Robust particle filters via sequential pairwise reparameterized Gibbs sampling

Liam Paninski, Kamiar Rahnama Rad, and Michael Vidne

***Abstract*—Sequential Monte Carlo ("particle filtering") methods provide a powerful set of tools for recursive optimal Bayesian filtering in state-space models. However, these methods are based on importance sampling, which is known to be non-robust in several key scenarios, and therefore standard particle filtering methods can fail in these settings. We present a filtering method which solves the key forward recursion using a reparameterized Gibbs sampling method, thus sidestepping the need for importance sampling. In many cases the resulting filter is much more robust and efficient than standard importance-sampling particle filter implementations. We illustrate the method with an application to a nonlinear, non-Gaussian model from neuroscience.**

I. Introduction and background

Sequential Monte Carlo ("particle filtering") methods have become quite popular over the last two decades [1], largely because these methods offer a general recipe for (approximate) optimal Bayesian inference on nonlinear, non-Gaussian time series data, in a recursive and computationally tractable form.

However, it is well-known that these methods can perform unreliably in a number of important cases. The basic problem is that standard particle filters rely fundamentally on importance sampling, which is known to be unreliable in many high-dimensional settings [2], [3] (though see [4] for an alternative view) and more generally in cases where it is difficult to construct accurate proposal densities [5]. In these cases the particle filter will often choose particle locations that provide a poor match to the data, leading to rapid particle depletion and a highly suboptimal approximation of the target posterior distributions. As a consequence, particle filters can be very non-robust with respect to outliers or model misspecifications.

A number of potential solutions to this problem have been proposed. To discuss these ideas, we need to introduce some notation. The basic filtering problem is to estimate the conditional probability $p(q_t|Y_{1:t})$ of the Markovian hidden state variable q_t, given all observed data $Y_{1:t} = \{y_0, ..., y_t\}$ in the time interval $[0, t]$, under the standard hidden Markov model assumption that each observed data point y_t depends directly only on the state variable q_t at time t. We assume that the observation probability, $p(y_t|q_t)$, and the transition probability, $p(q_t|q_{t-1})$, are known. Due to space limitations, we will not review basic particle filtering methodology here; see e.g. [1], [6] for background.

The "auxiliary" particle filter (APF) introduced by [7] is one effective method for incorporating the observation y_t into the sampler for q_t, therefore leading to a much more reliable filter. (Related approaches have been discussed more recently by [8] and [9].) The APF is highly effective if we can (1) compute a good approximation to the marginal likelihood $\int p(y_t|q_t)p(q_t|q_{t-1})dq_t = p(y_t|q_{t-1})$, and then (2) efficiently sample from the conditional distribution $p(q_t|y_t, q_{t-1})$. As we will discuss below, in many important cases the sampling step (2) is feasible. However, computation of marginal likelihoods is a notoriously difficult problem [10] (for some further discussion, see e.g. [11]); in practice, often importance sampling methods are used to approximate $p(y_t|q_{t-1})$, and our goal here is to avoid importance sampling methods entirely. (Though see the discussion section for a brief consideration of some approaches for improving the importance sampling step directly.)

Markov chain Monte Carlo (MCMC) methods provide a natural alternative to the importance sampling approach. The influential paper [12] proposes MCMC methods to sample from $p(q_t, q_{t-1}, \ldots q_{t-L}|y_t, y_{t-1}, \ldots y_{t-L}, q_{t-L-1})$, for some time lag L. (See [13] for a related idea that uses importance sampling instead of MCMC methods; [14] discuss a different stepwise MCMC-based approach that replaces the importance sampling step with a related independence Metropolis-Hastings sampler.) This approach is quite powerful in principle, since it allows us to correct "mistakes" — i.e., samples q_{t-n} that poorly match the observed data $Y_{1:t}$ — up to L time steps in the past, given the new observation y_t. In practice, it may be difficult to construct a rapidly-mixing MCMC chain to sample from $p(q_t, q_{t-1}, \ldots q_{t-L}|y_t, y_{t-1}, \ldots y_{t-L}, q_{t-L-1})$. Our main goal here is to develop more efficient MCMC chains, focusing on the $L = 1$ case.

II. Our approach: pairwise reparameterized Gibbs sampling

Assume that at time $t-1$ we have a weighted particle representation of the forward distribution,

$$p(q_{t-1}, Y_{0:t-1}) \approx \sum_{i=1}^{N} w_{t-1}^{(i)} \delta\left(q_{t-1} - q_{t-1}^{(i)}\right); \qquad (1)$$

here N is the number of particles, and $w_{t-1}^{(i)}$ and $q_{t-1}^{(i)}$ denote the weight and location, respectively, of the i-th particle at time $t-1$.

