

Unsupervised Learning of Sampling Distributions for Particle Filters

Fernando Gama , *Member, IEEE*, Nicolas Zilberstein , *Graduate Student Member, IEEE*, Martin Sevilla ,
Richard G. Baraniuk , *Fellow, IEEE*, and Santiago Segarra , *Senior Member, IEEE*

Abstract—Accurate estimation of the states of a nonlinear dynamical system is crucial for their design, synthesis, and analysis. Particle filters are estimators constructed by simulating trajectories from a sampling distribution and averaging them based on their importance weight. For particle filters to be computationally tractable, it must be feasible to simulate the trajectories by drawing from the sampling distribution. Simultaneously, these trajectories need to reflect the reality of the nonlinear dynamical system so that the resulting estimators are accurate. Thus, the crux of particle filters lies in designing sampling distributions that are both easy to sample from and lead to accurate estimators. In this work, we propose to *learn* the sampling distributions. We put forward four methods for learning sampling distributions from observed measurements. Three of the methods are parametric methods in which we learn the mean and covariance matrix of a multivariate Gaussian distribution; each method exploits a different aspect of the data (generic, time structure, graph structure). The fourth method is a nonparametric alternative in which we directly learn a transform of a uniform random variable. All four methods are trained in an unsupervised manner by maximizing the likelihood that the states may have produced the observed measurements. Our computational experiments demonstrate that learned sampling distributions exhibit better performance than designed, minimum-degeneracy sampling distributions.

Index Terms—Machine learning, unsupervised learning, particle filtering, neural networks, graph neural networks.

I. INTRODUCTION

NONLINEAR dynamical systems serve as models for a wide range of problems in science and engineering. For example, due to the natural hysteresis of the material, nonlinear dynamical systems are used to describe electrical circuits

involving ferromagnetic inductors [2]. Mostly, they have been very popular tools in control theory [3]. They play a key role in designing and synthesizing controllers for spacecraft systems [4], in managing energy consumption of electrical vehicles [5], in reducing ripple in wind power systems [6], and even in real-time bidding for programmatic advertising [7].

Due to their practical relevance, research on nonlinear dynamical systems has a long history [8]. Topics such as the existence and uniqueness of solutions, dependence on initial conditions, stability of the systems, as well as perturbation analysis have dominated the field [9]. Recent results concern dissipativity and its connection to stability [10], identification [11], and oscillations [12].

In this work, we focus on estimating the states of the system [13], [14]. Almost any decision to be made in the synthesis, design, or analysis of nonlinear dynamical systems needs to rely on a solid knowledge of the system state. An inaccurate modeling of the system or an impossibility to access the state – and instead being able to measure some function of the state – are the major hindrances in the task of estimation.

Many estimators have been developed, making different assumptions on the system to reach different levels of accuracy guarantees. For instance, assuming linear dynamics with Gaussian noise leads to the linear least-squares estimator [15]. Another example is assuming the model is nonlinear but follows the Markov property on the conditionality of its transition probabilities [16]. In this scenario, oftentimes the maximum *a posteriori* estimate can be obtained. These approaches, however, either oversimplify the model or can become computationally intractable due to the high-dimensional integrals involved.

Particle filtering has risen as an algorithmic tool that is capable of estimating the state in a computationally efficient way [17], [18], [19]. In essence, particle filtering consists of simulating plausible trajectories of the system of interest, and then carrying out a weighted average of these trajectories to obtain accurate estimators [20]. The challenge in particle filtering is in designing a sampling distribution that is simultaneously good at generating realistic trajectories and easy to sample from [21], [22].

Particle filters leverage the law of large numbers to provide certain guarantees on the accuracy of the estimators and convergence to the true value of the state if enough particles are simulated. However, these results oftentimes rely on using specific sampling distributions that may only be

Manuscript received 2 February 2023; revised 1 August 2023 and 19 September 2023; accepted 2 October 2023. Date of publication 13 October 2023; date of current version 26 October 2023. This work was supported in part by the NSF under Award CCF-2008555. An earlier version of this paper was presented in part at the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [DOI: 10.1109/ICASSP43922.2022.9747290]. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ya-Feng Liu. (*Corresponding author: Nicolas Zilberstein.*)

Fernando Gama was with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA. He is now with Pendulum Systems, Inc., Denver, CO 80128 USA (e-mail: fgama@ieee.org).

Nicolas Zilberstein, Martin Sevilla, Richard G. Baraniuk, and Santiago Segarra are with the Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA (e-mail: nzilberstein@rice.edu; msevilla@rice.edu; richb@rice.edu; segarra@rice.edu).

Digital Object Identifier 10.1109/TSP.2023.3324221

computationally efficient in limited scenarios [23]. Furthermore, particle filters suffer from weight degeneracy, a phenomenon that causes only a few of the simulated trajectories to be meaningful, severely impacting the accuracy of the resulting estimator [24]. Thus, most of the research on particle filtering has revolved around the appropriate design of sampling distributions [25], [26]. In this work, however, we leverage modern deep neural network techniques to learn the sampling distribution from observed measurements – without access to the true trajectories.

Deep neural networks consist of a cascade of blocks, each of which applies a linear transformation followed by an activation function that is, typically, nonlinear [27]. The blocks are known as layers, and the number of these layers in cascade determines the depth of the neural networks. The exact matrix values to be used in the linear transforms at each layer are typically determined by a gradient-descent algorithm (or a variant thereof) in an attempt to minimize some loss function over the observed data. The process of determining the actual value of the linear transforms is known as training [28].

In this work, we propose four different deep neural network frameworks that are used to learn the sampling distribution of particle filters, three of which are parametric – learning the mean and covariance matrix of a Gaussian distribution – and one which is non-parametric – learning an arbitrary transform of a uniform random variable. These frameworks lead, naturally, to distributions that are easy to sample from (either multivariate normal or uniform distributions). We train these deep neural networks in an unsupervised manner. This means that we only need access to a trajectory of measurements, but not to the true value of the states. By learning the deep neural network parameters that maximize the likelihood of the observed trajectories under the given model, we are able to obtain a sampling distribution that is capable of simulating good trajectories. In short, learning the sampling distribution is, generally, easier than designing it, while it also allows for more flexibility and adaptability.

Many works have proposed using learning to improve particle filtering. Typically, learning is used mostly to estimate the model or transform the variables into spaces more amenable for sampling [29], [30], [31]. Recently, conditional normalizing flows were proposed as parameterization of the proposal distribution [32]. Recurrent neural networks (RNNs) have also been used to learn the mean of a multivariate normal sampling distribution as well as the particle weights [33]. Most importantly, in all these cases, the neural networks are trained using supervised learning. This requires access to true trajectories of the system, which are typically unavailable, and it does not guarantee that the trajectories observed at test time will be similar enough to ensure generalization. We address this fundamental drawback by proposing a trainable particle filter based on unsupervised learning.

Article structure and overview of results. In Section II, we review the basics of particle filtering and introduce the notation. In Section III, we present the main contribution of the article, which is the unsupervised learning framework of sampling distributions. We explain first our framework in a

general way, and then we describe each of the four models that we propose to parameterize the sampling distribution.

- First, we assume the sampling distribution is a multivariate normal and we learn a time-dependent mean and covariance matrix using a fully-connected neural network (Section III-A).
- Second, we consider a recurrent neural network (RNN) architecture for learning the mean and covariance matrix (Section III-B). RNNs are good at keeping track of past values of the trajectory, potentially allowing the distribution to learn from samples that are located further in the past.
- Third, we consider architectures that exploit the data structure. In particular, we consider graph neural networks (GNNs) which are architectures tailored to process graph-based data (Section III-C). GNNs can be particularly useful when dealing with large distributed nonlinear dynamical systems, where the components have sparse connections between them.
- Fourth, we consider a non-parametric approach (Section III-D). More specifically, we sample from a uniform distribution, and we use a deep neural network to learn an arbitrary nonlinear transform between the uniform distribution and a random variable that represents the trajectory of states.

We close Section III by explaining the details of unsupervised learning through maximization of the likelihood of the model (Section III-E), and how the model is trained as well as a brief explanation of the reparameterization trick. In Section IV, we run a series of simulated examples to showcase the performance of learned sampling distributions as opposed to designed baselines. We consider linear Gaussian (Section IV-A), nonlinear Gaussian (Section IV-B), linear non-Gaussian (Section IV-C), and nonlinear non-Gaussian (Section IV-D) dynamical systems. In general, we observe that learned sampling distributions exhibit better performance than designed, minimum-degeneracy sampling distributions [17]. Finally, we draw conclusions in Section V.

II. PARTICLE FILTERING

Consider a dynamical system described by a sequence of states $\{\mathbf{x}_t\}_{t \geq 0}$ with $\mathbf{x}_t \in \mathbb{R}^N$ for all $t \in \mathbb{N}_0$. The transition between states in different time instants is considered to satisfy the Markov property and, thus, is completely characterized by the transition distribution given by

$$\mathbf{x}_t | \mathbf{x}_{t-1} \sim p(\mathbf{x}_t | \mathbf{x}_{t-1}) \quad (1)$$

for all $t \geq 1$. The initial state is distributed following

$$\mathbf{x}_0 \sim p(\mathbf{x}_0). \quad (2)$$

These states are considered unobservable. Instead, there is a sequence of measurements $\{\mathbf{y}_t\}_{t \geq 0}$, with $\mathbf{y}_t \in \mathbb{R}^M$ for all $t \in \mathbb{N}_0$, that is accessible. Given the current value of the state, the measurements are distributed as follows

$$\mathbf{y}_t | \mathbf{x}_t \sim p(\mathbf{y}_t | \mathbf{x}_t) \quad (3)$$

for all $t \geq 0$. These three distributions (1)–(3) are considered known.

The objective is to estimate a target quantity \mathbf{z}_t that depends on the true states of the system $\mathbf{x}_{0:t} = \{\mathbf{x}_0, \dots, \mathbf{x}_t\}$. We denote this as $\mathbf{z}_t = \mathbf{f}_t(\mathbf{x}_{0:t})$ for some (possibly time-varying) mapping \mathbf{f}_t . To estimate \mathbf{z}_t from a sequence of observations $\mathbf{y}_{0:t} = \{\mathbf{y}_0, \dots, \mathbf{y}_t\}$, we use the conditional expectation to construct an estimator $\tilde{\mathbf{z}}_t$ as

$$\tilde{\mathbf{z}}_t = \mathbb{E}[\mathbf{z}_t | \mathbf{y}_{0:t}] = \int_{-\infty}^{+\infty} \mathbf{f}_t(\mathbf{x}_{0:t}) p(\mathbf{x}_{0:t} | \mathbf{y}_{0:t}) d\mathbf{x}_{0:t}. \quad (4)$$

Using Bayes' rule, the posterior distribution of the state trajectory given the measurements $p(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$ can be written recursively as

$$p(\mathbf{x}_{0:t} | \mathbf{y}_{0:t}) = p(\mathbf{x}_{0:t-1} | \mathbf{y}_{0:t-1}) \frac{p(\mathbf{y}_t | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{x}_{t-1})}{p(\mathbf{y}_t | \mathbf{y}_{0:t-1})}. \quad (5)$$

Note that the numerator of the update rule, i.e. $p(\mathbf{y}_t | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{x}_{t-1})$, can be computed directly from (1)–(3). The denominator $p(\mathbf{y}_t | \mathbf{y}_{0:t-1})$ can also be computed from (1)–(3) by marginalizing over all possible states \mathbf{x}_t and \mathbf{x}_{t-1} at times t and $t-1$ and using the previous step in the recursion $p(\mathbf{x}_{0:t-1} | \mathbf{y}_{0:t-1})$. Then, since all the distributions are known, it should be technically possible to compute the conditional expectation estimator (4). However, this entails high-dimensional integrals with are typically intractable. Therefore, (4) cannot usually be used in practice.

The law of large numbers can be leveraged to suggest a practical estimator consisting of taking K samples $\{\mathbf{x}_{0:t}^{(k)}\}_{k=1}^K$, independently, identically distributed as $\mathbf{x}_{0:t}^{(k)} \sim p(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$, and then averaging them. Note, however, that even if we had access to the posterior $p(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$, it may still be intractable to sample from it.

