Variational Bayesian Filtering

Václav Šmídl, *Member, IEEE,*, Anthony Quinn, *Member, IEEE,*

***Abstract*—The use of the Variational Bayes (VB) approximation in Bayesian filtering is studied, both as a means to accelerate marginalized particle filtering, and as a deterministic local (one-step) approximation. The VB method of approximation is reviewed, together with restrictions that allow various computational savings to be achieved. These variants provide a range of algorithms that can be used in a principled trade-off between quality of approximation and computational cost. In combination with marginalized particle filtering, they generalize previously published work on variational filtering, and they extend currently available methods for speeding up stochastic approximations in Bayesian filtering. In particular, the free-form nature of the VB approximation allows optimal selection of moments which summarize the particles. Other Bayesian filtering schemes are developed by replacing the marginalization operator in Bayesian filtering with VB-marginals. This leads to further computational savings at the cost of quality of approximation. The performance of the various VB filtering schemes is illustrated in the context of a Gaussian model with a nonlinear sub-state, and a hidden Markov model.**

***Index Terms*—Bayesian filtering, Variational Bayes, particle filtering, EM algorithm, hidden Markov model.**

I. Introduction

In this paper, we are concerned with the classical problem of inferring the state variables which parameterize a sequence of observation models in the following manner:

$$d_t \sim f\left(d_t|\theta_t\right), \quad \theta_t \sim f\left(\theta_t|\theta_{t-1}\right). \tag{1}$$

Here, θ_t is a vector known as the state variable and d_t are the observations. By *Bayesian filtering*, we mean the recursive evaluation of the filtering distribution, $f\left(\theta_t|D_t\right)$, using Bayes' rule [9], [21], [29]:

$$f\left(\theta_t|D_{t-1}\right) \equiv f\left(\theta_1\right),\ t=1,$$
$$f\left(\theta_t|D_{t-1}\right) = \int f\left(\theta_t|\theta_{t-1}\right) f\left(\theta_{t-1}|D_{t-1}\right)\, d\theta_{t-1},\ t=2,..., \tag{2}$$

$$f\left(\theta_t|D_t\right) \propto f\left(d_t|\theta_t\right) f\left(\theta_t|D_{t-1}\right),\ \ t=1,2,.... \tag{3}$$

Here, $f\left(\theta_1\right)$ is the prior distribution, and $D_t = [d_1, \ldots, d_t]$ denotes the aggregated set of observations. The integration in (2), and elsewhere in this paper, is over the whole support of the integrand. (2) is referred to as the time update, and (3) as the data update, as illustrated in Fig. 1.

Bayesian filtering, as defined above, is just one formulation of the classical stochastic nonlinear filtering problem [2]. Other formulations exist, such as stochastic differential equations. Bayesian filtering has been used in signal processing for tasks such as blind deconvolution [15] in communication systems, speech recognition [3], and the design of navigation systems [25].

Bayesian filtering is analytically tractable if (i) the marginalization over θ_{t-1} in (2) is analytically tractable, and (ii) the resulting marginal distribution, $f\left(\theta_t|D_t\right)$, has the same functional form as the previous step, $f\left(\theta_{t-1}|D_{t-1}\right)$, allowing the procedure to be iterated. (i) and (ii) are satisfied for only a very limited class of models [7]. Typically, therefore, methods of distributional approximation are required. Specifically, we must replace $f\left(\theta_t|D_t\right)$ by a distributional approximation, $\tilde{f}\left(\theta_t|D_t\right)$.

Once again, we require that the functional form of the approximate distribution be preserved during the update. If this can be achieved via the exact Bayesian filtering steps (2,3), then the scheme is known as a *global* approximation. If, however, $\tilde{f}(\theta_t|D_t)$ is in a different functional class from $\tilde{f}(\theta_{t-1}|D_{t-1})$, then iterative approximation is required to restore the functional form at *each* time, t. Iterative approximation of this kind is known as a *local* approximation. The distinction between global and local approximations was first introduced in a non-Bayesian context [31], while a Bayesian interpretation was given in [14]. Global approximations for Bayesian filtering include point mass filters [5] and sequential Monte Carlo methods [9]. Local approximations include the extended Kalman filter [2] and the unscented filter [13]. Stochastic techniques have proved particularly popular for global approximation since they can achieve arbitrarily high accuracy. However their computational cost can be prohibitive. In contrast, local approximations tend to be deterministic, involving functional expansions of the distributions. Typically, these are computationally efficient but the error of approximation can accumulate with time. An active subject of research is to explore the combination of local and global approximations in an effort to improve computational efficiency without significant reduction in the accuracy of approximation [18], [26], [27]. Our aim in this paper is to explore the deterministic Variational Bayes approximation as a local approximation and in combination with global techniques.

The Variational Bayes (VB) approximation was developed in statistical physics—where it is known as the naïve mean-field approximation [20]—for non-recursive (off-line) inference of time-invariant parameters. One of the earliest applications was in off-line inference of hidden Markov models [17], and off-line signal processing applications of VB are now common [22]. The use of VB in recursive inference of time-invariant parameters was presented in [24], and an application to identification of extended autoregressive models appears in [23], [28]. Inference of the state trajectory in the linear

V. Šmídl is with the Institute of Information Theory and Automation, Prague, Czech Republic.

A. Quinn is with the Department of Electronic & Electrical Engineering, Trinity College Dublin, Ireland.

Manuscript received August 08, 2007; revised December 6, 2007.

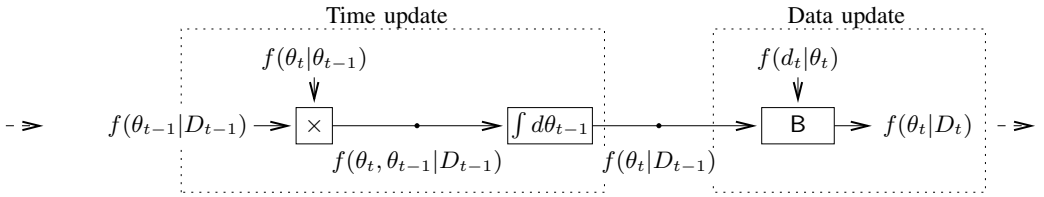

Fig. 1. An operator diagram illustrating Bayesian filtering. The '$\times$' operator denotes multiplication of arguments; '$\int d\cdot$' denotes marginalization of the stated parameter as in (2); and 'B' denotes an application of Bayes' rule as in (3).

Gaussian model defining the Kalman filter with unknown parameters was presented in [4], [33], but the inference was off-line (*i.e.* from a fixed number of data), avoiding marginalization over θ_{t-1} (2). The application of the VB approximation to the fundamental problem of Bayesian filtering—*i.e.* the requirement to marginalize over θ_{t-1} (2)—was studied for a constrained model class in [32], and also in [29]. Inconsistency of the VB approximation in the Kalman filter context was noted in both off-line [33] and on-line [29] scenarios.

The theory of the VB approximation is reviewed in Section II, including its simplified variants, namely functionally-constrained VB and restricted VB. These variants allow VB to be combined with other approximations, such as stochastic sampling which is briefly reviewed in Section III. The resulting *Variational Bayesian particle filtering* scheme is presented in Section IV. In this scheme, VB acts as a local approximation of the data update step in marginalized particle filtering, and therefore requires the time update step to be analytically tractable (Fig. 1). This requirement is relaxed in Section V, where VB is used to approximate both the time and data update steps.

II. The Variational Bayes Approximation

The Variational Bayes (VB) approximation is a deterministic, free-form technique for optimal distributional approximation, in the sense given by the following theorem [29].

Theorem 1: Let $f(\theta|D)$ be the posterior distribution of multivariate parameter, θ, and let $\theta = [\theta_1', \theta_2']'$ be a chosen partition into sub-vectors (where $'$ denotes transposition). Let $\breve{f}(\theta|D)$ be an approximating distribution restricted to the set of conditionally independent distributions:

$$\breve{f}(\theta|D) = \breve{f}(\theta_1, \theta_2|D) = \breve{f}(\theta_1|D)\,\breve{f}(\theta_2|D). \tag{4}$$

Any minimum of the following Kullback-Leibler divergence (KLD) from $\breve{f}(\cdot)$ to $f(\cdot)$,

$$KL(\breve{f}(\theta|D)\,||f(\theta|D)) = \int \breve{f}(\theta|D) \ln \frac{\breve{f}(\theta|D)}{f(\theta|D)} d\theta, \tag{5}$$

is achieved when $\breve{f}(\cdot) = \tilde{f}(\cdot)$, such that

$$\tilde{f}(\theta_i|D) \propto \exp\left(\mathsf{E}_{\tilde{f}(\theta_{/i}|D)}\left[\ln\left(f(\theta, D)\right)\right]\right),\ i = 1, 2. \tag{6}$$

Here, $\theta_{/i}$ denotes the complement of θ_i in θ. ■

We will refer to the $\tilde{f}(\theta_i|D)$ (6) as the *VB-marginals*. Also here, and throughout the paper, $\mathsf{E}_{f(\theta)}[g(\theta)]$ denotes the expected value of the function $g(\theta)$ with respect to the distribution $f(\theta)$. Since the functional form of each VB-marginal (6) emerges from a functional optimization, rather than being imposed beforehand, we will refer to VB as a *free-form* optimization technique. The theorem also extends to partitions of θ into several sub-vectors.

