



AN OVERVIEW OF DIFFERENTIABLE PARTICLE FILTERS FOR DATA-ADAPTIVE SEQUENTIAL BAYESIAN INFERENCE

XIONGJIE CHEN^{✉1} AND YUNPENG LI^{✉1}

¹University of Surrey, United Kingdom

ABSTRACT. By approximating posterior distributions with weighted samples, particle filters (PFs) provide an efficient mechanism for solving non-linear sequential state estimation problems. While the effectiveness of particle filters has been recognised in various applications, their performance relies on the knowledge of dynamic models and measurement models, as well as the construction of effective proposal distributions. An emerging trend involves constructing components of particle filters using neural networks and optimising them by gradient descent, and such data-adaptive particle filtering approaches are often called differentiable particle filters. Due to the expressiveness of neural networks, differentiable particle filters are a promising computational tool for performing inference on sequential data in complex, high-dimensional tasks, such as vision-based robot localisation. In this paper, we review recent advances in differentiable particle filters and their applications. We place special emphasis on different design choices for key components of differentiable particle filters, including dynamic models, measurement models, proposal distributions, optimisation objectives, and differentiable resampling techniques.

1. Introduction. In many signal processing and control problems, we are often interested in inferring posterior distributions of latent variables. These variables cannot be directly observed but manifest themselves through an observation process conditioned on the unobserved variables. The posterior distribution of latent variables is defined by two system elements, i.e. a prior distribution of the unobserved variables and a likelihood function that describes the relationship between latent variables and observations. When dealing with sequential data for both latent variables and observations, this task is known as the Bayesian filtering problem, where one recursively updates the posterior distribution of latent variables. In simple systems, such as linear-Gaussian state-space models, the evolving posterior distribution can be exactly derived with analytical expressions using Kalman filters [1]. However, real-world problems often involve non-linear and non-Gaussian systems, where posterior distributions are not analytically available due to the computation of intractable high-dimensional integrals. As a result, many approximation-based approaches have been proposed, including the extended Kalman filter [2, 3], the unscented Kalman filter [4], and grid-based filters [5]. Despite these developments, they often perform poorly in high-dimensional, highly non-linear cases [6].

2020 *Mathematics Subject Classification.* Primary: 62M20, 62M45, 62M05.

Key words and phrases. Sequential Monte Carlo, differentiable particle filters, parameter estimation, machine learning.

*Corresponding author: Xiongjie Chen.

One popular alternative solution to the non-linear and non-Gaussian filtering problem is particle filters, also known as sequential Monte Carlo (SMC) methods [7, 8, 9]. In particle filters, the target posterior distributions are recursively approximated by a sequence of empirical distributions associated with a set of particles, i.e. weighted samples. Particle filters do not assume linearity or Gaussianity on the considered state-space model, and convergence results on particle filtering methods have been established [10, 11, 12, 13, 14]. Due to these favourable properties, a variety of particle filtering methods have been developed following the seminal work on the bootstrap particle filter (BPF) introduced in [7]. Examples include auxiliary particle filters (APFs) [15, 16, 17] to guide particles into regions where the posterior distribution has higher densities. The Gaussian sum particle filters (GSPFs) [18, 19, 20] approximate posterior distributions with Gaussian mixture models by building banks of Gaussian particle filters. The unscented particle filters (UPFs) [21, 22, 23] incorporate the latest observations in the construction of proposal distributions generated by unscented Kalman filters (UKFs). Several other variants of particle filters, including Rao-Blackwellised particle filters [24, 25], multiple particle filters [26, 27], particle flow particle filters [28, 29, 30], have been designed for more efficient sampling in high-dimensional spaces. Although these variants of particle filters have shown to be effective in many applications [31, 32, 33, 34, 35], practitioners often have to design hand-crafted transition kernels of latent states and likelihood functions, which is often time-consuming and relies on extensive domain knowledge.

To address this challenge, various techniques have been proposed for parameter estimation in state-space models and particle filters [36, 37]. The particle Markov chain Monte Carlo (PMCMC) method and its variants provide a mechanism to perform parameter estimation for sequential Monte Carlo methods using Markov chain Monte Carlo (MCMC) sampling [38, 39]. The SMC² algorithm [40] is a particle equivalent of the PMCMC method, which simultaneously tracks the posterior of model parameters and latent variables using two layers of particle filters. Based on a similar nested filtering framework as in the SMC² algorithm, a series of nested hybrid filters were proposed within a purely recursive structure and thus are more suitable for jointly inferring both static parameters of the system and latent posteriors in on-line settings [41, 42, 43]. In [44, 45, 46], the transition matrix and the covariance matrix in linear-Gaussian state-space models were interpreted as a directed graph, and a GraphEM method based on the expectation-maximisation (EM) algorithm was proposed to learn the directed graph.

An emerging trend in this active research area is to construct and learn components of particle filters using machine learning techniques, such as neural networks and gradient descent. Such particle filters are often named differentiable particle filters (DPFs) [47, 48, 49, 50, 51, 52]. In particular, differentiable particle filters construct their dynamic models, measurement models, and proposal distributions using expressive neural networks that can model complex high-dimensional non-linear non-Gaussian systems. One main obstacle to the development of fully-differentiable particle filters is the resampling step which is known to be non-differentiable in classical resampling methods. To this end, different resampling strategies have been proposed to address this issue, e.g. no resampling [47], soft resampling [48], entropy-regularised optimal transport resampling [52], and transformer-based resampling [53]. Another key ingredient to developing data-adaptive particle filters is

a loss function, which is minimised using gradient descent to optimise parameters of neural networks used in the construction of differentiable particle filters.

Despite the surging interest in differentiable particle filters, as far as we know, there is an absence of a comprehensive review of recent advances in this field. Therefore, the main objective of this paper is to provide an overview of the design of differentiable particle filters. We place special emphasis on different design choices of key ingredients in differentiable particle filters, and aim to answer the following questions:

- How to model the transition of latent states in state-space models with neural networks?
- How to estimate the likelihood of observations given states using neural networks?
- How can we construct proposal distributions that lead to better approximations of the true posterior in complex environments?
- What loss functions should we use when training differentiable particle filters?
- How to perform resampling for differentiable particle filters?

The remainder of this paper is organised as follows. In Section 2, we provide the necessary background knowledge for introducing differentiable particle filters. In Section 3, we detail different design choices of sampling distributions of latent states in differentiable particle filters, i.e. their dynamic models and proposal distributions. Section 4 discusses the construction of differentiable particle filters’ measurement models. In Section 5, we list examples of resampling schemes used in differentiable particle filters. Loss functions used for optimising differentiable particle filters are introduced in Section 6. We provide implementation details of some differentiable particle filters in Section 7. We conclude the paper in Section 8.

2. Preliminaries.

2.1. State-Space Models. In this section, we detail the problem we consider in this paper, including a brief introduction to state-space models and the general goal of Bayesian filtering approaches.

State-space models (SSMs) refer to a class of sequential models that consist of two discrete-time random sequences, the latent state sequence $(x_t)_{t \geq 0}$ defined on $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, and the observed measurement sequence $(y_t)_{t \geq 1}$ defined on $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$. The latent state sequence $(x_t)_{t \geq 0}$ is characterised by a Markov process with an initial distribution $\pi(x_0)$ and a transition kernel $p(x_t|x_{t-1})$ for $t \geq 1$. The observation y_t is conditionally independent given the current latent state x_t . In this review, we focus on parameterised state-space models defined as follows:

$$x_0 \sim \pi(x_0; \theta), \quad (1)$$

$$x_t|x_{t-1} \sim p(x_t|x_{t-1}; \theta) \text{ for } t \geq 1, \quad (2)$$

$$y_t|x_t \sim p(y_t|x_t; \theta) \text{ for } t \geq 1, \quad (3)$$

where $\theta \in \mathbb{R}^{d_\theta}$ is the parameter set of interest, $\pi(x_0; \theta)$ is the initial distribution of the latent state, $p(x_t|x_{t-1}; \theta)$ is the dynamic model that describes the transition of the latent state, and $p(y_t|x_t; \theta)$ is the measurement model estimating the conditional likelihood of observation y_t given x_t . Denote by $x_{0:t} := \{x_0, \dots, x_t\}$ and $y_{1:t} := \{y_1, \dots, y_t\}$ respectively the sequence of latent states and the sequence of observations up to time step t , the parameterised state-space model defined by

Eqs. (1) - (3) can also be defined through the following joint distribution of $x_{0:t}$ and $y_{1:t}$:

$$p(x_{0:t}, y_{1:t}; \theta) = \pi(x_0; \theta) \prod_{k=1}^t p(x_k | x_{k-1}; \theta) p(y_k | x_k; \theta), \quad (4)$$

with $p(x_0, y_{1:0}; \theta) := \pi(x_0; \theta)$.

In Bayesian filtering problems, the goal is to infer recursively in time the joint posterior distribution $p(x_{0:t}|y_{1:t}; \theta)$, the marginal posterior distribution $p(x_t|y_{1:t}; \theta)$, or the expectation $\mathbb{E}_{p(x_{0:t}|y_{1:t}; \theta)}[\psi_t(x_{0:t})]$ of a function $\psi_t(\cdot) : \mathcal{X}^{t+1} \rightarrow \mathbb{R}^{d_{\psi_t}}$ w.r.t. the joint posterior distribution or the log evidence $L_t(\theta) := \log p(y_{1:t}; \theta)$.

We use the joint posterior and the log evidence as two examples to show how they can be updated recursively, other quantities of interest can be derived similarly. At the t -th time step, the joint posterior distribution $p(x_{0:t}|y_{1:t}; \theta)$ can be written as follows:

$$p(x_{0:t}|y_{1:t}; \theta) = \frac{p(y_{1:t}|x_{0:t}; \theta) p(x_{0:t}; \theta)}{p(y_{1:t}; \theta)} \quad (5)$$

$$= \frac{p(y_{1:t}|x_{0:t}; \theta) p(x_{0:t}; \theta)}{\int_{\mathcal{X}^{t+1}} p(x_{0:t}, y_{1:t}; \theta) dx_{0:t}}, \quad (6)$$

with $p(x_0|y_{1:0}; \theta) := \pi(x_0; \theta)$. In addition to Eq. (5), which directly computes the joint posterior by applying Bayes' theorem, we can also obtain a recursive formula for the joint posterior $p(x_{0:t}|y_{1:t}; \theta)$:

$$p(x_0|y_{1:0}; \theta) := \pi(x_0; \theta), \quad (7)$$

$$p(x_{0:t}|y_{1:t-1}; \theta) = p(x_{0:t-1}|y_{1:t-1}; \theta) p(x_t|x_{t-1}; \theta), \quad t \geq 1, \quad (8)$$

$$p(x_{0:t}|y_{1:t}; \theta) = \frac{p(y_t|x_t; \theta) p(x_{0:t}|y_{1:t-1}; \theta)}{p(y_t|y_{1:t-1}; \theta)} \quad (9)$$

$$= p(x_{0:t-1}|y_{1:t-1}; \theta) \frac{p(y_t|x_t; \theta) p(x_t|x_{t-1}; \theta)}{p(y_t|y_{1:t-1}; \theta)}, \quad t \geq 1, \quad (10)$$

with $p(y_1|y_{1:0}; \theta) := p(y_1; \theta)$. The log evidence $L_t(\theta) := \log p(y_{1:t}; \theta)$ also satisfies the following recursion:

$$L_t(\theta) = \sum_{k=1}^t l_t(\theta) = L_{t-1}(\theta) + l_t(\theta), \quad (11)$$

$$l_t(\theta) := \log p(y_t|y_{1:t-1}; \theta) = \log \int_{\mathcal{X}^{t+1}} p(y_t|x_t; \theta) p(x_{0:t}|y_{1:t-1}; \theta) dx_{0:t}. \quad (12)$$

However, except for only a limited class of state-space models such as linear-Gaussian models, obtaining $p(x_{0:t}|y_{1:t}; \theta)$ and $\log p(y_{1:t}; \theta)$ from $p(x_{0:t-1}|y_{1:t-1}; \theta)$ and $\log p(y_{1:t-1}; \theta)$ using Eqs. (7) – (12) are usually analytically intractable since they involve the computation of complex high-dimensional integrals over \mathcal{X}^{t+1} . As an alternative, particle filters resort to Monte Carlo integration techniques to approximate the posterior distributions and other quantities of interest. We present more details about particle filters in the next section.

2.2. Particle Filters. Particle filters (PFs), also known as sequential Monte Carlo (SMC) methods, are a family of Monte Carlo algorithms designed for approximating the latent posterior distributions sequentially in state-space models [6]. In particular, particle filters approximate the analytically intractable joint posterior $p(x_{0:t}|y_{1:t}; \theta)$ with an empirical distribution consisting of a set of weighted samples $\{W_t^i, x_{0:t}^i\}_{i \in [N_p]}$:

$$p(x_{0:t}|y_{1:t}; \theta) \approx \sum_{i=1}^{N_p} W_t^i \delta_{x_{0:t}^i}(\cdot), \quad (13)$$

where $[N_p] := \{1, \dots, N_p\}$, N_p is the number of particles, $\delta_{x_{0:t}^i}(\cdot)$ denotes the Dirac delta function located in $x_{0:t}^i$, and $W_t^i \geq 0$ with $\sum_{i=1}^{N_p} W_t^i = 1$ is the normalised importance weight of the i -th particle at the t -th time step. By representing the intractable posterior distributions of latent variables with sets of weighted samples, particle filters can provide efficient and asymptotically unbiased estimates of the posterior distributions, i.e. they are able to approximate the posterior distributions arbitrarily well under weak assumptions when the number of particles $N_p \rightarrow +\infty$ [10, 11, 12, 13, 14].