Now the standard forward recursion for hidden Markov models [15] is

$$\begin{aligned} p(q_t, Y_{0:t}) &= \int p(q_t, q_{t-1}, Y_{0:t}) dq_{t-1} \qquad (2) \\ &= \int p(y_t|q_t) p(q_t|q_{t-1}) p(q_{t-1}, Y_{0:t-1}) dq_{t-1}. \end{aligned}$$

To appear in the Proceedings of the 46th annual Conference on Information Science and Systems (CISS 2012), as part of an invited session on Computational Neuroscience. The authors are with the Department of Statistics and Center for Theoretical Neuroscience, Columbia University. Contact: liam@stat.columbia.edu. Revised: February 24, 2012.

Plugging in eq. (1), we have

$$p(q_t, q_{t-1}, Y_{0:t}) \approx \sum_{i=1}^{N} w_{t-1}^{(i)} p(y_t|q_t) p(q_t|q_{t-1}^{(i)}) \delta\left(q_{t-1} - q_{t-1}^{(i)}\right), \tag{3}$$

where we have abbreviated $p(q_t|q_{t-1}^{(i)}) = p(q_t|q_{t-1} = q_{t-1}^{(i)})$. Now the basic idea is that, according to eq. (2), if we can draw N samples from the pairwise conditional distribution

$$p(q_t, q_{t-1}|Y_{0:t}) \propto p(q_t, q_{t-1}, Y_{0:t}) \tag{4}$$

via the approximation (3), then we can obtain an (unweighted) particle approximation to the desired conditional forward distribution at time t, $p(q_t|Y_{0:t})$, simply by discarding (marginalizing) the q_{t-1} component of our samples.

Thus we will focus on developing fast methods for sampling from the pairwise distribution $p(q_t, q_{t-1}|Y_{0:t})$. It is helpful to rewrite eq. (3) as $p(q_t, q_{t-1}, Y_{0:t}) \approx$

$$\sum_{i=1}^{N} w_{t-1}^{(i)} p(y_t|q_{t-1}^{(i)}) p(q_t|q_{t-1}^{(i)}, y_t) \delta\left(q_{t-1} - q_{t-1}^{(i)}\right). \tag{5}$$

This clarifies the connection to the APF approach: if we can compute $p(y_t|q_{t-1}^{(i)}) = \int p(y_t|q_t) p(q_t|q_{t-1} = q_{t-1}^{(i)}) dq_t$ and then sample from $p(q_t|q_{t-1}^{(i)}, y_t) = p(q_t|q_{t-1} = q_{t-1}^{(i)}, y_t)$, then we can sample directly from our pairwise target distribution (3). As we mentioned in the introduction, the latter task is often relatively easy. For example, in many applications both the observation distribution $p(y_t|q_t)$ and the transition distribution $p(q_t|q_{t-1})$ are log-concave densities; we will assume that this is the case throughout this paper. This implies that the conditional density $p(q_t|q_{t-1}, y_t)$ is also log-concave; a number of effective rejection or MCMC methods exist for sampling from log-concave (and therefore unimodal) densities [5]. However, as emphasized above, computing $p(y_t|q_{t-1}^{(i)})$ is more challenging, and we do not attempt to do this directly.

Instead, we will apply a transformed Gibbs method to sample from (3). The key feature of our problem is that it is relatively easy to make MCMC moves in the q_t direction (since we have assumed that the conditional distribution $p(q_t|q_{t-1}, y_t)$ is log-concave), but moving in the q_{t-1} direction is typically harder, since in many cases, for any fixed q_t, $p(q_t|q_{t-1}^{(i)}) p(y_t|q_t)$ may be a sharply-peaked function of i (since the densities $p(q_t|q_{t-1}^{(i)}, y_t)$ may have minimal overlap for different values of i), implying that the Gibbs chain mixes slowly (or not at all, if the densities $p(q_t|q_{t-1}^{(i)}, y_t)$ have disjoint support for different indices i). See the left panel of Fig. 1 for an illustration.

Eq. (3) (or equivalently (5)) can be treated as a mixture, where each particle i indexes a different mixture component. The fact that Gibbs sampling often mixes slowly in mixture settings (because it can be a challenge to jump efficiently between mixture components) is well-known [16], [17], and a number of strategies have been suggested for dealing with this problem. Some examples include so-called "tempering" methods [18], [19], [20], which replace the original challenging mixture distribution with a sequence of flatter densities which are easier to sample from, or the "darting" approaches discussed, e.g., in [21]), where "darting" regions are defined

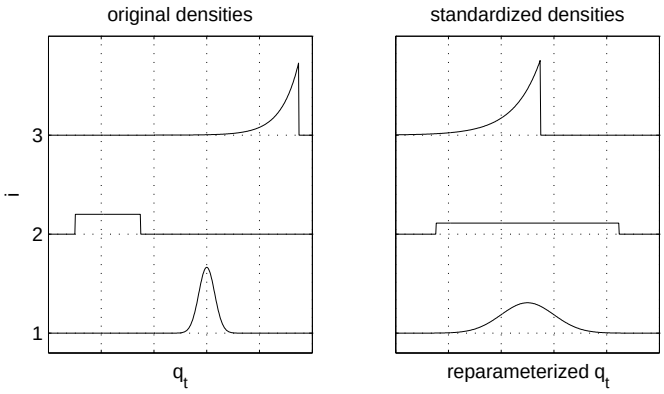

Fig. 1. A schematic illustration of the standardization idea. **Left**: three hypothetical densities $p(q_t|q_{t-1}^{(i)}, y_t)$, for $i = 1, 2, 3$. Note that these densities have nearly disjoint support, and therefore Gibbs jumps in the "vertical" direction (the i-direction) will be rare. **Right**: standardized versions of the densities shown on the left. Note that these standardized densities have much more overlap, greatly increasing the Gibbs mixing rate.