Particle filtering consists of sampling $\{\mathbf{x}_{0:t}^{(k)}\}_k$ from some other distribution $\mathbf{x}_{0:t}^{(k)} \sim \pi(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$ and computing the estimate as

$$\hat{\mathbf{z}}_t = \sum_{k=1}^K \hat{w}_t^{(k)} \mathbf{f}_t(\mathbf{x}_{0:t}^{(k)}) \quad \hat{w}_t^{(k)} = \frac{\tilde{w}_t^{(k)}}{\sum_{k'=1}^K \tilde{w}_t^{(k')}}, \quad (6)$$

where the normalized weights $\hat{w}_t^{(k)}$ are computed from the set of unnormalized weights $\{\tilde{w}_t^{(k)}\}_k$, each of which is given by $\tilde{w}_t^{(k)} = p(\mathbf{y}_{0:t} | \mathbf{x}_{0:t}^{(k)}) p(\mathbf{x}_{0:t}^{(k)}) / \pi(\mathbf{x}_{0:t}^{(k)} | \mathbf{y}_{0:t})$. For the estimate in (6) to be tractable, sampling from $\pi(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$ has to be computationally feasible. If no further restrictions are imposed on the sampling distribution $\pi(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$, then the particle filtering method receives the name of Bayesian Importance Sampling, the distribution π is known as the importance function, and the weights are known as the importance weights [34]. The samples $\{\mathbf{x}_{0:t}^{(k)}\}_k$ are often referred to as particles or trajectories.

To further facilitate computational tractability in particle filtering, the sampling distribution π is typically restricted to have the form

$$\pi(\mathbf{x}_{0:t} | \mathbf{y}_{0:t}) = \pi(\mathbf{x}_0 | \mathbf{y}_0) \prod_{\tau=1}^t \pi(\mathbf{x}_\tau | \mathbf{x}_{0:\tau-1}, \mathbf{y}_{0:\tau}). \quad (7)$$

This implies that $\pi(\mathbf{x}_{0:t} | \mathbf{y}_{0:t})$ can be computed recursively over time. Then, for each time t , it suffices to sample $\mathbf{x}_t^{(k)} \sim \pi(\mathbf{x}_t | \mathbf{x}_{0:t-1}, \mathbf{y}_{0:t})$. The unnormalized weights can be computed recursively as well, following

$$\tilde{w}_t^{(k)} = \tilde{w}_{t-1}^{(k)} \frac{p(\mathbf{y}_t | \mathbf{x}_t^{(k)}) p(\mathbf{x}_t^{(k)} | \mathbf{x}_{t-1}^{(k)})}{\pi(\mathbf{x}_t^{(k)} | \mathbf{x}_{0:t-1}, \mathbf{y}_{0:t})}. \quad (8)$$

Particle filtering with sampling distributions of the form (7) is often known as Sequential Importance Sampling [20].

While computationally convenient, adopting a sampling distribution as in (7) causes the particle filtering to suffer from weight degeneracy. This means that the unconditional variance of the weights, considering the observations $\mathbf{y}_{0:t}$ as random variables, increases over time [35]. The practical implications are that, over time, only one particle carries all the weight while the rest become insignificant. This affects the quality of the estimator (6) as it virtually relies on a single trajectory. The variance of the weights can be minimized, conditional upon $\mathbf{x}_{0:t-1}^{(k)}$ and $\mathbf{y}_{0:t}$, if the sampling distribution is chosen to be such that [36]

$$\pi(\mathbf{x}_t | \mathbf{x}_{0:t-1}^{(k)}, \mathbf{y}_{0:t}) = p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t). \quad (9)$$

In this case, the unnormalized weight updates become $\tilde{w}_t^{(k)} = \tilde{w}_{t-1}^{(k)} p(\mathbf{y}_t | \mathbf{x}_{t-1}^{(k)})$. Sampling from $p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t)$, however, is generally intractable.

Weight degeneracy can be minimized [cf. (9)] but it cannot be avoided completely. Thus, resampling is typically used to reduce its impact on the estimator (6). In short, resampling consists of randomly sampling trajectories according to their weight, thus giving more importance to the trajectories that carry larger weights. More specifically, the level of weight degeneracy is measured by

$$\hat{K}_t^{\text{eff}} = \frac{1}{\sum_{k=1}^K (\hat{w}_t^{(k)})^2}, \quad (10)$$

which is a proxy for the number of particles that can be considered to be effectively contributing to the estimator. If this number of effective particles drop below a certain threshold \hat{K}^{thres} , then the trajectories are resampled following a distribution that assigns a probability $\hat{w}_t^{(k)}$ of choosing trajectory k . After sampling K times, a new set of K trajectories is obtained (some of them likely to be repeated), and the weights are reset to be $1/K$ for all particles. It is evident that using resampling affects the i.i.d. assumption on the particles, and thus many theoretical results, such as convergence, no longer hold [37].

The particle filter is a computationally simple estimator for nonlinear systems that has shown significant success. For it to yield good results, however, the sampling distribution π has to be carefully designed in such a way that it is both easy to sample from and leads to good estimators (6). Many design methods have been proposed [21]. In what follows, instead of designing it, we propose and discuss several architectures to learn the sampling distribution from data, in an unsupervised manner.

III. UNSUPERVISED LEARNING OF SAMPLING DISTRIBUTIONS

Designing a good sampling distribution π for the particle filter that is simultaneously easy to sample from and yields acceptable performance is a challenging problem. In what follows, we propose to use the sequence of measurements $\{\mathbf{y}_t\}_{t \geq 0}$ to learn a suitable sampling distribution. First, we parametrize the sampling distribution with a multivariate normal and use algorithm unrolling to learn the mean and covariance matrix from data (Section III-A). Second, we use a recurrent neural network (RNN) that learns a hidden state that keeps track of past values of the trajectory (Section III-B). Third, we use a graph neural network that exploits the data structure of the data improving the scalability of the model (Section III-C). Fourth, we learn an arbitrary transform comprised of multi-layer perceptrons (Section III-D). Finally, we discuss how to learn these sampling distributions in an unsupervised manner using only the sequence of available measurements (Section III-E). Simulation results in a myriad of different scenarios can be found in Section IV.

A. Multivariate Normal Parametrization

One distribution that is easy to sample from is the multivariate normal distribution. Therefore, we choose to use this distribution to parametrize the sampling distribution π . The multivariate normal distribution is completely characterized by a mean vector and a covariance matrix. Since it is necessary for the sampling distribution π to depend on the trajectory $\{\mathbf{x}_\tau^{(k)}\}_{\tau=0}^{t-1}$ and the measurements $\{\mathbf{y}_\tau\}_{\tau=0}^t$ up to the current time t , we propose to learn a mapping between these and the mean and covariance matrix of the multivariate normal distribution.

In particular, we consider fully-connected neural networks (also known as multi-layer perceptrons; MLPs) that, inspired by (9), take as input the previous state $\mathbf{x}_{t-1}^{(k)}$ and the current measurement \mathbf{y}_t and return the mean and covariance matrix of the multivariate normal distribution. Namely,

$$\mathbf{x}_t^{(k)} \sim \pi(\mathbf{x}_t^{(k)} | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t), \quad (11)$$

with mean vector given by

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_t(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) \quad (12)$$

and covariance matrix given by

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t). \quad (13)$$

The equalities in both (12) and (13) are used to represent that the mean vector and covariance matrix actually depend on the previous value of the state $\mathbf{x}_{t-1}^{(k)}$ for the k th trajectory and on the measurement \mathbf{y}_t at time t .

The mapping between $\mathbf{x}_{t-1}^{(k)}$ and \mathbf{y}_t and the mean $\boldsymbol{\mu}_t$ at time t is given by a fully-connected neural network NN_t^μ

$$\boldsymbol{\mu}_t(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \text{NN}_t^\mu(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t). \quad (14)$$

This is a cascade of blocks, each of which applies an affine transform characterized by matrix $\mathbf{W}_{t,\ell}^\mu$ and offset vector $\mathbf{b}_{t,\ell}^\mu$ for block ℓ , followed by an activation function $\rho: \mathbb{R} \rightarrow \mathbb{R}$ that is

applied element-wise to the output of the affine transform [28, Ch. 6]. This can be compactly written as

$$\text{NN}_t^\mu(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \mathbf{z}_{t,L}^\mu \text{ with } \mathbf{z}_{t,\ell}^\mu = \rho(\mathbf{W}_{t,\ell}^\mu \mathbf{z}_{t,\ell-1}^\mu + \mathbf{b}_{t,\ell}^\mu) \quad (15)$$

for L blocks, so that $\ell = 1, \dots, L$. Essentially, each block (the affine transform followed by the activation function) is applied to the output of the previous block, forming a cascade. The input $\mathbf{z}_{t,0}^\mu$ to the first block is given by the concatenation of the previous state and the current measurement $[(\mathbf{x}_{t-1}^{(k)})^\top, \mathbf{y}_t^\top]^\top \in \mathbb{R}^{N+M}$. This implies that the matrix in the first affine transform $\mathbf{W}_{t,1}^\mu$ is of size $F_1 \times (M+N)$ where the value of F_1 is known as the number of (hidden) features at the output of block 1. We also have that $\mathbf{b}_{t,1}^\mu \in \mathbb{R}^{F_1}$. In general, $\mathbf{W}_{t,\ell}^\mu \in \mathbb{R}^{F_\ell \times F_{\ell-1}}$ and $\mathbf{b}_{t,\ell}^\mu \in \mathbb{R}^{F_\ell}$, so that each block transforms the $F_{\ell-1}$ input features into F_ℓ output features. The output of the multi-layer perceptron is the output of the last layer $\mathbf{z}_{t,L}^\mu \in \mathbb{R}^{F_L}$. Note that, since this output represents the value of the mean $\mathbf{z}_{t,L}^\mu = \boldsymbol{\mu}_t$, it has to hold that $F_L = N$ which is the size of the sampled state $\mathbf{x}_t^{(k)}$. The activation function $\rho: \mathbb{R} \rightarrow \mathbb{R}$ is applied elementwise to the output of each affine transform, and therefore does not alter the dimensions.

In machine learning, the set of matrices and vectors that form the affine transforms of each block $\boldsymbol{\Theta}_t^\mu = \{\mathbf{W}_{t,\ell}^\mu, \mathbf{b}_{t,\ell}^\mu\}_{\ell=1}^L$ are called the parameters of the fully-connected neural network. These are typically learned from data by solving an optimization problem [28, Ch. 8]. See Section III-E for more details. The number of blocks L , and the number of features F_ℓ at the output of each block $\ell = 1, \dots, L-1$ are design choices and are known as hyperparameters. While there exist methods for choosing hyperparameters [38], they are typically determined by experimentation. The activation function ρ is also a design choice and is typically a nonlinear function such as a rectified linear unit $\rho(x) = \text{ReLU}(x) = \max\{x, 0\}$ or a hyperbolic tangent $\rho(x) = \tanh(x)$.

We remark that, in (15), we are choosing to model the mapping from the previous state value $\mathbf{x}_{t-1}^{(k)}$ and the measurement \mathbf{y}_t to the target mean value $\boldsymbol{\mu}_t$ by means of a different fully-connected neural network for each time instant, as indicated by the subscript t . This approach is known as the unrolling of the algorithm [39].

There are two main reasons for choosing a fully-connected neural network to parametrize the mapping from the previous state value $\mathbf{x}_{t-1}^{(k)}$ and the current measurement \mathbf{y}_t to the target mean value $\boldsymbol{\mu}_t = \text{NN}_t^\mu(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t)$. From a theoretical perspective, fully-connected neural networks can approximate any Borel measurable function with an arbitrary degree of accuracy if the number of features is large enough. This is known as the universal approximation theorem [40], [41]. From a practical perspective, neural networks are somewhat easy to train with gradient-based methods due to the fact that their learnable parameters are in the linear operation of the architecture, and not in the nonlinear one. This makes the optimization problems easier to solve [28, Ch. 8].

To map the previous state and the current measurement to the covariance matrix, we propose to first leverage a fully-connected neural network to learn a representation of the

data, and then build a distance-based kernel from it. More specifically,

$$\Sigma_t = \mathbf{C} \mathbf{K}(\mathbf{z}_t) \mathbf{C}^\top \quad (16)$$

where $\mathbf{C} \in \mathbb{R}^{N \times N}$ is a matrix that is learnable from data, and where $\mathbf{K}(\mathbf{z}_t) \in \mathbb{R}^{N \times N}$ is a distance-based kernel matrix such as

$$[\mathbf{K}(\mathbf{z}_t)]_{ij} = \exp(-([\mathbf{z}_t]_i - [\mathbf{z}_t]_j)^2) \quad (17)$$

for the representation \mathbf{z}_t obtained from a neural network as

$$\mathbf{z}_t = \mathbf{z}(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \text{NN}^\Sigma(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t). \quad (18)$$

Basically, we first transform the previous state $\mathbf{x}_{t-1}^{(k)}$ and the current measurement \mathbf{y}_t into a representation vector $\mathbf{z}_t \in \mathbb{R}^N$ as in (18). Then we compute the distance-based kernel matrix $\mathbf{K}(\mathbf{z}_t) \in \mathbb{R}^{N \times N}$ as in (17). Finally, we learn matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ to account for possible changes in direction and rotations of the variance components as specified in (16). Overall, the resulting covariance matrix Σ_t is guaranteed to be non-negative definite.