As an importance sampling-based algorithm [54, 55, 56, 57], the performance of particle filters highly relies on the distribution where particles are sampled from, which is often called the proposal distribution. An ideal proposal distribution should allow efficient sampling of particles and induce tractable importance weights. In bootstrap particle filters (BPFs), the proposal distribution is simply the system dynamic $p(x_t|x_{t-1}; \theta)$ [7]. Denote by w_t^i unnormalised importance weights of particles, in bootstrap particle filters, particle weights are updated according to the likelihood of observation y_t given state x_t^i :

$$w_t^i = w_{t-1}^i p(y_t|x_t^i; \theta), \quad (14)$$

with $w_0^i = 1$. An obvious drawback of bootstrap particle filters is that sampling from the system dynamic $p(x_t|x_{t-1}; \theta)$ often results in trivially small weights for the majority of particles, especially when dealing with long sequences. This is known as the weight degeneracy problem, where the Monte Carlo approximation of the posterior is dominated by only a few particles with large weights [58]. Weight degeneracy leads to poor estimates of posteriors, and most of the computational cost will be wasted on particles with negligible weights. Therefore, to obtain better approximations of the posterior, information from the latest observations is usually utilised in the construction of proposal distributions, which gives rise to the guided particle filters [59]. Compared to bootstrap particle filters, by selecting an appropriate proposal distribution, guided particle filters can propose particles located in regions where the posterior has higher densities and thus yield more uniformly distributed particle weights. Denote the proposal distributions by $q(x_t|y_t, x_{t-1}; \phi)$ for $t \geq 1$, in guided particle filters, the unnormalised importance weight of the i -th particle is updated recursively through:

$$w_t^i = w_{t-1}^i \frac{p(y_t|x_t^i; \theta)p(x_t^i|x_{t-1}^i; \theta)}{q(x_t^i|y_t, x_{t-1}^i; \phi)}, \quad (15)$$

with $w_0^i = 1$.

However, even with proposal distributions constructed with information from observations, the weight degeneracy problem is still a challenging issue since the dimension of the joint posterior $p(x_{0:t}|y_{1:t}; \theta)$ increases over time. Therefore, particle

resampling [60], a critical step of particle filters, was introduced to further mitigate the weight degeneracy problem. It has been proved in the literature, both empirically and theoretically, that resampling steps play a crucial role in avoiding the weight degeneracy problem and stabilising the Monte Carlo error over time [61, 62].

There are different choices of resampling schemes, including multinomial resampling, residual resampling, stratified resampling, and systematic resampling [63, 64]. Resampling is often triggered when a pre-defined condition is met. One commonly used threshold metric for resampling is the effective sample size (ESS). The effective sample size in particle filters can be interpreted as the number of particles needed from the true posterior to produce the same variance in its estimates as that of the N_p samples drawn from the proposal distribution [65]. Since the actual value of the ESS is intractable in general, an approximation of the ESS is often used instead:

$$\text{ESS}_t \left(\{W_t^i\}_{i \in [N_p]} \right) = \frac{1}{\sum_{i=1}^{N_p} (W_t^i)^2}. \quad (16)$$

When using the effective sample size as the threshold metric, resampling steps are triggered only when $\text{ESS}_t \leq \text{ESS}_{\min}$, where ESS_{\min} refers to a predefined threshold for effective sample sizes. While the performance of particle filters is not sensitive to the value of ESS_{\min} , one common choice is to set $\text{ESS}_{\min} = \frac{N_p}{2}$. Notably, it has been shown that the estimator of the ESS defined by Eq. (16) provides an accurate estimation of the true ESS only in specific cases and may fail to detect certain issues with the particle approximation [66, 67]. As a result, alternatives to Eq. (16) have recently been proposed [66, 67, 68, 69].

A pseudocode for particle filtering approaches is provided in Algorithm 1. Note that the presented pseudocode is a standard implementation of particle filters. Some variants of particle filtering approaches may not fit in the framework, e.g. auxiliary particle filters [15, 16, 17] and Rao-Blackwellised particle filters [24, 25]. However, we do not extend it to more generic particle filtering algorithms because most existing differentiable particle filtering frameworks follow the structure of Algorithm 1.

2.3. Differentiable Particle Filters. While parameter estimation techniques for particle filters have long been an active research area [32, 70, 71], many existing approaches are restricted to scenarios where a subset of model parameters is known or the model has to follow certain structures [37]. Recently, an emerging trend in this direction is the so-called differentiable particle filters (DPFs) [47, 48, 49, 52, 53, 72, 73]. The main idea of differentiable particle filters is to develop data-adaptive particle filters by constructing particle filters' components through neural networks, and optimise them by minimising a loss function through gradient descent. With the expressiveness of neural networks, differentiable particle filters have shown their effectiveness in complicated applications, e.g. vision-based robot localisation, where the observation can be high-dimensional unstructured data like images [47, 48, 49, 52]. We provide in the remainder of this section an introduction of key components in differentiable particle filters, before we delve into more detailed discussions in the following sections.

In differentiable particle filters, given a particle x_{t-1}^i at the $(t-1)$ -th time step, we can generate a new particle x_t^i for the t -th time step as follows [47]:

$$x_t^i = g_\theta(x_{t-1}^i, \alpha_t^i) \sim p(x_t | x_{t-1}^i; \theta), \quad (17)$$

where $\alpha_t^i \in \mathbb{R}^{d_\alpha}$ simulates process noise, $g_\theta(\cdot) : \mathcal{X} \times \mathbb{R}^{d_\alpha} \rightarrow \mathcal{X}$ is a function that is differentiable w.r.t. both x_{t-1}^i and α_t^i . Regarding the measurement model, several

Algorithm 1: Pseudocode for standard particle filtering algorithms.

```

1: Require: A state-space model, resampling threshold  $\text{ESS}_{\min}$ , particle number
    $N_p$ , proposal distributions  $q(x_t|y_t, x_{t-1}; \phi)$ , a resampling scheme
    $\mathcal{R}(\{W_t^i\}_{i \in [N_p]})$  that outputs indices  $A_t^i$  of selected particles;
2: Initialisation: Sample  $x_0^i \stackrel{\text{i.i.d.}}{\sim} \pi(x_0; \theta)$  for  $\forall i \in [N_p]$ ;
3: Set weights  $w_0^i = 1$  for  $\forall i \in [N_p]$ ;
4: Set normalise weights  $W_0^i = \frac{1}{N_p}$  for  $\forall i \in [N_p]$ ;
5: Set  $L_0(\theta) = 1$ ;
6: Set the effective sample size  $\text{ESS}_0(\{W_0^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_0^i)^2} = N_p$ ;
7: for  $t = 1$  to  $T$  do
8:   if  $\text{ESS}_t(\{W_{t-1}^i\}_{i \in [N_p]}) < \text{ESS}_{\min}$  then
9:      $A_t^{1:N_p} \leftarrow \mathcal{R}(\{W_{t-1}^i\}_{i \in [N_p]})$ ;
10:     $\tilde{w}_{t-1}^i \leftarrow 1$  for  $\forall i \in [N_p]$ ;
11:   else
12:      $A_t^i \leftarrow i$  for  $\forall i \in [N_p]$ ;
13:      $\tilde{w}_{t-1}^i \leftarrow w_{t-1}^i$  for  $\forall i \in [N_p]$ ;
14:   end if
15:   Sample  $x_t^i \stackrel{\text{i.i.d.}}{\sim} q(x_t|y_t, x_{t-1}^{A_t^i}; \phi)$  for  $\forall i \in [N_p]$ ;
16:    $w_t^i = \tilde{w}_{t-1}^i \frac{p(y_t|x_t^i; \theta)p(x_t^i|x_{t-1}^{A_t^i}; \theta)}{q(x_t^i|y_t, x_{t-1}^{A_t^i}; \phi)}$ ;
17:   (optional) Estimate the log-likelihood of observations  $L_t(\theta)$ :
18:      $l_t(\theta) = \log \frac{\sum_{i=1}^{N_p} w_t^i}{\sum_{i=1}^{N_p} \tilde{w}_{t-1}^i}, L_t(\theta) = L_{t-1}(\theta) + l_t(\theta)$ ;
19:   Normalise weights  $W_t^i = w_t^i / \sum_{n=1}^{N_p} w_t^n$  for  $\forall i \in [N_p]$ ;
20:   Compute the effective sample size:  $\text{ESS}_t(\{W_t^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_t^i)^2}$ ;
end for

```

differentiable particle filters estimate the conditional likelihood of an observation y_t given x_t^i as follows [47, 48, 49, 52]:

$$p(y_t|x_t^i; \theta) \propto l_\theta(y_t, x_t^i), \quad (18)$$

where $l_\theta(\cdot) : \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$ is a function that is differentiable w.r.t. both y_t and x_t^i . In addition, proposal distributions can be specified through a function $f_\phi(\cdot) : \mathcal{X} \times \mathcal{Y} \times \mathbb{R}^{d_\beta} \rightarrow \mathcal{X}$ that are differentiable w.r.t. x_{t-1}^i , y_t , and β_t^i [50]:

$$x_t^i = f_\phi(x_{t-1}^i, y_t, \beta_t^i) \sim q(x_t^i|y_t, x_{t-1}^i; \phi), \quad (19)$$

where $\beta_t^i \in \mathbb{R}^{d_\beta}$ is the noise term used to generate proposal particles.

As a neural network-based machine learning model, the components of differentiable particle filters need to be trained often by minimising a given loss function $\mathcal{L}(\theta, \phi)$, e.g. the prediction error between the predicted latent states and the ground-truth latent states if they are available or the evidence lower bound (ELBO) of observation log-likelihoods. To minimise $\mathcal{L}(\theta, \phi)$ via gradient descent, one needs to compute the gradient of $\mathcal{L}(\theta, \phi)$ w.r.t. the system parameter θ and the proposal parameter ϕ , and we call a component differentiable if it can be interpreted as a differentiable function, i.e. it has a deterministic output and the derivative of its

output w.r.t. its input exists at each point in its domain – it is allowed in a machine learning context to have finitely many points where the derivatives do not exist. However, neither generating new particles nor the resampling steps in particle filters are differentiable in their vanilla forms, as they both lead to stochastic outputs. In addition, the output of the resampling step changes in a discrete manner as small changes in its input (particle weights) can lead to discrete changes in its output (particle indices). We introduce below how to construct differentiable components for particle filters such that the parameters θ and ϕ can be optimised using gradient descent.

Although they are not specifically designed for differentiable particle filters, different strategies have been proposed to achieve differentiable sampling, including the REINFORCE gradient estimator [74] and the reparameterisation trick [75], among which the reparameterisation trick is the one that most commonly used in differentiable particle filters. The essence of the reparameterisation trick is to rewrite samples from the target distribution as the output of a deterministic and differentiable function parameterised by distribution parameters such that the gradient of samples w.r.t. the parameters exists. In particular, consider a target distribution μ_β parameterised by β and defined on \mathbb{R}^{d_x} , the reparameterisation trick rewrites μ_β as a push-forward probability measure $\mu_\beta = G_\beta \# \nu$ of a base distribution ν defined on \mathbb{R}^{d_ϵ} , where $G_\beta(\cdot) : \mathbb{R}^{d_\epsilon} \rightarrow \mathbb{R}^{d_x}$ is a differentiable deterministic function parameterised by β . Samples following the target distribution μ_β can be obtained via sampling from the base distribution ν and transforming base samples $\epsilon \sim \nu$ through $x = G_\beta(\epsilon)$, with $x \sim \mu_\beta$. The gradient of an expectation of a function $\psi(\cdot) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_\psi}$ can now be rewritten as:

$$\nabla_\beta \mathbb{E}_{\mu_\beta} [\psi(x)] = \nabla_\beta \mathbb{E}_\nu [\psi(G_\beta(\epsilon))] \quad (20)$$

$$= \mathbb{E}_\nu [\nabla_\beta \psi(G_\beta(\epsilon))] \quad (21)$$

$$\approx \frac{1}{K} \sum_{k=1}^K \nabla_\beta \psi(G_\beta(\epsilon_k)), \quad (22)$$

where $\epsilon_k \stackrel{i.i.d.}{\sim} \nu$ are i.i.d samples from ν , and K is the number of samples used to estimate the required gradient. Gradient estimators obtained by using the reparameterisation trick were shown to be unbiased [75]. Note that not all distributions are reparameterisable. Some examples of reparameterisable distributions, e.g. Gaussian distributions, are listed in [75].

