-specified model. However, the interaction term is:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N} \sum_{k \in [N]} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{(\|\mathbf{m}_n - \mathbf{m}_k\| \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| \neq 0) + \bar{\mathbf{z}}_n \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| = 0))^2}{2\phi^2}\right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1}.$$

as explained in the main paper. We can notice that the big advantage of this formulation, as for Model 2, is the computational cost indeed:

$$\begin{aligned} \boldsymbol{\eta}_{n,t-1} &= \frac{1}{N} \sum_{k \in [N]} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{(\|\mathbf{m}_n - \mathbf{m}_k\| \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| \neq 0) + \bar{\mathbf{z}}_n \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| = 0))^2}{2\phi^2}\right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1} \\ &= \frac{1}{N} \sum_i \sum_{k \in \mathbf{a}_i} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{(\|\mathbf{m}_n - \mathbf{m}_k\| \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| \neq 0) + \bar{\mathbf{z}}_n \mathbb{I}(\|\mathbf{m}_n - \mathbf{m}_k\| = 0))^2}{2\phi^2}\right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1} \\ &= \frac{1}{N} \sum_i \sum_{k \in \mathbf{a}_i} \begin{bmatrix} 0 \\ \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{(\|\mathbf{m}_j^c - \mathbf{m}_i^c\| \mathbb{I}(\|\mathbf{m}_j^c - \mathbf{m}_i^c\| \neq 0) + \bar{\mathbf{z}}_j^c \mathbb{I}(\|\mathbf{m}_j^c - \mathbf{m}_i^c\| = 0))^2}{2\phi^2}\right\} \end{bmatrix}^\top \mathbf{x}_{k,t-1} \end{aligned}$$

where $\mathbf{m}_i^c$ is the centroid of community i, $\bar{\mathbf{z}}_i^c$ is the mean distance within community i, $\mathbf{a}_i$ is the set of individuals within community i, and we assume $n \in \mathbf{a}_j$.

The considered metrics for comparison are:

Algorithm 2 Previous guess

Require: $\mathbf{W}, \mathbf{Y}_{1:T}, g$
  Initialize $\boldsymbol{\pi}^{g}_{n,0}$ and $\boldsymbol{\pi}^{guess}_{n,0}$ with $1_M \oslash M$ for all $n \in [N]$
  for $t \in 1, \dots, T$ **do**
    for $n \in [N]$ **do**
      if $\mathbf{y}^{(M+1)}_{n,t} = 0$ **then**
        $\boldsymbol{\pi}^{g}_{n,t} = \mathbf{y}^{(1:M)}_{n,t}$
        $\boldsymbol{\pi}^{guess}_{n,t} = g \odot \mathbf{y}^{(1:M)}_{n,t} + \frac{1-g}{M-1} \odot \left(1 - \mathbf{y}^{(1:M)}_{n,t}\right)$
      else
        $\boldsymbol{\pi}^{g}_{n,t} = \boldsymbol{\pi}^{guess}_{n,t-1}$
        $\boldsymbol{\pi}^{guess}_{n,t} = \boldsymbol{\pi}^{guess}_{n,t-1}$
      end if
    end for
  end for

- cross-entropy loss: $-\frac{1}{NT}\sum_{t=1}^{T}\sum_{n\in[N]}\mathbf{x}^{\top}_{n,t}\log(\boldsymbol{\pi}_{n,t})$;
- accuracy: $\left[\frac{1}{NT}\sum_{t=1}^{T}\sum_{n\in[N]}\mathbb{I}\left(\text{argmax}_i\,\mathbf{x}_{n,t} = \text{argmax}_i\,\boldsymbol{\pi}_{n,t}\right)\right]\cdot 100\%$.

We now define the baseline classifiers. To build a classifier we need to create a vector of probabilities, which represents the probability of estimating the different states, e.g. for the CAL this is $\pi_{n,t}$. Let us start with the "Random" classifier, here our vector of probabilities for estimating is: $\boldsymbol{\pi}^{g}_{n,t} = \mathbb{I}\left(\mathbf{y}^{(M+1)}_{n,t} = 1\right)(1_3 \oslash 3) + \mathbb{I}\left(\mathbf{y}^{(M+1)}_{n,t} = 0\right)\mathbf{y}_{n,t}$, meaning that we estimate at random if the individual is unreported, otherwise we estimate with what is reported. "Prev. uncertain" and "Prev. certain" are more complicated and we define them via Algorithm 2. Here, if the individual is reported, we estimate the individual's state with what is reported, otherwise, we have a confidence parameter g which tells us how confident we are with predicting the nth individual at t with their latest observed state. If $g = 0.34$ we have "Prev. uncertain", while if $g = 0.99$ we have "Prev. certain".

D.4.1 Changing the parameters values

Suppose now that we want to consider different parameters values, for instance we want $q_R = 0$. This can be easily done by looking at the tutorial_SIR.ipynb file in the GitHub repository LorenzoRimella/CAL and by changing the configuration:

```
parameters = {
"prior_infection":tf.convert_to_tensor([1-0.5, 0.5, 0.0], dtype = tf.float32),
"log_beta":tf.math.log(tf.convert_to_tensor([3.0], dtype = tf.float32)),
"b_S":tf.convert_to_tensor([+0.5], dtype = tf.float32),
"b_I":tf.convert_to_tensor([+1.0], dtype = tf.float32),
"log_gamma":tf.math.log(tf.convert_to_tensor([0.05], dtype = tf.float32)),
```

```
"b_R":tf.convert_to_tensor([-0.1], dtype = tf.float32),
"log_phi":tf.math.log(tf.convert_to_tensor([1.5], dtype = tf.float32)),
"log_epsilon":tf.math.log(tf.convert_to_tensor([0.0001], dtype = tf.float32)),
"logit_prob_testing":logit(tf.convert_to_tensor([0.1, 0.2, 0.5],
dtype = tf.float32))
}
```

For the case $q_R = 0$ the above would change to:

```
parameters = {
"prior_infection":tf.convert_to_tensor([1-0.5, 0.5, 0.0], dtype = tf.float32),
"log_beta":tf.math.log(tf.convert_to_tensor([3.0], dtype = tf.float32)),
"b_S":tf.convert_to_tensor([+0.5], dtype = tf.float32),
"b_I":tf.convert_to_tensor([+1.0], dtype = tf.float32),
"log_gamma":tf.math.log(tf.convert_to_tensor([0.05], dtype = tf.float32)),
"b_R":tf.convert_to_tensor([-0.1], dtype = tf.float32),
"log_phi":tf.math.log(tf.convert_to_tensor([1.5], dtype = tf.float32)),
"log_epsilon":tf.math.log(tf.convert_to_tensor([0.0001], dtype = tf.float32)),
"logit_prob_testing":logit(tf.convert_to_tensor([0.1, 0.2, 0.0],
dtype = tf.float32))
}
```

Rerunning the simulation leads to similar conclusion to Section 5.3, with parameter estimates close to the truth and similar cross-entropy loss and accuracy, see Table 5.

