# ACCELERATED PARTICLE FILTERING USING THE VARIATIONAL BAYES APPROXIMATION

*Václav Šmídl*

Academy of Sciences, Prague, Czech Republic

*Anthony Quinn*

Trinity College Dublin, Ireland

**ABSTRACT**

In Bayesian filtering, the model may allow analytical marginalization over a subset, $\theta_{1,t}$, of the parameters. The Marginalized (Rao-Blackwellized) Particle Filter (MPF) exploits this, by requiring stochastic sampling only in the remaining parameters, $\theta_{2,t}$, with the potential for major computational and convergence speed-ups. The marginalized filtering distribution in $\theta_{1,t}$ is expressed as a mixture of $n$ analytical components, each conditioned on one of the $n$ particle trajectories in $\theta_{2,t}$; *i.e.* sufficient statistics must be stored and updated for *each* particle trajectory. In this paper, the Variational Bayes (VB) approximation is used as a one-step approximation to extract necessary moments from the $n$ particles in a principled manner, yielding a single-component marginalized filtering distribution. This formalizes and extends a recently reported certainty equivalence approach to accelerating MPFs. The comparative performance of the full and accelerated MPFs is explored via a scalar nonlinear filtering example.

Keywords: Bayesian filtering, Variational Bayes, marginalized particle filtering, nonlinear filtering.

## 1. BAYESIAN FILTERING

Consider a sequence of (observed) data, $D_t = [d_1, \ldots, d_t]$, and an associated sequence of (unobserved) parameters, $\Theta_t = [\theta_1, \ldots, \theta_t]$. We assume that $\theta_t$ is a *state* of the model:

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

We are concerned with *Bayesian Filtering (BF), i.e.* recursive evaluation of the filtering distribution, $f(\theta_t|D_t)$, using Bayes' rule. BF is analytically tractable if (i) marginalization over $\theta_{t-1}$ is analytically tractable, and (ii) the resulting marginal distribution, $f(\theta_t|D_t)$, is functionally invariant, $\forall t$. (i) and (ii) are satisfied for only a limited class of models.

### 1.1. Particle Filtering (PF)

*Particle filtering (PF)* [1] refers to a range of sequential Monte Carlo (MC) techniques for Bayesian filtering in intractable contexts, generating a sequence of empirical approximations:

$$f(\Theta_t|D_t) \approx \frac{1}{n}\sum_{i=1}^{n} \delta\left(\Theta_t - \Theta_t^{(i)}\right). \tag{2}$$

Here, $\Theta_t^{(i)}$ are i.i.d. samples from $f(\Theta_t|D_t)$, and $\delta(\cdot)$ denotes the Dirac $\delta$-function. In the (typical) case when sampling from the exact posterior is impossible, we can, instead, draw samples, $\Theta_t^{(i)} \sim q(\Theta_t|D_t)$, from a chosen proposal distribution (importance function), $q(\cdot)$, as follows:

$$f(\Theta_t|D_t) = \frac{f(\Theta_t|D_t)}{q(\Theta_t|D)} q(\Theta_t|D_t) \tag{3}$$

$$\approx \frac{f(\Theta_t|D_t)}{q(\Theta_t|D_t)} \frac{1}{n}\sum_{i=1}^{n} \delta\left(\Theta_t - \Theta_t^{(i)}\right). \tag{4}$$

Using the sifting property of $\delta(\cdot)$, (4) can be written in the form of a *weighted* empirical distribution:

$$f(\Theta_t|D_t) \approx \sum_{i=1}^{n} w_t^{(i)} \delta\left(\Theta_t - \Theta_t^{(i)}\right), \tag{5}$$

$$w_t^{(i)} \propto \frac{f\left(\Theta_t^{(i)}|D_t\right)}{q\left(\Theta_t^{(i)}|D_t\right)}. \tag{6}$$

Under this *importance sampling* procedure, the true posterior, $f(\cdot)$, need only be evaluated point-wise. Furthermore, normalizing constants of $f(\cdot)$ and $q(\cdot)$ are not required, since (5) can be normalized trivially via the constant $c = \sum_{i=1}^{n} w_t^{(i)}$.

