SEQUENTIAL MONTE CARLO OPTIMIZATION USING ARTIFICIAL STATE-SPACE MODELS

Joaquín Mínguez†, Cristina S. Mateć†, Petar M. Djurić‡ and Dan Crisan∗

†Dep. of Signal Theory and Communications, Universidad Carlos III de Madrid (Spain).
‡Dep. of Electrical and Computer Engineering, Stony Brook University (USA).
∗Department of Mathematics, Imperial College London (UK).

{jmiguez,csmaiz}@tsc.uc3m.es, djuric@ece.sunysb.edu, d.crisan@imperial.ac.uk

ABSTRACT

We introduce a method for sequential minimization of a certain class of (possibly non-convex) cost functions with respect to a high dimensional signal of interest. The proposed approach involves the transformation of the optimization problem into one of estimation in a discrete-time dynamical system. In particular, we describe a methodology for constructing an artificial state-space model which has the signal of interest as its unobserved dynamic state. The model is “adapted” to the cost function in the sense that the maximum a posteriori (MAP) estimate of the system state is also a global minimizer of the cost function. The advantage of the estimation framework is that we can draw from a pool of sequential Monte Carlo methods, for particle approximation of probability measures in dynamic systems, that enable the numerical computation of MAP estimates. We provide examples of how to apply the proposed methodology, including some illustrative simulation results.

Index Terms—Sequential optimization, Monte Carlo methods, stochastic optimization, particle smoothing

1. INTRODUCTION

Consider the problem of estimating an unobserved random signal \( \{ x_t \}_{t \in \mathbb{N}} \), where \( \mathbb{N}^* = \mathbb{N} \cup \{0\} \), from a sequence of related measurements, denoted \( \{ y_t \}_{t \in \mathbb{N}} \). Assume that, in the absence of an explicit probabilistic model for these random processes and their dependence, we are given an estimation criterion that consists in the minimization of a certain cost function, \( C_T(x_0:T, y_1:T) \), where \( T \) is the time of the last available observation and \( x_0:T = \{ x_0, x_1, \ldots, x_T \} \), \( y_1:T = \{ y_1, \ldots, y_T \} \). The function \( C_T(\cdot, \cdot) \) may be nonlinear, possibly exhibiting multiple minima, and analytically intractable for optimization. Therefore, we are in general led to use numerical techniques to find a sequence \( \hat{x}_{0:T} \) that (approximately) minimizes \( C_T \). However, even the numerical approach is, in general, difficult. Indeed, if \( x_t \) has fixed dimension \( n_x \) for all \( t \), then \( x_{0:T} \) is \( (T + 1)n_x \)-dimensional. Thus, when, e.g., \( T = 1000 \) and \( n_x = 4 \), we have 4000 variables to optimize. As a consequence, most global optimization techniques (such as the simple [1] or accelerated [2] random search, genetic-like algorithms [3] or simulated annealing methods [4]) will be subject to the so-called curse of dimensionality and become practically inapplicable.

A neat way to beat the curse is to design sequential optimization algorithms, i.e., procedures that process the observations one at a time, as they are collected, and build candidate solutions, \( x_{0:t} \), incrementally, for \( t = 0, 1, 2, \ldots \). Good examples are found in the field of digital communications. When the signals \( \{ x_t \} \) are symbols from a discrete alphabet and the measurements \( \{ y_t \} \) are observations at the output of a linear channel, the Viterbi algorithm [5] is a dynamic programming technique that yields the symbols \( \hat{x}_{0:T} \) with the highest likelihood given the observations. Message-passing algorithms used in turbo-decoding [6] are also sequential in the way they process the observations, although they are usually embedded in iterative procedures. Of course, these methods exploit the fact that \( x_t \) comes from a discrete space. If the signals of interest are continuous, the design of sequential techniques becomes harder. Dynamic optimization techniques [7] process the observations sequentially and can yield approximate solutions, but they are designed to track the minima of a time-evolving cost that depends on a single observation, \( y_t \), and is minimized with respect to a single signal vector, \( x_t \), at each time step. Hence, they do not necessarily converge to a global solution in terms of the complete series \( x_{0:T} \). Other optimization techniques, such as sequential convexification [8], consist in solving a sequence of adequately constructed sub-problems. Any solution of each sub-problem provides an improvement over the best previous candidate solution, in such a way that asymptotic convergence to the global minimum is achieved. However, each sub-problem involves the complete signal and measurement series, \( x_{0:T} \) and \( y_{1:T} \), respectively, so this approach does not overcome the curse of dimensionality.