Table 5: Cross-entropy loss (the lower the better) and accuracy (the higher the better) for the CAL well-specified and misspecified, along with some baselines. The predicted state for accuracy is the argmax of the probability vector.

Metric	Random	Prev. uncertain	Prev. certain	CAL
Cross-entropy	1.1	1.1	1.86	0.34
Accuracy	34.85%	41.65%	41.65%	86.09%

D.5 Comparing CAL with some baselines

In this section, we compare the run time and marginal likelihood values obtained by CAL against those from: SMC for individual-based models with approximate optimal proposals by Rimella et al. (2023), where α controls the number of future observations included in the lookahead scheme, and Simulation Based Composite Likelihood (SimBa-CL) (Rimella et al., 2025). The SMC proposed by Rimella et al. (2023) provides a suitable proxy for comparison as it uses proposal distribution that are informed by future observations avoiding particles degeneracy. It has been recently discovered that Rimella et al. (2023) likelihood estimator is biased due to the exclusion of the initial weights when performing calculations. However we expect this bias to disappear as P increases.

We consider a homogeneous-mixing SIS inspired by Ju et al. (2021). Here the initial distribution is given by a logistic regression on $\mathbf{c}_n$ with parameters $\mathbf{b}_0 \in \mathbb{R}^2$:

$$p_0(\mathbf{w}_n) = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\mathbf{c}_n^\top \mathbf{b}_0)} \\ \frac{1}{1+\exp(-\mathbf{c}_n^\top \mathbf{b}_0)} \end{bmatrix}.$$

We consider an interaction term $\boldsymbol{\eta}_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \mathbf{x}^{(2)}_{k,t-1}$ or equivalently:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \begin{bmatrix} 0 \\ 1 \end{bmatrix}^\top \mathbf{x}_{k,t-1},$$

as we are considering a homogeneous-mixing case, which is then used in the transition matrix:

$$K_{\eta_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} \exp\left(-h\frac{\boldsymbol{\eta}_{n,t-1}+\epsilon}{1+\exp\left(-\mathbf{b}_S^\top \mathbf{c}_n\right)}\right) & 1 - \exp\left(-h\frac{\boldsymbol{\eta}_{n,t-1}+\epsilon}{1+\exp\left(-\mathbf{b}_S^\top \mathbf{c}_n\right)}\right) \\ 1 - \exp\left(-\frac{h}{1+\exp\left(-\mathbf{b}_R^\top \mathbf{c}_n\right)}\right) & \exp\left(-\frac{h}{1+\exp\left(-\mathbf{b}_R^\top \mathbf{c}_n\right)}\right) \end{bmatrix},$$

where $\mathbf{b}_S, \mathbf{b}_I \in \mathbb{R}^2$ and $\epsilon > 0$ and . The observation model is then given by:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1 - q_S & q_S & 0 \\ 1 - q_I & 0 & q_I \end{bmatrix},$$

with $q_S, q_I \in [0, 1]$. We consider a population of 1000 individuals and a time horizon of 100, with parameters set to $\mathbf{b}_0 = [-\log(100-1), 0]^\top, \epsilon = 0.001, \mathbf{b}_S = [-1, 2]^\top, \mathbf{b}_R = [-1, -1]^\top, q_S = 0.6, q_I = 0.4, q_{Se} = 1.0, q_{Sp} = 1$.

Table 6: Log-likelihood means and standard deviations for the SIS model. We denote Rimella et al. (2023) with †, with α being the number of future observations included in the lookahead scheme. Log-likelihood results are averages and standard deviations over 100 runs. Running times are reported for a single run and as averages across particles.

Number of particles P	512	1024	2048	Time (sec)
† with $\alpha = 5$	-79551.92 (1.79)	-79552.24 (1.6)	-79552.81 (1.57)	3.78s
† with $\alpha = 10$	-79551.9 (1.81)	-79552.22 (1.47)	-79553.01 (1.56)	5.61s
SimBa-CL	-79612.74 (3.4)	-79612.31 (2.37)	-79612.34 (1.55)	1.03s
CAL			-79550.69	1.93s
CAL jit_compiled			-79550.69	0.003s

We consider $N = 1000, T = 100$, simulate from the model and, for that one realization of the data, estimate mean and standard deviation of the log-likelihood at the DGP for each method over 100 runs, the CAL requires a single run as it is a deterministic algorithm. The results are reported in Table 6. As expected, the method by Rimella et al. (2023) with the highest α has the lowest

variance, but it is computationally intensive as a single run requires more than 5s. SimBa-CL performs well computationally but it is more biased. The CAL is the fastest and the log-likelihood estimate is close to the one from Rimella et al. (2023), with a running time even dropping to 0.003s when considering just-in-time compilation, which is, as explained in Section 5, straightforward for the CAL.

The SEIR scenario is significantly more challenging as some transitions are not allowed, making the SMC more prone to particle impoverishment and degeneracy. In our SEIR we consider a homogeneous-mixing individual-based model with an initial distribution as in the aforementioned SIS, but an additional zero probability of being assigned to E, R at the beginning of the epidemic. The model is again inspired by Ju et al. (2021), now the initial distribution is:

$$p_0(\mathbf{w}_n) = \begin{bmatrix} 1 - \frac{1}{1+\exp(-\mathbf{b}_0^\top \mathbf{c}_n)} \\ 0 \\ \frac{1}{1+\exp(-\mathbf{b}_0^\top \mathbf{c}_n)} \\ 0 \end{bmatrix},$$

where $\mathbf{b}_0 \in \mathbb{R}^2$. As we are again considering a homogeneous-mixing scenario the interaction term is:

$$\boldsymbol{\eta}_{n,t-1} = \frac{1}{N} \sum_{k \in [N]} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}^\top \mathbf{x}_{k,t-1},$$

and used in the stochastic transition matrix:

$$K_{\eta_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} \exp\left(\frac{-h\boldsymbol{\eta}_{n,t-1}-h\epsilon}{1+\exp(-\mathbf{b}_S^\top \mathbf{c}_n)}\right) & 1 - \exp\left(\frac{-h\boldsymbol{\eta}_{n,t-1}-h\epsilon}{1+\exp(-\mathbf{b}_S^\top \mathbf{c}_n)}\right) & 0 & 0 \\ 0 & \exp(-h\rho) & 1-\exp(-h\rho) & 0 \\ 0 & 0 & \exp\left(\frac{-h}{1+\exp(-\mathbf{b}_R^\top \mathbf{c}_n)}\right) & 1 - \exp\left(\frac{-h}{1+\exp(-\mathbf{b}_R^\top \mathbf{c}_n)}\right) \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

with $\mathbf{b}_S, \mathbf{b}_I \in \mathbb{R}^2, \epsilon > 0$ and $\rho \in \mathbb{R}_+$. The observation matrix is:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1-q_S & q_S & 0 & 0 & 0 \\ 1-q_E & 0 & q_E & 0 & 0 \\ 1-q_I & 0 & 0 & q_I & 0 \\ 1-q_R & 0 & 0 & 0 & q_R \end{bmatrix},$$

with $q_S, q_E, q_I, q_R \in [0, 1]$.

