Gain-based Exploration: From Multi-armed Bandits to Partially Observable Environments

Bailu Si[1], J. Michael Herrmann[2], and Klaus Pawelzik[1]
[1]University of Bremen, Institute of Theoretical Neurophysics
Otto-Hahn-Allee 1, D-28359 Bremen, Germany
bailu,pawelzik@neuro.uni-bremen.de
[2]Göttingen University, Institute for Nonlinear Dynamics
Bernstein Center for Computational Neuroscience
Bunsenstr. 10, D-37073 Göttingen, Germany
michael@nld.ds.mpg.de

Abstract

We introduce gain-based policies for exploration in active learning problems. For exploration in multi-armed bandits with the knowledge of reward variances, an ideal gain-maximization exploration policy is described in a unified framework which also includes error-based and counter-based exploration. For realistic situations without prior knowledge of reward variances, we establish an upper bound on the gain function, resulting in a realistic gain-maximization exploration policy which achieves the optimal exploration asymptotically. Finally, we extend the gain-maximization exploration scheme towards partially observable environments. Approximating the environment by a set of local bandits, the agent actively selects its actions by maximizing discounted gain in learning local bandits. The resulting gain-based exploration not only outperforms random exploration, but also produces curiosity-driven behavior which is observed in natural agents.

1. Introduction

As a process of knowledge acquisition, exploration is crucially involved in a number of applications such as autonomous robots [1, 2, 3], human decision making [4], online convex optimization [5], and related problems in economics [6], such as project management and portfolio selection [7]. In learning tasks, e.g. classification, regression and clustering, exploration serves to actively select samples, so that a better overall performance is achieved on a finite data set [8, 9, 10]. The resulting learning paradigm implies active selection of training samples according to criteria based on partial information that is acquired during the learning process. The performance of an active data selection strategy is quantified by an objective function that is minimized during learning [11].

As a result of learning the agent is able to exploit domain knowledge, e.g. to maximize the expected reward. In reward maximization, exploration is necessary and any realistic solution of the reward maximization problem will involve exploration strategies. In this setting, other than trading exploitation for exploration, exploitation is often undertaken only after sufficient exploration is executed [12].

There are three classes of exploration strategies often used in active learning and reinforcement learning systems, namely, counter-based, error-based and gain-based exploration. In counter-based exploration, a register for each action is kept updated in order to evenly test all actions, so that redundancy is avoided. Error-based exploration prefers states with a large estimation error [13, 9]. This heuristic assumes that the larger the estimation error is, the more samples are needed. Finally, gain-based exploration strategies choose the data or action which imply maximal performance improvement, measured by some optimality measure [11]. For example, the entropy reduction [8, 14], variance reduction [15], or error reduction [16, 10] have been considered.

In this paper, we first analyze the optimal exploration in the generic setting of the multi-armed bandit [17] in Section 2. After a brief review of well-known counter-based and error-based exploration policies, a general gain-maximization exploration strategy is derived and is shown to unify three classes of policies minimizing different norms of the mean squared errors of reward estimations. The resulting ideal gain-maximization exploration policy can be achieved by a realistic policy without variance information. In Section 3 we generalize the gain-maximization exploration to partially observable environments. Section 4 presents results from numerical experiments which demonstrate the validity of the generalization. Finally, in Section 5 our work is related to previous results.

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE

2. Exploration in multi-armed bandits

The multi-armed bandit is a classic model to analyze exploration, because it captures the essential aspects of reinforcement learning and active learning, for example decision making under uncertainty, and exploration-exploitation trade-off [18]. Formally, a multi-armed bandit is a one-state Markov Decision Processes with the following elements:

- A set of actions (arms) $\mathcal{A} = \{1, 2, \cdots, A\}$.
- An action selection policy $\pi_t(a)$ specifying an action a_t at each time step.
- A family of distributions $\rho\,(r; a)$ from which reward r_t is drawn after performing action a_t.

