Solving Adaptive Sampling Problems in Graphical Models using Markov Decision Processes

Mathieu BONNEAU[1,2], Nathalie PEYRARD[1] and Régis SABBADIN[1]

[1] INRA-Toulouse ; UBIA UR875 ; BP 52627 – 31326 Castanet-Tolosan, France
[2] INRA, UMR1210 Biologie et Gestion des Adventices, F-21065 Dijon
{mathieu.bonneau,nathalie.peyrard,regis.sabbadin}@toulouse.inra.fr

Abstract. In environmental management problems, decision should ideally rely on knowledge of the whole system. However, due to limited budget, in practice only a small part of the system is sampled and the complete system state is reconstructed from the sampled observations. In this article we consider the situation where the biological system under study is structured and can be modeled as a graphical model. Optimal sampling in such models still raises some methodological questions, like adaptive sampling, or the measure of the quality of a sample in terms of quality of reconstruction. Here, we present a way to formalise these two questions. The sample is chosen as the one which maximises the expected utility of information brought by the observations minus the sample cost. The utily is derived from the notion of *Maximum a Posteriori*. This problem is known to be **NP**-hard. We present how to modelit as a *Markov decision process* in order to build approximate solution methods based on Reinforcement Learning.

1 Introduction

In many environmental management problems, decision should ideally rely on knowledge of the whole state of the system. For instance, for weeds management in an agricultural area, occurence maps of the different species is helpful for controling their spread. In most cases constructing such maps is difficult because it is too expensive to explore the whole area to map. Another example is the control of an epidemics on a social network. Usually, observations are available only for a sample of the population. A good sample must reach a trade-off between quality of the reconstruction of the whole system and sample cost.
In this paper we present preliminary work on the question of optimal sampling in a structured system. This question is classically studied in statistics [4], [2] and computer science [1], [3] communities. Still some issues remain open, like the quantification of the information brought by the sampled observations when the objective is the reconstruction of the whole system, or the modeling and resolution of optimal adaptive sampling. In adaptive sampling, the set of sampled sites is not chosen once and for all. Sample is sequential and observations from

previous sample steps are taken into account to select the next sites to explore. We consider here the situation where the system can be modeled as a graphical model. First, we present a formalisation of the problem of optimal sequential sampling based on the *Maximum A Posteriori* criterion to measure the quality of a sample. Then we propose a modelisation of the problem as a *Markov decision process* [6]. This will enable us to build approximate solution methods based on Reinforcement Learning [7].

2 Optimal adaptive sampling in graphical models

Let us consider a non directed graphical model (or Markov random field) on a set $X = (X_1, \ldots, X_n)$ of discrete random variables taking values in $\Omega^n = \{1, \ldots, K\}^n$. The model is defined by a graph $G = (V, E)$ where $V = \{1, \ldots, n\}$ and $E \in V^2$, and by a set of potential functions Ψ_c, defined on the cliques of G and parametrised by θ[1]. Then the joint distribution of X can be factorised as $\mathbb{P}(X = x \mid \theta) \propto \prod_{c \in \mathcal{C}} \Psi_c(x_c, \theta)$, where $\mathcal{C}$ is the set of cliques of V and $x_c = (x_i)_{i \in c}$.
Our goal is to reconstruct the vector X on a specified subset $R \subseteq V$. We can acquire observations only on a subset $O \subseteq V$ such that $R \cup O = V$. The intersection between O and R can be non empty. The problem is to select a set of *sites* $A \subseteq O$, named a *sample*, where X will be observed in order to build a "good"reconstruction of X_R. The sample choice procedure will also take sampling costs into account. An example of sampling for weeds occurrence map reconstruction is illustrated on Figure 1.

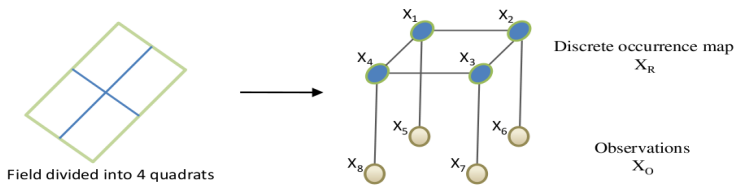

Fig. 1. Sampling for weeds occurence map reconstruction. False negative are possible, since a weed can be missed. Left: discretisation of the field. Right: the associated graphical model is that of a hidden Markov random field, $\Omega = \{0, 1\}^8$, $X_R = \{X_1, \ldots, X_4\}$, $X_O = \{X_4, \ldots, X_8\}$

.