Theorem 1 provides a powerful tool for approximating probability (density) functions (p(d)fs) that exhibit a *separable form* [29]:

$$\ln f(\theta_1, \theta_2, D) = g(\theta_1, D)' h(\theta_2, D). \tag{7}$$

Here, $g(\theta_1, D)$ and $h(\theta_2, D)$ are finite-dimensional vectors. Using (7) in (6), the VB-marginals become

$$\tilde{f}(\theta_1|D) \propto \exp\left(g(\theta_1, D)'\, \widehat{h(\theta_2, D)}\right), \tag{8}$$

and similarly for θ_2. In (8), $\widehat{h(\cdot)} \equiv \mathsf{E}_{\tilde{f}(\theta_2|D)}[h(\cdot)]$ are the *necessary VB-moments* of $\tilde{f}(\theta_2|D)$. An *Iterative VB (IVB)* [29] moment-swapping algorithm is implied.

Algorithm 2.1 (Iterative VB (IVB) algorithm): Cyclic iteration of the following steps, $n = 1, 2, 3, \ldots$, monotonically decreases the KLD (5):

1) Compute the current update of the VB-marginal of θ_2 at iteration n, via (6):
$$\tilde{f}^{[n]}(\theta_2|D) \propto \exp\left\{\mathsf{E}_{\tilde{f}^{[n-1]}(\theta_1|D)}\left[g(\theta_1, D)'\right] h(\theta_2, D)\right\}. \tag{9}$$
2) Use the result of 1) to compute the current update of the VB-marginal of θ_1 at iteration n, via (6):
$$\tilde{f}^{[n]}(\theta_1|D) \propto \left\{g(\theta_1, D)'\, \mathsf{E}_{\tilde{f}^{[n]}(\theta_2|D)}[h(\theta_2, D)]\right\}. \tag{10}$$

Here, the initializer, $\tilde{f}^{[0]}(\theta_1|D)$, may be chosen as any tractable distribution. Convergence of the algorithm to fixed VB-marginals, $\tilde{f}^{[\infty]}(\theta_i|D)$, $\forall i$, was proved in [24].

In many nonlinear cases of g and/or h, the VB-marginals (6) will be non-standard in form, and so the required VB-moments will be difficult to evaluate. This underlines the free-form nature of the approximation (Section 1). It may be necessary in such cases to replace any non-standard VB-marginal—for example $\tilde{f}(\theta_2|D)$—with a tractable alternative. In the next two sub-sections, we present variants of the VB approximation designed to achieve this aim.

Remark 1: Constraints on the separable form of $f(\theta_1, \theta_2, D)$ (7) can be imposed as a way to ensure tractability of the VB-marginals (6), and such constraints yield special cases of Algorithm 2.1. An important constraint is the extended exponential family (EEF) assumption, which yields the propagation algorithm of [10]. When Algorithm 2.1 is applied to Bayesian networks with EEF nodes, the variational message-passing algorithm [34] emerges. However, in the

on-line (Bayesian filtering) context, the EEF constraint is too restrictive, since the filtering distribution is required to be functionally invariant under the marginalization in (2) (Section I). We relax the EEF assumption in the Variational Bayesian Filtering techniques that we develop in Sections IV and V, and instead we achieve tractability using the VB variants of the next two sub-sections.

A. The Functionally-Constrained VB Approximation

In this case, an extra step is introduced within each IVB cycle. Specifically, $\tilde{f}(\theta_2|D)$ is projected into a tractable alternative, $\hat{f}(\theta_2|D)$, via a subsidiary projection rule. It is the moments of *this* distribution that are fed back to $\tilde{f}(\theta_1|D)$ (8). In fact, the popular Expectation-Maximization (EM) algorithm [8] is a case in point, where

$$\hat{f}(\theta_2|D) \equiv \delta\left(\theta_2 - \hat{\theta}_2\right), \tag{11}$$

and $\hat{\theta}_2 = \arg\max_{\theta_2} \tilde{f}(\theta_2|D)$ is the *mode* of the second VB-marginal. In this case, $\tilde{f}(\theta_1|D)$ becomes

$$\tilde{f}(\theta_1|D) \equiv f\left(\theta_1|\hat{\theta}_2, D\right),$$

via (8) and the sifting property of the Dirac δ-function, $\delta(\cdot)$.

B. The Restricted VB (RVB) Approximation

We replace $\breve{f}(\theta_2|D)$ by a tractable *fixed* distribution, $\overline{f}(\theta_2|D)$. Then, by Theorem 1:

$$\tilde{f}(\theta_1|D) \propto \exp\left(\mathsf{E}_{\overline{f}(\theta_2|D)}\left[\ln\left(f(\theta, D)\right)\right]\right). \tag{12}$$

Hence, a single substitution of necessary moments from $\overline{f}(\cdot)$ is required and IVB cycles are avoided. It is interesting to note that a number of popular distributional approximations are special cases of (12): (i) *certainty equivalence*, where $\overline{f} \equiv \delta(\theta_2 - \hat{\theta}_2)$ for some fixed $\hat{\theta}_2$, in which case (12) becomes $f\left(\theta_1|\hat{\theta}_2, D\right)$, i.e. the conditional; and (ii) the *Quasi-Bayes approximation*, where $\overline{f} \equiv f(\theta_2|D)$, the exact marginal [29]. If $h(\cdot)$ in (7) is linear, then (i) and (ii) are equivalent for the choice $\hat{\theta}_2 = \mathsf{E}_{f(\theta_2|D)}[\theta_2]$.

We conclude this section by listing the following distributional objects, which will be used in this paper:

$f(\theta\|D)$	the exact distribution/model;
$\tilde{f}(\theta\|D)$	a VB-based (free-form) optimizer;
$\hat{f}(\theta\|D)$	a functionally-constrained projection of $\tilde{f}(\theta\|D)$;
$f_\delta(\theta\|D)$	special case of $\hat{f}(\cdot)$, being an empirical approximation of $f(\theta\|D)$ (Section III-A);
$\overline{f}(\theta\|D)$	a fixed distribution.

C. The choice of restrictions

The two variants of the VB scheme presented above allow other distributional approximations to be combined with VB. We will be particularly interested in the rôle of the empirical approximation as a functionally-constrained distribution (see Section II-A). As we will see in Section IV, the resulting computational scheme is closely related to Sequential Importance Sampling, which we will now review.

III. Sequential Importance Sampling

A. Particle Filtering

Particle filtering (PF) [9] refers to a range of techniques for generating an empirical approximation of $f(\Theta_t|D_t)$, where $\Theta_t = [\theta_1, \ldots, \theta_t]$ is the state trajectory:

$$f(\Theta_t|D_t) \approx f_\delta(\Theta_t|D_t) \equiv \frac{1}{n}\sum_{i=1}^{n} \delta\left(\Theta_t - \Theta_t^{(i)}\right). \tag{13}$$

In this paper, we reserve the symbol $f_\delta(\cdot)$ for the (possibly weighted) empirical approximation of $f(\cdot)$. In (13), $\Theta_t^{(i)}$, $i = 1, \ldots, n$, are i.i.d. samples from the posterior, and so this approach is feasible only if we can sample from the exact posterior, $f(\Theta_t|D_t)$. If this is not the case (as commonly arises), we can draw samples from a chosen proposal distribution (importance function), $q(\Theta_t|D_t)$, as follows:

$$\begin{aligned} f(\Theta_t|D_t) &= \frac{f(\Theta_t|D_t)}{q(\Theta_t|D)} q(\Theta_t|D_t) \\ &\approx \frac{f(\Theta_t|D_t)}{q(\Theta_t|D_t)} \frac{1}{n}\sum_{i=1}^{n} \delta\left(\Theta_t - \Theta_t^{(i)}\right). \end{aligned} \tag{14}$$

In this case, $\Theta_t^{(i)} \sim q(\cdot)$. Using the sifting property of the Dirac δ-function, (14) can be written in the form of a *weighted* empirical approximation, as follows:

$$f(\Theta_t|D_t) \approx f_\delta(\Theta_t|D_t) \equiv \sum_{i=1}^{n} w_t^{(i)} \delta\left(\Theta_t - \Theta_t^{(i)}\right), \tag{15}$$

$$w_t^{(i)} \propto \frac{f\left(\Theta_t^{(i)}|D_t\right)}{q\left(\Theta_t^{(i)}|D_t\right)}. \tag{16}$$

Under this *importance sampling* procedure, the true posterior distribution, $f(\cdot)$, need only be evaluated point-wise. Furthermore, the normalizing constant of $f(\cdot)$ is not required, since (15) is normalized trivially via the constant $c = \sum_{i=1}^{n} w_t^{(i)}$.

