In *IEEE WCCI International Joint Conference on Neural Networks 2012, pp. 675–682. IEEE, 2012*
The original publication is available at *ieeexplore.ieee.org*

Towards Improved Theoretical Problems for Autonomous Discovery

Chris Lovell
Electronics and Computer Science
University of Southampton
Southampton, UK
Email: cjl3@ecs.soton.ac.uk

Steve Gunn
Electronics and Computer Science
University of Southampton
Southampton, UK
Email: srg@ecs.soton.ac.uk

***Abstract*—Active learning and experimental data acquisition address the same problems, understanding a system under investigation with as few resources as possible. However there are few instances where the theoretically principled techniques in active learning or sequential experimental design have been applied to managing data acquisition in physical experimentation. Partly this is due to fundamental differences between the problems investigated within active learning and the issues faced in much physical experimentation. From a previous study we conducted into autonomous experimentation, where we developed a system capable of automatically designing experiments and proposing potential hypotheses, we aim to investigate and highlight the differences between theoretical active learning and the requirements of experimentalists. We also propose an update of the multi-armed bandit problem that provides a theoretical problem more closely aligned to that found in physical experimentation. We believe that for active learning techniques to be used more widely as tools within physical experimentation, a greater focus of research has to be placed on theoretical problems that have assumptions more closely aligned to those found commonly within physical experimentation. Assumptions such as extremely limited resources, more so than typically considered in active learning problems, along with erroneous observations or noisy oracles, should become standard features of active learning problems, as in experimentation there are rarely enough resources available to be certain about the validity of the data obtained and the quality of the hypotheses produced.**

I. Introduction

In many discovery or experimentation problems, there are large numbers of possible experiments that could be performed, but the amount of experiments that can be performed is usually heavily restricted by some cost or resource availability. Take for example biological response characterisation, where there are a large number of potential chemical combinations that could be used to form an experiment, however the cost involved in each experiment can be large. Alternatively consider medical diagnosis, where a patient can undergo a wide range of different tests, but each test will have a monetary cost, along with a potential cost to patient health particularly in cancer diagnosis. Therefore, algorithms for minimising the resource usage whilst maximising the information gained are highly sort after.

In previous work we have investigated the creation of a system that can autonomously discover, which was tested with an experimental laboratory problem [1]. The purpose of this work was to develop a system that closes the loop between autonomous intelligent experiment selection and an automated experimentation platform, to allow for fully autonomous experimentation, as shown in Fig. 1. This work brought insight into discovery problems and highlighted discrepancies between current active learning research and problems faced within laboratory experimentation. In particular it was noted that whilst active learning provides mathematically rigorous techniques, they often overlook issues within physical experimentation, meaning ad-hoc techniques are often added to account for the limitations of the active learning methods. Investigations into autonomous discovery have previously had several successes using more ad-hoc techniques [2], [3], [4], although more mathematically rigorous techniques have also been utilised, albeit within an extremely small problem area with a large amount of previous knowledge [5].

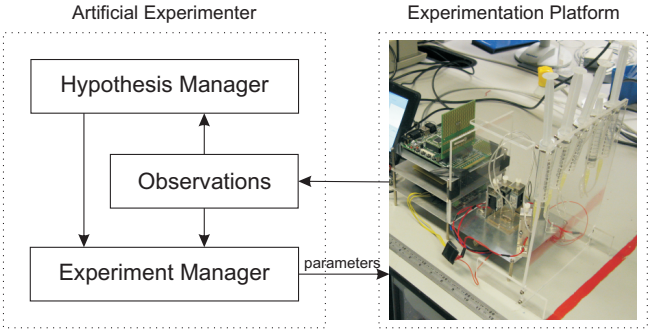

Fig. 1. Overview of autonomous experimentation. Algorithms for automatic hypothesis proposal and experiment selection interact with an automated experimentation platform. The platform shown on the right is a microfluidic system currently in development [6].