We propose a novel approach to sequential optimization that exploits the theory of Monte Carlo methods for particle approximation of probability measures in dynamic systems [9, 10, 11]. To do so, we build a dynamic state-space model in which the random signal \( \{ x_t \} \) is the sequence of unobserved states and the measurements \( \{ y_t \} \) are the associated observations. The model specifies a prior probability distribution for \( x_0 \), a Markov transition distribution that characterizes \( x_t \) conditional on \( x_{t-1} \), and a likelihood that describes the dependence of \( x_t \) on \( y_t \). We introduce a method to select these probability distributions in such a way that
the maximum a posteriori (MAP) estimate of \( x_{0:T} \) given \( y_{0:T} \) under
the resulting state-space model coincides with the cost minimizer
\( x_{0:T} \). Therefore, we convert the problem of cost minimization
into one of MAP estimation. The advantage of this transformation
is that we can draw from a pool of existing sequential Monte
Carlo (SMC) methods for the approximation of posterior probability
measures. In particular, we apply the particle smoothing technique
of [11], that yields an asymptotically convergent approximation of
the probability measure of \( x_{0:T} \) given an arbitrary fixed sequence
\( y_{0:T} \), and, from this measure, we build the desired MAP estimates.

The rest of this paper is organized as follows. In Section 2
we formally describe the class of cost functions we consider and
introduce artificial state-space models as auxiliary tools to tackle
the cost minimization problem. In Section 3, we provide sufficient
conditions to ensure that an artificial state-space model is “adapted”
 to a cost function, meaning that the MAP estimates resulting from
the model coincide with the cost minimizers. These conditions can
be explicitly used to build adapted artificial models for a variety of
cost functions, as exemplified in Section 4. In Section 5 we describe
the basics of particle approximations of probability measures and the
specific SMC technique we propose to apply. Section 6 illustrates
the use of our approach for a stochastic volatility estimation problem
and, finally, Section 7 is devoted to the conclusions.

2. COSTS AND MODELS

Let \( \{x_t\}_{t \in \mathbb{N}^+] \) and \( \{y_t\}_{t \in \mathbb{N}^+} \) be discrete-time random sequences in
\( \mathbb{R}^{n_x} \) and \( \mathbb{R}^{n_y} \), respectively. Given a finite and fixed time horizon,
\( T > 0 \), we wish to estimate the signal \( x_{0:T} \) from the observations
\( y_{1:T} \) by solving the minimization problem

\[
X_T^* = \arg \min_{x_{0:T} \in \mathbb{R}^{(T+1)n_x}} C_T(x_{0:T}, y_{1:T}),
\]

where \( C_T : \mathbb{R}^{(T+1)n_x} \times \mathbb{R}^{Tn_y} \rightarrow \mathbb{R}^+ \) is a cost function and \( X_T^* \)
denotes the set of sequences in the signal space that minimize \( C_T \).

We assume that the cost can be constructed sequentially, i.e.,
the function \( C_t(\cdot, \cdot) \) can be obtained from the function \( C_{t-1}(\cdot, \cdot) \) by
some known mechanism. In particular, we write

\[
C_t(x_{0:t}, y_{1:t}) \triangleq H(C_{t-1}(x_{0:t-1}, y_{1:t-1}), c(x_{t-1:t}, y_{t})),
\]

\( 0 < t \leq T \), where \( H : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+ \) denotes the cost update
mechanism\(^1\), \( c : \mathbb{R}^{2n_x} \times \mathbb{R}^{n_y} \rightarrow \mathbb{R}^+ \) is the partial cost at time \( t \)
and the recursion is initialized with a prior cost \( C_0(x_0) \) which does not
depend on any observation.

Example 1 In econometrics, the standard deviation of a return time
series is called volatility, a variable that is modeled and estimated
with great difficulty [12]. Let the signal \( \{x_t\} \) be the logarithm of
the stochastic volatility of the series \( \{y_t\} \), i.e., \( \operatorname{Var}(y_t) = \exp(x_t) \).
We wish to estimate the random sequence \( x_{0:T} \) from a fixed series
of observations \( y_{0:T} \), but the true mechanism from which the signals
are generated (i.e., the dynamics of the stochastic process \( \{x_t\} \)
and the distribution of \( y_t \)) is unknown. Instead, we are given the cost
function

\[
\begin{align*}
C_0(x_0) & = s_0|x_0|, \\
C_t(x_{0:t-1}, y_{1:t-1}) & = C_{t-1}(x_{0:t-1}, y_{1:t-1}) + s_1|x_t - \alpha - \beta x_{t-1}| + s_2 |y_t - \exp(x_t)|,
\end{align*}
\]

for fixed and known parameters \( \alpha, \beta \in \mathbb{R} \) and scale parameters
\( s_i > 0, i = 0, 1, 2 \), and we aim to estimate \( x_{0:T} \) by solving

\[
X_T^* = \arg \min_{x_{0:T} \in \mathbb{R}^{(T+1)n_x}} C_T(x_{0:T}, y_{1:T}).
\]