In Bayesian filtering (2,3), the challenge is to generate the samples, $\theta_t^{(i)}$, and evaluate the importance weights, $w_t^{(i)}$, recursively. Using (1) and standard Bayesian calculus, (16) can be written in the following recursive form:

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_t^{(i)}\right) f\left(\theta_t^{(i)}|\theta_{t-1}^{(i)}\right)}{q\left(\theta_t^{(i)}|\Theta_{t-1}^{(i)}, D_t\right)} w_{t-1}^{(i)}. \tag{17}$$

Now, $\theta_t^{(i)}$ are drawn from the denominator of (17), which can be chosen as $f(\theta_t|\theta_{t-1})$ (1). Thus, the weighted empirical form of the posterior distribution (15) is preserved during each Bayesian filtering update (Fig. 1), as is characteristic of a *global* approximation. Successful implementation of the particle filter involves other considerations such as re-sampling, appropriate choice of the importance function, *etc.* [9].

B. Marginalized Particle Filtering

The main advantage of importance sampling is its generality. However, in cases where θ_t is high-dimensional, it may be computationally prohibitive to generate the required particles, $\theta_t^{(i)}$. Furthermore, it is necessary to generate large numbers

of such particles in these high-dimensional cases, in order to achieve an acceptable error of approximation. These problems can be overcome when the structure of the model (1) allows analytical marginalization over a subset, $\theta_{1,t}$, of the full state vector $\theta'_t = [\theta'_{1,t}, \theta'_{2,t}]$ [9], [25]. Therefore, we consider the factorization

$$f(\Theta_t|D_t) = f(\Theta_{1,t}|\Theta_{2,t}, D_t)\, f(\Theta_{2,t}|D_t), \tag{18}$$

where $f(\Theta_{1,t}|\Theta_{2,t}, D_t)$ is analytically tractable, while $f(\Theta_{2,t}|D_t)$ is not. We replace the latter by a weighted empirical approximation, in analogy to (14), yielding

$$f(\Theta_t|D_t) \approx \sum_{i=1}^{n} w_t^{(i)} f\left(\Theta_{1,t}|\Theta_{2,t}^{(i)}, D_t\right) \delta\left(\Theta_{2,t} - \Theta_{2,t}^{(i)}\right), \tag{19}$$

$$w_t^{(i)} \propto \frac{f\left(\Theta_{2,t}^{(i)}|D_t\right)}{q\left(\Theta_{2,t}^{(i)}|D_t\right)}. \tag{20}$$

Note that we now need to sample only from the space of $\theta_{2,t}$. The weights can, once again, be evaluated recursively:

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_{2,t}^{(i)}\right) f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}\right)}{q\left(\theta_{2,t}^{(i)}|\Theta_{2,t-1}^{(i)}, D_t\right)} w_{t-1}^{(i)}. \tag{21}$$

From (19), we note that $f(\Theta_{1,t}|D_t)$ is a mixture (convex combination) of conditional Bayesian filters. Hence, the model (1) must admit a partition, $\left[\theta'_{1,t}, \theta'_{2,t}\right]$, for which $\theta_{1,t-1}$ can be integrated analytically in (2) and the resulting $f(\theta_{1,t}|\Theta_{2,t}, D_t)$ (3) is in the same form as for the previous step. The marginalized particle filter (19–21) can then be evaluated exactly. This requirement is always fulfilled if the model can be decomposed into linear and nonlinear parts [25], and may even be possible for a wider class of models [7]. Under these conditions, the form of (19) is preserved under an exact Bayesian filtering update (empirical in $\Theta_{2,t}$, mixture of functionally-invariant conditional filters in $\Theta_{1,t}$), and is therefore a global approximation (Section I). (19)–(21) is sometimes referred to as the Rao-Blackwellized particle filter [9].

C. Accelerating the Marginalized Particle Filter (MPF)

The mixture in (19) requires n parallel conditional Bayesian filtering updates, i.e. sufficient statistics are required for each particle trajectory $\Theta_{2,t}^{(i)}$, as displayed in Fig. 2 (left). This is computationally inefficient if the particle trajectories are similar. A Certainty Equivalence (CE) approach to reducing the computational cost of the MPF was reported in [18]. The idea was to replace the n trajectories by their weighted arithmetic mean, $\hat{\Theta}_{2,t} = \sum_{i=1}^{n} w_t^{(i)} \Theta_{2,t}^{(i)}$, and perform a single Bayesian filtering update (see Fig. 2 (right) for the implied flow-of-control). This corresponds to replacing each of the n conditional Bayesian filters in the mixture (19) by a single component, as follows:

$$f\left(\Theta_{1,t}|\Theta_{2,t}^{(i)}, D_t\right) \approx f\left(\Theta_{1,t}|\hat{\Theta}_{2,t}, D_t\right), \quad i = 1, \ldots, n. \tag{22}$$

From (19):

$$f(\Theta_t|D_t) \approx f\left(\Theta_{1,t}|\hat{\Theta}_{2,t}, D_t\right) \sum_{i=1}^{n} w_t^{(i)} \delta\left(\Theta_{2,t} - \Theta_{2,t}^{(i)}\right). \tag{23}$$

Note that this has a conditional independence structure, as in (4). One step of particle filtering, using (23), will again generate n components in $\Theta_{1,t}$, and so the proposed conflation into one component is required after *each* update, as is characteristic of a local approximation. In [18], this local approximation was chosen heuristically, *i.e.* conditioning on the weighted mean, $\hat{\Theta}_{2,t}$, as explained above. Note that the empirical form in $\Theta_{2,t}$ is preserved during each update (see Remark 4 to follow). The associated weights, $w_t^{(i)}$, were evaluated in [18] using a modified version of (21):

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_{1,t}^{(i)}, \theta_{2,t}^{(i)}\right) f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}\right)}{q\left(\theta_{2,t}^{(i)}|\Theta_{2,t-1}^{(i)}, D_t\right)} w_{t-1}^{(i)}. \tag{24}$$

Here, $\theta_{1,t}^{(i)}$ are samples from $f(\theta_{1,t}|\theta_{2,t}, D_t)$. The idea is closely related to the mean-field approach [20], as noted in [30]. A considerable extension is therefore possible using other mean-field approximations, and, in particular, the Variational Bayes approximation, introduced in Section II. In the next Section, we will find that this leads to a principled approach to the problem of concentrating the n components in the mixture $f(\Theta_{1,t}|D_t)$ (19) into a single component such as that achieved in (23).

IV. Variational Bayesian Particle Filtering

We now develop a local approximation of the posterior distribution, $f(\theta_t|D_t)$ (3), via the VB approximation. To achieve this, we once again partition the parameters into $\theta_t = \left[\theta'_{1,t}, \theta'_{2,t}\right]'$, such that

$$\tilde{f}(\theta_{t-1}|D_{t-1}) = \tilde{f}(\theta_{1,t-1}|D_{t-1})\, \tilde{f}(\theta_{2,t-1}|D_{t-1}). \tag{25}$$

Here we have assumed that the VB approximation (4,6) has been applied at the previous time step, $t-1$. In common with the MPF (Section III), we assume that the following marginal is available analytically:

$$f(\theta_{1,t}, \theta_{2,t}|\theta_{2,t-1}, D_t) \propto \int f(d_t|\theta_t)\, f(\theta_t|\theta_{t-1})\, \tilde{f}(\theta_{1,t-1}|D_{t-1})\, d\theta_{1,t-1}. \tag{26}$$

We now apply the VB approximation as a local approximation at time t, as follows:

$$f(\theta_{1,t}, \theta_{2,t}|\theta_{2,t-1}, D_t) \approx \tilde{f}(\theta_{1,t}, \theta_{2,t}|\theta_{2,t-1}, D_t) = = \tilde{f}(\theta_{1,t}|\theta_{2,t-1}, D_t)\, \tilde{f}(\theta_{2,t}|\theta_{2,t-1}, D_t). \tag{27}$$

As before, the necessary VB-moments of both VB-marginals, $\tilde{f}(\theta_{1,t}|\theta_{2,t-1}, D_t)$ and $\tilde{f}(\theta_{2,t}|\theta_{2,t-1}, D_t)$, must be available, $\forall t$. However, since the VB approximation is a free-form optimization, as explained in Section II, one or both of these VB-marginals may not be in standard form. Assuming, therefore, that the VB-moments of $\tilde{f}(\theta_{2,t}|\theta_{2,t-1}, D_t)$ are not available analytically, we now explore the variants of the VB approximation introduced in Sections II-A and II-B.

