The Strategic Student Approach for Life-Long Exploration and Learning

Manuel Lopes, Pierre-Yves Oudeyer

The Strategic Student Approach for Life-Long Exploration and Learning

Manuel Lopes and Pierre-Yves Oudeyer
INRIA, Bordeaux Sud-Ouest, France
Email: {manuel.lopes, pierre-yves.oudeyer}@inria.fr

***Abstract*—This article introduces the strategic student metaphor: a student has to learn a number of topics (or tasks) to maximize its mean score, and has to choose strategically how to allocate its time among the topics and/or which learning method to use for a given topic. We show that under which conditions a strategy where time allocation or learning method is chosen from the easier to the more complex topic is optimal. Then, we show an algorithm, based on multi-armed bandit techniques, that allows empirical online evaluation of learning progress and approximates the optimal solution under more general conditions. Finally, we show that the strategic student problem formulation allows to view in a common framework many previous approaches to active and developmental learning.**

I. Introduction

Life-long learning of multiple tasks in real world robots poses challenging problems. Since time, physical and cognitive resources are limited, learning requires multiple kinds of choices to be made by the learner or by its teacher. For example, one has to choose how to allocate time to the practice and learning of each task, and/or to choose which learning methods to use for a given task (there are in general multiple learning methods available, and a given one might be more suited to certain tasks). These choices generate a particular allocation of time among the topics. This learning trajectory can have a major impact on both what is learned and how efficiently it is learned.

In the computational learning literature, various approaches have been taken to study the mechanisms that can generate these trajectories as well as to study their impact. A first family of approaches have considered mechanisms where ordered training examples are provided by an external teacher, such as in optimal teaching techniques [1] or in curriculum learning [2], [3]. A second family of approaches have considered internal mechanisms of maturation, which control the evolution of the complexity of percepts and actions by progressively increasing the number or spatio-temporal resolution of motor channels [4], [5] or sensori channels [5]–[7].

A third family of approaches have considered internal mechanisms for active choice of either training examples [8]–[10], tasks to be explored [11], learning methods to be used [12], [13] or questions to pose to a teacher [14]. Two broad strategies have been investigated. In the optimal experiment design and statistical active learning literature [15], [16], methods have been devised around the strategy consisting in first exploring what is maximally complex (e.g. uncertain [8], high prediction errors [17], least visited [18]) and then progressively explore what is more simple. In the developmental robot learning literature, the opposite strategy has been explored, consisting in first exploring what is simpler, and then progressively exploring what is more complex [11], [19]–[21]. To realize such a developmental exploration, mechanisms of intrinsically motivated exploration and learning were introduced [19], [20], [22] modeling aspects of human spontaneous exploration and motivation [23]–[25]. In particular, the notion of *learning progress*, measuring how learning performance *improve* over time or over learning methods, has been proposed as a measure to maximize during exploration [19], [20], and automatically resulting in this developmental exploration from simple to complex [20].

While a number of qualitative arguments have been proposed to argue for the general higher suitability of the developmental learning approach for life-long multi-task learning in the real world [11], [19]–[21], a precise and formal characterization of learning problems where this can be precisely shown to be the case is still lacking [1]. In this article, we aim at such a characterization, introducing and formalizing a general class of learning problems, which we call the Strategic Student Problem. First, section II introduces a simple version of the problem, focusing on the multi-task setting, and explaining intuitively the strategic student metaphor. Then, through a numerical optimization experiment, we exhibit the optimal solution to an instance of this problem, and show that it has a developmental structure: simple tasks are explored first, and then a progressive and ordered transition towards more complicated tasks is observed. Then, section III presents a more general formalization of the underlying class of multi-task learning problems. We then show that, if certain general underlying properties of the learning problems are present (sub-modularity), an optimal developmental solution to this class of problem can be achieved through greedy maximization of learning progress. We discuss the consequences of relaxing

The authors are with the Flowers Team, an INRIA and ENSTA-Paristech joint-lab. Work (partially) supported by INRIA, Conseil Régional d'Aquitaine and the ERC grant EXPLORERS 24007.

[1]Formal study of active learning in general is also quite recent [15], [16]. The main problem is that most theory on learning relies on the assumptions that the learning data is acquired randomly, i.e. with the same distribution as the future encounters, and in active learning the agent itself chooses which data to sample. Recent development from machine learning, mainly from active learning and multi-armed bandits, started to contribute to a formal view on the complexity of learning agents that choose their own samples: optimal experimental design and *active learning* [8], [15], [26]–[28], n-armed bandits [29] and the general exploration-exploitation dilemma in RL [18]

sub-modularity assumptions, see section III-B. In section IV, we then discuss how such theoretical optimal solutions can be approximated by real online algorithms, in particular in the case where the learning progress needs to be estimated empirically. We then present a novel algorithm, combining bandit techniques [29] together with aspects of intrinsic motivation systems presented in [11], [30], and present initial experiments to evaluate its performance. Finally, section VI shows that the strategic student problem formulation allows to view in a common framework many previous approaches to active and developmental learning.