As for the SIS scenario, we consider a population of 1000 individuals and a time horizon of 100. We set $\mathbf{b}_0 = [-\log(100-1), 0]^\top, \epsilon = 0.001, \mathbf{b}_S = [-1, 2]^\top, \rho = 0.2, \mathbf{b}_R = [-1, -1]^\top, q_S = 0, q_E = 0, q_I = 0.4, q_R = 0.6$ and we simulate from the model. The log-likelihood mean and standard deviation are then estimated over multiple runs. As expected, the method by Rimella et al. (2023)

Table 7: Log-likelihood means and log-likelihood standard deviations for the individual-based SEIR model with $N = 1000$. We denote Rimella et al. (2023) with †, with α being the number of future observations included in the lookahead scheme. Log-likelihood results are averages and standard deviation over 100 runs. Running times are reported for a single run.

Number of particles P	512	1024	2048	Time (sec)
† with $\alpha = 5$	-43447.56 (52.04)	-43419.52 (51.08)	-43391.0 (52.41)	4.44s
† with $\alpha = 20$	-43004.55 (5.38)	-43001.9 (4.65)	-42999.76 (3.7)	11.08s
† with $\alpha = 50$	-42999.93 (3.44)	-42998.13 (2.72)	-42996.74 (2.39)	20.88s
SimBa	-43683.85 (9.54)	-43683.67 (7.35)	-43683.76 (5.16)	1.25s
CAL			-43454.97	1.89s
CAL jit_compiled			-43454.97	0.003s

requires several future observations to achieve low variance, which implies a significant increase in the running time. SimBa-CL is computationally more efficient, but it shows higher variance and a significant bias. The CAL is closer to Rimella et al. (2023) with $\alpha = 50$ compared to the other baselines. Again, if just-in-time compilation is considered, the CAL runs in about 0.003s.

D.6 Comparing CAL with SMC

For the experiment of Section 5.4 in the main manuscript, we use the parameter settings $p_0 = 0.01, \beta = 0.2, \mathbf{b}_I = 0.3, \mathbf{b}_S = -0.3, \gamma = 0.1, \mathbf{b}_R = 0.2, q_S = 0.2, q_I = 0.5, q_{Se} = 0.9, q_{Sp} = 0.95$.

We consider a batched Block APF to reduce memory consumption. Our implementation loads everything into memory to exploit GPU parallel computing, resulting in a memory blow-up when resampling is performed on all the individuals. Switching from an algorithm that is fully parallelized over both individuals and particles to a sequential version where parallelization is applied only within batches offers a trade-off between memory usage and running time.

In the batched Block APF, the particles are always batched into four subgroups: instead of resampling P particles in parallel, we sequentially sample $\frac{P}{4}$ particles and then concatenate the resulting samples. Individuals are batched according to the population size N. Specifically, there is no batching for $N = 10$; a batching size of 10 for $N = 100$; a batching size of 100 for $N = 1000$; and again a batching size of 100 for $N = 10000$.

From Table 3 it appears that when $N, P \to \infty$ the (batched) Block APF has a similar asymptotic behavior to the CAL as $N \to \infty$. We therefore provide a graphical illustration of this behavior in Figure 9. It can be observed that as N increases the SMC methods require more and more particles to control the variance. Moreover, for $N = 1000$ and $N = 10000$, we notice that when P increases the (batched) Block APF gets closer and closer to CAL. We expect this to be a consequence of the large-population limit of the model and the resulting decoupling of the individuals. Indeed, the CAL becomes exact as N goes to infinity, and we expect the Block APF to have similar properties as it considers a similar approximation. This perspective differs from that taken by Rebeschini and Van Handel (2015), where the quality of the block particle filter was measured on coarser and coarser partitions rather than increasing dimension, and may be of independent interest.

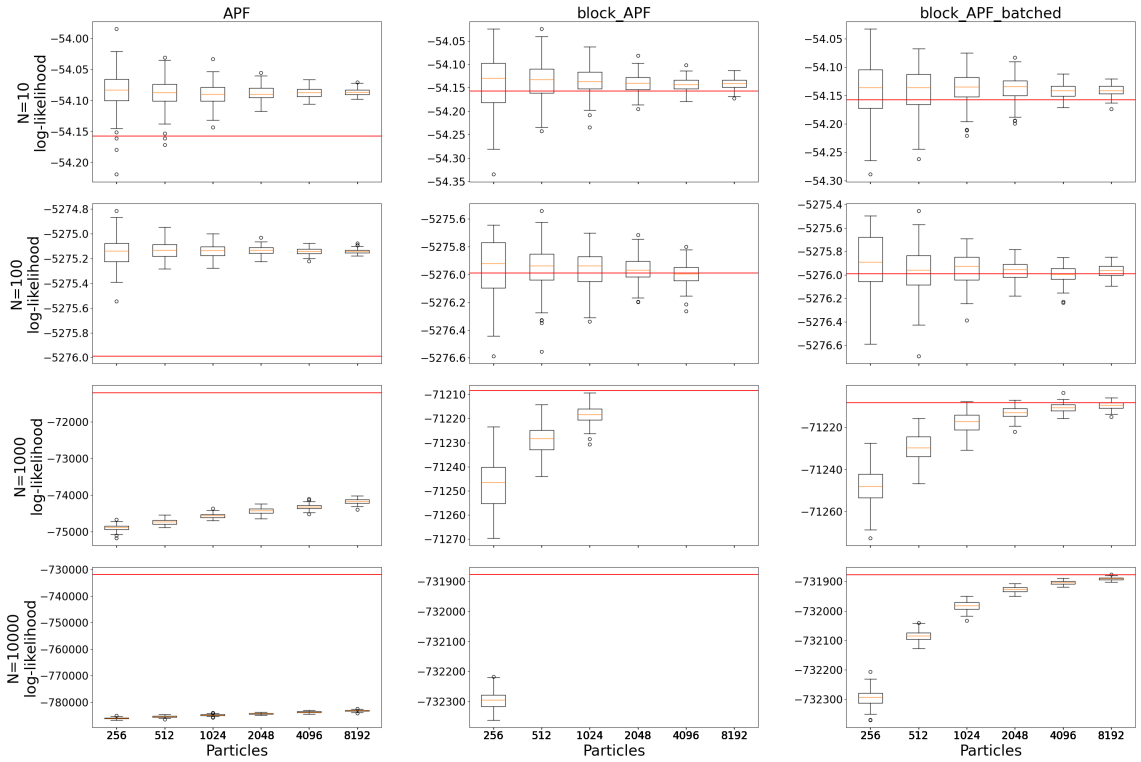

Figure 9: Log-likelihood boxplots for APF, Block APF, and batched Block APF when N and P increases for the SIS model from Section 1.1. The CAL log-likelihood is reported as a horizontal red line.