Adaptive policy In adaptive sampling, the sample A is chosen sequentially. The sampling procedure is divided into $H + 1$ steps. A^h is the sample (set of sites) explored at step h and X_{A^h} (resp. x_{A^h}) are the random variables (resp. observations) corresponding to the sample output at step h. The choice of sample

[1] We don't focus on the estimation of θ and consider that it is known.

A^h depends on the previous samples and samples outputs. An adaptive sampling policy $\delta = (\delta_0, \ldots, \delta_H)$ can then be defined by an initial sample A^0 and functions δ_h specifiying the sample at step h, depending on the results of the previous steps: $\delta_0 = A^0,\ \delta_1((A^0, x_{A^0})) = A^1 \ldots \delta_H((A^0, x_{A^0}), \ldots, (A^{H-1}, x_{A^{H-1}})) = A^H$. An *history* is a trajectory of the policy $\delta : (A^0, x_{A^0}), \ldots, (A^H, x_{A^H})$. If $A = \cup_h A^h$ and $x_A = \cup_h x^h$, an history will also be denoted (A, x_A). (Here we assumed that a site can be visited only once.) The set of all histories of a policy δ is τ_δ. In the following we will consider the set Δ_L of adaptive policies such that at each step the sample size is less than or equal to L.

Reconstruction When a sample A and a sample output x_A are available, we can use a criterion classically used in Image Analysis for deriving an estimator x_R^* of x_R, named *Maximum A Posteriori* :

$$x_R^* = \arg\max_{x_R \in \Omega^{|R|}} \mathbb{P}(x_R \mid (A, x_A), \theta)$$

More precisely, the reconstruction x_R^* is the most likely configuration of X_R given the whole history (A, x_A).

Sample cost The modeling of a cost function is a question as its own. Here we illustrate this notion with a simple one where sample costs are additive. For a given history (A, x_A), with $A = \{A^0, \ldots, A^H\}$, the cost of A is

$$c(A) = \sum_{h=0}^{H} \Big(\sum_{a \in A^h} c_a \Big), \quad \text{with} \quad c_a \in \mathbb{R}^+$$

Quality of a sampling policy The quality of a policy δ is measured from the quality of the estimators x_R^* that can be obtained from δ. This quality is measured by the value of the criterion optimised to get x_R^*. More precisely we first define the quality of an history (A, x_A):

$$U\big((A, x_A)\big) = \max_{x_R \in \Omega^{|R|}} \Big[\mathbb{P}(x_R \mid (A^0, x_{A^0}), \ldots, (A^H, x_{A^H}), \theta) \Big] \tag{1}$$

The quality of a sampling policy δ is then defined as the expectation over all possible histories of the quality of an history minus the cost of the history :

$$Q(\delta) = \sum_{(A, x_A) \in \tau_\delta} \mathbb{P}\big(x_{A^0}, \ldots, x_{A^H} \mid \theta\big) \big[U\big((A, x_A)\big) - \alpha c(A) \big]$$

The constant α allows to homogenize the scales of restoration quality and cost. Finally the problem of optimal adaptive sampling amounts to finding the policy of highest quality : $\delta^* = \text{argmax}_{\delta \in \Delta_L} Q(\delta)$.
Exact optimization of the optimal sampling policy is intractable in practice [5]. We must turn to approximate solution methods. In the next section we present a modeling of the optimal sampling problem as a Markov Decision Process. This will enable us to consider powerfull approximate methods from the family of Reinforcement Learning.

