Generalised Bayesian Filtering via Sequential Monte Carlo

Anonymous Author(s)
Affiliation
Address
email

Abstract

We introduce a framework for inference in general state-space hidden Markov models (HMMs) under likelihood misspecification. In particular, we leverage the loss-theoretic perspective of Generalized Bayesian Inference (GBI) to define generalised filtering recursions in HMMs, that can tackle the problem of inference under model misspecification. In doing so, we arrive at principled procedures for robust inference against observation contamination by utilising the β-divergence. Operationalising the proposed framework is made possible via sequential Monte Carlo methods (SMC), where most standard particle methods, and their associated convergence results, are readily adapted to the new setting. We apply our approach to object tracking and Gaussian process regression problems, and observe improved performance over both standard filtering algorithms and other robust filters.

1 Introduction

Estimating the hidden states in dynamical systems is a long-standing problem in many fields of science and engineering. This can be formulated as an inference problem of a general state-space hidden Markov model (HMM) defined via two processes, *the hidden process* $(\mathbf{x}_t)_{t\geq 0}$, and *the observation process* $(\mathbf{y}_t)_{t\geq 1}$. More precisely, we consider the general state-space hidden Markov models of the form

$$\mathbf{x}_0 \sim \pi_0(\mathbf{x}_0), \quad (1) \qquad \mathbf{x}_t|\mathbf{x}_{t-1} \sim f_t(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad (2) \qquad \mathbf{y}_t|\mathbf{x}_t \sim g_t(\mathbf{y}_t|\mathbf{x}_t), \quad (3)$$

where $\mathbf{x}_t \in \mathsf{X}$ for $t \geq 0$, $y_t \in \mathsf{Y}$ for $t \geq 1$, f_t is a Markov kernel on X and $g_t : \mathsf{Y} \times \mathsf{X} \to \mathbb{R}_+$ is the likelihood function. We assume $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\mathsf{Y} \subseteq \mathbb{R}^{d_y}$ for convenience; however, the extension to general Polish spaces follows directly. The key inference problem in this model class is estimating is the *filtering distributions*, i.e. the posterior distributions of the hidden states $(\mathbf{x}_t)_{t\geq 0}$ given the observations $\mathbf{y}_{1:t}$ denoted as $(\pi_t(\mathbf{x}_t|\mathbf{y}_{1:t}))_{t\geq 1}$ — commonly known as *Bayesian filtering* [1, 2].

Under assumptions of linearity and Gaussianity, the inference problem for the hidden states of HMMs can be solved analytically via the Kalman filter [3]. However, inference for general HMMs of the form (1)–(3) with nonlinear, non-Gaussian transitions and likelihoods lacked a general, principled solution until the arrival of the particle filtering schemes [4]. Particle filters (PFs) have become ubiquitous for Bayesian filtering in the general setting. In short, the PFs retain a weighted collection of Monte Carlo samples representing the filtering distribution $\pi_t(\mathbf{x}_t|\mathbf{y}_{1:t})$ and recursively approximate the sequence of distributions $(\pi_t)_{t\geq 0}$ using a particle mutation-selection scheme [5].

While PFs (and other inference schemes for HMMs) implicitly assume that the assumed model is well-specified, it is important to consider whether the proposed model class includes the true data-generating mechanism (DGM). In particular, for general state-space HMMs, misspecification can occur if the true dynamics of the hidden process significantly differ from the assumed model

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

f_t, or if the true observation model is markedly different from the assumed likelihood model g_t, e.g. corruption by heavy tailed noise. The latter case is of widespread interest within the field of *robust statistics* [6] and has recently attracted significant interest in the machine learning community [7]. It is the setting that this paper seeks to address.

When the true DGM cannot be modelled, one principled approach to address misspecification is Generalized Bayesian Inference (GBI) [8]. This approach views classical Bayesian inference as a loss minimisation procedure in the space of probability measures, a view first developed by [9]. In particular, the standard Bayesian update can be derived from this view, where a loss function is constructed using the Kullback-Leibler (KL) divergence from the empirical distribution of the observations to the assumed likelihood [8]. The KL divergence is sensitive to outliers [10], hence the overall inference procedure is not robust to observations that are incompatible with the assumed model. A principled remedy is to replace the KL divergence with alternative discrepancy, such as the β-divergence, which makes the overall procedure more robust [11] while retaining interpretability.

Previous work on robust particle filters have been done for handling outliers, sensor failures and misspecification of the transition model [12, 13, 14, 15, 16, 17, 18, 19]. However, these approaches are either based on problem-specific heuristic outlier detection schemes, or make strong assumptions about the DGM in order to justify the use of heavy-tailed distributions [15]. This requires knowledge of the contamination mechanism that is implicitly embedded in the likelihood.

In this work we propose a principled approach to robust filtering that does not impose additional modelling assumptions. We adapt the GBI approach of [8] to the Bayesian filtering setting and develop sequential Monte Carlo (SMC) methods for inference. We illustrate the performance of this approach, using the β-divergence, to mitigate the effect of outliers. We show that this approach significantly improves the PF performance in settings with contaminated data, while retaining a general and principled approach to inference. We provide empirical results that demonstrate improvement over Kalman and particle filters for both linear and non-linear HMMs. We further provide comparisons with various robust schemes against heavy-tailed noise, including t-based likelihoods [15] or auxiliary particle filters (APFs) [12]. Finally, exploiting the state-space representations of Gaussian processes (GPs) [20], we demonstrate our framework on London air pollution data using robust GP regression which has linear time-complexity in the number of observations.