II. The strategic student problem

In this Section we present an analogy providing an initial intuition for our problem. We imagine a student that is going to be tested in K different topics and there are still N days before the exam. The student is *strategic* in the sense that the motivation is to have the best total score choosing carefully what and how to study[2]. The problem for the student is to decide, each day, how to allocate the time to study each particular topic. For now we assume that the topics are learnable at different rates and with different expected final scores. The learning curve for the different subjects is thus different and this leads to two effects: a) for the same allocated time on a given topic the expected score differs and b) the learning rate decreases with the total allocated time. The problem can then be defined as finding the time allocation that maximize the average scores on all the different subjects with the restriction that the total time allocated is equal to the total time until the exam N. If we represent n_i the total amount of time taken with topic i and by $q_i(n_i)$ a function that describes the expected score on topic i if n_i time is used to study it. The strategic student has to solve the following problem:

$$\max_{n_1,\ldots,n_K} \sum_{i=1}^{K} q_i(n_i)$$
$$s.t. \sum_i n_i = N \ , \ n_i \geq 0$$

An interesting aspect is that when studying for some topics some chapters might be common with some other topics. Due to this aspect, a correlation between the scores is expected as studying one topic also corresponds to studying part of another. We can make a more complex model that represents this correlation between topics. We can make the total score per subject depend on the time taken on all subjects. For this we defined a new function $q_i(\sum_{j=1}^{K} \phi_{ij} n_j)$ where the weights ϕij represent how the different topics are related, To solve this problem students rely on different heuristics: divide the time equally among topics, finish learning the easier topics first, concentrate on hard topics, at each day pick the one with least expected score, among many others. More formally we can select, at each time instant, based on the following measures:

- topic with less time allocated
- topic with lower expected score
- topic with score closest to the maximum
- topic with maximum expected improvement in score

The first measure corresponds to an uniform time allocation, as we know that the same time allocated will yield different scores we hope that there exists other strategies that are better than this one. The second and third criteria might be worse than an uniform allocation because the rate of learning is different and also the best possible score on each topic is not the same. Because of this assuming that using time with the less explored topic, or the topic with the less expected score might result in no progress at all in terms of the total score. The last criteria is the one the most promising one because it balances exploration accordingly to the progress done and the progress that is still possible.

As we will see latter in the document, the structure of the scoring function $q(.)$ has a huge impact on the algorithms we can use. We will start by considering a simplified scoring function that can be solved efficiently and allows to verify the qualitative behavior of the optimal solution. We will be able to characterize the properties of the optimal solution to the problem and better understand the required properties of a solution algorithm for our problem.

We consider K topics each one with a standard power law for the scores. Defining p_i as the difficulty of topic i, the expect total score is the sum of the subject's score, $q_i = C_i \left(1 - e^{-\frac{n_i}{p_i}}\right) + B_i$. The coefficients C and B are written to make explicit that it is not possible to achieve the same score in all topics and that some topics weight more than others. For a known time frame N, the strategic student problem can then be defined as:

$$\max_{n_i} \sum_i C_i \left(1 - e^{-\frac{n_i}{p_i}}\right) + B_i$$
$$s.t. \sum_i n_i = N \ , \ n_i \geq 0$$

In that formalization there is an abuse of notation as the number of hours of study must be integer. Note that for a particular time N the problem is to find how many hours per topic to study.

This problem can be solved efficiently with any convex optimization tool, and if required to have integer times, a branch and bound step can be used. The parameters used were $C = 1, 1, 0, 1, 1$, $B = 0, 0, 0.5, 0.5, 0$ and $p_i = .1, 1, 10, 20, 50$ for respectively tasks $i = 1 \ldots 5$. We can see that these parameters ensure that we have tasks that are learned faster than others and that some of them have a higher impact in the total score. Figure 1 shows the result of this optimization process. The most interesting aspect to note is that the optimal strategy is non-stationary in the sense that for different time-frames, the percentage of time applied to each topic is different. We can see that there is a developmental progression from learning

[2]There are three main types of learning styles [31]: *surface learning*, someone that just memorizes and does not understand the concepts in depth, *deep learning*, someone that tries to understand all the fundamental concepts and *strategic learners* that use whatever learning style is required to achieve best quantitative results.

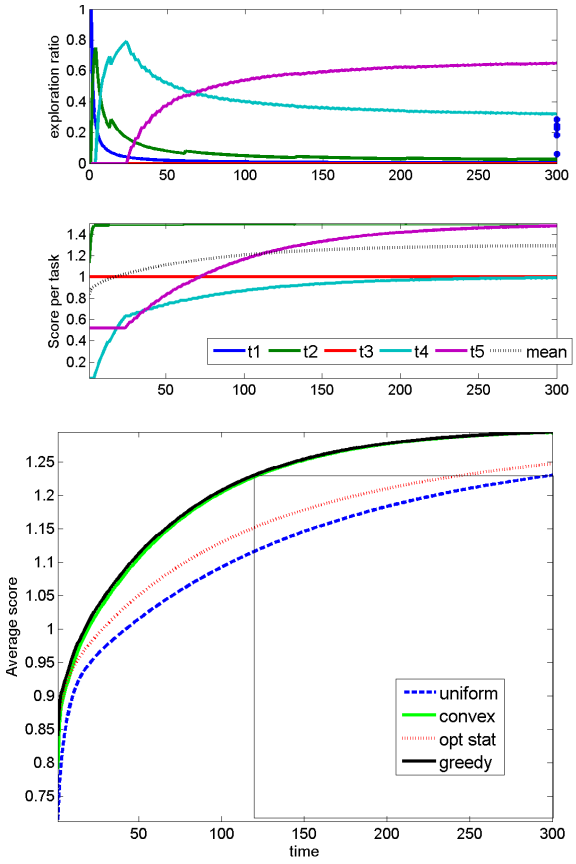

Fig. 1. Evolution of learning with the time available to learn. For each learning length the figure show respectively: (top) optimal cumulative percentage of time used in each task, (middle) evolution of score when this optimal strategy is used and (bottom) comparison between different sampling methods: random, optimal solution, optimal stationary solution and greedy (see main text for definition). We can note that the optimal strategy is non-stationary and developmental: an ordered structure appears where easier tasks are sample first, and then progressively more complicated tasks.