As for the resampling step, it is inherently non-differentiable because of the stochastic and discrete nature of multinomial resampling. Specifically, when using the inverse cumulative distribution function (CDF) algorithm to resample particles according to their importance weights, a small change in weights can lead to discrete changes in the resampling output. Recent efforts in differentiable particle filters have proposed various solutions to address this issue [47, 48, 52, 53, 73]. Several differentiable particle filtering approaches bypass resampling steps in gradient computation or simply do not use resampling [47, 48]. It was proposed in [47] to avoid backpropagating gradients across time steps, and the whole sequence was decomposed into multiple independent steps by truncating backpropagation after every single time step. It was suggested in [48] to use the same particles as the resampling output, i.e. no resampling, and this strategy was found to be effective in

low uncertainty environments. In standard resampling steps, the weights of survived particles \tilde{w}_t^i are set to be a constant, implying the gradient $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$ is always zero. To prevent this, several resampling schemes are proposed for differentiable particle filters to generate non-zero gradients $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$. For instance, in [73], the sum of unnormalised weights after resampling was kept the same as before resampling: $\tilde{w}_t^i = \frac{1}{N_p} \sum_{j=1}^{N_p} w_t^j$, such that the gradient $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$ becomes non-zero. In [48], it was proposed to resample particles from a multinomial distribution designed as a linear interpolation between the original multinomial distribution and a multinomial distribution with uniform weights. Biases caused by resampling from this interpolation are corrected by modifying the importance weights after resampling as a function of the original weights. This implies that the importance weights of particles after resampling now depend on the importance weights before resampling, hence $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$ becomes non-zero.

However, the aforementioned resampling techniques do not lead to fully differentiable resampling because they still rely on multinomial resampling, which is inherently non-differentiable as explained above. One solution to fully differentiable particle resampling is to consider it as an optimal transport (OT) problem [76]. Particularly, an entropy-regularised optimal transport-based resampler was proposed in [52]. This OT-based resampler reformulates non-differentiable resampling steps as deterministic and differentiable operations by solving the optimal transport problem in resampling steps with the Sinkhorn algorithm [77, 78, 79]. The OT-based resampler leads to an asymptotically consistent differentiable estimator of observation log-likelihoods, but it is biased due to the introduced non-zero regularisation term. Neural networks were utilised in the particle transformer proposed in [53], which is designed based on the set transformer architecture [80, 81]. The particle transformer is permutation invariant and scale equivariant by design. The output of the particle transformer is a set of new particles with equal weights. The particle transformer is trained by maximising the likelihood of particles before resampling in a Gaussian mixture distribution consisting of Gaussian distributions with new particles as their means. However, a particle transformer trained on one task does not adapt to other tasks, implying that one needs to train a new particle transformer from scratch each time there is a new task.

There are different strategies to train differentiable particle filters, which can be grouped into end-to-end training and those that involve individual training. In [47, 48, 49, 82, 83, 84], differentiable particle filters are trained in an end-to-end manner where all components of differentiable particle filters are jointly trained by minimising an overall loss function via gradient descent. In contrast, in [47, 85, 86, 87], it was proposed to first perform pre-training for independent modules of differentiable particle filters, then fine-tune these models jointly towards a task-specific target.

To optimise differentiable particle filters' parameters via gradient descent, a loss function or a training objective is needed. There are two major classes of loss functions that are often employed in training differentiable particle filters. The first class is the likelihood-based loss functions. In [88, 89, 90], an evidence lower bound (ELBO) of observation log-likelihood was derived for particle filters. The particle filter-based ELBO is a surrogate objective to the marginal log-likelihood and can be used to learn both model and proposal parameters [52, 89]. Besides, a pseudo-likelihood loss was proposed in [49] to enable semi-supervised learning

for differentiable particle filters when part of the ground-truth latent states is not provided. Another type of loss functions involves task-specific objectives. For example, in [47, 48, 49, 52, 82], differentiable particle filters are used in visual robot localisation tasks so the prediction error between estimates and the ground-truth latent states (RMSE or the log density of ground-truth latent states) are included in their loss functions. In [84], the state space was discretised and Gaussian blur was applied to the ground-truth latent states in order to minimise the KL divergence between the ground-truth posteriors and the particle estimate of posteriors. In [91], differentiable particle filters were used to guide the sampling process of the individualisation and refinement (IR) in graph neural networks (GNN), and the differentiable particle filters are learned by minimising the expected colouring error. A discriminative particle filter reinforcement learning (DPFRL) method was proposed in [83], where a differentiable particle filter is jointly trained end-to-end with a value network and a policy network by minimising a task-specific reinforcement learning loss. It was reported in [82] that combining task-specific loss with negative log-likelihood loss gives the best empirical performance in numerical simulations.

Having described the overall structure and design choices of differentiable particle filters, we now provide more detailed discussions on the key components of differentiable particle filters in the following sections.

3. Sampling Distributions of Differentiable Particle Filters. In particle filters, sampling distributions are the distributions from which particles are generated. The sampling distribution can be identical to the dynamic model $p(x_t|x_{t-1}; \theta)$ (bootstrap particle filters), or designed as a proposal distribution $q(x_t|y_t, x_{t-1}; \phi)$ (guided particle filters) conditioned on both the state x_{t-1} and the observation y_t . One of the key challenges in designing particle filters is the construction of sampling distributions that are easy to sample from, close to posterior distributions, and admit efficient weight update steps. We summarise below different design choices of sampling distributions in differentiable particle filters.

3.1. Gaussian dynamic models. In several differentiable particle filter approaches [47, 48, 49, 51, 52, 84, 87], the transition of latent states from x_{t-1} to x_t is modelled by a deterministic function $g_\theta(\cdot) : \mathcal{X} \times \mathbb{R}^{d_x} \rightarrow \mathcal{X}$ with the previous state x_{t-1} and a Gaussian noise term α_t as inputs and the new state x_t as outputs. We name such dynamic models as Gaussian dynamic models. Generally speaking, Gaussian dynamic models used in differentiable particle filters can be described as follows:

$$x_t^i \sim \mathcal{N}(\mu_\theta(x_{t-1}^i), \Sigma_{d_x}), \quad (23)$$

$$x_t^i = g_\theta(x_{t-1}^i, \alpha_t^i) = \mu_\theta(x_{t-1}^i) + \alpha_t^i, \quad (24)$$

where $\alpha_t^i \sim \mathcal{N}(\mathbf{0}, \Sigma_{d_x})$, and $\mu_\theta(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^{d_x}$ is a differentiable function of x_{t-1}^i [47, 48, 49, 52, 72]. The covariance matrix Σ_{d_x} can be either designed manually [47, 48, 49, 52] or parameterised and learned from data [72]. For data-adaptive covariance matrix Σ_{d_x} , two different ways to learn the covariance matrix Σ_{d_x} were proposed in [72], the constant noise model and the heteroscedastic noise model. In constant noise models, the covariance matrix $\Sigma_{d_x} = \text{diag}(\sigma_1^2, \dots, \sigma_{d_x}^2)$ is a diagonal matrix and keeps the same among different time steps. The diagonal elements $(\sigma_1^2, \dots, \sigma_{d_x}^2)$ of the covariance matrix Σ_{d_x} are set as learnable parameters and updated using gradient descent. For heteroscedastic noise models, the covariance matrix Σ_{d_x} is also a diagonal matrix, but it is predicted according to state values in heteroscedastic

noise models. Specifically, for the transition from time step $t - 1$ to time step t , a covariance matrix is computed for each particle:

$$(\sigma_1^i, \dots, \sigma_{d_x}^i) = \gamma_\theta(x_{t-1}^i), \quad (25)$$

$$\Sigma_{d_x}(x_{t-1}^i) = \text{diag}\left((\sigma_1^i)^2, \dots, (\sigma_{d_x}^i)^2\right), \quad (26)$$

where $\gamma_\theta(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^{d_x}$ is a neural network. The noise term α_t^i is sampled from $\mathcal{N}(\mathbf{0}, \Sigma_{d_x})$, where the covariance matrix $\Sigma_{d_x} = \sum_{i=1}^{N_p} W_t^i \Sigma_{d_x}(x_{t-1}^i)$ is computed as the weighted sum of individual covariance matrices $\Sigma_{d_x}(x_{t-1}^i)$, with W_t^i the normalised importance weight of x_{t-1}^i . With this setup, the covariance matrix Σ_{d_x} changes across time steps and hence the name heteroscedastic noise models.

Here we only introduced Gaussian dynamic models with additive Gaussian noise which are most commonly used in differentiable particle filters [47, 48, 49, 52, 72]. One can also use multiplicative Gaussian noise and other reparameterisable distributions depending on applications [75, 88].

3.2. Dynamic models constructed with normalising flows. A limitation of Gaussian dynamic models is their inability to adapt to complex systems where the latent state evolves according to complicated non-Gaussian dynamics. To address this, it was proposed in [50] to construct dynamic models in differentiable particle filters using normalising flows. Normalising flows are a family of invertible transformations that provides a flexible mechanism for constructing complex probability distributions [92]. By applying a series of invertible mappings on samples drawn from simple distributions, e.g. uniform or Gaussian distributions, normalising flows are able to transform simple distributions into arbitrarily complex distributions under some mild conditions [92]. In particular, it was proposed in [50] to build flexible dynamic models $p(x_t|x_{t-1}; \theta)$ upon base dynamic models $\tilde{p}(\tilde{x}_t|x_{t-1}; \theta)$, e.g. Gaussian dynamic models. Denote by $\{\tilde{x}_t^i\}_{i=1}^{N_p}$ a set of particles sampled from $\tilde{p}(\tilde{x}_t|x_{t-1}; \theta)$, in [50] the particles \tilde{x}_t^i , $i \in \{1, \dots, N_p\}$, are transformed by a normalising flow $\mathcal{T}_\theta(\cdot) : \mathcal{X} \rightarrow \mathcal{X}$:

$$x_t^i = \mathcal{T}_\theta(\tilde{x}_t^i) \sim p(x_t|x_{t-1}^i; \theta), \text{ where } \tilde{x}_t^i \sim \tilde{p}(\tilde{x}_t|x_{t-1}^i; \theta). \quad (27)$$

The probability density function $p(x_t|x_{t-1}^i; \theta)$ evaluated at x_t^i can be computed through the change of variable formula:

$$p(x_t^i|x_{t-1}^i; \theta) = \frac{\tilde{p}(\tilde{x}_t^i|x_{t-1}^i; \theta)}{\left| \det J_{\mathcal{T}_\theta}(\tilde{x}_t^i) \right|}, \text{ where } \tilde{x}_t^i = \mathcal{T}_\theta^{-1}(x_t^i) \sim \tilde{p}(\tilde{x}_t|x_{t-1}^i; \theta), \quad (28)$$

where $\det J_{\mathcal{T}_\theta}(\tilde{x}_t^i)$ refers to the determinant of the Jacobian matrix $J_{\mathcal{T}_\theta}(\tilde{x}_t^i) = \frac{\partial \mathcal{T}_\theta(\tilde{x}_t^i)}{\partial \tilde{x}_t^i}$ evaluated at \tilde{x}_t^i .

Because normalising flows can model complex non-Gaussian distributions, dynamic models constructed with normalising flows are theoretically more flexible than Gaussian dynamic models and require less prior knowledge of the system. In addition, experimental results reported in [50] provided empirical evidence that using normalising flows-based dynamic models in differentiable particle filters can produce better object tracking performance than using Gaussian dynamic models.

3.3. Dynamic models conditioned on observations. In several differentiable particle filtering approaches, dynamic models are constructed as $p(x_t|y_{1:t}, x_{t-1}; \theta)$ and used to generate new particles [82, 85]. This strategy has been successfully applied in a wide range of applications including robot localisation and sequence prediction tasks [82, 85]. For instance, the dynamic model of the differentiable particle filter proposed in [85] employs a neural network to model the displacement of robots. The neural network takes observation images (y_{t-1}, y_t) as its inputs, and outputs the mean and the variance of a Gaussian distribution that models the distribution of the robot's displacement. New particles $\{x_t^i\}_{i=1}^{N_p}$ are obtained by adding Gaussian noise drawn from the learned Gaussian distribution to $\{x_{t-1}^i\}_{i=1}^{N_p}$. A similar dynamic model is used in [83] to assist the training of a reinforcement learning algorithm. In [82], recurrent neural networks (RNNs) are used to construct dynamic models in differentiable particle filters. Specifically, it was proposed in [82] to model the evolution of the latent state using LSTM networks [93] or GRU networks [94], hence the name particle filter recurrent neural networks (PFRNNs). In the PFRNN method, particles at the t -th time step are generated through $x_t^i = \text{RNN}_\theta(x_{t-1}^i, y_t, \beta_t^i)$. β_t^i is sampled from a Gaussian distribution with learnable mean and variance which are the output of a function of x_{t-1}^i and y_t . Note that the latent state x_t in the PFRNN may not be the state of direct interest, and the weighted average of particles $\bar{x}_t = \sum_i^{N_p} W_t^i x_t^i$ is taken as the input of a task-dependent function to predict the state of interest. For example, in the robot localisation experiment presented in [82], the task-dependent function maps \bar{x}_t to robot locations, and in classification tasks, the outputs of the task-dependent function are probability vectors of categorical distributions.

3.4. Proposal distributions. To the best of our knowledge, the majority of existing differentiable particle filter approaches are restricted to the bootstrap particle filtering framework which does not utilise observations to construct proposal distributions. When latent states evolve according to a simple dynamics and the relationship between x_t and y_t can be easily speculated with prior knowledge, optimal proposals can be explicitly derived or approximated with closed-form expressions [59]. However, in more complex environments such as vision-based robot localisation, it is intractable to explicitly derive or approximate optimal proposals and therefore one needs to learn proposal distributions from data for differentiable particle filters.