near modes of the target distribution, and the chain is allowed to make jumps between these regions. We will discuss applications of these methods further below.

The approach we propose here is simple and less generally applicable than the tempering method, but in many cases leads to a fast, effective, and easy-to-code algorithm. The basic idea (quite common in the Gibbs sampling literature) is to apply a reparameterization so that the densities $p(q_t|q_{t-1}^{(i)}, y_t)$ have greater overlap in the rescaled space; see the right panel of Fig. 1 for an illustration. Then we can apply standard Gibbs (or Metropolis-Hastings-within-Gibbs) sampling in the rescaled space, and finally map our reparameterized samples back to the original space.

What kind of reparameterization should we use? Computational efficiency is a key consideration here, so while nonlinear reparameterizations could certainly be useful in some applications, we will limit our attention to linear (affine) reparameterizations here. In a sense we want to "standardize" each of the distributions $p(q_t|q_{t-1}^{(i)}, y_t)$, recentering and rescaling each of these distributions so that the high-probability region in each of these distributions coincides as much as possible. Since we are restricting our attention to linear mappings, it is sufficient to define ellipses for each distribution indexed by i, where the i-th ellipse corresponds (in some sense) to the high-probability region of the i-th distribution (recalling that each of these distributions is assumed to be unimodal); then our reparameterization consists simply of the standardization mapping each of the N ellipses onto the unit sphere. (Of course, we must keep track of the volume of each of the original ellipses, so that we can apply the standard determinant change-of-measure formula to each of the transformed distributions.)

Before discussing further practical implementation details, it is worth noting connections to a couple other ellipse-based methods. First, a very common (and often quite useful) method for constructing proposal densities in standard importance sampling-based filters is to form a Laplace approximation of $p(q_t|q_{t-1}^{(i)}, y_t)$, then use the resulting Gaussian densities (or heavier-tailed densities with the same location and scatter

parameters) as proposal densities. The scatter matrix of the i-th proposal density constructed in this manner is clearly related to the i-th ellipse discussed above. However, we emphasize that we will not use this importance-sampling approach here (since in many cases of interest, $p(q_t|q_{t-1}^{(i)}, y_t)$ may have sharp corners, or may be high-dimensional, and in either case the resulting proposal densities may not match the target densities closely enough to be useful); we only use the ellipses to define a reparameterization that improves the mixing of our Gibbs sampler. Similarly, the "darting" methods mentioned above [21] construct ellipses near each mode of the target distribution to define a "mode-hopping" MCMC algorithm: once the sampler enters one of these ellipses it is allowed to jump to one of the other ellipses, under an appropriate Metropolis-Hastings acceptance probability. In our proposed Gibbs-based algorithm, note again that the ellipses are used only to define the reparameterization, and we do not restrict our sampler to jump only within these ellipses — the Gibbs sampler can potentially jump from $q_{t-1}^{(i)}$ to $q_{t-1}^{(j)}$ for any i and j, given any value of q_t.

III. Implementation details

To specify our algorithm in any concrete model setting, we need to specify a computationally-efficient method for constructing the reparameterizations, or equivalently for approximating the N high-probability ellipses, one for each of the conditional distributions $p(q_t|q_{t-1}^{(i)}, y_t)$. This step is highly problem-dependent.

In some cases we can define good reparameterizations analytically. See section IV below for an example in which the reparameterized densities $p(q_t|q_{t-1}^{(i)}, y_t)$ turn out to be exactly equal for a suitably chosen linear reparameterization, leading to very fast mixing of the Gibbs chain.

More generally, we can apply standard approximation methods, such as Laplace approximation or expectation propagation [22]. These methods are iterative, but good initializations are often available analytically, and in many cases only a few iterations will suffice. For example, consider the fairly broad class of nonlinear state-space models with linear observations,

$$q_t = f_t(q_{t-1}) + \epsilon_t \quad (6)$$

$$y_t = B_t q_t + \eta_t, \quad (7)$$

where the $f_t(.)$ are arbitrary (potentially nonlinear) functions, each B_t is a linear operator, and ϵ_t and η_t are stochastic terms. The dynamics noise $p(\epsilon_t|q_{t-1})$ can depend on q_{t-1} (and t), but for simplicity assume that the observation noise $p(\eta_t)$ is independent of q_t (though the ideas here can be easily extended to the case that y_t is generated, e.g., by a generalized linear model given q_t). By assumption, $\log p(\epsilon_t|q_{t-1})$ and $\log p(\eta_t)$ are concave functions of ϵ_t and η_t, respectively; if we further assume that these functions are smooth and not too non-quadratic, then Laplace approximations to $p(q_t|q_{t-1}^{(i)}, y_t)$ can be initialized analytically.