In this paper we will highlight the lessons learned from our work in developing an autonomous experimentation system. This will lead towards potential new abstract problems that could be used to benchmark future work on theoretical autonomous discovery.

II. Lessons Learned from Autonomous Experimentation

Autonomous experimentation is a union between a computational system that can design experiments and propose hypotheses, with an automated laboratory platform. In previous work we investigated how response characterisation of biological systems could be conducted through autonomous experimentation [1], [7]. The approach taken was to investigate the computational decision making aspect of the problem, by understanding and then mimicking how a successful human experimenter makes decisions about the experiments to perform. Whilst this approach led us to investigate the philosophical and practical considerations of experimentation, in a vein similar to early work on computational scientific discovery [8], we also aimed to include mathematically sound components to the design of the algorithms. In doing so we were able to determine ways to improve upon concepts contained within computational scientific discovery by applying more mathematical grounding to them, whilst also identifying the limitations of current active learning research and its use within physical experimentation. In the following we consider the limitations that were identified to exist within active learning research.

A. Limited Resources

Experimentation is restricted by the resources available. Typically these resources are extremely small in comparison to the dimensionality of the parameter spaces that can be explored. In active and sequential learning, the problem of learning from limited numbers of data points is addressed, however generally the number of experiments used are not as low as may be expected in physical experimentation.

Take for example the examination of sequential experimental design techniques through the multi-armed bandit problem [9]. Generally the multi-armed bandit problem is considered where the number of experiments performed, or levers pulled, may be many times larger than the total number of unique experiments possible. Some of the best techniques, such as upper confidence bounds, rely on performing each experiment once to build an initial model of the rewards available, then subsequent experiments are used to obtain the highest reward [10]. In many physical systems however, the number of experiments available will not be so vast. In fact they are likely to be many times smaller than the total number of possible experiments. Therefore the limited number of experiments brings with it a large amount of uncertainty, as there is no possible way of performing all experiments.

More concrete uses of active learning within particular problems have also not adequately addressed the high cost of experimentation. A different investigation into autonomous experimentation used active learning to minimise the cost required to fill in the gaps of understanding within a limited problem domain, however it disregarded the high cost required in experimentally obtaining the vast amount of prior information to initialise the system [11]. Whilst another investigation considered active learning within regression problems similar to how experiments would be chosen within response characterisation experiments, however the techniques were evaluated over several hundred experiments per parameter dimension [12], far beyond that realistically available in much physical experimentation.

B. Experimental Noise and Erroneous Observations

Very few experimental domains exist where observations are obtained noise free. In most cases, performing the same experiment will result in an observation that is slightly different. Often this noise can be considered as an adjustment on the true behaviour by an additive noise model, for example Gaussian or some other known noise distribution. Additionally, there is the notion of erroneous observations. Erroneous observations are the results from experiments where something undetectable goes wrong. For example the reactants are contaminated or there is an equipment failure, which causes a different experiment to be performed than was requested. In these situations the observation returned is unrepresentative of the true behaviour you would expect to see and can be described as being erroneous. However, it is important to note that not all experiments will result in erroneous observations, instead only a minority of experiments will be erroneous.

In a regression problem, we can consider the effect of such noise in the following equation:

$$y = f(x) + \epsilon + \phi \tag{1}$$

where the observation obtained is adjusted by standard experimental noise, ϵ, and some, but not all experiments, may have additional adjustments through experimental error ϕ. Whilst ϵ can generally be considered as a Gaussian distribution, a general distribution for ϕ is not known, but could be considered as a normal distribution with large standard deviation or non-zero mean. Colloquially, ϕ may be considered as shock-noise, which provides a sharp adjustment to the actual observation.