We may realistically expect that problem (1) be hard to solve
in practical scenarios. Indeed, \( C_T(x_{0:T}, y_{1:T}) \) may be analytically
intractable and present multiple minima. Due to the high dimension
of the unknown, \( x_{0:T} \in \mathbb{R}^{(T+1)n_x} \), it may also be hard to devise
a convergent numerical algorithm with acceptable computational
complexity. However, if we can recast problem (1) as one of tracking
the state of a dynamic state-space model, then there is a pool of Monte
Carlo techniques that we can apply to numerically find approximate solutions. Therefore, we consider an artificial
state-space model where the signal of interest, \( \{x_t\} \), plays the role
of the system state and the measurements, \( \{y_t\} \), are the associated observations. Such model can be probabilistically described by the relationships

\[
\begin{align*}
\text{(state prior)} & \quad x_0 \sim \pi(x_0), \\
\text{(state transition)} & \quad x_t \sim \pi(x_t|x_{t-1}) \quad \text{and} \\
\text{(observation)} & \quad y_t \sim \pi(y_t|x_{0:t-1}, y_{1:t-1}).
\end{align*}
\]

where \( \pi(x_t|x_{t-1}) \) and \( \pi(y_t|x_{0:t-1}, y_{1:t-1}) \) are the artificial transition
(Markov) probability density function (pdf) and the artificial
likelihood, respectively, and \( \pi(x_0) \) is the artificial prior pdf. All
through the paper we use letter \( \pi \) to denote pdf’s. From Eqs. (5)-(7),
we can write down the artificial posterior pdf as

\[
\begin{align*}
\pi(x_{0:t}|y_{1:t}) & \propto \pi(y_t|x_{0:t-1}, y_{1:t-1})\pi(x_t|x_{t-1}) \\
& \quad \times \pi(x_{0:t-1}|y_{1:t-1}) \\
& = \pi(x_0) \prod_{k=1}^t \pi(y_k|x_{0:k}, y_{1:k-1})\pi(x_k|x_{k-1}).
\end{align*}
\]

If we choose the model adequately, we can make the MAP state
estimates resulting from it (for a fixed series of observations \( y_{1:T} \)) to coincide
with the signal sequence that minimizes the cost function
\( C_T \). In that case, we say that the artificial model is “adapted”
to the cost.

3. ADAPTED ARTIFICIAL MODELS

In this section we seek a general characterization of models adapted
to a given cost function. To avoid lengthy notations, in the sequel we
will use the following shorthand:

\[
\begin{align*}
\pi_t(x_{0:t}) & \triangleq \pi(x_{0:t}|y_{1:t}), \\
\pi_t(y_{t}) & \triangleq \pi(y_{t}|x_{0:t}, y_{1:t-1}), \\
\pi_t(x_{t}) & \triangleq \pi(x_{t}|x_{t-1}), \\
C_t(x_{0:t}) & \triangleq C_{t}(x_{0:t}, y_{1:t-1}), \quad \text{and} \\
c_t(x_{t-1:t}) & \triangleq c_{t}(x_{t-1:t}, y_{t}).
\end{align*}
\]

We first define formally when the state space model (5)-(7) is
adapted to

\[
\text{Definition 1 Let} \ y_{0:T} \text{ be a fixed sequence of observations. A}
\text{discrete-time random dynamical system specified by the prior pdf,}
\pi(x_0), \ \text{the likelihood function,} \ \pi_t(y_{t}), \ \text{and the transition pdf,}
\]

\[
\pi_t(x_{0:t}|y_{1:t}),
\]

\[
\pi_t(y_{t}|x_{0:t}, y_{1:t-1}),
\]

\[
\pi_t(x_{t}|x_{t-1}),
\]

\[
C_{t}(x_{0:t}, y_{1:t-1}) \text{, and}
\]

\[
c_{t}(x_{t-1:t}, y_{t}).
\]

\[
\text{is adapted to} \ C_{t}(x_{0:t}).
\]

\[
\pi_t(x_{0:t}) \triangleq \pi(x_{0:t}|y_{1:t}),
\]

\[
\pi_t(y_{t}) \triangleq \pi(y_{t}|x_{0:t}, y_{1:t-1}),
\]

\[
\pi_t(x_{t}) \triangleq \pi(x_{t}|x_{t-1}),
\]

\[
C_t(x_{0:t}) \triangleq C_{t}(x_{0:t}, y_{1:t-1}), \quad \text{and}
\]

\[
c_t(x_{t-1:t}) \triangleq c_{t}(x_{t-1:t}, y_{t}).
\]

\[1\] Simple examples may be \( H(a, b) = a + b \) or even \( H(a, b) = b \).
\( \pi_t(x_t), \) according to Eqs. (5), (6) and (7), respectively, is adapted to the cost function \( C_t(x_{0:t}) \) if, and only if,

\[
\arg \max_{x_{0:t} \in \mathbb{R}^{(t+1)n_x}} \pi_t(x_{0:t}) = \arg \min_{x_{0:t} \in \mathbb{R}^{(t+1)n_x}} C_t(x_{0:t}). \tag{15}
\]

for all \( t \in \{0, 1, \ldots, T\}. \)

For conciseness, we let \( X_t^* \triangleq \arg \max_{x_{0:t} \in \mathbb{R}^{(t+1)n_x}} \pi_t(x_{0:t}) \) and rewrite (15) as \( X_t^* = X_t^r. \)

Definition 1 is possibly too generic to test directly whether a state-space model and a cost function are adapted or not. Instead, we wish to have sufficient conditions that we can check in terms of the basic building blocks, namely \( \pi(x_t), \pi(y|t) \) and \( \pi(x_t) \) for the state-space system and \( c_t(x_{t-1:t}) \) and \( H(\cdot, \cdot) \) for the cost. The following result provides these conditions.