simpler topics to more complex ones. Even at the extreme cases where with little amount of time some topics are not studied at all. We compare the results with an uniform time allocation, that as expected is worse, and with an asymmetric stationary policy. This means that we run the optimization for a large N to compute the percentage of time allocated in each topic. Then we run a simulation using always such ratios of study. This approach is better than an uniform allocation and converges to the optimal one, and thus we call it "optimal stationary policy".

In this problem we can even go a step further an test if a simple heuristic such as learning progress is able to solve the problem. That is, at each time instant the student chooses the task where the expected progress is maximum: $argmax_i q_i(n_i + 1) - q_i(n_i)$. We see that this greedy strategy provides identical results to the theoretical optimal non-stationary solution.

This results confirms the heuristics of learning progress given by [19] and [20]. Both works considered that at any time instant the learner must sample the task that has given a larger benefit in the recent past. For the case at hand we can see that the solution is to probe, at any time instant, the task whose learning curve has a higher derivative, and for smooth learning curves both criteria are equivalent. In Section III we will show the conditions under which this heuristic is optimal.

III. Problem Definition

We can now define the problem in its more general setting. We assume that we want to estimate a function $f : \Re^l \rightarrow \Re^q$ of which we make the assumption that is composed of a sum of k functions $g_\theta^k : X_g \subset \Re^l \rightarrow \Re^q$ (the different topics) potentially with different characteristics θ_k (e.g. different bandwidths, noise levels, order). To simplify notation and to make the locality property of the functions explicit we assume that g is defined in the complete domain but multiplied by a mixture function $b_k : \Re^l \rightarrow [0 \dots 1]$. If the topics are independent then the different functions are indicator functions without overlapping domains, i.e. for a given input only one function is one. We assume we have a dataset $D = \{(x_i, y_i), i = 1 \dots n\}$. The target function can be defined as $f(x) = \sum_{k=1}^{K} b_k(x) g_\theta^k(x)$. We want to find an estimator for this target function[3]:

$$\hat{f}(x) = \sum_{k=1}^{K} b_k(x) \hat{g}_\theta^k(x; D)$$

where in $\hat{g}_\theta^k(x; D)$ we expressed explicitly the dependence on the dataset. We define a measure of how good is our estimator on the general function $h(x; D) = h(f(x), \hat{f}(x; D))$. For now $h(.)$ can be considered a metric but latter on we will see which properties this function must fulfill. The global target function can be defined as:

$$G(D) = \int_x h(x; D) dx$$

Our learning task is to probe the system for N examples $D_{1:N}$ in order to maximize G.

Problem 1: The Strategic Student Problem (SSP)

$$\max_D G(D)$$
$$s.t. \# D = N$$

We do not consider the case of selecting a set of samples from the ambient space but to select among a finite set of possibilities (choices). In the Strategic Student example we consider which topic to study from. Here we start by considering that we choose how many samples we must sample from each region. Inside each region the samples can be generated in any way, consider randomly for now. Later on we will show how different algorithms can be derived by changing the meaning of the choices.

Solving the *Strategic Student Problem* amounts to finding the best set of choices, with repetitions allowed, to maximize $G(.)$. In general, this is a non-convex combinatorial problem that is very hard to solve.

[3]The functional perspective taking here is made for clarity. As we will see in the experimental section the structure can be more general and consider non-euclidean spaces.

A. Submodular Costs

We will start by solving the optimization problem assuming that we can accurately evaluate the score function G. Solving Problem 1 is a combinatorial problem that can only be solved exactly for some specific classes of cost functions. For a *simple* example, we saw before that the simple greedy algorithm provides the optimal solution for the SSP. Intuitively, at each time step it selects the topic that gives a higher improvement in the score resulting that easier tasks are learned first. We now define more precisely the greedy algorithm, see Alg. 1. It is

Algorithm 1 Greedy optimization algorithm

```
D = {}
while G(D) < ε do
    x = arg max_{x∈X} (G({x} ∪ D) − G(D))
    D = D ∪ {x}
end while
return D
```

not difficult to show that this behavior is optimal only for very simple score functions, e.g. independent topics and convex functions. Our goal now is to generalize this heuristic and see in which more general conditions can the greedy algorithm still be (*quasi-*) optimal. Fortunately there is another class of functions where the SSP can be solved efficiently. A theorem from [32] says that for monotonic submodular functions, the value of the function for the set obtained with the greedy algorithm $G(D_g)$ is close, $(1 - 1/e)$, to the value of the optimal set $G(D_{OPT})$. This means that if we would solve the combinatorial problem, the solution we get with the greedy algorithm is at most 33% below the true optimal solution. Note that for the general case the greedy algorithm can be exponentially bad.