Notation. We denote the space of bounded, Borel measurable functions on X as $B(\mathsf{X})$. We denote the Dirac measure located at y as $\delta_{\mathbf{y}}(\mathrm{d}\mathbf{x})$ and note that $f(\mathbf{y}) = \int f(\mathbf{x})\delta_{\mathbf{y}}(\mathrm{d}\mathbf{x})$ for $f \in B(\mathsf{X})$. We denote the Borel subsets of X as $\mathcal{B}(\mathsf{X})$ and the set of probability measures on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ as $\mathcal{P}(\mathsf{X})$. For a probability measure $\mu \in \mathcal{P}(\mathsf{X})$ and $\varphi \in B(\mathsf{X})$, we write $\mu(\varphi) := \int \varphi(\mathbf{x})\mu(\mathrm{d}\mathbf{x})$. Given a probability measure μ, we abuse the notation denoting its density with respect to the Lebesgue measure as $\mu(\mathbf{x})$.

2 Background

2.1 Generalized Bayesian Inference (GBI)

Bayesian inference implicitly assumes that the generative model is well-specified, in particular, the observations are generated from the assumed likelihood model. When this assumption is not expected to hold in real-world scenarios, one may wish to take into account the discrepancy between the true DGM and the assumed likelihood. GBI [8] is an approach to deal with such cases. Here, we present the main idea of GBI and refer the reader to the appendix for a more detailed description and to the original reference for a full-treatment.

For the simple Bayesian updating setup, consider a prior π_0 and the assumed likelihood function $g(\mathbf{y}|\mathbf{x})$. The posterior $\pi(\mathbf{x}|\mathbf{y}) =: \pi(\mathbf{x})$ is given by Bayes rule $\pi(\mathbf{x}) = \pi_0(\mathbf{x})\frac{g(\mathbf{y}|\mathbf{x})}{Z}$, where $Z := \int g(\mathbf{y}|\mathbf{x})\pi_0(\mathbf{x})\mathrm{d}\mathbf{x}$. [9] and [8] showed that this update can be seen as a special case of a more general update rule, which can be described as a solution of an optimisation problem in the space of measures. This leads to a more general belief updating rule given by

$$\pi(\mathbf{x}) = \pi_0(\mathbf{x})\frac{G(\mathbf{y}|\mathbf{x})}{Z}, \tag{4}$$

with $G(\mathbf{y}|\mathbf{x}) := \exp(-\ell(\mathbf{x}, \mathbf{y}))$ where $\ell(\mathbf{x}, \mathbf{y})$ is a loss function connecting the observations to the model parameters. Specifying $\ell(\mathbf{x}, \mathbf{y})$ as the cross-entropy (from the KL-divergence) of the assumed likelihood relative to the empirical distribution of the data recovers the standard Bayes update.

As noted before, the standard Bayes update is not robust to outliers due to the properties of KL divergence [10]. Hence, substituting the cross-entropy with a more robust loss such as the β-cross-entropy [7], based on the β-divergence, can make the inference more robust. Specifically, in this setting the generalised Bayes update for the likelihood $g(\mathbf{y}|\mathbf{x})$ is written as $\pi(\mathbf{x}) = \pi_0(\mathbf{x})\frac{G^\beta(\mathbf{y}|\mathbf{x})}{Z_\beta}$, where

$$G^\beta(\mathbf{y}|\mathbf{x}) = \exp\left(\frac{1}{\beta}g(\mathbf{y}|\mathbf{x})^\beta - \frac{1}{\beta+1}\int g(\mathbf{y}'|\mathbf{x})^{\beta+1}\mathrm{d}\mathbf{y}'\right). \tag{5}$$

One can consider $G^\beta(\mathbf{y}|\mathbf{x})$ as a generalised likelihood, resulting from the use of a different loss function compared to the standard Bayes procedure. Here β is a hyperparameter that needs to be selected depending on the degree of misspecification. In general $\beta \in (0, 1)$ and $\lim_{\beta\to 0} G^\beta(\mathbf{y}|\mathbf{x}) = g(\mathbf{y}|\mathbf{x})$. Thus, intuitively, small β values are suitable for mild model misspecification and large β values are suitable when the assumed model is expected to significantly deviate from the true model. In the experimental section, we devote some attention to the selection of β and sensitivity analysis.

Generalised Bayesian updating is more robust against outliers if a suitable divergence is chosen [21, 22, 10]. We note that GBI is conceptually different from approximate Bayesian methods with alternative divergences such as [23, 24, 25, 26]. While these methods target approximate posteriors that minimise some discrepancy from the true posterior and are not necessarily robust, GBI methods change the inference target from the standard Bayesian posterior (obtained by setting $\ell(\mathbf{x}, \mathbf{y})$ to the KL divergence) to a different target distribution with more desirable properties such as robustness to outliers. Later, we demonstrate how the GBI approach can be used to construct robust PF procedures.

2.2 Sequential Monte Carlo for HMMs

Let $\mathbf{x}_{1:T}$ be a hidden process with $\mathbf{x}_t \in \mathsf{X}$ and $\mathbf{y}_{1:T}$ an observation process with $\mathbf{y}_t \in \mathsf{Y}$. Our goal is to conduct inference in HMMs of the form (1)–(3) where $\pi_0(\cdot)$ is a prior probability distribution on the initial state $\mathbf{x}_0$, $f_t(\mathbf{x}|\mathbf{x}')$ is a Markov transition kernel on X and $g_t(\mathbf{y}_t|\mathbf{x}_t)$ is the likelihood for observation $\mathbf{y}_t$. The observation sequence $\mathbf{y}_{1:T}$ is assumed to be fixed but otherwise arbitrary.

