Particle filters with auxiliary Markov transition. Application to crossover and to multitarget tracking

Audrey Cuillery, François Le Gland

Particle filters with auxiliary Markov transition
Application to crossover and to multitarget tracking

Audrey Cuillery
université de Rennes 1
campus de Beaulieu
35042 Rennes Cédex, France
email: acuillery@gmail.com

François Le Gland
INRIA Rennes and IRMAR
campus de Beaulieu
35042 Rennes Cédex, France
email: francois.le_gland@inria.fr

Abstract

This work introduces a new class of particle filters, that include an auxiliary Markov transition in their design. Actually, it was motivated by potential application to multitarget tracking, but the solution provided may be of practical interest elsewhere.

This work also shows how to include a crossover step in sequential Monte Carlo methods. It is well known that sequential Monte Carlo methods can be interpreted in terms of implementing *selection* and *mutation* steps, using the language of evolutionary algorithms. However, most general evolutionary algorithms also include a *crossover* step that has not been considered so far in sequential Monte Carlo methods.

A prototypical situation where crossover is needed, or at least could be useful, is multitarget tracking. In multitarget tracking, it may happen that some targets in a multitarget particle are good proxies, but are not going to be selected just because the other targets in the same multitarget particle are bad proxies. This is unfair, and a better design would be to produce shuffled multitarget particles such that the particle for each different target can be replicated from a different multitarget particle.

An efficient solution has been proposed in the literature under a *posterior independence* assumption that unfortunately is almost never met in practical situations. This work provides another solution that does not rely on the posterior independence assumption and that is based on introducing an auxiliary Markov transition in the design. This approach can be seen as an extension of the auxiliary particle filter, and may be of independent interest, outside the application to crossover and to multitarget tracking. Optimization of the design parameters is also addressed.

Keywords: Bayesian filtering, particle filtering, mutation/selection/crossover, optimal design, convex optimization, stochastic approximation

1 Introduction

To fix ideas, consider a hidden Markov model, i.e. a hidden (not observed) Markov chain $\{X_k\}$ with model transition kernel

$$Q_k(x, dx') = \mathbb{P}[X_k \in dx' \mid X_{k-1} = x] \ ,$$

and an observation sequence $\{Y_k\}$ related to hidden states, e.g.

$$Y_k = h_k(X_k) + V_k \qquad \text{with} \qquad V_k \sim q_k^V(v)\, dv \ ,$$

which defines a likelihood function

$$g_k(x') = q_k^V(Y_k - h_k(x')) \ ,$$

up to a normalizing constant. The solution to the Bayesian or MMSE estimation relies on the Bayesian filter

$$\mu_k(dx') = \mathbb{P}[X_k \in dx' \mid Y_0 \cdots Y_k] \ ,$$

i.e. on the conditionnal distribution of the current state X_k given past observations $(Y_0, \cdots, Y_k)$. The Bayesian filter satisfies a simple recurrent equation

$$\mu_{k-1} \xrightarrow{\text{prediction}} \eta_k = \mu_{k-1}\, Q_k \xrightarrow{\text{correction}} \mu_k = g_k \cdot \eta_k$$

where the predictor

$$\eta_k(dx') = \mu_{k-1}\, Q_k(dx') = \int_E \mu_{k-1}(dx)\, Q_k(x, dx') \ ,$$

is obtained by acting the transition kernel $Q_k(x, dx')$ on the filter $\mu_{k-1}(dx)$, and the filter

$$\mu_k(dx') = (g_k \cdot \eta_k)(dx') \propto g_k(x')\, \eta_k(dx') \ ,$$

is obtained by applying the Bayes rule with the predictor $\eta_k(dx')$ and the likelihood function $g_k(x')$. The idea behind particle filtering, or sequential Monte Carlo methods, is to look for an approximation of the Bayesian filter as a weighted empirical probability distribution

$$\mu_k \approx \mu_k^N = \sum_{i=1}^N w_k^i\, \delta_{\xi_k^i} \qquad \text{with} \qquad \sum_{i=1}^N w_k^i = 1 \ ,$$

associated with a population of N particles characterized by their *positions* $(\xi_k^1, \cdots, \xi_k^N)$ in E, and their non–negative normalized *weights* $(w_k^1, \cdots, w_k^N)$. Plugging the particle approximation

$$\mu_{k-1}^N = \sum_{i=1}^N w_{k-1}^i\, \delta_{\xi_{k-1}^i} \ ,$$

in the recurrent equation, one obtains the finite mixture

$$\mu_{k-1}^N\, Q_k(dx') = \sum_{i=1}^N w_{k-1}^i\, Q_k(\xi_{k-1}^i, dx') \ ,$$

as an approximation of the predictor $\eta_k = \mu_{k-1}\, Q_k$, and the probability distribution (up to a normalizing constant)

$$g_k(x')\, \sum_{i=1}^N w_{k-1}^i\, Q_k(\xi_{k-1}^i, dx') \ ,$$

as an approximation of the filter $\mu_k = g_k \cdot \eta_k$. To further approximate the 'target' probability distribution

$$\underbrace{g_k(x')}_{g(x')} \underbrace{\sum_{i=1}^N w_{k-1}^i \, Q_k(\xi_{k-1}^i, dx')}_{\eta(dx')} \qquad \text{with} \qquad \sum_{i=1}^N w_{k-1}^i = 1 \ ,$$

seen as a Boltzmann–Gibbs probability distribution, where $\eta(dx')$ is a mixture probability distribution, the simplest importance sampling approximation is obtained as

$$\mu_k^N = \sum_{i=1}^N w_k^i \, \delta_{\xi_k^i} \ ,$$

where independently for any $i = 1, \cdots, N$ the i–th particle ξ_k^i is sampled from $\eta(dx')$ and is assigned a weight w_k^i using $g(x')$. This is just a possible approximation, and it is well known that most sequential Monte Carlo methods exhibit a similar pattern and can be interpreted in terms of implementing *selection* and *mutation* steps, using the language of evolutionary algorithms. However, most general evolutionary algorithms also include a *crossover* step that is not present in sequential Monte Carlo methods. It is the objective of this work to make it possible to include a crossover step in sequential Monte Carlo methods.

A prototypical situation where crossover is needed, or at least could be useful, is multitarget tracking. In multitarget tracking, many particle approximations are available to sample from the filtering probability distribution, with the effect that a new multitarget particle is obtained by replicating *globally* an existing multitarget particle, i.e. in the new multitarget particle, the particle for each different target is replicated from the same multitarget particle. However, it may happen that a multitarget particle provides a good proxy for some targets and a bad proxy for the other targets, i.e. it may happen that some targets in a multitarget particle are good proxies, but are not going to be selected just because the other targets in the same multitarget particle are bad proxies. In other words, a multitarget particle is likely to be selected only if jointly all the targets deserve to be selected. This is unfair, and a better design would be to produce shuffled multitarget particles such that the particle for each different target can be replicated from a different multitarget particle. An efficient solution has been proposed by Ubéda–Medina et al. [5] under a *posterior independence* assumption, see Yi et al. [6], that unfortunately is almost never met in practical situations. The objective of this work is to propose another solution that does not rely on the posterior independence assumption and that is based on introducing an auxiliary Markov transition. This approach can be seen as an extension of the auxiliary particle filter, see Pitt and Shephard [4], and may be of independent interest, outside the application to crossover and to multitarget tracking.

This paper contains several sections, with many repetitions between the first four sections. Section 2 reviews the auxiliary particle filter. The next two sections, Section 3 and Section 4, consider the effect of adding a Markov transition in the design, using an auxiliary transition matrix and an auxiliary transition kernel, respectively. Section 5 uses this additional degree of freedom so as to introduce crossover between particles in association with a given partition of the state vector. Each of these four sections is organized along the same lines

- interpret the target probability distribution as the marginal of another probability distribution defined on a suitable *augmented* state space,

- design a class of particle filters on the augmented state space, based on a suitable *importance decomposition* and on the *importance sampling* principle,
- check that the more general class of particle filters contains known (and less general) classes of particle filters as special cases, i.e. *generalization* holds,
- check that the particle approximation of the normalization constant is an unbiased estimator, provide an expression of the variance of the approximation, characterize the *optimum design parameters*, and evaluate the associated minimum variance.

The last two sections, Section 6 and Section 7, consider two special cases where optimization of the design parameters is made simple, if not explicit, and the existence of a *unique* minimizer can be proved.

2 Basics on the auxiliary particle filter

Consider the probability distribution

$$\mu(dx') \propto \underbrace{g_k(x')}_{g(x')} \underbrace{\sum_{i=1}^N w_{k-1}^i \, Q_k(\xi_{k-1}^i, dx')}_{\eta(dx')} \qquad \text{with} \qquad \sum_{i=1}^N w_{k-1}^i = 1 \ ,$$

defined on E, seen as a Boltzmann–Gibbs probability distribution, where $\eta(dx')$ is a mixture probability distribution, and its importance sampling approximation

$$\mu^N = \sum_{i=1}^N w_k^i \, \delta_{\xi_k^i} \ ,$$

where independently for any $i = 1, \cdots, N$ the i–th particle ξ_k^i is sampled from $\eta(dx')$ and is assigned a weight w_k^i using $g(x')$. In practice, independently for any $i = 1, \cdots, N$

- the index a_k^i is sampled from the probability vector $(w_{k-1}^1, \cdots, w_{k-1}^N)$,

then, setting $a = a_k^i$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^a, dx')$,
- and it receives the weight $w_k^i \propto g(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \ .$$

The idea behind auxiliary particle filtering is to see $\mu(dx')$ as the marginal probability distribution on E of the joint probability distribution

$$\mu_{\text{aux}}^a(dx') \propto g_k(x') \, w_{k-1}^a \, Q_k(\xi_{k-1}^a, dx') \ ,$$

defined on the augmented space $\{1, \cdots, N\} \times E$. Indeed, summation with respect to the index $a \in \{1, \cdots, N\}$ yields

$$g_k(x') \sum_{a=1}^{N} w_{k-1}^a \, Q_k(\xi_{k-1}^a, dx') = g(x') \, \eta(dx') \; , \tag{1}$$

an identity for unnormalized distributions, that carries over to normalized probability distributions. Introducing the auxiliary probability vector $(\lambda_{\text{aux}}^1, \cdots, \lambda_{\text{aux}}^N)$, the components of which are known as the *auxiliary weights*, makes it possible to define the importance decomposition

$$\mu_{\text{aux}}^a(dx') \propto \underbrace{g_k(x') \, \frac{w_{k-1}^a}{\lambda_{\text{aux}}^a}}_{g_{\text{aux}}^a(x')} \; \underbrace{\lambda_{\text{aux}}^a \, Q_k(\xi_{k-1}^a, dx')}_{\eta_{\text{aux}}^a(dx')} \; , \tag{2}$$

the Monte Carlo approximation for the predictor

$$\eta_{\text{aux}}^N = \frac{1}{N} \sum_{i=1}^{N} \delta_{(a_k^i, \xi_k^i)} \qquad \text{hence} \qquad \langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle = \frac{1}{N} \sum_{i=1}^{N} g_{\text{aux}}^{a_k^i}(\xi_k^i) \; ,$$

and the importance sampling approximation for the filter

$$\mu_{\text{aux}}^N = \sum_{i=1}^{N} w_k^i \; \delta_{(a_k^i, \xi_k^i)} \qquad \text{and its marginal} \qquad \mu^N = \sum_{i=1}^{N} w_k^i \; \delta_{\xi_k^i} \; ,$$

where independently for any $i = 1, \cdots, N$ the i–th particle (a_k^i, ξ_k^i) is sampled from $\eta_{\text{aux}}^a(dx')$ and is assigned a weight w_k^i using $g_{\text{aux}}^a(x')$. In practice, independently for any $i = 1, \cdots, N$

- the index a_k^i is sampled from the auxiliary probability vector $(\lambda_{\text{aux}}^1, \cdots, \lambda_{\text{aux}}^N)$,

then, setting $a = a_k^i$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^a, dx')$,
- and it receives the weight $w_k^i \propto g_{\text{aux}}^a(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \; \frac{w_{k-1}^a}{\lambda_{\text{aux}}^a} \; ,$$

see Algorithm 1.

This class of particle filters depends upon $N-1$ design parameters, that are the components of the auxiliary probability vector λ_{aux} (a probability vector of dimension N) subject to normalization constraint. This class of particle filters contains the class of ordinary particle filters as a special case. Indeed, if $\lambda_{\text{aux}}^a = w_{k-1}^a$ for any $a \in \{1, \cdots, N\}$, then the resulting particle filter reduces to the ordinary (non auxiliary) particle filter.

Recall the particle approximation

$$\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle = \frac{1}{N} \sum_{i=1}^{N} g_{\text{aux}}^{a_k^i}(\xi_k^i) \; .$$

Algorithm 1: Particle filter with auxiliary weights

input : $\lambda_{\rm aux}$, $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\rm aux} = (\lambda_{\rm aux}^1, \cdots, \lambda_{\rm aux}^N)$
  Propagate the state vector
  Sample $\xi_k^i \sim Q_k(\xi_{k-1}^a, dx')$
  Evaluate the unnormalized weight
  Set $w_k^i = g_k(\xi_k^i) \, \dfrac{w_{k-1}^a}{\lambda_{\rm aux}^a}$
end
Normalize the weights

for the normalizing constant $\langle \eta_{\rm aux}, g_{\rm aux} \rangle$, where the random variables (a_k^i, ξ_k^i) for $i = 1, \cdots, N$ are i.i.d. with common probability distribution $\eta_{\rm aux}^a(dx')$, hence

$$\mathbb{E}[\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle] = \langle \eta_{\rm aux}, g_{\rm aux} \rangle \ ,$$

i.e. the approximation is unbiased, and

$$N \, \mathbb{E} |\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle - \langle \eta_{\rm aux}, g_{\rm aux} \rangle|^2 = {\rm var}(g_{\rm aux}, \eta_{\rm aux}) = \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle - \langle \eta_{\rm aux}, g_{\rm aux} \rangle^2 \ .$$

Remark 2.1 Using the identity (1) and the importance decomposition (2) yields

$$\sum_{a=1}^N \int_E g_{\rm aux}^a(x') \, \eta_{\rm aux}^a(dx') = g(x') \, \eta(dx') \ ,$$

an identity for unnormalized distributions, that implies identity for normalizing constants. Indeed, integration of both sides with respect to the variable $x' \in E$ shows that the normalizing constant $\langle \eta_{\rm aux}, g_{\rm aux} \rangle$ associated with the importance decomposition (2) coincides with the normalizing constant $\langle \eta, g \rangle$.

Remark 2.2 Actually, $\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle$ provides also an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, and the variance of the approximation error satisfies

$$N \, \mathbb{E} |\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle - \langle \eta, g \rangle|^2 = \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle - \langle \eta, g \rangle^2 \ ,$$

hence minimizing this expression w.r.t. the auxiliary weights, reduces to minimizing $\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle$ w.r.t. the auxiliary weights. Clearly, the minimum value is smaller than the value obtained for any choice of the auxiliary weights, and in particular is smaller than the valued obtained for the special choice corresponding to the ordinary (non auxiliary) particle filter, hence

$$\min_{\lambda_{\rm aux}} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \leq \langle \eta, g^2 \rangle \qquad \text{and} \qquad \min_{\lambda_{\rm aux}} {\rm var}(g_{\rm aux}, \eta_{\rm aux}) \leq {\rm var}(g, \eta) \ .$$

Lemma 2.3 *Let $x = (x_1, \cdots, x_N)$ be a non–zero vector of nonnegative real numbers. Minimizing the expression*

$$\sum_{a=1}^{N} \frac{x_a^2}{p_a} \ ,$$

over all probability vectors $p = (p_1, \cdots, p_N)$ has a unique solution p proportional to x, and the minimum value is

$$\min_p \sum_{a=1}^{N} \frac{x_a^2}{p_a} = [\sum_{a=1}^{N} x_a\,]^2 \ .$$

Proof. The problem is reformulated as a constrained optimization problem

$$\min_{p_1,\cdots,p_N} \sum_{a=1}^{N} \frac{x_a^2}{p_a} \qquad \text{subject to} \qquad \sum_{a=1}^{N} p_a = 1 \ ,$$

the Lagrange multiplier μ is introduced and the Lagrangian

$$\sum_{a=1}^{N} \frac{x_a^2}{p_a} + \mu\,(\sum_{a=1}^{N} p_a - 1) \ ,$$

is considered. First order optimality condition yields

$$-(\frac{x_a}{p_a})^2 + \mu = 0 \qquad \text{hence} \qquad p_a = c\,x_a \ ,$$

for any $a \in \{1, \cdots, N\}$. The constant c must be such that the constraint is satisfied, i.e.

$$p_a^{\text{opt}} = \frac{x_a}{\displaystyle\sum_{b=1}^{N} x_b} \ ,$$

for any $a \in \{1, \cdots, N\}$. Plugging this expression yields

$$\min_p \sum_{a=1}^{N} \frac{x_a^2}{p_a} = \sum_{a=1}^{N} \frac{x_a^2}{p_a^{\text{opt}}} = [\sum_{a=1}^{N} \frac{x_a^2}{x_a}\,]\,[\sum_{b=1}^{N} x_b\,] = [\sum_{a=1}^{N} x_a\,]^2 \ . \qquad \Box$$

Let

$$u_a = \{\int_E |g_k(x')|^2\ Q_k(\xi_{k-1}^a, dx')\}^{1/2} \ . \tag{3}$$

The optimal design, that minimizes the variance of $\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle$ seen as an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, is defined as follows.