Submodular functions are functions that observe the diminishing returns property, i.e. if $B \subset A$ then $F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B)$. This means that if we select a point sooner in the learning process we will always improve learning more, or equal, than if the same point is chosen later. Note that the submodularity property we defined is on the relation of G with the number of samples and not on the particular values that were observed.

A first question to ask is how frequent are functions with this property. In many standard cases it is the case that the function to minimize is indeed monotonic submodular (see [33] for the analysis of different measures of uncertainty).

A first example can be multi-armed bandits [29] that, following the analogy of slot-machines, consists in learning the value of m different arms by pulling them. The problem has mostly been studied in the exploration-exploitation setting where the gains acquired during learning are also taken into account. In this setting the learner is tested after a learning period and it has either to declare what is the best arm [34] or the value of all the arms [35]. We are more interested in the learning problem and we consider that we just want to estimate the mean of each arm. Let's consider that each arm is described by a gaussian distribution with an unknown mean μ but known covariance τ^{-1}. If we assume a prior distribution on the mean value $N(\mu_0, \tau_0)$, after n samples we have the following distribution on the mean (see for instance [36]):

$$\begin{aligned} p(\mu|X) &\propto p(X|\mu)p(\mu) = N(\mu, \tau)N(\mu_0, \tau_0) \\ &= N\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right) \end{aligned}$$

Here the gaussian distribution is parameterized by mean and the inverse of the variance. Noting that the variance of the posterior indicates the confidence, and the expected error we have on the mean, we see that the learning curve evolves as: $M(n) = \frac{1}{n\tau+\tau_0}$. So we see that the function is monotonic in the number of samples we have. For the case of gaussian processes we see that for a new point x_*, the variance we have on the estimation is given by: $var(y_*) = K_{**} - K_*K^{-1}K_*$ where K is the kernel matrix computed on the dataset and y is the predicted observation. As K is always positive and the inverse of a positive definite matrix is also positive definite, the variance on the estimation is also monotonicly decreasing and, in this special case we do not even consider the observations themselves.

In other cases we cannot know exactly the confidence on the estimator but we can measure it using PAC-Bounds. This formalism defines, for any given error level, the probability of achieving that error taking into account different learning parameters. In our case we are interested on the dependency on the amount of samples. A first case generalizes the first example for the case where each arm provides an object among k classes. For this we can model each arm as a multinomial distribution and we know that we can learn the task by making a number of queries bounded by [37]: $B(\epsilon, \delta) = \frac{n}{8\epsilon^2} ln \frac{2n}{\delta}$. A more complex example is when each model is itself a reinforcement learning problem. The learning curve has been shown to be polynomial ([18], [37]). This means that the expected accuracy of the algorithm is always increasing with the number of samples. We see that in all the previous examples, the error depends solely on the amount of data and we have a decreasing function on the size of the dataset.

B. Beyond submodular costs

Unfortunately there are problems that are not submodular. These might be caused by: a) incorrectly defined target functions, b) biased learners or c) problems that are not fit for a greedy strategy.

One example is to consider directly the prediction error, $|y - y_p|$, and select regions that have a high error. This measure will not converge to zero as it is bounded by Kramer-Rao bound. It means that if for a region the noise is high then, as the prediction error will never go below this level, the algorithm will not explore everything. If instead a change in the prediction error is made, enough exploration is going to be made.

A problem that is more difficult to address particularly in complex, non-linear problems, is that the learners might have a bias and/or due to the online aspect of learning get locked too early in local minima. In this situation the task-learner

is no longer submodular. The problem is that an initial good sample might guide the learning process to sub-optimal local solutions. In this situation it is clear that the sequence of data points is relevant to the learning task and so a greedy approach might be exponentially bad.

IV. SSP: A SOLUTION

In this Section we show how (parts) of the Strategic Student problem can be addressed. The previous result tells us that the simple greedy algorithm will provide a solution that is close to optimal for the case of submodular functions thus solving the combinatorial aspect of the problem. The two main difficulties we have are: a) the case of non-submodular functions and b) the score function needs to be estimated empirically. Both problems require a stochastic approach to either assure that early low learning progress is not due to ill behaved score functions and to balance between exploration, to estimate the progress, and exploitation, to select the topic with higher learning progress.

In short the previous sections tells us that using simply the learning progress is only optimal in very particular conditions. Even for well behaved functions if their shape is not known then a greedy algorithm can behave badly.

Several bandit algorithms have been proposed that are distribution free, i.e. that do not depend on particular assumptions on the cost functions, called the *adversarial bandit* setting. To have an initial solution to the SSP we can rely on the EXP4 algorithm [29].

We will construct an expert that tracks the gain that each task is currently giving. As this signal is very noisy and time varying we will keep track of the reward signal with a first order filter. For choice a, if we receive reward r, we update the quality q of that choice accordingly: $q_a \leftarrow q_a + \eta\,(r - q_a)$, being η a tunable parameter. We compute the reward using an empirical estimation of $\hat{r} = G(D) - G(D \backslash \{x_a, y_a\})$. We can now use this learning progress estimators directly by transforming it into probabilities. Taking into account that the rewards can be negative we compute the probabilities distribution suggested by the greedy expert ξ_g as:

$$\xi_g(a) \quad = \quad \frac{e^{\beta(q_i - \min(q))}}{\sum_j e^{\beta(q_j - \min(q))}}$$

where $\xi_g(a)$ is the probability of selecting choice a.