In this paper, we assume $\rho\,(r; a)$ to be normal with mean $\mu(a)$ and variance $\sigma^2(a)$. We estimate the parameters by

$$\hat{\mu}_t(a) = \frac{1}{\tau(a)} \sum_{k=1}^{t} r_k \delta_{a,a_k}, \qquad \tau(a) \geq 1, \tag{1}$$

$$\hat{\sigma}_t^2(a) = \frac{1}{\tau(a)-1} \sum_{k=1}^{t} \left[r_k - \hat{\mu}_t(a)\right]^2 \delta_{a,a_k}, \quad \tau(a) \geq 2. \tag{2}$$

δ_{a,a_k} is the Kronecker delta, $\tau(a) = \sum_{k=1}^{t} \delta_{a,a_k}$ is the number of times a was selected, and $\sum_{a=1}^{n} \tau(a) = t$.

For the purpose of reward estimation, exploration is an adaptive sampling behavior to achieve the minimal overall estimation error for the given learning task (1). The agent has to find a policy $\pi_t(a)$ that minimizes the expected total mean squared error

$$E(t) = \sum_{a=1}^{A} \left\langle \left[\hat{\mu}_t(a) - \mu(a)\right]^2 \right\rangle = \sum_{a=1}^{A} \frac{\sigma^2(a)}{\tau(a)}. \tag{3}$$

Counter-based exploration and error-based exploration have been used in the policies for multi-armed bandits. For example, in the *least-taken policy* [19], the action which has been selected least is chosen next. In the interval estimation method [20, 13], actions with a larger reward estimation error receive more exploration. We call the exploration policy underling the interval estimation method *the z-interval exploration policy* due to the fact that the interval is constructed for a Normally distributed variable z.

2.1. Optimality by gain maximization

We will initially assume that the variances $\sigma^2(a)$ of the reward distributions are known. For this ideal condition, an optimal exploration policy can be derived from the gain-maximization principle. The realistic situation of unknown variance is treated below in section 2.2.

Motivated by the objective function $E(t)$ (3), we consider the β-norm of the mean estimation errors as a generalized performance measure

$$G(t) = \Big(\sum_{a=1}^{A} \frac{\sigma^{2\beta}(a)}{\tau^{\beta}(a)}\Big)^{\frac{1}{\beta}}, \ \beta \in [0, \infty]. \tag{4}$$

When performing action a, the error reduction or the negative gradient $-\frac{\partial G(t)}{\partial \tau(a)}$ is the gain obtained in this step

$$g_t(a) = -\frac{\partial G(t)}{\partial \tau(a)} = \Big(\sum_{b=1}^{A} \frac{\sigma^{2\beta}(b)}{\tau^{\beta}(b)}\Big)^{\frac{1}{\beta}-1} \frac{\sigma^{2\beta}(a)}{\tau^{\beta+1}(a)}. \tag{5}$$

Note that Eq. 5 is based on the assumption that actions are independent and sampling numbers $\tau(a)$ are continuous, which will be made throughout this Section.

Directly maximizing the gain $g_t(a)$ and neglecting the common term for all a, result in a general gain-maximization exploration strategy

$$a_t = \arg\max_a \frac{\sigma^{2\beta}(a)}{\tau^{\beta+1}(a)} = \arg\max_a \frac{\sigma^{\frac{2\beta}{\beta+1}}(a)}{\tau(a)}. \tag{6}$$

Under the equilibrium condition, the action selection indices $\frac{\sigma^{\frac{2\beta}{\beta+1}}(a)}{\tau(a)}$ are kept at the same level, i.e. $\frac{\sigma^{\frac{2\beta}{\beta+1}}(a)}{\tau(a)} = \frac{\sigma^{\frac{2\beta}{\beta+1}}(b)}{\tau(b)}$, $\forall a, b \in \mathcal{A}$, which leads to the solution of the sampling numbers

$$\tau(a) = t \frac{\sigma^{\frac{2\beta}{\beta+1}}(a)}{\sum_{b=1}^{A} \sigma^{\frac{2\beta}{\beta+1}}(b)}, \tag{7}$$

where t denotes the current time.