In the most general settings, these classical approximations may still be inadequate. We have also assumed that we can sample easily from each $p(q_t|q_{t-1}^{(i)}, y_t)$. Thus a natural general approach is to simply use samples from each distribution indexed by i to define the i-th ellipse, e.g., by computing some robust estimates of the center and scale parameters from the samples [23], [24]. These sample-based ellipses can be computed using an initial run, before the reparameterized Gibbs sampler is started, or we can potentially update our ellipses as the Gibbs sampler produces more samples (though this may be computationally expensive, and makes the convergence analysis of the resulting time-varying Gibbs chain much more complicated).

Computationally, it is helpful to note that the Gibbs approach can be parallelized quite easily, simply by running multiple independent chains. We can also parallelize the numerical construction of the N reparameterizations, if analytic solutions are unavailable. It is also worth noting that in many cases it might be more efficient to use a Metropolis-within-Gibbs approach, rather than direct Gibbs. This is because each Gibbs step in the $q_{t-1}^{(i)}$ direction (holding the reparameterized q_t fixed) requires us to compute $w_{t-1}^{(i)} p(y_t|q_t) p(q_t|q_{t-1}^{(i)})$ for each of the N possible values of $q_{t-1}^{(i)}$. In many cases of interest, many of these values will be negligible. For example, $w_{t-1}^{(i)} p(y_t|q_{t-1}^{(i)})$ may be near zero for many values of i; this is typically the situation when the observation y_t is highly unlikely given the collection $q_{t-1}^{(i)}$, which is one of the major cases that we are interested in here. Thus if we have some approximate estimates for $p(y_t|q_{t-1}^{(i)})$ (e.g., via the Laplace approximation approach outlined above), it is much more efficient to focus the sampler's attention on the values of i for which $w_{t-1}^{(i)} p(y_t|q_{t-1}^{(i)})$ is large; for example, we can use an independence Metropolis-Hastings sampler to sample in the i direction, using a normalized version of our estimate of $w_{t-1}^{(i)} p(y_t|q_{t-1}^{(i)})$ as a proposal distribution.

In some particularly difficult cases (e.g., if the state variable q_t is high-dimensional, and the transition and/or observation densities have many "sharp corners") it may not be possible to align the conditional distributions $p(q_t|q_{t-1}^{(i)}, y_t)$ via a linear transformation. In these cases even the reparameterized Gibbs chain will mix slowly. This can in principle be diagnosed directly if we are running two or more Gibbs chains in parallel: if we find that two particles end up moving exclusively on different subsets of indices i, then we know we have a mixing problem[1]. If poor mixing is encountered it is necessary to switch to a more general, more computationally expensive approach, such as using linked importance sampling [25] to estimate $p(y_t|q_{t-1}^{(i)})$ (and then using the APF to generate samples q_t), or simulated tempering methods [18], [19], [20] to sample from the pairwise distribution (5) directly.

Finally, note that it will often be unnecessary to apply our method at every time step, since in many cases the standard particle filtering methods will work well, if the observations y_t are consistent with the prior distribution of the state variable q_t. Thus a natural approach would be to augment a standard particle filter with an outlier check [26], [27] at each time

[1]Note that it is not necessarily a problem if all of the particles stay on one index i, since as discussed above, in many cases $w_{t-1}^{(i)} p(y_t|q_{t-1}^{(i)})$ might be sharply concentrated on just one or a few values of i. But if there are two such high-probability indices i that do not communicate with each other, this indicates a mixing problem.

step (where "outlier" observations y_t can be detected based on an estimate of the marginal probability of y_t, or of the effective sample size following the incorporation of the y_t), only invoking our method at the subset of time points at which an outlier is detected, where we expect standard particle filter methods to return an inaccurate estimate of the posterior $p(q_t|Y_{1:t}$.