Table 8: Log-likelihood means and standard deviations, over 100 runs, for the SIS model from Ju et al. (2021). Running times are reported in seconds for a single run and as averages across the 100 runs.

N	P	CAL	time(s)	CAL compiled	time(s)	APF	time(s)	Block APF	time(s)	batched Block APF	time(s)
10	256	-67.13	1.354	-67.13	0.001	-67.01(0.03)	0.11	-67.1(0.05)	0.13	-67.11(0.06)	2.43
10	512					-67.01(0.02)	0.12	-67.1(0.03)	0.13	-67.1(0.04)	2.42
10	1024					-67.01(0.01)	0.11	-67.1(0.02)	0.13	-67.1(0.03)	2.41
10	2048					-67.01(0.01)	0.12	-67.1(0.02)	0.13	-67.1(0.02)	2.41
10	4096					-67.01(0.01)	0.12	-67.1(0.01)	0.16	-67.1(0.01)	2.41
10	8192					-67.01(0.0)	0.12	-67.1(0.01)	0.31	-67.1(0.01)	2.31
100	256	-7101.4	2.128	-7101.4	0.001	-7107.8(4.94)	1.12	-7102.37(1.82)	1.26	-7102.82(1.72)	23.53
100	512					-7105.34(3.21)	1.08	-7102.38(1.3)	1.31	-7102.25(1.24)	23.0
100	1024					-7102.46(2.92)	1.11	-7101.92(0.9)	1.46	-7102.02(0.94)	23.02
100	2048					-7100.6(2.37)	1.11	-7101.78(0.61)	2.35	-7101.7(0.67)	23.2
100	4096					-7099.73(1.98)	1.14	-7101.6(0.5)	5.93	-7101.61(0.53)	23.26
100	8192					-7099.1(1.47)	1.26	Out of memory		-7101.57(0.33)	37.83
1000	256	-82562.3	2.207	-82562.3	0.001	-84153.91(38.25)	1.17	-82584.62(7.16)	1.47	-82582.9(7.13)	25.75
1000	512					-84043.09(35.49)	1.22	-82573.03(4.79)	2.09	-82573.19(4.48)	25.43
1000	1024					-83963.92(37.44)	1.31	-82567.92(3.23)	4.51	-82567.3(2.98)	25.75
1000	2048					-83869.86(33.91)	1.47	Out of memory		-82565.28(2.24)	30.94
1000	4096					-83792.78(28.68)	1.86	Out of memory		-82563.78(1.88)	66.6
1000	8192					-83721.4(27.97)	2.71	Out of memory		-82562.98(1.16)	208.23
10000	256	-822379.25	2.435	-822379.25	0.001	-847562.6(144.4)	1.44	-822586.25(21.59)	3.84	-822587.7(19.98)	217.59
10000	512					-847122.5(152.59)	1.94	Out of memory		-822482.56(13.75)	218.71
10000	1024					-846751.25(141.99)	3.01	Out of memory		-822429.75(10.21)	220.13
10000	2048					-846343.2(130.99)	5.08	Out of memory		-822405.5(7.72)	283.91
10000	4096					-845975.6(137.44)	9.7	Out of memory		-822393.2(4.76)	676.48
10000	8192					-845647.8(132.39)	19.12	Out of memory		-822385.6(3.37)	2105.17

We also replicated the same experiment under the Ju et al. (2021) model from the previous subsection with parameters set to $\mathbf{b}_0 = [-\log(100-1), 0]^\top, \epsilon = 0.001, \mathbf{b}_S = [-1, 2]^\top, \mathbf{b}_R =$

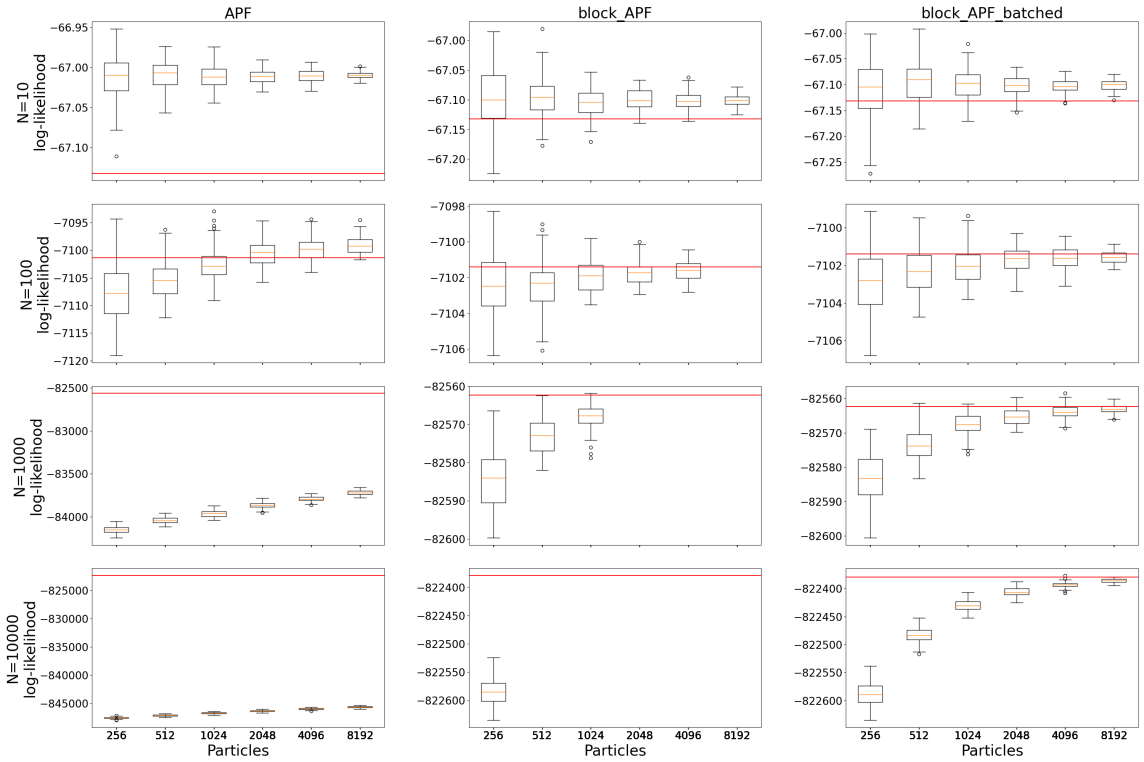

Figure 10: Log-likelihood boxplots for APF, Block APF, and batched Block APF when N and P increases for the SIS model by Ju et al. (2021). The CAL log-likelihood is reported as a horizontal red line.