When the value of β is set to 0, the general gain-maximization exploration strategy is the least-taken policy. At $\beta \to \infty$, the resulting exploration policy is the z-interval policy. When β is equal to 1, the general gain-maximization exploration strategy is an optimal policy minimizing the total mean squared error $E(t)$, and we call it *the ideal gain-maximization exploration policy*

$$a_t = \arg\max_a \frac{\sigma^2(a)}{\tau^2(a)}. \tag{8}$$

Eq. 4 gives thus rise to a family of policies. It is, however, the policy 8 that minimizes the objective function $E(t)$ (3), which can be proved by solving the constrained minimization problem using the method of Lagrange multipliers

$$L \ = \ \sum_{a=1}^{A} \frac{\sigma^2(a)}{\tau(a)} + \lambda[\sum_{a=1}^{A} \tau(a) - t]. \tag{9}$$

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE

2.2. A policy from confidence intervals

In order to transfer the above results to practical situations, we have to apply an estimator $\hat{\sigma}_t^2(a)$ for the generally unknown variance $\sigma^2(a)$. Following the idea of optimism in the face of uncertainty [18], we replace $\sigma^2(a)$ by its upper bound $\frac{\tau(a)-1}{\chi^2_{\alpha,\tau(a)-1}}\hat{\sigma}_t^2(a)$ [21], leading to *the realistic gain-maximization exploration policy*, which has a similar form as (8)

$$a_t = \arg\max_a \frac{m_{\alpha,\tau(a)-1}}{\tau(a)}\hat{\sigma}_t(a). \tag{10}$$

Here we define $m^2_{\alpha,\tau(a)-1} = (\tau(a)-1)/\chi^2_{\alpha,\tau(a)-1}$.

The realistic gain-maximization exploration policy approaches the ideal gain-maximization exploration policy asymptotically, since $m_{\alpha,\tau(a)-1}$ approaches 1 from above as $\tau(a)$ towards infinity.

3. Exploration with partial observability

In order to study exploration in an autonomous agent or robot, we extend the gain-based exploration for multi-armed bandits to partially observable environments. First, we define a checkerboard maze as a benchmark model for robot exploration task [22].

3.1. Checkerboard maze

The checkerboard maze is a 2-D stochastic environment with x-y coordinate system. The environment is divided into a grid with unitary squares as cells, indexed by z. Each cell of the grid has an occupancy probability, indicating its probability of being occupied by some object. Fig. 1 shows one instance of the environments.

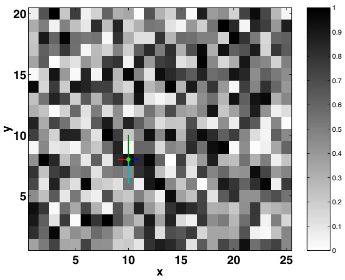


Figure 1. An agent with short-range sensors in a checkerboard maze.

The task of the agent is to localize itself and to learn a map of the environment by active exploration. We assume that the agent is allowed to go to any cell freely, neglecting the problem of object avoidance. The agent has four short-range sensors, detecting whether or not there are some objects in the four neighboring cells. Due to the stochastic nature of the environment, at each location z the agent observes one of the $S = 2^4 = 16$ binary sensory patterns denoted by $s \in \{1, \cdots, S\}$. The agent has five actions, namely to go left, right, forward, backward, and stay. The first four actions are subject to additive noises both in x and y directions independently sampled from a uniform distribution over the range $[-\sigma_M/2, \sigma_M/2]$.

3.2. Evaluating the prediction capability

The learning performance of the agent is quantified by its prediction capability. The sensory predictions of the agent are recorded by an ideal observer into a *prediction record* defined by

$$q_t(s|z) = \begin{cases} \frac{1}{S}, & z \text{ has never been visited} \\ \hat{p}_{\tau(z)}(s), & z \text{ was visited last time at } \tau(z) \end{cases} \tag{11}$$

where $\hat{p}_{\tau(z)}(s)$ is the agent's latest prediction in cell z when it was there at a time $\tau(z) \in [1, t]$. $q_t(s|z)$ is independent of the internal representations that the agent uses for its predictions. A quadratic loss function is used to measure the agent's prediction performance

$$E_{SQ}(t) = \sum_{s,z}\Big(p(s|z) - q_t(s|z)\Big)^2, \tag{12}$$

where $p(s|z)$ is the true observation probability, given by the probability of observing sensory pattern s at location z.

3.3. Learning a POMDP for exploration

The checkerboard maze is a partially observable environment. To explore efficiently, the agent actively learns a Partially Observable Markov Decision Process (POMDP) model for its sensory prediction and action selection.