3 Markov Decision Process Model

3.1 Markov Decision Processes and Reinforcement Learning

Markov Decision Processes (MDP) [6] provide a mathematical framework and efficient optimisation algorithms for sequential decision under uncertainty. In this section, we show how the problem of Optimal Adaptive Sampling in Graphical Models (OASGM) can be modelled as the resolution of a MDP which state and action spaces sizes are exponential in the size of the original problem. This exponential space representation cannot be avoided since MDP are known to be solvable in polynomial time in their representation, while graphical models optimal sampling problems are known to be NP-hard to solve, even in their non-adaptive version [5].
Explicit representation of the problem (and its solution policy) can be avoided, thanks to the use of Reinforcement Learning (RL) algorithms [7]. The RL family provides simulation-based MDP solution algorithms, which (i) do not require an explicit representation of the MDP transition and reward models, (ii) do not compute "a priori" the (exponentially large) optimal policy of the MDP but compute "on-line" the action to execute in the current state. So, the MDP encoding of the OASGM problem which we provide here will not be used to compute tabular representations of transition and reward functions, but rather be used for simulating state transitions and rewards within RL algorithms.
A finite-horizon Markov Decision Process is a 5-tuple $< S, D, T, p, r >$, where

- S is a finite set of systems *states*.
- D is a finite set of available *decisions* (or *actions*).
- $T = \{0, \ldots, T_e\}$ is a finite set of decision steps, termed *horizon*.
- $p_d^t(s'|s)$ are *transition functions*. $p_d^t(s'|s)$ indicates the probability that state s' results when decision d is implemented at time $t \in \{0, \ldots, T_e - 1\}$, when the system is in state s.
- $r^t(s, d)$ are *reward functions*. Reward $r^t(s, d) \in \mathbb{R}$ is obtained when the system is in state s at time t and decision d is applied, for all s, t. Note that rewards can be positive or negative (modeling costs).

A *decision policy* (or *policy*) $\delta = \{\delta_0, \ldots, \delta_{T_e-1}\}$ is a set of decision functions $\delta_t : S \to D$. Once a decision policy is fixed, the MDP dynamics becomes that of a finite Markov chain over S. The *value function* $V^\delta : S \times T \to \mathbb{R}$ of a policy δ is defined as the expectation of the sum of future rewards obtained, from the current state and time step, when following the Markov chain defined by δ.

$$\forall s \in s, t \in T, \quad V^\delta(s, t) = \mathbb{E}_\delta \left[\sum_{t'=t \ldots T_e} r^{t'} \mid s \right]$$

Solving a MDP amounts to finding an *optimal policy* δ^* which value is maximum in every states and decision steps ($V^{\delta^*}(s, t) \geq V^\delta(s, t), \forall \delta, s, t$).
The *backwards induction* algorithm [6] can be applied to compute δ^*. Alternately, RL algorithms [7] (such as *Q-learning*) can be applied to compute optimal actions (provided that sufficient computational effort is allocated) *on-line* in states actually encountered.

3.2 MDP encoding of the OASGM problem

In this section, we provide a MDP encoding of the OASGM problem.

State space The curent state at time t , $s^t \in S$, summarises the current information about variables indexed in O. This information is obtained from the sequence of past samples and outputs and thus $s^t \simeq \{(A^0, x_{A^0}), \ldots, (A^{t-1}, x_{A^{t-1}})\}$. It may be convenient to model s^t as a $|O| \times t$ matrix s, where $s_{i,j} = k > 0$ if and only the i^{th} variable of O was observed in state $k \in \{1, \ldots K\}$ during the j^{th} sample step. If this variable was not observed during this sample step, $s_{i,j} = 0$. An example state for $t = 1$, $K = 2$ and $O = \{1, 2, 3\}$ is : $s^1 = (0, 1, 2)'$. During the first sampling step we observed variable X_2 (value 1), variable X_3 (value 2) and X_1 was not observed. Note that s^0 is the empty matrix.
With this encoding, the total number of possible states of the system is exponential in the OASGM representation size (upper-bounded by $H \times (K+1)^{|O|}$), as mentioned above.

Action space A decision $d^t \in D$ will simply be the sample action A^t chosen at time step t. As such, it can be modelled, for example, as a length $|O|$ vector, with at most L entries equal to 1 (the remaining ones being equal to 0). With this encoding, the total number of actions is at most $|\{A^t \mid A^t \subseteq O,\ |A| \leq L\}|$.

