Campagne 2012-2015
Unité de recherche : dossier unique

Fiche équipe : ASPI

1) General presentation

The ASPI team addresses the design, practical implementation and mathematical analysis of interacting Monte Carlo methods, also known as particle methods or sequential Monte Carlo methods. These are sequential simulation methods, in which particles *explore* the state space by mimicking the evolution of some underlying random process, *learn* the environment by evaluating a fitness function, and *interact* so that only the most successful particles (in view of the value of the fitness function) are allowed to survive and to get offsprings at the next generation. This includes (i) statistical inference in hidden Markov models, using particle filters, and (ii) rare event simulation and global optimization using multi-level branching methods. In general, particle methods provide approximation of Feynman-Kac distributions, a pathwise generalization of Gibbs-Boltzmann distributions. The currently addressed applications are in localization, navigation and tracking (indoor localization, terrain-aided navigation), in evaluation of small probabilities (collision risk in aerospace, false alarm probabilities in protection of digital content), and in optimization of position and activation of sensors.

2) Team composition (01/01/2010)

- Frédéric Cérou (CR INRIA), Arnaud Guyader (MCF université de Rennes 2), François Le Gland (DR INRIA)
- 7 PhD students : Sébastien Beyou, Adrien Ickowicz, Nordine El Baraka (CIFRE Thalès), Rudy Pastel (ONERA), Paul Bui Quang (ONERA), Francis Céleste (DGA/CEP), Renaud Cariou (DGA/CELAR)
- 1 post-doc : Anindya Goswami

3) Summary of research activities (2006 – mid 2010)

- ### Major research results and impact

The following objectives have been listed in the proposal submitted in March 2004:
- mathematical analysis of particle methods,
- simulation of rare events,
- particle filtering, with applications in
  - localization, navigation and tracking,
  - sequential data assimilation.
An additional topic, that was not present in the initial list of objectives, has emerged during the evaluation period:
- classification in infinite dimension.

Mathematical analysis of particle methods

ASPI is essentially concerned with methodological issues, i.e. with the mathematical analysis of popular algorithmic variants and implementations of interacting Monte Carlo methods, some of which have been proposed and used in different communities, to solve different problems.

- **Nonasymptotic variance.** This is collaboration with Pierre Del Moral (EPI ALEA, INRIA Bordeaux --- Sud Ouest). A nonasymptotic theorem for interacting particle approximations of unnormalized Feynman-Kac distributions has been obtained, using recently developed coalescent tree-based functional representations of particle block distributions. Regularity conditions have been introduced under which the L^2 relative error of these weighted particle distributions grows linearly with respect to the time horizon.

- **Adaptive sample size and non-extinction.** This is collaboration with Nadia Oudjane (EDF RD, department OSIRIS). In the special case of a Feynman-Kac distribution, where the selection functions can possibly take the zero value, it can happen that the evaluation of the selection function returns the zero value for all the particles generated at the end of the mutation step, i.e. the particle systems dies out and the algorithm cannot go on. A sequential particle algorithm, already proposed in the PhD thesis of Nadia Oudjane, has been further studied, and a central limit theorem has been obtained, with two alternative proofs. By construction, this sequential algorithm automatically keeps the particle system alive, i.e. it ensures its non-extinction, and it can be seen as a *fixed performance* policy, as opposed to the usual nonsequential algorithm which is a *fixed effort* policy.

- **Combining weighting and resampling.** A particle approximation has been studied, that combines SIS and SIR algorithms in the sense that only a fraction of the importance weights is used for resampling, and two different approaches have been proposed to analyze its performance. The starting point is a factorization of the potential functions as the product of an importance weight function and a resampling weight function. In the first approach, based on a representation in path-space, the importance weight functions appear only in the test function, whereas in the second approach, based on a representation in terms of a multiplicative functional, importance weights that are not used for resampling are treated as particles.

- **Adaptive resampling.** This is collaboration with Élise Arnaud (université Joseph Fourier and EPI PERCEPTION, INRIA Grenoble --- Rhône Alpes). The first contribution has been to consider a design where resampling is performed at some intermediate fixed time instants, and to optimize the asymptotic variance of the estimation error w.r.t. the resampling time instants. The second contribution has been to prove a central limit theorem for particle methods with adaptive resampling, using an interpretation of particle methods where importance weights are treated as particles, as long as they are not used for resampling purpose, and to analyze the asymptotic variance in terms of the threshold.

- **Non-parametric learning of the optimal importance distribution.** This is collaboration with Nadia Oudjane (EDF RD, department OSIRIS). From the weighted empirical probability distribution associated with a first sample, a regularized probability distribution can be obtained, using a kernel method or an histogram, and can be used as an almost optimal importance distribution to estimate the original integral. The variance of the resulting estimator depends on the product of the inverse sample size by the chi-square distance between the almost optimal importance distribution and the optimal (zero variance) importance distribution. The contribution has been to provide an estimate of this chi-square distance, under mild assumptions.

