Online Appendix. A Conditional Particle Filter for Epidemic Processes

As we have shown in Sections 3, 4 and 5, epidemic processes are a suitable application domain for exploiting the factorization assumption for state distributions. In this appendix, we study the effect of not exploiting this factorization assumption. Without this assumption, the problem becomes one of tracking a state distribution on a product space with a particle filter, where the particle family giving the state distribution is a subset of the product space. As expected, we show such an approach only works for low-dimensional product spaces.

Next the effectiveness of the algorithms in Section 2 for filtering on graphs is empirically demonstrated. Consider an epidemic process spreading on a contact network, as described in Section 3. We employ conditional filtering to track states and estimate the parameters of epidemics on small contact networks. For this application, both the filter for parameters and the conditional filter for states are *particle* filters, i.e., the filter for parameters employs Algorithm 2** and the conditional filter for states employs Algorithm 5.

We simulate an infectious disease on two small graphs, which are social networks employed as contact networks among individuals. The first graph is a social network of 34 club members (Zachary, 1977), and the other graph is a social network of 62 dolphins (Lusseau et al., 2003). Let $C = \{\mathsf{S}, \mathsf{I}\}$. We assume the disease spreads over the contact network following the transitions of an SIS epidemic model (Eq. (37)) below with transmission probability $\beta = 0.2$ and average infectious period $\gamma^{-1} = 10$ days, and no immunity gained upon recovery.

For all $n \in \mathbb{N}$ and $k = 1, \ldots, L$, define

$$\breve{\tau}_n^{(k)} : C^L \to \mathcal{D}(C)$$

by

$$\breve{\tau}_n^{(k)}(s)(c) = \begin{cases} 1 - (1-\beta)^{d_k(s)} & \text{if } s_k = \mathsf{S} \text{ and } c = \mathsf{I} \\ (1-\beta)^{d_k(s)} & \text{if } s_k = \mathsf{S} \text{ and } c = \mathsf{S} \\ \gamma & \text{if } s_k = \mathsf{I} \text{ and } c = \mathsf{S} \\ 1 - \gamma & \text{if } s_k = \mathsf{I} \text{ and } c = \mathsf{I}, \end{cases} \tag{37}$$

where $s \in C^L$, $c \in C$, and $d_k(s)$ is the number of infected neighbours of node k.

The observation model is given by Eq. (38). For all $n \in \mathbb{N}$ and $k = 1, \ldots, L$,

$$\breve{\xi}_n^{(k)} : C \to \mathcal{D}(O)$$

is defined by

$$\begin{aligned} \breve{\xi}_n^{(k)}(\mathsf{S})(+) &= \alpha_{\mathsf{S}}\lambda_{FP}, & \breve{\xi}_n^{(k)}(\mathsf{I})(+) &= \alpha_{\mathsf{I}}(1-\lambda_{FN}), \\ \breve{\xi}_n^{(k)}(\mathsf{S})(-) &= \alpha_{\mathsf{S}}(1-\lambda_{FP}), & \breve{\xi}_n^{(k)}(\mathsf{I})(-) &= \alpha_{\mathsf{I}}\lambda_{FN}, \\ \breve{\xi}_n^{(k)}(\mathsf{S})(?) &= 1-\alpha_{\mathsf{S}}, & \breve{\xi}_n^{(k)}(\mathsf{I})(?) &= 1-\alpha_{\mathsf{I}}. \end{aligned} \tag{38}$$

Here, α_{S} and α_{I} are the probability that a susceptible and infected individual, respectively, is tested. Also, λ_{FP} is the false positive rate of the testing method and λ_{FN} is the false negative rate. We assume $\alpha_S = 0.1$ and $\alpha_I = 0.9$, which means that an infected individual is highly likely to be tested due to, for example, symptoms, and approximately 1 in 10 healthy (susceptible) individuals in the population will be tested as a result of large-scale random testing. And we further assume the testing method employed has a 10% false positive rate and a 10% false negative rate, i.e., $\lambda_{FP} = \lambda_{FN} = 0.1$.