Proposition 2.4 *For any $a \in \{1 \cdots N\}$, the optimal choice for the auxiliary weight is*

$$\lambda_{\text{opt}}^a \propto w_{k-1}^a\, u_a \ ,$$

and the minimum value is

$$\min_{\lambda_{\text{aux}}} \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle = [\sum_{a=1}^{N} w_{k-1}^a\, u_a\,]^2 \ .$$

PROOF. The variance of the approximation error is controlled by

$$\begin{aligned}\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle &= \sum_{a=1}^N \int_E |g_{\rm aux}^a(x')|^2 \, \eta_{\rm aux}^a(dx') \\ &= \sum_{a=1}^N \int_E |\frac{g_k(x') \, w_{k-1}^a}{\lambda_{\rm aux}^a}|^2 \, \lambda_{\rm aux}^a \, Q_k(\xi_{k-1}^a, dx') \\ &= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \int_E |g_k(x')|^2 \, Q_k(\xi_{k-1}^a, dx') \\ &= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \, |u_a|^2 \ ,\end{aligned}$$

and the result follows from Lemma 2.3. □

Remark 2.5 It follows from from the Cauchy–Schwartz inequality that

$$[\sum_{a=1}^N w_{k-1}^a \, u_a]^2 \leq \sum_{a=1}^N w_{k-1}^a \, |u_a|^2 \ ,$$

and note that

$$\sum_{a=1}^N w_{k-1}^a \, |u_a|^2 = \sum_{a=1}^N w_{k-1}^a \int_E |g_k(x')|^2 \, Q_k(\xi_{k-1}^a, dx') = \langle \eta, g^2 \rangle \ .$$

In other words, the optimal design for the auxiliary particle filter improves the performance (reduces the variance) over the classical particle filter.

Under the optimal design, independently for any $i \in \{1, \cdots, N\}$

- the index a_k^i is sampled from the optimal probability vector $(\lambda_{\rm opt}^1, \cdots, \lambda_{\rm opt}^N)$,

then, setting $a = a_k^i$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^a, dx')$,
- and it receives the weight $w_k^i \propto g_{\rm opt}^a(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \, \frac{w_{k-1}^a}{\lambda_{\rm opt}^a} \propto \frac{1}{u_a} \, g_k(\xi_k^i) \ ,$$

see Algorithm 2.

To implement this optimal design, the challenge is to compute

$$u_a = \{ \int_E |g_k(x')|^2 \, Q_k(\xi_{k-1}^a, dx') \}^{1/2} \ ,$$

which was defined in (3), so as to set the optimal probability vector $\lambda_{\rm opt}^a$, for any $a \in \{1, \cdots, N\}$.

Algorithm 2: Particle filter with optimal auxiliary weights

input : $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $a = 1 \cdots N$ **do**
  Optimize the design parameter
  Compute u_a
  Evaluate the unnormalized optimal auxiliary weight
  Set $\lambda_{\rm opt}^a = w_{k-1}^a \, u_a$
end
Normalize the optimal auxiliary weights
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the optimal probability vector $\lambda_{\rm opt} = (\lambda_{\rm opt}^1, \cdots, \lambda_{\rm opt}^N)$
  Propagate the state vector
  Sample $\xi_k^i \sim Q_k(\xi_{k-1}^a, dx')$
  Evaluate the unnormalized weight
  Set $w_k^i = \dfrac{1}{u_a} \, g_k(\xi_k^i)$
end
Normalize the weights

3 Particle filter with an auxiliary transition matrix

Introducing the auxiliary $N \times N$ transition probability matrix $\pi_{\rm aux} = (\pi_{\rm aux}^{i,j})$ from $\{1, \cdots, N\}$ to $\{1, \cdots, N\}$, equivalently seen as a collection indexed by $i \in \{1, \cdots, N\}$ of probability vectors on $\{1, \cdots, N\}$, the idea is to see $\mu(dx')$ as the marginal probability distribution on E of the joint probability distribution

$$\mu_{\rm aux}^{a,c}(dx') \propto g_k(x') \; w_{k-1}^a \; \pi_{\rm aux}^{a,c} \; Q_k(\xi_{k-1}^a, dx') \; ,$$

defined on the augmented space $\{1, \cdots, N\} \times \{1, \cdots, N\} \times E$. Indeed, summation with respect to the indices $a \in \{1, \cdots, N\}$ and $c \in \{1, \cdots, N\}$ yields

$$g_k(x') \; \sum_{a=1}^N w_{k-1}^a \; [\sum_{c=1}^N \pi_{\rm aux}^{a,c}] \; Q_k(\xi_{k-1}^a, dx') = g(x') \; \eta(dx') \; , \tag{4}$$

an identity for unnormalized distributions, that carries over to normalized probability distributions. Assuming that the model transition kernels have a density, i.e. assuming that

$$Q_k(x, dx') = q_k(x, x') \; dx' \; ,$$

makes it possible to define the importance decomposition

$$\mu_{\rm aux}^{a,c}(dx') \propto \underbrace{g_k(x') \; \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')} \; \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}}_{g_{\rm aux}^{a,c}(x')} \; \underbrace{\lambda_{\rm aux}^a \; \pi_{\rm aux}^{a,c} \; Q_k(\xi_{k-1}^c, dx')}_{\eta_{\rm aux}^{a,c}(dx')} \; , \tag{5}$$

the Monte Carlo approximation for the predictor

$$\eta_{\rm aux}^N = \frac{1}{N} \sum_{i=1}^N \delta_{(a_k^i,\, c_k^i,\, \xi_k^i)} \qquad \text{hence} \qquad \langle \eta_{\rm aux}^N, g_{\rm aux} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\rm aux}^{a_k^i, c_k^i}(\xi_k^i) \ ,$$

and the importance sampling approximation for the filter

$$\mu_{\rm aux}^N = \sum_{i=1}^N w_k^i \; \delta_{(a_k^i,\, c_k^i,\, \xi_k^i)} \qquad \text{and its marginal} \qquad \mu^N = \sum_{i=1}^N w_k^i \; \delta_{\xi_k^i} \ ,$$

where independently for any $i = 1, \cdots, N$ the i–th particle (a_k^i, c_k^i, ξ_k^i) is sampled from $\eta_{\rm aux}^{a,c}(dx')$ and is assigned a weight w_k^i using $g_{\rm aux}^{a,c}(x')$. In practice, independently for any $i = 1, \cdots, N$

- the index a_k^i is sampled from the auxiliary probability vector $(\lambda_{\rm aux}^1, \cdots, \lambda_{\rm aux}^N)$,

then, setting $a = a_k^i$

- the index c_k^i is sampled from the probability vector $(\pi_{\rm aux}^{a,1}, \cdots, \pi_{\rm aux}^{a,N})$, the a–th row of the auxiliary transition probability matrix $\pi_{\rm aux}$,

then, setting $c = c_k^i$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^c, dx')$,
- and it receives the weight $w_k^i \propto g_{\rm aux}^{a,c}(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \; \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^c, \xi_k^i)} \; \frac{w_{k-1}^a}{\lambda_{\rm aux}^a} \ ,$$

see Algorithm 3.

This class of particle filters depends upon $(N-1) + N\,(N-1) = N^2 - 1$ design parameters, that are the components of the auxiliary probability vector $\lambda_{\rm aux}$ (a probability vector of dimension N), and the components of the N rows of the auxiliary transition matrix $\pi_{\rm aux}$ (N probability vectors of dimension N each), subject to normalization constraints. This class of particle filters contains the class of auxiliary particle filters — and a fortiori it contains the class of ordinary (non auxiliary) particle filters — as a special case. Indeed

- if for any $a \in \{1, \cdots, N\}$ the a–th row of the auxiliary transition matrix $\pi_{\rm aux}$ satisfies $\pi_{\rm aux}^{a,c} = 1$ if $c = a$ and $\pi_{\rm aux}^{a,c} = 0$ otherwise, then the resulting particle filter belongs to the class of auxiliary particle filters described in Section 2,
- moreover, if $\lambda_{\rm aux}^a = w_{k-1}^a$ for any $a \in \{1, \cdots, N\}$, then it further reduces to the ordinary (non auxiliary) particle filter algorithm.

Algorithm 3: Particle filter with auxiliary transition matrix

input : λ_{aux}, π_{aux}, $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $i = 1 \cdots N$ **do**
- Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\text{aux}} = (\lambda_{\text{aux}}^1, \cdots, \lambda_{\text{aux}}^N)$
- Sample the shuffling index
  Sample c from the a–th row of the auxiliary transition matrix π_{aux}, i.e. from the probability vector $(\pi_{\text{aux}}^{a,1}, \cdots, \pi_{\text{aux}}^{a,N})$
- Propagate the state vector
  Sample $\xi_k^i \sim Q_k(\xi_{k-1}^c, dx')$
- Evaluate the unnormalized weight
  Set $w_k^i = g_k(\xi_k^i)\, \dfrac{w_{k-1}^a}{\lambda_{\text{aux}}^a}\, \dfrac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^c, \xi_k^i)}$

end
Normalize the weights

Recall the particle approximation

$$\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\text{aux}}^{a_k^i, c_k^i}(\xi_k^i) \ ,$$

for the normalizing constant $\langle \eta_{\text{aux}}, g_{\text{aux}} \rangle$, where the random variables (a_k^i, c_k^i, ξ_k^i) for $i = 1, \cdots, N$ are i.i.d. with common probability distribution $\eta_{\text{aux}}^{a,c}(dx')$, hence

$$\mathbb{E}[\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle] = \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle \ ,$$

i.e. the approximation is unbiased, and

$$N\ \mathbb{E}|\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle - \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle|^2 = \text{var}(g_{\text{aux}}, \eta_{\text{aux}}) = \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle - \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle^2 \ .$$

Remark 3.1 Using the identity (4) and the importance decomposition (5) yields

$$\sum_{a=1}^N \sum_{c=1}^N g_{\text{aux}}^{a,c}(x')\, \eta_{\text{aux}}^{a,c}(dx') = g(x')\, \eta(dx') \ ,$$

an identity for unnormalized distributions, that implies identity for normalizing constants. Indeed, integration of both sides with respect to the variable $x' \in E$ shows that the normalizing constant $\langle \eta_{\text{aux}}, g_{\text{aux}} \rangle$ associated with the importance decomposition (5) coincides with the normalizing constant $\langle \eta, g \rangle$.

Remark 3.2 Actually, $\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle$ provides also an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, and the variance of the approximation error satisfies

$$N\ \mathbb{E}|\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle - \langle \eta, g \rangle|^2 = \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle - \langle \eta, g \rangle^2 \ .$$

hence minimizing this expression w.r.t. the auxiliary weights and the auxiliary transition matrix, reduces to minimizing $\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle$ w.r.t. the auxiliary weights and the auxiliary transition matrix. Clearly, the minimum value is smaller than the value obtained for any choice of the auxiliary weights and the auxiliary transition matrix, and in particular is smaller than the value obtained for the special choice corresponding to the auxiliary particle filter or the ordinary (non auxiliary) particle filter, hence

$$\min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \leq \langle \eta, g^2 \rangle \qquad \text{and} \qquad \min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \text{var}(g_{\rm aux}, \eta_{\rm aux}) \leq \text{var}(g, \eta) \ .$$

Let

$$u_a^c = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')} |^2 \, Q_k(\xi_{k-1}^c, dx') \}^{1/2} \ , \tag{6}$$

and note that in the special case where $c = a$, this expression reduces to

$$u_a^a = \{ \int_E |g_k(x')|^2 \, Q_k(\xi_{k-1}^a, dx') \}^{1/2} = u_a \ ,$$

which was defined in (3). The optimal design, that minimizes the variance of $\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle$ seen as an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, is defined as follows.

Proposition 3.3 *For any $a \in \{1 \cdots N\}$, let $u_a^\bullet$ denote the minimum value of u_a^c when $c \in \{1 \cdots N\}$ and assume that there exists a unique minimizer, i.e. a unique index $c_a^\bullet$ for which the minimum value is achieved. Then for any $a \in \{1 \cdots N\}$ the optimal choice for the a–th row of the auxiliary transition matrix is such that*

$$\pi_{\rm opt}^{a,c} = 1 \ \textit{if}\ c = c_a^\bullet \ \textit{and}\ \pi_{\rm opt}^{a,c} = 0 \ \textit{otherwise},$$

the optimal choice for the auxiliary weight is

$$\lambda_{\rm opt}^a \propto w_{k-1}^a \, u_a^\bullet \ ,$$

and the minimum value is

$$\min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle = [\sum_{a=1}^N w_{k-1}^a \, u_a^\bullet]^2 \ .$$

Proof. The variance of the approximation error is controlled by

$$\begin{aligned}
\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle &= \sum_{a=1}^N \sum_{c=1}^N \int_E |g_{\rm aux}^{a,c}(x')|^2 \, \eta_{\rm aux}^{a,c}(dx') \\
&= \sum_{a=1}^N \sum_{c=1}^N \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')} \, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a} |^2 \, \lambda_{\rm aux}^a \, \pi_{\rm aux}^{a,c} \, Q_k(\xi_{k-1}^c, dx') \\
&= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \sum_{c=1}^N \pi_{\rm aux}^{a,c} \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')} |^2 \, Q_k(\xi_{k-1}^c, dx') \\
&= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \sum_{c=1}^N \pi_{\rm aux}^{a,c} \, |u_a^c|^2 \ .
\end{aligned}$$

Under the assumptions, it holds

$$\sum_{c=1}^{N} \pi_{\rm aux}^{a,c} \; |u_a^c|^2 \geq |u_a^\bullet|^2 \; ,$$

and the lower bound is achieved if the a–th row of the auxiliary transition matrix $\pi_{\rm aux}$ charges the minimizer only, i.e. if it satisfies $\pi_{\rm aux}^{a,c} = 1$ if $c = c_a^\bullet$ and $\pi_{\rm aux}^{a,c} = 0$ otherwise. Plugging this expression yields

$$\begin{aligned}
\min_{\pi_{\rm aux}} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \; &= \min_{\pi_{\rm aux}} \sum_{a=1}^{N} \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \; \sum_{c=1}^{N} \pi_{\rm aux}^{a,c} \; |u_a^c|^2 \\
&= \sum_{a=1}^{N} \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \; \sum_{c=1}^{N} \pi_{\rm opt}^{a,c} \, |u_a^c|^2 \\
&= \sum_{a=1}^{N} \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \; |u_a^\bullet|^2 \; ,
\end{aligned}$$

and it follows from Lemma 2.3 that the minimum w.r.t. the auxiliary probability vector is achieved for

$$\lambda_{\rm opt}^a \propto w_{k-1}^a \, u_a^\bullet \; ,$$

and

$$\begin{aligned}
\min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \; \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \; &= \min_{\lambda_{\rm aux}} \; \min_{\pi_{\rm aux}} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \\
&= \min_{\lambda_{\rm aux}} \sum_{a=1}^{N} \frac{|w_{k-1}^a \, u_a^\bullet|^2}{\lambda_{\rm aux}^a} \\
&= [\sum_{a=1}^{N} w_{k-1}^a \, u_a^\bullet \,]^2 \; . \qquad \Box
\end{aligned}$$

Remark 3.4 Recall that $u_a^\bullet$ denotes the minimum value of u_a^c when $c \in \{1, \cdots, N\}$ and in particular $u_a^\bullet \leq u_a^a = u_a$, hence

$$[\sum_{a=1}^{N} w_{k-1}^a \, u_a^\bullet \,]^2 \leq [\sum_{a=1}^{N} w_{k-1}^a \, u_a \,]^2 \; ,$$

for any $a \in \{1, \cdots, N\}$. In other words, the optimal design for the particle filter with an auxiliary transition matrix improves the performance (reduces the variance) over the optimal design for the auxiliary particle filter, and a fortiori it improves the performance over the ordinary (non auxiliary) particle filter.

Under the optimal design, independently for any $i \in \{1, \cdots, N\}$

- the index a_k^i is sampled from the optimal probability vector $(\lambda_{\rm opt}^1, \cdots, \lambda_{\rm opt}^N)$, then, setting $a = a_k^i$ and $c = c_a^\bullet$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^c, dx')$,
- and it receives the weight $w_k^i \propto g_{\rm opt}^{a,c}(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \, \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^c, \xi_k^i)} \, \frac{w_{k-1}^a}{\lambda_{\rm opt}^a} \propto \frac{1}{u_a^\bullet} \, g_k(\xi_k^i) \, \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^c, \xi_k^i)} \; .$$

see Algorithm 4.