Simulation of rare events

During the evaluation period, a major application area of rare event simulation has been the reliability of protection mechanisms for digital documents, i.e. the evaluation of (small) probabilities of false alarm in *watermarking* or *fingerprinting* schemes. This activity has started on the occasion of the ANR project NEBBIANO, within the SETIN (sécurité et informatique) programme.

- **Design and optimization of Tardos probabilistic fingerprinting codes.** This is collaboration with Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). Gábor Tardos was the first to give a construction of a fingerprinting code whose length meets the lowest known bound. This was a real breakthrough because the construction is very simple. Its efficiency comes from its probabilistic nature. However, although Tardos almost gave no argument of his rationale, many parameters of his code are precisely fine-tuned. We have proposed this missing rationale supporting the code construction. The key idea is to make the statistics of the scores as independent as possible from the collusion process. Tardos optimal parameters are rediscovered. This interpretation allows small improvements when some assumptions hold on the collusion process.

- **Estimating the minimal length of Tardos code.** This is collaboration with Luis Pérez-Feire from GRADIANT (Galician R&D Center in Advanced Telecommunications, Vigo, Spain) and with Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). We are interested in the minimal length of a binary probabilistic traitor tracing code. We consider the code construction proposed by Gábor Tardos in 2003, with the symmetric accusation function as improved by Boris Skoric et al. The length estimation is based on two pillars. First, we consider the worst-case attack that a group of colluders can lead. This attack minimizes the mutual information between the code sequence of a colluder and the pirated sequence. Second, an algorithm pertaining to the field of rare event analysis is constructed in order to estimate the probabilities of error: the probability that an innocent user is framed, and the probabilities that all colluders are missed. Therefore, for a given collusion size, we are able to estimate the minimal length of the code satisfying some error probability constraints. This estimation is far lower than the known lower bounds.

- **Experimental assessment of reliability for watermarking and fingerprinting.** This is collaboration with Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). The concept of reliability in watermarking has been introduced as the ability of assessing that a probability of false alarm is very low and below a given significance level. We propose an iterative and adaptive algorithm, which estimates very low probabilities of error. It performs much quicker and more accurately than a classical Monte Carlo estimator. The article finishes with applications to zero-bit watermarking (probability of false alarm, error exponent), and to probabilistic fingerprinting codes (probability of wrongly accusing a given user, code length estimation). Prior to publication, a patent has been jointly submitted in August~2008 by INRIA and Bretagne Valorisation.

- **Estimating the probability of false alarm for zero-bit watermarking.** This again is collaboration with Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). Assessing that a probability of false alarm is below a given significance level is a crucial issue in watermarking. We have proposed an iterative and adaptive algorithm, which estimates very low probabilities of error. Some experimental investigations validate its performance for a rare detection scenario where there exists a close form formula of the probability of false alarm. Our algorithm appears to be much quicker and more accurate than a classical Monte Carlo estimator. It even allows the experimental measurement of error exponents.

The remaining part is devoted to presenting scientific achievements in the methodology of splitting methods.

- **Characterization of optimal importance function and optimal threshold.** The *multi-level splitting method* consists in (i) introducing a decreasing sequence of intermediate, more and more critical, regions in the state space, (ii) counting the fraction of trajectories that reach an intermediate region before final time, given that the previous intermediate region has been reached before final time, and (iii) regenerating the population at each stage, through redistribution. In addition to their non-intrusive behavior, the splitting methods make it possible to learn the probability distribution of typical critical trajectories, which reach the critical region before final time, an important feature that methods based on importance sampling usually miss. Many variants have been proposed, whether
  - the branching rate (number of offsprings allocated to a successful trajectory) is fixed, which allows for depth-first exploration of the branching tree, but raises the issue of controlling the population size,
  - the population size is fixed, which requires a breadth-first exploration of the branching tree, with random (multinomial) or deterministic allocation of offsprings, etc.

  In this way, the algorithm learns
  - the transition probability between successive levels, hence the probability of reaching each intermediate level,
  - and the entrance probability distribution of the Markov process in each intermediate region.

  Our contribution has been to study the impact of the shape of the intermediate regions (selection of the importance function), of the thresholds (levels), of the population size, on the asymptotic variance, obtained through a central limit theorem. For instance, assuming that the optimal importance function is used, the levels should be selected in such a way that the transition probability to reach the next level before final time, given that the current level has been reached before final time, does not depend on the level.

- **Adaptive threshold selection.** Besides importance sampling, a widespread technique to estimate rare event probabilities is multi-level splitting, which requires at least some knowledge of the system, to decide where to place the intermediate level sets. This is not always possible, and an adaptive algorithm has been proposed to cope with this problem, and has been analyzed thoroughly in the one-dimensional case. Assuming that the problem is to estimate the probability P that a one-dimensional Markov process reaches some critical level before returning to the origin, and taking advantage of the findings about optimal threshold selection, the adaptive algorithm fixes beforehand the probability q to reach the next level before returning to the origin. At each stage, N excursions are simulated that try to approach the final (and critical) level, and the $k = \lfloor qN \rfloor$ excursions approaching the final level most closely are selected, which implicitly defines the next level, and are given offsprings so that the population size N remains constant. The algorithm continues until generation L where the current (virtual) level exceeds the final level, i.e. L denotes the first generation where at least one excursion reaches the final level, and let k_{L+1} denote the exact number of such excursions. Then, the rare event probability is estimated as

$$P \approx \frac{k_{L+1}}{N} \left(\frac{k}{N} \right)^L$$

  and the number of (virtual) intermediate levels is L. In this one-dimensional framework, almost sure convergence and asymptotic normality of the estimator has been proved, with the same variance as other algorithms that use given intermediate levels. It has also been showed on numerical examples that this method can be used for multidimensional problems as well, even if there is still no convergence result in this case.