Different ways to learn proposal distributions in differentiable particle filters have been proposed [50, 95, 96]. In [95, 96], the proposal distribution is constructed as a Gaussian distribution whose mean and covariance matrix are given by a neural network with the estimate of the previous latent state and the current observation as inputs. To provide a mechanism for constructing proposal distributions more flexible than Gaussian distributions, it was proposed in [50] to construct proposal distributions for differentiable particle filters with conditional normalising flows [97]. Specifically, proposal distributions in [50] are constructed by stacking a conditional normalising flow $\mathcal{G}_\phi(\cdot) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X}$ upon the dynamic model:

$$x_t^i = \mathcal{G}_\phi(\tilde{x}_t^i, y_t) \sim q(x_t|y_t, x_{t-1}^i; \phi), \text{ where } \tilde{x}_t^i \sim p(\tilde{x}_t|x_{t-1}^i; \theta). \quad (29)$$

The conditional normalising flow $\mathcal{G}_\phi(\cdot)$ is an invertible function of the particle $\tilde{x}_t^i \sim p(\tilde{x}_t|x_{t-1}^i; \theta)$ when the observation y_t is given. Since information from the latest observation is taken into account, the conditional normalising flow $\mathcal{G}_\phi(\cdot)$ has the potential to migrate samples drawn from the prior to regions that are closer to the

posterior. Similar to Eq. (28), the proposal density can be evaluated by applying the change of variable formula:

$$q(x_t^i|y_t, x_{t-1}^i; \phi) = \frac{p(\check{x}_t^i|x_{t-1}^i; \theta)}{\left| \det J_{\mathcal{G}_\phi}(\check{x}_t^i) \right|}, \text{ where } \check{x}_t^i = \mathcal{G}_\phi^{-1}(x_t^i, y_t) \sim p(\check{x}_t^i|x_{t-1}^i; \theta). \quad (30)$$

$\det J_{\mathcal{G}_\phi}(\check{x}_t^i)$ refers to the determinant of the Jacobian matrix, i.e., $J_{\mathcal{G}_\phi}(\check{x}_t^i) = \frac{\partial \mathcal{G}_\phi(\check{x}_t^i, y_t)}{\partial \check{x}_t^i}$ evaluated at \check{x}_t^i .

4. Measurement Models of Differentiable Particle Filters. In this section we briefly review different designs of measurement models in the differentiable particle filter literature. We group measurement models used in differentiable particle filters into four categories. The first category estimates $p(y_t|x_t; \theta)$ using known distributions with unknown parameters [52, 88], e.g. parameterised Gaussian distributions or Poisson distributions. The second category relies on neural networks to learn a scalar function to approximate $p(y_t|x_t; \theta)$ [47, 48]. The third category measures the discrepancy or similarity between observations and latent states in a learned feature space [49, 50]. The last category utilises conditional normalising flows to learn the generating process of y_t and consequently the density of y_t given x_t [51]. More details on these different types of measurement models are provided below.

4.1. Measurement models built with known distributions. We first introduce measurement models used in differentiable particle filters that capture the relationship between x_t and y_t using known distributions with unknown parameters. An example of this type of measurement model can be found in [89], where differentiable particle filters are used to learn the parameters of a linear-Gaussian state-space model:

$$x_0 \sim \mathcal{N}(0, 1), \quad (31)$$

$$x_t|x_{t-1} \sim \mathcal{N}(\theta_1 x_{t-1}, 1), \text{ for } t \geq 1, \quad (32)$$

$$y_t|x_t \sim \mathcal{N}(\theta_2 x_t, 0.1), \text{ for } t \geq 1, \quad (33)$$

where $\theta = \{\theta_1, \theta_2\}$ are parameters to be estimated. In the above example, the likelihood of y_t given x_t is established as the probability density function of a Gaussian distribution with learnable parameters.

Measurement models built with known distributions are limited to cases where the relationship between y_t and x_t can be explicitly inferred. In more complex systems where observations are high-dimensional unstructured data like images, one often needs to design measurement models using more powerful tools such as neural networks.

4.2. Measurement models as scalar functions built with neural networks. In scenarios where it is non-trivial to model $p(y_t|x_t; \theta)$ with explicit closed-form expressions, several differentiable particle filtering approaches employ neural networks to approximate $p(y_t|x_t; \theta)$ [47, 48]. A straightforward idea is to learn a scalar function $l_\theta(\cdot) : \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$ constructed by neural networks and consider its output as a measure of the compatibility between x_t and y_t . In [47], observations are first mapped into a feature space through $e_t = E_\theta(y_t) \in \mathbb{R}^{d_e}$, where $E_\theta(\cdot) : \mathcal{Y} \rightarrow \mathbb{R}^{d_e}$ is a neural network. Then another neural network $h_\theta(\cdot) : \mathbb{R}^{d_e} \times \mathcal{X} \rightarrow \mathbb{R}$ with e_t and x_t^i

as inputs is employed, whose outputs are considered as the unnormalised likelihood of y_t given x_t^i :

$$e_t = E_\theta(y_t), \quad (34)$$

$$\begin{aligned} p(y_t|x_t^i; \theta) &\propto l_\theta(y_t, x_t^i) \\ &= h_\theta(E_\theta(y_t), x_t^i). \end{aligned} \quad (35)$$

In [48], differentiable particle filters are applied in robot localisation tasks where a global 2D map is provided and observations are 3D local images. For this environment, spatial transformer networks [98] are used in [48] to generate local maps $M_{x_t^i} \in \mathbb{R}^h \times \mathbb{R}^w$ based on the global map $\mathbb{M} \in \mathbb{R}^H \times \mathbb{R}^W$ and particle locations x_t^i , where h , w , H , and W are the height and width of the local and global maps, respectively. Then, the 2D local map $M_{x_t^i}$ is compared with the 3D local image y_t through a neural network $h_\theta(\cdot) : \mathcal{Y} \times (\mathbb{R}^h \times \mathbb{R}^w) \rightarrow \mathbb{R}$:

$$M_{x_t^i} = \mathbf{ST}_\theta(\mathbb{M}, x_t^i), \quad (36)$$

$$p(y_t|x_t^i; \theta) \propto l_\theta(y_t, M_{x_t^i}) \quad (37)$$

$$= h_\theta(y_t, M_{x_t^i}), \quad (38)$$

where $\mathbf{ST}_\theta(\cdot) : (\mathbb{R}^H \times \mathbb{R}^W) \times \mathcal{X} \rightarrow \mathbb{R}^h \times \mathbb{R}^w$ refers to a spatial transformer network [98].

One open question for this category of measurements is how to normalise the learned unnormalised likelihood of y_t given x_t^i . When the loss function used to train differentiable particle filters is maximum likelihood-based, a measurement model built with a neural network and scalar, unnormalised output may assign arbitrarily high likelihood to any input, leading to poor generalisation performance of differentiable particle filters with these setups.

4.3. Measurement models as feature similarities. Instead of solely relying on neural networks to build measurement models for differentiable particle filters, there are other differentiable particle filtering approaches that consider neural networks as feature extractors and employ user-defined metric functions to compare observation features and state features. In [49], two neural networks $E_\theta(\cdot) : \mathcal{Y} \rightarrow \mathbb{R}^{d_e}$ and $O_\theta(\cdot) : \mathcal{X} \rightarrow \mathbb{R}^{d_e}$ are used to extract features from observations and states respectively, then $p(y_t|x_t^i; \theta)$ is assumed to be proportional to the reciprocal of the cosine distance between $E_\theta(y_t)$ and $O_\theta(x_t^i)$:

$$e_t = E_\theta(y_t), o_t^i = O_\theta(x_t^i), \quad (39)$$

$$c(e_t, o_t^i) = 1 - \frac{e_t \cdot o_t^i}{\|e_t\|_2 \|o_t^i\|_2}, \quad (40)$$

$$p(y_t|x_t^i; \theta) \propto \frac{1}{c(e_t, o_t^i)}, \quad (41)$$

where $c(\cdot) : \mathbb{R}^{d_e} \times \mathbb{R}^{d_e} \rightarrow \mathbb{R}$ refers to the cosine distance, $e_t \cdot o_t^i$ denotes the inner product of features vectors, and $\|\cdot\|_2$ is the L_2 norm of feature vectors. In the robot localisation experiment of [52], the relationship between observation features and state features is modelled as follows:

$$e_t = E_\theta(y_t) \sim \mathcal{N}(o_t^i, \sigma_{\text{obs}}^2 \mathbf{I}). \quad (42)$$

In other words, the conditional likelihood of y_t given x_t is computed as the Gaussian density of the observation feature e_t with $o_t^i = O_\theta(x_t^i)$ and $\sigma_{\text{obs}}^2 \mathbf{I}$ as its mean and variance.

To ensure that the observation feature extractor $E_\theta(\cdot)$ maintains essential information from y_t , it was suggested to add an auto-encoder loss into the training objective [49, 52]:

$$\mathcal{L}_{\text{AE}} = \sum_{t=0}^T \|D_\theta(E_\theta(y_t)) - y_t\|_2^2, \quad (43)$$

where T is the length of sequences, $E_\theta(\cdot) : \mathcal{Y} \rightarrow \mathbb{R}^{d_e}$ and $D_\theta(\cdot) : \mathbb{R}^{d_e} \rightarrow \mathcal{Y}$ are the observation encoder and the observation decoder, respectively.

4.4. Measurement models built with conditional normalising flows. A conditional normalising flow-based measurement model was proposed for differentiable particle filters in [51]. Conditional normalising flow-based measurement models provide a mechanism to compute valid conditional probability density of y_t given x_t . In particular, the measurement model proposed in [51] computes $p(y_t|x_t^i; \theta)$ via:

$$p(y_t|x_t^i; \theta) = p_z(\mathcal{F}_\theta(y_t, x_t^i)) \left| \det \frac{\partial \mathcal{F}_\theta(y_t, x_t^i)}{\partial y_t} \right|, \quad (44)$$

where $\mathcal{F}_\theta(\cdot) : \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}^{d_y}$ is a parameterised conditional normalising flow, and $\left| \det \frac{\partial \mathcal{F}_\theta(y_t, x_t^i)}{\partial y_t} \right|$ is the determinant of the Jacobian matrix $\frac{\partial \mathcal{F}_\theta(y_t, x_t^i)}{\partial y_t}$ evaluated at y_t .

The base distribution $p_z(\cdot)$ defined on \mathbb{R}^{d_y} can be user-specified and is often chosen as a simple distribution such as an isotropic Gaussian.

In scenarios where observations y_t are high-dimensional data like images, the evaluation of Eq. (44) with raw observations can be computationally expensive. In addition, when observations contain too much irrelevant information, the construction of measurement models with raw observations even becomes more challenging. As an alternative solution, it was proposed in [51] to map observations y_t to a lower-dimensional feature space through a neural network $E_\theta(\cdot) : \mathcal{Y} \rightarrow \mathbb{R}^{d_e}$. With the assumption that the conditional probability density $p(e_t|x_t^i; \theta)$ of observation features $e_t = E_\theta(y_t)$ given state is an approximation of the actual measurement likelihood $p(y_t|x_t^i; \theta)$, in [51] the conditional likelihood of observations is computed by:

$$p(y_t|x_t^i; \theta) \approx p(e_t|x_t^i; \theta) \quad (45)$$

$$= p_z(\mathcal{F}_\theta(e_t, x_t^i)) \left| \det \frac{\partial \mathcal{F}_\theta(e_t, x_t^i)}{\partial e_t} \right|. \quad (46)$$

5. Differentiable Resampling. Resampling is the key operation that distinguishes sequential Monte Carlo (SMC) methods from sequential importance sampling (SIS). It is also a major obstacle to achieve fully differentiable particle filters. Due to the fact that the resampling output changes in a discrete manner, it has been widely documented that the resampling step is not differentiable [47, 48, 52, 73]. Additionally, since the importance weights of particles are set to be a constant after resampling, the gradients of particle weights at the t -th time step w.r.t. particle weights at time step $t' < t$ are always zero, rendering the operation non-differentiable in a machine learning context [48].

To estimate the gradients of resampling steps, different solutions have been proposed in [48, 52, 53] to achieve differentiable resampling. We review these techniques below.

5.1. Soft resampling. The soft resampling method proposed in [48] attempts to generate non-zero values for the gradients $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$, where $\partial \tilde{w}_t^i$ and ∂w_t^i are unnormalised particle weights after and before resampling, respectively. Denote by $\{W_t^i\}_{i=1}^{N_p}$ the set of normalised particle weights before resampling, the soft resampling constructs an importance multinomial distribution $\text{Mult}(\tilde{W}_t^1, \dots, \tilde{W}_t^{N_p})$ defined with $\{\tilde{W}_t^i\}_{i=1}^{N_p}$:

$$\tilde{W}_t^i = \lambda W_t^i + (1 - \lambda) \frac{1}{N_p}. \quad (47)$$

The indices of selected particles are sampled from the importance multinomial distribution $\text{Mult}(\tilde{W}_t^1, \dots, \tilde{W}_t^{N_p})$. The importance distribution $\text{Mult}(\tilde{W}_t^1, \dots, \tilde{W}_t^{N_p})$ can be interpreted as a linear interpolation between the original multinomial distribution $\text{Mult}(W_t^1, \dots, W_t^{N_p})$ and a multinomial distribution $\text{Mult}(\frac{1}{N_p}, \dots, \frac{1}{N_p})$ with equal weights $\frac{1}{N_p}$. To correct the bias caused by sampling from the importance distribution, soft resampling updates particle weights after resampling as:

$$\tilde{w}_t^i = \frac{W_t^i}{\tilde{W}_t^i} = \frac{W_t^i}{\lambda W_t^i + (1 - \lambda) 1/N_p}. \quad (48)$$

Since the modified weights in Eq. (48) is a function of the original weight $W_t^i = \frac{w_t^i}{\sum_{j=1}^{N_p} w_t^j}$, the gradient $\frac{\partial \tilde{w}_t^i}{\partial w_t^i}$ yields non-zero values with the soft resampling method.