IV. Application

In this section we describe a simple illustrative nonlinear, nonstationary, non-Gaussian example from neuroscience. The Fitzhugh-Nagumo model [28] is a two-dimensional model of excitable media, used in computational neuroscience as a simplified model for action potential generation. While much more realistic models are available, we chose this two-dimensional model for its ease of visualization. In this context the two state variables $q_t = (v_t, w_t)$ are interpreted as a membrane voltage (v_t) and an auxiliary variable w_t that controls the excitability of the neuron. We used the state dynamics

$$dv_t = \left(v_t - v_t^3/3 - w_t + I_t\right) dt + \sqrt{dt} e_t^v \quad (8)$$

$$dw_t = 0.4\left(v_t + 0.7 - 0.8 w_t\right) dt + \sqrt{dt} e_t^w, \quad (9)$$

where e_t^w and e_t^v were chosen as independent Laplacian (double-exponential) variables, with scale $1/5$ and $1/1000$, respectively; I_t was a 3 Hz sinusoidal input of amplitude 0.3, and $dt = 0.1$. (We used the Laplacian distribution rather than the more common Gaussian here to explore a slightly heavier-tailed noise distribution.) These parameters put the model neuron in a weakly stochastically resonant regime: when the noise e_t^w and e_t^v is set to zero, the neuron does not emit any action potentials, but in the presence of noise the neuron fired action potentials roughly synchronized to the periodic input signal I_t. The results described below do not depend strongly on the model parameters, as long as the firing frequency of the neuron is fairly small (but nonzero) and the scale of the voltage noise e_t^v was significantly greater than that of the auxiliary noise e_t^w. (We will discuss the relevance of these conditions at more length below.)

We examined the performance of the filter given some very simple simulated observed data Y: no action potentials are observed for the first 0.5 second of the experiment, and then a single action potential is observed at $t = 0.5$. Observations were considered to be binary variables, indicating the presence or absence of an action potential. We use a very basic deterministic voltage threshold-crossing observation model:

$$p(y_t = 1|q_t) = 1 \text{ if } v_t > 1 \quad (10)$$

$$= 0 \text{ otherwise.} \quad (11)$$

Note w_t is not observed directly; i.e., observations $\{y_t\}$ are conditionally independent of $\{w_t\}$ given $\{v_t\}$. After $t = 0.5$, we allowed the state variable q_t to evolve according to its dynamics, with no further observations taken for $t > 0.5$.

Our choice of models here has several important implications. First, the "perfectly-adapted" APF (in the terminology of [7]) can be computed explicitly — i.e., we can analytically compute $p(y_t|q_{t-1})$ and sample directly from $p(q_t|q_{t-1}, y_t)$ — and we use this analytical filter as a "gold standard" for comparison. (Of course, more generally the PA-APF will not be easily available analytically, and therefore we have to approximate it using the methods described above.)

Second, because $p(w_t|q_{t-1})$ has small variance relative to the variance of $w_{t-1}^{(i)}$ in this model (since the latter quantity depends on the past history of v_t, which has larger conditional variance than w_t; see Fig. 4 below for an illustration), we expect that the conditional distributions $p(q_t|q_{t-1}^{(i)}, y_t)$ will have small overlap and therefore straightforward Gibbs sampling (with no standardization) will mix poorly when applied to the pairwise distribution $p(q_t, q_{t-1}|y_t)$ (recall eqs. 3 and 5). This is indeed the case, as demonstrated in Figs. 2-3 below.

Finally, in contrast, we expect the reparameterized Gibbs approach to perform quite well here, since it turns out that the standardized $p(q_t|q_{t-1}, y_t)$ densities turn out to all be equal in this case, as can be demonstrated with a direct calculation (omitted for brevity). This makes the reparameterized Gibbs chain mix optimally (since the target mixture distribution (5) can be written as a product distribution as a function of the indices i and the reparameterized q_t variables), ensuring the method's computational efficiency in this case. Obviously the densities $p(q_t|q_{t-1}, y_t)$ will not overlap so nicely in general; however, we have focused on this special case here to demonstrate that this simple reparameterized Gibbs approach can often perform quite well, even in strongly non-Gaussian examples where methods based on importance sampling can break down[2].

Figs. 2 and 3 compare the performance of several of the filters we have discussed above. The perfectly adapted APF is labeled "PA-APF"; the reparameterized Gibbs method we presented above is "Standardized Gibbs"; and "Gibbs" refers to straightforward, non-reparameterized Gibbs applied to the pairwise distribution $p(q_t, q_{t-1}|y_t)$. Finally, "Weare" refers to the method introduced by [9], which can be seen as a version of the APF in which $p(q_t|q_{t-1}, y_t)$ is sampled exactly, but $p(y_t|q_{t-1}^{(i)})$ is estimated via a simple importance sampling estimate (where $p(q_t|q_{t-1}^{(i)})$ is used as a proposal density). The number of particles $N = 20$ in each case.

The performance of each of the four filters was similar for $t < 0.5$ (data not shown), where all of the observations y_t were set to be zero and the prior density specified by the dynamics provided a good match to the conditional density given the data (recall that the model dynamics chosen here lead to a small probability of a voltage threshold crossing, consistent with $y_t = 0$). Therefore, to enable a fair comparison, we used the PA-APF to generate the particle trajectories for $t < 0.5$, for all four methods. This is indicated in black in Fig. 2.