- **Rare event simulation for a static distribution.** This is collaboration with Pierre Del Moral (EPI ALEA, INRIA Bordeaux --- Sud Ouest) and with Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). The key issue is to learn as fast as possible regions of the input space which contribute most to the computation of the target quantity. The proposed splitting methods consists in (i) introducing a sequence of intermediate regions in the input space, implicitly defined by exceeding an increasing sequence of thresholds or levels, (ii) counting the fraction of samples that reach a level given that the previous level has been reached already, and (iii) improving the diversity of the selected samples, usually using an artificial Markovian dynamics. In this way, the algorithm learns
  - the transition probability between successive levels, hence the probability of reaching each intermediate level,
  - and the probability distribution of the input random variable, conditioned on the output variable reaching each intermediate level.

  Our contribution has been to study the impact of design parameters, such as the intermediate levels or the Metropolis kernel introduced in the mutation step, on the asymptotic variance obtained through a central limit theorem. This work has been recently extended, in collaboration with Nicolas Hengartner (LANL) and Éric Matzner-Løber (université de Rennes 2).

Particle filtering

During the evaluation period, several applications in localization, navigation and tracking have been studied
- information fusion for indoor localization,
- geolocalization and tracking in a wireless network,
- terrain-aided navigation,

which are described in the following section on contracts and technology transfer.

The remaining part is devoted to presenting scientific achievements in sequential data assimilation, where the favorite application domain is oceanology and meteorology. The popular method here is the ensemble Kalman filter, in which ensemble elements (i) are propagated independently according to some underlying physical model, (ii) all contribute to the evaluation of an empirical covariance matrix, and (iii) are updated when a new observation occurs, using a Kalman-like step in which the prediction covariance matrix is replaced by the empirical covariance matrix.

- **Asymptotics of the ensemble Kalman filter (EnKF).** This is collaboration with Valérie Monbet (université de Bretagne Sud). Very little was known about the asymptotic behavior of the ensemble Kalman filter, whereas on the other hand, the asymptotic behavior of many different classes of particle filters is well understood, as the number of particles goes to infinity. Interpreting the ensemble elements as a population of particles with mean-field interactions, and not only as an instrumental device producing an estimation of the hidden state as the ensemble mean value, it has been possible to prove the convergence of the ensemble Kalman filter, with a rate of order $1/\sqrt{N}$, as the number N of ensemble elements increases to infinity. In addition, the limit of the empirical distribution of the ensemble elements has been exhibited, which differs from the usual Bayesian filter. Several cases have been investigated, from the simple case where the drift coefficient is bounded and globally Lipschitz continuous, to the more realistic case where the drift coefficient is locally Lipschitz continuous, with polynomial growth. In all these cases, the observation coefficient was assumed linear, so that the analysis step for each ensemble element has exactly the same structure as the analysis step of the usual Kalman filter. The next step is to study the asymptotic normality of the estimation error, i.e. to prove a central limit theorem. It is somehow expected that the asymptotic variance for the ensemble Kalman filter would be smaller than the known asymptotic variance for the different brands of particle filters, just because the ensemble Kalman filter follows essentially a parametric approach, where only the first two empirical moments are propagated, whereas the particle filters follows a fully nonparametric approach.

- **Particle filter for mean-field models.** This is collaboration with Christophe Baehr (Météo-France, centre national de recherche météorologique). The motivating application is the estimation of Lagrangian velocity in a turbulent flow: to filter out the observation noise a Bayesian approach is used, with a simplified Langevin model proposed by Stephen Pope as the prior for the Lagrangian velocity. This model involves local means of the Eulerian velocity field, which can be expressed in terms of the probability distribution of the Lagrangian velocity. Other nonlinear terms in the model, such as the mean pressure gradient, the turbulent kinetic energy and its dissipation rate are either considered as unknown random variables, with a somehow arbitrary prior probability distribution, or are related with local Eulerian means and can then be expressed in terms of the probability distribution of the Lagrangian velocity. In other words, the proposed simplified Langevin model is a special example of a nonlinear McKean model with mean-field interactions, where the drift coefficient depends on the probability distribution of the solution. The original estimation problem reduces to the estimation of the hidden state in a nonlinear Markov model, and numerical approximations have been studied, with two populations of particles: the first population of particles with mean-field interactions learns the unconditional probability distribution of the hidden state, whereas the second population of particles approximates the Bayesian filter, i.e. the conditional probability distribution of the hidden state given the observations. Alternatively, since noisy observations are available, the local Eulerian means can be expressed in terms of the conditional probability distribution of the Lagrangian velocity given the observations. This results in a much simpler model, where a single population of particles is sufficient to approximate the Bayesian filter.