However, soft resampling still relies on sampling from multinomial distributions, thus it does not change the discrete nature of the resampling step. In other words, the soft resampling outputs may still change in a discrete manner with slight changes in W_t^i , i.e. the non-differentiable part of resampling is ignored in soft resampling [52]. Nonetheless, soft resampling has been widely used in the differentiable particle filter literature, due to its simple implementation and competitive empirical performance in practice [49, 82, 83, 85, 86].

5.2. Differentiable resampling via entropy-regularised optimal transport. An optimal transport (OT) map-based differentiable resampling algorithm was proposed in [52], which we refer to as OT-resampler. The OT-resampler finds the optimal transport map between an equally weighted empirical distribution $\rho(x_t) = \frac{1}{N_p} \sum_{i=1}^{N_p} \delta_{x_t^i}(x_t)$ and the target empirical distribution $\hat{\rho}(x_t|y_{1:t}; \theta) = \sum_{i=1}^{N_p} W_t^i \delta_{x_t^i}(x_t)$. Since the original optimal transport problem is computationally expensive and not differentiable, the optimal transport map is obtained by solving an entropy-regularised version of optimal transport problems [99, 100] with the Sinkhorn algorithm [77, 78, 79] which is efficient and differentiable.

Let us first consider the application of optimal transport in particle resampling without regularisation. Ideally, we want to find a transportation map $T : \mathcal{X} \rightarrow \mathcal{X}$ such that the resulting distribution $T\# \rho(x_t) = \frac{1}{N_p} \sum_{i=1}^{N_p} \delta_{\tilde{x}_t^i}(x_t)$ with $\tilde{x}_t^i = T(x_t^i)$ exactly matches the target distribution $\hat{\rho}(x_t|y_{1:t}; \theta) = \sum_{i=1}^{N_p} W_t^i \delta_{x_t^i}(x_t)$. However, this is not possible because the resulting distribution has equal weights while the target

distribution does not. An approximation $\hat{T} : x_t^i \in \mathcal{X} \rightarrow N_p \sum_{j=1}^{N_p} P^{\text{OT}}(i, j) x_t^j \in \mathcal{X}$ of the optimal transportation map T is the so-called barycentric projection [52, 79]:

$$\hat{x}_t^i = N_p \sum_{j=1}^{N_p} P^{\text{OT}}(i, j) x_t^j, \quad (49)$$

$$\hat{X}_t = N_p P^{\text{OT}} X_t, \text{ where } P^{\text{OT}} \in \mathbb{R}^{N_p \times N_p}. \quad (50)$$

The matrix P^{OT} is the optimal coupling that produces the minimal transportation cost $\sum_{i,j}^{N_p} P^{\text{OT}}(i, j) \mathcal{D}(x_t^i, x_t^j)$ for a predefined distance metric $\mathcal{D}(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, and satisfies $P^{\text{OT}} \mathbf{1}_{N_p} = (\frac{1}{N_p}, \dots, \frac{1}{N_p})^\top$, $(P^{\text{OT}})^\top \mathbf{1}_{N_p} = (W_t^1, \dots, W_t^{N_p})^\top$. We use $P^{\text{OT}}(i, j)$ to denote the (i, j) -th element of P^{OT} and $\mathbf{1}_{N_p} \in \mathbb{R}^{N_p}$ a N_p -dimensional column vector with all entries set to 1. In Eq. (50), the i -th row of $\hat{X}_t \in \mathbb{R}^{N_p \times d_x}$ and $X_t \in \mathbb{R}^{N_p \times d_x}$ are the i -th transported particle \hat{x}_t^i and the i -th original particle x_t^i , respectively. The transported particle \hat{x}_t^i is the barycentre of the empirical distribution $\tau_t^i(x_t) = N_p \sum_{j=1}^{N_p} P^{\text{OT}}(i, j) \delta_{x_t^j}(x_t)$, hence the name barycentric projection. The distribution $\hat{T} \# \rho(x_t) = \frac{1}{N_p} \sum_{i=1}^{N_p} \delta_{\hat{x}_t^i}(x_t)$ is considered as the resampling output of the OT-resampler.

However, as aforementioned, the original OT problem leads to high computational overhead and is not differentiable. In [52], to make the OT-based resampling efficient and differentiable, it was proposed to solve the dual form of an entropy-regularised version of the original OT problem with the Sinkhorn algorithm [77]. Using $P_\xi^{\text{OT}} \in \mathbb{R}^{N_p \times N_p}$ to refer to the entropy-regularised coupling with regularisation coefficient ξ , resampled particles are obtained by applying the corresponding barycentric projection $\hat{T}_\xi : x_t^i \in \mathcal{X} \rightarrow N_p \sum_{j=1}^{N_p} P_\xi^{\text{OT}}(i, j) x_t^j \in \mathcal{X}$:

$$\hat{x}_{t,\xi}^i = N_p \sum_{j=1}^{N_p} P_\xi^{\text{OT}}(i, j) x_t^j, \quad (51)$$

$$\hat{X}_{t,\xi} = N_p P_\xi^{\text{OT}} X_t. \quad (52)$$

While the OT-resampler is fully differentiable, the regularisation coefficient ξ introduces biases into the OT-resampler, and the bias only vanishes when $\xi \rightarrow 0$. Despite the biasedness of the OT-resampler, the OT-resampler has been shown to outperform soft resampling in experiments conducted in [52].

5.3. Differentiable resampling via neural networks. The particle transformer proposed in [53] is a neural network architecture designed for performing differentiable particle resampling. Since the inputs of the resampling step are point sets, the particle transformer is designed to be permutation-invariant and scale-equivariant by utilising the set transformer [80]. The input of the particle transformer is a set of weighted particles $\{W_t^i, x_t^i\}_{i=1}^{N_p}$, and the output of the particle transformer is a set of new particles $\{\tilde{x}_t^i\}_{i=1}^{N_p}$:

$$\{\tilde{x}_t^i\}_{i=1}^{N_p} = \mathbf{PT}_\theta(\{W_t^i, x_t^i\}_{i=1}^{N_p}), \quad (53)$$

where $\mathbf{PT}_\theta(\cdot)$ denotes the particle transformer. The output particles are assigned with equal weights $\tilde{W}_t^i = \frac{1}{N_p}$.

Since it is a neural network-based resampler, a training objective is needed to train the particle transformer. In [53], the particle transformer is optimised by

maximising the likelihood of original samples $\{x_t^i\}_{i=1}^{N_p}$ in the distribution after resampling, and the likelihood of a particle x_t^i is weighted by its importance weight W_t^i . To convert empirical distributions into continuous distributions with proper probability density functions, it was proposed in [53] to construct a Gaussian mixture distribution based on the empirical distribution after resampling (one Gaussian kernel per particle). Components of the Gaussian mixture model are weighted equally, since particles after resampling have uniform weights. The loss function used to train the particle transformer is defined as the weighted likelihood of the original samples in the Gaussian mixture distribution:

$$-\sum_{i=1}^{N_p} W_t^i \log \sum_{j=1}^{N_p} \frac{1}{\sqrt{|\Sigma|}} \exp \left(-\frac{1}{2} (x_t^i - \tilde{x}_t^j)^\top \Sigma^{-1} (x_t^i - \tilde{x}_t^j) \right), \quad (54)$$

where Σ is a specified covariance matrix of the Gaussian mixture model, and $|\Sigma|$ refers to the determinant of Σ . After training the particle transformer individually, the particle transformer was integrated into differentiable particle filters and trained jointly by minimising a task-specific loss function [53].

Particle transformer requires pre-training, implying that one needs to collect training data beforehand, which may not be practical in some applications. In addition, a particle transformer trained for a specific task may not fit into other tasks and requires re-training to deploy it in a new task. Additionally, experimental results presented in [53] show that the best performance is achieved if one does not backpropagate through the particle transformer when training differentiable particle filters in the reported experiment.

6. Loss Functions for Training Differentiable Particle Filters. With all the components we have described, we need a loss function to train neural networks used to build the above key components of differentiable particle filters. In the differentiable particle filter literature, loss functions can be mainly grouped into three categories, supervised losses, semi-supervised losses, and variational objectives, as described below.

6.1. Supervised losses. Supervised losses are a popular choice in differentiable particle filters when ground-truth latent states are available [47, 48, 50, 51, 52], which we use x_t^* to denote the ground-truth latent states. Training differentiable particle filters with supervised losses implies the ground-truth latent state x_t^* is available, therefore we can evaluate the similarity between the estimated state \bar{x}_t given by differentiable particle filters and the ground-truth x_t^* . There are two major classes of supervised losses used in differentiable particle filters, i.e., the root mean square error (RMSE) and the negative (state) likelihood loss.

We first introduce the RMSE loss. The RMSE loss computes the distance between the predicted states \bar{x}_t and the ground-truth latent states x_t^* :

$$\bar{x}_t = \sum_{i=1}^{N_p} W_t^i x_t^i, \quad (55)$$

$$\mathcal{L}_{\text{RMSE}} := \sqrt{\frac{1}{T+1} \sum_{t=0}^T \|x_t^* - \bar{x}_t\|_2^2}, \quad (56)$$

where $\|\cdot\|_2$ denotes the L_2 norm, and $T + 1$ is the total number of time steps. Minimising the RMSE loss is equivalent to maximising the log-likelihood of the ground-truth latent state x_t^* in the Gaussian distribution with mean \bar{x}_t and an identity covariance matrix.

In contrast to the RMSE loss which considers only the final estimates, the so-called data likelihood loss takes into account all the particles [47]. The log-likelihood loss \mathcal{L}_{LL} computes the logarithm of probability density of ground-truth latent states x_t^* in a Gaussian mixture distribution defined by a set of weighted particles $\{W_t^i, x_t^i\}_{i=1}^{N_p}$:

$$\mathcal{L}_{\text{LL}} := -\frac{1}{T+1} \sum_{t=0}^T \log \sum_{i=1}^{N_p} \frac{W_t^i}{\sqrt{|\Sigma|}} \exp \left(-\frac{1}{2} (x_t^* - x_t^i)^\top \Sigma^{-1} (x_t^* - x_t^i) \right), \quad (57)$$

where Σ is a specified covariance matrix of the Gaussian mixture model, and $|\Sigma|$ refers to the determinant of Σ .

Despite that supervised losses have been empirically proven to be effective among various differentiable particle filtering methods [47, 48, 50, 51, 52], one limitation of supervised losses targeting either the posterior mean or the largest posterior mode is that they may hinder the particle filter's ability to capture less prominent modes or the tail of the distribution.

Another limitation of training differentiable particle filters with supervised losses is that they require sufficient training data with ground-truth latent state information, which can be expensive to collect and are often unavailable in real-world applications. We discuss alternative choices of training loss functions below.

6.2. Semi-supervised loss. When ground-truth latent states are not provided, a common choice of training objectives for differentiable particle filters is the marginal likelihood of observations $p(y_{1:t}; \theta)$. In [101], it was proposed to search for point estimates of static parameters in particle filters by maximising the log-likelihood $L_t(\theta) = \log p(y_{1:t}; \theta)$ using noisy gradient algorithms. However, this approach requires the estimation of sufficient statistics of joint distributions whose dimension increases over time. An alternative is to use a pseudo-likelihood [71]. Specifically, it was proposed in [71] to split the data into slices and consider the product of each block's likelihood as the pseudo-likelihood of the whole trajectory. Inspired by [71], a pseudo-likelihood loss was introduced in [49] to enable semi-supervised learning in differentiable particle filters.

To introduce the semi-supervised loss, one first derives a logarithmic pseudo-likelihood $\mathcal{Q}(\theta)$. Consider m slices $X_b := x_{bL:(b+1)L-1}$ and $Y_b := y_{bL:(b+1)L-1}$ of $\{x_t\}_{t=0}^T$ and $\{y_t\}_{t=1}^T$, where L is the time length of each block, X_b and Y_b are the b -th block of states and observations, respectively. The log pseudo-likelihood for m blocks of observations is defined as:

$$\mathcal{Q}(\theta) := \sum_{b=0}^{m-1} \log p(Y_b; \theta). \quad (58)$$

Compared to the true log-likelihood $L_t(\theta) = \log p(y_{1:t}; \theta)$, the pseudo-likelihood overcomes the increasing dimension problem by ignoring the dependency between blocks.