More specifically, the POMDP model consists of the following elements:

- A set of states $i \in \{1, 2, \cdots, N\}$, representing places in the environment.
- A set of actions $a \in \{1, 2, \cdots, A\}$.
- A set of sensory observations $s \in \{1, 2, \cdots, S\}$.
- The agent's belief of its location $p_t(i) := p(i|s_1, \cdots, s_t, a_1, \cdots, a_{t-1})$, which is a probability distribution over states i.
- The sensor model $\hat{p}_t(s|i)$ is the internal estimation of the true observation probability $p(s|z)$. The sensor model is initialized as a uniform distribution, and is learned by the agent during its exploration in the environment. $\hat{p}_t(s|i)$ is a functional representation of the map of the environment.

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE

- Motion model $T^a(i,j) := \hat{p}(i|a,j)$ for each action a. It is the estimation of the true transition probability $p(i|a,j)$, which is the probability that the agent goes from state j to state i under action a. We assume that the agent has already learned the true transition probabilities with a similar gradient descent method presented in this section.

- Intrinsic reward signal $r_t \in R$ for the action a_t chosen by the agent at time t. r_t is measured by the gain in learning $\hat{p}_t(s|i)$ (cf. Eq. 24).

- Action preferences $\hat{Q}_t(a,i)$ which are given by the estimated utility of action a when the agent believes to be in state i.

At time step t, the agent selects a greedy action $a_t \in \{1, 2, \cdots, A\}$ according to the predicted preferences

$$\hat{Q}_t(a) = \sum_i \hat{Q}_t(a,i) p_t(i), \tag{13}$$

$$a_t = \arg\max_a \hat{Q}_t(a). \tag{14}$$

Then the internal belief is propagated by the motion model to track the agent's location

$$\begin{aligned} \hat{p}_{t+1}(i) &:= p(i|s_1, \cdots, s_t, a_1, \cdots, a_t) \\ &= \sum_j T^{a_t}(i,j) p_t(j). \end{aligned} \tag{15}$$

The sensory input at the new position is predicted by

$$\hat{p}_{t+1}(s) = \sum_i \hat{p}_t(s|i) \hat{p}_{t+1}(i). \tag{16}$$

As a result of action a_t, the agent reaches a new location, and perceives new sensory pattern s_{t+1}. The prediction loss is given by the square distance

$$\ell_{SQ}(t+1) = \frac{1}{2} \sum_s \Big(\hat{p}_{t+1}(s) - p_{t+1}(s)\Big)^2, \tag{17}$$

where $p_{t+1}(s) = \delta_{s,s_{t+1}}$ is the observed sensory distribution.

Combining Eq. 16 and 17, we find the gradient of the square loss with respect to $\hat{p}_t(s|i)$

$$\frac{\partial \ell_{SQ}(t+1)}{\partial \hat{p}_t(s|i)} = \Big(\hat{p}_{t+1}(s) - p_{t+1}(s)\Big) \hat{p}_{t+1}(i). \tag{18}$$

To minimize the prediction loss, the additive gradient descent rule is used (called the *additive rule* hereafter)

$$\hat{p}_{t+1}(s|i) = \hat{p}_t(s|i) - \eta \Big(\hat{p}_{t+1}(s) - p_{t+1}(s)\Big) \hat{p}_{t+1}(i), \tag{19}$$

where η is a fixed positive learning rate.

In order to optimize the learning speed, learning rates are decreased by summing the believes, giving rise to the *counting rule*

$$\begin{aligned} \hat{p}_{t+1}(s|i) &= \hat{p}_t(s|i) - \eta_t(i) \Big(\hat{p}_{t+1}(s) - p_{t+1}(s)\Big) \hat{p}_{t+1}(i) \\ \eta_t(i) &= \frac{\mu}{1 + \sum_{\tau=1}^{t+1} \hat{p}_\tau(i)} \end{aligned} \tag{20}$$

where μ is a positive learning rate.

Both the additive rule (19) and the counting rule (20) are self-normalizing. However, $\hat{p}_{t+1}(s|i)$ can become negative or larger than one if the learning rate is too large. When this happens, $\hat{p}_{t+1}(s|i)$ are projected to the boundary values 0 or 1 respectively and are renormalized in order to conserve the probability.