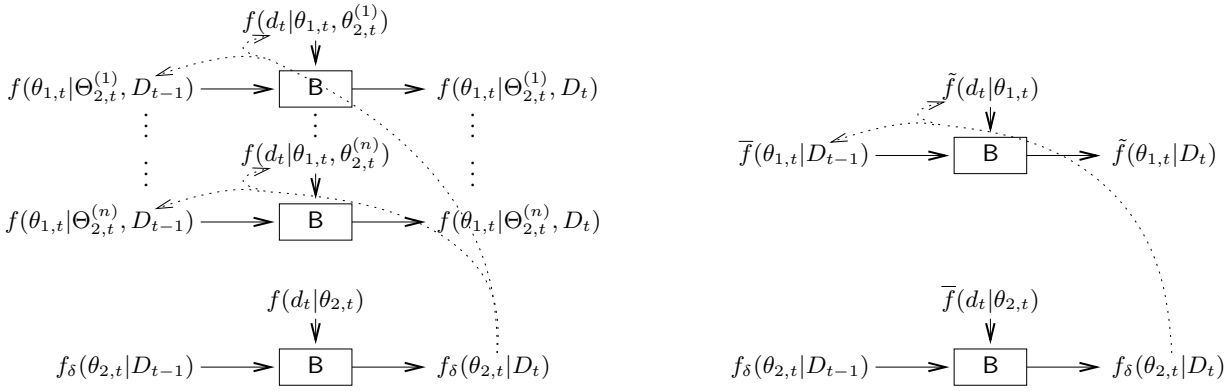

Fig. 2. **Left:** operator diagram for the data-update step in marginalized particle filtering, involving n parallel particle-conditioned updates. The conditioning on particles is represented by dotted arrows. **Right:** the restricted VB (*i.e.* RVB) particle filter (Section IV-B), involving a single substitution of VB-moments (dotted arrows). Under the certainty equivalence (CE) approach (see Section III-C and Remark 5), only first-order moments are substituted.

A. VB Particle Filtering

Using the functionally-constrained VB approximation (Section II-A), we project $\tilde{f}\left(\theta_{2,t}|\theta_{2,t-1},D_t\right)$ into a (weighted) empirical approximation,

$$\hat{f}\left(\theta_{2,t}|\theta_{2,t-1},D_t\right) \equiv$$
$$\equiv f_\delta\left(\theta_{2,t}|\theta_{2,t-1},D_t\right)=\sum_{i=1}^{n}\omega_t^{(i)}\delta\left(\theta_{2,t}-\theta_{2,t}^{(i)}\right), \quad (28)$$

$$\omega_t^{(i)} \propto \frac{\tilde{f}\left(\theta_{2,t}^{(i)}|\theta_{2,t-1},D_t\right)}{q\left(\theta_{2,t}^{(i)}|\theta_{2,t-1},D_t\right)}, \quad (29)$$

where $\theta_{2,t}^{(i)}\sim q(\cdot)$. The necessary VB-moments (8,10) of (28) are now readily available:

$$\mathsf{E}_{f_\delta(\theta_{2,t}|\theta_{2,t-1},D_t)}\left[h(\theta_{2,t},\theta_{2,t-1},D_t)\right]=$$
$$=\sum_{i=1}^{n}\omega_t^{(i)}h\left(\theta_{2,t}^{(i)},\theta_{2,t-1},D_t\right). \quad (30)$$

Note that the terms in (29) and (30) are conditioned on $\theta_{2,t-1}$. The following adaptation of the IVB algorithm (Section 2) is implied:

Algorithm 4.1 (VB Particle Filtering):

1. Draw samples $\theta_{2,t}^{(i)}$ from $q\left(\theta_{2,t}|\theta_{2,t-1}^{(i)},D_t\right)$.
2. Evaluate moments $\mathsf{E}_{\tilde{f}(\theta_{1,t}|\theta_{2,t-1},D_t)}\left[g(\theta_{1,t},\theta_{2,t-1},D_t)\right]\equiv\widehat{g(\theta_{1,t})}$, and generate
$$\tilde{f}\left(\theta_{2,t}|\theta_{2,t-1},D_t\right)\propto\exp\left\{\widehat{g(\theta_{1,t})}h(\theta_{2,t},\theta_{2,t-1},D_t)\right\},$$
using (26)—written in separable form (7)—in (9).
3. Evaluate weights $\omega_t^{(i)}$ using (29).
4. Evaluate moments $\mathsf{E}_{f_\delta(\theta_{2,t}|\theta_{2,t-1},D_t)}\left[h(\theta_{2,t},\theta_{2,t-1},D_t)\right]\equiv\widehat{h(\theta_{2,t})}$, using (30), and generate
$$\tilde{f}\left(\theta_{1,t}|\theta_{2,t-1},D_t\right)\propto\exp\left\{g(\theta_{1,t},\theta_{2,t-1},D_t)\widehat{h(\theta_{2,t})}\right\},$$
using (26)—again written in separable form (7)—in (10).
5. If not converged, go to step 2.
6. Generate the empirical marginal, $f_\delta\left(\theta_{2,t}|D_t\right)$, using samples $\theta_{2,t}^{(i)}$ from Step 1 and the following (marginal) weights, calculated via Step 3:
$$w_t^{(i)}=\omega_t^{(i)}w_{t-1}^{(i)}. \quad (31)$$
7. Generate the following marginal via Steps 4 and 6:
$$\tilde{f}\left(\theta_{1,t}|D_t\right)\propto\exp\left(\mathsf{E}_{f_\delta(\theta_{2,t-1}|D_{t-1})}\left[g\left(\theta_{1,t},\theta_{2,t-1},D_t\right)\right.\right.$$
$$\left.\left.\mathsf{E}_{f_\delta(\theta_{2,t}|\theta_{2,t-1},D_t)}\left[h(\theta_{2,t},\theta_{2,t-1},D_t)\right]\right]\right). \quad (32)$$

Remark 2: The marginal (32) (Step 7) is required in (26). Further computational savings may be achieved using $\tilde{f}\left(\theta_{1,t}|D_t\right)$ as a functional constraint in (27); *i.e.*

$$f\left(\theta_{1,t},\theta_{2,t}|\theta_{2,t-1},D_t\right)\approx\tilde{f}\left(\theta_{1,t}|D_t\right)\tilde{f}\left(\theta_{2,t}|\theta_{2,t-1},D_t\right),$$

eliminating the need for Step 7.

Remark 3: VB particle filtering (Algorithm 4.1) imposes no modelling assumptions beyond those of the marginalized particle filter (19)–(21), namely the requirement for analytical marginalization over a subset, $\theta_{1,t-1}$ (26). The approximated distribution (27) therefore depends in general on the remaining (non-analytical) parameters, $\theta_{2,t-1}$. In this way, the empirical approximation (28) involves a complete particle filtering update, with re-sampling, generating particle trajectories, $\Theta_{2,t}^{(i)}$. The joint posterior distribution is generated at each time by moment-swapping (IVB cycles, Algorithm 2.1) between this particle filter and the VB-marginal in $\theta_{1,t}$ (Step 4).

A related algorithm for tracking—known as variational sequential estimation—was introduced in [32], but there the model imposed independence assumptions which avoided the intractable marginalization over $\theta_{2,t-1}$. Sampling was used only for moment evaluation in these parameters, without any relation to the previous samples, leaving the advantages of sequential sampling (such as re-sampling) unused. From a practical point-of-view, VB particle filtering (Algorithm 4.1) is applicable to a wider class of models, including models with non-linear transformations, such as the one in Section IV-C, to follow.

Remark 4: The VB approximation acts as a local approximation only in respect of $\theta_{1,t}$, serving to conflate the n components of the mixture in (19) into a single component (32) at

each step. From (27) and (32), we note that neither the particle trajectories, $\Theta_{2,t}^{(i)}$ (19), nor associated sufficient statistics, need to be stored. However, an empirical approximation for $\theta_{2,t}$ (28) is generated at each update of the procedure, consistent with a global approximation (Section I). In this sense, VB particle filtering is a *semi-global* approximation.

B. Restricted VB (RVB) Particle Filtering

In cases where the un-normalized analytical marginal, $f\left(\theta_{2,t}|\theta_{2,t-1}, D_t\right)$, from (26), can be evaluated pointwise, we can replace $\tilde{f}\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}, D_t\right)$ by $f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}, D_t\right)$ in (29), giving

$$\omega_t^{(i)} \propto \frac{f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}, D_t\right)}{q\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}, D_t\right)}. \tag{33}$$

Hence, IVB iterations are avoided, as is characteristic of any RVB approximation (Section II-B). The following non-iterative algorithm is implied.

Algorithm 4.2 (RVB Particle Filtering):

1. Draw samples $\theta_{2,t}^{(i)}$ from $q\left(\theta_{2,t}|\theta_{2,t-1}^{(i)}, D_t\right)$.
2. Evaluate weights $\omega_t^{(i)}$ using (33).
3. Evaluate moments $\mathsf{E}_{f_\delta(\theta_{2,t}|\theta_{2,t-1},D_t)}\left[h(\theta_{2,t}, \theta_{2,t-1}, D_t)\right]$ using (30).
4. Generate $\tilde{f}\left(\theta_{1,t}|D_t\right)$, using (32), and $f_\delta(\theta_{2,t}|D_t)$, using (31).