Note that the covariance matrix Σ_t is different for each time instant because the input to the neural network (18) used to build the representation \mathbf{z}_t changes with time. However, the learning architectures are the same for all time instants. The set of parameters to learn from data is comprised of the matrices $\mathbf{W}_\ell^\Sigma \in \mathbb{R}^{F_\ell \times F_{\ell-1}}$ and offset vectors $\mathbf{b}_\ell^\Sigma \in \mathbb{R}^{F_\ell}$ of each layer of the neural network in (18), as well as the matrix \mathbf{C} in (16). Thus, the set of learnable parameters $\Theta^\Sigma = \{\{\mathbf{W}_\ell^\Sigma, \mathbf{b}_\ell^\Sigma\}_{\ell=1}^L, \mathbf{C}\}$ is the same for all time instants. For the sake of completeness, we note that, similar to (15), the input to the neural network (18) is the concatenation of the previous state and the current measurement $\mathbf{z}_{t,0}^\Sigma = [(\mathbf{x}_{t-1}^{(k)})^\top, \mathbf{y}_t^\top]^\top$ so that $F_0 = N + M$. Likewise, the representation in (18) is collected as the output of the last layer $\mathbf{z}_t = \mathbf{z}_{t,L}^\Sigma$ so that $F_L = N$. The decision to make the architecture that learns the covariance matrix fixed with time is to avoid the number of parameters growing proportionally to both time and the square of the dimension of the state (through the learnable matrix \mathbf{C}). This is different from the architecture for learning the mean, in that the latter which grows proportionally to time and the dimension of the state – not quadratically with it.

B. Recurrent Neural Networks

Parametrizing the sampling distribution π with a multivariate normal distribution makes it easy to sample from. Learning the mean and covariance matrix as described in Section III-A only takes into account the current measurement and the immediate past value of the state. While this is suggested by the minimum-degeneracy sampling distribution (9), it may be the case that including past information beyond the immediate one is helpful. To do so in a way that avoids ever-increasing dimensionality, we consider recurrent neural networks (RNNs) [28, Ch. 10], [42].

RNNs are machine learning architectures conceived to learn from sequential data. Given an input sequence, they learn an internal representation known as a hidden state. This hidden state is expected to capture past information from the sequence that is relevant for the task at hand. More specifically, given

the input sequence, which in our case is $\{[(\mathbf{x}_{t-1}^{(k)})^\top, \mathbf{y}_t^\top]^\top\}_t$, a sequence of internal states $\{\mathbf{z}_t\}_t$ for $\mathbf{z}_t \in \mathbb{R}^H$ is computed as

$$\mathbf{z}_t = \rho\left(\mathbf{A}\mathbf{z}_{t-1} + \mathbf{B} \begin{bmatrix} \mathbf{x}_{t-1}^{(k)} \\ \mathbf{y}_t \end{bmatrix}\right). \quad (19)$$

The matrices $\mathbf{A} \in \mathbb{R}^{H \times H}$ and $\mathbf{B} \in \mathbb{R}^{H \times (M+N)}$ are the parameters learned from data. Note that these matrices are the same for all time instants, and thus, unlike the method in Section III-A, the number of parameters to learn does not depend on the length of the sequence.

The computation of the hidden state as in (19) may be unable to capture long-term dependencies [43], [44]. To overcome this, we actually employ long short-term memory (LSTM) architectures. These architectures add a series of gating strategies to be able to control the influx of present information with respect to past values of the input sequence. Formally, the internal states are computed as

$$\mathbf{z}_t = \rho\left(\underbrace{\mathbf{f}_t \mathbf{s}_{t-1} + \mathbf{g}_t \rho\left(\mathbf{A}\mathbf{z}_{t-1} + \mathbf{B} \begin{bmatrix} \mathbf{x}_{t-1}^{(k)} \\ \mathbf{y}_t \end{bmatrix}\right)}_{\mathbf{s}_t}\right), \quad (20)$$

where \mathbf{s}_t is an LSTM cell's internal state, \mathbf{f}_t is a forget gate unit, defined in (21), and \mathbf{g}_t is an input gate unit, defined in (22)

$$\mathbf{f}_t = \rho\left(\mathbf{A}^f \mathbf{z}_{t-1} + \mathbf{B}^f \begin{bmatrix} \mathbf{x}_{t-1}^{(k)} \\ \mathbf{y}_t \end{bmatrix}\right), \quad (21)$$

$$\mathbf{g}_t = \rho\left(\mathbf{A}^g \mathbf{z}_{t-1} + \mathbf{B}^g \begin{bmatrix} \mathbf{x}_{t-1}^{(k)} \\ \mathbf{y}_t \end{bmatrix}\right). \quad (22)$$

Notice that the cells of LSTM are RNN. Thus, the main difference between the LSTM and a vanilla RNN is that the former has an additional internal recurrence – a self-loop – in addition to the outer recurrence of the RNN, which is (20). More details can be found in [45], [46]. Gated recurrent units (GRUs) constitute an alternative architecture to LSTMs, and their description can be found in [47], [48].

The mean of the multivariate normal can then be learned at each time instant by computing an affine transform on the hidden state

$$\boldsymbol{\mu}_t = \mathbf{W}_{\text{RNN}}^\mu \mathbf{z}_t + \mathbf{b}_{\text{RNN}}^\mu, \quad (23)$$

with $\mathbf{W}_{\text{RNN}}^\mu \in \mathbb{R}^{N \times H}$ and $\mathbf{b}_{\text{RNN}}^\mu \in \mathbb{R}^N$ being learnable parameters. For the covariance matrix, we compute the kernel of an affine transform of the hidden state [cf. (16)]

$$\Sigma_t = \mathbf{C} \mathbf{K}(\mathbf{W}_{\text{RNN}}^\Sigma \mathbf{z}_t + \mathbf{b}_{\text{RNN}}^\Sigma) \mathbf{C}^\top, \quad (24)$$

where $\mathbf{W}_{\text{RNN}}^\Sigma \in \mathbb{R}^{N \times H}$, $\mathbf{b}_{\text{RNN}}^\Sigma \in \mathbb{R}^N$ and $\mathbf{C} \in \mathbb{R}^{N \times N}$ are all learnable parameters.

The set of learnable parameters for the RNN is $\Theta^{\text{RNN}} = \{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{W}_{\text{RNN}}^\mu, \mathbf{b}_{\text{RNN}}^\mu, \mathbf{W}_{\text{RNN}}^\Sigma, \mathbf{b}_{\text{RNN}}^\Sigma\}$. Note that this set is independent of time, which allows for scalability to arbitrarily long sequences. The information about past states and measurements is captured by the sequence of hidden states $\{\mathbf{z}_t\}$. Thus, learning the mean and covariance by means of (19)–(24) has the potential to leverage information that is not directly

accessible to the method proposed in Section III-A, albeit in compressed form.

Finally, we would like to note that the expressivity of the hidden state can be enhanced by considering deep RNNs, instead of single-layer ones. The same holds for the mean vector and the covariance matrix, where the affine transforms in (23) and (24) can be replaced by fully-connected neural networks.

C. Exploiting Data Structure: Graph Neural Networks

The fully-connected NN architecture used to map the previous state and the current measurement to the mean vector and covariance matrix of a multivariate normal (Section III-A) learns an affine transform from $(N + M)$ -dimensional vectors to N -dimensional ones. The RNN architecture (Section III-B) maps analogous dimensions but does so in a way that exploits the time structure by learning a hidden state that keeps track of past values of the measurement sequence and the simulated trajectory. The number of learnable parameters in both architectures depends on the size of the state N and on the size of the measurements M . Oftentimes, the states or measurements may present additional structure that can further regularize the number of learnable parameters to be independent of either N or M , improving scalability.

One such particular case is that of graph neural networks (GNNs) [49], [50], [51]. GNNs consider the input to be a graph signal [52], [53], [54]. Given a graph support $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is a set of N nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, a graph signal associates an F -dimensional vector $\mathbf{x}_n \in \mathbb{R}^F$ to each node, $n = 1, \dots, N$. The collection of N vectors of dimension F can be compactly written in a matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$ where each row is given by the vector \mathbf{x}_n^T . Graph signals can be used to model measurements that arise in distributed plants [55], power grids [56], communication networks [57], brain activity [58], and teams of autonomous agents [59], among many others [60].

To exploit the underlying graph structure, we require an operation that only relates measurements if they are connected by an edge. Towards this end, let $\mathbf{S} \in \mathbb{R}^{N \times N}$ be a matrix description of the graph, i.e. it satisfies that $[\mathbf{S}]_{ij} = 0$ whenever $(j, i) \notin \mathcal{E}$ for $i \neq j$. The most popular choices of graph matrix descriptions in the literature include the adjacency matrix [52], [54], the Laplacian matrix [53], [60], the random walk matrix [61], and their normalized counterparts. Then, we can define the graph convolutional filter [62] as a linear operation on the input graph signal \mathbf{X} whose output is computed by means of a D -order polynomial on the graph matrix description \mathbf{S}

$$\mathbf{Y} = \mathbf{W}(\mathbf{X}; \mathbf{S}) = \sum_{d=0}^D \mathbf{S}^d \mathbf{X} \mathbf{W}_d. \quad (25)$$

The operation $\mathbf{S}^d \mathbf{X}$ in (25) gathers information located at the d -hop neighborhood of each node by means of d successive exchanges with one-hop neighbors. Multiplying $\mathbf{S}^d \mathbf{X}$ by the filter coefficients in $\mathbf{W}_d \in \mathbb{R}^{F \times G}$ determines how much weight to assign to the information at each d -hop neighborhood. Note that the output is a graph signal $\mathbf{Y} \in \mathbb{R}^{N \times G}$ consisting of G -dimensional vectors at each node, $\mathbf{y}_n \in \mathbb{R}^G$. The set of

$(D + 1)$ filter coefficients $\{\mathbf{W}_k\}_{k=0}^D$ amounts to $FG(D + 1)$ parameters which may be learned from data.

A GNN is a particular case of the fully-connected neural network, where the affine transform is replaced by a graph convolutional filter (25). Then, the mean vector can be learned as $\boldsymbol{\mu}_t = \boldsymbol{\mu}_t(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \text{GNN}_t^\mu(\mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t; \mathbf{S}) = \mathbf{Z}_{t,L}^\mu$ where

$$\mathbf{Z}_{t,\ell}^\mu = \rho(\mathbf{W}_{t,\ell}^\mu(\mathbf{Z}_{t,\ell-1}^\mu; \mathbf{S}) + \mathbf{B}_{t,\ell}^\mu). \quad (26)$$

Here, the graph convolutional filter at each layer $\mathbf{W}_{t,\ell}^\mu$ is of order $D_{t,\ell}$ and maps $F_{t,\ell-1}$ -dimensional input graph signals into $F_{t,\ell}$ -dimensional output graph signals. Thus, each graph filter is characterized by a set of $(D_{t,\ell} + 1)$ filter coefficients $\{\mathbf{W}_{t,\ell,d}^\mu\}_{d=0}^{D_{t,\ell}}$, where $\mathbf{W}_{t,\ell,d}^\mu \in \mathbb{R}^{F_{t,\ell-1} \times F_{t,\ell}}$. The off-set matrix $\mathbf{B}_{t,\ell}^\mu \in \mathbb{R}^{N \times F_{t,\ell}}$ is actually computed as $\mathbf{B}_{t,\ell}^\mu = [b_{t,\ell,1} \mathbf{1}_N, \dots, b_{t,\ell,F_{t,\ell}} \mathbf{1}_N]$ where the f th column is $b_{t,\ell,f} \mathbf{1}_N$ with $\mathbf{1}_N \in \mathbb{R}^N$ a vector of all ones and $b_{t,\ell,f}$ the learnable coefficient, for $f = 1, \dots, F_{t,\ell}$. The input to the GNN is given by $\mathbf{Z}_{t,0}^\mu = [\mathbf{x}_{t-1}^{(k)}, \mathbf{A}_t^\mu \mathbf{y}_t] \in \mathbb{R}^{N \times 2}$ so that $F_0 = 2$. The matrix $\mathbf{A}_t^\mu \in \mathbb{R}^{N \times M}$ is used to adapt the potentially different dimensions of the measurement and the state. The number of features at the output of the GNN is $F_{t,L} = 1$ so that $\mathbf{Z}_{t,L}^\mu \in \mathbb{R}^{N \times F_{t,L}}$ becomes an N -dimensional vector that is used as the mean vector for the multivariate normal distribution $\mathbf{z}_{t,L}^\mu = \boldsymbol{\mu}_t$. Finally, we note that the set of learnable parameters for the GNN-based architecture for learning the mean vector is given by $\Theta_t^{\mu, \text{GNN}} = \{\{\mathbf{W}_{t,\ell,d}^\mu\}_{d=0}^{D_{t,\ell}}, \{b_{t,\ell,f}^\mu\}_{f=1}^{F_{t,\ell}}\}_{\ell=1}^L$. This amounts, for each time t , to $\sum_{\ell=1}^L (F_{t,\ell} F_{t,\ell-1} (D_{t,\ell} + 1) + F_{t,\ell})$ learnable parameters, a quantity determined by design choices and independent of the size of the measurement N . Thus, the dimensionality of the optimization landscape is also independent of N , allowing for scalability.