The challenge for on-line algorithms is therefore to generate recursively the samples (particles), $\{\theta_t^{(i)}\}$, and the importance weights (6), $\{w_t^{(i)}\}$. From (1) and (6),

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_t^{(i)}\right) f\left(\theta_t^{(i)}|\theta_{t-1}^{(i)}\right)}{q\left(\theta_t^{(i)}|\Theta_{t-1}^{(i)}, D_t\right)} w_{t-1}^{(i)}, \tag{7}$$

where, now, $\theta_t^{(i)}$ are drawn from the denominator of (7), typically chosen as $f(\theta_t|\theta_{t-1})$ (1). Implementation issues—such as the appropriate choice of the importance function, resampling, *etc.*—are addressed, for example, in [1].

Support from grants MŠMT 1M0572 and AVCR 1ET 100 750 401 is gratefully acknowledged.

### 1.2. Marginalized Particle Filtering (MPF)

The general importance sampling framework of the previous section is widely applicable, but sampling is required from the *joint* state space. This may be computationally prohibitive in high dimensions, and large numbers of particles are required in such cases to satisfy convergence criteria [1]. Consider factorization $f(\Theta_t|D_t) = f(\Theta_{1,t}|\Theta_{2,t}, D_t) f(\Theta_{2,t}|D_t)$, with replacement of the final term by a weighted empirical distribution of the type in (5). Following (3)–(6):

$$f(\Theta_t|D_t) \approx \sum_{i=1}^{n} w_t^{(i)} f\left(\Theta_{1,t}|\Theta_{2,t}^{(i)}, D_t\right) \delta\left(\Theta_{2,t} - \Theta_{2,t}^{(i)}\right), \tag{8}$$

where sampling is now required only in the reduced dimensions of $\theta_{2,t}$. Here, $w_t^{(i)} \propto \frac{f\left(\Theta_{2,t}^{(i)}|D_t\right)}{q\left(\Theta_{2,t}^{(i)}|D_t\right)}$, with recursive form

$$w_t^{(i)} \propto \frac{f\left(d_t|\theta_{2,t}^{(i)}\right) f\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}\right)}{q\left(\theta_{2,t}|\Theta_{2,t-1}^{(i)}, D_t\right)} w_{t-1}^{(i)}. \tag{9}$$

Hence, the model (1) must admit a partition, $\theta_t = [\theta_{1,t}, \theta_{2,t}]$, for which $\theta_{1,t}$ can be integrated analytically. This is also the requirement for $f\left(\Theta_{1,t}|\Theta_{2,t}^{(i)}, D_t\right)$, *i.e.* the *conditional* filtering distribution, to be available analytically in (8). Only a limited class of models—such as Gaussian state-space models [2]—satisfy this requirement. The MPF scheme (8)–(9) is sometimes called the Rao-Blackwellized particle filter [1].

### 1.3. Accelerating the MPF via Certainty Equivalence (CE)

The mixture in (8) requires $n$ parallel conditional Bayesian filtering updates; *i.e.* sufficient statistics are required for *each* particle trajectory, $\Theta_{2,t}^{(i)}$. This is computationally inefficient if the particle trajectories are similar. A simple Certainty Equivalence (CE) approach to reducing the computational cost of the MPF was reported in [3]. The idea was to replace the $n$ components in (8) by a single component,

$$f(\Theta_t|D_t) \approx f\left(\Theta_{1,t}|\hat{\Theta}_{2,t}, D_t\right) \delta\left(\Theta_{2,t} - \hat{\Theta}_{2,t}\right), \tag{10}$$

where $\hat{\Theta}_{2,t} = \sum_{i=1}^{n} w_t^{(i)} \Theta_{2,t}^{(i)}$, and the $w_t^{(i)}$ were evaluated using a modified version of (9). The idea is closely related to the mean-field approach to distributional approximation [4], but considerable extensions are possible via full application of such approximations. We will now emphasize the Variational Bayes (VB) approximation [5], and find that this leads to a principled approach to the problem of concentrating the $n$ components in (8) into a single component in (10).