Algorithm 4: Particle filter with optimal auxiliary transition matrix

```
input  : (ξ¹_{k-1}, ··· , ξᴺ_{k-1}), (w¹_{k-1}, ··· , wᴺ_{k-1}), Y_k
output: (ξ¹_k, ··· , ξᴺ_k), (w¹_k, ··· , wᴺ_k)
for a = 1 ··· N do
    Optimize the design parameters
    Compute the minimum value u_a^• and the minimizer c_a^•
    Evaluate the unnormalized optimal auxiliary weight
    Set λ^a_opt = w^a_{k-1} u_a^•
end
Normalize the optimal auxiliary weights
for i = 1 ··· N do
    Sample the auxiliary variable
    Sample a from the optimal probability vector λ_opt = (λ¹_opt, ··· , λᴺ_opt)
    Propagate the state vector
    Set c = c_a^• and sample ξⁱ_k ~ Q_k(ξᶜ_{k-1}, dx')
    Evaluate the unnormalized weight
    Set wⁱ_k = (1/u_a^•) g_k(ξⁱ_k) q_k(ξᵃ_{k-1}, ξⁱ_k) / q_k(ξᶜ_{k-1}, ξⁱ_k)
end
Normalize the weights
```

To implement this optimal design, the challenge is to compute

$$u_a^c = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')}|^2 \; Q_k(\xi_{k-1}^c, dx') \}^{1/2} \; ,$$

which was ddefined in (6), for any index $a \in \{1, \cdots, N\}$ and any index $c \in \{1, \cdots, N\}$, so as to compute the minimum value $u_a^\bullet$, find the minimizer $c_a^\bullet$ and set the optimal probability vector $\lambda_{\rm opt}^a$, for any $a \in \{1, \cdots, N\}$. Introduce the mapping

$$u_a(z) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}|^2 \; Q_k(z, dx') \}^{1/2} \; , \tag{7}$$

and note that in the special case where $z = \xi_{k-1}^c$ this expression reduces to

$$u_a(\xi_{k-1}^c) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^c, x')}|^2 \; Q_k(\xi_{k-1}^c, dx') \}^{1/2} = u_a^c \; .$$

Remark 3.5 The problem of minimizing $u_a^c = u_a(\xi_{k-1}^c)$ with respect to the index $c \in \{1, \cdots, N\}$ can only be solved by exhaustive search within a finite (but large) set. If this combinatorial minimization problem would be replaced by the wider problem of minimizing $u_a(z)$ with respect to the variable $z \in E$, as if there would be a continuum of possible choices and not a finite (but large) number N of possible choices, then it would be possible to use numerical optimization procedures instead of exhaustive search.

This motivates the introduction in Section 4 of another (and larger) class of particle filters, with an auxiliary transition kernel.

4 Particle filter with an auxiliary transition kernel

Introducing the auxiliary transition probability kernel $\kappa_{\rm aux} = (\kappa_{\rm aux}^i(dz))$ from $\{1, \cdots, N\}$ to E, equivalently seen as a collection indexed by $i \in \{1, \cdots, N\}$ of probability distributions on E, the idea is to see $\mu(dx')$ as the marginal probability distribution on E of the joint probability distribution

$$\mu_{\rm aux}^a(dz, dx') \propto g_k(x')\; w_{k-1}^a\; \kappa_{\rm aux}^a(dz)\; Q_k(\xi_{k-1}^a, dx') \;,$$

defined on the augmented space $\{1, \cdots, N\} \times E \times E$. Indeed, summation with respect to the index $a \in \{1, \cdots, N\}$ and integration with respect to the variable $z \in E$ yields

$$g_k(x') \sum_{a=1}^N w_{k-1}^a \; [\int_E \kappa_{\rm aux}^a(dz)]\; Q_k(\xi_{k-1}^a, dx') = g(x')\; \eta(dx') \;, \tag{8}$$

an identity for unnormalized distributions, that carries over to normalized probability distributions. Assuming that the model transition kernels have a density, i.e. assuming that

$$Q_k(x, dx') = q_k(x, x')\; dx' \;,$$

makes it possible to define the importance decomposition

$$\mu_{\rm aux}^a(dz, dx') \propto \underbrace{g_k(x')\; \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}\; \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}}_{g_{\rm aux}^a(z, x')}\; \underbrace{\lambda_{\rm aux}^a\; \kappa_{\rm aux}^a(dz)\; Q_k(z, dx')}_{\eta_{\rm aux}^a(dz, dx')} \;, \tag{9}$$

the Monte Carlo approximation for the predictor

$$\eta_{\rm aux}^N = \frac{1}{N} \sum_{i=1}^N \delta_{(a_k^i, \zeta_k^i, \xi_k^i)} \qquad \text{hence} \qquad \langle \eta_{\rm aux}^N, g_{\rm aux} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\rm aux}^{a_k^i}(\zeta_k^i, \xi_k^i) \;,$$

and the importance sampling approximation for the filter

$$\mu_{\rm aux}^N = \sum_{i=1}^N w_k^i\; \delta_{(a_k^i, \zeta_k^i, \xi_k^i)} \qquad \text{and its marginal} \qquad \mu^N = \sum_{i=1}^N w_k^i\; \delta_{\xi_k^i} \;,$$

where independently for any $i = 1, \cdots, N$ the i–th particle $(a_k^i, \zeta_k^i, \xi_k^i)$ is sampled from $\eta_{\rm aux}^a(dz, dx')$ and is assigned a weight w_k^i using $g_{\rm aux}^a(z, x')$. In practice, independently for any $i = 1, \cdots, N$

- the index a_k^i is sampled from the auxiliary probability vector $(\lambda_{\rm aux}^1, \cdots, \lambda_{\rm aux}^N)$,

then, setting $a = a_k^i$

- the particle ζ_k^i is sampled from the probability distribution $\kappa_{\rm aux}^a(dz)$, the a–th row of the auxiliary transition probability kernel $\kappa_{\rm aux}$,

then, setting $z = \zeta_k^i$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(z, dx')$,
- and it receives the weight $w_k^i \propto g_{\rm aux}^a(z, \xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i)\, \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)}\, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a} \;,$$

see Algorithm 5.

Algorithm 5: Particle filter with auxiliary transition kernel

---

input : $\lambda_{\rm aux}$, $\kappa_{\rm aux}$, $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\rm aux} = (\lambda_{\rm aux}^1, \cdots, \lambda_{\rm aux}^N)$
  Sample the shuffled state vector
  Sample z from the a–th 'row' of the auxiliary transition kernel $\kappa_{\rm aux}$, i.e. from the probability distribution $\kappa_{\rm aux}^a(dz)$
  Propagate the state vector
  Sample $\xi_k^i \sim Q_k(z, dx')$
  Evaluate the unnormalized weight
  Set $w_k^i = g_k(\xi_k^i)\, \dfrac{w_{k-1}^a}{\lambda_{\rm aux}^a}\, \dfrac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)}$
end
Normalize the weights

This class of particle filters depends upon several design parameters, that are the components of the auxiliary probability vector $\lambda_{\rm aux}$ (a probability vector of dimension N) subject to normalization constraint, and the N 'rows' of the auxiliary transition kernel $\kappa_{\rm aux}$ (N probability distributions defined on E each). This class of particle filters contains the class of particle filters with an auxiliary transition matrix — and a fortiori it contains the class of auxiliary particle filters and the class of ordinary (non auxiliary) particle filters — as a special case. Indeed

- if for any $a \in \{1, \cdots, N\}$ the a–th 'row' of the auxiliary transition kernel $\kappa_{\rm aux}$ charges the finite set of available particles only, i.e. if

$$\kappa_{\rm aux}^a(dz) = \sum_{c=1}^{N} \pi_{\rm aux}^{a,c}\, \delta_{\xi_{k-1}^c}(dz) \;,$$

as a finite mixture, where the vector of mixture weights is given as the a–th row of some transition matrix $\pi_{\rm aux}$ of dimension $N \times N$, then the resulting particle filter belongs to the class of particle filters with an auxiliary transition matrix described in Section 3,

- moreover, if for any $a \in \{1, \cdots, N\}$ the a–th row of the auxiliary transition matrix $\pi_{\rm aux}$ satisfies $\pi_{\rm aux}^{a,c} = 1$ if $c = a$ and $\pi_{\rm aux}^{a,c} = 0$ otherwise, then the a–th 'row' of the auxiliary transition kernel $\kappa_{\rm aux}$ charges the particle ξ_{k-1}^a only, i.e.

$$\kappa_{\rm aux}^a(dz) = \delta_{\xi_{k-1}^a}(dz) \ ,$$

  and the resulting particle filter belongs to the class of auxiliary particle filters described in Section 2,

- moreover, if $\lambda_{\rm aux}^a = w_{k-1}^a$ for any $a \in \{1, \cdots, N\}$, then it further reduces to the ordinary (non auxiliary) particle filter.

Recall the particle approximation

$$\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\rm aux}^{a_k^i}(\zeta_k^i, \xi_k^i) \ .$$

for the normalizing constant $\langle \eta_{\rm aux}, g_{\rm aux} \rangle$, where the random variables $(a_k^i, \zeta_k^i, \xi_k^i)$ for $i = 1, \cdots, N$ are i.i.d. with common probability distribution $\eta_{\rm aux}^a(dz, dx')$, hence

$$\mathbb{E}[\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle] = \langle \eta_{\rm aux}, g_{\rm aux} \rangle \ ,$$

i.e. the approximation is unbiased, and

$$N \ \mathbb{E}|\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle - \langle \eta_{\rm aux}, g_{\rm aux} \rangle|^2 = {\rm var}(g_{\rm aux}, \eta_{\rm aux}) = \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle - \langle \eta_{\rm aux}, g_{\rm aux} \rangle^2 \ .$$

Remark 4.1 Using the identity (8) and the importance decomposition (9) yields

$$\sum_{a=1}^N \int_E g_{\rm aux}^a(z, x') \, \eta_{\rm aux}^a(dz, dx') = g(x') \, \eta(dx') \ ,$$

an identity for unnormalized distributions, that implies identity for normalizing constants. Indeed, integration of both sides with respect to the variable $x' \in E$ shows that the normalizing constant $\langle \eta_{\rm aux}, g_{\rm aux} \rangle$ associated with the importance decomposition (9) coincides with the normalizing constant $\langle \eta, g \rangle$.

Remark 4.2 Actually, $\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle$ provides also an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, and the variance of the approximation error satisfies

$$N \ \mathbb{E}|\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle - \langle \eta, g \rangle|^2 = \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle - \langle \eta, g \rangle^2 \ ,$$

therefore minimizing this expression w.r.t. the auxiliary weights and the auxiliary transition kernel, reduces to minimizing $\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle$ w.r.t. the auxiliary weights and the auxiliary transition kernel. Clearly, the minimum value is smaller than the value obtained for any choice of the auxiliary weights and the auxiliary transition kernel, and in particular is smaller than the value obtained for the special choice corresponding to the particle filter with an auxiliary transition matrix, the auxiliary particle filter or the ordinary (non auxiliary) particle filter, hence

$$\min_{(\lambda_{\rm aux},\kappa_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \leq \langle \eta, g^2 \rangle \qquad \text{and} \qquad \min_{(\lambda_{\rm aux},\kappa_{\rm aux})} {\rm var}(g_{\rm aux}, \eta_{\rm aux}) \leq {\rm var}(g, \eta) \; .$$

Let

$$u_a(z) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}|^2 \, Q_k(z, dx') \}^{1/2} \; ,$$

which was defined in (7). The optimal design, that minimizes the variance of $\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle$ seen as an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, is defined as follows.

Proposition 4.3 *For any $a \in \{1 \cdots N\}$, let $u_a^{\bullet\bullet}$ denote the minimum value of $u_a(z)$ when $z \in E$ and assume that there exists a unique minimizer, i.e. a unique vector $z_a^{\bullet\bullet}$ for which the minimum value is achieved. Then for any $a \in \{1 \cdots N\}$ the optimal choice for the a–th 'row' of the auxiliary transition kernel is*

$$\kappa_{\rm opt}^a(dz) = \delta_{z_a^{\bullet\bullet}}(dz) \; ,$$

the optimal choice for the auxiliary weight is

$$\lambda_{\rm opt}^a \propto w_{k-1}^a \, u_a^{\bullet\bullet} \; ,$$

and the minimum value is

$$\min_{(\lambda_{\rm aux},\kappa_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle = [\sum_{a=1}^N w_{k-1}^a \, u_a^{\bullet\bullet} \,]^2 \; .$$

Proof. The variance of the approximation error is controlled by

$$\begin{aligned} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle &= \sum_{a=1}^N \int_E \int_E |g_{\rm aux}^a(z, x')|^2 \, \eta_{\rm aux}^a(dz, dx') \\ &= \sum_{a=1}^N \int_E \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')} \, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}|^2 \, \lambda_{\rm aux}^a \, \kappa_{\rm aux}^a(dz) \, Q_k(z, dx') \\ &= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \int_E [\int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}|^2 \, Q_k(z, dx') \,] \, \kappa_{\rm aux}^a(dz) \\ &= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\rm aux}^a} \int_E |u_a(z)|^2 \, \kappa_{\rm aux}^a(dz) \; . \end{aligned}$$

Under the assumptions, it holds

$$\int_E |u_a(z)|^2 \, \kappa_{\rm aux}^a(dz) \geq |u_a^{\bullet\bullet}|^2 \; ,$$

and the lower bound is achieved if the a–th 'row' of the transition probability kernel $\kappa_{\rm aux}$ charges the minimizer only, i.e. if it satisfies $\kappa^a_{\rm aux}(dz) = \delta_{z_a^{\bullet\bullet}}(dz)$. Plugging this expression yields

$$\begin{aligned}\min_{\kappa_{\rm aux}} \langle \eta_{\rm aux}, g^2_{\rm aux} \rangle &= \min_{\kappa_{\rm aux}} \sum_{a=1}^N \frac{|w^a_{k-1}|^2}{\lambda^a_{\rm aux}} \int_E |u_a(z)|^2 \, \kappa^a_{\rm aux}(dz) \\ &= \sum_{a=1}^N \frac{|w^a_{k-1}|^2}{\lambda^a_{\rm aux}} \int_E |u_a(z)|^2 \, \kappa^a_{\rm opt}(dz) \\ &= \sum_{a=1}^N \frac{|w^a_{k-1}|^2}{\lambda^a_{\rm aux}} \, |u_a^{\bullet\bullet}|^2 \; ,\end{aligned}$$

and it follows from Lemma 2.3 that the minimum w.r.t. the auxiliary probability vector is achieved for

$$\lambda^a_{\rm opt} \propto w^a_{k-1} \, u_a^{\bullet\bullet} \; ,$$

and

$$\begin{aligned}\min_{(\lambda_{\rm aux}, \kappa_{\rm aux})} \langle \eta_{\rm aux}, g^2_{\rm aux} \rangle &= \min_{\lambda_{\rm aux}} \min_{\kappa_{\rm aux}} \langle \eta_{\rm aux}, g^2_{\rm aux} \rangle \\ &= \min_{\lambda_{\rm aux}} \sum_{a=1}^N \frac{|w^a_{k-1} \, u_a^{\bullet\bullet}|^2}{\lambda^a_{\rm aux}} \\ &= [\sum_{a=1}^N w^a_{k-1} \, u_a^{\bullet\bullet} \,]^2 \; . \qquad \square\end{aligned}$$

Remark 4.4 Recall that $u_a^{\bullet\bullet}$ denotes the minimum value of $u_a(z)$ when $z \in E$, while $u_a^{\bullet}$ denotes the minimum value of $u_a^c = u_a(\xi^c_{k-1})$ when $c \in \{1, \cdots, N\}$, hence $u_a^{\bullet\bullet} \leq u_a^{\bullet}$ for any $a \in \{1, \cdots, N\}$ and

$$[\sum_{a=1}^N w^a_{k-1} \, u_a^{\bullet\bullet} \,]^2 \leq [\sum_{a=1}^N w^a_{k-1} \, u_a^{\bullet} \,]^2 \; .$$

In other words, the optimal design for the particle filter with an auxiliary transition kernel improves the performance (reduces the variance) over the optimal design for the particle filter with an auxiliary transition matrix, and a fortiori it improves the performance over the optimal design for the auxiliary particle filter and over the ordinary (non auxiliary) particle filter.