To learn the covariance matrix, we follow the same scheme as in (16)–(18), except that we replace (18) with a GNN. The input to the first layer, then, is analogous to that of the GNN-based architecture for learning the mean vector, except it may potentially have a different adaptation matrix $\mathbf{A}^\Sigma \in \mathbb{R}^{M \times N}$, i.e. $\mathbf{Z}_{t,0}^\Sigma = [\mathbf{x}_{t-1}^{(k)}, \mathbf{A}^\Sigma \mathbf{y}_t] \in \mathbb{R}^{N \times 2}$ so that $F_0 = 2$. The output $\mathbf{Z}_{t,L}^\Sigma \in \mathbb{R}^{N \times F_L}$ is set to be a vector $\mathbf{z}_{t,L}^\Sigma$, meaning $F_L = 1$, and then is fed into (17) and later into (16). The GNN is independent of time in the same way that the fully-connected NN in (18) is. This makes the set of learnable parameters to be $\Theta^{\Sigma, \text{GNN}} = \{\{\mathbf{W}_{\ell,d}^\Sigma\}_{d=0}^{D_\ell}, \{b_{\ell,f}^\Sigma\}_{f=1}^{F_\ell}\}_{\ell=1}^L$. Note, however, that since the matrix \mathbf{C} in (16) is of size $N \times N$, the overall number of learnable parameters required for the covariance matrix does depend on N . Additionally, the matrices $\{\{\mathbf{A}_t^\mu\}, \mathbf{A}^\Sigma\}$ can either be learned or designed to reflect some sort of topological structure on the measurements as well. Note, however, that if they are learned, then the number of learnable parameters will depend on N , potentially hindering scalability.

GNNs exploit the assumption that the measurements exhibit a graph-based structure. They are naturally distributed architectures, meaning that each node in the graph can compute its output separately, requiring only to communicate with nearby neighbors. They are also better at generalizing for graph-based data, since they are permutation equivariant and stable to deformations of the graph support [63], [64]. There is

a vast body of literature on GNNs, whereby graph convolutional filters (25) can be replaced by non-convolutional filters [65] or where the pointwise activation functions can be replaced by graph-based activation functions [66]. Also, graph RNNs have been developed to handle sequences of graph signals [67]. The key conceptual aspect to highlight is that, if the measurements present certain data structures, it could be convenient to consider learning architectures that exploit such structures [28, Ch. 9].

D. Arbitrary Transform

All the learning methods discussed in the previous sections parameterize the sampling distribution as a multivariate normal. Then, each of them provides different ways of learning the mean vector and covariance matrix that take into account different aspects of the data. In this section, we consider an arbitrary transform of a uniform distribution, increasing the expressivity of the attainable sampling distribution.

A well-known result from probability theory states that, given the cumulative distribution function F_X of a random variable \mathbf{x} , then the random variable $\mathbf{u} = F_X(\mathbf{x})$ is uniformly distributed [68]. Therefore, defining F_X^{-1} as the generalized inverse of F_X , we can obtain samples from any random variable \mathbf{x} by taking samples from a uniform \mathbf{u} and transforming those samples through F_X^{-1} . Computing F_X^{-1} , however, is typically intractable.

We propose to learn a transform $\Psi_t : \mathbb{R}^N \rightarrow \mathbb{R}^N$ that maps a uniform random variable \mathbf{u} into the sample $\mathbf{x}_t^{(k)}$ for each time t . To do this, we parametrize Ψ_t by means of a neural network. In order to include the information from the past state and the current measurement, we consider the following architecture

$$\mathbf{x}_t^{(k)} = \Psi_t(\mathbf{u}; \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t) = \text{NN}_t^\Psi(\rho(\mathbf{A}_t^\Psi \mathbf{u} + \mathbf{B}_t^\Psi \mathbf{x}_{t-1}^{(k)} + \mathbf{C}_t^\Psi \mathbf{y}_t)) \quad (27)$$

where $\mathbf{u} \sim \mathcal{U}([0, 1]^N)$, $\mathbf{A}_t^\Psi \in \mathbb{R}^{N \times N}$, $\mathbf{B}_t^\Psi \in \mathbb{R}^{N \times N}$ and $\mathbf{C}_t^\Psi \in \mathbb{R}^{N \times M}$.

Each sampled state $\mathbf{x}_t^{(k)}$ in the trajectory [cf. (27)] is distributed according to some probability distribution $\mathbf{x}_t^{(k)} \sim \pi_t^\Psi(\mathbf{x}_t^{(k)} | \mathbf{x}_{t-1}^{(k-1)}, \mathbf{y}_t)$. A means to evaluate values of π_t^Ψ is required in order to be able to compute the weight associated to each trajectory [cf. (8)]. Observing that Ψ_t is continuous, and since \mathbf{u} is a continuous random variable, then $\mathbf{x}_t^{(k)}$ is continuous as well, and thus it can be characterized by a continuous probability density function (pdf). Provided that Ψ_t is invertible, this pdf can be computed as follows

$$\pi_t^\Psi(\mathbf{x}_t^{(k)} | \mathbf{x}_{t-1}^{(k-1)}, \mathbf{y}_t) = |\det(\mathbf{J}_{\Psi_t^{-1}}(\mathbf{x}_t^{(k-1)}))| \quad (28)$$

since $f_X(\Psi_t^{-1}(\mathbf{x}_t^{(k)})) = 1$, and where $\mathbf{J}_{\Psi_t^{-1}} \in \mathbb{R}^{N \times N}$ is the Jacobian of Ψ_t^{-1} . This Jacobian can be computed by noting that

$$\Psi^{-1}(\mathbf{x}_t^{(k)}) = (\mathbf{A}_t^\Psi)^{-1}(\mathbf{z}_{t,0}^\Psi - \mathbf{B}_t^\Psi \mathbf{x}_{t-1}^{(k)} - \mathbf{C}_t^\Psi \mathbf{y}_t) \quad (29)$$

with

$$\mathbf{z}_{t,\ell-1}^\Psi = (\mathbf{W}_{t,\ell}^\Psi)^{-1} \rho^{-1}(\mathbf{z}_{t,\ell}^\Psi - \mathbf{b}_{t,\ell}^\Psi) \quad (30)$$

for $\ell = 1, \dots, L$ the number of layers in the neural network NN_t^Ψ , and where $(\mathbf{W}_{t,\ell}^\Psi, \mathbf{b}_{t,\ell}^\Psi)$ are the parameters of the affine

transform of layer ℓ . Note that $\mathbf{z}_{t,L}^\Psi = \mathbf{x}_t^{(k)}$. In other words, Ψ^{-1} has a structure analog to that of a neural network, and thus its Jacobian can be computed by an algorithm analog to backpropagation [69].

While we assumed that Ψ is invertible from the moment we decided to use it to learn F_X^{-1} , neural networks are not necessarily invertible, and thus we need to guarantee this in the design. The simplest way to do so is to choose square matrices \mathbf{W}_ℓ^Ψ for all ℓ , and thus $F_\ell = N$ for all ℓ . This curtails the ability of the designer to control the representation capability of the neural network. We note that a square matrix does not necessarily guarantee invertibility. One way to approximately compute the inverse, then, would be to add a small identity matrix, which is a common practice in ridge regression estimators [70]. Also, the activation function has to be invertible, which is the case for the hyperbolic tangent, but not for the ReLU. More specific solutions can be found in the field of normalizing flows [71], [72], [73].

The set of learnable parameters for each time instant t is given by $\Theta_t^\Psi = \{\{\mathbf{W}_{t,\ell}^\Psi, \mathbf{b}_{t,\ell}^\Psi\}_{\ell=1}^L, \mathbf{A}_t^\Psi, \mathbf{B}_t^\Psi, \mathbf{C}_t^\Psi\}$. Note that, in (27), we have chosen to learn a different neural network for each time instant t . This makes the number of learnable parameters a function of t , and thus requires more data as the trajectories get longer. Alternatively, we can fix a single neural network for all t , and use the values of $\mathbf{x}_{t-1}^{(k)}$ and \mathbf{y}_t to keep track of the changes in the system across time.

E. Likelihood of the Model

All of the architectures presented so far are determined by a set of parameters. These parameters are updated iteratively during the training phase by computing gradient descent steps towards optimizing some objective function [28, Ch. 8]. We train the architectures by choosing to maximize the likelihood of the model. This results in an unsupervised regime, since only the sequence of measurements $\{\mathbf{y}_t\}$ is required for training. Notice that we consider a sequence of measurements $\{\mathbf{y}_t\}$ from one single trajectory.

The model of the nonlinear dynamical system under study is characterized by the transition distribution $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ and the measurement distribution $p(\mathbf{y}_t | \mathbf{x}_t)$, both considered known. Hence, we learn the sampling distributions π in an unsupervised way by maximizing the likelihood of the model

$$\max_{\Theta} \sum_{k=1}^K p(\mathbf{x}_t^{(k)} | \Theta) p(\mathbf{x}_{t-1}^{(k)} | \Theta) p(\mathbf{y}_t | \mathbf{x}_t^{(k)} | \Theta). \quad (31)$$

By the notation $\mathbf{x}_t^{(k)} | \Theta$ we have explicitly indicated that each sample depends on the learnable parameters Θ through the learned sampling distribution Θ . By updating the parameters Θ towards maximizing (31), we attempt to generate samples that are reasonable in light of the system dynamics. For instance, if one sampled value $\mathbf{x}_t^{(k)}$ would not be reasonable in light of the measurement \mathbf{y}_t , as dictated by the distribution $p(\mathbf{y}_t | \mathbf{x}_t)$, then the parameters Θ are going to be updated, subsequently reducing the probability of sampling such $\mathbf{x}_t^{(k)}$. In this way, we can train the architectures using only knowledge of the

system dynamics and the measurements \mathbf{y}_t , but with no true knowledge of the states \mathbf{x}_t , which amounts to an unsupervised training regime. Notice that to compute (31) and train in an unsupervised regime, we assume that we *know both the system dynamics and the measurement processes*. This is different from [74], [75], where they parameterize both the dynamic and measurement processes as Gaussian distributions with mean and covariance defined as learnable Gaussian processes. Our approach also differs from [76], where they assume that the system dynamics are deterministic, and they approximate them with a polynomial.

The parameters are typically updated by means of some iterative algorithm based on stochastic gradient ascent [77], [78]. We note that the objective is not necessarily to optimize the function, but rather to learn to solve the task at hand. For this reason, usually a fixed number of iterations are done, instead of following some stopping criterion based on the value of the objective function. Oftentimes, it may be convenient to maximize the log-likelihood, especially when the distributions of the system dynamics belong to the exponential parametric family. Finally, we remark that since $\mathbf{x}_t^{(k)}$ is actually a sample drawn from the distribution π , which is the one we want to learn, we use the reparametrization trick to be able to estimate the gradients required for the stochastic gradient ascent algorithm [79]. This trick allows us to rewrite an expectation with respect to the sampling distribution π as a differentiable Monte Carlo estimator. In a nutshell, the idea is to rewrite the random variable that we want to sample as a deterministic variable that depends on an auxiliary random variable, which typically is a Gaussian random variable.

IV. NUMERICAL EXPERIMENTS

Let us consider a dynamical system given by

$$\begin{aligned}\mathbf{x}_t &= \phi(\mathbf{F}\mathbf{x}_{t-1}) + \mathbf{v}_t, \\ \mathbf{y}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{q}_t,\end{aligned}\quad (32)$$

where $\mathbf{x}_t \in \mathbb{R}^N$ is the state, $\mathbf{y}_t \in \mathbb{R}^M$ is the measurement, $\mathbf{F} \in \mathbb{R}^{N \times N}$ is the state transition matrix and $\mathbf{H} \in \mathbb{R}^{M \times N}$ is the measurement matrix. The function $\phi: \mathbb{R}^N \rightarrow \mathbb{R}^N$ may be nonlinear (depending on the simulation scenario). The state noise $\mathbf{v}_t \in \mathbb{R}^{N \times N}$ has mean $\mathbb{E}[\mathbf{v}_t] = \mathbf{0}$ and covariance matrix $\mathbb{E}[\mathbf{v}_t \mathbf{v}_t^T] = \sigma_v^2 \mathbf{I}_N \in \mathbb{R}^{N \times N}$ for all t . Likewise for the measurement noise $\mathbf{q}_t \in \mathbb{R}^M$, where $\mathbb{E}[\mathbf{q}_t] = \mathbf{0}$ and $\mathbb{E}[\mathbf{q}_t \mathbf{q}_t^T] = \sigma_w^2 \mathbf{I}_M \in \mathbb{R}^{M \times M}$ for all t . Noise vectors are independent for any pair t, t' such that $t \neq t'$, and all the random vectors in the sequence $\{\mathbf{v}_t\}_{t \geq 0}$ are independent from those in the sequence $\{\mathbf{q}_t\}_{t \geq 0}$.