$[-1,-1]^\top, q_S = 0.6, q_I = 0.4, q_{Se} = 1.0, q_{Sp} = 1$. We report numerical results in Table 8 and a graphical representation in Figure 10.

D.7 2001 UK foot-and-mouth disease outbreak

Local authorities meta-population model. For the local authorities we consider data.gov.uk (2023), reporting digital vector boundaries of the UK's local authority districts in December 2023. We did not find any open-source digital vector boundaries from 2001, which would have been ideal. Local authorities with less than five farms were excluded from the study, e.g. London and Birmingham. The farm-specific covariates are then $\mathbf{w}_n = [\mathbf{m}_n, \bar{\mathbf{z}}_n, \mathbf{c}_n]$, where $\mathbf{m}_n$ is the centroid (in EPSG:27700, the projected coordinate system for the UK) of the local authority individual n is assigned to, $\bar{\mathbf{z}}_n$ is the mean-distance (in km) across farms within the local authority, and $\mathbf{c}_n$ is a bi-dimensional vector containing the log-number of cattle and the log-number of sheep. The components of $\mathbf{m}_n$ are further divided by 1000 so when computing Euclidean distances across local authorities the resulting distances are in Km. Observe that $\mathbf{c}_n$ is still at an individual-level, while we have aggregated the spatial component.

Model. We consider a heterogeneous-mixing individual-based SIR model, where transitions from S to R are also allowed, representing the culling/quarantine of healthy farms to create containment zones around infected farms. We consider two interaction terms:

$$\boldsymbol{\eta}^I_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \begin{bmatrix} 0 \\ \frac{\exp\{\mathbf{c}_k^\top \mathbf{b}_I\}}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{\|\mathbf{m}_n-\mathbf{m}_k\|^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|\neq 0)+\bar{\mathbf{z}}_n^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|=0)}{2\phi^2}\right\} \\ 0 \end{bmatrix}^\top \mathbf{x}_{k,t-1}, \tag{59}$$

$$\boldsymbol{\eta}^C_{n,t-1} = \frac{1}{N}\sum_{k\in[N]} \begin{bmatrix} 0 \\ \frac{1}{\sqrt{2\pi\psi^2}} \exp\left\{-\frac{\|\mathbf{m}_n-\mathbf{m}_k\|^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|\neq 0)+\bar{\mathbf{z}}_n^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|=0)}{2\psi^2}\right\} \\ 0 \end{bmatrix}^\top \mathbf{x}_{k,t-1}, \tag{60}$$

where $\mathbf{b}_I \in \mathbb{R}^2, \phi > 0, \psi > 0$, and we also define:

$$\boldsymbol{\eta}^0_n \coloneqq \frac{1}{N}\sum_{k\in[N]} \exp\{\mathbf{c}_k^\top \mathbf{b}_I\} \frac{\exp\left\{-\frac{\|\mathbf{m}_n-\mathbf{m}_k\|^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|\neq 0)+\bar{\mathbf{z}}_n^2\mathbb{I}(\|\mathbf{m}_n-\mathbf{m}_k\|=0)}{2\phi^2}\right\}}{\sqrt{2\pi\phi^2}}\tau, \tag{61}$$

where $\tau > 0$ can be interpreted as the probability of being infected before $t = 0$. We consider an initial distribution:

$$p_0(\mathbf{w}_n) = \begin{bmatrix} \exp\left(-\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}^0_n - \epsilon\right) \\ 1 - \exp\left(-\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}^0_n - \epsilon\right) \\ 0 \end{bmatrix}.$$

We then define the culling/quarantine probability of farm n by $P^C_n \coloneqq 1-\exp\left(-h\rho\boldsymbol{\eta}^C_{n,t-1}\right)$ where $\rho > 0$, the infection probability of a non-culled/quarantine farm by $P^I_n \coloneqq 1-\exp\left(-h\beta \exp\{\mathbf{c}_n^\top \mathbf{b}_S\}\boldsymbol{\eta}^I_{n,t-1} - h\epsilon\right)$ where $\beta > 0$ and $\mathbf{b}_S \in \mathbb{R}^2$, and the recovery probability $P^R_n \coloneqq 1 - \exp(-h\gamma)$ where $\gamma > 0$. The stochastic transition matrix is then given by:

$$K_{\eta^I_{n,t-1},\eta^R_{n,t-1}}(\mathbf{w}_n) = \begin{bmatrix} (1-P^C_n)(1-P^I_n) & (1-P^C_n)P^I_n & P^C_n \\ 0 & (1-P^C_n)(1-P^R_n) & P^C_n + (1-P^C_n)P^R_n \\ 0 & 0 & 1 \end{bmatrix}.$$

For the observation model, we do not allow for misreporting and we assume only infected are observable:

$$G(\mathbf{w}_n) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1-q_I & 0 & q_I & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

The aforementioned individual-based model has two interaction terms and an interaction term inside the initial distribution and does not strictly belong to the class of models that satisfy our assumptions. Nevertheless, the terms (59),(60),(61) follow a law of large numbers, as they are averages, hence our saturation theory can be developed for such individual-based models with multidimensional interaction terms. Consistency then follows using the same techniques of Section 4.

Param.	$\log(\tau)$	$\log(\beta)$	$\mathbf{b}_S^{(1)}$	$\mathbf{b}_S^{(2)}$	$\mathbf{b}_I^{(1)}$	$\mathbf{b}_I^{(2)}$
Mean	-12.52	4.83	0.37	0.11	0.23	0.09
95% CI	[-13.00,-11.97]	[4.49,5.14]	[0.35,0.40]	[0.10,0.13]	[0.18,0.30]	[0.05,0.12]

Param.	$\log(\phi)$	$\log(\gamma)$	$\log(\rho)$	$\log(\psi)$	$\log(\epsilon)$	$logit(q_I)$
Mean	3.36	3.0	12.52	1.25	-12.83	0.67
95% CI	[3.32,3.40]	[2.47,3.49]	[12.08,12.99]	[0.92,1.44]	[-13.14,-12.54]	[0.51,0.85]

Table 9: HMC posterior means and 95% credible intervals on the foot-and-mouth model.