Under the optimal design, independently for any $i \in \{1, \cdots, N\}$

- the index a^i_k is sampled from the optimal probability vector $(\lambda^1_{\rm opt}, \cdots, \lambda^N_{\rm opt})$,

then, setting $a = a^i_k$ and $z = z_a^{\bullet\bullet}$

- the particle ξ^i_k is sampled from the probability distribution $Q_k(z, dx')$,

- and it receives the weight $w_k^i \propto g_{\rm opt}^a(z, \xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \, \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)} \, \frac{w_{k-1}^a}{\lambda_{\rm opt}^a} \propto \frac{1}{u_a^{\bullet\bullet}} \, g_k(\xi_k^i) \, \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)} \; ,$$

see Algorithm 6.

Algorithm 6: Particle filter with optimal auxiliary transition kernel

---

input : $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $a = 1 \cdots N$ **do**
  Optimize the design parameters
  Compute the minimum value $u_a^{\bullet\bullet}$ and the minimizer $z_a^{\bullet\bullet}$
  Evaluate the unnormalized optimal auxiliary weights
  Set $\lambda_{\rm opt}^a = w_{k-1}^a u_a^{\bullet\bullet}$
end
Normalize the optimal auxiliary weights
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\rm opt} = (\lambda_{\rm opt}^1, \cdots, \lambda_{\rm opt}^N)$
  Propagate the state vector
  Set $z = z_a^{\bullet\bullet}$ and sample $\xi_k^i \sim Q_k(z, dx')$
  Evaluate the unnormalized weight
  Set $w_k^i = \dfrac{1}{u_a^{\bullet\bullet}} \, g_k(\xi_k^i) \, \dfrac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)}$
end
Normalize the weights

To implement this optimal design, the challenge is to compute

$$u_a(z) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}|^2 \, Q_k(z, dx') \}^{1/2} \; ,$$

which was defined in (7), for any $a \in \{1, \cdots, N\}$ and any $z \in E$, so as to compute the minimum value $u_a^{\bullet\bullet}$, find the minimizer $z_a^{\bullet\bullet}$ and set the optimal weight $\lambda_{\rm opt}^a$, for any $a \in \{1, \cdots, N\}$. As noticed earlier, it is not really necessary to compute $u_a(z)$ for any $z \in E$, it is only necessary to compute the minimum value and find the minimizer, which can be achieved after a few iterations of a numerical optimization procedure. This is illustrated in Section 6 in the special case where the hidden state is modelled as a linear Gaussian system, a special case of practical interest in applications.

5 Application to crossover and to multitarget tracking

It is well known that sequential Monte Carlo methods can be interpreted in terms of implementing *selection* and *mutation* steps, using the language of evolutionary algorithms. However, most general evolutionary algorithms also include a *crossover* step that is not present in sequential Monte Carlo methods. The approach presented in the previous two sections can be adapted to include such a crossover step.

Assume that the state space $E = E_1 \times \cdots \times E_T$ is decomposed as a product space. For instance in multitarget tracking, each different subspace would be the state space for a different target, and the multitarget state variable $x = (x_1, \cdots, x_T)$ would be the concatenation of individual target state variables, with $x_t \in E_t$ the state variable of the t–th target for any $t = 1, \cdots, T$.

The approach presented in Section 3, that uses an auxiliary transition matrix, can be adapted as follows (conceptually, the only difference is that the index c should be replaced by the multi–index $(c_1, \cdots, c_T)$). Introducing the auxiliary $N \times (N \times \cdots \times N)$ transition probability matrix $\pi_{\rm aux} = (\pi_{\rm aux}^{i,(j_1,\cdots,j_T)})$ from $\{1, \cdots, N\}$ to $\{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$, equivalently seen as a collection indexed by $i \in \{1, \cdots, N\}$ of probability vectors on $\{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$, the idea is to see $\mu(dx')$ as the marginal probability distribution on E of the joint probability distribution

$$\mu_{\rm aux}^{a,(c_1,\cdots,c_T)}(dx') \propto g_k(x')\, w_{k-1}^a\, \pi_{\rm aux}^{a,(c_1,\cdots,c_T)}\, Q_k(\xi_{k-1}^a, dx')\ ,$$

defined on the augmented space $\{1, \cdots, N\} \times \{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\} \times E$. Indeed, summation with respect to the index $a \in \{1, \cdots, N\}$ and to the multi–index $(c_1, \cdots, c_T) \in \{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$, yields

$$g_k(x') \sum_{a=1}^N w_{k-1}^a\, [\sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\rm aux}^{a,(c_1,\cdots,c_T)}]\, Q_k(\xi_{k-1}^a, dx') = g(x')\, \eta(dx')\ , \tag{10}$$

an identity for unnormalized distributions, that carries over to normalized probability distributions.

Definition 5.1 For any multi–index $(c_1, \cdots, c_T) \in \{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$, the shuffled multitarget particle $\xi_{k-1}^{(c_1,\cdots,c_T)} = (\xi_{k-1,1}^{c_1}, \cdots, \xi_{k-1,T}^{c_T})$ is obtained by taking its first component from the c_1–th particle, its second component from the c_2–th particle, and so forth up to taking its last component from the c_T–th particle.

Assuming that the model transition kernels have a density, i.e. assuming that

$$Q_k(x, dx') = q_k(x, x')\, dx'\ ,$$

makes it possible to define the importance decomposition

$$\mu_{\rm aux}^{a,(c_1,\cdots,c_T)}(dx') \propto \underbrace{g_k(x')\, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, x')}\, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}}_{g_{\rm aux}^{a,(c_1,\cdots,c_T)}(x')}\ \underbrace{\lambda_{\rm aux}^a\, \pi_{\rm aux}^{a,(c_1,\cdots,c_T)}\, Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')}_{\eta_{\rm aux}^{a,(c_1,\cdots,c_T)}(dx')}\ , \tag{11}$$

the Monte Carlo approximation for the predictor

$$\eta_{\rm aux}^N = \frac{1}{N} \sum_{i=1}^N \delta_{(a_k^i, c_{k,1}^i, \cdots, c_{k,T}^i, \xi_k^i)} \qquad \text{hence} \qquad \langle \eta_{\rm aux}^N, g_{\rm aux} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\rm aux}^{a_k^i,(c_{k,1}^i,\cdots,c_{k,T}^i)}(\xi_k^i)\ ,$$

and the importance sampling approximation for the filter

$$\mu_{\rm aux}^N = \sum_{i=1}^N w_k^i \; \delta_{(a_k^i, c_{k,1}^i, \cdots, c_{k,T}^i, \xi_k^i)} \qquad \text{and its marginal} \qquad \mu^N = \sum_{i=1}^N w_k^i \; \delta_{\xi_k^i} \;,$$

where independently for any $i = 1, \cdots, N$ the i–th particle $(a_k^i, c_{k,1}^i, \cdots, c_{k,T}^i, \xi_k^i)$ is sampled from $\eta_{\rm aux}^{a,(c_1,\cdots,c_T)}(dx')$ and is assigned a weight w_k^i using $g_{\rm aux}^{a,(c_1,\cdots,c_T)}(x')$. In practice, independently for any $i = 1, \cdots, N$

- the index a_k^i is sampled from the auxiliary probability vector $(\lambda_{\rm aux}^1, \cdots, \lambda_{\rm aux}^N)$,

then, setting $a = a_k^i$

- the multi–index $(c_{k,1}^i, \cdots, c_{k,T}^i)$ is sampled from the $N \times \cdots \times N$–dimensional joint probability vector $(\pi_{\rm aux}^{a,(1,\cdots,1)}, \cdots, \pi_{\rm aux}^{a,(N,\cdots,N)})$, the a–th row of the auxiliary transition matrix $\pi_{\rm aux}$,

then, setting $(c_1, \cdots, c_T) = (c_{k,1}^i, \cdots, c_{k,T}^i)$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')$,
- and it receives the weight $w_k^i \propto g_{\rm aux}^{a,(c_1,\cdots,c_T)}(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \; \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, \xi_k^i)} \; \frac{w_{k-1}^a}{\lambda_{\rm aux}^a} \;,$$

see Algorithm 7.

Remark 5.2 With this approach, multitarget particles such as $\xi_{k-1}^a = (\xi_{k-1,1}^a, \cdots, \xi_{k-1,T}^a)$ that are present in the population initially, can be replaced by shuffled multitarget particles such as $\xi_{k-1}^{(c_1,\cdots,c_T)} = (\xi_{k-1,1}^{c_1}, \cdots, \xi_{k-1,T}^{c_T})$. It is the role of the auxiliary transition matrix $\pi_{\rm aux}$ to make this crossover between the target particles possible.

This class of particle filters depends upon $(N-1)+N(N^T-1) = N^{T+1}-1$ design parameters, that are the components of the auxiliary probability vector $\lambda_{\rm aux}$ (a probability vector of dimension N) and the components of the N rows of the auxiliary transition matrix $\pi_{\rm aux}$ (N probability vectors of dimension N^T each), subject to normalization constraints. This class of particle filters contains the class of auxiliary particle filters as a special case. Indeed

- if for any $a \in \{1, \cdots, N\}$, the a–th row of the auxiliary transition matrix $\pi_{\rm aux}$ satisfies $\pi_{\rm aux}^{a,(c_1,\cdots,c_T)} = 1$ if $c_1 = \cdots = c_T = a$ and $\pi_{\rm aux}^{a,(c_1,\cdots,c_T)} = 0$ otherwise, then the resulting particle filter belongs to the class of auxiliary particle filters described in Section 2.

Algorithm 7: Multitarget particle filter with auxiliary transition matrix

input : λ_{aux}, π_{aux}, $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $i = 1$ *to* N **do**
  Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\text{aux}} = (\lambda_{\text{aux}}^1, \cdots, \lambda_{\text{aux}}^N)$
  Sample the shuffling multi-index
  Sample $(c_1, \cdots, c_T)$ from the a–th row of the auxiliary transition matrix π_{aux}, i.e. from the probability vector $(\pi_{\text{aux}}^{a,(1,\cdots,1)}, \cdots, \pi_{\text{aux}}^{a,(N,\cdots,N)})$
  Propagate the multitarget state vector
  Sample $\xi_k^i \sim Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')$
  Evaluate the unnormalized weights
  Set $w_k^i = g_k(\xi_k^i)\, \dfrac{w_{k-1}^a}{\lambda_{\text{aux}}^a}\, \dfrac{q_{k,t}(\xi_{k-1}^a, \xi_k^i)}{q_{k,t}(\xi_{k-1}^{(c_1,\cdots,c_T)}, \xi_k^i)}$
end
Normalize the weights

Recall the particle approximation

$$\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle = \frac{1}{N} \sum_{i=1}^N g_{\text{aux}}^{a_k^i,(c_{k,1}^i,\cdots,c_{k,T}^i)}(\xi_k^i) \ ,$$

for the normalizing constant $\langle \eta_{\text{aux}}, g_{\text{aux}} \rangle$, where the random variables $(a_k^i, c_{k,1}^i, \cdots, c_{k,T}^i, \xi_k^i)$ for $i = 1, \cdots, N$ are i.i.d. with common probability distribution $\eta_{\text{aux}}^{a,(c_1,\cdots,c_T)}(dx')$, hence

$$\mathbb{E}[\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle] = \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle \ ,$$

i.e. the approximation is unbiased, and

$$N\ \mathbb{E}|\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle - \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle|^2 = \text{var}(g_{\text{aux}}, \eta_{\text{aux}}) = \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle - \langle \eta_{\text{aux}}, g_{\text{aux}} \rangle^2 \ .$$

Remark 5.3 Using the identity (10) and the importance decomposition (11) yields

$$\sum_{a=1}^N \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N g_{\text{aux}}^{a,(c_1,\cdots,c_T)}(x')\, \eta_{\text{aux}}^{a,(c_1,\cdots,c_T)}(dx') = g(x')\, \eta(dx') \ ,$$

an identity for unnormalized distributions, that implies identity for normalizing constants. Indeed, integration of both sides with respect to the variable $x' \in E$ shows that the normalizing constant $\langle \eta_{\text{aux}}, g_{\text{aux}} \rangle$ associated with the importance decomposition (11) coincides with the normalizing constant $\langle \eta, g \rangle$.

Remark 5.4 Actually, $\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle$ provides also an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, and the variance of the approximation error satisfies

$$N\ \mathbb{E}|\langle \eta_{\text{aux}}^N, g_{\text{aux}} \rangle - \langle \eta, g \rangle|^2 = \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle - \langle \eta, g \rangle^2 \ ,$$

therefore minimizing this expression w.r.t. the auxiliary weights and the auxiliary transition matrix, reduces to minimizing $\langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle$ w.r.t. the auxiliary weights and the auxiliary transition matrix. Clearly, the minimum value is smaller than the value obtained for any choice of the auxiliary weights and the auxiliary transition matrix, and in particular is smaller than the value obtained for the special choice corresponding to the ordinary (non auxiliary) particle filter, hence

$$\min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle \leq \langle \eta, g^2 \rangle \qquad \text{and} \qquad \min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \mathrm{var}(g_{\rm aux}, \eta_{\rm aux}) \leq \mathrm{var}(g, \eta) \; .$$

Let

$$u_a^{(c_1, \cdots, c_T)} = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, x')}|^2 \, Q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, dx') \}^{1/2} \; , \tag{12}$$

and note that in the special case where $c_1 = \cdots = c_T = a$, this expression reduces to

$$u_a^{(a, \cdots, a)} = \{ \int_E |g_k(x')|^2 \, Q_k(\xi_{k-1}^a, dx') \}^{1/2} = u_a \; ,$$

which was defined in (3). The optimal design, that minimizes the variance of $\langle \eta_{\rm aux}^N, g_{\rm aux} \rangle$ seen as an unbiased approximation of the normalizing constant $\langle \eta, g \rangle$, is defined as follows.

Proposition 5.5 *For any $a \in \{1 \cdots N\}$, let $u_a^\bullet$ denote the minimum value of $u_a^{(c_1, \cdots, c_T)}$ when $(c_1, \cdots, c_T) \in \{1 \cdots N\} \times \cdots \times \{1 \cdots N\}$ and assume that there exists a unique minimizer, i.e. a unique multi–index $(c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$ for which the minimum value is achieved. Then for any $a \in \{1 \cdots N\}$ the optimal choice for the a–th row of the auxiliary transition matrix is such that*

$$\pi_{\rm opt}^{a, (c_1, \cdots, c_T)} = 1 \textit{ if } (c_1, \cdots, c_T) = c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet) \textit{ and } \pi_{\rm opt}^{a, (c_1, \cdots, c_T)} = 0 \textit{ otherwise,}$$

the optimal choice for the auxiliary weight is

$$\lambda_{\rm opt}^a \propto w_{k-1}^a \, u_a^\bullet \; ,$$

and the minimum value is

$$\min_{(\lambda_{\rm aux}, \pi_{\rm aux})} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle = [\sum_{a=1}^N w_{k-1}^a \, u_a^\bullet]^2 \; .$$

Proof. The variance of the approximation error is controlled by

$$\begin{aligned} \langle \eta_{\rm aux}, g_{\rm aux}^2 \rangle &= \sum_{a=1}^N \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \int_E |g_{\rm aux}^{a, (c_1, \cdots, c_T)}(x')|^2 \, \eta_{\rm aux}^{a, (c_1, \cdots, c_T)}(dx') \\ &= \sum_{a=1}^N \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, x')} \, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}|^2 \\ &\qquad\qquad \lambda_{\rm aux}^a \, \pi_{\rm aux}^{a, (c_1, \cdots, c_T)} \, Q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, dx') \end{aligned}$$

$$= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\text{aux}}^a} \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\text{aux}}^{a,(c_1,\cdots,c_T)}$$

$$\int_E |g_k(x') \; \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, x')}|^2 \; Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')$$

$$= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\text{aux}}^a} \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\text{aux}}^{a,(c_1,\cdots,c_T)} \; |u_a^{(c_1,\cdots,c_T)}|^2 \; .$$

Under the assumptions, it holds

$$\sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\text{aux}}^{a,(c_1,\cdots,c_T)} \; |u_a^{(c_1,\cdots,c_T)}|^2 \geq |u_a^\bullet|^2 \; ,$$

and the lower bound is achieved if the a–th row of the $N \times (N \times \cdots \times N)$ transition matrix π_{aux} charges the minimizer only, i.e. if it satisfies $\pi_{\text{aux}}^{a,(c_1,\cdots,c_T)} = 1$ if $(c_1, \cdots, c_T) = (c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$ and $\pi_{\text{aux}}^{a,(c_1,\cdots,c_T)} = 0$ otherwise. Plugging this expression yields

$$\begin{aligned}
\min_{\pi_{\text{aux}}} \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle \; &= \; \min_{\pi_{\text{aux}}} \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\text{aux}}^a} \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\text{aux}}^{a,(c_1,\cdots,c_T)} \; |u_a^{(c_1,\cdots,c_T)}|^2 \\
&= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\text{aux}}^a} \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\text{opt}}^{a,(c_1,\cdots,c_T)} \; |u_a^{(c_1,\cdots,c_T)}|^2 \\
&= \sum_{a=1}^N \frac{|w_{k-1}^a|^2}{\lambda_{\text{aux}}^a} \; |u_a^\bullet|^2 \; ,
\end{aligned}$$

and it follows from Lemma 2.3 that the minimum w.r.t. the auxiliary weights is achieved for

$$\lambda_{\text{opt}}^a \propto w_{k-1}^a \; u_a^\bullet \; ,$$

and

$$\begin{aligned}
\min_{(\lambda_{\text{aux}}, \pi_{\text{aux}})} \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle \; &= \; \min_{\lambda_{\text{aux}}} \min_{\pi_{\text{aux}}} \langle \eta_{\text{aux}}, g_{\text{aux}}^2 \rangle \\
&= \min_{\lambda_{\text{aux}}} \sum_{a=1}^N \frac{|w_{k-1}^a \; u_a^\bullet|^2}{\lambda_{\text{aux}}^a} \\
&= [\sum_{a=1}^N w_{k-1}^a \; u_a^\bullet \,]^2 \; . \qquad \Box
\end{aligned}$$

Remark 5.6 Recall that $u_a^\bullet$ denotes the minimum value of $u_a^{(c_1,\cdots,c_T)}$ when $(c_1, \cdots, c_T) \in \{1 \cdots N\} \times \cdots \times \{1 \cdots N\}$ and in particular $u_a^{(c_1,\cdots,c_T)} \leq u_a^{(a,\cdots,a)} = u_a$, hence

$$[\sum_{a=1}^N w_{k-1}^a \; u_a^\bullet \,]^2 \leq [\sum_{a=1}^N w_{k-1}^a \; u_a \,]^2 \; .$$

In other words, the optimal design using an additional crossover step in the auxiliary particle filter improves the performance (reduces the variance) over the optimal design for the auxiliary particle filter.