Erroneous observations would generally appear as outliers in regression problems. By assuming there are two possible distributions of distortions being applied to an observation, techniques such as a robust Gaussian process could be applied to effectively ignore the erroneous observations [13, Ch. 5]. However, the limited resources mean that there will be very few observations, making outlier identification difficult. Additionally, with only a small number of observations available, the uncertainty in the model space will be large, meaning that if an outlier could be identified, it would be unclear if the outlier is due to an erroneous observation or due to the prediction of the hypothesis being incorrect. Therefore anything that is suspected to be an erroneous observation needs to be examined to determine if it is erroneous or whether the current hypothesis is wrong. The uncertainties presented by the combination of limited resources and erroneous observations may best be handled by ensemble based approaches [1].

Query-by-committee is an ensemble approach to determining the most likely hypothesis [14], which is similar in how philosophers of science would argue that multiple hypotheses should be considered in experimentation to ensure a range

of different ideas are kept in consideration [15]. In a manner similar to falsification [16], the ensemble of hypotheses are then used to determine experiments by selecting where the committee disagrees the most. However, query-by-committee considers experimental observations to be noise-free [14]. By assuming observations to be noise free, query-by-committee does not build alternate hypotheses that actively question the validity of observations, instead hypotheses are generally created randomly. Additionally, query-by-committee does not refine hypotheses that are weakened by an experiment to propose better hypotheses, as is required for falsification. These two problems mean that first obtaining accurate hypotheses will be slow if all adjustments to the hypotheses occur randomly, whilst secondly the technique does not match well with how a scientist may consider hypotheses within a lab meaning that experimenters may not trust or value their use.

There have been a small number of investigations into active learning with noisy oracles, or where the labels obtained may be inaccurate, however the assumptions made of the oracles are not the same as that found in experimentation, where the errors have no well defined occurrence mechanism [17].

C. Incomplete Model Space

A Popperian view of experimentation is that the true hypothesis can never be found and that there will always be improvements that can be made. In many experimentation problems this view is easily demonstrated, as the limited resources and erroneous observations ensure that there will be a high amount of uncertainty within any hypothesis developed. However, in many active learning problems, the possible classes of outcomes are already known and is often reduced to a binary problem [14], [18]. Whilst an autonomous experimentation machine has been developed that uses active learning for classification where the available model space is known [5], a vast amount of prior information was required and the quality of the discoveries made were subject to error if there were any mistakes with the prior assumptions.

As the nature of discovery is to find things that were not known before experimentation began, there will be periods where a representative hypothesis will not exist within those hypotheses under consideration. Take for example a case in response characterisation, where only one experiment in a particular region of the parameter space has been performed, which yielded an erroneous observation. With only the one experiment in that region, using a single distribution to predict the response across the parameter space would state the observation to be indicative of the true underlying behaviour. However, as the observation is erroneous, the prediction is incorrect and the true hypothesis is not in consideration. This is a problem not addressed in core active learning research, where there is often an assumption that either the current distribution can be shrunk down to the target model or that the true hypothesis exists within the set of possible hypotheses [19], [20]. Essentially these techniques can be thought of as exploitation only, as they use the models and data available to repeatedly identify the differences to determine which of the hypotheses are the most likely. Such techniques will perform poorly if a mistake has been made in assuming the validity of the observations and the set of hypotheses under consideration does not contain a representative hypothesis. Instead techniques are required to also explore the parameter space to determine if there are features of the behaviour or system being investigated that are not captured by the hypotheses. Therefore a suitable trade-off between exploration and exploitation is required, where many existing techniques have been devised through investigating the multi-armed bandit problem and as such are designed to work on performing much larger numbers of experiments than will be typically available.

D. Experimentation is not Always Classification

A smaller consideration for making active learning techniques more accepted within physical experimentation is the class of problem that they address. Active learning considers mostly classification problems [17] and there have been examples of active learning in laboratory problems [18], [5]. However, many problems within laboratory discovery are not classification problems and bring with them their own set of additional issues, such as an incomplete model space as discussed above. Another problem is that few physical experimenters will have the mathematical background to take a solution considered in classification and apply it to regression. By ensuring problems are addressed that match those found in experimentation, there may be a wider acceptance and understanding of the mathematically principled techniques for experiment selection.