Horizon Decision steps in the MDP correspond to decision steps in the OASGM problem. Thus, $T = \{0, \ldots, H+1\}$. After decision d^H has been implemented in state s^H at decision step H, a final state s^{H+1} will be reached.

Transition functions Note that in the MDP representation, the state s^t can be identified to $\big((A^0, x_{A^0}), \ldots, (A^{t-1}, x_{A^{t-1}})\big)$, for $t = \{0, \ldots, H\}$. So, it is easy to check that :

$$\forall t \in \{0, \ldots, H\} \quad p^t_{d^t}\big(s^{t+1} \mid s^t\big) = \mathbb{P}\big(x_{A^t} \mid (A^0, x_{A^0}), \ldots, (A^{t-1}, x_{A^{t-1}}), A^t, \theta\big).$$

These transition probabilities can be computed online (for fixed s^t and d^t), but only at high computational cost, since their computation is equivalent to marginal probabilities computation in a graphical model. However, the dynamics of the model can be efficiently *simulated*, by simulating values of the variables in R and values of the sample output x_{A^t}, from past sample outputs and current sample A^t.

Reward functions We set $r^0(s^0) = 0$ and for all $t \in \{1, \ldots, H\}$, rewards represent sampling costs. For state s^t, we have :

$$r^t(s^t, d^t) = r^t(s^t) = -\alpha c(A^{t-1})$$

After decision d^H has been applied at decision step H, a final state $s^{H+1} = \big((A^0, x_{A^0}), \ldots, (A^H, x_{A^H})\big)$ is reached, in which the reward $r^{H+1}(s^{H+1})$ is the quality of the reconstruction X^*_R minus the cost of action A^H :

$$r^{H+1}(s^{H+1}) = U\big((A^1, x_{A^1}), \ldots, (A^H, x_{A^H})\big) - \alpha c(A^H),$$

where U is defined by equation (1).
The optimal policy for the above-defined MDP is a set of functions associating samples to lists of past observation outputs. It thus has the same structure as a OASGM sampling policy. More than that, we can show the following proposition:

Proposition 1 (OASGM-MDP optimal policies equivalence). *An optimal policy for the MDP model of a OASGM is optimal for the initial OASGM.*

4 Conclusion

In this article, we have presented the problem of adaptive sampling in graphical models and its modeling as a Markov decision process. The size of the state space is exponential in the size of the set O of sites which can be sampled and exact computation of the value function is intractable. Next, we will exploit Reinforcement Learning methods and simulation techniques in graphical models to develop approximate sampling methods. It is worth noting that the problem we have described could also be modeled as a (structured) *Partially Observed MDP* (POMDP), eventhough it is a strict subclass (no evolution of the system and no action which modifies the system). The corresponding POMDP will be of exponential complexity, so that existing algorithms would be difficult to apply. For this reasons, the MDP approach seems more suited to us.
A straightforward application of adaptive sampling in graphical models is that of spatial sampling in Hidden Markov Random Fields. In this case, the set R is the set of sites of the hidden image, O is the set of observation nodes and $V = R \cup O$. Based on the framework presented here, we aim at modeling and solving more general sampling problems, with any graph structure and any choice of the sets O and R.

References

1. Das, A., Kempe, D.: Algorithms for subset selection in linear regression. In: Symposium on the Theory of computing (2008)
2. Gruijter, Brus, B., Knotters: Sampling for Natural Resource Monitoring. Springer (2006)
3. Krause, A.: Optimizing Sensing, Theory and Applications. Ph.D. thesis, School of computer Science, Carnegie Mellon University Pittsburg, PA 15213 (2008)
4. M. J. Dobbie, B.H., D.L Stevens, J.: Sparse sampling : Spatial designe for monitoring stream networks. Statistics Surveys 2, 113–153 (2007)
5. Peyrard, N., Sabbadin, R., Niaz, U.F.: Decision-theoretic optimal sampling with hidden markov random fields. In: European Conference of Artificial Intelligence (ECAI) (2010)
6. Puterman, M.: Markov Decision Processes : Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc (1994)
7. Sutton, R.S., Barto, A.: Reinforcement Learning : An Introduction. MIT Press (1998)