Under the optimal design, independently for any $i \in \{1, \cdots, N\}$

- the index a_k^i is sampled from the optimal probability vector $(\lambda_{\text{opt}}^1, \cdots, \lambda_{\text{opt}}^N)$,

then, setting $a = a_k^i$ and $(c_1, \cdots, c_T) = (c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$

- the particle ξ_k^i is sampled from the probability distribution $Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')$,
- and it receives the weight $w_k^i \propto g_{\text{opt}}^{a,(c_1,\cdots,c_T)}(\xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i) \; \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, \xi_k^i)} \; \frac{w_{k-1}^a}{\lambda_{\text{opt}}^a} \propto \frac{1}{u_a^\bullet} \; g_k(\xi_k^i) \; \frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, \xi_k^i)} \; ,$$

see Algorithm 8.

Algorithm 8: Multitarget particle filter with optimal auxiliary transition matrix

---

input : $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $a = 1 \cdots N$ **do**
  Optimize the design parameters
  Compute the minimum value $u_a^\bullet$ and the minimizer $c_a^\bullet = (c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$
  Evaluate the unnormalized optimal weights
  Set $\lambda_{\text{opt}}^a = w_{k-1}^a u_a^\bullet$
end
Normalize the optimal weights
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the optimal probability vector $\lambda_{\text{opt}} = (\lambda_{\text{opt}}^1, \cdots, \lambda_{\text{opt}}^N)$
  Propagate the multitarget state vector
  Set $(c_1, \cdots, c_T) = (c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$ and sample $\xi_k^i \sim Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx')$
  Evaluate the unnormalized weight
  Set $w_k^i = \dfrac{1}{u_a^\bullet} \; g_k(\xi_k^i) \; \dfrac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, \xi_k^i)}$
end
Normalize the weights

To implement this optimal design, the challenge is to compute

$$u_a^{(c_1,\cdots,c_T)} = \{ \int_E |g_k(x') \; \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, x')}|^2 \; Q_k(\xi_{k-1}^{(c_1,\cdots,c_T)}, dx') \}^{1/2} \; ,$$

which was defined in (12), for any index $a \in \{1, \cdots, N\}$ and any multi–index $(c_1, \cdots, c_T) \in \{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$, so as to compute the minimum value $u_a^\bullet$, find the minimizer $(c_{a,1}^\bullet, \cdots, c_{a,T}^\bullet)$ and set the optimal weight $\lambda_{\rm opt}^a$, for any $a \in \{1, \cdots, N\}$. Recall the expression of the mapping

$$u_a(z) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(z, x')}|^2 \; Q_k(z, dx') \}^{1/2} \; ,$$

which was defined in (7) and note that in the special case where $z = \xi_{k-1}^{(c_1, \cdots, c_T)}$ this expression reduces to

$$u_a(\xi_{k-1}^{(c_1, \cdots, c_T)}) = \{ \int_E |g_k(x') \, \frac{q_k(\xi_{k-1}^a, x')}{q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, x')}|^2 \; Q_k(\xi_{k-1}^{(c_1, \cdots, c_T)}, dx') \}^{1/2} = u_a^{(c_1, \cdots, c_T)} \; .$$

Remark 5.7 The problem of minimizing $u_a^{(c_1, \cdots, c_T)} = u_a(\xi_{k-1}^{(c_1, \cdots, c_T)})$ with respect to the multi–index $(c_1, \cdots, c_T) \in \{1, \cdots, N\} \times \cdots \times \{1, \cdots, N\}$ considers all possible shuffled multitarget particles and can only be solved by exhaustive search within a finite (but huge) set. If this combinatorial minimization problem would be replaced by the wider problem of minimizing $u_a(z)$ with respect to the variable $z \in E$, as if there would be a continuum of possible choices and not a finite (but huge) number $N \times \cdots \times N$ (T–times) of possible choices, then it would be possible to use numerical optimization procedures instead of exhaustive search.

The larger class of particle filters already considered in Section 4, with an auxiliary transition kernel, would make this possible and could be implemented as such, virtually without any adaptation.

Remark 5.8 With the approach presented in Section 4, multitarget particles such as $\xi_{k-1}^a = (\xi_{k-1,1}^a, \cdots, \xi_{k-1,T}^a)$ that are present in the population initially, can be replaced by practically *any* multitarget particles, not limited to shuffled multitarget particles such as $\xi_{k-1}^{(c_1, \cdots, c_T)} = (\xi_{k-1,1}^{c_1}, \cdots, \xi_{k-1,T}^{c_T})$. It is the role of the auxiliary transition kernel $\kappa_{\rm aux}$ to make this proposition of new multitarget particles possible.

This class of particle filters depends upon several design parameters, that are the components of the auxiliary probability vector $\lambda_{\rm aux}$ (a probability vector of dimension N) subject to normalization constraint and the N 'rows' of the auxiliary transition kernel $\kappa_{\rm aux}$ (N probability distributions defined on E each). This class contains the class of particle filters defined above, with an auxiliary transition matrix — and a fortiori it contains the class of auxiliary particle filters — as a special case. Indeed

- if for any $a \in \{1, \cdots, N\}$ the a–th 'row' of auxiliary transition kernel $\kappa_{\rm aux}$ charges the finite set of all possible shuffled multitarget particles $\xi_{k-1}^{(c_1, \cdots, c_T)}$ only, i.e. if

$$\kappa_{\rm aux}^a(dz) = \sum_{c_1=1}^N \cdots \sum_{c_T=1}^N \pi_{\rm aux}^{a,(c_1, \cdots, c_T)} \; \delta_{\xi_{k-1}^{(c_1, \cdots, c_T)}}(dz) \; ,$$

  as a finite mixture, where the vector of mixture weights is given as the a–th row of some transition matrix $\pi_{\rm aux}$ of dimension $N \times (N \times \cdots \times N)$, then the resulting particle filter belongs to the class of particle filters with an auxiliary transition matrix described at the beginning of this section.

6 Special case: linear Gaussian model transitions

Explicit calculations, and the proof of the existence of a *unique* minimizer for the mapping $u_a(z)$ defined in (7), are available in the special case where the hidden state is modelled as a linear Gaussian system

$$X_k = F_k\, X_{k-1} + W_k \ ,$$

where W_k is a Gaussian random vector with zero mean and invertible covariance matrix Σ_k, so that the model transition densities are Gaussian densities of the form

$$q_k(x,x') \propto \exp\{-\tfrac{1}{2}\,(x' - F_k\, x)^*\, \Sigma_k^{-1}\, (x' - F_k\, x)\} \ ,$$

for any $x, x' \in E$. For simplicity, the notation $m_a = \xi_{k-1}^a$ is used throughout this and the next sections.

Proposition 6.1 *For any $a \in \{1, \cdots, N\}$, the mapping*

$$u_a(z) = \{\int_E |g_k(x')\, \frac{q_k(m_a,x')}{q_k(z,x')}|^2\; Q_k(z,dx')\}^{1/2}$$

which was defined in (7), can be expressed as $u_a(z) = v_a(F_k\,(m_a - z))$ in terms of the new variable $\theta = F_k\,(m_a - z)$ and the reduced mapping

$$v_a(\theta) = \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\; \{\int_E |g_k(x'' + F_k\, m_a)|^2\; \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma_k^{-1}\,(x'' - \theta)\}\, dx''\}^{1/2} \ , \quad (13)$$

defined up to a multiplicative constant that does not depend on $\theta \in E$ nor on $a \in \{1, \cdots, N\}$.

Proof. For simplicity, write $F_k = F$ and $\Sigma_k = \Sigma$, i.e. the subscript index k is dropped throughout the proof. It follows from (the first part of) Lemma A.1 that

$$\begin{aligned} \frac{q_k(m_a,x')}{q_k(z,x')} &= \frac{\exp\{-\frac{1}{2}\,(x' - F\,m_a)^*\,\Sigma^{-1}\,(x' - F\,m_a)\}}{\exp\{-\frac{1}{2}\,(x' - F\,z)^*\,\Sigma^{-1}\,(x' - F\,z)\}} \\ &= \exp\{(m_a - z)^*\,F^*\,\Sigma^{-1}\,(x' - F\,m_a)\}\; \exp\{\tfrac{1}{2}\,(m_a - z)^*\,F^*\,\Sigma^{-1}\,F\,(m_a - z)\} \ , \end{aligned}$$

hence

$$\begin{aligned} g_{\rm aux}^a(z,x') &= g_k(x')\, \frac{q_k(m_a,x')}{q_k(z,x')}\, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a} \\ &= g_k(x')\, \frac{w_{k-1}^a}{\lambda_{\rm aux}^a}\; \exp\{(m_a - z)^*\,F^*\,\Sigma^{-1}\,(x' - F\,m_a)\} \\ &\qquad\qquad \exp\{\tfrac{1}{2}\,(m_a - z)^*\,F^*\,\Sigma^{-1}\,F\,(m_a - z)\} \ , \end{aligned}$$

for the weight function. It follows from (the second part of) Lemma A.1 that

$$
\begin{aligned}
& | \frac{q_k(m_a, x')}{q_k(z, x')} |^2 \; q_k(z, x') \\
& \quad = | \frac{\exp\{ -\frac{1}{2} \, (x' - F \, m_a)^* \, \Sigma^{-1} \, (x' - F \, m_a) \}}{\exp\{ -\frac{1}{2} \, (x' - F \, z)^* \, \Sigma^{-1} \, (x' - F \, z) \}} |^2 \; \exp\{ -\frac{1}{2} \, (x' - F \, z)^* \, \Sigma^{-1} \, (x' - F \, z) \} \\
& \quad = \exp\{ -\frac{1}{2} \, (x' - F \, (m_a + (m_a - z)))^* \, \Sigma^{-1} \, (x' - F \, (m_a + (m_a - z))) \} \\
& \qquad\qquad \exp\{ (m_a - z)^* \, F^* \, \Sigma^{-1} \, F \, (m_a - z) \} \; ,
\end{aligned}
$$

hence integration with respect to the variable $x' \in E$ yields

$$
\begin{aligned}
u_a(z) \; & = \; \{ \int_E |g_k(x') \; \frac{q_k(m_a, x')}{q_k(z, x')} |^2 \; Q_k(z, dx') \}^{1/2} \\
& = \; \exp\{ \tfrac{1}{2} \, (m_a - z)^* \, F^* \, \Sigma^{-1} \, F \, (m_a - z) \} \; \{ \int_E |g_k(x')|^2 \\
& \qquad \exp\{ -\tfrac{1}{2} \, (x' - F \, (m_a + (m_a - z)))^* \, \Sigma^{-1} \, (x' - F \, (m_a + (m_a - z))) \} \, dx' \}^{1/2} \\
& = \; \exp\{ \tfrac{1}{2} \, (m_a - z)^* \, F^* \, \Sigma^{-1} \, F \, (m_a - z) \} \; \{ \int_E |g_k(x'' + F \, m_a)|^2 \\
& \qquad \exp\{ -\tfrac{1}{2} \, (x'' - F \, (m_a - z))^* \, \Sigma^{-1} \, (x'' - F \, (m_a - z)) \} \, dx'' \}^{1/2} \; ,
\end{aligned}
$$

up to a multiplicative constant that does not depend on $z \in E$ nor on $a \in \{1, \cdots, N\}$. Notice that the model transition density $q_k(z, x')$ depends on z only through $F \, z$ and the two functions $g^a_{\rm aux}(z, x')$ and $u_a(z)$ depends on z only through $F \, (m_a - z)$. Indeed, $u_a(z) = v_a(F \, (m_a - z))$ and in particular $u_a(m_a) = v_a(0)$ with

$$
v_a(\theta) = \exp\{ \tfrac{1}{2} \, \theta^* \, \Sigma^{-1} \, \theta \} \; \{ \int_E |g_k(x'' + F \, m_a)|^2 \; \exp\{ -\tfrac{1}{2} \, (x'' - \theta)^* \, \Sigma^{-1} \, (x'' - \theta) \} \, dx'' \}^{1/2} \; ,
$$

and

$$
\int_E |u_a(z)|^2 \; \kappa^a_{\rm aux}(dz) = \int_E |v_a(F \, (m_a - z))|^2 \; \kappa^a_{\rm aux}(dz) = \int_{{\rm range}(F)} |v_a(\theta)|^2 \; \kappa^a_{\rm aux} \circ F_a^{-1}(d\theta) \; ,
$$

where $\kappa^a_{\rm aux} \circ F_a^{-1}(d\theta)$ denotes the image (or pushforward) probability measure of $\kappa^a_{\rm aux}(dz)$ under the change of variable $\theta = F_a(z) = F \, (m_a - z)$. In other words, this is the probability distribution of the random variable $F_a(Z) = F \, (m_a - Z)$ where the random variable Z has probability distribution $\kappa^a_{\rm aux}(dz)$. □

It follows from Proposition A.2 that the mapping $\theta \mapsto v_a(\theta)$ is strongly log–convex (indeed, $(\log v_a)''(\theta) \geq c \, I$, where the positive constant c does not depend on $\theta \in E$ nor on $a \in \{1, \cdots, N\}$) hence it has a unique minimizer, and so does the mapping $z \mapsto u_a(z) = v_a(F_k \, (m_a - z))$ provided that the square matrix F_k is invertible. However, if the square matrix F_k is not invertible, then

- the mapping $z \mapsto u_a(z) = v_a(F_k\,(m_a - z))$ is still log–convex but not strongly log–convex and it cannot have a unique minimizer (indeed, if z is a minimizer, then $z + z_0$ is another minimizer, for any z_0 in the null space of F_k),
- when restricted to the range space of F_k, the mapping $\theta \mapsto v_a(\theta)$ is strongly log–convex, hence it has a unique minimizer.