III. Towards a Better Framework for Discovery

From the lessons learned in autonomous experimentation, we consider a new abstract problem for active learning. For simplicity of the description we consider learning in a discovery problem using an ensemble of hypotheses, which was how our previous work considered the problem and how experimenters and philosophers of science would consider the problem. Although the translation to a system using a distribution based single hypothesis could be made. Here we focus on the problem of experiment selection and in particular the trade-off between exploration and exploitation within a discovery system.

The multi-armed bandit problem has provided a platform for understanding sequential experimental design. The problem consists of a number of different arms, or experiments, which can be performed to obtain some reward. In the original multi-armed bandit problem the rewards obtained from a lever were normally distributed amongst some predefined mean and standard deviation for that lever [9]. Although the multi-armed bandit problem has been extended to allow alterations such as the rewards at the levers changing over time [10], or to allow the rewards at each lever to be dependent on neighbouring levers [21]. The problem has been used to develop theoretical solutions that can then be applied to practical problems. For our abstract depiction of experiment selection in discovery, we believe the multi-armed bandit problem can be used as

the basis. The advantage of using the multi-armed bandit as a basis for this problem, is that the problem is already widely accepted as a means to understanding experimental design, and also as experiments are independent to each other, it allows for the dimensionality of multi-parameter discovery to be reduced down to a single parameter, the choice of experiment.

To enable the extension we consider how the multi-armed bandit problem matches with a discovery problem. First we consider the reward metric. In the multi-armed bandit problem the reward obtained is notionally defined as a monetary reward. Whilst in a discovery problem there is no direct monetary reward to performing an experiment, rather each experiment will provide information that can be used to discriminate between existing hypotheses and allow improved predictions of the behaviour under investigation to be made. Therefore the reward obtained will be the information the experiment provides. In a discovery system the expected information can be predicted, by either determining the degree to which hypotheses disagree with each other within a multiple hypotheses approach, or measuring the uncertainty or error bar within a single hypothesis. The expected information will be maximal where the uncertainty is greatest, which in a multiple hypotheses system could be considered as the maximum disagreement between good hypotheses. However, the actual reward obtained may be higher in some instances where the hypotheses are incorrect in their prediction or where an erroneous observation is obtained. In the autonomous experimentation problem we considered previously, a key part of the problem was to decide when to exploit the information held within the hypotheses, to evaluate them and differentiate between them, with when to explore the parameter space to discover new features of the behaviour not yet discovered. With information as the reward, the proposed framework for discovery captures this trade-off.

Next we consider how this reward changes over time. If we repeat the same experiment multiple times and obtain the same result, it is clear that the information the experiment provides will decrease, as we would expect to be able to predict the result accurately by the later experiments. Therefore the reward an experiment provides should generally decrease on subsequent performances of that experiment. However, there are instances whereby the information may increase on a repeated performance of the experiment. For example, consider the case where an observation disagrees with the consensus of the hypotheses. This disagreement will be due to either none of the hypotheses being suitable, or the observation being erroneous. This observation will provide some information but a subsequent observation will provide more information through either confirming the observation to be true and the hypotheses false, or by confirming the observation to be erroneous. Therefore we consider the reward obtained for a particular experiment x, to be relative to the last reward obtained for that experiment. The adjustment in information obtained is handled through a simple approximation β, which is a variable scalar modelled through some distribution that allows the next reward to be higher or lower than the current reward. The reward that will be obtained the next time an experiment is performed, $I_{\bar{t}+1}(x)$, is the reward obtained when the experiment was last performed, $I_{\bar{t}}(x)$, scaled by β, and hard bounded between zero and one:

$$I_{\bar{t}+1}(x) = \begin{cases} \beta I_{\bar{t}}(x) & \text{if } 0 \leq \beta I_{\bar{t}}(x) \leq 1, \\ 1 & \text{if } \beta I_{\bar{t}}(x) > 1, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

where time $\bar{t}$ is independent for each unique experiment parameter, to demonstrate how the information obtained for an experiment adjusts based on the last time that particular experiment was performed. There are open choices for choosing β, here we choose a normal distribution for simplicity, with $\mu = 0.8$ and $\sigma^2 = 0.4$, to ensure that the information will decrease in most instances but still has the ability to increase. The maximum and minimum values permitted from the normal distribution are hard limited by the requirement that $0 \leq I_{\bar{t}+1}(x) \leq 1$.