The typical interest in probabilistic models is the estimation of expectations of general test functions with respect to the posterior distribution, in this case, of the hidden process $\pi_t(\mathbf{x}_t|\mathbf{y}_{1:t})$ and the associated joint distributions $\mathsf{p}_t(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$. More precisely, given a bounded test function $\varphi \in B(\mathsf{X})$, we are interested in estimating integrals of the form

$$\pi_t(\varphi) = \int \varphi(\mathbf{x}_t)\pi_t(\mathbf{x}_t|\mathbf{y}_{1:t}). \tag{6}$$

Kalman filtering [3, 1] can be used to obtain closed form expressions for $(\pi_t, \mathsf{p}_t)_{t\geq 0}$ if f_t and g_t are linear-Gaussian. However, for non-linear or non-Gaussian cases, the target distributions are almost always intractable, requiring an alternative approach, such as SMC methods [5, 27]. Known as Particle Filters (PFs) when employed in the HMM setting, SMC methods combine importance sampling and resampling algorithms tailored to approximate the solution of the filtering and smoothing problems.

In a typical iteration, a PF method proceeds as follows: given a collection of samples $\{\mathbf{x}_{t-1}^{(i)}\}_{i=1}^N$ representing the posterior $\pi_{t-1}(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$, it first samples from a (possibly observation dependent) proposal $\bar{\mathbf{x}}_t^{(i)} \sim q_t(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_{1:t})$. It then computes weights for each sample (particle) $\bar{\mathbf{x}}_{t-1}^{(i)}$ in the collection for a given observation $\mathbf{y}_t$, evaluating its fitness with respect to the likelihood g_t as $\mathsf{w}_t^{(i)} \propto g_t(\mathbf{y}_t|\bar{\mathbf{x}}_t^{(i)})\frac{f_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_t)}$, where $\sum_{i=1}^N \mathsf{w}_t^{(i)} = 1$. Finally, an optional resampling step [1] is used to prevent degeneracy, leading to $\mathbf{x}_t^{(i)} \sim \sum_{i=1}^N \mathsf{w}_t^{(i)}\delta_{\bar{\mathbf{x}}_t^{(i)}}(\mathrm{d}\mathbf{x}_t)$. One can then construct the empirical measure $\pi_t^N(\mathrm{d}\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_t^{(i)}}(\mathrm{d}\mathbf{x}_t)$, and the estimate of $\pi_t(\varphi)$ in (6) is given by

$$\pi_t^N(\varphi) = \frac{1}{N}\sum_{i=1}^N \varphi(\mathbf{x}_t^{(i)}). \tag{7}$$

[1] In the simplest form, drawing N times with replacement from the weighted empirical measure to obtain an unweighted sample whose empirical distribution approximates the same target; see [28] for an overview of resampling schemes and their properties.

Algorithm 1 The generalised particle filter

Input: Observation sequence $\mathbf{y}_{1:T}$, number of samples N, proposal distributions $q_{1:T}(\cdot)$.
Initialize: Sample $\{\bar{\mathbf{x}}_0^{(i)}\}_{i=1}^N$ for the prior $\pi_0(\mathbf{x}_0)$.
for $t = 1$ **to** T **do**
  Sample: $\bar{\mathbf{x}}_t^{(i)} \sim q_t(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_t), \quad \text{for} \quad i = 1, \dots, N.$
  Weight: $\mathsf{w}_t^{(i)} \propto \exp(-\ell(\bar{\mathbf{x}}_t^{(i)}, \mathbf{y}_t)) \frac{f_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_t)}, \quad \text{for} \quad i = 1, \dots, N.$
  Resample: $\mathbf{x}_t^{(i)} \sim \sum_{i=1}^N \mathsf{w}_t^{(i)} \delta_{\bar{\mathbf{x}}_t^{(i)}}(\mathrm{d}\mathbf{x}_t), \quad \text{for} \quad i = 1, \dots, N.$
end for

If the proposal is chosen as the transition density, i.e., $q_t(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_t) = f_t(\mathbf{x}_t|\mathbf{x}_{t-1}^{(i)})$, we obtain the bootstrap particle filter (BPF) [4]. This corresponds to the simple procedure of sampling $\bar{\mathbf{x}}_t^{(i)}$ from $f_t(\mathbf{x}_t|\mathbf{x}_{t-1}^{(i)})$, and setting its weight $\mathsf{w}_t^{(i)} \propto g_t(\mathbf{y}_t|\bar{\mathbf{x}}_t^{(i)})$.