At the b -th block, denote by θ_b the current estimate of parameter value and θ^* the optimal parameter value, $\mathcal{Q}(\theta, \theta_b)$ is maximised over θ , which is equivalent to

maximising the pseudo-likelihood $\mathcal{Q}(\theta)$, where $\mathcal{Q}(\theta, \theta_b)$ is defined as follows:

$$\mathcal{Q}(\theta, \theta_b) := \int_{\mathcal{X}^L \times \mathcal{Y}^L} \log(p(X_b, Y_b; \theta)) \cdot p(X_b|Y_b; \theta_b) p(Y_b; \theta^*) dX_b dY_b. \quad (59)$$

Particularly, $\mathcal{Q}(\theta, \theta_b)$ is maximised through gradient descent with the initial value of θ set to be θ_b . With the pseudo-likelihood objective, we can train differentiable particle filters in a semi-supervised manner as follows:

$$\theta = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(\lambda_1 \mathcal{L}(\theta) - \frac{\lambda_2}{m} \mathcal{Q}(\theta, \theta_b) \right), \text{ if ground-truth latent states are available,} \quad (60)$$

$$\theta = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(-\frac{\lambda_2}{m} \mathcal{Q}(\theta, \theta_b) \right), \text{ if ground-truth latent states are not available,} \quad (61)$$

where the supervised loss $\mathcal{L}(\theta)$ can be selected from supervised losses introduced in previous sections and λ_1, λ_2 are hyperparameters.

6.3. Variational objectives. In variational inference (VI) problems, using the evidence lower bound (ELBO) as a surrogate objective for the maximum likelihood estimation (MLE) enables one to perform model learning and proposal learning simultaneously and has demonstrated its effectiveness in the literature [75, 102, 103, 104, 105, 106]. There are two main components in VI, i.e., a parametric family of variational distributions and the ELBO as the training objective. Following the framework of VI, in [88, 89, 90], the components of particle filters were optimised with a filtering variational ELBO, which was proposed based on the estimator of the marginal likelihood given by particle filters.

To present the filtering variational objective, we first introduce a unified perspective on ELBOs used in VI. Given a set of observations $\{y_t\}_{t=1}^T$, the ELBO of the logarithmic marginal likelihood $\log p(y_{1:T}; \theta)$ can be derived by:

$$\log p(y_{1:T}; \theta) = \log \int_{\mathcal{X}^T} p(y_{1:T}, x_{0:T}; \theta) dx_{0:T} \quad (62)$$

$$= \log \int_{\mathcal{X}^T} \frac{p(y_{1:T}, x_{0:T}; \theta)}{q(x_{0:T}|y_{1:T}; \phi)} q(x_{0:T}|y_{1:T}; \phi) dx_{0:T} \quad (63)$$

$$= \log \mathbb{E}_q \left[\frac{p(y_{1:T}, x_{1:T}; \theta)}{q(x_{0:T}|y_{1:T}; \phi)} \right] \quad (64)$$

$$\geq \mathbb{E}_q \left[\log \frac{p(y_{1:T}, x_{1:T}; \theta)}{q(x_{0:T}|y_{1:T}; \phi)} \right], \quad (65)$$

where Jensen's inequality is applied from Eq. (64) to Eq. (65), and $q(x_{0:T}|y_{1:T}; \phi)$ is the proposal distribution. It can be observed from Eq. (62) to Eq. (65) that as long as we can derive an unbiased estimator of the observation likelihood $p(y_{1:T}; \theta)$, e.g. $\mathbb{E}_q \left[\frac{p(y_{1:T}, x_{0:T}; \theta)}{q(x_{0:T}|y_{1:T}; \phi)} \right]$ in Eq. (64), we can apply Jensen's inequality to obtain an ELBO of the logarithmic marginal likelihood $\log p(y_{1:T}; \theta)$. Following this pattern, the importance weighted auto-encoder proposes to use the estimator $\frac{1}{N_p} \sum_{i=1}^{N_p} \frac{p(y_{1:T}, x_{0:T}^i; \theta)}{q(x_{0:T}^i|y_{1:T}; \phi)}$ by drawing multiple samples from the proposal:

$$\log p(y_{1:T}; \theta) = \log \mathbb{E}_{\mathcal{Q}_{\text{IWAE}}} \left[\frac{1}{N_p} \sum_{i=1}^{N_p} \frac{p(y_{1:T}, x_{0:T}^i; \theta)}{q(x_{0:T}^i|y_{1:T}; \phi)} \right] \quad (66)$$

$$\geq \mathbb{E}_{\mathcal{Q}_{\text{IWAE}}} \left[\log \left(\frac{1}{N_p} \sum_{i=1}^{N_p} \frac{p(y_{1:T}, x_{0:T}^i; \theta)}{q(x_{0:T}^i | y_{1:T}; \phi)} \right) \right], \quad (67)$$

where $\mathcal{Q}_{\text{IWAE}} = \prod_{i=1}^{N_p} q(x_{0:T}^i | y_{1:T}; \phi)$. It was proved in [103] that with $N_p \geq 1$, Eq. (67) is a tighter bound than the ELBO defined by Eq. (65). Similarly, the filtering ELBO for SMC methods can be derived as:

$$\log p(y_{1:T}; \theta) \geq \mathbb{E}_{\mathcal{Q}_{\text{SMC}}} [\log \hat{p}(y_{1:T}; \theta)] \quad (68)$$

$$= \int_{\mathcal{X}_{0:T}} \mathcal{Q}_{\text{SMC}}(x_{0:T}^{1:N_p}, A_{0:T-1}^{1:N_p}) \log \hat{p}(y_{1:T}; \theta) dx^{1:N_p}, \quad (69)$$

where $\mathcal{Q}_{\text{SMC}}(x_{0:T}^{1:N_p}, A_{0:T-1}^{1:N_p})$ is the sampling distribution of SMC [89], $x_{0:T}^{1:N_p}$ and $A_{0:T-1}^{1:N_p}$ respectively refer to samples from the proposal distribution and resampling indices, and $\hat{p}(y_{1:T}; \theta)$ is the SMC estimates of the marginal likelihood:

$$\hat{p}(y_{1:T}; \theta) = \prod_{t=0}^T \left[\frac{1}{N_p} \sum_{i=1}^{N_p} w_t^i \right].$$

7. Examples of pseudocode for differentiable particle filters. With the above sections introducing different components of differentiable particle filters, we provide in this section pseudocode for two recent differentiable particle filtering frameworks to reveal sufficient implementation details for interested readers to implement differentiable particle filters in their applications.

7.1. Particle filter networks. We first present the implementation detail of the particle filter network (PFNet), a differentiable particle filter proposed in [48]. The PFNet was developed for a robot localisation task, where the goal in the task is to localise a robot with image observations collected from onboard cameras. The soft resampler introduced in Section 5.1 was applied in the PFNet. The latent state in the PFNet is defined as $x_t = [s_t^{(1)}, s_t^{(2)}, \eta_t]$, where $[s_t^{(1)}, s_t^{(2)}]$ and η_t are the location and the orientation of the robot respectively. At each time step, the actions $a_t = [v_t^{(1)}, v_t^{(2)}, \omega_t]$ of the robot are given. Given actions, the transition of the robot location is as follows:

$$\hat{\eta}_t = \eta_t + \alpha_t^{(3)}, \quad \eta_{t+1} = \hat{\eta}_t + \omega_t, \quad (70)$$

$$s_{t+1}^{(1)} = s_t^{(1)} + v_t^{(1)} \cos(\hat{\eta}_t) + v_t^{(2)} \sin(\hat{\eta}_t) + \alpha_t^{(1)}, \quad (71)$$

$$s_{t+1}^{(2)} = s_t^{(2)} + v_t^{(1)} \sin(\hat{\eta}_t) - v_t^{(2)} \cos(\hat{\eta}_t) + \alpha_t^{(2)}, \quad (72)$$

where $\alpha_t^{(1)} \sim \mathcal{N}(0, \sigma_1^2)$, $\alpha_t^{(2)} \sim \mathcal{N}(0, \sigma_2^2)$, and $\alpha_t^{(3)} \sim \mathcal{N}(0, \sigma_3^2)$ are the noise term of the dynamic model. The PFNet uses its dynamic model to propose new particles, i.e. the PFNet is a bootstrap particle filter. Global 2D maps of different environments are given, and a spatial transformer network is used in the PFNet to transform global maps into local 2D maps according to the robot's location and orientation. The local 2D maps are then compared with the onboard 3D images from robot cameras to evaluate the conditional likelihood of observations given states, as introduced in Section 4.2. The loss function used to train the PFNet is the mean square error (MSE) between the predicted state and the ground-truth latent state.

The pseudocode for the PFNet can be found in Algorithm 2. The PFNet has achieved impressive performance in a challenging robot localisation task and generalises well to unseen environments.

Algorithm 2: Particle filter network (PFNet) [48].

1: **Inputs:** Initial distribution $\pi(x_0; \theta)$, resampling threshold ESS_{\min} , particle number N_p , dynamic model $p(x_t|x_{t-1}; \theta)$, soft resampler $\mathcal{R}_\lambda(\{W_t^i\}_{i \in [N_p]})$ that outputs indices A_t^i of selected particles, spatial transformer network $\mathbf{ST}_\theta(\cdot)$, global 2D map \mathbb{M} , local 2D map generated by the spatial transformer network $M_{x_t^i}$, scalar function $h_\theta(\cdot)$ parameterised by neural networks, ground-truth robot location $x_t^* = [s_t^{(1)*}, s_t^{(2)*}, \eta_t^*]$, hyperparameters λ and ξ , learning rate ζ , $[N_p] := \{1, \dots, N_p\}$;

2: **Initialisation:** Sample $x_0^i \stackrel{\text{i.i.d.}}{\sim} \pi(x_0; \theta)$ for $\forall i \in [N_p]$;

3: Set weights $w_0^i = 1$ for $\forall i \in [N_p]$;

4: Normalise weights $W_0^i = w_0^i / \sum_{n=1}^{N_p} w_0^n$ for $\forall i \in [N_p]$;

5: Compute the effective sample size: $\text{ESS}_0(\{W_0^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_0^i)^2}$;

6: **for** $t = 1$ to T **do**

7: **if** $\text{ESS}_t(\{W_{t-1}^i\}_{i \in [N_p]}) < \text{ESS}_{\min}$ **then**

8: $\tilde{W}_{t-1}^i = \lambda W_{t-1}^i + (1 - \lambda) \frac{1}{N_p}$ (Eq. (47)).

9: $A_t^{1:N_p} \leftarrow \mathcal{R}(\{\tilde{W}_{t-1}^i\}_{i \in [N_p]})$, $\tilde{w}_{t-1}^i \leftarrow \frac{W_{t-1}^i}{\tilde{W}_{t-1}^i}$ for $\forall i \in [N_p]$ (Eq. (48));

10: **else**

11: $A_t^i \leftarrow i$, $\tilde{w}_{t-1}^i \leftarrow w_{t-1}^i$ for $\forall i \in [N_p]$;

12: **end if**

13: Sample $x_t^i \stackrel{\text{i.i.d.}}{\sim} p(x_t|x_{t-1}^{A_t^i}; \theta)$ for $\forall i \in [N_p]$ using Eqs. (70) – (72);

14: Generate local 2D maps : $M_{x_t^i} = \mathbf{ST}_\theta(\mathbb{M}, x_t^i)$;

15: Compute unnormalised conditional likelihood: $p(y_t|x_t^i; \theta) \propto h_\theta(y_t, M_{x_t^i})$;

16: Update weights $w_t^i = \tilde{w}_{t-1}^i h_\theta(y_t, M_{x_t^i})$;

17: Normalise weights: $W_t^i = w_t^i / \sum_{n=1}^{N_p} w_t^n$ for $\forall i \in [N_p]$;

18: Estimate robot location $\bar{x}_t := [\bar{s}_t^{(1)}, \bar{s}_t^{(2)}, \bar{\eta}_t]$: $\bar{x}_t = \sum_{i=1}^{N_p} W_t^i x_t^i$.

19: Compute the effective sample size: $\text{ESS}_t(\{W_t^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_t^i)^2}$;

20: **end for**

21: Loss function $\mathcal{L} = \sum_{t=0}^T (\bar{s}_t^{(1)} - s_t^{(1)*})^2 + (\bar{s}_t^{(2)} - s_t^{(2)*})^2 + \xi(\bar{\eta}_t - \eta_t^*)^2$;

22: Update θ by gradient descent: $\theta \leftarrow \theta - \zeta \nabla_\theta \mathcal{L}$.

7.2. Normalising flow-based differentiable particle filters. Differentiable particle filters proposed in [50] and [51] are constructed using normalising flows. In [50] normalising flows are used to construct dynamic models and proposal distributions, and in [51] normalising flows are used to construct measurement models. Measurement models built with conditional normalising flows can be trained with likelihood-based training objectives as they admit valid probability densities by construction. Additionally, the normalising flow-based differentiable particle filters presented in Algorithm 3 also provide a flexible mechanism to build complex dynamic models and proposal distributions. The pseudocode presented in Algorithm 3 provides the implementation details of a differentiable particle filter whose dynamic model, proposal distribution, and measurement model are built with (conditional) normalising flows. In Algorithm 3, the entropy-regularised OT resampler is employed, following the setup in [51].

Algorithm 3: Normalising flow-based differentiable particle filters [50, 51].