After adapting the sensor model, the agent updates its belief into a posterior distribution through Bayes' rule to include the newest information

$$\begin{aligned} p_{t+1}(i) &:= p(i|s_1, \cdots, s_t, s_{t+1}, a_1, \cdots, a_t) \\ &= \frac{\hat{p}_{t+1}(s_{t+1}|i) \hat{p}_{t+1}(i)}{\sum_j \hat{p}_{t+1}(s_{t+1}|j) \hat{p}_{t+1}(j)}. \end{aligned} \tag{21}$$

3.4. Gain as intrinsic reward

According to Eq. 8, the gain function is the ratio of the variance over the squared sampling number. For a Bernoulli distribution, the probability of success p is estimated by the proportion of successes in the total trials. The Wilson interval $[L(\hat{p}_n, n), U(\hat{p}_t, t)]$ bounds p with confidence level $(1-\alpha)$ [21]

$$U(\hat{p}_t, t) = \frac{\hat{p}_t + \frac{z^2_{\alpha/2}}{2t} + z_{\alpha/2} \sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{t} + \frac{z^2_{\alpha/2}}{4t^2}}}{1 + \frac{z^2_{\alpha/2}}{t}}, \tag{22}$$

$$L(\hat{p}_t, t) = \frac{\hat{p}_t + \frac{z^2_{\alpha/2}}{2t} - z_{\alpha/2} \sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{t} + \frac{z^2_{\alpha/2}}{4t^2}}}{1 + \frac{z^2_{\alpha/2}}{t}}. \tag{23}$$

Here t is the total number of trials. $z_{\alpha/2}$ denotes the variate value from the standard normal distribution such that the area to the right of the value is $\alpha/2$.

If the true variance $p(1-p)$ is replaced by its upper bound, the gain function is the intrinsic reward $r_t := \hat{g}_{t+1}$ for action a_t

$$r_t = \frac{\sum_s U\big(\hat{p}_{t+1}(s), \hat{\tau}_{t+1}\big)\Big(1 - L\big(\hat{p}_{t+1}(s), \hat{\tau}_{t+1}\big)\Big)}{\hat{\tau}^2_{t+1}}, \tag{24}$$

where $\hat{\tau}_{t+1} = \sum_i \tau_{t+1}(i) \hat{p}_{t+1}(i)$ is the averaged sampling number. $\tau_{t+1}(i) = \tau_t(i) + \hat{p}_{t+1}(i)$ is the counter of the believes and is initialized by 1.

Following a similar method as Q-learning [23], the agent adapts its action preferences by

$$\Delta \hat{Q}_t(a,i) = \varepsilon\Big(r_t + \gamma \max_a \hat{Q}_{t+1}(a) - \hat{Q}_t(a_t)\Big)e_t(a,i), \tag{25}$$

where $\varepsilon \in [0,1]$ is a positive learning rate and $\gamma \in [0,1)$ a discount rate. $\hat{Q}_{t+1}(a) = \sum_i \hat{Q}_t(a,i)p_{t+1}(i)$ is the predicted action preference for the new belief $p_{t+1}(i)$ according to Eq. 13. $e_t(i,a)$ is the eligibility trace, memorizing how recently the states are visited

$$e_t(a,i) = \gamma\lambda e_{t-1}(a,i) + p_t(i)\delta_{a,a_t}. \tag{26}$$

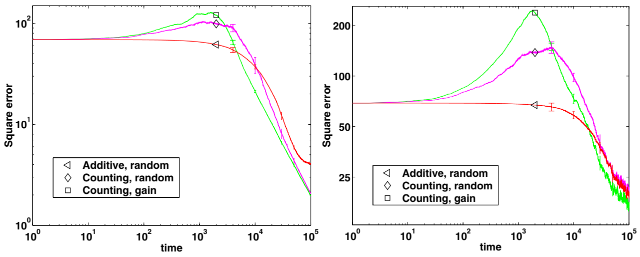
4. Numerical results

Gain-based exploration and random exploration with different learning rules are compared. Five checkerboard mazes are randomly generated by choosing the occupancy probabilities from the uniform distribution on the interval $[0,1]$. The agent explores each maze starting from two different initial positions, adopting random exploration or gain-based exploration. Without knowing the true observation probability, the agent starts with a prior sensor models (uniform distributions), and updates the sensor model during exploration according to the learning rule 19 or 20. The average performance of the ten runs is compared with respect to $E_{SQ}(t)$ (12).