**Proposition 1** Let \( y_{k:T} \) be an arbitrary, but fixed, sequence of observations and let the pdf’s \( \pi_k(x_{0:k}), \pi_k(y_k), \pi_k(x_k) \), for \( k = 1, \ldots, T \), be proper. If we assume that:

(i) There exists a monotonically decreasing function \( F_0 : \mathbb{R}^+ \to \mathbb{R} \) such that \( C(x_0) = F_0(\kappa_0 \pi(x_0)) \), with proportionality constant \( \kappa_0 \) independent of \( x_0 \).

(ii) There exists a monotonically decreasing function \( F : \mathbb{R}^+ \to \mathbb{R} \) such that the inverse function \( F^{-1} \) factorizes the cost function as

\[
F^{-1}(H(C_{t-1}(x_{t-1:t}), c_t(x_{t-1:t}))) = F^{-1}(C_{t-1}(x_{t-1:t})) \times f(C_{t-1}(x_{t-1:t}), c_t(x_{t-1:t})),
\]

where \( 1 \leq t \leq T \) and \( f : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+ \) is a positive function.

(iii) Function \( f \) is such that

\[
f(C_{t-1}(x_{t-1:t}), c_t(x_{t-1:t})) \propto \pi_t(y_t)\pi_t(x_t),
\]

with a proportionality constant independent of \( x_{t-1:t} \).

Then \( C_t(x_{0:t}) = F(\kappa_t \pi_t(x_{0:t})) \), for all \( t \in \{1, \ldots, T\} \) and some constant \( \kappa_t \) independent of \( x_{0:t} \) and \( X_t^* = X_t^r \).

Since \( F \) is monotonically decreasing, \( C_t(x_{0:t}) = F(\kappa_t \pi_t(x_{0:t})) \) implies \( X_t^r = X_t^r. \) Function \( f \) intuitively represents that updates the cost function, \( C_{t-1}(x_{t-1:t}) \), with the updates of the posterior pdf, \( \pi_{t-1}(x_{0:t-1}) \), at time \( t \), when a new observation, \( y_t \), is collected. The argument of the proof is sketched below.

**Proof (outline):** The proof proceeds by induction in \( t \). At time \( t = 0 \), by assumption (i) we have \( C_0(x_0) = F_0(\kappa_0 \pi(x_0)) \), hence \( X_0^* = X_0^r \). At time \( t - 1 \), we assume that \( C_{t-1}(x_{t-1:t}) = F(\kappa_{t-1} \pi_{t-1}(x_{0:t-1})) \) and, as a consequence, \( \pi_{t-1}(x_{0:t-1}) \propto F^{-1}(C_{t-1}(x_{t-1:t})). \)

At time \( t \), we apply assumption (ii) to obtain

\[
F^{-1}(C_t(x_{0:t})) = F^{-1}(H(C_{t-1}(x_{t-1:t}), c_t(x_{t-1:t}))) =
F^{-1}(C_{t-1}(x_{t-1:t})) \times f(C_{t-1}(x_{0:t-1}), c_t(x_{t-1:t})).
\]

By the induction hypothesis,

\[
F^{-1}(C_{t-1}(x_{0:t-1})) \propto \pi_{t-1}(x_{0:t-1})
\]

and, by assumption (iii),

\[
f(C_{t-1}(x_{0:t-1}), c_t(x_{t-1:t})) \propto \pi_t(y_t)\pi_t(x_t),
\]

hence we can substitute (19) and (20) into Eq. (18) to yield

\[
F^{-1}(C_t(x_{0:t})) \propto \pi_{t-1}(x_{0:t-1})\pi_t(y_t)\pi_t(x_t) \propto \pi_t(x_t)
\]

and, as a consequence, \( C_t(x_{0:t}) = F(\kappa_t \pi_t(x_{0:t})) \), for some constant \( \kappa_t \), and \( X_t^r = X_t^r \), since \( F \) is a monotonically decreasing function. \( \square \)

### 4. FAMILIES OF COST FUNCTIONS

Let us put Proposition 1 to work. We consider three types of cost functions: purely additive forms, \( H(a, b) = a + b \), additive functions with a forgetting factor, \( H(a, b) = \theta a + b \), where \( 0 < \theta < 1 \), and the nonlinearity \( H(a, b) = \max(a, b) \). In the three cases, we explicitly show the functions \( F \) and \( f \) that relate \( C_t(x_{0:t}) \) with \( \pi_t(x_{0:t}). \)

#### 4.1. Additive cost

The additive form \( C_t(x_{0:t}) = C_{t-1}(x_{0:t-1}) + c_t(x_{t-1:t}) \) can be related to the posterior pdf easily by means of the monotonically decreasing functions \( F_0(a) = F(a) = -\log(a) \) and \( f(a, b) = f(b) = \exp(-b) \), which yield

\[
\begin{align*}
F^{-1}(C_{t-1}(x_{0:t-1}) + c_t(x_{t-1:t})) &= \exp\{-C_{t-1}(x_{0:t-1})\} \exp\{-c_t(x_{t-1:t})\} = \\
&\exp\left\{-C_0 - \sum_{k=1}^{t} c_k(x_{k-1:k})\right\}.
\end{align*}
\]

For this formal decomposition to be valid, we require integrability of the terms \( F^{-1}(C(x_0)) \) and \( f(c_k(x_{k-1:k})) \), i.e.,

\[
\int \cdots \int \exp\{-C(x_0)\} dx_0 < \infty \quad \text{and} \quad \int \cdots \int \exp\{-c_k(x_{k-1:k})\} dx_k < \infty
\]

for each \( k \in \{1, \ldots, t\} \).