Then we combine this expert with another uniform expert $\xi_u = \frac{\gamma}{m}$ to ensure a good exploration of all choices. We can now introduce the Strategic Bandit, show in Alg. 2, as a variant of the EXP4 presented by [29]. This algorithm computes online the weight w that it should give to each expert to ensure that it will converge to the best one. At each step it asks the advice vectors for each expert ξ_g and ξ_u that give the probability of selecting each choice. All the experts are combined, weighted by w. From this combination of experts results a probability distribution p on the available choices and the next choice is made by sampling proportionally to this distribution. By observing a new sample acquired with such choice the learning machines are updated and a reward function is computed. This reward will lead to a reweighting of the experts according to:

$$w_i \leftarrow w_i exp\left(\gamma \xi_i(a) \frac{r}{p(a)m}\right)$$

where the expert that gave higher probability to the choice that was actually made will be updated more strongly.

Algorithm 2 Strategic Bandit (SB)

Require: Set of topics C and choices a
Require: Initialize $D \leftarrow \emptyset$
Initialize $w_g = 1$ $w_u = 1$
Initialize experts: uniform $\xi_u = \frac{\gamma}{m}$ and greedy: $\xi_g = 0$
while *learning* **do**
  $p = w_g \xi_g + w_u \xi_u$
  Select choice a proportional to p
  Draw sample x_a using choice a
  Observe output $y_a \sim (C_a, x_a)$ using a and x_a
  $D = D \cup \{x_a, y_a\}$
  $r = \hat{G}(D) - \hat{G}(D \backslash \{x_a, y_a\})$
  $w_i \leftarrow w_i exp\left(\gamma \xi_i(a) \frac{r}{p(a)m}\right)$
  Update greedy expert:
  $q_a \leftarrow q_a + \eta\,(r - q_a)$
  $\xi_g(a) = \frac{e^{\beta(q_a - \min(q))}}{\sum_j e^{\beta(q_j - \min(q))}}$
end while

The algorithm requires selecting a sample x_a using a given choice a. That is when we choose a region/topic to sample from it is still necessary to pick a specific sample from there. For now we just assume that that sample is chosen randomly, but we could imagine to use an active learning approach as long as it is ensure a good behavior of G with the number of samples.

This approach shares several ideas with other approaches. The algorithms shown in [13], [38] were introduced to choose among different active learning methods in a classification task and also relies on the EXP4. Another work, presented in [11], was introduced to improve upon previous methods, e.g. [20], that were not robust to the noise in the progress estimation and uses the idea of a stochastic selection approach. In this presentation we are considering a very naive approach to compute the learning progress, other authors have considered efficient data structures to make such computation [11] and how empirical measures of progress can generalize exploration methods from RL [39], [40].

V. EXPERIMENTS

Although a general solution for the SSP problem is outside the scope of this paper, we present two examples showing that the proposed algorithm can address two problems that at first hand can be considered different. The solution proposed is used and compared against fixed and uniform exploration.

We want to learn simultaneously the dynamic models of different environments (topics). This can be understood directly as learning how to behave in different environments or consider that each topic is a different task/option in the same

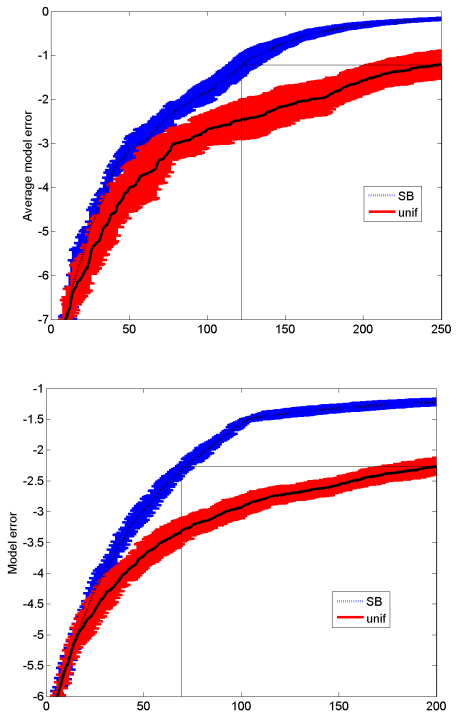

Fig. 2. Comparison between the use of the strategic bandit and a random uniform sampling approach. (Top) Multiple tasks. (Bottom) Single task with multiple learning approaches: uniform, $R-max$ and $\epsilon-greedy$. The Strategic Bandit is able to learn faster and to choose the best learning approach.

environment. We consider a *markov decision process* (MDP) defined by the state S and action A spaces, the actions being move up, down, left or right and stop. The dynamical model of the environment that defines the transitions probabilities $T^a_{ss'} = p(s'|s,a)$ is not known and has to be estimated. We consider that there is no external reward function. The cost function is: $D = \sum_i var\left(\hat{p}(y_i|x_i,a_i)\right)$.

Figure 2 shows the results for a problem with four environments with 4, 9, 16 and 49 states respectively. With $\gamma = 0.2$, $\alpha = 0.2$ and $\beta = 3$. The results show the behavior we expected where more complex tasks are explored later in the learning process. Our method provides the best results when compared with a random exploration.