1: **Inputs:** Initial distribution $\pi(x_0; \theta)$, resampling threshold ESS_{\min} , particle number N_p , dynamic model $p(x_t|x_{t-1}; \theta)$, proposal distribution $q(x_t|y_t, x_{t-1}; \phi)$, base dynamic model $\tilde{p}(\tilde{x}_t|x_{t-1}; \theta)$, entropy-regularised OT resampler $\mathcal{R}_\xi(\{x_t^i\}_{i \in [N_p]}, \{W_t^i\}_{i \in [N_p]})$, dynamic model normalising flow $\mathcal{T}_\theta(\cdot)$, proposal normalising flow $\mathcal{G}_\phi(\cdot)$, measurement normalising flow $\mathcal{F}_\theta(\cdot)$, neural network $E_\theta(\cdot)$, learning rate ζ , $p_z(\cdot)$ probability density function of standard Gaussian distribution, hyperparameters ξ , $[N_p] := \{1, \dots, N_p\}$;

2: **Initialisation:** Sample $x_0^i \stackrel{\text{i.i.d.}}{\sim} \pi(x_0; \theta)$ for $\forall i \in [N_p]$;

3: Set weights $w_0^i = 1$ for $\forall i \in [N_p]$;

4: Normalise weights $W_0^i = w_0^i / \sum_{n=1}^{N_p} w_0^n$ for $\forall i \in [N_p]$;

5: Compute the effective sample size: $\text{ESS}_0(\{W_0^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_0^i)^2}$;

6: **for** $t = 1$ to T **do**

7: **if** $\text{ESS}_t(\{W_{t-1}^i\}_{i \in [N_p]}) < \text{ESS}_{\min}$ **then**

8: $\{x_{t-1}^i\}_{i=1}^N \leftarrow \mathcal{R}_\xi(\{x_{t-1}^i\}_{i \in [N_p]}, \{W_{t-1}^i\}_{i \in [N_p]})$, $\tilde{w}_{t-1}^i \leftarrow 1$, $\forall i \in [N_p]$;

9: **else**

10: $\tilde{w}_{t-1}^i \leftarrow w_{t-1}^i$, $\forall i \in [N_p]$;

11: **end if**

12: Sample from the dynamic model (Eq. (27)):
 $\check{x}_t^i = \mathcal{T}_\theta(\tilde{x}_t^i) \stackrel{\text{i.i.d.}}{\sim} p(\check{x}_t|x_{t-1}^i; \theta)$ where $\tilde{x}_t^i \stackrel{\text{i.i.d.}}{\sim} \tilde{p}(\tilde{x}_t|x_{t-1}^i; \theta)$, $\forall i \in [N_p]$;

13: Generate particles (Eq. (29)): $x_t^i = \mathcal{G}_\phi(\check{x}_t^i, y_t) \sim q(x_t|y_t, x_{t-1}^i; \phi)$, $\forall i \in [N_p]$;

14: Compute particle densities in the proposal distribution (Eq. (30)):
 $q(x_t^i|y_t, x_{t-1}^i; \phi) = \frac{p(x_t^i|x_{t-1}^i; \theta)}{|\det J_{\mathcal{G}_\phi}(\check{x}_t^i)|} = \frac{\tilde{p}(\tilde{x}_t^i|x_{t-1}^i; \theta)}{|\det J_{\mathcal{G}_\phi}(\check{x}_t^i)||\det J_{\mathcal{T}_\theta}(\tilde{x}_t^i)|}$, $\forall i \in [N_p]$;

15: Compute particle densities in the dynamic model (Eq. (28)):
 $p(x_t^i|x_{t-1}^i; \theta) = \frac{\tilde{p}(\mathcal{T}^{-1}(x_t^i)|x_{t-1}^i; \theta)}{|\det J_{\mathcal{T}_\theta}(\mathcal{T}^{-1}(x_t^i))|}$, $\forall i \in [N_p]$;

16: Compute conditional likelihood of observations given states (Eq. (44)):
 $p(y_t|x_t^i; \theta) = p_z(z_t^i) \left| \det \frac{\partial z_t^i}{\partial y_t} \right|$, with $z_t^i = \mathcal{F}_\theta(y_t, x_t^i)$, $\forall i \in [N_p]$;

17: Update weights: $w_t^i = \tilde{w}_{t-1}^i \frac{p(y_t|x_t^i; \theta)p(x_t^i|x_{t-1}^i; \theta)}{q(x_t^i|y_t, x_{t-1}^i; \phi)}$, $\forall i \in [N_p]$;

18: Normalise weights: $W_t^i = w_t^i / \sum_{n=1}^{N_p} w_t^n$ for $\forall i \in [N_p]$;

19: Compute the effective sample size: $\text{ESS}_t(\{W_t^i\}_{i \in [N_p]}) = \frac{1}{\sum_{i=1}^{N_p} (W_t^i)^2}$;

20: **end for**

21: Compute the overall loss function $\mathcal{L}(\theta, \phi)$;

22: Update θ and ϕ : $\theta \leftarrow \theta - \xi \nabla_\theta \mathcal{L}(\theta, \phi)$, $\phi \leftarrow \phi - \xi \nabla_\phi \mathcal{L}(\phi, \phi)$.

8. Discussion & Summary. In this paper, we reviewed recent advances in differentiable particle filters, with a particular focus on five core components of differentiable particle filters: dynamic models, measurement models, proposal distributions, differentiable resampling techniques, and training objectives. Specifically, the above modules of differentiable particle filters can be summarised as follows:

1. Dynamic models used in many differentiable particle filter frameworks are Gaussian, where the transition of states is simulated by a Gaussian distribution whose mean and variance are predicted by a function of previous states.

Normalising flows can be used to transform simple distributions, e.g. Gaussian, into arbitrarily complex distributions under mild conditions, and they also enable the evaluation of densities of transformed particles. Dynamic models in several works are built with observations, and have shown competitive performances in several examined scenarios. However, this class of dynamic models does not fit into the problem setup of state-space models we consider in Bayesian filtering problems.

2. For the proposal distribution of differentiable particle filters, we can employ conditional normalising flows to incorporate information from observations into the construction of proposal distributions, with a straightforward evaluation of the proposal density.
3. In cases where the relationship between states and observations can be built with known distribution families, measurement models are often differentiable by themselves. In more complex scenarios, differentiable particle filters resort to build connections between observations and states in a feature space. Typical examples include: (i) measurement models as scalar functions built with neural networks, where the conditional likelihood of observation is given by neural networks with observation and state features as inputs; (ii) measurement models as feature similarities, where the conditional likelihood of observation is evaluated as the similarity between the observation feature and the state feature; (iii) measurement models built with conditional normalising flows, where observations are transformed into Gaussian samples through invertible neural networks and the conditional likelihood of observation is computed through the change of variable formula.
4. Different techniques for differentiable resampling have been proposed to address differentiability of the resampling step. In the soft resampling method, the gradients of the importance weights before resampling w.r.t. previous weights are non-zero. Soft resampling is not truly differentiable as the non-differentiable part of resampling is ignored. A fully-differentiable resampling scheme is achieved by considering the resampling operation through solving entropy-regularised optimal transport problems. The regularisation term in the OT-resampler introduces biases into resampling outcomes, and the biases only vanish if the regularisation coefficient approaches zero. A transformer-based resampler, particle transformer, was proposed based on the recent development of the transformer architecture and self-attention mechanism. One limitation of the particle transformer is that a particle transformer trained for one specific task may not be suitable for other tasks.
5. The training objective of differentiable particle filters include supervised losses, semi-supervised losses, and variational objectives. Supervised losses require ground-truth latent states to train differentiable particle filters. One type of semi-supervised losses includes a pseudo-likelihood loss to optimise parameters of differentiable particle filters when ground-truth latent states are unavailable. Variational objectives adopt an evidence lower bound (ELBO) of the marginal log-likelihood of observations as a surrogate objective for training differentiable particle filters.

Compared with traditional particle filters, differentiable particle filters are particularly suitable for performing sequential Bayesian inference in complex environments where neither the structure nor the parameters of the system of interest are known.

In such cases, traditional particle filters require hand-crafted designs of the components of particle filters, which can be challenging, often rely on extensive domain knowledge, and can lead to significant delays in the deployment of the filtering algorithm. The trade-off is that differentiable particle filters have to be trained before they can produce reasonable filtering results and therefore are more computationally expensive to train than traditional approaches. Nonetheless, differentiable particle filters provide a data-adaptive framework that holds the promise to significantly accelerate the development and adoption of particle filters in real-world sequential Bayesian filtering applications, with many open research questions awaiting to be explored.

REFERENCES

- [1] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *J .Basic Eng.*, vol. 82, no. 1, pp. 33–45, 1960.
- [2] B. Anderson and J. B. Moore, “Optimal filtering,” *Prentice-Hall Inform. and Syst. Sci. Ser.*, 1979.
- [3] F. E. Daum, “Extended Kalman filters.” *Encyclo. Syst. and Control*, 2015.
- [4] E. A. Wan and R. Van Der Merwe, “The unscented kalman filter for nonlinear estimation,” in *Proc. IEEE Adapt. Syst. Signal. Process. Commun. Control Symp.*, Oct., Lake Louise, Canada 2000.
- [5] R. S. Bucy and K. D. Senne, “Digital synthesis of non-linear filters,” *Automatica*, vol. 7, no. 3, pp. 287–298, 1971.
- [6] A. Doucet, D. N. Freitas, and N. Gordon, “An introduction to sequential Monte Carlo methods,” in *Sequential Monte Carlo methods in practice*. Springer, 2001, pp. 3–14.
- [7] N. Gordon, D. Salmond, and A. Smith, “Novel approach to nonlinear/non-Gaussian Bayesian state estimation,” in *IEE Proc. F (Radar and Signal Process.)*, vol. 140, 1993, pp. 107–113.
- [8] P. M. Djurić, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez, “Particle filtering,” *IEEE Signal Process. Mag.*, vol. 20, no. 5, pp. 19–38, 2003.
- [9] A. Doucet, N. De Freitas, N. J. Gordon *et al.*, *Sequential Monte Carlo methods in practice*. Springer, 2001, vol. 1, no. 2.
- [10] P. Del Moral and L. Miclo, “Branching and interacting particle systems. approximations of Feynman-Kac formulae with applications to non-linear filtering,” *Sémin. Probab. Strasbourg*, vol. 34, pp. 1–145, 2000.
- [11] D. Crisan and A. Doucet, “A survey of convergence results on particle filtering methods for practitioners,” *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 736–746, 2002.
- [12] P. Del Moral, “Measure-valued processes and interacting particle systems. application to nonlinear filtering problems,” *Ann. Appl. Probab.*, vol. 8, no. 2, pp. 438–495, 1998.
- [13] 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, 2017.
- [14] V. Elvira, J. Míguez, and P. M. Djurić, “On the performance of particle filters with adaptive number of particles,” *Stat. Comput.*, vol. 31, pp. 1–18, 2021.
- [15] M. K. Pitt and N. Shephard, “Filtering via simulation: Auxiliary particle filters,” *J. Amer. Statist. Assoc.*, vol. 94, no. 446, pp. 590–599, 1999.
- [16] V. Elvira, L. Martino, M. F. Bugallo, and P. M. Djurić, “Elucidating the auxiliary particle filter via multiple importance sampling,” *IEEE Signal Process. Mag.*, vol. 36, no. 6, pp. 145–152, 2019.
- [17] N. Branchini and V. Elvira, “Optimized auxiliary particle filters: adapting mixture proposals via convex optimization,” in *Proc. Conf. Uncertain. Artif. Intell. (UAI)*, 2021, pp. 1289–1299.
- [18] J. H. Kotecha and P. M. Djurić, “Gaussian sum particle filtering,” *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2602–2612, 2003.
- [19] ———, “Gaussian sum particle filtering for dynamic state-space models,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Salt Lake City, USA, May 2001.
- [20] ———, “Gaussian particle filtering,” *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2592–2601, 2003.