**Example 2** The cost function given in Example 1 has a purely additive form, with \( C_0(x_0) = s_0|x_0| \) and

\[
c_t(x_{t-1:t}) = s_1|x_t - \alpha - \beta x_{t-1}| + s_2 |y_t^2 - \exp\{x_t\}|.
\]

According to the assumptions of Proposition 1,

\[
\exp\{-c_t(x_{t-1:t})\} \propto \pi_t(y_t)\pi_t(x_t),
\]

**hence we can write**

\[
\pi_t(y_t) \propto \exp\{-s_2 |y_t^2 - \exp\{x_t\}|\},
\]

\[
\pi_t(x_t) \propto \exp\{-s_1 |x_t - \alpha - \beta x_{t-1}|\}.
\]

Both \( x_t \) and \( z_t = y_t^2 \) are conditionally Laplacian rv’s. To be specific, if we let \( \text{Lap}(.; m, h) \) denote the Laplace pdf with mean \( m \) and scale parameter \( h \), then

\[
\pi(z_t|x_t) = \text{Lap}(z_t; \exp(x_t), s_2^{-1})
\]

and

\[
\pi_t(x_t) = \text{Lap}(x_t; \alpha + \beta x_{t-1}, s_1^{-1}).
\]

Since \( F_0 = -\log, \pi(x_0) \propto \exp\{-s_0|x_0|\} \) and, as a consequence, \( \pi(x_0) = \text{Lap}(x_0; 0, s_0^{-1}). \)
Example 3 It is illustrative to consider the case of a quadratic cost to estimate the sequence \( x_{0:T} \) of Example 1. Specifically, let \( C_0(x_0) = s_2 x_0^2 \) and \( C_t(x_{t-1}) = C_{t-1}(x_{t-1}) + c_t(x_{t-1},t) \) where
\[
c_t(x_{t-1},t) = s_2 \left( y_t^2 - \exp(x_t) \right)^2 + s_1 \left( x_t - \alpha - \beta x_{t-1} \right)^2. \tag{31}
\]
The derivation of the adapted model is very similar to Example 2, but the resulting densities are Gaussian instead of Laplacian, i.e., \( \pi(x_0) = N \left( x_0; 0, \frac{1}{\pi_0} \right) \), \( \pi(x_t) = N \left( x_t; \alpha + \beta x_{t-1}, \frac{1}{\pi_{2,t}} \right) \) and \( p(z_t|x_t) = N \left( z_t; \exp(x_t), \frac{1}{\pi_{2,t}} \right) \), where the latter pdf implies that
\[
\pi_t(y_t) \propto \exp \left\{ -s_2 \left( y_t^2 - \exp(x_t) \right)^2 \right\}. \tag{32}
\]

4.2. Additive cost with a forgetting factor

We now consider the form \( C_t(x_{t-1}) = \theta C_{t-1}(x_{t-1}) + c_t(x_{t-1},t) \), where \( 0 < \theta < 1 \) is a constant. This class of cost functions is commonly used in signal processing applications and its physical interpretation is that “old” observations are “forgotten” in such a way that they have little effect on current signal estimates.

Assume a fixed time horizon \( T > 0 \). The complete expansion of the cost \( C_T(x_{0:T},y_{1:T}) \) in terms of the partial costs \( c_k(x_{k-1:k}) \), \( k = 1, \ldots, T \), and the prior \( C_0(x_0) \) is
\[
C_T(x_{0:T}) = \theta^T C_0(x_0) + \sum_{k=1}^{T} \theta^{T-k} c_k(x_{k-1:k}). \tag{33}
\]
Eq. (33) suggests that we can handle this family of costs in the same way as in Section 4.1 if we redefine
\[
\tilde{C}_0(x_0) = \theta^T C_0(x_0), \tag{34}
\]
\[
\tilde{C}_k(x_{0:k}) = \tilde{C}_{k-1}(x_{0:k-1}) + \theta^{T-k} c_k(x_{k-1:k}), \tag{35}
\]
which ensures that \( \tilde{C}_T(x_{0:T}) = C_T(x_{0:T}) \), and we choose \( F_0(a) = F(a) = -\log(a) \) and \( f(a,b) = \exp(-b) \). However, this approach is not fully satisfactory because the resulting state-space model is adapted to \( C_T \), but not to any other \( C_{T'} \), with \( T' \neq T \).