Next we consider how the information available is initially distributed across the possible experiments. In experimentation not all experiments will provide information. For example if all of the hypotheses predict the observation obtained for a particular experiment, then the experiment did not provide any additional information. However, the percentage of experiments that will provide information cannot be generalised to a particular parameter setting. Instead it would be of interest to examine the effect of adjusting the proportion of experiments that provide information with respect to the total number of possible experiments. Therefore a range of different percentages of experiments with information initially available should be tested, where the information is set arbitrarily high for some and arbitrarily low for others. Experiments capable of providing information should be randomly distributed through the possible experiment parameters.

Finally we consider the number of experiments available. Typically the number of experiments allowed will be low, however the exact value may depend on the cost of resources and dimensionality of the problem. However, we can assume an absolute maximum number of experiments to be the number of unique possible experiments, $|X|$, which can be performed, although in reality the number will generally be far smaller. The number of unique possible experiments can be considered as a discrete set, as equipment used in physical experimentation to perform the experiments will typically have limited precision. Therefore each trial of the problem will occur over $|X|$ experiments, with the performance of the techniques over time being considered in the evaluation, instead of simply performance after $|X|$ experiments, to account for situations where techniques have different rates of increase in performance over time.

To summarise the problem, we have designed the abstraction to meet the following issues:

- The reward is the information provided by the experiment.
- Not all experiments will provide information, it may be only a small minority of experiments that do.

- Repeating an experiment will generally yield lower rewards on the repeat experiments
- Sometimes erroneous observations will occur that may make an experiment appear to provide more information than it really does. In these cases the reward for that experiment will statistically drop rapidly on subsequent trials.
- Resources will be extremely limited, generally not enough to perform all experiments.

A. Evaluation

To evaluate the techniques there are several measures that can be made. First the cumulative reward, ω, obtained:

$$\omega = \sum_{t=1}^{T} \hat{i}_t \tag{3}$$

where T are the number of time steps performed and $\hat{i}_t$ is the actual reward obtained at time t. Second the regret, ρ, can be measured between the highest available reward and the selected reward at each time step:

$$\rho = \sum_{t=1}^{T} \left(\max_{x \in X} \{I_t(x)\} - \hat{i}_t \right) \tag{4}$$

where $\max_{x \in X}\{I_t(x)\}$ is the maximum reward possible at time t, and X is the set of possible experiments. Finally the mean of the actual reward obtained at each time step, I_t, over the repeated trials can be taken.

IV. Initial Techniques Used

A. Relative Information Gain Switching

In the previous work developing an artificial experimenter, the information obtained by an experiment was captured as the information gain between the confidences of the ensemble of hypotheses under consideration, before and after an experiment was performed [7]. The information measure, was captured as the KL-divergence:

$$\dot{I} = \sum_{i} C(h_i) \log \frac{C(h_i)}{C'(h_i)} \tag{5}$$

where $C(h_i)$ was the confidence of hypothesis i before the experiment was performed and $C'(h_i)$ was the confidence of the hypothesis after the experiment was performed. The motivation for using the KL-divergence in this way came from work by Itti and Baldi that produced a quantification of surprise based around the KL-divergence [22]. Having a quantification for surprise was desirable in developing an artificial experimenter, as surprise had been expressed in previous investigations into how successful human experimenters conduct experiments [23], [24].