The minimum value of $u_a(z)$ when $z \in E$, coincides with the minimum value of $v_a(F_k\,(m_a - z))$ when $z \in E$, hence the minimum value $u_a^{\bullet\bullet}$ can be interpreted as the minimum value of $v_a(\theta)$ subject to $\theta \in \mathrm{range}(F_k)$. Even though the minimizer of $u_a(z)$ when $z \in E$ is not necessarily unique, any minimizer z satisfies $F_k\,(m_a - z) = \theta_a^{\bullet\bullet}$ or equivalently $F_k\,z = F_k\,m_a - \theta_a^{\bullet\bullet}$, where $\theta_a^{\bullet\bullet}$ is the unique minimizer of $v_a(\theta)$ subject to $\theta \in \mathrm{range}(F_k)$. Then, setting $\theta = \theta_a^{\bullet\bullet}$

- the probability distribution $Q_k(z, dx')$ is a Gaussian probability distribution with mean $F_k\,z = F_k\,m_a - \theta$ and covariance matrix Σ_k,
- the particle ξ_k^i sampled from the probability distribution $Q_k(z, dx')$ can be defined as $\xi_k^i = F_k\,m_a - \theta + \Xi_k^i$, where Ξ_k^i is a Gaussian random vector with zero mean and covariance matrix Σ_k, hence $\xi_k^i - F_k\,m_a = -\theta + \Xi_k^i$,

and therefore

$$\begin{aligned}
\frac{q_k(m_a, \xi_k^i)}{q_k(z, \xi_k^i)} &= \exp\{(m_a - z)^*\,F_k^*\,\Sigma_k^{-1}\,(\xi_k^i - F_k\,m_a)\}\ \exp\{\tfrac{1}{2}\,(m_a - z)^*\,F_k^*\,\Sigma_k^{-1}\,F_k\,(m_a - z)\} \\
&= \exp\{\theta^*\,\Sigma_k^{-1}\,(-\theta + \Xi_k^i)\}\ \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta)\} \\
&= \exp\{\theta^*\,\Sigma_k^{-1}\,\Xi_k^i - \tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\ .
\end{aligned}$$

Under the optimal design, independently for any $i \in \{1, \cdots, N\}$

- the index a_k^i is sampled from the optimal probability vector $(\lambda_{\mathrm{opt}}^1, \cdots, \lambda_{\mathrm{opt}}^N)$,

then, setting $a = a_k^i$ and $\theta = \theta_a^{\bullet\bullet}$

- the particle ξ_k^i is defined as $\xi_k^i = F_k\,\xi_{k-1}^a - \theta + \Xi_k^i$ where Ξ_k^i is a Gaussian random vector with zero mean and covariance matrix Σ_k,
- and it receives the weight $w_k^i \propto g_{\mathrm{opt}}^a(z, \xi_k^i)$, i.e.

$$w_k^i \propto g_k(\xi_k^i)\,\frac{q_k(\xi_{k-1}^a, \xi_k^i)}{q_k(z, \xi_k^i)}\,\frac{w_{k-1}^a}{\lambda_{\mathrm{opt}}^a} \propto \frac{1}{u_a^{\bullet\bullet}}\,g_k(\xi_k^i)\ \exp\{\theta^*\,\Sigma_k^{-1}\,\Xi_k^i - \tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\ ,$$

see Algorithm 9.

To implement this optimal design, the challenge is to compute

$$v_a(\theta) = \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\ \{\int_E |g_k(x'' + F_k\,m_a)|^2\ \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma_k^{-1}\,(x'' - \theta)\}\,dx''\}^{1/2}\ ,$$

Algorithm 9: Particle filter with optimal transition kernel (Gaussian model)

input : $(\xi_{k-1}^1, \cdots, \xi_{k-1}^N)$, $(w_{k-1}^1, \cdots, w_{k-1}^N)$, Y_k
output: $(\xi_k^1, \cdots, \xi_k^N)$, $(w_k^1, \cdots, w_k^N)$
for $a = 1 \cdots N$ **do**
  Optimize the design parameters
  Compute the minimum value $u_a^{\bullet\bullet}$ and the minimizer $\theta_a^{\bullet\bullet}$
  Evaluate the unnormalized optimal auxiliary weights
  Set $\lambda_{\rm opt}^a = w_{k-1}^a u_a^{\bullet\bullet}$
end
Normalize the optimal auxiliary weights
for $i = 1 \cdots N$ **do**
  Sample the auxiliary variable
  Sample a from the auxiliary probability vector $\lambda_{\rm opt} = (\lambda_{\rm opt}^1, \cdots, \lambda_{\rm opt}^N)$
  Propagate the state vector
  Set $\theta = \theta_a^{\bullet\bullet}$ and sample $\Xi \sim \mathcal{N}(0, \Sigma_k)$
  Set $\xi_k^i = (F_k\, \xi_{k-1}^a - \theta) + \Xi$
  Evaluate the unnormalized weight
  Set $w_k^i = \dfrac{1}{u_a^{\bullet\bullet}}\, g_k(\xi_k^i)\, \exp\{\theta^*\, \Sigma_k^{-1}\, \Xi - \tfrac{1}{2}\, \theta^*\, \Sigma_k^{-1}\, \theta\}$
end
Normalize the weights

which was defined in (13), for any $a \in \{1, \cdots, N\}$ and any $\theta \in \text{range}(F_k)$, so as to compute the minimum value $u_a^{\bullet\bullet}$, find the minimizer $\theta_a^{\bullet\bullet}$ and set the optimal weight $\lambda_{\rm opt}^a$, for any $a \in \{1, \cdots, N\}$. Finding the unique minimizer of the strongly convex mapping $\theta \mapsto \log v_a(\theta)$ is routinely transformed into finding the unique zero of the mapping $\theta \mapsto (\log v_a)'(\theta)$, and this can be achieved using a stochastic approximation algorithm, see Appendix B.

7 Special case: linear Gaussian system

The purpose of this section is to consider a simple enough special case where it is possible for any $a \in \{1, \cdots, N\}$ to provide an explicit expression for the reduced mapping $v_a(\theta)$, and for the minimum value $u_a^{\bullet\bullet}$ and the unique minimizer $\theta_a^{\bullet\bullet}$. However this special case is of limited practical interest, since the Bayesian filter itself has an explicit expression in terms of the Kalman filter, and there is no need for a Monte Carlo numerical approximation in terms of particle filters.

Assume that the hidden state and the observations are modelled jointly as a linear Gaussian system

$$X_k \;=\; F_k\, X_{k-1} + W_k \;,$$

$$Y_k \;=\; H_k\, X_k + V_k \;,$$

where W_k and V_k are two independent Gaussian random vectors with zero mean and invertible covariance matrices Σ_k and R_k respectively, so that the model transition densities are Gaussian densities of the form

$$q_k(x, x') \propto \exp\{-\tfrac{1}{2}\,(x' - F_k\,x)^*\,\Sigma_k^{-1}\,(x' - F_k\,x)\}\ ,$$

for any $x, x' \in E$, the emission densities are Gaussian densities of the form

$$\exp\{-\tfrac{1}{2}\,(y - H_k\,x')^*\,R_k^{-1}\,(y - H_k\,x')\}\ ,$$

for any $x' \in E$ and $y \in \mathbb{R}^d$, and the likelihood function is defined as

$$g_k(x') = \exp\{-\tfrac{1}{2}\,(Y_k - H_k\,x')^*\,R_k^{-1}\,(Y_k - H_k\,x')\}\ ,$$

for any $x' \in E$. For simplicity, the notations $m_a = \xi_{k-1}^a$ and $I_a = Y_k - H_k\,F_k\,m_a$ for innovation are used throughout this section.

Proposition 7.1 *For any $a \in \{1, \cdots, N\}$, the mapping*

$$v_a(\theta) = \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\ \{\int_E |g_k(x'' + F_k\,m_a)|^2\ \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma_k^{-1}\,(x'' - \theta)\}\,dx''\}^{1/2}\ ,$$

which was defined in (13), has the following explicit expression

$$v_a(\theta)\ =\ \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta\}\ \exp\{-\tfrac{1}{4}\,(I_a - H_k\,\theta)^*\,[H_k\,\Sigma_k\,H_k^* + \tfrac{1}{2}\,R_k]^{-1}\,(I_a - H_k\,\theta)\}\ ,$$

up to a multiplicative constant that does not depend on $\theta \in E$ nor on $a \in \{1, \cdots, N\}$. Moreover, the symmetric matrix

$$A = \Sigma_k^{-1} - \tfrac{1}{2}\,H_k^*\,[H_k\,\Sigma_k\,H_k^* + \tfrac{1}{2}\,R_k]^{-1}\,H_k\ ,$$

associated with the quadratic form $\log v_a(\theta)$, is positive definite.

Equivalently

$$\log v_a(\theta)\ =\ \tfrac{1}{2}\,\theta^*\,\Sigma_k^{-1}\,\theta - \tfrac{1}{4}\,(I_a - H_k\,\theta)^*\,[H_k\,\Sigma_k\,H_k^* + \tfrac{1}{2}\,R_k]^{-1}\,(I_a - H_k\,\theta)\ ,$$

and in the special case where $\theta = 0$, this expression reduces to

$$\log u_a = \log v_a(0) = -\tfrac{1}{4}\,I_a^*\,[H_k\,\Sigma_k\,H_k^* + \tfrac{1}{2}\,R_k]^{-1}\,I_a\ ,$$

up to an additive constant that does not depend on $\theta \in E$ nor on $a \in \{1, \cdots, N\}$.

Proof. For simplicity, write $F_k = F$ and $\Sigma_k = \Sigma$, and also $H_k = H$ and $R_k = R$, i.e. the subscript index k is dropped throughout the proof. It holds

$$\begin{aligned} g_k(x'' + F\,m_a)\ &=\ \exp\{-\tfrac{1}{2}\,(Y_k - H\,(x'' + F\,m_a))^*\,R^{-1}\,(Y_k - H\,(x'' + F\,m_a))\} \\ &=\ \exp\{-\tfrac{1}{2}\,(I_a - H\,x'')^*\,R^{-1}\,(I_a - H\,x'')\}\ , \end{aligned}$$

hence

$$\begin{aligned} &|g_k(x'' + F\,m_a)|^2\ \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\} \\ &\quad = \exp\{-(I_a - H\,x'')^*\,R^{-1}\,(I_a - H\,x'')\}\ \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\} \\ &\quad = \exp\{-\tfrac{1}{2}\,(I_a - H\,x'')^*\,(\tfrac{1}{2}\,R)^{-1}\,(I_a - H\,x'')\}\ \exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\}\ . \end{aligned}$$

Notice that the expression

$$\exp\{-\tfrac{1}{2}\,(y - H\,x'')^*\,(\tfrac{1}{2}\,R)^{-1}\,(y - H\,x'')\}\;\exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\}\;, \tag{14}$$

can be interpreted (up to a normalizing constant) as the joint density of the random vector (X, Y) in the model $Y = H\,X + V$, where X and V are two independent Gaussian random vectors, with mean θ and 0 and with covariance matrix Σ and $\frac{1}{2}\,R$, respectively. Alternatively, this joint density can be factored as the product of the density of the random vector Y, a Gaussian density with mean $H\,\theta$ and with covariance matrix $H\,\Sigma\,H^* + \frac{1}{2}\,R$, and of the conditional density of the random vector X given $Y = y$, a Gaussian density with mean and covariance matrix

$$\widehat{X}(y) = \theta + K\,(y - H\,\theta) \qquad \text{and} \qquad P = (I - K\,H)\,\Sigma\;,$$

respectively, with the Kalman gain matrix

$$K = \Sigma\,H^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\;.$$

Therefore, the expression (14) factors as

$$\begin{aligned}
&\exp\{-\tfrac{1}{2}\,(y - H\,x'')^*\,(\tfrac{1}{2}\,R)^{-1}\,(y - H\,x'')\}\;\exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\}\\
&= \exp\{-\tfrac{1}{2}\,(y - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(y - H\,\theta)\} \qquad (15)\\
&\qquad \exp\{-\tfrac{1}{2}\,(x'' - \widehat{X}(y))^*\,P^{-1}\,(x'' - \widehat{X}(y)\}\;.
\end{aligned}$$

Actually, this identity does not hold up to a multiplicative constant only. Indeed, keeping track of normalizing constants, the normalizing constant for the left–hand side should be

$$\sqrt{\det(2\pi\tfrac{1}{2}R)}\;\sqrt{\det(2\pi\Sigma)}\;,$$

and the normalizing constant for the right–hand side should be

$$\sqrt{\det(2\pi(H\,\Sigma\,H^* + \tfrac{1}{2}\,R))}\;\sqrt{\det(2\pi P)}\;.$$

However, introducing the covariance block–matrix

$$\begin{pmatrix} \Sigma & \Sigma\,H^* \\ H\,\Sigma & H\,\Sigma\,H^* + \tfrac{1}{2}\,R \end{pmatrix}\;,$$

considering that

$$\Sigma - \Sigma\,H^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,H\,\Sigma = P\;,$$

is the Schur complement of $(H\,\Sigma\,H^* + \frac{1}{2}\,R)$ and that

$$(H\,\Sigma\,H^* + \tfrac{1}{2}\,R) - H\,\Sigma\,\Sigma^{-1}\,\Sigma\,H^* = \tfrac{1}{2}\,R\;,$$

is the Schur complement of Σ, and using the Schur determinant formula, yields

$$\det(\tfrac{1}{2}\,R)\;\det\Sigma = \det(H\,\Sigma\,H^* + \tfrac{1}{2}\,R)\;\det P\;.$$

In other words, the missing normalizing constant for the left–hand side is equal to the missing normalizing constant for the right–hand side. The factorization (15) yields

$$|g_k(x'' + F\, m_a)|^2 \;\exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\, \Sigma^{-1}\,(x'' - \theta)\}$$

$$= \exp\{-\tfrac{1}{2}\,(I_a - H\, x'')^*\,(\tfrac{1}{2}\,R)^{-1}\,(I_a - H\, x'')\}\;\exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\}$$

$$= \exp\{-\tfrac{1}{2}\,(I_a - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(I_a - H\,\theta)\}$$

$$\exp\{-\tfrac{1}{2}\,(x'' - \widehat{X}(I_a))^*\,P^{-1}\,(x'' - \widehat{X}(I_a))\}\;.$$

Integration with respect to the variable x'' provides the normalizing constant, and normalization provides the expression for the normalized density $p(\theta, x)$, a Gaussian density with mean vector $\widehat{X}(I_a)$ and with covariance matrix P. Indeed, keeping track of the normalizing constants

$$\int_E |g_k(x'' + F\,m_a)|^2\;\exp\{-\tfrac{1}{2}\,(x'' - \theta)^*\,\Sigma^{-1}\,(x'' - \theta)\}\,\frac{dx''}{\sqrt{\det(2\pi\Sigma)}}$$

$$= \frac{\sqrt{\det(2\pi P)}}{\sqrt{\det(2\pi\Sigma)}}\;\exp\{-\tfrac{1}{2}\,(I_a - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(I_a - H\,\theta)\}$$

$$\int_E \exp\{-\tfrac{1}{2}\,(x'' - \widehat{X}(I_a))^*\,P^{-1}\,(x'' - \widehat{X}(I_a))\}\,\frac{dx''}{\sqrt{\det(2\pi P)}}$$

$$= \frac{\sqrt{\det P}}{\sqrt{\det \Sigma}}\;\exp\{-\tfrac{1}{2}\,(I_a - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(I_a - H\,\theta)\}\;,$$

hence

$$v_a(\theta) \;=\; (\frac{\sqrt{\det P}}{\sqrt{\det \Sigma}})^{1/2}\;\exp\{\tfrac{1}{2}\,\theta^*\,\Sigma^{-1}\,\theta\}\;\exp\{-\tfrac{1}{4}\,(I_a - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(I_a - H\,\theta)\}\;,$$

and

$$\log v_a(\theta) \;=\; \tfrac{1}{2}\,\theta^*\,\Sigma^{-1}\,\theta - \tfrac{1}{4}\,(I_a - H\,\theta)^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,(I_a - H\,\theta) + \tfrac{1}{4}\,\log\frac{\det P}{\det \Sigma}\;.$$

Note that the expression for $\log v_a(\theta)$ is the *difference* of two symmetric quadratic forms, a definite positive quadratic form and a semi–definite positive quadratic form, and this difference could in principle be any kind of a symmetric quadratic form. It follows from Proposition A.2 that the mapping $\theta \mapsto \log v_a(\theta)$ is strongly convex, whatever the emission density could be. This general result can be obtained directly here, i.e. it can be shown that the symmetric matrix

$$A = \Sigma^{-1} - \tfrac{1}{2}\,H^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,H\;,$$

associated with the quadratic form $\log v_a(\theta)$, is positive definite. Indeed, using the matrix inversion lemma yields

$$[\Sigma + \Sigma\,H^*\,(\tfrac{1}{2}\,R)^{-1}\,H\,\Sigma]^{-1} = \Sigma^{-1} - H^*\,[H\,\Sigma\,H^* + \tfrac{1}{2}\,R]^{-1}\,H\;,$$

hence

$$
\begin{aligned}
A &= \Sigma^{-1} - \tfrac{1}{2}\, H^* \, [H\, \Sigma\, H^* + \tfrac{1}{2}\, R]^{-1}\, H \\
&= \tfrac{1}{2}\, \Sigma^{-1} + \tfrac{1}{2}\, \Sigma^{-1} - \tfrac{1}{2}\, H^* \, [H\, \Sigma\, H^* + \tfrac{1}{2}\, R]^{-1}\, H \\
&= \tfrac{1}{2}\, \Sigma^{-1} + \tfrac{1}{2}\, [\Sigma + \Sigma\, H^* \, (\tfrac{1}{2}\, R)^{-1}\, H\, \Sigma]^{-1} \geq \tfrac{1}{2}\, \Sigma^{-1} \; . \qquad \Box
\end{aligned}
$$

Rewriting

$$
\begin{aligned}
\log v_a(\theta) &= \tfrac{1}{2}\, \theta^* \, \Sigma_k^{-1}\, \theta - \tfrac{1}{4}\, (I_a - H_k\, \theta)^* \; [H_k\, \Sigma_k\, H_k^* + \tfrac{1}{2}\, R_k]^{-1}\, (I_a - H\, \theta) + \mathrm{cst} \\
&= \tfrac{1}{2}\, \theta^* \, A\, \theta + b^* \, \theta + e \\
&= \tfrac{1}{2}\, (\theta + A^{-1}\, b)^* \; A\; (\theta + A^{-1}\, b) + e - \tfrac{1}{2}\, b^* \, A^{-1}\, b \; ,
\end{aligned}
$$

where

$$
b = \tfrac{1}{2}\, H_k^* \, [H_k\, \Sigma_k\, H_k^* + \tfrac{1}{2}\, R_k]^{-1}\, I_a \; ,
$$

and

$$
e = -\tfrac{1}{4}\, I_a^* \; [H_k\, \Sigma_k\, H_k^* + \tfrac{1}{2}\, R_k]^{-1}\, I_a + \mathrm{cst} = \log u_a \; ,
$$

it appears that there is an explicit expression for both the minimizer

$$
\theta_a^{\bullet\bullet} = -A^{-1}\, b \; ,
$$

and the minimum value

$$
u_a^{\bullet\bullet} = \exp\{e - \tfrac{1}{2}\, b^* \, A^{-1}\, b\} = u_a \; \exp\{-\tfrac{1}{2}\, b^* \, A^{-1}\, b\} \leq u_a \; .
$$

Note that

References

[1] Yves F. Atchadé, Gersende Fort, and Éric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(10):1–33, 2017.