In this context, we consider to have access to a sequence of t measurements $\{\mathbf{y}_t\}_{t=0, \dots, T}$ and we want to estimate the value \mathbf{x}_t of the state at time t , so that the target quantity becomes $\mathbf{z}_t = \mathbf{f}_t(\mathbf{x}_{0:t}) = \mathbf{x}_t$. The baseline estimator is then given by $\hat{\mathbf{z}}_t$ as in (4). In what follows, we will obtain estimates $\hat{\mathbf{z}}_t$ using particle filtering as in (6), under the consideration of different sampling distributions $\pi(\mathbf{x}_t | \mathbf{x}_{0:t-1}^{(k)}, \mathbf{y}_{0:t})$ as discussed in Section III.

The baseline sampling distribution is given by the one that minimizes the degeneracy, i.e. $\pi(\mathbf{x}_t | \mathbf{x}_{0:t-1}^{(k)}, \mathbf{y}_{0:t}) = p(\mathbf{x}_t | \mathbf{x}_{t-1}^{(k)}, \mathbf{y}_t)$. This sampling distribution can be computed in

closed form for the dynamical system in (32) for the case in which the noise distributions for \mathbf{v}_t and \mathbf{q}_t are Gaussian, see [17] for details.

In all the following simulations, we set $M = N - 2$ with matrix \mathbf{F} being generated as the weighted adjacency matrix of a random geometric graph. The weights are computed using a Gaussian kernel on the distance, i.e. the weight between node i and node j is given by $\exp(-\|\mathbf{r}_i - \mathbf{r}_j\|^2)$ where $\mathbf{r}_i \in \mathbb{R}^2$ are the coordinates of node i on the $[0, 1]$ plane. A weighted, 3-nearest neighbor graph is constructed, and the adjacency matrix is normalized so that it has unit spectral norm. The measurement matrix is obtained from $\mathbf{H} = [\mathbf{I}_{M \times M} \quad \mathbf{I}_{M \times (N-M)}]$ where $\mathbf{I}_{P \times Q}$ is a $P \times Q$ matrix such that $[\mathbf{I}]_{ii} = 1$ and $[\mathbf{I}]_{ij} = 0$ for all $i \neq j$, normalized to have unit spectral norm. Such dynamical system is characteristic of diffusion processes in graphs, including rumor spreads, heat diffusion, and graph filtering [80].

The initial state \mathbf{x}_0 is drawn from a multivariate Gaussian with mean $\mathbb{E}[\mathbf{x}_0] = \boldsymbol{\mu}^0 = \mathbf{1}_N$ and covariance matrix $\mathbb{E}[(\mathbf{x}_0 - \boldsymbol{\mu}^0)(\mathbf{x}_0 - \boldsymbol{\mu}^0)^T] = \mathbf{I}_N$. In this setting, we define the state noise SNR as $10 \log_{10}(\|\boldsymbol{\mu}_0\|^2 / \sigma_v^2)$ and the measurement noise SNR analogously. For the simulations, we consider the SNR to be 5dB, fixing the value of both σ_v^2 and σ_w^2 .

We construct particle filter estimates drawing K particles, with resampling whenever K_t^{eff} is smaller than $\hat{K}^{\text{thres}} = K/3$. To account for the randomness in the generation of the particle filtering estimate, we repeat the process 100 times. That is, we sample K particles, construct the estimate, sample another K particles, construct another estimate, and so on for 100 repetitions. These estimates are then averaged to produce a single particle filtering estimate that accounts for the inherent randomness in the particle filter.

For the learnable sampling distributions leveraging fully-connected neural networks to learn the mean and covariance matrix of a multivariate distribution, we consider 4 layers, where the input to the first layer is of size $N + M$ and the output of the last layer is of size N , as explained in Section III-A. The number of hidden units is set to 256, 512 and 1,024 at the outputs of layers 1, 2, and 3, respectively. We use the same hyperparameters for the mean neural network and the covariance neural network (note that the hyperparameters are the same, but the parameters actually learned will be different). For the RNN model (Section III-B), the size of the hidden state is set to $H = 1024$. For the GNN (Section III-C) we consider the graph matrix description to be $\mathbf{S} = \mathbf{F}$ the state transition matrix, and we use 4 layers with dimensions $F_{t,1} = 256$, $F_{t,2} = 512$, $F_{t,3} = 1024$ and $F_{t,4} = 1$ for all t . The number of neighborhood exchanges is set to $D = 3$ for all filters involved. Finally, for the learnable sampling distribution capable of learning an arbitrary distribution (i.e. not the parameters of a multivariate normal) described in Section III-D, we consider 9 layers (recall that the number of hidden units is fixed by the requirement that the matrices be squared). In all learnable architectures, the activation function is set to be the hyperbolic tangent.

To train these learnable sampling distributions, we have access to a sequence of t measurements $\{\mathbf{y}_t\}_t$ and we find the parameters that maximize the likelihood of the model as described in Section III-E. To do so, we use an optimization

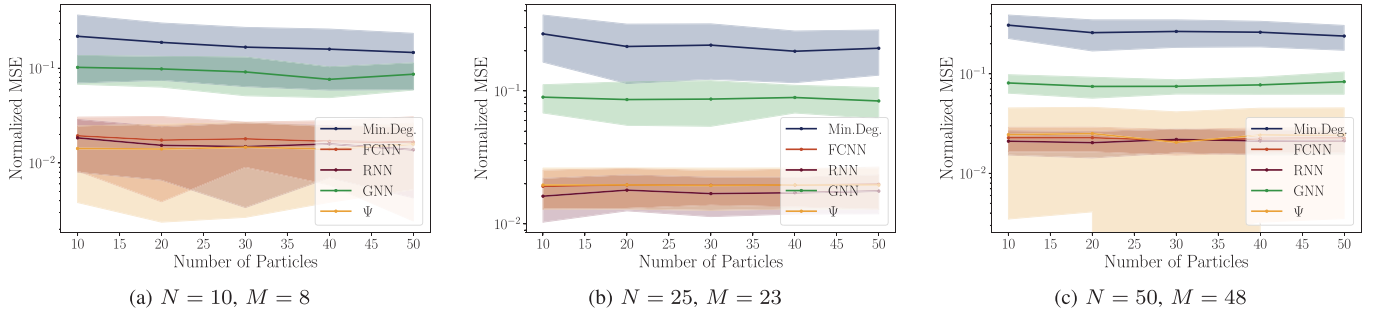


Fig. 1. Linear Gaussian dynamical system as a function of the number of particles K . (a)–(c) We vary the state dimension N and the measurement dimension M . The MSE of the particle filter estimators is normalized by the ground truth, which is the Kalman filter. All the sampling distributions that were learned from data significantly outperform the baseline of the designed, minimum-degeneracy sampling distribution.

algorithm known as Adam [78] which is a momentum-based variant of stochastic gradient descent. We use a learning rate of 0.001, and forgetting factors of 0.9 and 0.999. We carry out 200 training steps, where each step uses all data (also known as 200 epochs). To account for the randomness of the system dynamics (in generating the state transition matrices \mathbf{F}) and of the generated measurements $\{\mathbf{y}_t\}$ we repeat the whole learning and testing process 20 times. We report the median performance of each particle filter estimate, together with the corresponding standard deviation.

In the following simulations, we explore how the performance of particle filters with learnable sampling distributions changes as a function of the state dimension N (recall that the measurement dimension is set to $M = N - 2$), the trajectory length T , the number of particles K , the noise distribution, and the nonlinear function ϕ .

A. Linear Gaussian Dynamical System

First, we consider a linear dynamical system, where ϕ is the identity function

$$\begin{aligned}\mathbf{x}_t &= \mathbf{F}\mathbf{x}_{t-1} + \mathbf{v}_t, \\ \mathbf{y}_t &= \mathbf{H}\mathbf{x}_t + \mathbf{q}_t,\end{aligned}\quad (33)$$

and we also consider both the state noise and the measurement noise to be multivariate Gaussian. In such a scenario, we can compute the ground truth estimator in closed form $\mathbb{E}[\mathbf{x}_t | \mathbf{y}_{0:t}]$ since $\mathbf{x}_t | \mathbf{y}_{0:t}$ is also a multivariate Gaussian random variable. Notice that the ground truth estimator, which is the optimal Bayesian estimator, boils down to the Kalman filter for the linear Gaussian system.

We consider three different values of state dimension, i.e. $N = 10$, $N = 25$ and $N = 50$. The resulting, normalized mean squared error between the particle filter estimators and the ground truth is plotted in Fig. 1 as a function of the number of particles in the set $K \in \{10, 20, 30, 40, 50\}$. In all cases, we set the trajectory length to be $T = 12$.

For all the state dimensions, we observe that the learned sampling distributions considerably outperform the designed, minimum degeneracy sampling distribution. We note that there is no significant difference between using a fully connected neural network (Section III-A), a recurrent neural network (Section III-B) or an arbitrary learned linear transform. Interestingly enough, the arbitrary linear transform Ψ exhibits

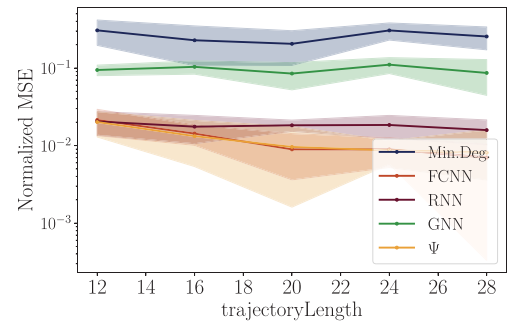


Fig. 2. Linear Gaussian dynamical system as a function of the trajectory length T .

comparable performance to the FCNN and the RNN, even though it relies on significantly less number of parameters. One difference is that the arbitrary linear transform Ψ exhibits higher variance, especially for the large case of $N = 50$. The sampling distribution based on the GRNN still performs better than the minimum-degeneracy designed sampling distribution, but considerably worse than the other learnable distributions. We believe this may be caused by overfitting due to already incorporating information on the underlying graph \mathbf{F} in the form of the matrix \mathbf{S} used in the graph filtering layers.

Next, we fix the state dimension to be $N = 10$, the number of measurements to be $M = 8$, and the number of particles to be $K = 10$, and we simulate the performance of the particle filters with learnable distributions as a function of $T \in \{12, 16, 20, 24, 28\}$.

Results are shown in Fig. 2. Again, we note that the particle filter with learned sampling distribution outperforms the use of a designed, minimum-degeneracy one, with the GNN performing worse than the other three alternatives. We note that the performance of the RNN seems independent of the trajectory length, probably due to the fact that this the parameters of the RNN are time-independent by design. Meanwhile, the performance of the FCNN and the arbitrary linear transform Ψ improve with trajectory length, likely because of the increased expressivity that is obtained by learning a different set of parameters for each time instant. In any case, the performance of these three architectures is comparable.

Finally, we fix the trajectory length to be $T = 12$ and the number of particles to be $K = 10$ and we simulate as a function

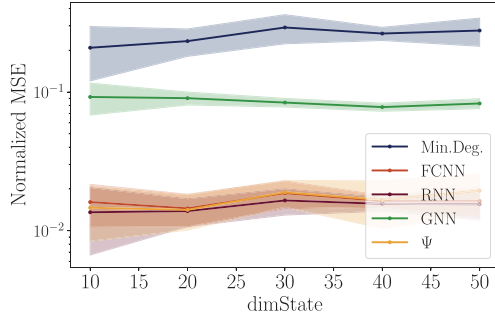


Fig. 3. Linear Gaussian dynamical system as a function of the state dimension N .

of the state dimension for $N \in \{10, 20, 30, 40, 50\}$ (recall that, for each case, $M = N - 2$).

Results shown in Fig. 3 exhibit a similar behavior as in the previous experiments, where all learned distributions perform better than the designed, minimum-degeneracy one, and where the GNN works worse than the rest. In this case, the performance seems to be independent of the dimension of the state N , with all three (FCNN, RNN and arbitrary transform Ψ) performing similarly.

B. Nonlinear Gaussian Dynamical System

Next, we consider a nonlinear dynamical system, where ϕ is set to be the absolute value. The distributions of the state noise and the measurement noise remain Gaussian. In this scenario, the ground truth $\mathbb{E}[\mathbf{x}_t | \mathbf{y}_{0:t}]$ can no longer be computed in closed-form. In this case, we use the simulated values of \mathbf{x}_t as ground truth to measure performance against, i.e. the results shown in the figures report the normalized mean squared error between the different methods and the estimated ground truth value \mathbf{x}_t .