- [21] R. Van Der Merwe, A. Doucet, N. De Freitas, and E. Wan, "The unscented particle filter," *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS)*, Dec. 2000.
- [22] Y. Rui and Y. Chen, "Better proposal distributions: Object tracking using unscented particle filter," in *IEEE Conf. Comp. Vis. and Pat. Recog. (CVPR)*. Kauai, USA.: IEEE, Dec. 2001.
- [23] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, 2004.
- [24] A. Doucet, N. de Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised particle filtering for dynamic Bayesian networks," in *Proc. Conf. Uncertain. Artif. Intell. (UAI)*, Stanford, USA, 2000, pp. 176–183.
- [25] N. De Freitas, "Rao-Blackwellised particle filtering for fault diagnosis," in *Proc. IEEE Aerosp. Conf.*, vol. 4, 2002, pp. 4–4.
- [26] P. M. Djurić, T. Lu, and M. F. Bugallo, "Multiple particle filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Honolulu, USA, Apr. 2007.
- [27] P. M. Djurić and M. F. Bugallo, "Particle filtering for high-dimensional systems," in *Proc. Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Saint Martin, France, Dec. 2013.
- [28] P. Bunch and S. Godsill, "Approximations of the optimal importance density using gaussian particle flow importance sampling," *J. Amer. Statist. Assoc.*, vol. 111, no. 514, pp. 748–762, 2016.
- [29] Y. Li and M. Coates, "Particle filtering with invertible particle flow," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4102–4116, 2017.
- [30] J. Heng, A. Doucet, and Y. Pokern, "Gibbs flow for approximate transport with applications to Bayesian computation," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 83, no. 1, pp. 156–187, 2021.
- [31] T. Zhang, C. Xu, and M.-H. Yang, "Multi-task correlation particle filter for robust object tracking," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recogn. (CVPR)*, Honolulu, Hawaii, July 2017.
- [32] S. Malik and M. K. Pitt, "Particle filters for continuous likelihood evaluation and maximisation," *J. Econometrics*, vol. 165, no. 2, pp. 190–209, 2011.
- [33] A. Gunatilake, S. Kodagoda, and K. Thiagarajan, "A novel UHF-RFID dual antenna signals combined with Gaussian process and particle filter for in-pipe robot localization," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6005–6011, 2022.
- [34] P. J. Van Leeuwen, H. R. Kiensch, L. Nerger, R. Potthast, and S. Reich, "Particle filters for high-dimensional geoscience applications: A review," *Q. J. R. Meteorol. Soc.*, vol. 145, no. 723, pp. 2335–2365, 2019.
- [35] C. Pozna, R.-E. Precup, E. Horváth, and E. M. Petriu, "Hybrid particle filter–particle swarm optimization algorithm and application to fuzzy controlled servo systems," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 10, pp. 4286–4297, 2022.
- [36] N. Kantas, A. Doucet, S. S. Singh, and J. M. Maciejowski, "An overview of sequential Monte Carlo methods for parameter estimation in general state-space models," *IFAC Proc. Vol.*, vol. 42, no. 10, pp. 774–785, 2009.
- [37] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin, "On particle methods for parameter estimation in state-space models," *Stat. Sci.*, vol. 30, no. 3, pp. 328–351, 2015.
- [38] C. Andrieu, A. Doucet, and R. Holenstein, "Particle Markov chain Monte Carlo methods," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 72, no. 3, pp. 269–342, 2010.
- [39] F. Lindsten, M. I. Jordan, and T. B. Schon, "Particle gibbs with ancestor sampling," *J. Mach. Learn. Res.*, vol. 15, pp. 2145–2184, 2014.
- [40] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos, "SMC2: an efficient algorithm for sequential analysis of state space models," *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, vol. 75, no. 3, pp. 397–426, 2013.
- [41] S. Pérez-Vieites, I. P. Mariño, and J. Míguez, "Probabilistic scheme for joint parameter estimation and state prediction in complex dynamical systems," *Phys. Rev. E*, vol. 98, no. 6, p. 063305, 2018.
- [42] D. Crisan and J. Míguez, "Nested particle filters for online parameter estimation in discrete-time state-space Markov models," *Bernoulli*, vol. 24, no. 4A, pp. 3039–3086, 2018.
- [43] S. Pérez-Vieites and J. Míguez, "Nested Gaussian filters for recursive Bayesian inference and nonlinear tracking in state space models," *Signal Proces.*, vol. 189, p. 108295, 2021.

- [44] E. Chouzenoux and V. Elvira, “GraphEM: EM algorithm for blind Kalman filtering under graphical sparsity constraints,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2020.
- [45] V. Elvira and É. Chouzenoux, “Graphical inference in linear-Gaussian state-space models,” *IEEE Trans. Signal Process.*, vol. 70, pp. 4757–4771, 2022.
- [46] E. Chouzenoux and V. Elvira, “Sparse graphical linear dynamical systems,” *arXiv preprint arXiv:2307.03210*, 2023.
- [47] R. Jonschkowski, D. Rastogi, and O. Brock, “Differentiable particle filters: end-to-end learning with algorithmic priors,” in *Proc. Robot.: Sci. and Syst. (RSS)*, Pittsburgh, Pennsylvania, July 2018.
- [48] P. Karkus, D. Hsu, and W. S. Lee, “Particle filter networks with application to visual localization,” in *Proc. Conf. Robot Learn. (CoRL)*, Zurich, Switzerland, Oct 2018.
- [49] H. Wen, X. Chen, G. Papagiannis, C. Hu, and Y. Li, “End-to-end semi-supervised learning for differentiable particle filters,” in *Proc. IEEE Int. Conf. Robot. Automat., (ICRA)*, Xi'an, China, May 2021.
- [50] X. Chen, H. Wen, and Y. Li, “Differentiable particle filters through conditional normalizing flow,” in *Proc. IEEE Int. Conf. Inf. Fusion (FUSION)*, Sun City, South Africa, Nov. 2021.
- [51] X. Chen and Y. Li, “Conditional measurement density estimation in sequential Monte Carlo via normalizing flow,” in *Proc. Euro. Sig. Process. Conf. (EUSIPCO)*, Belgrade, Serbia, Aug. 2022.
- [52] A. Corenflos, J. Thornton, G. Deligiannidis, and A. Doucet, “Differentiable particle filtering via entropy-regularized optimal transport,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, July 2021.
- [53] M. Zhu, K. Murphy, and R. Jonschkowski, “Towards differentiable resampling,” *arXiv preprint arXiv:2004.11938*, 2020.
- [54] V. Elvira and L. Martino, “Advances in importance sampling,” *Wiley StatsRef-Statistics Reference Online*, pp. 1–14, 2021.
- [55] T. Hesterberg, “Weighted average importance sampling and defensive mixture distributions,” *Technometrics*, vol. 37, no. 2, pp. 185–194, 1995.
- [56] S. T. Tokdar and R. E. Kass, “Importance sampling: a review,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 1, pp. 54–60, 2010.
- [57] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, “Importance sampling: Intrinsic dimension and computational cost,” *Statist. Sci.*, pp. 405–431, 2017.
- [58] P. Bickel, B. Li, and T. Bengtsson, “Sharp failure rates for the bootstrap particle filter in high dimensions,” *Pushing the limits of contemporary statistics: Contributions in honor of Jayanta K. Ghosh*, 2008.
- [59] N. Chopin and O. Papaspiliopoulos, *An introduction to sequential Monte Carlo*. Springer, 2020.
- [60] T. Li, M. Bolic, and P. M. Djuric, “Resampling methods for particle filtering: classification, implementation, and strategies,” *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 70–86, 2015.
- [61] M. Gerber, N. Chopin, and N. Whiteley, “Negative association, ordering and convergence of resampling methods,” *Ann. Statist.*, vol. 47, no. 4, pp. 2236–2260, 2019.
- [62] T.-c. Li, G. Villarrubia, S.-d. Sun, J. M. Corchado, and J. Bajo, “Resampling methods for particle filtering: identical distribution, a new method, and comparable study,” *Frontiers Inform. Tech. & Electron. Eng.*, vol. 16, no. 11, pp. 969–984, 2015.
- [63] R. Douc and O. Cappé, “Comparison of resampling schemes for particle filtering,” in *Proc. Int. Symp. Image and Signal Process. and Anal.*, Zagreb, Croatia, 2005.
- [64] M. Bolić, P. M. Djurić, and S. Hong, “Resampling algorithms for particle filters: A computational complexity perspective,” *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 15, pp. 1–11, 2004.
- [65] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for bayesian filtering,” *Stat. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
- [66] V. Elvira, L. Martino, and C. P. Robert, “Rethinking the effective sample size,” *Int. Stat. Rev.*, vol. 90, no. 3, pp. 525–550, 2022.
- [67] L. Martino, V. Elvira, and F. Louzada, “Effective sample size for importance sampling based on discrepancy measures,” *Signal Process.*, vol. 131, pp. 386–401, 2017.
- [68] J. H. Huggins and D. M. Roy, “Convergence of sequential Monte Carlo based sampling methods,” *arXiv preprint arXiv:1503.00966*, 2015.

- [69] Z. Li, P. Fan, and Y. Dong, “Flexible effective sample size based on the message importance measure,” *IEEE Open J. Signal Process.*, vol. 1, pp. 216–229, 2020.
- [70] A. Doucet and V. B. Tadić, “Parameter estimation in general state-space models using particle methods,” *Ann. Inst. Stat. Math.*, vol. 55, no. 2, pp. 409–422, 2003.
- [71] C. Andrieu, A. Doucet, and V. B. Tadic, “On-line parameter estimation in general state-space models,” in *Proc. IEEE Conf. Dec. and Contr. (CDC)*, Seville, Spain, Dec. 2005.
- [72] A. Kloss, G. Martius, and J. Bohg, “How to train your differentiable filter,” *Auto. Robot.*, vol. 45, no. 4, pp. 561–578, 2021.
- [73] C. Rosato, L. Devlin, V. Beraud, P. Horridge, T. B. Schön, and S. Maskell, “Efficient learning of the parameters of non-linear models using differentiable resampling in particle filters,” *IEEE Trans. Signal Process.*, vol. 70, pp. 3676–3692, 2022.
- [74] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Mach. Learn.*, vol. 8, no. 3-4, pp. 229–256, 1992.
- [75] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Scottsdale, Arizona, May 2013.
- [76] S. Reich, “A nonparametric ensemble transform method for Bayesian inference,” *SIAM J. Sci. Comput.*, vol. 35, no. 4, pp. A2013–A2024, 2013.
- [77] M. Cuturi, “Sinkhorn distances: Lightspeed computation of optimal transport,” in *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS)*, Lake Tahoe, USA, Dec. 2013.
- [78] J. Feydy *et al.*, “Interpolating between optimal transport and mmd using Sinkhorn divergences,” in *Proc. Int. Conf. Artif. Intell. Stat. (AISTAS)*, Naha, Japan, Apr. 2019.
- [79] G. Peyré and M. Cuturi, “Computational optimal transport,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.
- [80] J. Lee *et al.*, “Set transformer: A framework for attention-based permutation-invariant neural networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Baltimore, USA, June 2019.
- [81] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS)*, Long Beach, USA, Dec. 2017.
- [82] X. Ma, P. Karkus, D. Hsu, and W. S. Lee, “Particle filter recurrent neural networks,” in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, New York, USA, Feb. 2020.
- [83] X. Ma *et al.*, “Discriminative particle filter reinforcement learning for complex partial observations,” in *Proc. Int. Conf. Learn. Rep. (ICLR)*, New Orleans, USA, May 2019.
- [84] R. Chen, H. Yin, Y. Jiao, G. Dissanayake, Y. Wang, and R. Xiong, “Deep samplable observation model for global localization and kidnapping,” *IEEE Robot. Autom. Lett. (RAL)*, vol. 6, no. 2, pp. 2296–2303, 2021.
- [85] P. Karkus, S. Cai, and D. Hsu, “Differentiable slam-net: Learning particle slam for visual navigation,” in *Proc. IEEE Conf. Comp. Vis. and Pat. Recog. (CVPR)*, June 2021.
- [86] P. Karkus *et al.*, “Differentiable algorithm networks for composable robot learning,” in *Proc. Robot.: Sci. and Syst. (RSS)*, Messe Freiburg, Germany, June 2019.
- [87] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, “Multimodal sensor fusion with differentiable filters,” in *Proc. IEEE/RSJ Int. Conf. Intel. Robot. Sys. (IROS)*, Las Vegas, USA, Oct. 2020.
- [88] C. Naesseth, S. Linderman, R. Ranganath, and D. Blei, “Variational sequential Monte Carlo,” in *Proc. Int. Conf. Artif. Intel. and Stat. (AISTATS)*, Playa Blanca, Spain, Apr. 2018.
- [89] T. A. Le, M. Igl, T. Rainforth, T. Jin, and F. Wood, “Auto-encoding sequential Monte Carlo,” in *Proc. Int. Conf. Learn. Rep. (ICLR)*, Vancouver, Canada, Apr. 2018.
- [90] C. J. Maddison *et al.*, “Filtering variational objectives,” in *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS)*, Long Beach, USA, Dec. 2017.
- [91] M. H. Dupty, Y. Dong, and W. S. Lee, “PF-GNN: Differentiable particle filtering based approximation of universal graph representations,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021.
- [92] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, “Normalizing flows for probabilistic modeling and inference,” *J. Mach. Learn. Res.*, vol. 22, pp. 1–64, 2021.
- [93] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [94] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” *Syntax, Semant. and Struct. in Stat. Transl.*, p. 103, 2014.

- [95] 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. (ICASSP)*, Singapore, May 2022, pp. 5498–5502.
- [96] F. Gama, N. Zilberstein, M. Sevilla, R. Baraniuk, and S. Segarra, “Unsupervised learning of sampling distributions for particle filters,” *arXiv preprint arXiv:2302.01174*, 2023.
- [97] C. Winkler, D. Worrall, E. Hoogeboom, and M. Welling, “Learning likelihoods with conditional normalizing flows,” *arXiv preprint arXiv:1912.00042*, 2019.
- [98] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” in *Proc. Adv. Neur. Inf. Process. Sys. (NeurIPS)*, Montreal, Canada, Dec. 2015.
- [99] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003.
- [100] ———, *Optimal Transport: old and new*. Berlin, Germany: Springer Science & Business Media, 2008, vol. 338.
- [101] C. Andrieu, A. Doucet, S. Singh, and V. Tadic, “Particle methods for change detection, system identification, and control,” *Proc. IEEE*, vol. 92, no. 3, pp. 423–438, 2004.
- [102] D. Rezende and S. Mohamed, “Variational inference with normalizing flows,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, July 2015.
- [103] Y. Burda, R. B. Grosse, and R. Salakhutdinov, “Importance weighted autoencoders,” in *Proc. Int. Conf. Learn. Rep. (ICLR)*, San Juan, Puerto Rico, May. 2016.
- [104] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, “An introduction to variational methods for graphical models,” *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.
- [105] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London, University College London, United Kingdom, 2003.
- [106] D. J. Rezende, S. Mohamed, and D. Wierstra, “Stochastic backpropagation and approximate inference in deep generative models,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, June 2014.