[2] Arkadii Nemirovskii, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[3] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87 of *Applied Optimization*. Springer–Verlag, Boston, 2004.

[4] Michael K. Pitt and Neil Shephard. Filtering via simulations: auxiliary particle filter. *Journal of the American Statistical Association*, 94(446):590–599, June 1999.

[5] Luis Úbeda-Medina, Ángel F. García-Fernández, and Jesús Grajal. Adaptive auxiliary particle filter for track–before–detect with multiple targets. *IEEE Transactions on Aerospace and Electronic Systems*, AES–53(5):2317–2330, October 2017.

[6] Wei Yi, Mark R. Morelande, Lingjiang Kong, and Jianyu Yang. A computationally efficient particle filter for multitarget tracking using an independence approximation. *IEEE Transactions on Signal Processing*, SP–61(4):843–856, February 2013.

A Gaussian computations

To begin with, here are some results on ratios of two Gaussian densities with different mean vectors m_0 and m, and the same invertible covariance matrix Σ.

Lemma A.1 *Let $\Delta = m_0 - m$ denote the difference of the two mean vectors. Then*

$$\frac{\exp\{-\frac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)\}}{\exp\{-\frac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}} = \exp\{\tfrac{1}{2}\,\Delta^*\,\Sigma^{-1}\,\Delta\}\ \exp\{\Delta^*\,\Sigma^{-1}\,(x-m_0)\}\ ,$$

and

$$|\frac{\exp\{-\frac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)\}}{\exp\{-\frac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}}|^2\ \exp\{-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\} = \exp\{\Delta^*\,\Sigma^{-1}\,\Delta\}\ \exp\{-\tfrac{1}{2}\,(x-(m_0+\Delta))^*\,\Sigma^{-1}\,(x-(m_0+\Delta))\}\ .$$

Proof. Clearly

$$\frac{\exp\{-\frac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)\}}{\exp\{-\frac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}} = \exp\{-\tfrac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)+\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}\ ,$$

and note that

$$\begin{aligned}\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m) &= \tfrac{1}{2}\,(x-m_0+\Delta)^*\,\Sigma^{-1}\,(x-m_0+\Delta)\\ &= \tfrac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)+\Delta^*\,\Sigma^{-1}\,(x-m_0)+\tfrac{1}{2}\,\Delta^*\,\Sigma^{-1}\,\Delta\ ,\end{aligned} \tag{16}$$

which shows the first part. Then

$$|\frac{\exp\{-\frac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)\}}{\exp\{-\frac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}}|^2\ \exp\{-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\} = \exp\{\Delta^*\,\Sigma^{-1}\,\Delta\}\ \exp\{2\,\Delta^*\,\Sigma^{-1}\,(x-m_0)\}\ \exp\{-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}\ ,$$

and using (16) yields

$$-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m) + 2\,\Delta^*\,\Sigma^{-1}\,(x-m_0)$$

$$= -\tfrac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0) + \Delta^*\,\Sigma^{-1}\,(x-m_0) - \tfrac{1}{2}\,\Delta^*\,\Sigma^{-1}\,\Delta$$

$$= -\tfrac{1}{2}\,(x-m_0-\Delta)^*\,\Sigma^{-1}\,(x-m_0-\Delta)\ ,$$

which shows the second part. □

As a by–product, integration with respect to the variable x provides the expression for the χ^2–divergence between two Gaussian densities with different mean vectors m_0 and m, and the same invertible covariance matrix Σ. Indeed

$$\int_E (|\frac{\exp\{-\tfrac{1}{2}\,(x-m_0)^*\,\Sigma^{-1}\,(x-m_0)\}}{\exp\{-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}}|^2 - 1)\ \exp\{-\tfrac{1}{2}\,(x-m)^*\,\Sigma^{-1}\,(x-m)\}\ \frac{dx}{\sqrt{\det(2\pi\Sigma)}}$$

$$= \exp\{\Delta^*\,\Sigma^{-1}\,\Delta\} - 1\ .$$

Note that the point $\mu = m_0 + \Delta = 2m_0 - m$ can be interpreted as the symmetric of the point m with respect to the central point m_0, since (put in other words) the point $m_0 = \tfrac{1}{2}\,(m+\mu)$ is the middle of the segment joining the two points m and μ.

Proposition A.2 *For any function ϕ defined on E, the mapping*

$$\theta \mapsto \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma^{-1}\,\theta\}\ \{\int_E |\phi(x)|^2\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx\}^{1/2}\ ,$$

defined up to a multiplicative constant that does not depend on $\theta \in E$, is strongly log–convex.

Proof. For a given function ϕ defined on E, consider the two mappings

$$v(\theta) = \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma^{-1}\,\theta\}\ \{\int_E |\phi(x)|^2\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx\}^{1/2}\ ,$$

and

$$v_{\rm int}(\theta) = \int_E |\phi(x)|^2\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx\ .$$

The first two derivatives (Jacobian row vector and Hessian symmetric matrix) are defined as

$$v'_{\rm int}(\theta) = \int_E |\phi(x)|^2\ (x-\theta)^*\,\Sigma^{-1}\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx\ ,$$

and

$$v''_{\rm int}(\theta) = \int_E |\phi(x)|^2\ [\Sigma^{-1}\,(x-\theta)\ (x-\theta)^*\,\Sigma^{-1} - \Sigma^{-1}]\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx$$

$$= \int_E |\phi(x)|^2\ \Sigma^{-1}\ (x-\theta)\ (x-\theta)^*\ \Sigma^{-1}\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx - \Sigma^{-1}\ v_{\rm int}(\theta)\ ,$$

respectively. Therefore

$$(\log v_{\rm int})'(\theta) = \frac{v'_{\rm int}(\theta)}{v_{\rm int}(\theta)} = \int_E (x-\theta)^*\,\Sigma^{-1}\,p(\theta,x)\,dx\ ,$$

and

$$\begin{aligned}(\log v_{\rm int})''(\theta) &= \frac{v''_{\rm int}(\theta)}{v_{\rm int}(\theta)} - \left(\frac{v'_{\rm int}(\theta)}{v_{\rm int}(\theta)}\right)^* \frac{v'_{\rm int}(\theta)}{v_{\rm int}(\theta)} \\ &= \int_E \Sigma^{-1}\,(x-\theta)\,(x-\theta)^*\,\Sigma^{-1}\,p(\theta,x)\,dx \\ &\qquad -[\int_E \Sigma^{-1}\,(x-\theta)\,p(\theta,x)\,dx]\,[\int_E (x-\theta)^*\,\Sigma^{-1}\,p(\theta,x)\,dx] - \Sigma^{-1}\ ,\end{aligned}$$

with the probability density

$$p(\theta,x) \propto |\phi(x)|^2\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ ,$$

parametrized by θ and known up to the normalizing constant $v_{\rm int}(\theta)$, i.e.

$$p(\theta,x) = \frac{|\phi(x)|^2\ \exp\{-\frac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}}{\displaystyle\int_E |\phi(x')|^2\ \exp\{-\frac{1}{2}\,(x'-\theta)^*\,\Sigma^{-1}\,(x'-\theta)\}\,dx'}\ .$$

It follows from the Cauchy–Schwartz inequality that

$$(\log v_{\rm int})''(\theta) \geq -\Sigma^{-1} \qquad \text{hence} \qquad (\log v)''(\theta) = \Sigma^{-1} + \tfrac{1}{2}\,(\log v_{\rm int})''(\theta) \geq \tfrac{1}{2}\,\Sigma^{-1} \geq c\,I\ ,$$

in the sense of symmetric matrices (here, the positive constant c is half the smallest eigenvalue of the matrix Σ^{-1}, i.e. half the inverse of the largest eigenvalue $\sigma_{\rm max}$ of the covariance matrix Σ). The Hessian matrix $(\log v)''(\theta)$ is bounded from below by a definite positive matrix that does not depend on θ, hence the mapping $\theta \mapsto v(\theta)$ is strongly log–convex, and this property holds uniformly no matter what the function ϕ appearing in the definition of $v(\theta)$ may be. Notice also that

$$(\log v)'(\theta) = \theta^*\,\Sigma^{-1} + \tfrac{1}{2}\,(\log v_{\rm int})'(\theta) = \tfrac{1}{2}\int_E (x+\theta)^*\,\Sigma^{-1}\,p(\theta,x)\,dx\ . \qquad \Box$$

Lemma A.3 *If the function ϕ is bounded, then the mappings $v_{\rm int}(\theta)$ and $v_{\rm int}^{1/2}(\theta)$ are bounded and globally Lipschitz continuous.*

Proof. Clearly

$$0 \leq v_{\rm int}(\theta) = \int_E |\phi(x)|^2\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ \frac{dx}{\sqrt{\det(2\pi\Sigma)}} \leq \sup_{x\in E}|\phi(x)|^2\ .$$

Recall that, keeping track of the normalizing constant

$$v'_{\rm int}(\theta) = \int_E |\phi(x)|^2\ (x-\theta)^*\,\Sigma^{-1}\ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ \frac{dx}{\sqrt{\det(2\pi\Sigma)}}\ ,$$

hence

$$|v'_{\rm int}(\theta)| \;\leq\; \int_E |\phi(x)|^2 \; |(x-\theta)^* \, \Sigma^{-1}| \; \exp\{-\tfrac{1}{2}\,(x-\theta)^* \, \Sigma^{-1} \, (x-\theta)\} \; \frac{dx}{\sqrt{\det(2\pi\Sigma)}}$$

$$\leq \sup_{x \in E} |\phi(x)|^2 \; \int_E |u^* \, \Sigma^{-1}| \; \exp\{-\tfrac{1}{2}\, u^* \, \Sigma^{-1} \, u\} \; \frac{du}{\sqrt{\det(2\pi\Sigma)}} \;,$$

which shows the first statement. Using the Cauchy–Schwartz inequality yields

$$|v'_{\rm int}(\theta)| \;\leq\; \int_E |\phi(x)|^2 \; |(x-\theta)^* \, \Sigma^{-1}| \; \exp\{-\tfrac{1}{2}\,(x-\theta)^* \, \Sigma^{-1} \, (x-\theta)\} \; \frac{dx}{\sqrt{\det(2\pi\Sigma)}}$$

$$\leq \{\int_E |\phi(x)|^2 \; \exp\{-\tfrac{1}{2}\,(x-\theta)^* \, \Sigma^{-1} \, (x-\theta)\} \; \frac{dx}{\sqrt{\det(2\pi\Sigma)}}\}^{1/2}$$

$$\{\int_E |\phi(x)|^2 \; |(x-\theta)^* \, \Sigma^{-1}|^2 \; \exp\{-\tfrac{1}{2}\,(x-\theta)^* \, \Sigma^{-1} \, (x-\theta)\} \; \frac{dx}{\sqrt{\det(2\pi\Sigma)}}\}^{1/2}$$

$$\leq v^{1/2}_{\rm int}(\theta) \; \sup_{x \in E} |\phi(x)| \; \{\int_E |u^* \, \Sigma^{-1}|^2 \; \exp\{-\tfrac{1}{2}\, u^* \, \Sigma^{-1} \, u\} \; \frac{du}{\sqrt{\det(2\pi\Sigma)}}\}^{1/2} \;,$$

hence

$$|(v^{1/2}_{\rm int})'(\theta)| = \tfrac{1}{2} \, \frac{|v'_{\rm int}(\theta)|}{v^{1/2}_{\rm int}(\theta)} \leq \tfrac{1}{2} \, \sup_{x \in E} |\phi(x)| \; \{\int_E |u^* \, \Sigma^{-1}|^2 \; \exp\{-\tfrac{1}{2}\, u^* \, \Sigma^{-1} \, u\} \; \frac{du}{\sqrt{\det(2\pi\Sigma)}}\}^{1/2} \;,$$

which shows the second statement. □

Remark A.4 Clearly, the mapping $\theta \mapsto \exp\{\frac{1}{2}\,\theta^* \, \Sigma^{-1} \, \theta\}$ is locally bounded and Lipschitz continuous, and it follows from Lemma A.3 that the mapping $v(\theta) = \exp\{\frac{1}{2}\,\theta^* \, \Sigma^{-1} \, \theta\} \; v^{1/2}_{\rm int}(\theta)$ is locally Lipschitz continuous.

In many cases, the state variable x has two components, say $x_{\rm o}$ and $x_{\rm no}$, and the function $\phi(x) = \phi(x_{\rm o})$ depends on the component $x_{\rm o}$ only. In such a case, the minimization problem can be further simplified, i.e. its dimension can be reduced.

Proposition A.5 *Let*

$$x = \begin{pmatrix} x_{\rm o} \\ \\ x_{\rm no} \end{pmatrix} \;, \qquad \theta = \begin{pmatrix} \theta_{\rm o} \\ \theta_{\rm no} \end{pmatrix} \qquad \textit{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{\rm o} & \Sigma_{\rm o,no} \\ \Sigma_{\rm no,o} & \Sigma_{\rm no} \end{pmatrix} \;.$$

If the function $\phi(x) = \phi(x_{\rm o})$ depends on the component $x_{\rm o}$ only, then the unique minimizer of

$$v(\theta) = \exp\{\tfrac{1}{2}\,\theta^* \, \Sigma^{-1} \, \theta\} \; \{\int_E |\phi(x)|^2 \; \exp\{-\tfrac{1}{2}\,(x-\theta)^* \, \Sigma^{-1} \, (x-\theta)\} \; dx\}^{1/2} \;,$$

takes the form $\theta^{\bullet\bullet} = (\theta^{\bullet\bullet}_{\rm o}, \theta^{\bullet\bullet}_{\rm no})$ where $\theta^{\bullet\bullet}_{\rm o}$ denotes the unique minimizer of

$$v_{\rm o}(\theta_{\rm o}) = \exp\{\tfrac{1}{2}\,\theta^*_{\rm o} \, \Sigma^{-1}_{\rm o} \, \theta_{\rm o}\} \; \{\int_{E_{\rm o}} |\phi(x_{\rm o})|^2 \; \exp\{-\tfrac{1}{2}\,(x_{\rm o}-\theta_{\rm o})^* \, \Sigma^{-1}_{\rm o} \, (x_{\rm o}-\theta_{\rm o})\} \; dx_{\rm o}\}^{1/2} \;,$$

and where $\theta^{\bullet\bullet}_{\rm no} = \Sigma_{\rm no,o} \, \Sigma^{-1}_{\rm o} \, \theta^{\bullet\bullet}_{\rm o}$.