3 Generalised Bayesian filtering

3.1 A simple generalised particle filter

As explained in Section 2.1, given a standard probability model comprised of the prior $\pi_0(\mathbf{x})$ and a likelihood $g(\mathbf{y}|\mathbf{x})$, the general Bayes update defines an alternative, generalised likelihood $G(\mathbf{y}|\mathbf{x})$. The sequence of generalised likelihoods, denoted as $G_t(\mathbf{y}_t|\mathbf{x}_t)$ for $t \geq 1$, in an HMM yields a joint generalised posterior density which factorises as

$$\mathsf{p}_t(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \propto \pi_0(\mathbf{x}_0) \prod_{k=1}^t f_k(\mathbf{x}_k|\mathbf{x}_{k-1}) G_k(\mathbf{y}_k|\mathbf{x}_k), \tag{8}$$

where $G_t(\mathbf{y}_t|\mathbf{x}_t) := \exp(-\ell_t(\mathbf{x}_t, \mathbf{y}_t))$. Inference can be done via SMC applied to this sequence of twisted probabilities defining a Feynman-Kac flow in the terminology of [29].

Comparing the update rule in (4) to the standard Bayes update suggests a generalisation of the particle filter. In particular, under the model in (1)–(3), one can perform generalised inference using $(f_t)_{t\geq 1}$ as usual, but replacing the likelihood with $(G_t)_{t\geq 1}$. Hence, a generalised sequential importance resampling PF (given fully in Algorithm 1) keeps the sampling step intact, but applies a different weight computation step $\mathsf{w}_t^{(i)} \propto \exp(-\ell(\bar{\mathbf{x}}_t^{(i)}, \mathbf{y}_t)) \frac{f_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q_t(\bar{\mathbf{x}}_t^{(i)}|\mathbf{x}_{1:t-1}^{(i)}, \mathbf{y}_t)}$. Indeed, most PFs (including the APF, see Algorithm 3 in the appendix) and related algorithms can be adapted to the GBI context.

3.2 The β-BPF and the β-APF

The β-BPF is derived by selecting $\ell_t(\mathbf{x}_t, \mathbf{y}_t)$ as the β-divergence and applying the BPF procedure with the associated generalised likelihood. In this case, the loss is

$$\ell_t^\beta(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{\beta+1} \int g_t(\mathbf{y}_t'|\mathbf{x}_t)^{\beta+1} d\mathbf{y}_t' - \frac{1}{\beta} g_t(\mathbf{y}_t|\mathbf{x}_t)^\beta. \tag{9}$$

We can then construct the general β-likelihood as

$$G_t^\beta(\mathbf{y}_t|\mathbf{x}_t) \propto \exp(-\ell_t^\beta(\mathbf{x}_t, \mathbf{y}_t)). \tag{10}$$

In this instance, the use of the β-divergence provides the sampler with robust properties [11]. This can informally be seen from the form of the loss function in (9), where small values of β temper the likelihood extending its tails making the loss more forgiving to outliers. The β-BPF procedure is given in Algorithm 2 in the appendix. The β-APF (Algorithm 3 in the appendix) is an Auxiliary Particle Filter [12, 30] adapted to the GBI setting, and is derived similarly to the β-BPF.

Note that the integral term in (9) is independent of $\mathbf{x}_t$ and can be absorbed, without evaluation, into the normalising constant when $\mathbf{x}_t$ is a location parameter for a symmetric $g_t(\cdot)$ and Y is a linear subspace of $\mathbb{R}^{d_y}$. More generally, if $g_t(\cdot)$ is a member of the exponential family, the integral can be computed by identifying $g_t^\beta(\cdot)$ with the kernel of another member of the same family with canonical parameters scaled by β. The overhead of computing $G_t^\beta(\cdot)$ is negligible in this instance, which is not too restrictive in the context of misspecitfied models. For other likelihoods, unbiased estimators for $G_t^\beta(\cdot)$, e.g. Poisson estimator [31], can be used in a random weight particle filter framework [32], where the overhead of computing $G_t^\beta(\cdot)$ will depend on the variance of the estimator and the convergence results from this setting apply but as [32] demonstrate this cost need not be prohibitive.

3.3 Selecting β

It is often the case that the primary goal of inference, particularly in the presence of model misspecification, is prediction. Hence, we propose choosing divergence parameters that lead to maximally predictive posterior belief distributions. In particular, for the β-BPF and β-APF, define $\mathcal{L}_\beta(\mathbf{y}_t, \hat{\mathbf{y}}_t)$ as a loss function of the observations $\mathbf{y}_t$ and the predictions $\hat{\mathbf{y}}_t$. We propose to choose β as the solution to the following decision-theoretic optimisation problem:

$$\min_\beta \mathsf{agg}_{t=1}^{T}(\mathbb{E}_{p(\hat{\mathbf{y}}_t|\mathbf{y}_{1:t-1})}\mathcal{L}_\beta(\mathbf{y}_t, \hat{\mathbf{y}}_t)), \tag{11}$$

where agg denotes an aggregating function. This approach requires some training data to allow the selection of β. In filtering contexts, this can be historical data from the same setting or other available proxies. For offline inference one could also employ the actual data within this framework. Since, this proposal relies on the quality of the observations, which in the case of outlier contamination is violated by definition. To remedy this, we propose choosing robust versions for agg and $\mathcal{L}$, e.g. the median and the (standardised) absolute error respectively.

4 Theoretical guarantees

Theoretical guarantees for SMC methods can be extended to the generalised Bayesian filtering setting. Since the generalised Bayesian filters can be seen as a standard SMC methods with modified likelihoods, the same analytical tools can be used in this setting. We provide guarantees for the β-BPF but emphasise that essentially the same results can be obtained much more broadly (including for the β-APF via the approach of [30]). We denote the generalised filters and generalised posteriors for the HMM in the β-divergence setting as π_t^β and p_t^β respectively. Consequently, corresponding quantities constructed by the β-BPF are denoted as $\pi_t^{\beta,N}$ and $\mathsf{p}_t^{\beta,N}$.