Remark 5: In these VB scenarios, the n particles in (19) have been concentrated into $\tilde{f}\left(\theta_1|\theta_{2,t-1}, D_t\right)$ via (30) and (10), eliminating the need for n parallel Bayesian filtering steps, as illustrated in Fig. 2 (right). Note that (23), as proposed in [18], is a special case of RVB particle filtering above, assuming that (i) the weights, $w_t^{(i)}$, are fixed via (24), and (ii) $f_\delta(\theta_{2,t}|\cdot)$ is further constrained to the CE approximation, $\delta(\theta_{2,t} - \hat{\theta}_{2,t})$, where $\hat{\theta}_{2,t} = \sum_{i=1}^n w_t^{(i)}\theta_{2,t}^{(i)}$.

The general case of RVB particle filtering (Algorithm 4.2) therefore generalizes the CE approach by exploiting higher-order moments of the empirical approximation (30). The necessary VB-moments are, once again, those implied by the free-form nature of the VB approximation [29].

C. Illustrative Example: Gaussian model with nonlinear sub-state

We consider the following state-space model for multivariate observations, d_t, conditioned on multivariate hidden variables, x_t. It is similar to the model examined in [18]:

$$\begin{aligned} f\left(d_t|x_t, C_t\right) &= \mathcal{N}\left(C_t' x_t, R\right), \\ f\left(x_t|x_{t-1}\right) &= \mathcal{N}\left(Ax_{t-1}, Q\right), \\ f\left(c_{i,j,t}|c_{i,j,t-1}\right) &= \mathcal{U}\left(\gamma(c_{i,j,t-1}), 1 + \gamma(c_{i,j,t-1})\right), \forall i,j, \\ \gamma(c_{i,j,t}) &= 2 + \arctan\left(c_{i,j,t} - 2\right). \end{aligned} \tag{34}$$

Here, matrices R, A and Q are assumed to be known, and $c_{i,j,t}$ denotes the element (i,j) of unknown time-variant matrix, C_t. Essentially, this is a standard linear-Gaussian model with unknown time-variant C_t, for which a non-linear evolution model is defined. The full set of state variables is therefore $\theta_t = \{x_t, C_t\}$ (18). $\mathcal{U}(\cdot,\cdot)$ denotes the rectangular distribution on the indicated semi-closed interval.

Integration over x_{t-1} is possible using standard Kalman Filtering (KF) theory, yielding the following conditional filtering distribution for x_t:

$$f\left(x_t|C_t, D_t\right) = \mathcal{N}\left(\mu_t, \Omega_t^{-1}\right), \tag{35}$$

$$\begin{aligned} \Omega_t &= \left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1} + C_t R^{-1} C_t', \\ \mu_t &= \Omega_t^{-1}\left[\left(Q + A\Omega_{t-1}^{-1}A'\right)^{-1} A\mu_{t-1} + C_t R^{-1} d_t\right]. \end{aligned} \tag{36}$$

This is written in terms of precision matrix Ω_t for analytical convenience. Exact integration over C_{t-1} is intractable. A marginalized particle filter (MPF) (Section III-B) is obtained using (19)–(21).

1) VB Particle Filtering: The required distribution (26), *i.e.* $f\left(x_t, C_t|C_{t-1}, D_t\right)$, is obtained by multiplying (35) by (34). Application of the VB theorem (Section II) yields the following VB-marginals:

$$\tilde{f}\left(x_t|C_{t-1}, D_t\right) = \mathcal{N}\left(\tilde{\mu}_t, \tilde{\Omega}_t^{-1}\right), \tag{37}$$

$$\tilde{f}\left(C_t|C_{t-1}, D_t\right) \propto \mathcal{N}\left(\tilde{C}_t,\ \ I \otimes \tilde{\Sigma}_t\right)|\Omega_t(C_t)|^{\frac{1}{2}}. \tag{38}$$

The support of (38) is restricted to that of (34). I is the identity matrix of appropriate dimensions, $\otimes$ is the Kronecker product, $\Omega_t(C_t)$ is defined by (36), with the dependence on C_t now shown explicitly, and $|\cdot|$ denotes the matrix determinant. The dependence on the latter causes (38) to be intractable (for example, its normalizing constant and moments are not available in closed form). The shaping parameters of the distributions (37,38) are

$$\begin{aligned} \tilde{\Omega}_t &= \left(Q + A\tilde{\Omega}_{t-1}^{-1}A'\right)^{-1} + \mathsf{E}\left[C_t R^{-1} C_t'\right], \\ \tilde{\mu}_t &= \tilde{\Omega}_t^{-1}\left[\left(Q + A\tilde{\Omega}_{t-1}^{-1}A'\right)^{-1} A\tilde{\mu}_{t-1} + \widehat{C_t} R^{-1} d_t\right], \\ \tilde{\Sigma}_t &= \left(R^{-1}\widehat{x_t x_t'}\right)^{-1}, \\ \tilde{C}_t &= \tilde{\Sigma}_t\left(R^{-1} d_t \widehat{x}_t'\right). \end{aligned} \tag{39, 40}$$

$\widehat{C_t}$ and $\mathsf{E}\left[C_t R^{-1} C'_t\right]$ are the first and second moments, respectively, of the empirical approximation, $f_\delta\left(C_t|C_{t-1}, D_t\right)$, evaluated via (28,29), while $\widehat{x_t}$ and $\widehat{x_t x_t'}$ are the first and second moments of the tractable VB-marginal (37).

Remark 6: Under RVB particle filtering (Section II-B), the weights are evaluated via (33) rather than via (29). In order to study the influence of higher-order moments, we also consider a certainty equivalent restriction in the RVB scheme (RVB-CE), such that $\mathsf{E}\left[C_t R^{-1} C_t'\right]$ in (39) is replaced by $\widehat{C_t} R^{-1} \widehat{C_t}'$. This replacement yields the same form of Kalman filter as was used in [18].

2) Simulation study: The system (34) was simulated with d_t and x_t each in $\mathbb{R}^2$, $C_t \in \mathbb{R}^{2\times 2}$, and the following

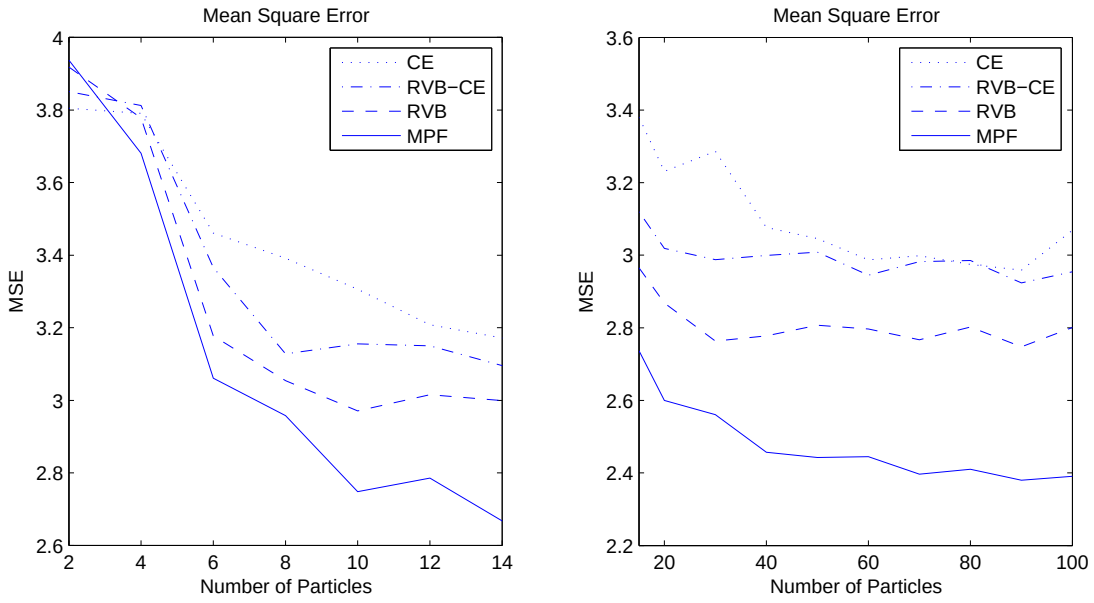

Fig. 3. Comparative MSE performance of the marginalized particle filter (MPF) and Restricted VB variants (RVB, RVB-CE, CE) for state estimation in the model (34). For each setting (*i.e.* number of particles), 100 Monte Carlo trials were undertaken.

parameters and initialization:

$$\begin{aligned} A &= \begin{bmatrix} 1 & -0.5 \\ 1 & 0 \end{bmatrix}, \qquad Q = 0.5 I_2, \qquad R = 0.5 I_2, \\ C_0 &= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \\ C_0^{(i)} &= C_0, \quad x_0^{(i)} = x_0, \quad \forall i = 1, \ldots, n. \end{aligned}$$

We use the same proposal density, $q\left(C_t|C_{t-1}\right) \equiv f(C_t|C_{t-1})$ (34), and the residual re-sampling scheme [16] for all tested methods. The aim is to illustrate the effect of the propagation of higher moments (30,39) within the VB particle filtering schemes. The performance was assessed via the following Mean Square Error (MSE) of the state estimate:

$$MSE = \frac{1}{t_u} \sum_{t=1}^{t_u} \left[||x_t - \widehat{x_t}||^2 + ||C_t - \widehat{C_t}||^2 \right].$$

Here, t_u is the total number of data, d_t, used in the simulation, x_t is the simulated trajectory of the linear sub-state, and $\widehat{x_t} = \tilde{\mu}_t$ is the mean of the VB-marginal, $\tilde{f}\left(x_t|C_{t-1}, D_t\right)$ (37). Similarly, C_t is the simulated trajectory of the nonlinear sub-state, and $\widehat{C_t}$ is its posterior mean, evaluated via each of the following four distributional approximations: **MPF:** marginalized particle filter (19)–(21); **RVB:** restricted VB approximation, evaluating the second-order moments in (39) via (30,33); **RVB-CE:** further approximation of the previous method, via the certainty-equivalent replacement of the second moment (Remark 6); and **CE:** the method presented in [18], using certainty equivalence and sampling from $f\left(x_t|x_{t-1}^{(i)}, C_t, D_t\right)$ (23,24). Results of a Monte Carlo study with $t_u = 100$, using 100 runs per setting, and for a varying number of particles, n, are displayed in Fig. 3.