We can provide a more compelling argument if we notice the recursive relationship between the artificial models adapted to \( C_t \) and \( C_{t-1} \). First, since the two models are different, we need a different notation for the posterior pdf’s resulting from each one of them. Hence, we let \( \pi(x_{t-1}) \) denote the smoothing density obtained from the model adapted to \( C_t \). Let us remark that the subscript \( (t) \) refers to the cost for which the state-space model is built, hence we have a complete family of pdf’s, \( \pi(x_0:k|y_{1:k}) \), \( 0 \leq k \leq t \), for each \( t \). Also, for fixed \( k \), \( \pi(x_{t-1:k}|y_{1:k}) \neq \pi(x_{t-1:k}|y_{1:k}) \) in general, whenever \( t_1 \neq t_2 \). The factorization of \( \pi(x_{t-1:k}|y_{1:k}) \) given by \( F(a) = -\log(a) \) and \( f(a,b) = \exp(-b) \) is
\[
\pi_t(x_{t-1:t}|y_{1:t}) \propto \exp \left\{ -\theta^T C_0(x_0) \right\} \prod_{k=1}^{t} \exp \left\{ -\theta^{T-k} c_k(x_{k-1:k}) \right\} \tag{36}
\]
and it is immediate to realize that
\[
\pi_{t-1}(x_{t-1:t}|y_{1:t}) \propto \pi_t(x_{t-1:t}|y_{1:t}) \exp \left\{ -c_k(x_{k-1:k}) \right\}. \tag{37}
\]
An interpretation of this result is that we can generate an adapted state-space model for each \( t \) and the recursive relationship of Eq. (37) enables the computation of the posterior pdf for the \( (t+1) \)-th model by simply updating the posterior pdf from the \( t \)-th model.

It is straightforward to rebuild Proposition 1 in order to make it valid for the recursive pdf decomposition
\[
\pi_t(x_{t-1:t}|y_{1:t}) \propto \pi_{t-1}(x_{t-1:t-1}|y_{1:t-1}) \pi_t(x_{t-1:t}|y_{1:t}) \times \pi_{t-1}(x_{t-1}). \tag{38}
\]
In particular, the choice \( F(a) = -\log(a) \) and \( f(a,b) = \exp(-b) \) still yields the adequate factorization that ensures \( C_t(x_{0:T}) = \int \pi_t(x_{0:T}|y_{1:T}) \, dx_T \) and \( \pi_0 = \pi_{t-1} \) for all \( t \in \{1, 2, \ldots, T\} \). Of course, the validity of this derivation requires that \( c_k(x_{k-1:k}) > 0 \), for all \( k \), and both \( \pi_0 = -\log(C_0(x_0)) \) and \( \exp(-c_k(x_{k-1:k})) \), \( k = 1, \ldots, t \), have finite integrals.

4.3. Costs with max(\( \cdot, \cdot \)) nonlinearities

The form \( C_t(x_{t-1}) = \max(C_{t-1}(x_{t-1}), c_t(x_{t-1},t)) \) can also be factorized by means of \( F_0(a) = F(a) = -\log(a) \). From this, we construct
\[
\begin{align*}
F^{-1} \left( \max(C_{t-1}(x_{t-1}), c_t(x_{t-1},t)) \right) &= \exp \left\{ \max \left( -C_{t-1}(x_{t-1}), c_t(x_{t-1},t) \right) \right\} \\
&= \exp \left\{ -C_{t-1}(x_{t-1}) \right\} \exp \left\{ -c_t(x_{t-1},t) \right\} \\
&\leq \min \left( \exp \left\{ -C_{t-1}(x_{t-1}) \right\}, \exp \left\{ -c_t(x_{t-1},t) \right\} \right).
\end{align*}
\]
By inspection of (39) we quickly notice that
\[
\begin{align*}
f(C_{t-1}(x_{t-1}), c_t(x_{t-1},t)) &= \exp \left\{ -c_t(x_{t-1},t) \right\} \\
&= \exp \left\{ \min \left( \exp \left\{ -C_{t-1}(x_{t-1}) \right\}, \exp \left\{ -c_t(x_{t-1},t) \right\} \right) \right\}.
\end{align*}
\]
Again, the validity of the formal derivation relies on the assumptions that \( c_k(x_{k-1:k}) > 0 \) and \( \exp(-c_k(x_{k-1:k})) \) be integrable for any \( k \).

Example 4 Let us consider the cost function
\[
C_t(x_{t-1}) = \max(C_{t-1}(x_{t-1}), c_t(x_{t-1},t)),
\]
which ensures that \( \tilde{C}_T(x_{0:T}) = C_T(x_{0:T}) \), and we choose \( F_0(a) = F(a) = -\log(a) \) and \( f(a,b) = \exp(-b) \). However, this approach is not fully satisfactory because the resulting state-space model is adapted to \( C_T \), but not to any other \( C_{T'} \), with \( T' \neq T \).

Combining assumption (iii) of Proposition 1 and Eq. (40), we readily arrive at
\[
\pi_t(y_t) \propto \exp \left\{ -s_2 \left( y_t^2 - \exp(x_t) \right) \right\} \min \left( \exp \left\{ -C_{t-1}(x_{t-1}) \right\}, \exp \left\{ -c_t(x_{t-1},t) \right\} \right).
\]

Note that \( y_t \to \pm \infty \) implies that \( y_t^2 - \exp(x_t) \to \pm \infty \) and, as a consequence, \( \lim_{y_t \to \infty} \pi_t(y_t) = \kappa \), where \( \kappa > 0 \) is a constant. Therefore, for \( \pi_t(y_t) \) to be a proper density the support of \( y_t \) must be a subset \( \mathbb{R} \subset \mathbb{R}^n \) with finite Lebesgue measure.