For this case, we fix $N = 10$, $M = 8$ and $T = 12$, and present simulations as a function of the number of particles $K \in \{10, 20, 30, 40, 50\}$. Results are shown in Fig. 4.

The first observation is that, while in general the learned sampling distributions outperform the designed, minimum-degeneracy one, the gap now is significantly smaller. Furthermore, the behavior of learned distributions such as the FCNN or the arbitrary linear transform Ψ is slightly more erratic as a function of the number of particles K , sometimes being comparable to the minimum-degeneracy particle filter, sometimes being slightly worse. This is likely due to the added complexity of the model, and the inability of these generic architectures of adequately capturing the information without proper regularization. As a matter of fact, the impact of regularization (i.e. choosing operations that reflect certain structure, being either time, like RNNs, or graphs, like GNNs) is evident in the fact that the performance of GNNs is now significantly better (in relative terms) and that both the GNN and the RNN offer a stable behavior as a function of K .

C. Linear Non-Gaussian Dynamical System

For the third simulation we consider a linear dynamical system (33), but with non-Gaussian state and measurement

noise. Again, we fix $N = 10$, $M = 8$ and $T = 12$, and present simulations as a function of the number of particles $K \in \{10, 20, 30, 40, 50\}$. Results are shown in Fig. 5 for exponential noise and in Fig. 6 for uniform noise. In both cases, the mean and covariance matrix are still given by the values described at the beginning of the section, but the noise samples are sampled from the corresponding distributions.

As in Section IV-B, the learned sampling distributions outperform the designed, minimum-degeneracy one, by a gap smaller than the one observed in the linear Gaussian case (Section IV-A). We note that the FCNN now exhibits a much more stable behavior than in the nonlinear Gaussian case (Section IV-B), suggesting that the linearity of the system is more important in the performance of the FCNN than the distribution of the noise. Here, the performance of the arbitrary linear transform Ψ is considerably more erratic, making it somewhat unreliable. Regularization techniques may come in handy to avoid this erratic behavior. Finally, we note that the RNN is the more consistent performer with the most stable behavior.

D. Realistic Dynamical Models

In this section we consider two mathematical models that are used to describe real phenomena. The first one consists of a model that is often employed to describe the dynamics of an epidemic: the SIR model. It is described by a system of ordinary differential equations [81] that can be discretized via the Euler method for our purposes. By doing so with a time step of size Δ , the discrete system is given by

$$\begin{aligned} S_{t+1} &= -\beta S_t I_t \Delta + S_t \\ I_{t+1} &= (\beta S_t I_t - \gamma I_t) \Delta + I_t \\ R_{t+1} &= \gamma I_t \Delta + R_t \end{aligned} \quad (34)$$

where S_t is the number of susceptible people at time t , I_t is the number of infected people at time t , and R_t represents the number of people that have been removed from the system (either by death or recovery) by time t . The parameters of the model β , γ , and Δ are assumed known. The state of the system is given by $\mathbf{x}_t = [S_t, I_t, R_t]^T$. Notice that the system described in (34) can be thought of as the nonlinear function ϕ within the framework defined in (32).

The second realistic system we focus on consists of an application to bioinformatics. In particular, we evaluate our algorithm for estimation and filtering within a heat shock response system, an essential cellular mechanism conserved in almost every organism [82]. The model we use corresponds to the heat shock model in the bacterium *Escherichia coli*. It can be expressed as a system of three differential equations

$$\begin{aligned} \dot{D}(t) &= 3 \frac{S(t)}{1 + \frac{0.05D(t)}{1+0.0254U(t)}} - 0.015D(t), \\ \dot{S}(t) &= 10 - 0.03S(t) - 3 \frac{\frac{D(t)}{1+0.0254U(t)}}{1 + \frac{0.05D(t)}{1+0.0254U(t)}} S(t), \\ \dot{U}(t) &= 40(2 \times 10^6 - U(t)) - 6040D(t) \end{aligned} \quad (35)$$

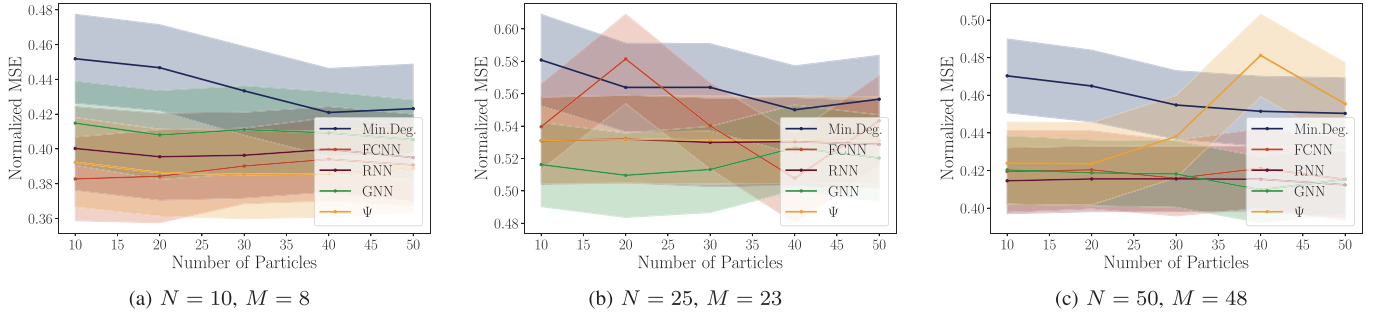


Fig. 4. Nonlinear Gaussian dynamical system as a function of the number of particles K . (a)–(c) We vary the state dimension N and the measurement dimension M . While, generally, the learned distributions outperform the minimum-degeneracy one, the relative gap is now smaller compared to Fig. 1. Additionally, for some distributions (like the FCNN or the arbitrary transform Ψ) the performance is comparable to the minimum degeneracy (and may be worse) for some values of K .

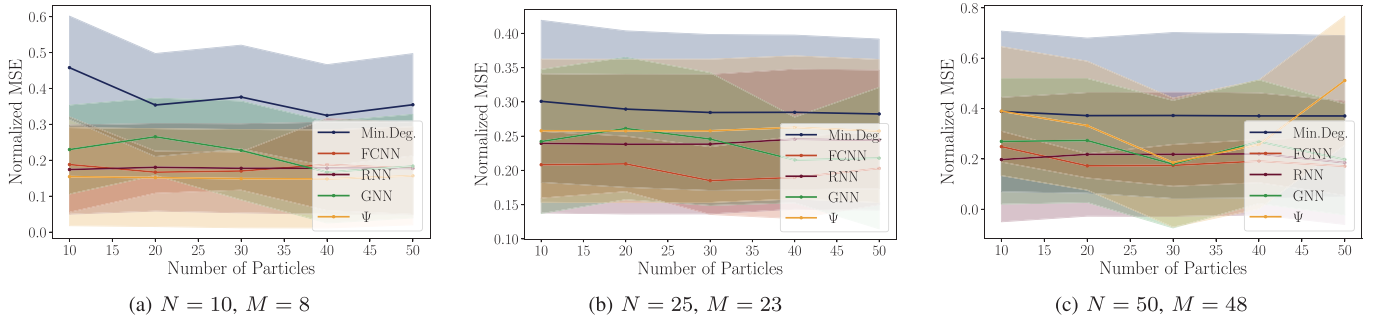


Fig. 5. Linear non-Gaussian dynamical system as a function of the number of particles K . (a)–(c) We vary the state dimension N and the measurement dimension M . The performance of the arbitrary distribution Ψ becomes comparable to the one of the minimum degeneracy as the dimension N of the system grows.

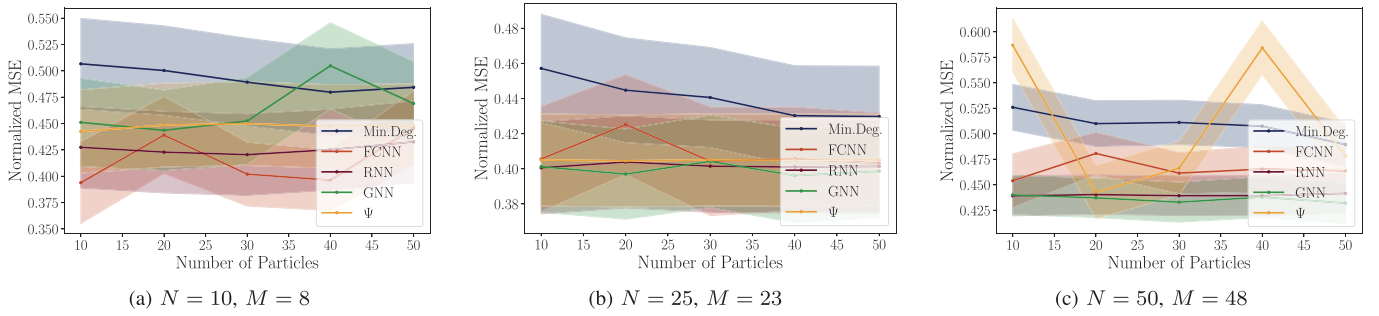


Fig. 6. Linear-exponential dynamical system as a function of the number of particles K . (a) System with state dimension $N = 10$ and measurement dimension $M = 8$. (b) System with state dimension $N = 25$ and measurement dimension $M = 23$. (c) System with state dimension $N = 50$ and measurement dimension $M = 48$. While, generally, the learned distributions outperform the minimum-degeneracy one, the gap is now smaller. Additionally, the performance of the arbitrary distribution Ψ becomes comparable to the one of the minimum degeneracy as the dimension N of the system grows.

where the constants come from replacing the model parameters by its values given in [82]. In the system of equations (35), $D(t)$ represents the number of molecules of chaperones, $S(t)$ is the number of molecules of an enzyme RNA polymerase bound to a key regulatory factor, and $U(t)$ describes the total number of unfolded proteins; see [82] for more biological details.

We first evaluate the performance for the SIR model. For this experiment, we fix the parameters to be $\beta = 5 \times 10^{-4}$ and $\gamma = 0.04$. We set a trajectory of 200 samples, using a time step $\Delta = 0.7$. We measure $M = 2$ states, with I_t being the state that we do not measure. More precisely, \mathbf{H} in (32) consists of the first and third row of an identity matrix of size three. The initial conditions are independently distributed and follow

an exponential distribution, with means $\mathbb{E}[S_0] = 997$, $\mathbb{E}[I_0] = 3$ and $\mathbb{E}[R_0] = 0$, and with variances $\text{Var}(S_0) = \text{Var}(I_0) = \text{Var}(R_0) = 500$. The measurement and state noises are also exponentially distributed, with variances $\text{Var}(\mathbf{q}_t) = 2500$ and $\text{Var}(\mathbf{v}_t) = 200$, respectively. The number of particles is fixed at $K = 300$. The architecture of the neural networks is changed as well, using 4 hidden layers in the case of the FCNN (with 512, 256, 128 and 32 hidden units each, respectively), setting $H = 2048$ in the case of the RNN, and using 10 layers for the arbitrary distribution Ψ . The GNN-based method is not used in this experiment, as there is no graphical structure in the system dynamics that could be exploited. The estimation of the unobserved state using the different methods is shown in

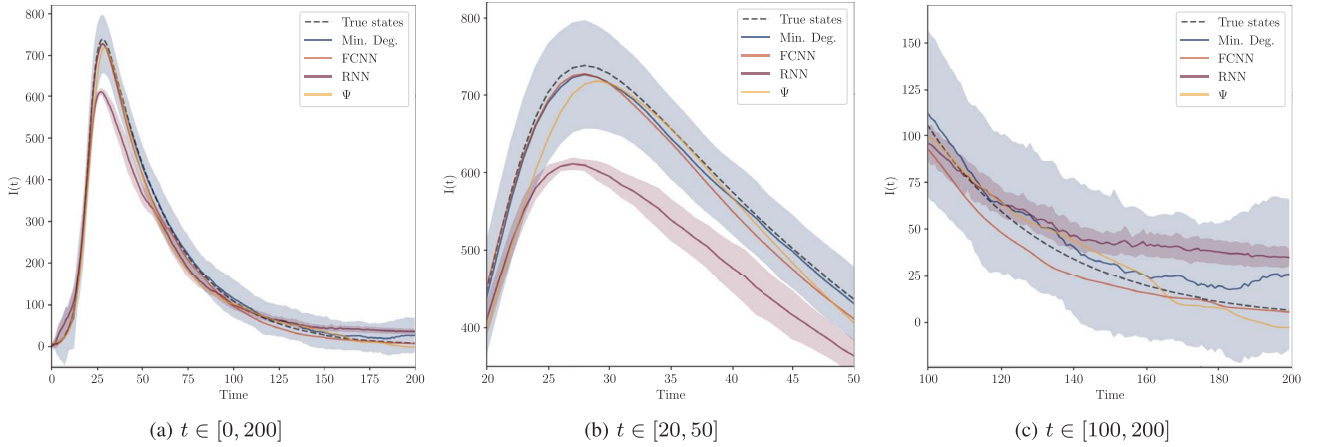


Fig. 7. The unobserved state $I(t)$ and the estimations using the proposed methods for the SIR model. (a) The state $I(t)$ during all the trajectory. (b) The state $I(t)$ zoomed to show the time window where $t \in [20, 50]$. (c) The state $I(t)$ zoomed to show the time window where $t \in [100, 200]$. The performance achieved by the FCNN-based method and the arbitrary distribution Ψ is higher than that of the minimum degeneracy.