Classification in infinite dimension

This additional topic was not present in the initial list of objectives, and has emerged during the evaluation period.

In pattern recognition and statistical learning, also known as machine learning, nearest neighbor (NN) algorithms are amongst the simplest but also very powerful available algorithms. Basically, given a training set of data, i.e. an N-sample of i.i.d. object-feature pairs (X_i, Y_i) for $i = 1, \cdots, N$, with real-valued features, the question is how to generalize, that is how to guess the feature Y associated with a new object X, with the same probability distribution as the X_i's. To achieve this, one chooses some integer k smaller than N, and takes the mean-value of the k features associated with the k objects that are nearest to the new object X, for some given metric. The asymptotic behavior when the sample size grows is well understood in finite dimension, but the situation is radically different in general infinite dimensional spaces, when the objects to be classified are functions, images, etc.

- **Nearest neighbor classification in infinite dimension.** In finite dimension, the k-nearest neighbor classifier g_N is universally consistent, i.e. its probability of error converges to the Bayes risk as N goes to infinity,

whatever the joint probability distribution of (X,Y), provided that the ratio k/N goes to zero. Unfortunately, this result is no longer valid in general metric spaces, and the objective is to find out reasonable sufficient conditions for the weak consistency to hold. Even in finite dimension, there are exotic distances such that the nearest neighbor does not even get closer (in the sense of the distance) to the point of interest, and the state space E needs to be complete for the metric d, which is the first condition. Some regularity on the regression function is required next. Clearly, continuity is too strong because it is not required in finite dimension, and a weaker form of regularity is assumed. The following consistency result has been obtained: if the metric space (E,d) is separable and if some *Besicovich condition* holds, then the nearest neighbor classifier is weakly consistent, i.e. $P[g_N \neq Y] \to L$, as $N \uparrow \infty$, where L denotes the Bayes probability of error, or Bayes risk. Note that the Besicovich condition is always fulfilled in finite dimensional vector spaces (this result is called the Besicovich theorem), and that a counterexample can be given in an infinite dimensional space with a Gaussian measure (in this case, the nearest neighbor classifier is clearly nonconsistent). Finally, a simple example has been found which verifies the Besicovich condition with a noncontinuous regression function.

- **Rates of convergence of the functional k-nearest neighbor estimator.** This is collaboration with Gérard Biau (université Pierre et Marie Curie, ENS Paris and EPI CLASSIC, INRIA Paris --- Rocquencourt). Motivated by a broad range of potential applications, such as regression on curves, rates of convergence of the k-nearest neighbor estimator of the regression function, based on N independent copies of the pair (X,Y), have been investigated when X is in a suitable ball in some functional space. Using compact embedding theory, we present explicit and general finite sample bounds on the expected squared difference between k-nearest neighbor estimator and Bayes regression function, in a very general setting. The results have also been particularized to classical function spaces such as Sobolev spaces, Besov spaces and reproducing kernel Hilbert spaces. The rates obtained are genuine nonparametric convergence rates, and up to our knowledge the first of their kind for k-nearest neighbor regression.

- **Rate of convergence of the bagged nearest neighbor estimator.** This is collaboration with Gérard Biau (université Pierre et Marie Curie, ENS Paris and EPI CLASSIC, INRIA Paris --- Rocquencourt). Bagging is a simple way to combine estimators in order to improve their performance. This method, suggested by Leo Breiman in 1996, proceeds by resampling from the original data set, constructing a predictor from each subsample, and decide by combining. By bagging an N-sample, the crude nearest neighbor regression estimator is turned into a consistent weighted nearest neighbor regression estimator, which is amenable to statistical analysis. Letting the resampling size k_N grow with N in such a way that $k_N \to \infty$ and $k_N/N \to 0$, we have shown that this estimator achieves optimal rate of convergence, independently from the fact that resampling is done with or without replacement. Since the estimator with the optimal rate of convergence depends on the unknown distribution of the observations, adaptation results by data-splitting are also obtained.

It is expected that this emerging activity should find some application domains, e.g. in the statistical analysis of recommendation systems, that would be a source of interesting problems and would provide some external support.

- ### Main contracts and technology transfer

On localization, navigation and tracking:

- **Geolocalization** (ALLOC 851, from 05/2005 to 08/2006), supported by France Télécom RD (now Orange Labs). The objective was to implement and assess the performance of particle filtering in localization and tracking of mobile terminals in a wireless network, using network measurements (received signal power strength and possibly TOA (time of arrival)) and a database of reference measurements of the signal power strength, available in a few points or in the form of a digital map (power attenuation map). Generic algorithms have been specialized to the indoor context (wireless local area network, e.g. WiFi) and to the outdoor context (cellular network, e.g. GSM). In particular, constraints and obstacles such as building walls in an indoor environment, street, road or railway networks in an outdoor environment, have been represented in a simplified manner, using a prior model on a graph, e.g. a Voronoi graph as in similar experiments in mobile robotics. The findings of this work is that localization in outdoor applications using measurements of the signal power strength alone is not yet operational, whereas the situation is much more favorable in indoor applications. This is because the digital maps available for GSM are obtained by running numerical propagation models that do not capture small-scale variations, and the solution would be to use additional measurements, such as TOA (time of arrival).