We continue in the same model as before but now we consider that the choices are different strategies to gather data including: i) selecting actions uniformly, ii) following the $R-max$ exploration method [18] and iii) $\epsilon-greedy$ strategy. We then compare again a random learner with a learner that chooses the strategic bandit method. Figure 2 again shows that the strategic student is able to learn the task faster than using a random approach and, more importantly, to select the best exploration method available. In this case it converges to using always the $R-max$ exploration.

VI. Instantiations of the SSP

The main contribution of this paper is to present a general formal perspective in which to understand important aspects of the various kinds of active choices that a curious agent has to make during learning. We consider that many different approaches and algorithms can be described in the framework of the *strategic student problem*. In Table I we show different instantiations of SSP describing what are the choices and topics for different works in the literature. We do not claim to be exhaustive and more importantly we do not consider that we capture all the problems already addressed in this literature, for this we refer to Sections VI-C and VI-D. Many approaches can be viewed under the SSP framework as follows, and along dimensions spanning the number of topics and the number of learning methods upon which active choices can be made. We also identify dimensions that follow from the SSP framework, but were not studied in the previous sections and thus shall be the subject of future work.

A. Multi-Topic

The simple way to see the different topics is to consider that each topic is a different task and the goal is to learn, simultaneously, many tasks while selecting which task to spend resources on. This perspective of **multi-task learning** has been addressed in several ways. A typical classification task was considered on the works of [41], [42] where active learning methods are used to improve not only one task, but the overall quality of the different tasks. In sequential problems, as in robotics, several works considered how to learn different tasks simultaneously. When viewed in the SSP setting, each topic can be either a local predictive forward model [20], [43], an option [22], or a region in a parameterized goal/option space [11], i.e. a local inverse model (see also [44] and [45] for similar ideas). In these works the tasks cannot be sampled simultaneously and a decision has to be made. The choice is to decide which topic to explore/sample next. The authors used different variants of improvement quality to bias exploration.

B. Single-Topic, Multiple Learning Methods

The other big perspective on SSP is to consider that the choices are the different methods that can be used to learn from the task, in this case a single-task is often considered. This **learning how to learn** approach makes explicit that a learning problem is extremely dependent on the method to collect the data and the algorithm used to learn the task. For instance, in the work of [13], the authors want to select among different active learning strategies due to the known fact that different strategies yield different results in different tasks. In this setting there is only one topic but the learner has the choice of selecting *how to* sample from the process, i.e. using different active learning methods. The results show that such online selection of strategies give faster learning than relying always on the same strategy. Other approaches include the choice among the different teachers that are available to be observed [46] where some of them might not even be cooperative [47], or even choose between looking/asking for a teacher demonstration or doing self-exploration [48]. Some authors started to combine the how and what to explore in a visual recognition tasks from object manipulation [48]. Another approach considers the problem of having different

TABLE I
FORMULATION OF SEVERAL MACHINE LEARNING PROBLEMS AS A STRATEGIC STUDENT PROBLEM.

Choices	Topics	References
n Regions	n Functions	[11]
n Environment	n Environments	[22], [34], [35], [43]
Control or Task Space	Direct/Inverse Model	[11], [44], [45]
Exploration strategies	1 Environment	[13], [38], [51]
n Teachers	1 Environment	[46], [47]
{Teacher,self-exploration}	1 Function	[48]
n Representations	1 Environment	[49], [50]
Exploration strategies	n Objects	[52]

representation and then the representation that is given more progress might be used more frequently [49], [50].

C. Dynamic and Hierarchical Topics

The previous methods started to address central difficulties of the *strategic student problem* but there are still some other important problems that arise in practical instantiations of the problem. There are still many open issues that need to be addressed before agents are able to learn how to learn [12].

An important issue is the assumption about the structure of the global score function and, related, the relation between topics and choices. In the optimization setting, the work of [51] considers already known regions and that each choice corresponds to sample from that region. Yet, in real world applications, the repertoire of topics to choose from might not be provided initially or might evolve dynamically. The aforementioned works of [11], [20] considers initially a single topic/task (a prediction task in the former and a control task in the latter) but then automatically and continuously constructs new topics, by sub-dividing or joining previous existing topics. At each instant it considers the standard SSP and iteratively refines the definition of topics and actively samples from them. The work of [22] constructs the different topics, represented as options by assuming that a different topic must be created to reach certain sub-spaces on a large markov-decision process. Heuristics to start such process include relevant point detection [22], bottlenecks [53], [54] among others. These works can also be seen as the creation of a hierarchy of skills where the SSP guides which skill to learn at each time step.

The repertoire of topics can also be itself hierarchical. For example, the SAGG-RIAC architecture [11] hierarchically and actively makes choices at two levels: in a goal space, it chooses what topic/region to sample (i.e. which goal to set), and in a control space, it chooses which motor commands to sample to improve its know-how towards goals chosen at the higher level.

Other kinds of abstract topics, integrated in a hierarchical architecture, could be used. For example, one can think about the generalization machines being used and study how the learning process itself should evolve. For instance by learning how to learn learning strategies [12] or by searching for the simplest still unexplained problem [40], [55].

D. Planning Topics

When the score functions are not any more sub-modular, approaches to planning the optimal choice of topics based on ideas such as maximization of learning progress, can also be considered. As a generalization of exploration methods in reinforcement learning, such as [18], ideas have been suggested such as planning to be surprised [56] or the combination of empirical learning progress with visit counts [39].