All of the RVB (mean-field) methods, *i.e.* RVB, RVB-CE and CE, perform comparably to the MPF when the number of particles is low (Fig. 3, left). However, with an increasing number of particles, the performance of the MPF improves significantly, while that of all RVB methods improves far more slowly (Fig. 3, right). The clear performance improvement of RVB over the CE variants (RVB-CE and CE) in all cases is due to the propagation of second-order moments in (39). The RVB methods are approximately twice as fast as the MPF with the same number of particles. Hence, RVB particle filtering (Section IV-B) is recommended in cases where only a small number of particles can be computed. These cases arise in computationally constrained environments, such as in embedded devices.

V. Local Variational Bayesian Filtering

The semi-global approach introduced in Section IV is only possible in cases where analytical integration over some subset, $\theta_{1,t-1}$, is tractable (26) (*i.e.* the time update step (2) is tractable in $\theta_{1,t}$). This is not the case for many models (Section I), and so we investigate the rôle of the VB approximation as a *local* approximation, replacing the exact marginalization in the time-update of Bayesian filtering (2) with a VB-marginal (6). Formally, we impose conditional independence between θ_t and θ_{t-1}:

$$\tilde{f}\left(\theta_t, \theta_{t-1}|D_t\right) = \tilde{f}(\theta_t|D_t)\, \tilde{f}(\theta_{t-1}|D_t)\,. \tag{41}$$

The joint distribution needed in (6) is

$$f\left(d_t, \theta_t, \theta_{t-1}|D_{t-1}\right) = f\left(d_t|\theta_t\right) f\left(\theta_t|\theta_{t-1}\right) \tilde{f}\left(\theta_{t-1}|D_{t-1}\right). \tag{42}$$

Application of Theorem 1 to (42) yields VB-marginals in the form of two parallel Bayes' rule updates:

$$\tilde{f}\left(\theta_t|D_t\right) \propto f\left(d_t|\theta_t\right) \tilde{f}\left(\theta_t|D_{t-1}\right), \tag{43}$$

$$\tilde{f}\left(\theta_{t-1}|D_t\right) \propto \tilde{f}\left(d_t|\theta_{t-1}\right) \tilde{f}\left(\theta_{t-1}|D_{t-1}\right). \tag{44}$$

The following approximate distributions are involved:

$$\tilde{f}\left(\theta_t|D_{t-1}\right) \propto \exp\left\{\mathsf{E}_{\tilde{f}(\theta_{t-1}|D_t)}\left[\ln f\left(\theta_t|\theta_{t-1}\right)\right]\right\}, \quad (45)$$

$$\tilde{f}\left(d_t|\theta_{t-1}\right) \propto \exp\left\{\mathsf{E}_{\tilde{f}(\theta_t|D_t)}\left[\ln f\left(\theta_t|\theta_{t-1}\right)\right]\right\}. \quad (46)$$

The resulting iterative scheme, involving two VB-moment substitutions (dotted arrows), and two Bayes' rule updates, is illustrated in Fig. 4 (left). From (45), we note that the functional form of $\tilde{f}\left(\theta_t|D_{t-1}\right)$ is determined by the form of the time-invariant parameter evolution model $f\left(\theta_t|\theta_{t-1}\right)$. Therefore, notwithstanding the fact that VB is a free-form approximation (Section I), the *same* functional form is recovered at each time, t, via (43). Thus, a constant computational cost is incurred per time step. This important property of the approximation scheme is known as VB-conjugacy [29].

A. Restricted Variational Bayes (RVB) Filtering

Once again, the functional forms of the free-form optimization (Theorem 1) may be intractable. In particular, $\tilde{f}\left(\theta_{t-1}|D_t\right)$ in (44), which is the result of two Bayes' rule updates, may be intractable. An obvious restriction (Section II-B) is to replace it by the fixed VB-filtering distribution from the previous step (43):

$$\overline{f}\left(\theta_{t-1}|D_t\right) \equiv \tilde{f}\left(\theta_{t-1}|D_{t-1}\right). \quad (47)$$

The resulting VB-marginal of θ_t has the same form as (43), but the expectations in (45) are now taken with respect to $\tilde{f}\left(\theta_{t-1}|D_{t-1}\right)$ instead of $\tilde{f}\left(\theta_{t-1}|D_t\right)$. IVB iterations are therefore avoided, as illustrated in Fig. 4 (right).

B. Illustrative Example: Classification with a Hidden Markov Model (HMM)

Off-line variational inference methods for the HMM have been known for a long time [17], [12]. Since the HMM satisfies the extended exponential family (EEF) assumptions referred to in Remark 1, its combination with other EEF members in a Bayesian network is straightforward [10]. Recently, off-line variational inference for a mixture of HMMs was reported [22]. In these off-line contexts, both the hidden field and the time-invariant transition matrix are inferred by iterative forward and backward passes through the data.

In the on-line context of this paper, the backward pass is replaced by the time update step, involving integration over θ_{t-1} (Fig. 1). In the example which now follows, we extend the HMM model to allow a time-variant transition matrix modelled as a Dirichlet random-walk. As we will see, this induces an analytically intractable VB-marginal, motivating the use of the functionally-constrained and restricted VB variants outlined in Section II.

Consider a HMM with the following two constituents: (i) a first-order Markov chain on the unobserved discrete variable l_t (the class label), with c possible states (classes); and (ii) a set of c known class-conditional observation models, as arises in classification. For analytical convenience, we denote each state of l_t by a c-dimensional elementary basis vector $\epsilon_c\left(i\right) = \left[\delta\left(i-1\right), \delta\left(i-2\right), \ldots, \delta\left(i-c\right)\right]'$; *i.e.* $l_t \in \left\{\epsilon_c\left(1\right), \ldots, \epsilon_c\left(c\right)\right\}$. The probability of transition from the jth to the ith state, $1 \leq i, j \leq c$, is

$$\Pr\left(l_t = \epsilon_c\left(i\right)|l_{t-1} = \epsilon_c\left(j\right)\right) = t_{i,j,t},$$

where $0 < t_{i,j,t} < 1$, $i, j \in [1, \ldots, c]$. These are aggregated into the transition probability matrix, T_t, such that the column sums are unity (stochastic matrix). T_t is modelled as a random walk process, *i.e.* the expected value of T_t at time t is set to T_{t-1}. The following Bayesian filtering model is consistent with the assumptions above:

$$f(T_t|T_{t-1}) = \mathcal{D}i\left(\kappa T_{t-1}\right), \quad (48)$$

$$f(l_t|l_{t-1}, T_t) = \mathcal{M}u\left(T_t l_{t-1}\right), \quad (49)$$

$$f(d_t|l_t) = f_1\left(d_t\right)^{l_{1,t}} \times \ldots \times f_c\left(d_t\right)^{l_{c,t}}. \quad (50)$$

Here, $\mathcal{M}u\left(\cdot\right)$ denotes the multinomial distribution, and $\mathcal{D}i\left(\cdot\right)$ denotes the Dirichlet distribution [29], with $\widehat{T}_t = T_{t-1}$. The scalar parameter κ is a concentration (precision) parameter controlling the dependence between T_{t-1} and T_t. For high values of κ, only slow evolution of T_t is possible, while low values of κ allow faster variations of T_t. $f_i(d_t)$ denotes the ith known class-conditional pdf, $i = 1, \ldots, c$.

1) Local VB Filtering: The problem of inferring the state, $\{l_t, T_t\}$, can be formalized via Bayesian filtering (1–3), the exact solution of which is computationally intractable. Instead, we will apply the local VB filtering scheme (43–46) to this problem. Furthermore, we will enforce conditional independence of the kind $f(l_t, T_t|D_t) \approx \tilde{f}(l_t|D_t)\tilde{f}(T_t|D_t)$.