We start a simulation by choosing an individual, that is, a node, uniformly at random from the set of all nodes in the graph and make it infected (i.e., the patient zero); all other nodes in the contact network are susceptible; we then evolve the transition model (Eq. (37)) for 600 time steps and generate observations according to the observation model (Eq. (38)) in each time step.

Experimental setup A parameter particle filter with 300 particles is employed, so that $N = 300$. The parameter particle family is initialized by sampling from uniform distributions as follows:

$$x^{(i)}_{n_*,1} \sim \mathcal{U}(0, 0.8), \quad x^{(i)}_{n_*,2} \sim \mathcal{U}(0, 0.8),$$

where $i \in \{1, \dots, N\}$ and $n_* \in \mathbb{N}$ is defined below.

The initial state distribution of an epidemic is usually unknown and we have to estimate it, for example, using information of the reported positive cases. Let n_* be the time step when the number of reported positive cases from testing surpasses, for the first time, a predefined threshold[1], and suppose that most individuals in the population are susceptible[2] at time n_*. The initial state distribution is then estimated by performing an observation update using the observation at time n_*. In particular, let $(p_k, 1 - p_k)$ denote the tuple of (estimated) probabilities for the compartments, in the order S, I, for node k at time n_*. The state particle families are initialized by sampling from categorical distributions as follows:

$$y^{(i,j)}_{n_*,k} \sim Cat(p_k, 1 - p_k)$$

for $k = 1, \dots, L$, $i = 1, \dots, N$ and $j = 1, \dots, M$. Here $Cat(p_k, 1 - p_k)$ is a categorical distribution for which the parameters are p_k and $1 - p_k$.

We determine the number of state particles M using a formula suggested by Snyder et al. (2008),

$$M = 10^{0.05d+0.78}, \tag{39}$$

where d is the state dimension. This formula was developed empirically from a simple application for a lower bound on the number of particles that is needed for filtering to perform reasonably well. Thus, for $d = 34, M \approx 300$ and, for $d = 62, M \approx 7600$. While the formula given by Eq. (39) is at most an approximate lower bound for a rather simpler problem than an SIS epidemic, it seemed worthwhile to try these values for the state particle family sizes for the graphs of 34 and 62 nodes in the experiments.

We also leverage an adaptive Gaussian jittering kernel in Algorithm 2** with covariance matrix

$$\Sigma_n = \max(ar^n, b)\mathbf{I}_2, \ n \in \mathbb{N}, \tag{40}$$

where $a = 10^{-4}, b = 9 \times 10^{-6}, r = 0.996$, and $\mathbf{I}_2$ is the identity matrix of size 2.

[1] The threshold is 3 in all experiments in this appendix.

[2] We assume the probabilities that an individual being susceptible and infected are 0.9 and 0.1, respectively.

Evaluation The performance of the filter in tracking states is now investigated. As explained in Section 3.5, it is sufficient to define a suitable metric ρ on C^L. Let $\kappa : C \times C \to \mathbb{R}$ be the discrete metric on C and $\rho : C^L \times C^L \to \mathbb{R}$ the metric on C^L defined by

$$\rho(y, z) = \frac{1}{L} \sum_{k=1}^{L} \kappa(y_k, z_k),$$

for all $y, z \in C^L$. Let $\widetilde{y}_n \in C^L$ be the ground truth state at time n. Suppose that

$$\widehat{(\nu_n \odot \mu_n)(h_n)} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta_{y_n^{(i,j)}}$$

is the filter's estimate of the state distribution at time n. Then

$$\begin{aligned}
&\mathit{Err}(\widehat{(\nu_n \odot \mu_n)(h_n)}) \\
&= \int_Y \lambda y.\rho(\widetilde{y}_n, y)\ d\widehat{(\nu_n \odot \mu_n)(h_n)} \\
&= \int_{C^L} \lambda y.\frac{1}{L} \sum_{k=1}^{L} \kappa(\widetilde{y}_{n,k}, y_k)\ d\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \delta_{y_n^{(i,j)}} \\
&= \frac{1}{NML} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L} \kappa(\widetilde{y}_{n,k}, y_{n,k}^{(i,j)}).
\end{aligned}$$