5. PARTICLE APPROXIMATIONS

We have recast the minimization of \( C_T(x_{0:T},y_{1:T}) \) with respect to \( x_{1:T} \) into a problem of MAP estimation for an artificial state-space model. However, this is also intractable for most models of practical interest (linear Gaussian systems being the exception) and we need to resort to numerical techniques in order to compute solutions. In this paper, we advocate the use of SMC methods to build a particle approximation of the a posteriori smoothing probability measure, from which MAP estimates can be computed.
5.1. Monte Carlo approximation of smoothing densities

Let \( \{z^{(i)}_{t}\}_{i=1}^{M} \) be a set of \( M \) independent and identically distributed (i.i.d.) samples from the smoothing distribution. These samples are often called particles and we seek a point-mass approximation of \( \pi(x_{0:T}|y_{1:T}) \) of the form

\[
\pi(x_{0:T}|y_{1:T}) \approx \hat{\pi}(x_{0:T}|y_{1:T}) = \frac{1}{M} \sum_{i=1}^{M} \delta_{i}(x_{0:T}),
\]

where \( \delta_{i}(\cdot) \) is the delta measure centered at \( z^{(i)}_{0:T} \). From (45) we can numerically approximate moments of the true smoothing distribution. E.g., let \( g \) be an integrable function with support in \( \mathbb{R}^{(T+1)n} \), then we can write the expected value of \( g(x_{0:T}) \) given \( y_{1:T} \) as

\[
E[g(x_{0:T})|y_{1:T}] = \int g(x_{0:T})\pi(x_{0:T}|y_{1:T})dx_{0:T} \\
\approx \int g(x_{0:T})\hat{\pi}(x_{0:T}|y_{1:T})dx_{0:T} \\
= \frac{1}{M} \sum_{i=1}^{M} g(z^{(i)}_{0:T}).
\]

There is an obvious difficulty in drawing the desired particles from \( \pi(x_{0:T}|y_{1:T}) \). In this paper, we propose to follow the procedure introduced in [11], which builds up a set \( \{z^{(i)}_{t}\}_{i=1}^{M} \), with each \( z^{(i)}_{t} \) approximately drawn from \( \pi(x_{0:T}|y_{1:T}) \), by sweeping the data \( y_{1:T} \) sequentially. In the first (forward) pass, we use a particle filtering algorithm to construct approximations of the densities, \( \pi(x_{t}|y_{1:t}) \). Specifically, the particle approximation of \( \pi(x_{t}|y_{1:t}) \) involves the computation of a set of weighted samples \( \{x_{t}^{(i)}|w_{t}^{(i)}\}_{i=1}^{N} \) for each \( t \), with \( \pi(x_{t}|y_{1:t}) \approx \hat{\pi}(x_{t}|y_{1:t}) = \sum_{i=1}^{N} w_{t}^{(i)} \delta_{i}(x_{t}) \). Then, in order to draw from the smoothing distribution, we perform a backward sweep over the data. The result is a stream \( \tilde{x}_{0:T} \), approximately drawn from \( \pi(x_{0:T}|y_{1:T}) \), where \( \tilde{x}_{t} \in \{x_{t}^{(i)}\}_{i=1}^{N} \) for all \( t \). By performing \( M \) backward sweeps we build the set of independent particles \( \{\tilde{x}^{(i)}_{0:T}\}_{i=1}^{M} \). The asymptotic convergence of the approximation is proved in [11].

5.2. MAP estimation

The discretization of the state space provided by the particle sets \( \{x_{t}^{(i)}|w_{t}^{(i)}\}_{i=1}^{N} \), for \( t = 0,1,\ldots,T \), enables the application of dynamic programming algorithms for MAP estimation [13]. In this paper, however, we investigate the approximation of the MAP estimate \( \tilde{x}^{\pi}_{0:T} \in X_{T}^{\pi} \) using Eq. (45), by means of the integral

\[
\tilde{x}^{\pi}_{0:T} = \int_{A} A_{x_{0:T}}\pi(x_{0:T}|y_{1:T})dx_{0:T} \\
\approx \int_{A} A_{x_{0:T}}\hat{\pi}(x_{0:T}|y_{1:T})dx_{0:T} \\
= \frac{\sum_{i=1}^{M} I_{A}(\tilde{x}^{(i)}_{0:T})\tilde{x}^{(i)}_{0:T}}{\sum_{k=1}^{M} I_{A}(\tilde{x}^{(k)}_{0:T})},
\]