On the other hand, the observation $y_t = 1$ at $t = 0.5$ is

[2]More generally, we can guarantee that the proposed method will perform well if the e_t^v and e_t^w random variables are independent, and $\log p(e_t^v)$ is concave and smoothly decaying. In this case, it is easy to show that the conditional distribution given an unlikely $y_t = 1$ observation can be well-approximated by a product of $p(e_t^w)$ and an exponential distribution in e_t^v offset so that the left edge of its support coincides exactly with the voltage threshold, exactly as in the case considered here. (This approximation becomes exact in the limit that the voltage threshold is very high.) Generalizations of this example are easy to construct. Thus it is clear that there is a wider range of cases for which the conditional distributions $p(q_t|q_{t-1}^{(i)}, y_t)$ are well-alignable in the sense required here.

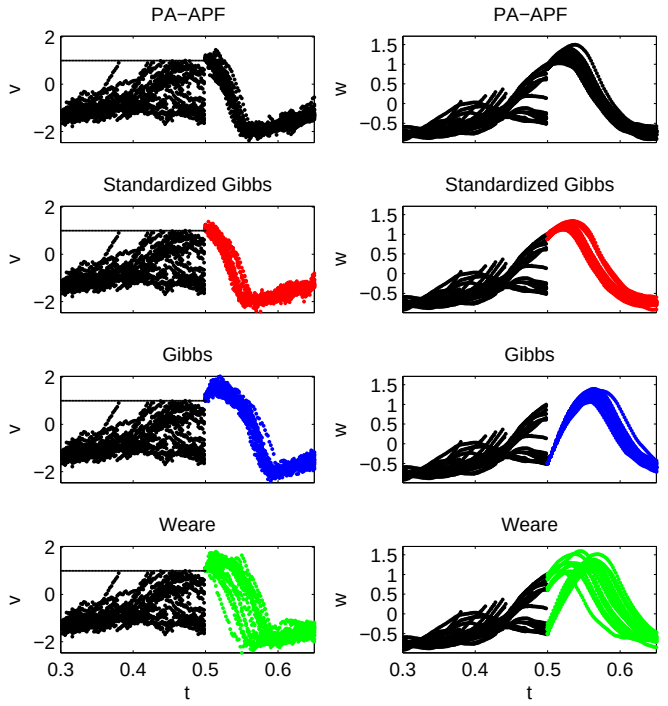

Fig. 2. Comparing four particle filters applied to thresholded voltage observations from a Fitzhugh-Nagumo model. Each row shows the performance of one filter (described in the text), where the filtered voltages v_t are displayed in the left column and the filtered auxiliary variable w_t are shown in the right column. Each dot corresponds to one particle location $q_t^{(i)}$; each of the methods shown here produce equally-weighted particles at all time points, i.e., $w_t^{(i)} = 1/N$ for all i and t. Line at $v = 1$ indicates voltage threshold; particles crossing the threshold prematurely (for $t < 0.5$) are killed. Note the discontinuity at $t = 0.5$, where y_t switches from 0 to 1. Colorscheme chosen to emphasize that the same filter is used for all four rows for $t < 0.5$, and then the indicated filters are used at $t = 0.5$, with the particles allowed to evolve according to the prior dynamics for $t > 0.5$ (see main text). Note that the reparameterized Gibbs method (red) matches the output of the "gold standard" perfectly adapted APF (black), while significant errors are evident for the other two filters. See Fig. 3 for summary statistics.

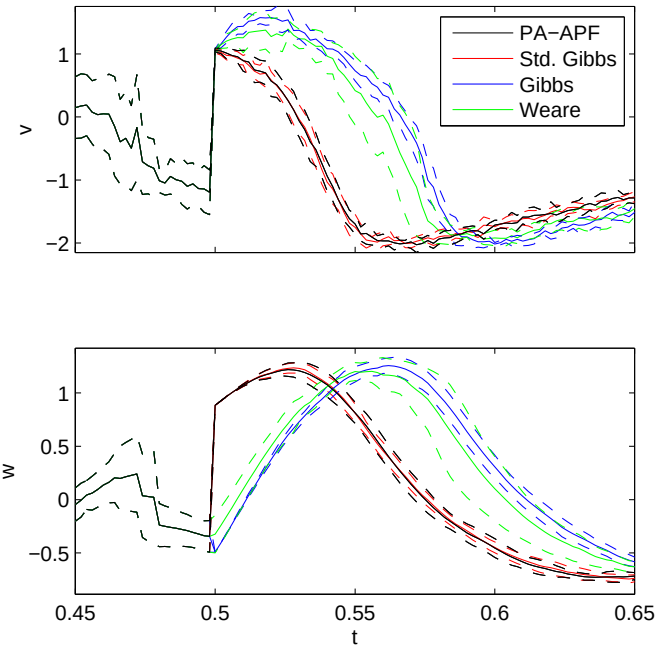

Fig. 3. Summary and zoom-in of the filtering results shown in Fig. 2: median (solid) $\pm$ median absolute deviation from the median (dashed) for each of the four filters presented in Fig. 2. Again, note that the reparameterized Gibbs method (red) matches the PA-APF output (black), while large errors are evident in both the v_t and w_t output for the other two filters.