- **Terrain–aided navigation** (ALLOC 2857, from 09/2007 to 08/2010). This collaboration with Thalès Communications is supported by DGA (délégation générale à l'armement) under the PEA project NCT on terrain–aided navigation, and is related with the supervision of the CIFRE thesis of Nordine El Baraka. The overall objective is to study innovative algorithms for terrain–aided navigation, and to demonstrate these algorithms on four different situations involving different platforms, inertial navigation units, sensors and geo–referenced databases.

- **FIL** (ALLOC 2856, from 01/2008 to 12/2010), ANR project within the *Télécommunications* programme, coordinated by Thalès Alenia Space. The overall objective is to study and demonstrate information fusion

algorithms for localization of pedestrian users in an indoor environment, where GPS solution cannot be used. The design combines (i) a pedestrian dead-reckoning (PDR) unit, providing noisy estimates of the linear displacement, angular turn, and possibly of the level change through an additional pressure sensor, (ii) range and / or proximity measurements provided by beacons at fixed and known locations, and possibly indirect distance measurements to access points, through a measure of the power signal attenuation, (iii) constraints provided by an indoor map of the building (map-matching), and (iv) collaborative localization when two users meet and exchange their respective position estimates.

On sequential data assimilation:

- **ADOQA** (from 01/2005 to 12/2006). This ARC (action de recherche coopérative INRIA) was coordinated by EPI CLIME, INRIA Paris --- Rocquencourt and CEREA/ENPC. One objective of ADOQA was to investigate sequential methods (as opposed to variational methods) for data assimilation of intrinsically nonlinear models, i.e. coupling of numerical models and measured data. In principle, a data assimilation algorithm should propagate uncertainties through the probability distribution of the state variables, whereas current sequential algorithms, such as the ensemble Kalman filter (EnKF), only propagate the first two moments. For large-scale systems (physical state of the atmosphere, of the ocean, chemical composition of the atmosphere, etc.), the direct implementation of sequential Monte Carlo methods seems impractical, and simplified reduced-order models should be used. Our contribution has been to better understand and compare on simple examples such as the three-dimensional Lorenz model, the qualitative behavior and the performance in terms of the sample size, of the ensemble Kalman filter and other sequential data assimilation methods, with the qualitative behavior and the performance of particle filters. The interest of ASPI for these questions started on this occasion, with the internship of Vu-Duc Tran, later continued in her thesis.

- **PREVASSEMBLE** (ALLOC 3767, from 01/2009 to 12/2011), ANR project within the COSINUS (conception and simulation) programme, coordinated by Institut Pierre-Simon Laplace. The contribution of ASPI to this project is to continue the comparison, initiated during the PhD thesis of Vu-Duc Tran, of sequential data assimilation methods, such as the ensemble Kalman filter (EnKF) and the weighted ensemble Kalman filter (WEnKF), with particle filters. This comparison will be made on the basis of asymptotic variances, as the ensemble or sample size goes to infinity, and also on the impact of dimension on small sample behavior.

On rare event simulation:

- **RARE** (from 01/2006 to 12/2007). This ARC (action de recherche cooperative INRIA) was coordinated by EPI ARMOR (now DIONYSOS), INRIA Rennes --- Bretagne Atlantique, with EDF RD and DGAC/DSNA as industrial partners, and with CWI (The Netherlands) and University of Bamberg (Germany) as international partners. The objective of RARE was to design and evaluate various Monte Carlo techniques (importance sampling, importance splitting, cross-entropy, etc.), for the simulation of rare but critical events, in several important domains of applications (communication networks, financial risk management, air traffic management, etc.). Our contribution has been to better understand the asymptotic behavior of importance splitting methods, where intermediate less critical events are introduced, and where trajectories that manage to reach an upper level are replicated into a number of offsprings. Splitting can be achieved in many different ways: in *fixed splitting* for instance, each successful trajectory is given a prescribed deterministic number (possibly depending on the generation number) of offsprings, whereas in *fixed effort* splitting, there is a prescribed deterministic number of trajectories alive at each generation, which amounts to sample with replacement from the successful trajectories at the current stage of the algorithm, and in *fixed performance* splitting a random number of trajectories is simulated, until a prescribed deterministic number of successful trajectories is obtained. It appears that importance splitting can be interpreted in terms of Feynman-Kac distributions, which makes it possible not only to approximate the probability of the rare but critical event, but also to learn which critical trajectories are responsible for the critical event to occur. Challenging issues that have been investigated include the automatic selection of the intermediate sets and their number: asymptotic results have been obtained in the one-dimensional case, while in the multi-dimensional case a preliminary objective would be the efficient choice of the importance function, used to define the intermediate sets as level sets. These contributions were collected in a survey article included in the monograph published on this occasion.