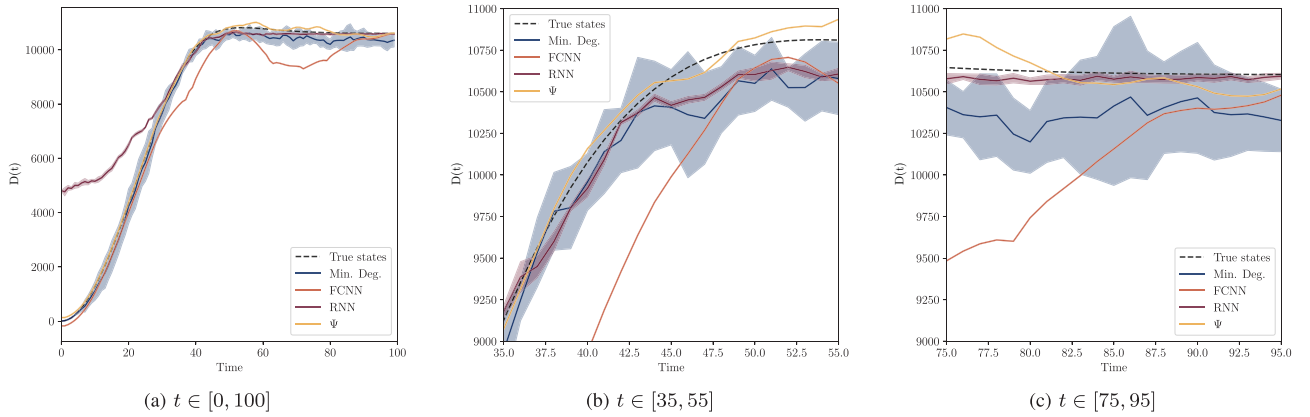


Fig. 8. The unobserved state $D(t)$ and the estimations using the proposed methods for the heat shock response. (a) The state $D(t)$ during all the trajectory. (b) The state $D(t)$ zoomed to show the time window where $t \in [35, 55]$. (c) The state $D(t)$ zoomed to show the time window where $t \in [75, 95]$.

Fig. 7. It is also of interest to evaluate how the learned sampling distributions perform in terms of filtering out the noise from the measured states. The MSE for each method can be seen in Table I.

The FCNN-based method results to be the one that achieves the best performance, not just at estimating the unknown state but also at filtering the noise from the measured ones. This method offers the lowest variance and also the lowest bias. The arbitrary distribution Ψ outperforms the minimum-degeneracy one too, although the bias is somewhat larger than the one attained using the FCNN. Nonetheless, the variance of the estimator is really low as well. The RNN-based method fails to provide a sampling distribution whose performance is comparable to the other two proposed methods. Not only is the variance much higher (although still lower than in the minimum-degeneracy case), but the bias is not desirable either. This can be clearly observed in Fig. 7(b), where the peak of the signal $I(t)$ is correctly captured by the other three methods but not by the RNN-based one.

Now we evaluate our approach for the heat shock response system in (35). In this setting, we assume that we only have access to noisy measurements of $S(t)$ and $U(t)$, and we want to not only filter out the noise from these two signals but also to

TABLE I
NORMALIZED MSE BETWEEN THE TRUE STATES AND THE ESTIMATES FOR THE SIR MODEL

	$S(t)$	$I(t)$	$R(t)$
Min. Deg.	0.082	0.161	0.029
FCNN	0.027	0.042	0.017
RNN	0.062	0.172	0.026
Ψ	0.063	0.077	0.016

estimate $D(t)$, the unobserved state. We discretize (35) via the Euler method, with a time step of $\Delta = 0.05$. The initial conditions are independently distributed and follow normal distributions, with means $\mathbb{E}[D_0] = \mathbb{E}[S_0] = 0$ and $\mathbb{E}[U_0] = 2 \times 10^6$, and with standard deviations $\sigma_d = 245$, $\sigma_s = 3$ and $\sigma_u = 44500$. The state noises are also normally distributed, with variances equal to the ones used in the respective initial conditions. The measurement noise standard deviations are $5\sigma_s$ and $5\sigma_u$ for the states S and U , respectively. We use $K = 500$ particles for this experiment.

The estimation of the unobserved state using the different methods is shown in Fig. 8. The normalized MSE for each method can be seen in Table II for all three states. The normalizing flow method results to be the one that achieves the

TABLE II
NORMALIZED MSE BETWEEN THE TRUE
STATES AND THE ESTIMATES FOR THE
HEAT SHOCK RESPONSE SYSTEM

	$D(t)$	$S(t)$	$U(t)$
Min. Deg.	0.036	0.046	0.053
FCNN	0.081	0.085	0.206
RNN	0.203	0.105	0.041
Ψ	0.010	0.029	0.040

best performance in this case. Nonetheless, contrary to the SIR example, the FCNN method fails to capture the correct dynamics of the system. We can conclude that when the dynamics is highly nonlinear and does not exhibit any underlying structure, the normalizing flow is the best option as it can successfully learn an arbitrary sampling distribution.

V. CONCLUSION

We proposed four different methods for learning the sampling distribution of a particle filter. The first three are parametric methods, which sample from a multivariate Gaussian distribution with mean and covariance matrix that are learned in different ways. First, we consider the generic case, in which the mean and the covariance matrix are obtained through representations learned by means of a fully-connected neural network applied to the current measurement and the previous simulated state. Second, we consider a recurrent neural network that is capable of capturing past information beyond the immediate previous simulated state. Third, we make a case for the possibility of using neural network architectures that exploit additional data structure, if available. In particular, we assume that the nonlinear dynamic system may be explained in terms of a distributed plant, and leverage graph neural networks to exploit this structure. The fourth and last method is a non-parametric one, in which we learn an arbitrary mapping between samples from a uniform distribution and the state simulation.

We ran several simulations studying the performance of the proposed method against the sampling distribution designed to minimize particle degeneracy. Overall, while the learned sampling distributions generally outperform the designed, minimum-degeneracy one, it is the RNN-based ones that consistently does so in a wide range of scenarios, with stable performance for a wide range of sweeping problem hyperparameters. Note, however, that in the case of complex nonlinear systems, the other methods can provide better estimations, as seen in the case of the SIR and the heat shock systems. In particular, the arbitrary distribution is the most flexible, allowing to learn sampling distributions that adjust better to each instant of time and thus lead to better results.

This article presents a first approach to learning sampling distributions in an unsupervised manner as opposed to designing them. There are many directions for improvement of the methods presented herein. First of all, more complex architectures can be considered, which may be better tailored for different specific nonlinear dynamical systems. Second, the training can be improved by including regularization techniques such as dropout or penalties. These may be particularly useful when the dimension of the systems is large. Third, normalizing flows

arise as an interesting direction to pursue. The fourth method presented here is a very elementary normalizing flow. Using more complex conditioning models to ensure the invertibility of the neural network may actually lead to increased representation capability and more sophisticated distributions.

REFERENCES

- [1] F. Gama, N. Zilberstein, R. G. Baraniuk, and S. Segarra, "Unrolling particles: Unsupervised learning of sampling distributions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 5498–5502.
- [2] S. Kumar and R. S. Williams, "Tutorial: Experimental nonlinear dynamical circuit analysis of a ferromagnetic inductor," *IEEE Circuits Syst. Mag.*, vol. 18, no. 2, pp. 28–34, Secondquarter 2018.
- [3] H. K. Khalil, *Nonlinear Control*. Harlow, U.K.: Pearson, 2015.
- [4] M. F. Mehrjardi, H. Sanusi, M. A. M. Ali, and M. Abdullah, "Integrated attitude-orbit dynamics and control of spacecraft systems: State of the art and future trends," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 33, no. 7, pp. 60–71, Jul. 2018.
- [5] S. A. Sajadi-Alamdari, H. Voos, and M. Darouach, "Ecological advanced driver assistance system for optimal energy management in electric vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 12, no. 4, pp. 92–109, Winter 2020.
- [6] S. Mahmodcherati, M. Elbuluk, and Y. Sozer, "Reducing ripple in wind power systems: A hybrid method formed using two power controllers," *IEEE Ind. Appl. Mag.*, vol. 25, no. 2, pp. 23–35, Mar./Apr. 2019.
- [7] N. Karlsson, "Feedback control in programmatic advertising: The frontier of optimization in real-time bidding," *IEEE Control Syst. Mag.*, vol. 40, no. 5, pp. 40–77, Oct. 2020.
- [8] A. Isidori, *Nonlinear Control Systems* (Series Communication and Control Engineering), 3rd ed. London, U.K.: Springer, 1995.
- [9] H. K. Khalil, *Nonlinear Systems*, 3rd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2002.
- [10] D. J. Hill and T. Liu, "Dissipativity, stability, and connections: Progress in complexity," *IEEE Control Syst. Mag.*, vol. 42, no. 2, pp. 88–106, Apr. 2022.
- [11] K. Batselier, "Low-rank tensor decompositions for nonlinear system identification: A tutorial with examples," *IEEE Control Syst. Mag.*, vol. 42, no. 1, pp. 54–74, Feb. 2022.
- [12] S. Koshkin and V. Jovanovic, "Swinging a playground swing: Torque controls for inducing sustained oscillations," *IEEE Control Syst. Mag.*, vol. 42, no. 2, pp. 18–34, Apr. 2022.
- [13] D. Simon, *Optimal State Estimation: Kalman, H_∞ , and Nonlinear Approaches*. Hoboken, NJ, USA: Wiley, 2006.
- [14] F. Sawo, *Nonlinear State and Parameter Estimation of Spatially Distributed Systems* (Karlsruhe Series on Intelligent Sensor-Actuator-System), vol. 5. Karlsruhe, Germany: Universitätsverlag Karlsruhe, 2009.
- [15] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation* (Prentice Hall Information and System Sciences Series). Upper Saddle River, NJ, USA: Prentice-Hall, 2000.
- [16] B. D. O. Anderson and J. B. Moore, *Optimal Filtering* (Prentice Hall Information and System Sciences Series). Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.
- [17] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Statist. Comput.*, vol. 10, no. 3, pp. 197–208, Jul. 2000.
- [18] P. M. Djurić et al., "Particle filtering," *IEEE Signal Process. Mag.*, vol. 20, no. 5, pp. 19–38, Sep. 2003.
- [19] S. Godsill, "Particle filtering: The first 25 years and beyond," in *Proc. 44th IEEE Int. Conf. Acoust., Speech Signal Process.* Brighton, U.K.: IEEE, May 12–17, 2019, pp. 7760–7764.
- [20] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Míguez, and P. M. Djurić, "Adaptive importance sampling: The past, the present, and the future," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 60–79, Jul. 2017.
- [21] V. Elvira and L. Martino, *Advances in Importance Sampling*. Hoboken, NJ, USA: Wiley, Feb. 17, 2021. [Online]. Available: <https://doi.org/10.1002/9781118445112.stat08284>
- [22] N. Branchini and V. Elvira, "Optimized auxiliary particle filters: Adapting mixture proposals via convex optimization," in *Proc. 37th Conf. Uncertainty Artif. Intell.*, vol. 161. Toronto, Canada: PMLR, May 27–29, 2021.
- [23] V. Elvira, J. Míguez, and P. M. Djurić, "Adapting the number of particles in sequential Monte Carlo methods through an online scheme