The log of the parameter evolution model (48,49) is

$$\ln f\left(l_t, T_t|l_{t-1}, T_{t-1}\right) \propto \mathrm{tr}\left(\ln(T_t)'(\kappa T_{t-1} - \mathbf{1}_{c,c})\right) + \\ + l_t' \ln(T_t) l_{t-1} - \sum_{i=1}^{c}\sum_{j=1}^{c} \ln\Gamma\left(\kappa t_{i,j,t-1}\right), \quad (51)$$

where $\mathbf{1}$ denotes a matrix of ones, of the stated dimensions. Under the assignment

$$\lambda_t \equiv [\ln f_1(d_t), \ln f_2(d_t), \ldots, \ln f_c(d_t)]', \quad (52)$$

the observation model (50) can be rewritten as follows:

$$f\left(d_t|l_t\right) = \exp\left(\lambda_t' l_t\right). \quad (53)$$

The approximate distributions, (45) and (46), are obtained by taking expectations of (51). Substituting the results into (43) and (44) respectively, together with (53), the resulting VB-marginals are

$$\tilde{f}\left(l_t|D_t\right) = \mathcal{M}u\left(\alpha_t\right), \tilde{f}\left(T_t|D_t\right) = \mathcal{D}i\left(Q_t\right), \quad (54)$$
$$\tilde{f}\left(l_{t-1}|D_t\right) = \mathcal{M}u\left(\beta_t\right), \tilde{f}\left(T_{t-1}|D_t\right) \propto \exp\left(\phi\left(T_{t-1}\right)\right),$$

involving shaping parameters, α_t, Q_t and β_t, and nonlinear function, $\phi\left(T_{t-1}\right)$, as follows:

$$\alpha_t = \exp\left(\lambda_t + \widehat{\ln T_t}\widehat{l_{t-1}}\right), \quad Q_t = \kappa\widehat{T_{t-1}} + \widehat{l_t}\widehat{l_{t-1}}, \quad (55)$$
$$\beta_t = \alpha_{t-1} \circ \exp\left(\widehat{\ln T_t}'\widehat{l_t}\right),$$

$$\phi(T_{t-1}) = \mathrm{tr}\left(\kappa\widehat{\ln T_t}'(T_{t-1} - \mathbf{1}_{c,c})\right) - \sum_{i=1}^{c}\sum_{j=1}^{c} \ln\Gamma\left(\kappa t_{i,j,t-1}\right).$$

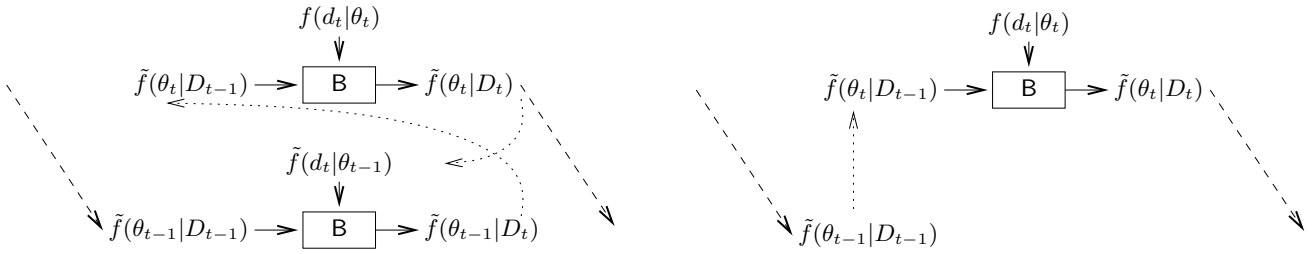

Fig. 4. Variational Bayesian filtering as a local approximation for Bayesian filtering. VB-moment substitutions are indicated by dotted arrows. Propagation of distributions between VB filtering steps is indicated by dashed arrows. **Left:** the full algorithm. **Right:** the Restricted VB (RVB) algorithm.

'$\circ$' above denotes the Hadamard product. The necessary VB-moments in (55) are

$$\widehat{l_t} \propto \alpha_t, \quad \widehat{l_{t-1}} \propto \beta_t, \tag{56}$$

$$\widehat{\ln(t_{i,j,t})} = \psi(q_{i,j,t}) - \psi\left(\mathbf{1}'_{c,1} Q_t \mathbf{1}_{c,1}\right). \tag{57}$$

The elements of both moments in (56) are probabilities, and so their sum in each case is unity, determining the constants of proportionality. $\widehat{\ln(t_{i,j,t})}$ is the (i,j)th element of matrix $\widehat{\ln T_t}$. $\psi(\cdot)$ is the digamma (psi) function [1], and $q_{i,j,t}$ is the (i,j)th element of Q_t. Since $\phi(T_{t-1})$ is a nonlinear function of its argument, no standard form for $\tilde{f}(T_{t-1}|D_t)$ can be found, and so the VB-moment, $\widehat{T_{t-1}}$, requires numerical integration. To avoid this, we make the RVB fixed distributional assignment (47), which is Dirichlet (54):

$$\overline{f}(T_{t-1}|D_t) \equiv \tilde{f}(T_{t-1}|D_{t-1}) = \mathcal{D}i_{T_{t-1}}(Q_{t-1}). \tag{58}$$

Furthermore, evaluation of the digamma function in the nonlinear moment, $\widehat{\ln T_t}$ (57), is computationally expensive. Significant savings are achieved by invoking the following functional constraint (11):

$$\hat{f}(T_t|D_t) = \delta(T_t - \hat{T}_t). \tag{59}$$

Here, $\hat{T}_t$ is the (analytical) mode of the VB-marginal in (54). In this case, $\widehat{\ln T_t}$ in (55) is replaced by $\ln \hat{T}_t$, which is computationally cheaper. This is equivalent to application of the EM algorithm (Section II-A).

2) Simulation study: A comparative Monte Carlo (MC) study was performed for binary classification ($c = 2$) of conditionally beta-distributed scalar observations:

$$f(d_t|l_t) = \mathcal{B}(l_{1,t}+1, l_{2,t}+1). \tag{60}$$

d_t is therefore a probability with $\hat{d}_t \in \{1, 0\}$. Hence, (48), (49) and (60) model a sequence of soft-bits exhibiting a Markov chain dependence.

For each setting of the following inference methods, 100 runs were performed: **MPF:** marginalized particle filtering, (19)–(21), using the parameter evolution model (51) as the importance function, $q(\cdot)$, and using residual re-sampling [16]; **VB:** local VB filtering (54) with numerical evaluation of $\widehat{T_{t-1}}$ on a grid of 50×50 points; **EM:** functionally-constrained VB-marginal (59); **RVB:** restricted VB-marginal (58); **Threshold:** l_t is inferred heuristically from the soft-bits by thresholding:

$$\hat{l}_t = [\hat{l}_{1,t}, \hat{l}_{2,t}]', \quad \hat{l}_{1,t} = \text{round}(d_t) = \begin{cases} 1 & \text{if } d_t > 0.5, \\ 0 & \text{if } d_t \leq 0.5. \end{cases} \tag{61}$$

This constitutes Maximum Likelihood (ML) estimation of l_t, ignoring the Markov chain model for l_t (49). For each method, performance was measured by the Mean Square Error (MSE) between the simulated and estimated labels, with $t_u = 1000$ (Section IV-C2).

The MPF has as a 'tuning knob' the number of particles, and the EM and VB approximations have the number of IVB iterations as a tuning knob. The computational cost of all the methods increases linearly with these tuning knobs. The performance of each inference algorithm as a function of execution time in Matlab (using a 1.7 GHz Intel Centrino processor) is plotted in Fig. 5. RVB and Thresholding have no tuning knob and their execution times are small (see left of each frame in Fig. 5).

The performance of VB and EM does not improve when the number of iterations is greater than five, but the performance of MPF steadily improves with an increasing number of particles, as is typical of these global stochastic approximations. For $n > 15$ particles, it outperforms all its competitors. This underlines the superiority of global approximations such as MPF, but it is achieved at ever-increasing computational cost. For example, we note that VB out-performs MPF in these simulations if the execution time is held below 3.5 secs. This emphasizes once again the fact that VB-based approximations are particularly attractive in contexts demanding low computational cost.

We note in this example that each additional constraint or restriction imposed on the VB scheme resulted in loss of performance. The performance degradation was less when the higher-order moment (57) was removed via the EM functional constraint (59), than when the RVB restriction was imposed (58). Note that the performances of all the methods greatly improve, and converge, for higher κ (*i.e.* slower variations in T_t).