On rare event simulation in protection of digital content:

- **NEBBIANO** (ALLOC 2229, from 01/2007 to 11/2009), ANR project, within the SETIN (sécurité et informatique) programme, initially coordinated by Teddy Furon (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique) and later by Arnaud Guyader. There are mainly two strategic axes in NEBBIANO: watermarking and independent component analysis, and watermarking and rare event simulations. To protect copyright owners, user identifiers are embedded in purchased content such as music or movie. This is basically what we mean by watermarking. This watermarking is to be *invisible* to the standard user, and as difficult to find as possible. When content is found in an illegal place (e.g. a P2P network), the right holders decode the hidden message, find a serial number, and thus they can trace the traitor, i.e. the client who has illegally broadcast their copy. However, the task is not that simple as dishonest users might collude. For security reasons, anti-collusion codes have to be employed. Yet, these solutions (also called weak traceability codes) have a non-zero probability of error defined as the probability of accusing an innocent. This probability should be, of course, extremely low, but it is also a very sensitive parameter: anti-collusion codes get longer (in terms of

the number of bits to be hidden in content) as the probability of error decreases. Fingerprint designers have to strike a trade-off, which is hard to conceive when only rough estimation of the probability of error is known. The major issue for fingerprinting algorithms is the fact that embedding large sequences implies also assessing reliability on a huge amount of data, which may be practically unachievable without using rare event analysis. Our task within this project is to adapt our methods for estimating rare event probabilities to this framework, and provide watermarking designers with much more accurate false detection probabilities than the bounds currently found in the literature. We have already applied these ideas to some randomized watermarking schemes and obtained much sharper estimates of the probability of accusing an innocent. A patent has been jointly submitted in August 2008 by INRIA and Bretagne Valorisation.

On rare event simulation in air traffic management:

- **iFLY** (ALLOC 2399, from 05/2007 to 08/2010). This FP6 project within the *Aeronautics and Space* programme is coordinated by the National Aerospace Laboratory (NLR) (The Netherlands), and can be seen as a follow-up of the previous FP5 project HYBRIDGE within the IST programme. Eighteen academic and industrial partners were involved, and actual collaboration occurred with Politecnico di Milano (Italy), NLR and University of Twente (The Netherlands), ETH Zürich (Switzerland), and DGAC/DSNA (France). The objective of iFLY is to develop both an advanced airborne self separation design and a highly automated air traffic management (ATM) design for en-route traffic, which takes advantage of autonomous aircraft operation capabilities and which is aimed to manage a three to six times increase in current en-route traffic levels. The contribution of ASPI to this project concerns the work package on accident risk assessment methods and their implementation using conditional Monte Carlo methods, especially for large scale stochastic hybrid systems: designing and studying variants suited for hybrid state space (resampling per mode, marginalization) are currently investigated.

On global optimization (seen as an application of rare event simulation):

- **Optimization of sensors position and activation** (ALLOC 4233, from 04/2009 to 03/2011). This project is supported by CTSN (centre technique des systèmes navals) a DGA (délégation générale à l'armement) entity. This collaboration with Sébastien Paris, from université Paul Cézanne, is related with the supervision of the PhD thesis of Mathieu Chouchane, and was initiated by Jean-Pierre Le Cadre. The objective of this project is to optimize the position and activation times of a few sensors deployed by a platform over a search zone, so as to maximize the probability of detecting a moving target. The underlying idea is to replace the problem of maximizing a cost function (the probability of detection) over the possible configurations (admissible position and activation times) by the apparently simpler problem of sampling a population according to a probability distribution depending on a small parameter, which asymptotically concentrates on the set of global maxima of the cost function, as the small parameter goes to zero.

The next series of ANR projects has been the occasion of a transfer to Alcatel Lucent of results and expertise in the field of signal processing for digital communications, especially equalization and synchronization, which belong to the former EPI SIGMA2.

- **COHDEQ40** (ALLOC 2205, from 12/2006 to 11/2009). This first ANR project within the *Télécommunications* programme was coordinated by Alcatel-Lucent. This is collaboration with Jean-Jacques Fuchs (EPI TEMICS, INRIA Rennes --- Bretagne Atlantique). The project COHDEQ40 intends to demonstrate the potential of coherent detection associated with digital signal processing for the next generation high-density 40Gb/s WDM systems optimized for transparency and flexibility. Key integrated optoelectronic components and specific algorithms will be developed and system evaluation performed. The INRIA task is to develop these signal processing algorithms needed to recover the message on the decoder side. This makes full use of the knowledge and expertise that belong to the former EPI SIGMA2 on equalization and synchronization techniques involved in digital communications. A patent has been jointly submitted in September 2008 by Alcatel Lucent and INRIA.

- **TCHATER** (ALLOC 2801, from 01/2008 to 12/2010). This second ANR project within the *Télécommunications* programme was coordinated by Alcatel-Lucent. The primary goal of the TCHATER project is to demonstrate a coherent terminal operating at 40Gb/s using *real-time* digital signal processing and efficient polarization division multiplexing. The terminal will benefit to next-generation high information-spectral density optical networks, while offering straightforward compatibility with current 10Gb/s networks. It will require that advanced high-speed electronic components, especially analog-to-digital converters, are designed within the project. Specific algorithms for polarization demultiplexing and forward error correction with soft decoding will also have to be developed.