Although the generalised likelihoods $G_t^\beta(\mathbf{y}_t|\mathbf{x}_t)$ are not normalised, they can be considered as potential functions [29]. Since $G_t^\beta(\mathbf{y}_t|\mathbf{x}_t) < \infty$ whenever $g_t(\mathbf{y}_t|\mathbf{x}_t) < \infty$ and β is fixed, we can adapt the standard convergence results into the generalised case.

Assumption 1. *For a fixed arbitrary observation sequence $\mathbf{y}_{1:T} \in \mathsf{Y}^{\otimes T}$, the potential functions $(G_t^\beta)_{t\geq 1}$ are bounded and $G_t^\beta(\mathbf{y}_t|\mathbf{x}_t) > 0, \quad \forall t \in \{1,\dots,T\}$ and $\mathbf{x}_t \in \mathsf{X}$.*

This assumption holds for most used likelihood functions and their generalised extensions.

Theorem 1. *For any $\varphi \in B(\mathsf{X})$ and $p \geq 1$, $\|\pi_t^{\beta,N}(\varphi) - \pi_t^\beta(\varphi)\|_p \leq \frac{c_{t,p,\beta}\|\varphi\|_\infty}{\sqrt{N}}$, where $c_{t,p,\beta} < \infty$ is a constant independent of N.*

The proof sketch and the constant $c_{t,p,\beta}$ are in the supplement. This L_p bound provides a theoretical guarantee on the convergence of particle approximations to generalised posteriors. The special case when $p = 2$ also provides the error bound for the mean-squared error. It is well known that Theorem 1 with $p > 2$ leads to to a law of large numbers via Markov's inequality and a Borel-Cantelli argument:

Corollary 1. *Under the setting of Theorem 1, $\lim_{N\to\infty} \pi_t^{\beta,N}(\varphi) = \pi_t^\beta(\varphi) \quad \text{a.s.,} \quad \text{for} \quad t \geq 1.$*

Finally, a central limit theorem for estimates of expectations with respect to the smoothing distributions can be obtained by considering the path space $\mathsf{X}^{\otimes t}$. Recall the joint posterior $\mathsf{p}_t^\beta(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ and consider a test function $\varphi_t : \mathsf{X}^{\otimes t} \to \mathbb{R}$. We denote $\overline{\varphi}_t^\beta := \int \varphi_t^\beta(\mathbf{x}_{1:t})\mathsf{p}_t^\beta(\mathbf{x}_{1:t}|\mathbf{y}_{1:t})$ and denote the β-BPF estimate of $\overline{\varphi}_t$ with $\overline{\varphi}_t^{\beta,N} := \int \varphi_t(\mathbf{x}_{1:t})\mathsf{p}_t^{\beta,N}(\mathrm{d}\mathbf{x}_{1:t})$.

Theorem 2. *Under the regularity conditions given in [33, Theorem 1], $\sqrt{N}\left(\overline{\varphi}_t^{\beta,N} - \overline{\varphi}_t^\beta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_{t,\beta}^2(\varphi_t)\right)$, as $N \to \infty$ where $\sigma_{t,\beta}^2(\varphi_t) < \infty$.*

The expression for $\sigma^2_{t,\beta}(\varphi_t)$ can be found in the appendix. These results illustrate that the standard guarantees for generic particle filtering methods extend to our case.

5 Experiments

In this section, we focus on β-BPF illustrating its the properties and empirically verifying its robustness. We include three experiments in the main text and another in Appendix D. Furthermore, we specifically investigate the β-APF in Section 5.2 comparing its behaviour to the β-BPF. Throughout, we report the *normalised mean squared error (NMSE)* and the *90% empirical coverage* as goodness-of-fit measures. The NMSE scores indicate the mean fit for the inferred posterior distribution and the empirical coverage measures the quality of its uncertainty quantification. We note that any claim in performance difference is based on the Wilcoxon signed-rank test. Further results and in-depth details on the experimental setup are given in the supplementary material.

5.1 A Linear-Gaussian state-space model

The Wiener velocity model [34] is a standard model in the target tracking literature, where the velocity of a particle is modelled as a Wiener process. The discretised version of this model can be represented as a Linear-Gaussian State-Space model (LGSSM),

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \boldsymbol{\nu}_{t-1}, \quad \boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \quad (12) \qquad \mathbf{y}_t = \mathbf{H}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad (13)$$

where $\mathbf{A}, \mathbf{Q}$ are state-transition parameters dictated by the continuous-time model and $\mathbf{H}$ is the observation matrix (see Appendix). We simulate this model in two-dimensions with $\Sigma = \mathbf{I}$, contaminating the observations with a large scale, zero-mean Gaussian, $\mathcal{N}(0, 100^2)$ with probability p_c. Our aim is to obtain the filtering density under the heavily-contaminated setting where optimal filters struggle to perform. We compare our scheme for a large range of β to the standard BPF with a Gaussian likelihood (BPF), as well as the (optimal) Kalman filter.

We shed light onto three questions on this simple setup: (a) Does the β-BPF produce accurate and well-calibrated posterior distributions in the presence of contaminated data? (b) Is it sensitive to the choice β? (c) Does the method described in Section 3.3 for selecting β return a near optimal result?