VI. Discussion

An important feature of the VB approximation revealed in this work is its free-form nature. The optimal approximation is found by a principled substitution of the necessary VB-moments into the model (6). Hence, the functional form of the approximation is deduced rather than imposed. This reveals possibilities beyond the standard Gaussian approximations that underlie many of the deterministic approximation methods for Bayesian filtering. New algorithmic flows of control are implied. Thus, for example, we have seen that the VB approximation generalizes the EM algorithm by taking into account higher-order moments. When the free-form optimizers are

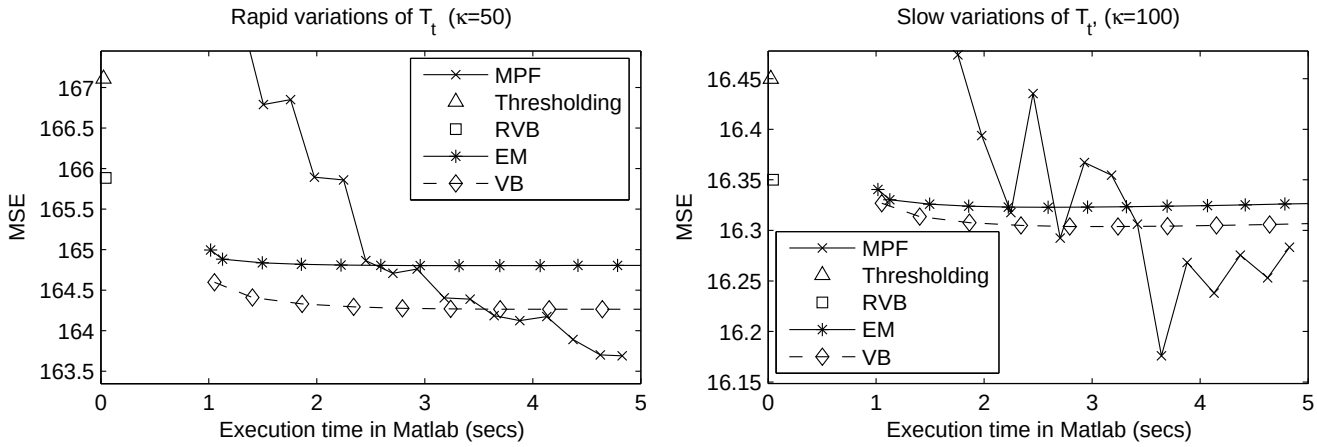

Fig. 5. Comparative Monte Carlo study (100 trials per setting) of MSE performance for the marginalized particle filter (MPF), VB, EM and RVB approximations for estimation of l_t in the HMM model (49). Performance is plotted against execution time in Matlab. '$\times$' denotes an increment by one in the number of particles in the MPF; '$*$' and '$\diamond$' denote an increment by one in the number of iterations of the EM or VB algorithms respectively. Both RVB and the heuristic thresholding method have a small, invariant execution time (see left of each frame).

intractable, as occurred in all examples in this paper, functional constraints and/or restrictions to tractable forms have been shown to be practicable means to proceed.

The underlying principle of the VB approximation is to force conditional independence between subsets of the model parameters. Clearly the VB approximation will not perform well in cases where there exists strong correlation between these subsets [29], [33], and so great care must be taken by the designer in choosing where to impose this conditional independence. The choice should be made in an effort to achieve tractability. Further unnecessary partitioning of the parameter set decreases the quality of approximation, as demonstrated in [11]. Furthermore, it may be possible to re-parameterize the model so that the conditional independence assumption is more appropriate in the transformed parameters. A proposal for future work is to find parameter transformations (via parameter orthogonalization techniques [6] for instance) for which approximate conditional independence is exhibited, and to apply the VB approximation to this natural partition.

Note that the VB approximation is an example of a mean-field approximation, being known as the naïve mean-field approximation in statistical physics [20]. Less severe functional constraints than the conditional independence assumption (4) underlying VB—such as TAP equations [19])—can be accommodated within a fuller mean-field theory. These merit further research as a basis for developing novel flows of control in signal processing.

VII. Conclusion

Stochastic approximations such as particle filtering are the golden standard in Bayesian filtering, but variants need to be considered in some on-line signal processing contexts in order to reduce the implied computational cost. On the evidence of this paper, the VB approximation has an important rôle to play in this area. Indeed, we have shown that the certainty equivalence approach is a variant of VB filtering. In this paper, a wider range of choices has been revealed, allowing a trade-off between accuracy of approximation and speed. The particular features of VB filtering are (i) utilization of higher-order moments, and (ii) iterative solution of the implicit equations to reach the optimal approximation. When combined with particle filtering (Section IV), we found that the higher-order moments can improve performance over certainty equivalence in situations where a small number of particles is necessitated. When used as a deterministic local approximation (Section V), the iterative scheme using certainty equivalence (such as the EM algorithm) improved on non-iterative certainty equivalence approaches. Furthermore, iterations involving higher-order moments yielded even better performance in the HMM example. These properties need to be investigated for broader classes of models. Clearly there is a strong motivation for exploring the VB approximation and its variants in Bayesian filtering, particularly in computationally-constrained environments.

Acknowledgement

This work was supported by grants MŠMT 1M0572.

References

[1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover Publications, 1972.
[2] B. Anderson and J. Moore, *Optimal Filtering*. New Jersey: Prentice Hall, 1979.
[3] H. Attias, J. C. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," in *Advances in Neural Information Processing Systems*, 2001, pp. 758–764.
[4] M. J. Beal and Z. Ghahramani, "The variational Kalman smoother," University College London, Tech. Rep., 2000.
[5] R. Bucy and K. Senne, "Digital synthesis on nonlinear filters," *Automatica*, vol. 7, pp. 287–298, 1971.
[6] D. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *Journal of Royal Statistical Society, B*, vol. 49, no. 1, pp. 1–39, 1987.
[7] E. Daum, "New exact nonlinear filters," in *Bayesian Analysis of Time Series and Dynamic Models*, J. Spall, Ed. New York: Marcel Dekker, 1988.
[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.
[9] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[10] Z. Ghahramani and M. Beal, "Propagation algorithms for variational Bayesian learning," *Advances in Neural Information Processing Systems*, vol. 13, pp. 507–513, 2001.
[11] A. Ilin and H. Valpola, "On the effect of the form of the posterior approximation in variational learning of ICA models," *Neural Processing Letters*, vol. 22, no. 2, pp. 183–204, 2005.
[12] S. Ji, B. Krishnapuram, and L. Carin, "Variational Bayes for continuous hidden Markov models and its application to active learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 522–532, 2006.
[13] S. Julier and J. Uhlman, "A new extension of the Kalman filter to nonlinear systems," in *Proc. AeroSense*, 1997.
[14] R. Kulhavý, *Recursive Nonlinear Estimation: A Geometric Approach*, ser. Lecture Notes in Control and Information Sciences. London: Springer-Verlag, 1996, vol. 216.
[15] G. Lee, S. Gelfand, and M. Fitz, "Bayesian techniques for blind deconvolution," *IEEE Transactions on Communications*, vol. 44, pp. 826–835, 1996.
[16] J. Liu and R. Chen, "Sequential Monte Carlo methods for dynamic systems," *Journal of the American Statistical Association*, vol. 93, 1998.
[17] D. J. C. MacKay, "Ensemble learning for hidden Markov models," University of Cambridge, Tech. Rep., 1997.
[18] F. Mustière, M. Bolić, and M. Bouchard, "A modified Rao-Blackwellised particle filter," in *Proceedings of the IEEE conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006.
[19] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice*. Cambridge, Massachusetts: The MIT Press, 2001.
[20] M. Opper and O. Winther, "From naive mean field theory to the TAP equations," in *Advanced Mean Field Methods*, M. Opper and D. Saad, Eds. The MIT Press, 2001.
[21] V. Peterka, "Bayesian approach to system identification," in *Trends and Progress in System identification*, P. Eykhoff, Ed. Oxford: Pergamon Press, 1981, pp. 239–304.
[22] Y. Qi, J. Paisley, and L. Carin, "Music Analysis Using Hidden Markov Mixture Models," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, pp. 5209–5224, 2007.
[23] S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.
[24] M. Sato, "Online model selection based on the variational Bayes," *Neural Computation*, vol. 13, pp. 1649–1681, 2001.
[25] T. Schön, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2279–2289, 2002.
[26] M. Šimandl, J. Královec, and T. Söderström, "Advanced point-mass method for nonlinear state stimation," *Automatica*, vol. 42, pp. 1133–1145, 2006.
[27] M. Šimandl and O. Straka, "Application of the EM algorithm in Gausssian sums methods," University of West Bohemia, Tech. Rep., 2000.
[28] V. Šmídl and A. Quinn, "Mixture-based extension of the AR model and its recursive Bayesian identification," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3530–3542, 2005. [Online]. Available: files/publ/tsp05.pdf
[29] ——, *The Variational Bayes Method in Signal Processing*. Springer, 2005. [Online]. Available: http://www.springer.com/east/home/engineering?SGWID=5-175-22-70903497-0
[30] ——, "The restricted variational Bayes approximation in Bayesian filtering," in *Proceedings of the IEEE nonlinear statistical signal processing workshop*, Cambridge, UK, 2006. [Online]. Available: files/publ/camb06.pdf
[31] H. Sorenson, "On development of practical nonlinear filters," *Information Sciences*, vol. 7, pp. 253–270, 1974.
[32] J. Vermaak, N. Lawrence, and P. Perez, "Variational inference for visual tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 773–780.
[33] B. Wang and D. Titterington, "Lack of consistency of mean field and variational Bayes approximations for state space models," *Neural Processing Letters*, vol. 20, pp. 151–170, 2004.
[34] J. Winn and C. Bishop, "Variational message passing," *The Journal of Machine Learning Research*, vol. 6, pp. 661–694, 2005.