In our previous work, the confidence was the likelihood measure of whether the data obtained during the experimentation agreed with the predictions for each hypothesis. As the prior and posterior distributions were not normalised and potentially different, the result of the KL-divergence could be negative. A negative information gain stated that the last experiment provided no new information to the hypotheses, whilst a positive value stated the experiment did provide new information as overall the observation disagreed with the most likely hypotheses at that time. In the referenced work, this information gain was equated to a notion of surprise, to allow it to fit within a framework of building machine learning techniques that mimic how human experimenters perceive the data they obtain. The most successful technique demonstrated in that study managed the exploration–exploitation trade-off by exploiting when the last experiment was surprising, or in other words when the information gain was increasing. The technique explored when the experiment was not surprising, or when the information gain was decreasing.

This leads to the first technique considered, where the next experiment to be performed, x^*_{t+1} is chosen as:

$$x^*_{t+1} = \begin{cases} \max_{x \in X} \hat{I}(x) & \text{if } I_t - I_{t-1} > 0, \\ \text{random(X)} & \text{otherwise} \end{cases} \tag{6}$$

where X is the set of possible experiments, $\hat{I}(x)$ is the prediction of the reward that will be obtained by performing experiment x, and I_t is the information obtained a time t. In words this method says, if the last experiment provided more information than the previous, exploit by choosing the experiment with the highest predicted information gain, otherwise explore.

B. Repeating Relative Information Gain Switching

The above selection method is a direct translation from previous work. A potential downside of this approach is that the experiment it chooses may not be the same as the last experiment performed that increased the relative information gain. In this second strategy we consider repeating an experiment if it increases the relative information gain:

$$x^*_{t+1} = \begin{cases} x^*_t & \text{if } I_t - I_{t-1} > 0, \\ \text{random(X)} & \text{otherwise} \end{cases} \tag{7}$$

C. Baseline Strategies

For reference, we provide several baseline strategies: random experiment selection; ϵ-first selection; and performing each experiment once. The ϵ-first selection will perform a number of initial exploration experiments, then perform greedy selection thereafter. Greedy selection performs where the highest predicted reward occurs:

$$x^*_{t+1} = \max_{x \in X} \hat{I}(x) \tag{8}$$

Performing each experiment once is provided as the absolute baseline strategy and is the first stage of upper confidence bound techniques, however it should be noted that the evaluation of the techniques should also occur when the number of experiments performed is far below the total number of different experiments.