PROOF. Clearly

$$\int_E |\phi(x)|^2 \ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx$$

$$= \int_{E_{\rm o}} |\phi(x_{\rm o})|^2 \ \exp\{-\tfrac{1}{2}\,(x_{\rm o}-\theta_{\rm o})^*\,\Sigma_{\rm o}^{-1}\,(x_{\rm o}-\theta_{\rm o})\}\ dx_{\rm o}\ ,$$

up to a multiplicative constant that does not depend on θ, or keeping track of normalizing constants

$$\int_E |\phi(x)|^2 \ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ \frac{dx}{\sqrt{\det(2\pi\Sigma)}}$$

$$= \int_{E_{\rm o}} |\phi(x_{\rm o})|^2 \ \exp\{-\tfrac{1}{2}\,(x_{\rm o}-\theta_{\rm o})^*\,\Sigma_{\rm o}^{-1}\,(x_{\rm o}-\theta_{\rm o})\}\ \frac{dx_{\rm o}}{\sqrt{\det(2\pi\Sigma_{\rm o})}}\ .$$

Introducing the Schur complement $\Delta = \Sigma_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no}$ of the matrix $\Sigma_{\rm o}$ in the block–matrix Σ, it holds

$$\Sigma = \begin{pmatrix} \Sigma_{\rm o} & \Sigma_{\rm o,no} \\ \Sigma_{\rm no,o} & \Sigma_{\rm no} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1} & I \end{pmatrix} \begin{pmatrix} \Sigma_{\rm o} & 0 \\ 0 & \Delta \end{pmatrix} \begin{pmatrix} I & \Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no} \\ 0 & I \end{pmatrix}\ ,$$

hence

$$\Sigma^{-1} = \begin{pmatrix} I & -\Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no} \\ 0 & I \end{pmatrix} \begin{pmatrix} \Sigma_{\rm o}^{-1} & 0 \\ 0 & \Delta^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -\Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1} & I \end{pmatrix}\ ,$$

and

$$\theta^*\,\Sigma^{-1}\,\theta \ = \ (\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})^*\,(\Sigma_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no})^{-1}\,(\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})$$

$$+\,\theta_{\rm o}^*\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o}\ .$$

Therefore

$$v(\theta) \ = \ \exp\{\tfrac{1}{2}\,\theta^*\,\Sigma^{-1}\,\theta\}\ \{\int_E |\phi(x)|^2 \ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\}\ dx\}^{1/2}$$

$$= \ \exp\{\tfrac{1}{2}\,(\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})^*\,(\Sigma_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no})^{-1}\,(\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})\}$$

$$\exp\{\tfrac{1}{2}\,\theta_{\rm o}^*\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o}\}\ \{\int_{E_{\rm o}} |\phi(x_{\rm o})|^2 \ \exp\{-\tfrac{1}{2}\,(x_{\rm o}-\theta_{\rm o})^*\,\Sigma_{\rm o}^{-1}\,(x_{\rm o}-\theta_{\rm o})\}\ dx_{\rm o}\}^{1/2}$$

$$= \ \exp\{\tfrac{1}{2}\,(\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})^*\,(\Sigma_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\Sigma_{\rm o,no})^{-1}\,(\theta_{\rm no} - \Sigma_{\rm no,o}\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o})\}\ v_{\rm o}(\theta_{\rm o})\ ,$$

where

$$v_{\rm o}(\theta_{\rm o}) \ = \ \exp\{\tfrac{1}{2}\,\theta_{\rm o}^*\,\Sigma_{\rm o}^{-1}\,\theta_{\rm o}\}\ \{\int_{E_{\rm o}} |\phi(x_{\rm o})|^2 \ \exp\{-\tfrac{1}{2}\,(x_{\rm o}-\theta_{\rm o})^*\,\Sigma_{\rm o}^{-1}\,(x_{\rm o}-\theta_{\rm o})\}\ dx_{\rm o}\}^{1/2}\ ,$$

up to a multiplicative constant that does not depend on θ. Clearly, the unique minimizer of $v(\theta)$ takes the form $\theta^{\bullet\bullet} = (\theta_{\mathrm{o}}^{\bullet\bullet}, \theta_{\mathrm{no}}^{\bullet\bullet})$ where $\theta_{\mathrm{o}}^{\bullet\bullet}$ denotes the unique minimizer of $v_{\mathrm{o}}(\theta_{\mathrm{o}})$ and where $\theta_{\mathrm{no}}^{\bullet\bullet} = \Sigma_{\mathrm{no,o}}\, \Sigma_{\mathrm{o}}^{-1}\, \theta_{\mathrm{o}}^{\bullet\bullet}$. □

Finding the unique minimizer of the strongly convex mapping $\theta \mapsto \log v(\theta)$ is routinely transformed into finding the unique zero of the mapping $\theta \mapsto (\log v)'(\theta)$, and this can be achieved using a stochastic approximation algorithm. Indeed, recall that

$$(\log v)'(\theta) = \tfrac{1}{2} \int_E (x + \theta)^* \, \Sigma^{-1} \, p(\theta, x) \, dx \ ,$$

hence the (column vector) gradient has the following integral expression

$$\nabla (\log v)(\theta) = \tfrac{1}{2} \, \Sigma^{-1} \int_E (x + \theta) \, p(\theta, x) \, dx \ ,$$

with the probability density

$$p(\theta, x) \propto |\phi(x)|^2 \ \exp\{-\tfrac{1}{2} \, (x - \theta)^* \, \Sigma^{-1} \, (x - \theta)\} \ ,$$

parametrized by θ and known up to the (usually untractable) normalizing constant

$$v_{\mathrm{int}}(\theta) = \int_E |\phi(x)|^2 \ \exp\{-\tfrac{1}{2} \, (x - \theta)^* \, \Sigma^{-1} \, (x - \theta)\} \, dx \ .$$

An iteration of the stochastic gradient algorithm would take the form

$$\theta_{n+1} = \theta_n - \gamma_n \, H_n \ ,$$

where H_n is some random possibly biased approximation of the gradient $\nabla(\log v)(\theta_n)$. This class of stochastic approximation algorithms is studied in Appendix B.

B Stochastic approximation

To summarize the situation, here are some features of the minimization problem presented at the end of Appendix A

- the objective function $\ell(\theta)$ is strongly convex (hence it has a unique minimizer θ_{opt}), but is does not have an integral expression,
- its gradient does have an integral expression of the form

$$\nabla \ell(\theta) = \int_E H(\theta, x) \, p(\theta, x) \, dx \ ,$$

- the probability density $p(\theta, x) = \pi(\theta, x) / Z(\theta)$ depends on the parameter θ, and it is known up to some untractable normalizing constant

$$Z(\theta) = \int_E \pi(\theta, x) \, dx \ .$$

Indeed, the connection with the minimization problem presented at the end of Appendix A is made through the following identification

$$\ell(\theta) = \log v(\theta) \ , \qquad H(x,\theta) = \tfrac{1}{2}\,\Sigma^{-1}\,(x+\theta) \ ,$$

$$\pi(\theta,x) = |\phi(x)|^2 \ \exp\{-\tfrac{1}{2}\,(x-\theta)^*\,\Sigma^{-1}\,(x-\theta)\} \qquad \text{and} \qquad Z(\theta) = v_{\rm int}(\theta) \ .$$

The key step in a stochastic gradient algorithm for this minimization problem is to provide at each iteration of the algorithm an approximation of the gradient $\nabla\ell(\theta_n)$ by some random variable H_n, for instance

A if an *oracle* is available, or if it is possible to sample directly from the current probability density $p(\theta_n,x)$, e.g. using an accept/reject method, then set

$$H_n = \frac{1}{N_n}\,\sum_{j=1}^{N_n} H(\theta_n,X_j) \ ,$$

an unbiased estimator of $\nabla\ell(\theta_n)$, where $(X_1,\cdots,X_{N_n})$ are i.i.d. random variables with common probability density $p(\theta_n,x)$,

B if it is not possible to sample directly from the current probability density $p(\theta_n,x)$, but it is possible to sample from a dominating probability density $\rho(\theta_n,x)$, then introduce the ratio

$$w(\theta_n,x) = \frac{\pi(\theta_n,x)}{\rho(\theta_n,x)} \qquad \text{for any } x \in E,$$

and set

$$H_n = \frac{\displaystyle\sum_{j=1}^{N_n} H(\theta_n,X_j)\; w(\theta_n,X_j)}{\displaystyle\sum_{j=1}^{N_n} w(\theta_n,X_j)} \qquad \text{and} \qquad Z_n = \frac{1}{N_n}\,\sum_{j=1}^{N_n} w(\theta_n,X_j) \ ,$$

a biased estimator of $\nabla\ell(\theta_n)$ and an unbiased estimator of $Z(\theta_n)$ respectively, where $(X_1,\cdots,X_{N_n})$ are i.i.d. random variables with common probability density $\rho(\theta_n,x)$,

C alternatively, if it is not possible to sample directly from the current probability density $p(\theta_n,x)$, but it is possible to sample from a probability transition density $M_n(x,x')$ that admits the current probability density $p(\theta_n,x)$ as its unique invariant density, e.g. using a Metropolis–Hastings algorithm, then set

$$H_n = \frac{1}{N_n}\,\sum_{j=1}^{N_n} H(\theta_n,X_j) \ ,$$

a biased estimator of $\nabla\ell(\theta_n)$, where $(X_1,\cdots,X_{N_n})$ is a Markov chain with some initial probability density and with probability transition density $M_n(x,x')$.

Nemirovskii et al. [2] considers the case A, Atchadé et al. [1] considers the case C, while the focus here is on the case B. The estimator is updated as

$$\theta_{n+1} = \theta_n - \gamma_n\, H_n \ ,$$

hence the design parameters are the gain sequence $\{\gamma_n\}$ and the Monte Carlo batch–size sequence $\{N_n\}$. Write

$$H_n = \nabla\ell(\theta_n) + \varepsilon_n \ ,$$

where ε_n denotes the approximation error, and let

$$B_n = \mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] \qquad \text{and} \qquad M_n = \mathbb{E}[\, |\varepsilon_n|^2 \mid \mathcal{F}_n] \ ,$$

denote the conditional bias and the conditional mean–square error, respectively. If importance sampling (case B) is used, then both the bias B_n and the mean–square error M_n are of order $O(1/N_n)$.

Bound on the gradient approximation Clearly

$$|H_n|^2 = |\nabla\ell(\theta_n)|^2 + 2\, \varepsilon_n^*\, \nabla\ell(\theta_n) + |\varepsilon_n|^2 \ .$$

If the approximation error is unbiased, i.e. if $\mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] = 0$, then

$$\mathbb{E}[\, |H_n|^2 \mid \mathcal{F}_n] = |\nabla\ell(\theta_n)|^2 + M_n \ ,$$

hence

$$\mathbb{E}|H_n|^2 = \mathbb{E}|\nabla\ell(\theta_n)|^2 + \mathbb{E}[M_n] \ .$$

Otherwise, if the approximation error is biased, then using the Young inequality yields

$$\begin{aligned} \mathbb{E}[\, |H_n|^2 \mid \mathcal{F}_n] \ &= \ |\nabla\ell(\theta_n)|^2 + 2\, B_n^*\, \nabla\ell(\theta_n) + M_n \\ &\leq \ (1+A)\, |\nabla\ell(\theta_n)|^2 + \frac{|B_n|^2}{A} + M_n \ , \end{aligned}$$

hence

$$\mathbb{E}|H_n|^2 \leq (1+A)\, \mathbb{E}|\nabla\ell(\theta_n)|^2 + \frac{\mathbb{E}|B_n|^2}{A} + \mathbb{E}[M_n] \ .$$

and taking $A = a/N_n$ yields

$$\mathbb{E}|H_n|^2 \leq (1 + \frac{a}{N_n})\, \mathbb{E}|\nabla\ell(\theta_n)|^2 + N_n\, \frac{\mathbb{E}|B_n|^2}{a} + \mathbb{E}[M_n] \ .$$

Approximation of the minimizer Clearly

$$\theta_{n+1} - \theta_{\rm opt} = \theta_n - \theta_{\rm opt} - \gamma_n\, H_n \qquad \text{with} \qquad H_n = \nabla\ell(\theta_n) + \varepsilon_n \ ,$$

and strong convexity, see Nesterov [3, Theorem 2.1.9], yields

$$(\theta_n - \theta_{\rm opt})^*\, \nabla\ell(\theta_n) = (\theta_n - \theta_{\rm opt})^*\, (\nabla\ell(\theta_n) - \nabla\ell(\theta_{\rm opt})) \geq c\, |\theta_n - \theta_{\rm opt}|^2 \ ,$$

hence

$$
\begin{aligned}
|\theta_{n+1} - \theta_{\rm opt}|^2 \;&=\; |\theta_n - \theta_{\rm opt}|^2 - 2\gamma_n\, (\theta_n - \theta_{\rm opt})^* \, \nabla \ell(\theta_n) \\
&\qquad - 2\gamma_n\, (\theta_n - \theta_{\rm opt})^* \, \varepsilon_n + \gamma_n^2\, |H_n|^2 \\
&\le\; (1 - 2\gamma_n\, c)\; |\theta_n - \theta_{\rm opt}|^2 - 2\gamma_n\, (\theta_n - \theta_{\rm opt})^* \, \varepsilon_n + \gamma_n^2\, |H_n|^2 \; .
\end{aligned}
$$

If the approximation error is unbiased, i.e. if $\mathbb{E}[\varepsilon_n \mid \mathcal{F}_n] = 0$, then

$$
\mathbb{E}[\, |\theta_{n+1} - \theta_{\rm opt}|^2 \mid \mathcal{F}_n] \le (1 - 2\gamma_n\, c)\; |\theta_n - \theta_{\rm opt}|^2 + \gamma_n^2\, \mathbb{E}[\, |H_n|^2 \mid \mathcal{F}_n] \; ,
$$

hence

$$
\mathbb{E}|\theta_{n+1} - \theta_{\rm opt}|^2 \le (1 - 2\gamma_n\, c)\; \mathbb{E}|\theta_n - \theta_{\rm opt}|^2 + \gamma_n^2\, \mathbb{E}|H_n|^2 \; .
$$

This is equation (2.8) in Nemirovskii et al. [2]. Otherwise, if the approximation error is biased, then using the Young inequality yields

$$
\begin{aligned}
\mathbb{E}[\, |\theta_{n+1} - \theta_{\rm opt}|^2 \mid \mathcal{F}_n] \;&\le\; (1 - 2c\, \gamma_n)\; |\theta_n - \theta_{\rm opt}|^2 \\
&\qquad - 2\gamma_n\, (\theta_n - \theta_{\rm opt})^* \, B_n + \gamma_n^2\, \mathbb{E}[\, |H_n|^2 \mid \mathcal{F}_n] \\
&\le\; (1 - (2c - A)\, \gamma_n)\; |\theta_n - \theta_{\rm opt}|^2 \\
&\qquad + \gamma_n\, \frac{|B_n|^2}{A} + \gamma_n^2\, \mathbb{E}[\, |H_n|^2 \mid \mathcal{F}_n] \; ,
\end{aligned}
$$

and taking $A = c$ yields

$$
\mathbb{E}|\theta_{n+1} - \theta_{\rm opt}|^2 \;\le\; (1 - \gamma_n\, c)\; \mathbb{E}|\theta_n - \theta_{\rm opt}|^2 + \gamma_n\, \frac{\mathbb{E}|B_n|^2}{c} + \gamma_n^2\, \mathbb{E}|H_n|^2 \; .
$$

Approximation of the minimum value It would be possible, following Nemirovskii et al. [2] or Atchadé at al. [1], to get bounds for the approximation of the minimum value of the mapping $\ell(\theta) = \log v(\theta)$, i.e. to bound the approximation error $\ell(\theta_n) - \ell(\theta_{\rm opt})$. However, it would still remain to explain how to compute $\ell(\theta_n)$ in practice, and besides, recall that the ultimate objective here is to get bounds for the approximation of the minimum value of the mapping $v(\theta) = \exp\{\theta^* \, \Sigma^{-1}\, \theta\}\; v_{\rm int}^{1/2}(\theta)$. Clearly

$$
\exp\{\tfrac{1}{2}\, \theta_n^* \, \Sigma^{-1}\, \theta_n\}\; Z_n^{1/2} - v(\theta_{\rm opt}) = \exp\{\tfrac{1}{2}\, \theta_n^* \, \Sigma^{-1}\, \theta_n\}\; (Z_n^{1/2} - v_{\rm int}^{1/2}(\theta_n)) + v(\theta_n) - v(\theta_{\rm opt}) \; ,
$$

and using the bound $(\sqrt{z} - \sqrt{v})^2 \le |z - v|$ yields

$$
\begin{aligned}
|\exp\{\tfrac{1}{2}\, \theta_n^* \, \Sigma^{-1}\, \theta_n\}\; Z_n^{1/2} - v(\theta_{\rm opt})|^2 \;&\le\; 2\, \exp\{\theta_n^* \, \Sigma^{-1}\, \theta_n\}\; |Z_n^{1/2} - v_{\rm int}^{1/2}(\theta_n)|^2 + 2\, |v(\theta_n) - v(\theta_{\rm opt})|^2 \\
&\le\; 2\, K^2\; |Z_n - v_{\rm int}(\theta_n)| + 2\, L^2\; |\theta_n - \theta_{\rm opt}|^2 \; ,
\end{aligned}
$$

hence

$$
\mathbb{E}|\exp\{\tfrac{1}{2}\, \theta_n^* \, \Sigma^{-1}\, \theta_n\}\; Z_n^{1/2} - v(\theta_{\rm opt})|^2 \;\le\; 2\, K^2\; \{\, \mathbb{E}|Z_n - v_{\rm int}(\theta_n)|^2\, \}^{1/2} + L^2\; \mathbb{E}|\theta_n - \theta_{\rm opt}|^2 \; .
$$