- for convergence assessment,” *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1781–1794, Apr. 2017.
- [24] V. Elvira, L. Martino, and C. P. Robert, “Rethinking the effective sample size,” *Int. Statist. Rev.*, vol. 90, no. 3, pp. 525–550, Dec. 2022.
- [25] V. Elvira, L. Martino, M. F. Bugallo, and P. Djurić, “In search for improved auxiliary particle filters,” in *Proc. 26th Eur. Signal Process. Conf. Rome, Italy: IEEE*, Sep. 3–7, 2018, pp. 1637–1641.
- [26] V. Elvira, L. Martino, M. F. Bugallo, and P. Djurić, “Elucidating the auxiliary particle filter via multiple importance sampling,” *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 145–152, Nov. 2019.
- [27] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (Series Adaptive Computation Machine Learning). Cambridge, MA, USA: MIT Press, 2012.
- [28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (Series Adaptive Computation Machine Learning). Cambridge, MA, USA: MIT Press, 2016.
- [29] R. Jonschkowski, D. Rastogi, and O. Brock, “Differentiable particle filters: End-to-end learning with algorithmic priors,” in *Proc. Robot. Sci., Syst.* Pittsburgh, PA, USA: RSS Foundation, Jun. 26–30, 2018, pp. 1–9.
- [30] P. Karkus, D. Hsu, and W. S. Lee, “Particle filter networks with application to visual localization,” in *Proc. Conf. Robot Learn. (CORL)*, 2018, pp. 169–178.
- [31] A. Corenflos, J. Thornton, G. Deligiannidis, and A. Doucet, “Differentiable particle filtering via entropy-regularized optimal transport,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2100–2111.
- [32] X. Chen, H. Wen, and Y. Li, “Differentiable particle filters through conditional normalizing flow,” in *Proc. IEEE Int. Conf. Inf. Fusion (FUSION)*. Piscataway, NJ, USA: IEEE, 2021, pp. 1–6.
- [33] X. Ma, P. Karkus, D. Hsu, and W. S. Lee, “Particle filter recurrent neural networks,” in *Proc. 34th AAAI Conf. Artif. Intell.*, vol. 34. New York, NY, USA: Assoc. Advancement Artif. Intell., Feb. 7–12, 2020, pp. 5101–5108.
- [34] J. Geweke, “Bayesian inference in econometrics models using Monte Carlo integration,” *Econometrica*, vol. 57, no. 6, pp. 1317–1339, Nov. 1989.
- [35] A. Kong, J. S. Liu, and W. H. Wong, “Sequential imputations and Bayesian missing data problems,” *J. Amer. Statist. Assoc.*, vol. 89, no. 425, pp. 278–288, Mar. 1994.
- [36] R. Chen and J. S. Liu, “Predictive updating methods with application to Bayesian classification,” *J. Roy. Statist. Soc. B*, vol. 58, no. 2, pp. 397–415, 1996.
- [37] J. S. Liu, “Metropolized independent sampling with comparison to rejection sampling and importance sampling,” *Statist. Comput.*, vol. 6, pp. 113–119, Jun. 1996.
- [38] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for hyperparameter optimization,” in *Proc. 25th Conf. Neural Inform. Process. Syst.* Granada, Spain: Neural Inform. Process. Syst. Found., Dec. 12–17, 2011, pp. 2546–2554.
- [39] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, Mar. 2021.
- [40] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural Netw.*, vol. 2, no. 5, pp. 359–366, 1989.
- [41] A. R. Barron, “Approximation and estimation bounds for artificial neural networks,” *Mach. Learn.*, vol. 14, pp. 115–133, Jan. 1994.
- [42] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks* (Series Studies in Computational Intelligence), vol. 385. Heidelberg, Germany: Springer, 2012.
- [43] Y. Bengio, P. Smard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
- [44] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Proc. 30th Int. Conf. Mach. Learn.*, vol. 28. Atlanta, GA, USA: PMLR, Jun. 16–21, 2013, pp. 1310–1318.
- [45] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 15, 1997.
- [46] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” in *Proc. 9th IEEE Int. Conf. Artif. Neural Netw.*, vol. 2. Edinburgh, U.K.: IEEE, Sep. 7–10, 1999, pp. 850–855.
- [47] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Gated feedback recurrent neural networks,” in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. Lille, France: PMLR, Jul. 6–11, 2015, pp. 2067–2075.
- [48] R. Jozefowicz, W. Zaremba, and I. Sutskever, “An empirical evaluation of recurrent network architectures,” in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37. Lille, France: PMLR, Jul. 6–11, 2015, pp. 2342–2350.
- [49] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, “Convolutional neural network architectures for signals supported on graphs,” *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1034–1049, Feb. 2019.
- [50] F. Gama, E. Isufi, G. Leus, and A. Ribeiro, “Graphs, convolutions, and neural networks: From graph filters to graph neural networks,” *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 128–138, Nov. 2020.
- [51] L. Ruiz, F. Gama, and A. Ribeiro, “Graph neural networks: Architectures, stability and transferability,” *Proc. IEEE*, vol. 109, no. 5, pp. 660–682, May 2021.
- [52] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs,” *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1644–1656, Apr. 2013.
- [53] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
- [54] A. Sandryhaila and J. M. F. Moura, “Discrete signal processing on graphs: Frequency analysis,” *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3042–3054, Jun. 2014.
- [55] F. Gama and S. Sojoudi, “Distributed linear-quadratic control with graph neural networks,” *Signal Process.*, vol. 196, Jul. 2022, Art. no. 108506.
- [56] D. Owerko, F. Gama, and A. Ribeiro, “Unsupervised optimal power flow using graph neural networks,” Oct. 17, 2022. [Online]. Available: <http://arxiv.org/abs/2210.09277>
- [57] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, “Unfolding WMMSE using graph neural networks for efficient power allocation,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6004–6017, Sep. 2021.
- [58] J. D. Medaglia et al., “Brain network efficiency is influenced by the pathologic source of corticobasal syndrome,” *Neurology*, vol. 89, no. 13, pp. 1373–1381, 2017.
- [59] F. Gama, Q. Li, E. Tolstaya, A. Prorok, and A. Ribeiro, “Synthesizing decentralized controllers with graph neural networks and imitation learning,” *IEEE Trans. Signal Process.*, vol. 70, pp. 1932–1946, 2022.
- [60] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: Going beyond Euclidean data,” *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.
- [61] A. Heimowitz and Y. C. Eldar, “A unified view of diffusion maps and signal processing on graphs,” in *Proc. Int. Conf. Sampling Theory Appl.*, Tallin, Estonia, Jul. 3–7, 2017, pp. 308–312.
- [62] S. Segarra, A. G. Marques, and A. Ribeiro, “Optimal graph-filter design and applications to distributed linear networks operators,” *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4117–4131, Aug. 1, 2017.
- [63] F. Gama, J. Bruna, and A. Ribeiro, “Stability properties of graph neural networks,” *IEEE Trans. Signal Process.*, vol. 68, pp. 5680–5695, 2020.
- [64] T. M. Roddenberry, F. Gama, R. G. Baraniuk, and S. Segarra, “On local distributions in graph signal processing,” *IEEE Trans. Signal Process.*, vol. 70, pp. 5564–5577, 2022.
- [65] E. Isufi, F. Gama, and A. Ribeiro, “EdgeNets: Edge varying graph neural networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 3862–3877, Nov. 2022.
- [66] L. Ruiz, F. Gama, A. G. Marques, and A. Ribeiro, “Invariance-preserving localized activation functions for graph neural networks,” *IEEE Trans. Signal Process.*, vol. 68, pp. 127–141, 2019.
- [67] L. Ruiz, F. Gama, and A. Ribeiro, “Gated graph recurrent neural networks,” *IEEE Trans. Signal Process.*, vol. 68, pp. 6303–6318, 2020.
- [68] R. Durrett, *Probability: Theory and Examples* (Cambridge Series in Statistical and Probabilistic Mathematics), 4th ed. New York, NY, USA: Cambridge Univ. Press, 2010.
- [69] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.
- [70] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, ser. Signal Process. Ser. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
- [71] I. Kobayez, S. J. D. Prince, and M. A. Brubaker, “Normalizing flows: An introduction and review of current methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3964–3979, Nov. 2020.
- [72] G. Papamakarios, E. Nalisnick, D. Jimenez Rezende, S. Mohamed, and B. Lakshminarayanan, “Normalizing flows for probabilistic modeling and inference,” *J. Mach. Learn. Res.*, vol. 22, pp. 1–64, Dec. 2021.

- [73] X. Wei, H. van Gorp, L. Gonzalez-Carabarin, D. Freedman, Y. C. Eldar, and R. J. G. van Sloun, "Deep unfolding with normalizing flow priors for inverse problems," *IEEE Trans. Signal Process.*, vol. 70, pp. 2962–2971, 2022.
- [74] J. Ko and D. Fox, "GP-BayesFilters: Bayesian filtering using Gaussian process prediction and observation models," *Auton. Robots*, vol. 27, pp. 75–90, Oct. 2009.
- [75] W. Aftab and L. Mihaylova, "A learning Gaussian process approach for maneuvering target tracking and smoothing," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 1, pp. 278–292, Feb. 2021.
- [76] T. Li, H. Chen, S. Sun, and J. M. Corchado, "Joint smoothing and tracking based on continuous-time target trajectory function fitting," *IEEE Trans. Autom. Sci. Eng.*, vol. 16, no. 3, pp. 1476–1483, Jul. 2019.
- [77] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 61, pp. 2121–2159, Jul. 2011.
- [78] D. P. Kingma and J. L. Ba, "ADAM: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015, pp. 1–15.
- [79] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, Banff, AB, Canada, Apr. 14–16, 2014, pp. 1–14.
- [80] F. Gama, E. Isufi, A. Ribeiro, and G. Leus, "Controllability of bandlimited graph processes over random time varying graphs," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6440–6454, Dec. 15, 2019.
- [81] R. Beckley, C. Weatherspoon, M. Alexander, M. Chandler, A. Johnson, and G. S. Bhatt, "Modeling epidemics with differential equations," 2013. Tennessee State University Internal Report. [Online]. Available: <https://www.tnstate.edu/mathema/cs/mathreu/filesreu/GroupProjectSIR.pdf>
- [82] H. El-Samad, S. Prajna, A. Papachristodoulou, J. Doyle, and M. Khammash, "Advanced methods and algorithms for biological networks analysis," *Proc. IEEE*, vol. 94, no. 4, pp. 832–853, Apr. 2006.



Fernando Gama (Member, IEEE) received the M.A. degree in statistics from Wharton School, in 2017, and the Ph.D. degree in electrical and systems engineering from the University of Pennsylvania, in 2020. He is a Machine Learning Scientist with Pendulum Systems, Inc. He had a Postdoctoral stints with the Electrical Engineering and Computer Sciences Department at the University of California, Berkeley, CA, USA, from 2020 to 2021, and the Electrical and Computer Engineering Department at Rice University, from 2021 to 2022. He has also

been a Research Intern with Facebook Artificial Intelligence Research at Montréal, in 2018, and a Visiting Scholar with TU Delft at The Netherlands, in 2017. Before joining Pendulum, he worked as a Machine Learning Researcher with Morgan Stanley, from 2022 to 2023. The core of his published research is related to machine learning and signal processing for graph-structured data, with applications in power grids, sensor networks, multiagent robotics, and distributed control. He has a background in statistics, time series analysis, and signal processing.



Nicolas Zilberstein (Graduate Student Member, IEEE) received the B.Sc. (Hons.) degree in electrical engineering from the Universidad de Buenos Aires, Argentina, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. His research interests are in the field of inverse problem and Bayesian methods.



in 2021 and receiving the "Ing. Isidoro Marín" award from the National Engineering Association at Argentina in 2019.



Richard G. Baraniuk (Fellow, IEEE) received the B.S. degree in electrical engineering from the University of Manitoba, Winnipeg, MB, Canada, in 1987, the M.S. degree from the University of Wisconsin–Madison, Madison, WI, USA, in 1988, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992. He is currently the C. Sidney Burrus Professor in electrical and computer engineering with Rice University, and the Founding Director of OpenStax (openstax.org). His research interests lie in new

theory, algorithms, and hardware for sensing, signal processing, and machine learning. He is a member of the U.S. National Academy of Engineering and American Academy of Arts and Sciences and a fellow of the National Academy of Inventors, American Association for the Advancement of Science. He received the DOD Vannevar Bush Faculty Fellow Award, the Harold W. McGraw Jr. Prize in Education, the IEEE James H. Mulligan Jr. Education Medal, and the IEEE Signal Processing Society, Technical Achievement, Education, Best Paper, Best Magazine Paper, and Best Column Awards. He holds 45 U.S. and eight foreign patents.



Santiago Segarra (Senior Member, IEEE) received the B.Sc. degree in industrial engineering with highest honors (Valedictorian) from the Instituto Tecnológico de Buenos Aires (ITBA), Argentina, in 2011, the M.Sc. in electrical engineering from the University of Pennsylvania (Penn), Philadelphia, PA, USA, in 2014, and the Ph.D. degree in electrical and systems engineering from Penn, in 2016. From September 2016 to June 2018, he was a Postdoctoral Research Associate with the Institute for Data, Systems, and Society at the Massachusetts Institute of Technology. Since July 2018, he has been a W. M. Rice Trustee Assistant Professor with the Department of Electrical and Computer Engineering at Rice University. His research interests include network theory, data analysis, machine learning, and graph signal processing. He received the ITBA's 2011 Best Undergraduate Thesis Award in Industrial Engineering, the 2011 Outstanding Graduate Award granted by the National Academy of Engineering of Argentina, the 2017 Penn's Joseph and Rosaline Wolf Award for Best Doctoral Dissertation in Electrical and Systems Engineering, the 2020 IEEE Signal Processing Society Young Author Best Paper Award, the 2021 Rice's School of Engineering Research + Teaching Excellence Award, and five best conference paper awards.