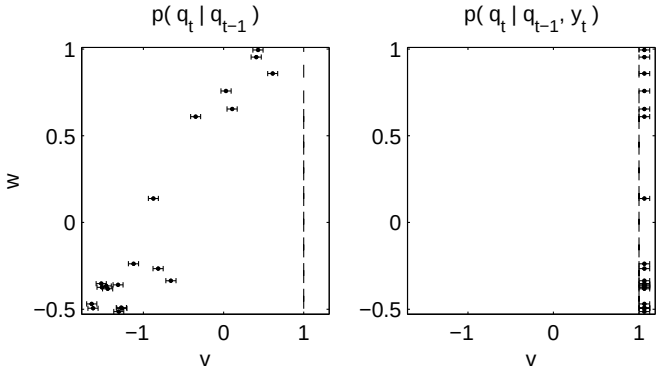

Fig. 4. The effective support of the densities $p(q_t|q_{t-1}^{(i)})$ (left) and $p(q_t|q_{t-1}^{(i)}, y_t)$ (right), evaluated at $t = 0.5$, for the particles shown in Fig. 2. Each horizontal errorbar shows the conditional standard deviation of $p(v_t|q_{t-1}^{(i)})$; vertical errorbars corresponding to the conditional standard deviation of $p(w_t|q_{t-1}^{(i)})$ are shown as well, but are too small to see here (recall that the conditional scale of w_t given q_{t-1} is much smaller than that of v_t in these simulations). The vertical line indicates the voltage threshold at $v = 1$. Note that $p(v_t \geq 1|q_{t-1}^{(i)})$ is very small for all values of i here; this explains why the approach of [9] breaks down in this case (c.f. Fig. 3; see main text for additional discussion). Additionally, note that almost all of the conditional densities $p(q_t|q_{t-1}^{(i)}, y_t)$ (right panel) have negligible overlap, leading to a very slow mixing rate of the non-reparameterized Gibbs chain, and the failure of the corresponding non-reparameterized Gibbs filter (Fig. 3).

quite unlikely given all of the preceding $q_{t-1}^{(i)}$ particles that happen to have been chosen here. Both the APF and the reparameterized Gibbs method are able to handle this *a priori* unlikely observation well; in fact, we find that the performance of these two filters is indistinguishable here.

However, both of the other approaches fail. The nonstandardized Gibbs filter (third panel of Fig. 2; blue traces) breaks down because the mixing rate of the Gibbs chain in this case is extremely slow, because of the small overlaps of the conditional distributions $p(q_t|q_{t-1}^{(i)}, y_t)$, as discussed above; see Fig. 4 for further details. Empirically, this means that the Gibbs chain gets stuck on whatever value of $q_{t-1}^{(i)}$ we happen to initialize the chain, and is never able to sample other values; this obviously sharply reduces the particle diversity (since all N selected values of $q_{t-1}^{(i)}$ are identical, with high probability), and leads to a bias due to the fact that the chain depends so strongly on its initial conditions. The approach of [9] (fourth panel of Fig. 2; green traces), on the other hand, fails because with high probability none of the importance samples from the prior proposal $p(q_t|q_{t-1}^{(i)})$ are consistent with the observed data $y_t = 1$ (see Fig. 4), which means that the estimates for $p(y_t|q_{t-1})$ are all set to zero. (To prevent catastrophic failure of the code in this case, we set $p(y_t|q_{t-1}) \propto 1/N$, which leads to the bias and overestimate of the variance for $t > 0.5$ shown in the green traces of Figs. 2-3.)

V. Conclusion

We have presented a sequential pairwise reparameterized Gibbs sampling approach that can significantly improve the robustness of particle filtering methods. Our approach is most effective when it is easy to draw samples from the conditional distribution $p(q_t|q_{t-1}, y_t)$ (e.g., via MCMC methods), and additionally these distributions can be easily linearly aligned, i.e., the high-probability regions of the distributions $p(q_t|q_{t-1}, y_t)$ and $p(q_t|q'_{t-1}, y_t)$ can be made to overlap via a linear transformation for any pair (q_{t-1}, q'_{t-1}). We provide an example application to a classical model from neuroscience in section IV, where the improved performance of the reparameterized Gibbs approach is especially clear — in particular, in this case we can sample and reparameterize the necessary conditional distributions exactly — but as noted above we expect similar performance gains even in many cases for which such simple analytical approaches are not available.

As we emphasized in the introduction, a number of approaches for improving the robustness of particle filtering methods have been introduced over the last two decades. We have already discussed a number of these, in the process of developing and explaining our approach. In addition, the recent papers [26], [27] discuss particle filtering methods which reject (or at least attenuate the effect of) outliers. It is worth noting that this is slightly distinct from our approach: we are not assuming that the observations are necessarily departures from our underlying state-space model, and therefore we do not wish to reject or attenuate the observed data y_t. Indeed, our goal is to compute the optimal filter $p(q_t|Y_{1:t})$ accurately, assimilating all of the observed data $Y_{1:t}$, given the state-space model parameters (a task that the standard importance-sampling methods are unable to accomplish in many cases, as emphasized here).