Thus

$$\mathit{Err}(\widehat{(\nu_n \odot \mu_n)(h_n)}) = \frac{1}{NML} \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{L} \kappa(\widetilde{y}_{n,k}, y_{n,k}^{(i,j)}).$$

For the parameters, the error is given by Eq. (16), so that

$$\mathit{Err}_j(\widehat{\nu_n(h_n)}) = \frac{1}{\widetilde{x}_j} \frac{1}{N} \sum_{i=1}^{N} |\widetilde{x}_j - x_{n,j}^{(i)}|,$$

for $j = 1, 2$. Here $\widetilde{x}_1 = \beta$ and $\widetilde{x}_2 = \gamma$.

Results and discussion Figure 19 shows the results of tracking states, and Figures 20 and 21 the results of estimating parameters using the conditional particle filter (Algorithms 5 and 2**) on the two small contact networks with 34 and 62 nodes.

We have seen the state particle family sizes suggested by Eq. (39) proved to be good lower bounds – filtering worked well enough with those sizes for both graphs. Further, we note that the accuracy of state tracking and parameter estimation are remarkably similar between the two cases, the reason might be the difference in dimension is counterbalanced by the number of state particles. We remark that with smaller state particle families, the experimental results were not as good, and unsurprisingly, with bigger state particle families, the experimental results were better, for each graph.

On the other hand, according to Eq. (39), for $d = 80$, M is approximately $60,000$, but for $d = 100$, M will exceed $600,000$, which is not all that feasible. This suggests that for contact

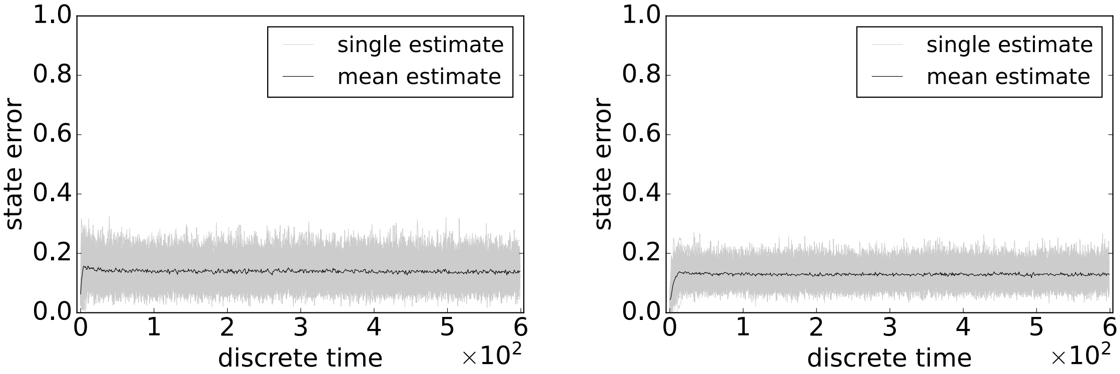

Figure 19: State errors of tracking SIS epidemics on small contact networks using the conditional particle filter (i.e., Algorithms 2** and 5) in 100 independent runs in which the disease does not die out in 600 time steps. The true values of transmission probability $\beta = 0.2$ and average recovery period $\gamma^{-1} = 10$ days. Left: using a contact network with 34 nodes (Zachary, 1977), and the number of particles employed are $N = 300, M = 300$. Right: using a contact network with 62 nodes (Lusseau et al., 2003), and the number of particles employed are $N = 300, M = 7,600$. Light gray: state errors from individual runs. Dark solid: mean state error averaged over 100 independent runs.