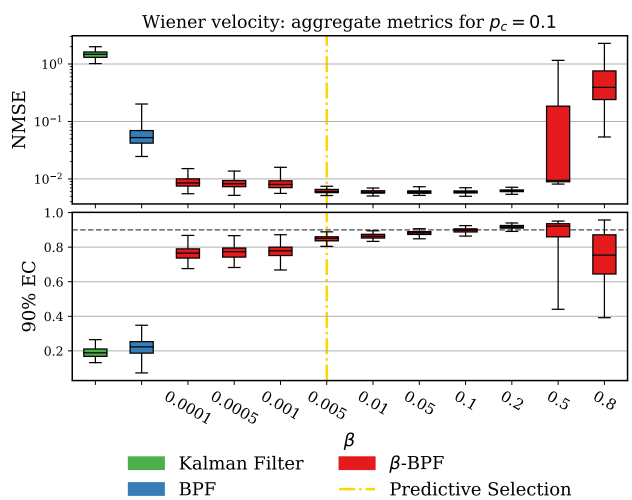

Figure 1: The mean metrics over state dimensions for the Wiener velocity example with $p_c = 0.1$. The top panel presents the NMSE results (lower is better) and the bottom panel presents the 90% empirical coverage results (higher is better), on 100 runs. The vertical dashed line in gold indicate the value of β chosen by the selection criterion in Section 3.3. The horizontal dashed line in black in the lower panel indicates the 90% mark for the coverage.

Figure 1 shows the results for $p_c = 0.1$. We observe that (a) the β-BPF outperforms the Kalman filter and the standard BPF for $\beta \leq 0.2$ while producing well-calibrated posteriors accounting for the uncertainty (for $\beta \in [0.01, 0.2]$ the coverage approaches the 90% threshold), (b) we see drastic performance gains (with median NMSE scores around $10\times$ smaller than the BPF and $100\times$ smaller that the Kalman filter) for a large range of β values, (c) we also see that the β-choice heuristic [2] chooses a well-performing β (gold vertical

[2] We apply this choice criterion on an alternative dataset that is obtained from the same simulation but with 90% fewer observations.

lines in Figure 1). Note that, for most values of β, the β-BPF significantly outperforms both the Kalman filter and the standard BPF predictively. The full set of results for the predictive performance are presented in Table 4 in Appendix F.1.

5.2 Terrain Aided Navigation

Terrain Aided Navigation (TAN) is a challenging estimation problem, where the state evolution is defined as in (12) (in three dimensions), but with a highly non-linear observation model, $\mathbf{y}_t = h(\mathbf{x}_t) + \boldsymbol{\epsilon}_t$, where $h(\cdot)$ is a non-linear function, typically including a non-analytic Digital Elevation Map (DEM). This problem simulates the trajectory of an aeroplane or a drone over a terrain map, where we observe its elevation over the terrain and its distance from its take-off hub from on-board sensors (see supplement for more details). We simulate transmission failure of the measurement system as impulsive noise on the observations, i.e., i.i.d. draws from a Student's t distribution with $\nu = 1$ degrees of freedom. In other words, we define $\boldsymbol{\epsilon}_t \sim (1 - p_c)\mathcal{N}(0, 20^2) + p_c t_{\nu=1}(0, 20^2)$.

We apply both the β-BPF and the β-APF to this problem and compare them to the standard BPF with the Gaussian (BPF). We also compare against two other robust PF methods from the literature: Student's t (t-BPF) [15] and the APF [12]. We set the degrees of freedom for the t-BPF to the same value as the contamination $\nu = 1$.

From Figure 2, we observe that for low contamination, both the β-BPF and the β-APF outperform the standard Gaussian BPF, the t-BPF and the APF. This shows that the use of t-distribution for the low contamination setting is inappropriate. This gap in the performance tightens, naturally, as p_c grows since t-distribution becomes a good model for the observations. Notably, the performance gaps between the standard PFs and their β-robustified counterparts are similar, indicating that the use of the β-divergence in a particle filtering procedure does indeed robustify the inference.

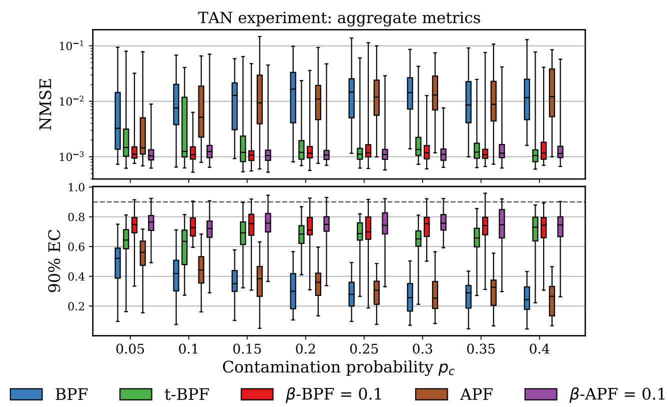

Figure 2: The mean metrics over state dimensions for the TAN example for different p_c. The top panel presents the NMSE results (lower is better) and the bottom panel presents the 90% empirical coverage results (higher is better), both evaluated on 50 runs. The horizontal dashed line in black in the lower panel indicate the 90% mark for the coverage.