Fig. 2(a) depicts the agent's performance when it has accurate odometry ($\sigma_M = 0$). The curves marked by triangles and diamonds correspond to the performance of random exploration when the additive rule ($\eta = 0.02$) or the counting rule ($\mu = 1$) is used respectively. In the beginning the counting rule is worse than the additive rule. This is because the counting rule has initially large learning rates. However, after the initial phase (about 10^4 steps), the counting rule outperforms the additive rule by about 27%. Asymptotically, the counting rule keeps a decaying trend in prediction error. As a contrast, the asymptotic performance of the additive rule saturates at a finite level due to its fixed learning rate.

The curve marked by squares in Fig. 2a is the performance of the gain-based exploration with the counting rule ($\mu = 1, \theta = 0.1, \gamma = 0.95, \lambda = 0.1$). Using the same counting rule, the agent improves its performance by about 25% by directing its exploration by the gain in learning.

When the agent is subject to 20% motion noise ($\sigma_M = 0.2$), the advantage of the gain-based exploration is still evident (Fig. 2b). The gain-based exploration with the counting rule (marked by squares, $\eta = 0.02, \mu = 3, \theta = 0.1, \gamma = 0.9, \lambda = 0.1$) is about 22% better than random exploration using the same counting rule (marked by diamonds). The gain-based exploration improve the learning performance of the counting rule, which otherwise has a similar performance as the additive rule under random exploration.



(a) $\sigma_M = 0$.

(b) $\sigma_M = 0.2$.

Figure 2. Average square error $E_{SQ}(t)$ as a function of learning time t according to Eq. 19 or 20 under random exploration (triangle and diamond markers) or gain-based exploration (square markers, Eqs. 24-26).

Interestingly, with the gain-based exploration, the agent produces an exploration behavior similar to the frontier-based exploration [24]. Fig. 3 shows some snapshots of the exploration process in the checkerboard maze shown in Fig. 1 with 20% motion noise ($\sigma_M = 0.2$). With gain-based exploration, the agent gradually expands its territory due to the fact that the gain in learning diminishes with respect to the visiting frequency. However, in random exploration, the exploration is not systematic, and the resulting exploration frontier is loosely connected and irregularly scattered[1].

5. Discussion

In this paper, we first study the exploration in the multi-armed bandit, which is a one-state Markov Decision Process. A general gain-maximization exploration strategy is derived, resulting in a unification of a new gain-maximization exploration policy, minimizing the mean squared error, and classical error-based and counter-based exploration policies, minimizing different norms of the mean squared errors.

The proposed gain-maximization exploration is extended to gain-based exploration in partially observable environments. The gain in learning Partially Observable Markov Decision Process is defined as the intrinsic reward for the agent to maximize. The resulting gain-based exploration produces curiosity-driven exploratory behavior.

Intrinsic reward has been considered previously to produce curiosity-driven behaviors. In [16, 10] curiosity is modelled to guide the agent in data or action selection. Singh et al. combined intrinsic motivation with extrinsic rewards of the traditional reinforcement learning framework, and demonstrated that the agent improved the efficiency in developing hierarchical skills [25]. The intrinsic reward de-

[1] videos are available from `http://www.neuro.uni-bremen.de/~bailu/video/index.html`

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE

IEEE
COMPUTER
SOCIETY

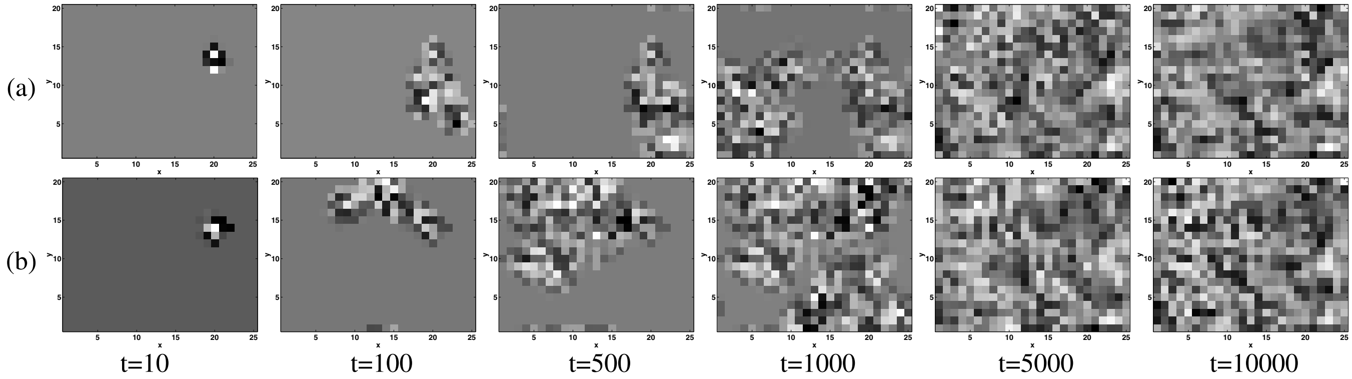


Figure 3. The evolution of the map learned by the agent with the counting rule ($\mu = 3, \sigma_M = 0.2$). (Row a) Random exploration; (Row b) Gain-based exploration.