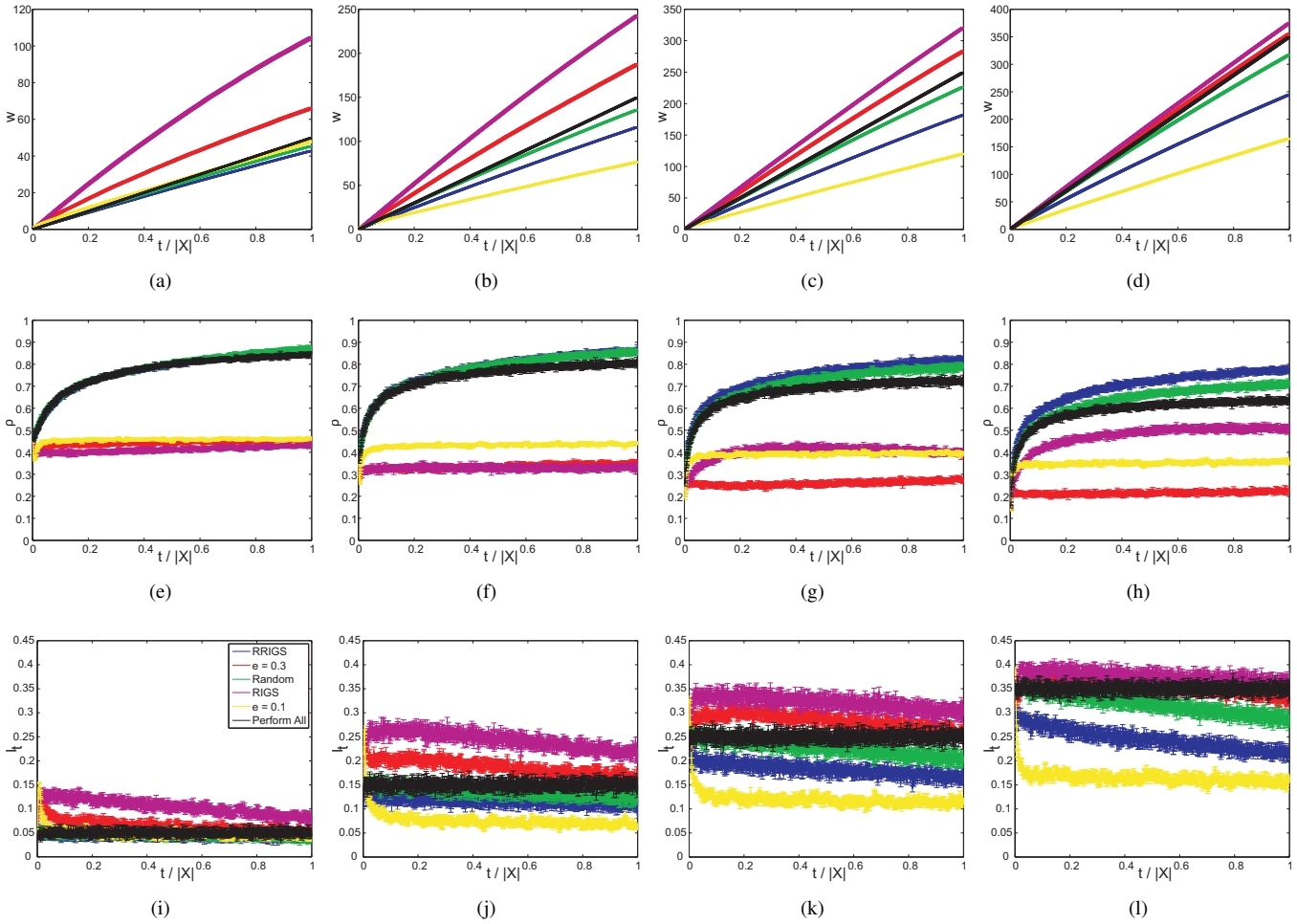

Fig. 2. Performance of techniques tested in terms of their mean cumulative reward ω, regret ρ and reward for each experiment I_t. In (a–d), mean total reward ω over proportion of experiments performed ($t/|X|$) when 10%, 30%, 50% and 70% of the experiments are initially interesting respectively. In (e–h), the regret ρ for those same trials is shown. In (i–l), the mean of the actual reward collected at each time step is shown for those same trials. The techniques shown are: Perform each experiment (Perform All) – black; Random – green; ϵ-greedy ($\epsilon = 0.1$) – yellow; ϵ-greedy ($\epsilon = 0.3$) – red; Relative information gain switching (RIGS)– magenta; Repeating relative information gain switching (RRIGS) – dark blue;

V. Preliminary Results

To evaluate the different techniques, the three evaluation methods described in Section III-A were used. Additionally the proportion of experiments that initially provided information was varied, as the amount of information available, or discoveries possible, in a discovery problem is unlikely to be known a priori and will vary between different discovery problems. This variation meant we were not fixed to an invalid a priori assumption within our framework problem. Finally the performance of the techniques was monitored over the number of experiments performed, with a range of zero to the number of experiment performed being equal to the number of unique experiments available.

Presented here are the results for when 10%, 30%, 50% and 70% of the possible experiments, X, initially yield a high amount of information. Each technique was tested over 1000 trials, where each trial had a maximum number of 1000 experiments that could be afforded, $n = 1000$, and there were 1000 different experiments possible, $|X| = 1000$. The experiments that would provide a high amount of information were randomly selected, with each being given an arbitrarily high initial reward of $I(x) = 0.5$, which allows future rewards for that experiment to grow or shrink. All other experiments were set with zero information reward available, which would remain zero throughout the trial. Throughout, the predicted reward $\hat{I}(x)$ is the value of the last reward obtained, or zero if not previously performed. In Fig. 2, the total reward and regret are shown over the proportion of experiments performed.