- **STRADE** (ALLOC 4402, from 11/2009 to 10/2012). This third ANR project within the *Réseaux du Futur et Services* programme was coordinated by Alcatel-Lucent Bell Labs France. The focus of this project is to reduce the impact of nonlinear effect. The objective is twofold: specify, design, realize and evaluate fibers of reduced nonlinear effects by firstly increasing the effective area to unprecedented values and secondly, by splitting optical power along two modes, using bimodal propagation. While the first step is ambitious but primarily relies in the evolution of current fiber technologies, the second is disruptive and requires not only deep changes in fiber technologies but also new advanced transmitter / receiver equipment, preferably based on coherent detection. Naturally, bimodal propagation also brings another key advantage, namely a twofold increase of system capacity.

- **Main international collaboration**

Arnaud Guyader is collaborating with Nicolas Hengartner (LANL) on adaptive multi-level splitting method for rare event simulation and on Monte Carlo estimation of extreme quantiles of posterior probabilities. He has made several short visits to Los Alamos since the summer of 2009.

- **Visibility**

  - Awards

Jean-Pierre Le Cadre has received the *2006 Barry Carlton Award*, awarded by the IEEE Aerospace and Electronic Systems Society (AESS) to the best paper published in the IEEE Transactions on Aerospace and Electronic Systems in 2006, for his work on closed-form posterior Cramér-Rao bounds for bearings-only tracking, co-authored with his former PhD student Thomas Bréhard.

  - Conference organization

Arnaud Guyader and Frédéric Cérou have organized, together with Éric Matzner-Løber from université de Rennes 2, the workshop JSTAR (journées de statistique de Rennes) held at INRIA Rennes --- Bretagne Atlantique in October 2009.

  - Number of membership of PhD and HDR committees

François Le Gland has participated in 1 HDR committee: Bruno Tuffin, and in 8 PhD committees: Jaroslav Krystul (Twente), Olivier Rabaste (Télécom Bretagne), Anne Cuzol, Agnès Lagnoux (Toulouse), Boujemaa Ait El Fquih (Évry), Kari Heine (Tampere), Christophe Baehr (Toulouse), Benoît Landelle (Orsay).

- **Training (teaching at master level, number of PhD and HDR defenses in the team, etc.)**

As an associate professor, Arnaud Guyader (yearly total, 192 hours) is teaching analysis, statistics and probability at université de Rennes 2, and he is also a member of the committee of *oraux blancs d'agrégation de mathématiques* for ENS Cachan at Ker Lann.

François Le Gland (yearly total, 52 hours) gives several courses: on *Bayesian filtering and particle approximation* at ENSTA (école nationale supérieure de techniques avancées, 21 hours), on *hidden Markov models* at Télécom Bretagne (6 hours), on *linear and nonlinear filtering* at ENSAI (école nationale de la statistique et de l'analyse de l'information, 10 hours) and on *Kalman filtering and hidden Markov models* at université de Rennes 1, master in electronics and telecommunications, track in signal and image processing, embedded systems, and control (15 hours).

3 PhD defenses: Vu-Duc Tran (June 2009), Francis Céleste (February 2010), Adrien Ickowicz (May 2010).

4) Publications : the ten major publications of the evaluation period

OS-06-0011   ASPI   OS   2006   A sequential algorithm that keeps the particle system alive   Stochastic Hybrid Systems : Theory and Safety Critical Applications   Le Gland, François:Oudjane, Nadia

AP-08-0012   ASPI   AP   2008   A nonasymptotic variance theorem for unnormalized Feynman-Kac particle models   Cérou, Frédéric:Del Moral, Pierre:Guyader, Arnaud

ACL-06-0022   ASPI   ACL   2006   Genetic genealogical models in rare event analysis   ALEA, Latin American Journal of Probability and Mathematical Statistics   Cérou, Frédéric:Del Moral, Pierre:Le Gland, François:Lezaud, Pascal   Paper 01-08

ACL-07-0014   ASPI   ACL   2007   Adaptive multilevel splitting for rare event analysis   Stochastic Analysis and Applications   Cérou, Frédéric:Guyader, Arnaud

OS-08-0006   ASPI-TEMICS   OS   2008   On the design and optimization of Tardos probabilistic fingerprinting codes   10th International Workshop on Information Hiding, Santa Barbara   Furon, Teddy:Guyader, Arnaud:Cérou, Frédéric

AP-09-0033   ASPI-TEMICS   AP   2009   Rare event simulation for a static distribution   Cérou, Frédéric: Del Moral, Pierre:, Guyader, Arnaud: Furon, Teddy

OS-10-0002   ASPI   OS   2010   Large sample asymptotics for the ensemble Kalman filter   Handbook on Nonlinear Filtering   Le Gland, François:Monbet, Valérie:Tran, Vu-Duc

ACL-06-0023   ASPI   ACL   2006   Nearest neighbor classification in infinite dimension   ESAIM : Probability and Statistics   Cérou, Frédéric:Guyader, Arnaud