In Figure 3, we plot the filtering distributions for the sixth state dimension (vertical velocity) obtained from an illustrative run with $p_c = 0.1$. The top panel shows the filtering distributions from the (Gaussian) BPF (up) and the β-BPF (down). The locations of the most prominent outliers are marked with dashed vertical lines in black. Figure 3 displays the significant difference between the two approaches: while the uncertainty for the standard BPF collapses when it meets the outliers, e.g. around $t = 1700$, the β-BPF does not suffer from this problem. This performance difference is partly related to the stability of the weights. The lower panel in Figure 3 demonstrates the effective sample size (ESS) with time for the two filters showing that the β-BPF consistently exhibits larger ESS values, avoiding particle degeneracy. The ESS values for the BPF, on the other hand, sharply decline when it meets outliers. A similar result is observed for the APF versus the β-APF in the figures in the Appendix F.2. Further results on predictive performance can be found in Appendix F.2.

5.3 London air quality Gaussian process regression

To measure air quality, London authorities use a network of sensors around the city recording pollutant measurements. Sensor measurements are susceptible to significant outliers due to environmental effects, manual calibration and sensor deterioration. In the experiment, we use Gaussian process (GP) regression to infer the underlying signal from a PM2.5 sensor.

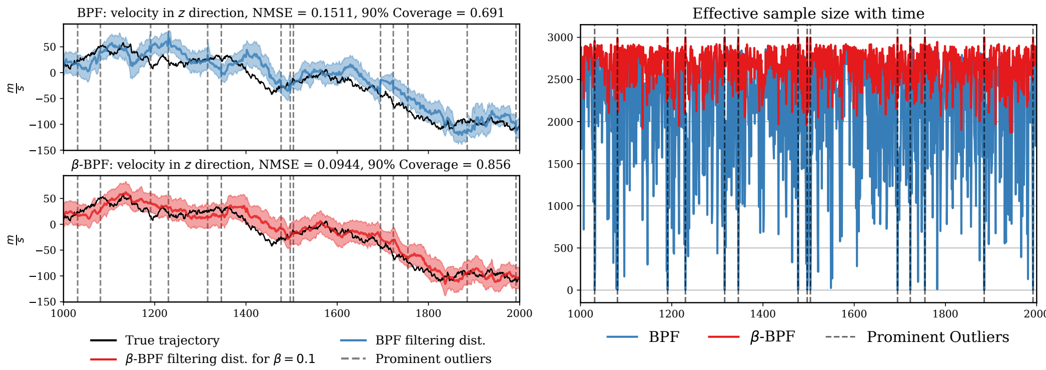

Figure 3: The left panel shows the inferred marginal filtering distributions for the velocity in the z direction for the BPF and β-BPF with $\beta = 0.1$. The right panel shows the effective sample size with time. The locations of the most prominent (largest deviation) outliers are shown as dashed vertical lines in black in both panels.

For 1-D time series data, GP inference [35] can be accelerated to linear time in the number of observations by formulating an equivalent stochastic differential equation whose solution precisely matches the GP under consideration [3] [20]. The resulting model is a LGSSM of the form (12)–(13) where the smoothing distribution matches the GP marginals at discrete-times. One can then apply smoothing algorithms, such as Rauch Tung Striebel (RTS) [36] or Forward Filters Backward Smoothing (FFBS) [37], to obtain the GP posterior. These require a forward filtering step with the Kalman filter for RTS or a PF for FFBS. Here, we fit a Matérn 5/2 GP with known hyperparameters to a time series from one of the sensors. We plot the median of the signals from the wider sensor network to obtain a simple approximation of the ground truth.

To further investigate the GP solution of the β-BPF (FFBS), we show the fit for $\beta = 0.1$ and compare it with Kalman (RTS) smoothing. In Figure 24 we see that the latter is sensitive to outliers forcing the GP mean towards them while the β-BPF is robust and ignores them.

Table 1: GP regression NMSE (higher is better) and 90% empirical coverage for the credible intervals of the posterior predictive distribution, on 100 runs. **Bold** indicates statistically significant best result from Wilcoxon signed-rank test. All presented results are statistically different from each other according to the test.

	median (IQR)	
Filter (Smoother)	NMSE	EC
Kalman (RTS)	$0.144(0)$	$0.685(0)$
BPF (FFBS)	$0.116(0.015)$	$0.650(0.020)$
$(\beta = 0.1)$-BPF (FFBS)	$0.061(0.003)$	$0.760(0.015)$
$(\beta = 0.2)$-BPF (FFBS)	$\mathbf{0.059(0.002)}$	$\mathbf{0.803(0.020)}$

Table 1 compares results with a Gaussian likelihood for GP regression with Kalman (RTS) smoothing, the standard BPF (FFBS) and two runs for the β-BPF (FFBS) ($\beta = 0.1$ by predictive selection as Section 3.3 and $\beta = 0.2$ by overall best performance). For both choices of β, the β-BPF outperforms all other methods on both metrics .

6 Conclusions

We provided a generalised filtering framework based on GBI, which tackles likelihood misspecification in general state-space HMMs. Our approach leverages SMC methods, where we extended some analytical results to the generalised case. We presented the β-BPF, a simple instantiation of our approach based on the the β-divergence, developed an APF for this setting, and showed performance gains compared to other standard algorithms on a variety of problems and contamination settings.

This work opens up many exciting avenues for future research. Principle among which is online learning for model parameters (system identification) in the presence of misspecification. Our framework can directly incorporate most estimators found in the SMC literature and the computation of derivatives can be tackled with automatic differentiation tools.

[3]The SDE representation of a GP depends on the form of the covariance function. In this paper we use a GP with the Mateŕn 5/2 kernel, which admits a dual SDE representation.