Maximum CAL estimation. We consider 100 different random initializations of $\tau, \beta, \mathbf{b}_S$, $\mathbf{b}_I, \phi, \gamma, \rho, \psi, \epsilon, q_I$ and run Adam optimizer for each of them for 10000 gradients steps using auto-differentiation in TensorFlow. We then select the one with the highest log-CAL. We then perform a sequential optimization with Adam where every 50000 iteration we change the value of the step size and reset the Adam optimizer. Specifically, we consider the step sizes $0.1, 0.01, 0.01, 0.001, 0.001$, $0.001, 0.00001, 0.00001$, resulting in a total of 400000 gradient steps.

Credible intervals. We then run an HMC using the maximum CAL estimation as a warm start. We considered a step size of 0.00274625, which was obtained via pre-tuning on the acceptance rate, 10 leapfrog steps, a vague prior for $\beta, \mathbf{b}_S, \mathbf{b}_I, \phi, \epsilon, q_I$ given by a Gaussian with mean 0 and standard deviation 100, and an informative prior on τ, γ, ρ, ψ given by Gaussians with means $12.5, 3, 12.5, 1.2$ and standard deviation 0.25. The informative prior choice was needed to improve mixing of the HMC. The mean of the prior was set to the maximum CAL estimation we used as a warm start, while the standard deviation was tuned to ensure proper mixing. Posterior means and 95% credible interval are reported in Table 9

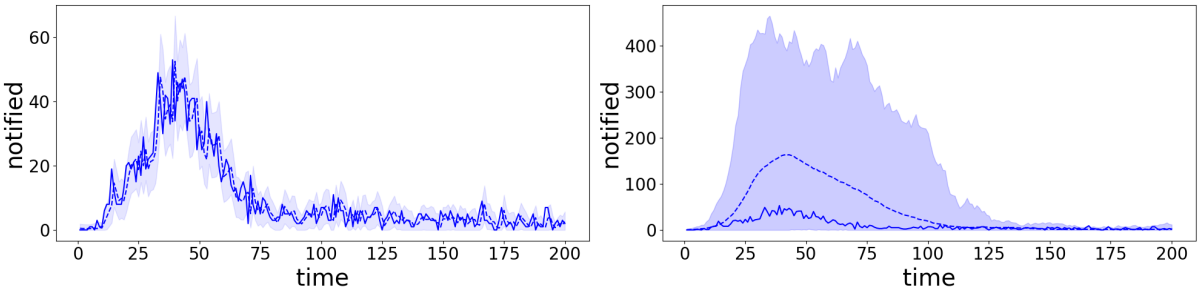

Figure 11: On the left, the CAL filter mean (dashed blue line) with 95% bands based on the CAL filter variance. On the right, the Monte Carlo mean (dashed blue line) and 2.5%, 97.5% quantile of the predictive distribution. Solid lines are used for the observations.

In Figure 11 we report the CAL filter and the predictive distribution under the optimized parameters for the total number of notified. The CAL filter is simply obtained by running the CAL on the optimized parameters and by using the Categorical approximation to estimate mean and variance of $\sum_{n\in[N]} \mathbf{y}_{n,t}$. The predictive distribution is obtained via multiple simulations from the

model with the DGP set to the optimized parameters, which are then used to get Monte Carlo estimates of mean and quantiles. In both cases we can observe that we get good coverage of the observations, showing that the optimized parameters are valid estimates.

The spatial location of a single farm cannot be disclosed for privacy, hence the left-hand side of Figure 4 and the whole Figure 6 cannot be reproduced. However, we are allowed to disclose information about local authorities and, in particular, the spread of the disease within the local authorities, which can be found in our GitHub repository. The equivalent of Figure 6 for the local authorities is Figure 12. Even though this aggregated version is less informative compared to the fully spatial one, we can still recognize the same patterns in terms of the disease's spread of the infection and distribution of the removed.

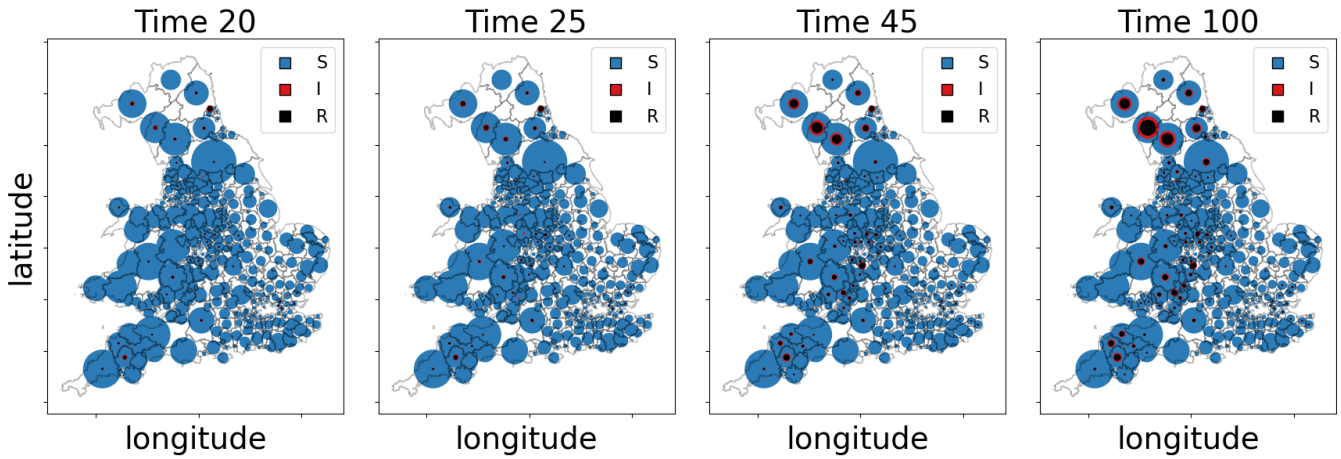

Figure 12: The CAL prediction over time of susceptible farms (blue), infected farms (red), and removed farms (black) for the local authority model. Black dots are double the radius and red dots are four times the radius for visual purposes.