ACL-10-0006   ASPI   ACL   2010   On the rate of convergence of the bagged nearest neighbor estimate   Journal of Machine Learning Research   Biau, Gérard:Cérou, Frédéric:Guyader, Arnaud

ACL-10-0007   ASPI   ACL   2010   Rates of convergence of the functional k-nearest neighbor estimator   IEEE Transactions on Information Theory   Biau, Gérard:Cérou, Frédéric:Guyader, Arnaud

5) Self assessment

ASPI is obviously a small research team in size, and one could get the wrong impression, from the many applications that are addressed here, that the team is insufficiently focused. It is therefore important to stress here that ASPI is essentially concerned with methodological issues, i.e. with the mathematical analysis of popular algorithmic variants and implementations of interacting Monte Carlo methods that have been proposed and used in different communities, to solve different problems. This implies collaboration with experts from these different communities. In this respect, there is indeed a common feature to most of the activities lead in ASPI. The most notorious exceptions to this explanation are (i) the ANR projects with Alcatel Lucent on signal processing for digital communications, which is truly the transfer of an expertise that belongs to the former EPI SIGMA2, and which as such is expected to slowly decrease, and (ii) the recent activity on classification in infinite dimension, the future of which is still open.

6) Research perspectives for the next four years

The proposition for the next four years can be divided in three parts.

The first part consists in the continuation of ongoing works, some of which are connected with workpackages of the european project iFLY or the ANR projects FIL and PREVASSEMBLE. To this first group belong studies on:

- comparison of parametric and nonparametric learning of the optimal importance distribution,
- asymptotic behavior of rare event simulation methods adapted to hybrid stochastic systems,
- asymptotic behavior of KLD-sampling, an adaptive sample size Monte Carlo method, proposed in the mobile robotics community,
- large sample asymptotics of the ensemble Kalman filter and its weighted version, introduced by Étienne Mémin (EPI FLUMINANCE, INRIA Rennes --- Bretagne Atlantique) and his group.

The second part addresses the statistical analysis of recommendation systems, and can be seen as an opportunity to apply recent results and expertise obtained by ASPI, and already described in a previous section, on classification in infinite dimension. This would build upon the research project of the (unsuccessful) proposition MESSY submitted to the ARC (action de recherche coopérative INRIA) programme, and it would also contribute to the theme of *web of knowledge and services*, which has been identified by INRIA in its strategic plan as one of its priority for the next period.

The third and last part deals with two new research directions related to rare event simulation: global optimization and molecular simulation.

- **Global optimization.** An issue closely related to rare event simulation is global optimization. Indeed, the difficult problem of finding the set M of global minima of a real-valued function V can be replaced by the apparently simpler problem of sampling a population from a probability distribution depending on a small parameter, and asymptotically supported by the set M as the small parameter goes to zero. The popular approach here is to use the cross-entropy method, which relies on learning the optimal importance distribution within a prescribed parametric family. On the other hand, multi-level splitting methods developed in ASPI could provide an alternate nonparametric approach to this problem, and it is worth comparing these with new algorithms proposed recently. Besides theoretical studies, interest in random optimization algorithms such as the cross-entropy method to solve operations research problems can be seen as a legacy of our colleague Jean-Pierre Le Cadre, who passed away in July 2009: indeed, trajectory optimization for a mobile robot localization with map uncertainties, was addressed in the thesis of Francis Céleste (DGA/CEP), trajectory optimization to minimize the detection of an aircraft is currently addressed in the doctoral project of Renaud Cariou (DGA/CELAR), and optimization of sensors position and activation is the object of a contract with DGA/CTSN and is currently addressed in the doctoral project of Mathieu Chouchane, a PhD student of Sebastien Paris.

- **Molecular simulation.** It has recently occurred, through discussions with Tony Lelièvre (CERMICS and EPI MICMAC, INRIA Paris --- Rocquencourt), that multi-level splitting methods developed and studied in ASPI for the simulation of rare events could be of interest in molecular simulation. To be specific, consider the problem of computing the (small) probability that a molecule changes from a metastable state A (the local minimum of a given potential function) to another metastable state B, under some stochastic dynamics. This can be seen as the problem of computing a rare event, and variance reduction methods, such as importance sampling, are popular in this context. However, multi-level splitting methods have the advantage that not only the transition probability can be evaluated accurately, but also the typical critical trajectories that lead from A to B are automatically exhibited, as a by-product of the method. This is a useful piece of information, since it can explain the mechanism used by the molecule to reconfigure itself. There is an obvious interest in starting this collaboration: exporting our expertise on interacting Monte Carlo methods on one hand, and in return getting insight on a new application domain with new problems, and importing solutions that have been imagined in a different community, that could be useful in our more usual application domains. The collaboration should also involve Erwan Faou (EPI IPSO, INRIA Rennes --- Bretagne Atlantique), Mathias Rousset (EPI SIMPAF, INRIA Lille --- Nord Europe) and Florent Malrieu (IRMAR) who has recently applied for a *délégation* (temporary research position) at INRIA with a related research project.