## 2. THE VARIATIONAL BAYES APPROXIMATION

The VB approximation is an optimal deterministic distributional approximation, as shown by the following theorem.

**Theorem 1** *Let $f(\theta|D)$ be the posterior distribution of multivariate parameter, $\theta = [\theta_1', \theta_2']'$, and $\breve{f}(\theta|D)$ be an approximate distribution with conditional independence restriction:*

$$\breve{f}(\theta|D) = \breve{f}(\theta_1, \theta_2|D) = \breve{f}(\theta_1|D) \breve{f}(\theta_2|D). \tag{11}$$

*Any minimum of the Kullback-Leibler divergence from $\breve{f}(\cdot)$ to $f(\cdot)$ is achieved when $\breve{f}(\cdot) = \tilde{f}(\cdot)$, where*

$$\tilde{f}(\theta_i|D) \propto \exp\left(\mathsf{E}_{\tilde{f}(\theta_{/i}|D)}\left[\ln\left(f(\theta, D)\right)\right]\right), \ i = 1, 2. \tag{12}$$

*Here $\theta_{/i}$ denotes the complement of $\theta_i$ in $\theta$. We will refer to $\tilde{f}(\theta_i|D)$ (12) as the* VB-marginals*.*

Theorem 1 provides a powerful tool for approximation of joint pdfs of *separable form* [5]:

$$\ln f(\theta_1, \theta_2, D) = g(\theta_1, D)' h(\theta_2, D). \tag{13}$$

Here, $g(\theta_1, D)$ and $h(\theta_2, D)$ are finite-dimensional vectors. Using (13) in (12),

$$\tilde{f}(\theta_1|D) \propto \exp\left(g(\theta_1, D)' h\left(\widehat{\theta_2, D}\right)\right). \tag{14}$$

$\widehat{h(\cdot)} = \mathsf{E}_{\tilde{f}(\theta_2|D)}[h]$ are the *VB-moments* of $\theta_2$ (ditto for $\theta_1$), giving an Iterative VB (IVB) [5] moment-swapping algorithm.

In nonlinear cases of $g$ and/or $h$, the VB-marginals (12) may have nonstandard form, with VB-moments (14) difficult to evaluate. It may be necessary to replace such non-standard VB-marginals—*e.g.* $\tilde{f}(\theta_2|D)$—with a tractable alternative. In the next two Sections, we present two such modifications.

### 2.1. The Functionally Constrained VB Approximation

Here, an extra step is introduced within each IVB cycle; *i.e.* $\tilde{f}(\theta_2|D)$ is projected into a tractable alternative, $\hat{f}(\theta_2|D)$. The moments of *this* distribution are fed back via (14) to generate $\tilde{f}(\theta_1|D)$. The EM algorithm is a case in point. Here,

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

with $\hat{\theta}_2 = \arg\max_{\theta_2} \tilde{f}(\theta_2|D)$, and $\tilde{f}(\theta_2|D)$ is the VB-marginal given by (12). It follows from (14) that

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

### 2.2. The Restricted VB Approximation

In this case, we replace $\breve{f}(\theta_2|D)$ in (11) by a tractable *fixed* distribution, $\overline{f}(\theta_2|D)$. Using Theorem 1:

$$\tilde{f}(\theta_1|D) \propto \exp\left(\mathsf{E}_{\overline{f}(\theta_2|D)}\left[\ln\left(f(\theta, D)\right)\right]\right). \tag{15}$$