D.7.1 Benchmarking

Consider $\widetilde{\mathbf{y}}_{n,t}$ which takes values $1, 0$ depending on whether the farm is reported as infected or not. We model $p(\widetilde{\mathbf{y}}_{n,1} = 1), p(\widetilde{\mathbf{y}}_{n,t}|\widetilde{\mathbf{y}}_{n,t-1})$ with an AR logistic regression:

$$p(\widetilde{\mathbf{y}}_{n,1} = 1) = \frac{\gamma}{1 + e^{-\mathbf{b}^\top \mathbf{w}_n}}, \quad p(\widetilde{\mathbf{y}}_{n,t} = 1|\widetilde{\mathbf{y}}_{n,t-1}) = \frac{\gamma}{1 + e^{-\beta \widetilde{\mathbf{y}}_{n,t-1} - \mathbf{b}^\top \mathbf{w}_n}},$$

where $\gamma \in [0, 1]$, $\beta \in \mathbb{R}$ and $\mathbf{b} \in \mathbb{R}^{C+1}$, with C number of covariates for each individual ($+1$ for the intercept). This results in the likelihood:

$$p(\widetilde{\mathbf{y}}_{n,1:T} = 1) = \prod_{n \in [N]} \left(\frac{\gamma}{1 + e^{-\mathbf{b}^\top \mathbf{w}_n}} \right)^{\mathbf{y}_{n,1}} \left(1 - \frac{\gamma}{1 + e^{-\mathbf{b}^\top \mathbf{w}_n}} \right)^{1 - \mathbf{y}_{n,1}}$$
$$\cdot \prod_{t=2}^{T} \left(\frac{\gamma}{1 + e^{-\beta \widetilde{\mathbf{y}}_{n,t-1} - \mathbf{b}^\top \mathbf{w}_n}} \right)^{\mathbf{y}_{n,t}} \left(1 - \frac{\gamma}{1 + e^{-\beta \widetilde{\mathbf{y}}_{n,t-1} - \mathbf{b}^\top \mathbf{w}_n}} \right)^{1 - \mathbf{y}_{n,t}}.$$

Algorithm 3 CAL within bootstrap particle filter for "shared" overdispersion

Require: $\mathbf{W}, \mathbf{Y}_{1:T}, p_0(\cdot), K_\cdot(\cdot), G(\cdot), P$
  Initialize $\boldsymbol{\pi}^P_{n,0}$ with $p_0(\mathbf{w}_n)$ for all $n \in [N]$ and $p \in [P]$
  for $t \in 1, \dots, T$ **do**
    $\boldsymbol{\Pi}^p_{t-1} = (\boldsymbol{\pi}^p_{1,t-1}, \dots, \boldsymbol{\pi}^p_{N,t-1})$ for all $p \in [P]$
    Sample ξ^p_t from the prior for all $p \in [P]$
    for $n \in [N]$ **do**
      $\widetilde{\boldsymbol{\eta}}^p_{n,t-1} = \eta(\mathbf{w}_n, \mathbf{W}, \boldsymbol{\Pi}^p_{t-1})$
      $\boldsymbol{\pi}^p_{n,t|t-1} = \left[(\boldsymbol{\pi}^p_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}_{n,t-1}}(\mathbf{w}_n, \xi^p_t)\right]^\top$
      $\boldsymbol{\mu}^p_{n,t} = \left[(\boldsymbol{\pi}^p_{n,t|t-1})^\top G(\mathbf{w}_n)\right]^\top$
      $\boldsymbol{\pi}^p_{n,t} = \boldsymbol{\pi}^p_{n,t|t-1} \odot \left\{\left[G(\mathbf{w}_n) \oslash \left(1_M(\boldsymbol{\mu}^p_{n,t})^\top\right)\right] \mathbf{y}_{n,t}\right\}$
    end for
    Set $w^p_t = \prod_{n\in[N]} \mathbf{y}^\top_{n,t} \boldsymbol{\mu}^p_{n,t}$ for all $p \in [P]$
    Resample the particles $\xi^p_t, (\boldsymbol{\pi}^p_{n,t})_{n\in[N]}$ according to $\bar{w}^p_t \propto w^p_t$
  end for
  Return the approximate likelihood $\prod_{t=1}^T \frac{1}{P} \sum_{p\in[P]} w^p_t$

We optimize $\gamma, \beta, \mathbf{b}$ using gradient ascent on the log-likelihood and Adam optimizer with 10000 gradient steps and a learning rate of 0.1, resulting in a log-likelihood of -20858.994.

We then consider the same AR logistic regression but with local authority specific parameters. Specifically, we have

$$p(\widetilde{\mathbf{y}}_{n,1} = 1) = \frac{\gamma}{1 + e^{-\mathbf{b}_k^\top \mathbf{w}_n}}, \quad p(\widetilde{\mathbf{y}}_{n,t} = 1 | \widetilde{\mathbf{y}}_{n,t-1}) = \frac{\gamma_k}{1 + e^{-\beta_k \widetilde{\mathbf{y}}_{n,t-1} - \mathbf{b}_k^\top \mathbf{w}_n}},$$

where $\gamma_k \in [0,1]$, $\beta_k \in \mathbb{R}$ and $\mathbf{b}_k \in \mathbb{R}^{C+1}$ and with k being the local authority of individual n.

We similarly optimize $(\gamma_k, \beta_k, \mathbf{b}_k)_k$ using gradient ascent on the log-likelihood and Adam optimizer with 10000 gradient steps and a learning rate of 0.1, resulting in a log-likelihood of -18866.531.

D.7.2 Dealing with overdispersion

Suppose that we want to include overdispersion in our foot-and-mouth model. This can be done by considering some stochastic parameters when the interaction term enters P^I_n.

Model overdispersed. We consider the model as above but we set:

$$P^I_n \coloneqq 1 - \exp\left(-h\beta\xi_{n,t} \exp\{\mathbf{c}_n^\top \mathbf{b}_S\} \boldsymbol{\eta}^I_{n,t-1} - h\epsilon\right)$$

where $\log \xi_{n,t}$ is Gaussian with mean μ_o and standard deviation σ_o. Here $\xi_{n,t}$ can be common across all the individuals, hence ξ_t, or across local authorities, hence $\xi_{B,t}$, resulting in what we call "shared" overdispersion and "local authority" overdispersion. $K_\cdot(\cdot)$ now depends on the stochastic

Algorithm 4 CAL within block bootstrap particle filter for "local authority" overdispersion