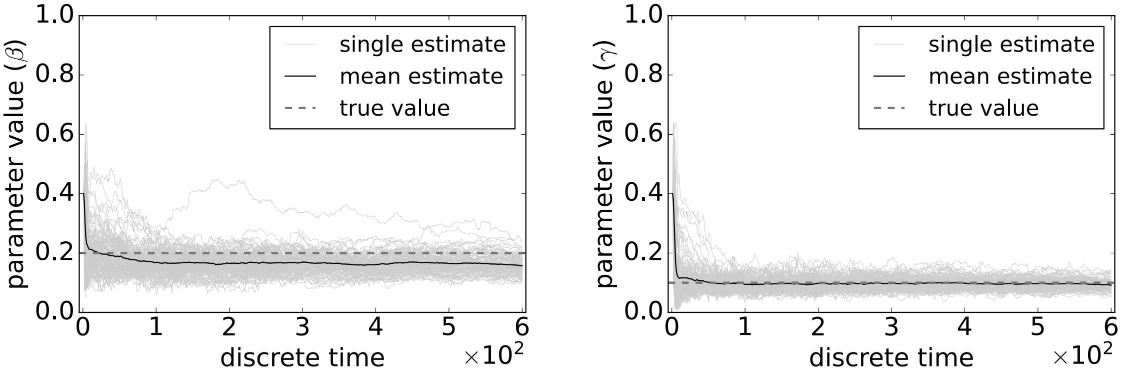

Figure 20: Estimating parameters of an SIS epidemic on a small contact network with 34 nodes (Zachary, 1977) using the conditional particle filter (i.e., Algorithms 2** and 5) in 100 independent runs in which the disease does not die out in 600 time steps. The number of particles employed are $N = 300, M = 300$, and the true values of transmission probability $\beta = 0.2$ and average recovery period $\gamma^{-1} = 10$ days. Light gray: estimated value from individual runs. Dark solid: mean estimation averaged over 100 independent runs. Dark dash: true values of parameters.

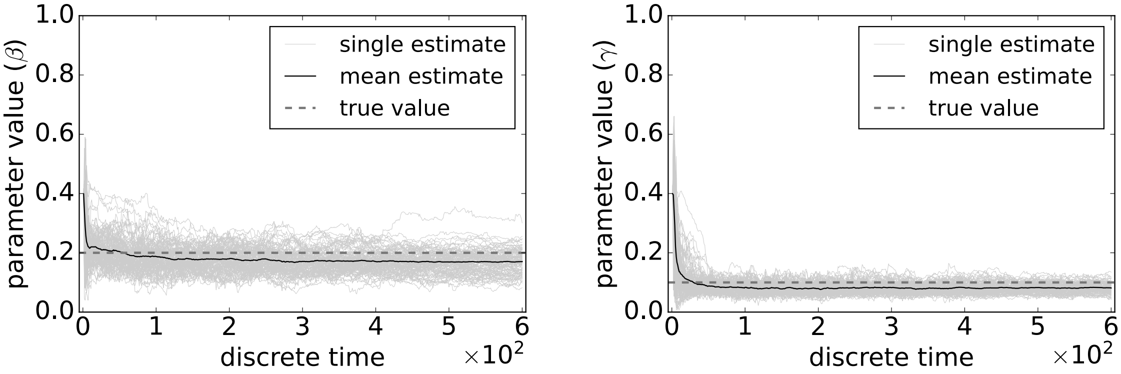

Figure 21: Estimating parameters of an SIS epidemic on a small contact network with 62 nodes (Lusseau et al., 2003) using the conditional particle filter (Algorithms 2^{**} and 5) in 100 independent runs in which the disease does not die out in 600 time steps. The number of particles employed are $N = 300, M = 7,600$, and the true values of transmission probability $\beta = 0.2$ and average recovery period $\gamma^{-1} = 10$ days. Light gray: estimated value from individual runs. Dark solid: mean estimation averaged over 100 independent runs. Dark dash: true values of parameters.

networks much larger than those considered in this appendix, it is necessary to exploit the structural properties of epidemics, such as factorization. Our experiments with larger contact networks, not using factorization, produced poor results and thus confirmed this view.