Therefore, only a *single* substitution of moments from $\overline{f}(\cdot)$ is required to generate the approximation, avoiding IVB cycles. Under assumption (13), the moments which need to be

substituted are $\widehat{h(\cdot)} = \mathsf{E}_{\overline{f}(\theta_2|D)}[h(\cdot)]$. It is interesting to note that a number of popular distributional approximations are special cases of (15): (i) *certainty equivalence*, where $\overline{f} \equiv \delta(\theta_2 - \hat{\theta}_2)$ for some chosen point estimate, $\hat{\theta}_2$, in which case $\tilde{f}(\theta_1|D) = f(\theta_1|\hat{\theta}_2, D)$ in (15); and (ii) the *Quasi-Bayes (QB) approximation*, where $\overline{f} \equiv f(\theta_2|D)$, the exact marginal, if is is available. If $h(\cdot)$ in (13) is *linear*, then (i) and (ii) are equivalent under assignment $\hat{\theta}_2 = \mathsf{E}_{f(\theta_2|D)}[\theta_2]$ [5].

## 3. VARIATIONAL BAYESIAN FILTERING (VBF)

We now develop a one-step VB approximation of the Bayesian filtering distribution, $f(\theta_t|D_t)$ (Section 1). We impose separability into $\theta_t = [\theta_{1,t}, \theta_{2,t}]$, such that (Theorem 1)

$$\tilde{f}(\theta_{t-1}|D_{t-1}) = \tilde{f}(\theta_{1,t-1}|D_{t-1})\,\tilde{f}(\theta_{2,t-1}|D_{t-1}), \quad (16)$$

requiring the following marginal to be available analytically:

$$f(\theta_{1,t}, \theta_{2,t}|D_t) \propto \int f(d_t|\theta_t)\, f(\theta_t|\theta_{t-1})\, \tilde{f}(\theta_{1,t-1}|D_{t-1})\, d\theta_{1,t-1}. \quad (17)$$

The VB approximation of (17) is then, once again, obtained via Theorem 1, completing a step of VB Filtering (VBF) [5]. For conciseness, we have suppressed the conditioning of the distributions in (16) and (17) on the trajectory, $\Theta_{2,t}$.

As before, the necessary VB moments of *both* VB-marginals on the righthand-side of (16) must be available, $\forall t$. Assuming, for example, that the VB-moments of $\tilde{f}(\theta_{2,t}|D_t)$ are not available analytically, we next explore the modified VB approximations of Sections 2.1 and 2.2 in this VBF context.

### 3.1. VB Particle Filtering

Using the functionally-constrained VB approximation of Section 2.1, we project $\tilde{f}(\theta_{2,t}|D_t)$ into an empirical distribution:

$$\hat{f}(\theta_{2,t}|D_t) = \sum_{i=1}^{n} w_t^{(i)} \delta\left(\theta_{2,t} - \theta_{2,t}^{(i)}\right), \quad (18)$$

$$w_t^{(i)} \propto \frac{\tilde{f}\left(\theta_{2,t}^{(i)}|D_t\right)}{q\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}, D_t\right)}. \quad (19)$$

The necessary VB-moments $h(\widehat{\theta_{2,t}, D_t})$ needed for evaluation of $\tilde{f}(\theta_{1,t}|D_t)$ (14) are now *always* available:

$$h(\widehat{\theta_{2,t}, D_t}) = \sum_{i=1}^{n} w_t^{(i)} h\left(\theta_{2,t}^{(i)}, D_t\right). \quad (20)$$

The implied IVB algorithm (Section 2) is as follows:

**Algorithm 3.1 (VB Particle Filtering)**
*0. Draw samples $\theta_{2,t}^{(i)}$ from $q\left(\theta_{2,t}|\theta_{2,t-1}^{(i)}, D_t\right)$.*
*1. Evaluate moments $g(\widehat{\theta_{1,t}, D_t})$ needed in numerator of (19).*
*2. Evaluate weights $w_t^{(i)}$ via (19).*
*3. Evaluate moments $h(\widehat{\theta_{2,t}, D_t})$ via (20), and generate $\tilde{f}(\theta_{1,t}|D_t)$ via (14).*
*4. If not converged go to step 1.*

### 3.2. Quasi-Bayes (QB) Particle Filtering

Given (16), and assuming that (1) factorizes as

$$f(\theta_t|\theta_{t-1}) = f(\theta_{1,t}|\theta_{1,t-1})\, f(\theta_{2,t}|\theta_{2,t-1}), \quad (21)$$

then the *unnormalized* analytical marginal, $f(\theta_{2,t}|D_t)$, of (17) can always be evaluated pointwise. Hence, we can replace (19) by $w_t^{(i)} \propto \frac{f\left(\theta_{2,t}^{(i)}|D_t\right)}{q\left(\theta_{2,t}^{(i)}|\theta_{2,t-1}^{(i)}, D_t\right)}$, avoiding IVB iterations.