VII. DISCUSSION

In this work we presented a general formal framework to understand aspects of exploration and learning when the learner has the choice to select among different tasks, learning algorithms or data acquisition methods. Initially, we framed a simple version of the SSP problem as a convex optimization problem and were able to show that the optimal solution is non-stationary. We discussed under what conditions the commonly used learning progress heuristics can be a quasi-optimal solution to the SSP problem.

When the learning progress cannot be measured easily and accurately or when the score function is not benign we had to rely on a stochastic approach that no longer deterministically chooses the most promising choice but instead chooses it probabilistically. This algorithm was motivated from well-established algorithms from the adversarial multi-armed bandit setting [13], [29], as well as from experimental investigations of intrinsic motivation systems [11], [30].

With this general formalism we presented several works from the literature in a common formalism. This unified view enables a better understanding of the different problems and the required properties of the algorithms used to address them. We discussed several components of previous methods that are not modeled in the current formalism and that will require a much harder theoretical study.

Several question arose and it is clear that a pure data-driven approach might not be enough to deal with all complexity of real systems. Artificial development [57] will require particular structures that will guide exploration and learning beyond what can be addressed by pure measures of learning progress. These mechanisms include maturational constraints [5], [6], [58], the development of intrinsic rewards [24], [59], pre-dispositions to detect meaningful salient events, among many other aspects.

REFERENCES

[1] M. Cakmak and M. Lopes, "Algorithmic and human teaching of sequential decision tasks," in *AAAI Conference on Artificial Intelligence (AAAI'12)*, Toronto, Canada, 2012.

[2] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–9, 1993.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *International Conference on Machine Learning (ICML '09)*, 2009.

[4] L. Berthouze and M. Lungarella, "Motor skill acquisition under environmental perturbations: On the necessity of alternate freezing and freeing of degrees of freedom," *Adaptive Behavior*, vol. 12, no. 1, pp. 47–64, 2004.

[5] A. Baranes and P. Oudeyer, "The interaction of maturational constraints and intrinsic motivations in active motor development," in *International Conference on Development and Learning (ICDL'11)*. IEEE, 2011.

[6] M. Lee, Q. Meng, and F. Chao, "Staged competence learning in developmental robotics," *Adaptive Behavior*, vol. 15, no. 3, pp. 241–255, 2007.

[7] M. Luciw, V. Graziano, M. Ring, and J. Schmidhuber, "Artificial curiosity with planning for autonomous perceptual and cognitive development," in *International Conference on Development and Learning (ICDL'11)*, 2011.

[8] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[9] R. Martinez-Cantin, M. Lopes, and L. Montesano, "Body schema acquisition through active learning," in *IEEE International Conference on Robotics and Automation (ICRA'10)*, Alaska, USA, 2010.

[10] L. Montesano and M. Lopes, "Active learning of visual descriptors for grasping using non-parametric smoothed beta distributions," *Robotics and Autonomous Systems*, no. 60(3), pp. 452–462, 2012.

[11] A. Baranes and P. Oudeyer, "Active learning of inverse models with intrinsically motivated goal exploration in robots," *Robotics and Autonomous Systems*, 2012.

[12] J. Schmidhuber, "On learning how to learn learning strategies," Fakultaet fuer Informatik, Technische Universitaet Muenchen, Tech. Rep. FKI-198-94, 1995.

[13] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *The Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.

[14] M. Lopes, F. S. Melo, and L. Montesano, "Active learning for reward estimation in inverse reinforcement learning," in *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'09)*, 2009.

[15] S. Dasgupta, "Two faces of active learning," *Theoretical computer science*, vol. 412, no. 19, pp. 1767–1781, 2011.

[16] R. Nowak, "The geometry of generalized binary search," *Information Theory, Transactions on*, vol. 57, no. 12, pp. 7893–7906, 2011.

[17] S. Thrun, "Exploration in active learning," *Handbook of Brain Science and Neural Networks*, pp. 381–384, 1995.

[18] R. Brafman and M. Tennenholtz, "R-max - a general polynomial time algorithm for near-optimal reinforcement learning," *The Journal of Machine Learning Research*, vol. 3, pp. 213–231, 2003.

[19] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers." in *From Animals to Animats: First International Conference on Simulation of Adaptive Behavior*, Cambridge, MA, USA, 1991, pp. 222 – 227.

[20] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[21] J. Schmidhuber, "Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts," *Connection Science*, vol. 18, no. 2, pp. 173 – 187, June 2006.

[22] A. Barto, S. Singh, and N. Chentanez, "Intrinsically motivated learning of hierarchical collections of skills," in *International Conference on development and learning (ICDL'04)*, San Diego, USA, 2004.

[23] D. Berlyne, *Conflict, arousal, and curiosity*. McGraw-Hill Book Company, 1960.

[24] G. Baldassarre, "What are intrinsic motivations? a biological perspective," in *International Conference on Development and Learning (ICDL'11)*, 2011.

[25] S. Singh, R. Lewis, and A. Barto, "Where do rewards come from?" in *Annual Conference of the Cognitive Science Society*, 2009.

[26] D. Angluin, "Queries and concept learning," *Machine Learning*, vol. 2, pp. 319–342, 1988.

[27] D. Golovin, A. Krause, and D. Ray, "Near-optimal bayesian active learning with noisy observations," in *Proc. Neural Information Processing Systems (NIPS)*, December 2010.