where \( A \subset \mathbb{R}^{(T+1)n} \) is a set with high probability, according to the approximation \( \hat{\pi}(x_{0:T}|y_{1:T}) \), and \( I_{A}(z) \) is an indicator function that yields \( I_{A}(z) = 1 \) if \( z \in A \) and zero otherwise. We construct \( A \) as an open ball centered at the trajectory \( \tilde{x}^{(i)}_{0:T} \), with the highest posterior density. Specifically, if we let \( \tilde{x}_{0:T}^{\pi} = \arg\max_{x_{0:T} \in \{x_{0:T}^{(i)}\}_{i=1}^{M}} \pi(x_{0:T}|y_{1:T}) \), then

\[
A = \left\{ x_{0:T} \in \{x_{0:T}^{(i)}\}_{i=1}^{M} : d(x_{0:T},\tilde{x}_{0:T}^{\pi}) < r \right\},
\]

for some radius \( r > 0 \) and a proper distance \( d \). We note that \( \pi(\tilde{x}_{0:T}^{\pi}|y_{1:T}) \) can be computed sequentially in the backward pass for each \( i \).

6. COMPUTER SIMULATIONS

We present some computer simulation results to illustrate the performance of the proposed approach. Let us consider the stochastic volatility estimation problem described in Example 1, the additive cost function of Example 2 (with scale parameters \( s_{1} = 1, s_{2} = 1, 2 \)) and the obtained adapted model. For the simulation, we generate the random signal \( \{x_{t}\} \) and the measurements \( \{y_{t}\} \) using the dynamic system

\[
x_{t} = \alpha + \beta x_{t-1} + u_{t},
\]

\[
y_{t} = \exp(\frac{x_{t}}{2}) v_{t},
\]

where \( \alpha = -2.5, \beta = 0.5 \) and the state and observation noise components have mixture Gaussian distributions, \( u_{t} \sim cN(u_{t};0,\frac{1}{2}) + (1-c)N(u_{t};0,5) \) and \( v_{t} \sim \nu N(u_{t};0,\frac{1}{2}) + (1-\nu)N(u_{t};0,3) \), respectively, with \( c = \frac{9}{10} \) and \( \nu = \frac{8}{10} \). The prior pdf is standard Gaussian, \( x_{0} \sim N(x_{0};0,1) \). We note that \( \text{Var}(v_{t}) = 1 \), hence \( \text{Var}(y_{t}) = \exp(x_{t}) \), as stated in Example 1. Let us also remark that system (49)-(50) is used here for the purpose of simulating sample trajectories \( x_{0:T} \) and \( y_{0:T} \). This system is different from the adapted model derived in Example 2, which is used for particle smoothing and MAP estimation.

We first illustrate the transformation of the cost function into a smoothing density for the simplest case, \( T = 1 \), and a fixed observation \( y_{1} = -0.5 \). Figure 1 shows the 3-dimensional plot of \( C_{t}(x_{0:1},y_{1}) = 20 + |x_{0:1}| + |x_{1} - \alpha - \beta x_{0} + |y_{1} - \exp(x_{t})| \) (the addition of 20 is for the sake of visibility) together with the contour plot of the posterior pdf, \( \pi(x_{0:1}|y_{1}) \), for the adapted model of Example 2. The minimum of \( C_{t} \) and the maximum of \( \pi(x_{0:1}|y_{1}) \) are marked and shown to coincide.

In order to numerically test the performance of our approach, we have generated \( K = 50 \) independent realizations of \( y_{0:T} \), with \( T = 80 \), using the system (49)-(50) and have approximated the cost minimizer \( \hat{x}_{0:T} = \arg\min_{x_{0:T} \in \mathbb{R}^{T+1}} C_{t}(x_{0:T},y_{1:t}) \) using the accelerated random search (ARS) technique of [2] for \( t = 20, 40, 60 \) and 80. We have also tackled the same optimization problems by means of the proposed technique, with \( N = 1000 \) particles in the bootstrap filter applied in the forward sweep and \( M = 1000 \) trajectories in the particle smoother. The MAP estimates, \( \hat{x}_{0:T}^{\pi} \), are approximated by the method in Section 5.2, with the set \( A \) constructed using Euclidean distance and radius \( r = 1.5 \).

Figure 2 shows the results. We plotted the mean and standard deviation of the costs of the minimizers provided by the ARS and SMC techniques for \( t = 20, 40, 60 \) and 80. We observe that the solutions computed by ARS exhibited a lower cost for \( t = 20 \). Then, the SMC algorithm clearly beat the ARS technique for \( t \geq 40 \). The latter is a consequence of its superior efficiency when the signal of interest is high dimensional.
Fig. 1. Representation of the cost function $C_1(x_0, -0.5) = 20 + |x_0| + |x_1 - 0.5 - 0.25 - \exp(x_1)|$ (mesh plot) and the posterior pdf for the adapted state-space model, $\pi(x_0|y_1 = -0.5)$ (contour plot), that illustrates the coincidence of their extrema.

7. CONCLUSIONS

We have introduced a novel method for the sequential optimization of a certain class of cost functions in high-dimensional spaces. The proposed approach casts the optimization problem into one of MAP estimation in a properly constructed dynamical system. The latter problem can be handled using convergent SMC techniques, which perform efficiently when the dimension of the signal of interest is large. The performance of our method is illustrated by computer simulations for a problem of stochastic volatility estimation.

8. REFERENCES