**Remark 1** *Under these VB scenarios, the $n$ particles have been concentrated into $\tilde{f}(\theta_1|D_t)$ via (20), eliminating the need for $n$ parallel Bayesian filtering steps. (10), as proposed in [3], is a special case of QB particle fitering having assumed a* linear $h(\cdot)$ *(13). QB particle filtering has the advantage of allowing higher-order moments (20) of the empirical distribution (18) to be exploited.*

## 4. SCALAR NONLINEAR FILTERING EXAMPLE

Consider the following model [3]:

$$\begin{aligned} f(x_t|x_{t-1}) &= \mathcal{N}(Ax_{t-1}, Q), \\ f(C_t|C_{t-1}) &= \mathcal{N}(\arctan(C_{t-1}), P), \\ f(d_t|x_t, C_t) &= \mathcal{N}(C_t' x_t, R). \end{aligned} \quad (22)$$

Essentially, this is a standard linear-Gaussian model with unknown non-stationary $C_t$, for which a nonlinear evolution model is defined. Here, integration over $x_{t-1}$ is possible using standard Kalman Filtering (KF) theory, yielding the following conditional posterior of $x_t$:

$$f(x_t|C_t, D_t) = \mathcal{N}\left(\mu_t, \Omega_t^{-1}\right), \quad (23)$$

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

This is written in terms of precision matrix $\Omega_t$ for analytical convenience. Exact integration over $C_{t-1}$ is intractable. A MPF (Section 1.2) is obtained using (8)–(9) with assignments $\theta_{1,t} = x_t$ and $\theta_{2,t} = C_t$.

### 4.1. VB Particle Filtering

The required distribution (17), i.e. $f\left(x_t, C_t | C_{t-1}, D_t\right)$, is obtained by multiplying (23) by (22). The VB-marginals are

$$\tilde{f}\left(x_t | C_{t-1}, D_t\right) = \mathcal{N}\left(\tilde{\mu}_t, \tilde{\Omega}_t^{-1}\right), \tag{24}$$

$$\tilde{f}\left(C_t | C_{t-1}, D_t\right) \propto \mathcal{N}\left(\tilde{C}_t, \Sigma_t\right) |\Omega_t(C_t)|^{\frac{1}{2}}, \tag{25}$$

with shaping parameters

$$\tilde{\Omega}_t = \left(Q + A\tilde{\Omega}_{t-1}^{-1}A'\right)^{-1} + \mathsf{E}\left[C_t' R^{-1} C_t\right], \tag{26}$$

$$\tilde{\mu}_t = \tilde{\Omega}_t^{-1}\left[\left(Q + A\tilde{\Omega}_{t-1}^{-1}A'\right)^{-1} A\tilde{\mu}_{t-1} + \widehat{C_t}' R^{-1} d_t\right],$$

$$\Sigma_t = \left(R^{-1}\widehat{x_t' x_t} + P^{-1}\right)^{-1},$$

$$\tilde{C}_t = \Sigma_t\left(R^{-1} d_t \hat{x}_t' + P^{-1}\arctan(C_{t-1})\right).$$

Here, $\widehat{C_t}$, and $\mathsf{E}\left[C_t' R^{-1} C_t\right]$ are first and second moments of the empirical distribution, $\hat{f}\left(C_t | D_t\right)$ (18), and $\hat{x}_t$ and $\widehat{x_t' x_t}$ are moments of the VB-marginal (24).