[28] D. Golovin and A. Krause, "Adaptive submodularity: A new approach to active learning and stochastic optimization," in *Proc. International Conference on Learning Theory (COLT)*, 2010.

[29] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2003.

[30] A. Baranès and P.-Y. Oudeyer, "R-iac: Robust intrinsically motivated exploration and active learning," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 3, pp. 155–169, 2009.

[31] N. Entwistle, "Promoting deep learning through teaching and assessment: conceptual frameworks and educational contexts," in *TLRP Conference*, Leicester, UK, 2000.

[32] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[33] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Uncertainty in AI*, 2005.

[34] Victor Gabillon, Alessandro Lazaric, Mohammad Ghavamzadeh, and S. Bubeck, "Multi-bandit best arm identification," in *Neural Information Processing Systems (NIPS'11)*, 2011.

[35] A. Carpentier, M. Ghavamzadeh, A. Lazaric, R. Munos, and P. Auer, "Upper confidence bounds algorithms for active learning in multi-armed bandits," in *Algorithmic Learning Theory*, 2011.

[36] C. M. Bishop, *Pattern recognition and machine learning*. Springer, New York, 2006.

[37] L. Li, M. Littman, T. Walsh, and A. Strehl, "Knows what it knows: a framework for self-aware learning," *Machine learning*, vol. 82, no. 3, pp. 399–443, 2011.

[38] M. Hoffman, E. Brochu, and N. de Freitas, "Portfolio allocation for bayesian optimization," in *Uncertainty in artificial intelligence*, 2011, pp. 327–336.

[39] T. Hester and P. Stone, "Intrinsically motivated model learning for a developing curious agent," in *AAMAS Workshop on Adaptive Learning Agents*, 2012.

[40] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer, "Exploration in model-based reinforcement learning by empirically estimating learning progress," in *Neural Information Processing Systems (NIPS'12)*, Tahoe, USA, 2012.

[41] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang, "Two-dimensional active learning for image classification," in *Computer Vision and Pattern Recognition (CVPR'08)*, 2008.

[42] R. Reichart, K. Tomanek, U. Hahn, and A. Rappoport, "Multi-task active learning for linguistic annotations," *ACL08*, 2008.

[43] P.-Y. Oudeyer, F. Kaplan, V. Hafner, and A. Whyte, "The playground experiment: Task-independent development of a curious robot," in *AAAI Spring Symposium on Developmental Robotics*, 2005, pp. 42–47.

[44] L. Jamone, L. Natale, K. Hashimoto, G. Sandini, and A. Takanishi, "Learning task space control through goal directed exploration," in *International Conference on Robotics and Biomimetics (ROBIO'11)*, 2011.

[45] M. Rolf, J. Steil, and M. Gienger, "Online goal babbling for rapid bootstrapping of inverse models in high dimensions," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, 2011.

[46] B. Price and C. Boutilier, "Accelerating reinforcement learning through implicit imitation," *J. Artificial Intelligence Research*, vol. 19, pp. 569–629, 2003.

[47] A. P. Shon, D. Verma, and R. P. N. Rao, "Active imitation learning," in *AAAI Conference on Artificial Intelligence (AAAI'07)*, 2007.

[48] S. Nguyen, A. Baranes, and P. Oudeyer, "Bootstrapping intrinsically motivated learning with human demonstration," in *International Conference on Development and Learning (ICDL'11)*, 2011.

[49] G. Konidaris and A. Barto, "Sensorimotor abstraction selection for efficient, autonomous robot skill acquisition," in *International Conference on Development and Learning (ICDL'08)*, 2008.

[50] O. A. Maillard, R. Munos, and D. Ryabko, "Selecting the state-representation in reinforcement learning," in *Advances in Neural Information Processing Systems*, 2011.

[51] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.

[52] S. M. Nguyen, S. Ivaldi, N. Lyubova, A. Droniou, D. Gérardeaux-Viret, D. Filliat, V. Padois, O. Sigaud, and P.-Y. Oudeyer, "Learning to recognize objects through curiosity-driven manipulation," in *under review*, 2013.

[53] A. McGovern and A. G. Barto, "Automatic discovery of subgoals in reinforcement learning using diverse density," in *International Conference on Machine Learning (ICML'01)*, San Francisco, CA, USA, 2001.

[54] O. Şimşek and A. G. Barto, "Using relative novelty to identify useful temporal abstractions in reinforcement learning," in *International Conference on Machine Learning*, 2004.

[55] J. Schmidhuber, "Powerplay: Training an increasingly general problem solver by continually searching for the simplest still unsolvable problem," http://arxiv.org/abs/1112.5309, Tech. Rep., 2011.

[56] Y. Sun, F. Gomez, and J. Schmidhuber, "Planning to be surprised: Optimal bayesian exploration in dynamic environments," *Artificial General Intelligence*, pp. 41–51, 2011.

[57] J. Elman, *Rethinking innateness: A connectionist perspective on development*. The MIT press, 1997, vol. 10.

[58] M. Lapeyre, O. Ly, and P. Oudeyer, "Maturational constraints for motor learning in high-dimensions: the case of biped walking," in *International Conference on Humanoid Robots (Humanoids'11)*, 2011, pp. 707–714.

[59] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg, "Intrinsically motivated reinforcement learning: an evolutionary perspective," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 2, 2010.