7 Broader Impact

Robust inference in the context of misspecified models is a topic of broad interest. However, there are a few robust generally-applicable methods which can be employed in the context of online inference in time series settings. This paper provides a principled solution to this problem within a formal framework backed by theoretical guarantees and opening up the benefits to multiple application domains. The illustrative applications demonstrate the potential improvements in settings including navigation and Gaussian process regression, which, if realised more widely, could have wide-reaching impact. We hope that this inspires the community to build-on or apply our work to other challenging real-world scenarios.

Of particular interest is the application of Robust SMC methods, like the β-BPF and the auxiliary counterpart which were developed in this work, to impactful data-streaming applications in environmental monitoring and forecasting. Indeed, our research in this area was motivated by a real-world application in which existing techniques were inadequate (see *anonymized reference* for more details). We have demonstrated the benefits such methods in proof-of-concept work and are incorporating the resulting algorithms into a fully-developed platform, that has been in development for approximately three years. We are partnering with local authorities to help in directly informing policy makers and ultimately the general public.

More widely, this work provides an additional illustration that the GBI framework can provide good solutions to challenging problems in the world of misspecified framework and hence provides additional motivation to further investigate this extremely promising but rather new direction.

References

[1] Brian D O Anderson and John B Moore. *Optimal filtering*. Englewood Cliffs, N.J. Prentice Hall, 1979.

[2] Simo Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

[3] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[4] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.

[5] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.

[6] Peter J Huber. *Robust statistics*. Springer, 2011.

[7] Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 813–822, 2018.

[8] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

[9] Arnold Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

[10] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized Variational Inference: Three arguments for deriving new Posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

[11] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[12] Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

[13] Cristina S Maiz, Joaquin Miguez, and Petar M Djuric. Particle filtering in the presence of outliers. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 33–36. IEEE, 2009.

[14] Cristina S Maiz, Elisa M Molanes-Lopez, Joaquín Miguez, and Petar M Djuric. A particle filtering scheme for processing time series corrupted by outliers. *IEEE Transactions on Signal Processing*, 60(9):4611–4627, 2012.

[15] Dingjie Xu, Chen Shen, and Feng Shen. A robust particle filtering algorithm with non-Gaussian measurement noise using student-t distribution. *IEEE Signal Processing Letters*, 21(1):30–34, 2013.

[16] Laurent E. Calvet, Veronika Czellar, and Elvezio Ronchetti. Robust filtering. *Journal of the American Statistical Association*, 110(512):1591–1606, 2015.

[17] Francisco Curado Teixeira, João Quintas, Pramod Maurya, and António Pascoal. Robust particle filter formulations with application to terrain-aided navigation. *International Journal of Adaptive Control and Signal Processing*, 31(4):608–651, 2017.

[18] Xiao-Li Hu, Thomas B Schon, and Lennart Ljung. A robust particle filter for state estimation—with convergence results. In *46th IEEE Conference on Decision and Control*, pages 312–317. IEEE, 2007.

[19] Ömer Deniz Akyildiz and Joaquín Míguez. Nudging the particle filter. *Statistics and Computing*, 30:305–330, 2020.

[20] Simo Särkkä, Arno Solin, and Jouni Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

[21] Abhik Ghosh and Ayanendranath Basu. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

[22] Jeremias Knoblauch, Jack E Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data with β-divergences. In *Advances in Neural Information Processing Systems*, pages 64–75, 2018.

[23] Tom Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

[24] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

[25] Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

[26] Dilin Wang, Hao Liu, and Qiang Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, pages 5737–5747, 2018.

[27] Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan and B. Rozovskiĭ, editors, *The Oxford Handbook of Nonlinear Filtering*, pages 656–704. Oxford University Press, 2011.

[28] Mathieu Gerber, Nicolas Chopin, and Nick Whiteley. Negative association, ordering and convergence of resampling methods. *Annals of Statistics*, 47(4):2236–2260, 2019.

[29] Pierre Del Moral. *Feynman-Kac formulae: Genealogical and interacting particle systems with applications*. Springer, 2004.

[30] Adam M Johansen and Arnaud Doucet. A note on the auxiliary particle filter. *Statistics and Probability Letters*, 78(12):1498–1504, September 2008.

[31] Alexandros Beskos, Omiros Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of the Royal Statistical Society, Series B*, 68(3):333–382, 2006.

[32] Paul Fearnhead, Omiros Papaspiliopoulos, and Gareth O. Roberts. Particle filters for partially-observed diffusion. *Journal of the Royal Statistical Society, Series B*, 70(4):755–777, 2008.

[33] Nicolas Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.

[34] Simo Särkkä and Arno Solin. *Applied Stochastic Differential Equations*, volume 10. Cambridge University Press, 2019.

[35] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

[36] Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal*, 3(8):1445–1450, 1965.

[37] Mark Briers, Arnaud Doucet, and Simon Maskell. Smoothing algorithms for state space models. *Annals of the Institute of Statistical Mathematics*, 62(1):61–89, 2010.

[38] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1947.

[39] Jack Jewson, Jim Q Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

[40] Pieralberto Guarniero, Adam M Johansen, and Anthony Lee. The iterated auxiliary particle filter. *Journal of the American Statistical Association*, 112(520):1636–1647, 2017.

[41] Joaquín Míguez, Dan Crisan, and Petar M Djurić. On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Statistics and Computing*, 23(1):91–107, 2013.