**Remark 2** *Under QB particle filtering, the weights are evaluated via (23) rather than via (25). In [3], (10) has the same form as (24) with* $\mathsf{E}\left[C_t' R^{-1} C_t\right]$ *replaced by* $\widehat{C_t}' R^{-1} \widehat{C_t}$.

### 4.2. Simulation study

A system (22) with two-dimensional state, $x_t$, and scalar output, $d_t$, is simulated with parameters $A = 1$, $Q = 1$, $P = 1$, $R = 1$. The aim is to illustrate the effect of propagation of higher moments (20) within the VB particle filtering schemes. We use the same proposal, $q\left(C_t | C_{t-1}\right) = f(C_t | C_{t-1})$ (22), and the same re-sampling scheme for all tested methods. Performance is assessed via two measures, the first being the Mean Square Error (MSE) of the state estimate,

$$MSE = \frac{1}{T}\sum_{t=1}^{T} \left\|x_t - \hat{x}_t\right\|^2.$$

$T$ is the number of samples in the simulation, $x_t$ the simulated value of the state, and $\hat{x}_t$ the mean value of the posterior distribution $\tilde{f}\left(x_t | C_{t-1}, D_t\right)$ (24) under each approximation. The second measure is the log of the marginal likelihood:

$$\log f\left(D_t\right) = \sum_{t=1}^{T} \log f\left(d_t | D_{t-1}\right),$$

$$f\left(d_t | D_{t-1}\right) = \int f\left(d_t | \theta_t\right) \tilde{f}\left(\theta_t | D_{t-1}\right) d\theta_t.$$

We compared (i) the Marginalized Particle Filter (MPF), (ii) the Certainty Equivalence (CE) approach of [3], and (iii) QB particle filtering which substitutes 2nd-order moments (26)

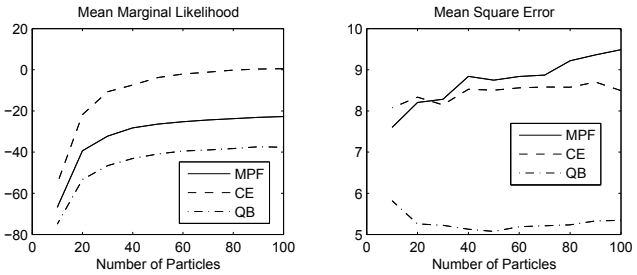


**Fig. 1**. Performance of approximate filtering in a MC study.

(QB). A Monte Carlo (MC) study, with $T = 1000$ and 1000 realizations per setting, was undertaken (Fig. 1). The CE method appears to provide the best posterior approximation (*i.e.* highest marginal likelihood values). The 2nd-order moments of the QB particle filter result in a flatter QB posterior distribution (24), *i.e.* a lower marginal likelihood. However, this wider covariance may permit better coverage of the parameter space, yielding the best MSE performance.

## 5. CONCLUSION

Ideas from mean-field theory have been used to accelerate marginalized particle filters, via principled substitution of moments from the weighted empirical distribution. The VB particle filter requires iterative re-evaluation of weights before moment calculations, while QB particle filtering is non-iterative. They have in common with certainty equivalence approaches—which they justify—the advantage of just one Bayes filter update per time-step, but the added advantage of communicating higher-order moments from the empirical distribution, improving accuracy. The framework is formal and there is scope for issues such as (i) the class of models amenable to VB particle filtering, and (ii) other mean-field approximations.

## 6. REFERENCES


[1] A. Doucet, N. de Freitas, and N. Gordon, eds., *Sequential Monte Carlo Methods in Practice*. Springer, 2001.

[2] T. Schön, F. Gustafsson, and P.-J. Nordlund, "Marginalized particle filters for mixed linear/nonlinear state-space models," *IEEE Trans. Sig. Process.*, vol. 53, 2002.

[3] F. Mustière, M. Bolić, and M. Bouchard, "A modified Rao-Blackwellised particle filter," in *Proc. IEEE Conf. on Acoustics, Speech, and Signal Process.*, 2006.

[4] 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.

[5] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*. Springer, 2005.