fined here is the gain in local learning. The gain function is appropriate to model curiosity because it measures the efficiency in learning [10].

The work presented in [13] is also related to ours. The difference is that we use multi-armed bandits to approximate a POMDP, and the gain in learning local bandits is propagated. However, Meuleau et al. approximated a Markov Decision Process by local bandits, and used local estimation error to direct exploration, resulting in error-based exploration. Their work can be considered as a special case of the general approach presented here.

References

[1] A. Makarenko, S. Williams, F. Bourgault, and H. Durrant-Whyte. An experiment in integrated exploration. In *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2002.

[2] B. Si, K. Pawelzik, and J. M. Herrmann. Robot exploration by subjectively maximizing objective information gain. In *Proc. of the IEEE International Conference on Robotics and Biomimetics*, 2004.

[3] C. Stachniss, G. Grisetti, and W. Burgard. Information gain-based exploration using rao-blackwellized particle filters. In *Proc. of Robotics: Science and Systems (RSS)*, 2005.

[4] N. D. Daw, J. P. O'Doherty, P. Dayan, B. Seymour, and R. J. Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

[5] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proc. of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394, 2005.

[6] D. Bergemann and J. Valimaki. Bandit problems. In S. Durlauf and L. Blume, editors, *The New Palgrave Dictionary of Economics, 2nd ed*. Macmillan Press, London, 2006.

[7] C. H. Loch and S. Kavadias. Dynamic portfolio selection of npd programs using marginal returns. *Manage. Sci.*, 48(10):1227–1241, 2002.

[8] H. S. Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proc. of the Fifth Annual ACM Conference on Computational Learning Theory*, pages 287–294, 1992.

[9] S. B. Thrun and K. Möller. Active exploration in dynamic environments. In *Advances in Neural Information Processing Systems 4*, pages 531–538, 1992.

[10] J. M. Herrmann. Dynamical systems for predictive control of autonomous robots. *Theory in Biosciences*, 120(3-4):241–251, 2001.

[11] D. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, 1992.

[12] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(2/3):1079–1105, 2006.

[13] N. Meuleau and P. Bourgine. Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35(2):117–154, 1999.

[14] B. Anderson and A. Moore. Active learning for hidden markov models: objective functions and algorithms. In *Proc. of the 22nd International Conf. on Machine Learning*, pages 9–16, New York, NY, USA, 2005. ACM Press.

[15] D. A. Cohn. Neural network exploration using optimal experiment design. *Neural Network.*, 9(6):1071–1083, 1996.

[16] J. Schmidhuber. Curious model-building control systems. In *IEEE International Joint Conference on Neural Networks (5th IJCNN'91)*, volume 2, pages 1458–1463, Singapore, 1991. IEEE. CU Boulder.

[17] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[18] L. P. Kaelbling, M. L. Littman, and A. P. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[19] J. Vermorel and M. Mohri. Multi-armed bandit algorithms and empirical evaluation. In *European Conference on Machine Learning*, pages 437–448, Berlin, 2005. Springer.

[20] L. P. Kaelbling. *Learning in embedded systems*. MIT Press, Cambridge, MA, USA, 1993.

[21] J. E. Freund and R. E. Walpole. *Mathematical statistics (4th ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1986.

[22] K. P. Murphy. Bayesian map learning in dynamic environments. In *Advances in Neural Information Processing Systems 12*, 2000.

[23] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

[24] B. Yamauchi. A frontier-based approach for autonomous exploration. In *Proc. of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 146–151, Washington, DC, USA, 1997. IEEE Computer Society.

[25] S. P. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17*, 2005.

Third International Conference on Natural Computation (ICNC 2007)
0-7695-2875-9/07 $25.00 © 2007 IEEE