Require: $\mathbf{W}, \mathbf{Y}_{1:T}, p_0(\cdot), K_\cdot(\cdot), G(\cdot), P$
  Initialize $\boldsymbol{\pi}^P_{n,0}$ with $p_0(\mathbf{w}_n)$ for all $n \in [N]$ and $p \in [P]$
  for $t \in 1, \ldots, T$ **do**
    $\boldsymbol{\Pi}^p_{t-1} = (\boldsymbol{\pi}^p_{1,t-1}, \ldots, \boldsymbol{\pi}^p_{N,t-1})$ for all $p \in [P]$
    for $B \in$ local authorities **do**
      Sample $\xi^p_{B,t}$ from the prior for all $B \in$ local authorities and $p \in [P]$
      for $n \in B$ **do**
        $\widetilde{\boldsymbol{\eta}}^p_{n,t-1} = \eta(\mathbf{w}_n, \mathbf{W}, \boldsymbol{\Pi}^p_{t-1})$
        $\boldsymbol{\pi}^p_{n,t|t-1} = \left[(\boldsymbol{\pi}^p_{n,t-1})^\top K_{\widetilde{\boldsymbol{\eta}}_{n,t-1}}(\mathbf{w}_n, \xi^p_{B,t})\right]^\top$
        $\boldsymbol{\mu}^p_{n,t} = \left[(\boldsymbol{\pi}^p_{n,t|t-1})^\top G(\mathbf{w}_n)\right]^\top$
        $\boldsymbol{\pi}^p_{n,t} = \boldsymbol{\pi}^p_{n,t|t-1} \odot \left\{\left[G(\mathbf{w}_n) \oslash \left(1_M(\boldsymbol{\mu}^p_{n,t})^\top\right)\right] \mathbf{y}_{n,t}\right\}$
      end for
      Set $w^p_{B,t} = \prod_{n \in B} \mathbf{y}^\top_{n,t} \boldsymbol{\mu}^p_{n,t}$
      Resample the particles $\xi^p_{B,t}, (\boldsymbol{\pi}^p_{n,t})_{n \in B}$ according to $\bar{w}^p_{B,t} \propto w^p_{B,t}$
    end for
  end for
  Return the approximate likelihood $\prod_{t=1}^T \prod_B \frac{1}{P} \sum_{p \in [P]} w^p_{B,t}$

parameters, meaning that we denote it with $K_{\eta^I_{n,t-1}, \eta^R_{n,t-1}}(\mathbf{w}_n, \xi_t)$ for "shared" overdispersion and with $K_{\eta^I_{n,t-1}, \eta^R_{n,t-1}}(\mathbf{w}_n, \xi_{B,t})$ for "local authority" overdispersion. To make the notation lighter we refer to $K_{\eta^I_{n,t-1}, \eta^R_{n,t-1}}(\mathbf{w}_n, \cdot)$ with just $K_{\eta_{n,t-1}}(\mathbf{w}_n, \cdot)$.

Likelihood computation. To provide likelihood estimate we nest the CAL within an SMC (Whitehouse et al., 2023). Precisely, we consider a bootstrap particle filter for "shared" overdispersion and a block bootstrap particle filter for "local authority" overdispersion. Here a particle approximation of $\xi_{n,t}|\mathbf{Y}_{1:t}$ is provided recursively via the SMC, where $\xi_{n,t}$ might be shared across different individuals. Algorithm 3 provides the pseudo-code for the bootstrap particle filter for the "shared" overdispersion, while Algorithm 4 provides the pseudo-code for the block bootstrap particle filter for the "local authority" overdispersion.

Maximum CAL estimation. Ideally we would like to optimize the parameters based on the likelihood approximations from both Algorithm 3 and Algorithm 4. However, the current version of the algorithms is not suitable to automatic differentiation and it be JIT compiled only within a time step and not across time steps because of the resampling procedure. For computational reasons we hence optimize the parameters μ_o, σ_o on a grid while keeping the other parameters fixed to the maximum CAL estimator obtained without overdispersion. We decided to optimize also μ_o to ensure that shifts in the transmission rate are also possible if required. We consider $\mu_o \in \{-2, -1, -0.5, -0.25, 0, 0.25, 0.5, 1, 2\}$ and $\sigma_o \in \{0.05, 0.125, 0.25, 0.5, 1, 2, 5\}$, and we get $\mu_o =$

$0, \sigma_o = 0.25$ as maximum on the grid, see Figure 13 for a graphical illustration. We then consider $\mu_o = 0, \sigma_o = 0.25$ and rerun Algorithm 3 100 times to get an estimate of the Monte Carlo variability. As we reported in the main manuscript we get a log-likelihood estimate of -17926.988 ± 0.4619721. Considering Algorithm 4, we set again $\mu_o = 0, \sigma_o = 0.25$ and run the algorithm 100 times. Here we obtain -17907.555 ± 0.835302 as a log-likelihood estimate.

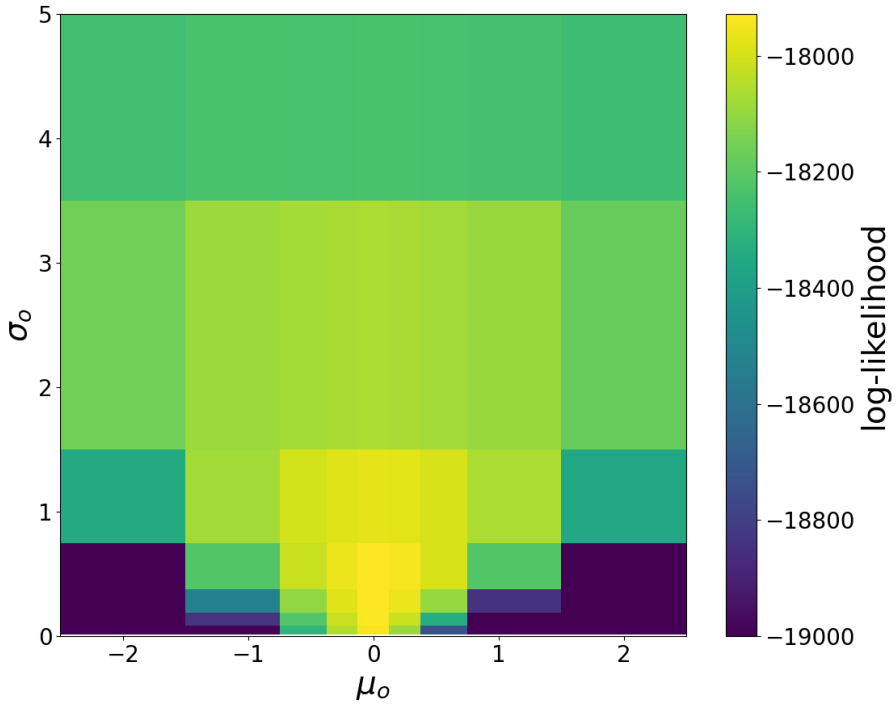

Figure 13: Log-likelihood estimates from Algorithm 3 on a grid of μ_o, σ_o.