Finally, it is natural to ask if our methods extend beyond the pairwise case to the problem of sampling from the L-lag conditional distributions $p(q_t, q_{t-1}, \ldots, q_{t-L}|Y_{1:t})$, as considered in [12], [13]. In some cases a similar reparameterization approach will be effective, but in much less generality than in the pairwise setting, since typically the L-lag conditional densities are not log-concave in the q variables, even if $p(q_t|y_t, q_{t-1})$ is log-concave in q_t. This makes the L-lag distributions much harder to sample in general, since local optima can trap the sampler, leading to slow mixing.

Acknowledgment

This work was supported by an NSF CAREER award, a McKnight Scholar award, and by the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Jack Judy, through the Space and Naval Warfare Systems Center, Pacific Grant/Contract No. N66001-11-1-4205.

References

[1] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. Springer Verlag, 2001.

[2] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, "Obstacles to high-dimensional particle filtering," *Monthly Weather Review*, vol. 136, no. 12, pp. 4629–4640, 2008.

[3] P. Bickel, B. Li, and T. Bengtsson, "Sharp failure rates for the bootstrap particle filter in high dimensions," *IMS Collections, Vol. 3, 318-329*, 2008.

[4] A. Beskos, D. Crisan, and A. Jasra, "On the stability of sequential Monte Carlo methods in high dimensions," *Arxiv preprint arXiv:1103.3965*, 2011.

[5] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2005.

[6] A. Doucet and A. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," *The Oxford Handbook of Nonlinear Filtering*, 2009.

[7] M. Pitt and N. Shephard, "Filtering via simulation: Auxiliary particle filters," *Journal of the American Statistical Association*, vol. 94, no. 446, pp. 590–599, 1999.

[8] C. Carvalho, M. Johannes, H. Lopes, and N. Polson, "Particle learning and smoothing," *Statistical Science*, vol. 25, no. 1, pp. 88–106, 2010.

[9] J. Weare, "Particle filtering with path sampling and an application to a bimodal ocean current model," *Journal of Computational Physics*, vol. 228, no. 12, pp. 4312–4331, 2009.

[10] R. Kass and A. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, pp. 773–795, 1995.

[11] R. Neal, "http://radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever/," *Radford Neal's blog*, 2008.

[12] W. Gilks and C. Berzuini, "Following a moving target Monte Carlo inference for dynamic Bayesian models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 1, pp. 127–146, 2001.

[13] A. Doucet, M. Briers, and S. Sénécal, "Efficient block sampling strategies for sequential Monte Carlo methods," *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 693–711, 2006.

[14] A. Brockwell, P. Del Moral, and A. Doucet, "Sequentially interacting Markov chain Monte Carlo methods," *The Annals of Statistics*, vol. 38, no. 6, pp. 3387–3411, 2010.

[15] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, 1989.

[16] G. McLachlan and D. Peel, *Finite mixture models*. Wiley-Interscience, 2000, vol. 299.

[17] S. Fruhwirth-Schnatter, *Finite Mixture and Markov Switching Models*. Springer, 2006.

[18] E. Marinari and G. Parisi, "Simulated tempering: a new Monte Carlo scheme," *Europhysics Letters*, vol. 19, p. 451, 1992.

[19] R. Neal, "Sampling from multimodal distributions using tempered transitions," *Statistics and Computing*, vol. 6, pp. 353–366, 1996.

[20] S. Godsill and T. Clapp, "Improvement strategies for Monte Carlo particle filters," *Sequential Monte Carlo methods in practice*, pp. 139–158, 2001.

[21] C. Sminchisescu and M. Welling, "Generalized darting Monte Carlo," *Pattern Recognition*, vol. 44, pp. 2738–2748, 2011.

[22] T. Minka, "A family of algorithms for approximate Bayesian inference," Ph.D. dissertation, MIT, 2001.

[23] S. Verboven and M. Hubert, "Libra: a matlab library for robust analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 2, pp. 127–136, 2005.

[24] M. Hubert, P. Rousseeuw, and T. Verdonck, "A deterministic algorithm for robust location and scatter," *In press, Journal of Computational and Graphical Statistics*, 2011.

[25] R. Neal, "Estimating ratios of normalizing constants using linked importance sampling," *Technical Report, University of Toronto*, 2005.

[26] C. Maiz, J. Miguez, and P. Djuric, "Particle filtering in the presence of outliers," in *IEEE/SP 15th Workshop on Statistical Signal Processing, SSP'09*, 2009, pp. 33–36.

[27] R. Kumar, D. Castanón, E. Ermis, and V. Saligrama, "A new algorithm for outlier rejection in particle filters," in *Information Fusion (FUSION), 2010 13th Conference on*. IEEE, 2010, pp. 1–7.

[28] W. Gerstner and W. Kistler, *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, 2002.