First we consider a general overview of the results shown. When the number of information rewarding experiments are low, the passive baseline techniques perform poorly in terms of total reward, except for the ϵ-greedy technique with $\epsilon = 0.3$. However, as the proportion of experiments that provide a reward increases, the *random* and *perform each experiment* techniques perform similarly to the more effective active strategies. This is interesting because the problems where the number of informative experiments available is low in proportion to the total number of experiments that can be afforded,

would appear to be the hardest and arguably most realistic problems. These results demonstrate that active learning is able to provide an advantage over more passive approaches and confirms that the techniques applied in the previous autonomous discovery work would provide an advantage over more baseline techniques. Interestingly, in these results the rate of increase on the actual reward obtained appears to remain largely constant across t.

Throughout, the regret for the two ϵ-greedy strategies is lower than the majority of the alternate strategies, but their cumulative rewards do not illustrate them to be the most rewarding strategies. This is due to how the reward for an experiment can increase and decrease over repeated experiments. In the ϵ-greedy techniques the majority of the experiments will be exploitation, which means that when a rewarding experiment is found, the experiment will most likely be repeated immediately, obtaining all of the reward possible for that experiment. As rewarding experiments have an initial reward of 0.5 and reward will tend to zero over repeated performances, the most rewarding experiment available will generally have a value no greater than 0.5 for these strategies. However, in the alternate strategies, such as *random*, a rewarding experiment may be found but not immediately repeated. As $I_{t+1}(x)$ can be larger than $I_t(x)$, the maximum reward available at any particular time may grow beyond 0.5 for those alternate strategies. This means that the mean regret can be higher for those alternate techniques, even if the mean reward obtained is also higher, as can be seen in Fig. 2 (e–h) and (i–l). Therefore, future studies would need to ensure that regret is not the sole evaluator used within this problem.

The ϵ-greedy strategies have a reasonably constant regret, as any rewarding experiments are immediately repeated, preventing the rate from increasing, whilst not performing enough exploration to identify all of the experiments with rewards, which prevents the regret from decreasing. The *relative information gain switching* acts in a similar manner of immediately repeating the most rewarding experiment, except that it performs a larger amount of exploration, caused by the strategy exploring each time the information obtained is less than was obtained in the previous experiment, then the technique reverts to exploitation when the information obtained increases over the previous experiment. Whilst the other strategies all identify rewarding experiments without repeating them to reduce their information available. The *repeating relative information gain switching* performs far worse than the *relative information gain switching*, as the repeating version will stop repeating a particular experiment as soon as the reward decreases and not return to that experiment, even if the predicted reward for it is the highest, until it randomly selects the experiment again through exploration.

Throughout, the relative information gain switching (RIGS) technique outperforms the alternate techniques, as it did in our previous work on physical automated discovery. The benefit this technique provides can be seen most clearly in the actual reward obtained evaluators (ω and I_t) shown in Fig. 2(a–d) and (i–l). An interesting new insight into the RIGS technique made possible by this new framework, is its performance compared to other techniques with respect to the proportion of experiments that yield informative and rewarding observations. The results indicate that as the proportion of experiments that are rewarding decreases, the benefit of the RIGS technique over alternate and baseline techniques increases. In other words, as the problem becomes harder, our active learning technique becomes more beneficial.

VI. Conclusion

Autonomous discovery and learning are important and growing fields of interest. They provide challenges in machine and active learning that have not yet received appropriate attention and can lead to solutions that are widely applicable in many domains of discovery. In this work we have considered the current mismatch between active learning solutions and the physical experimentation domains that would most benefit from them. The biggest mismatch is the issue of resources. Whilst most active learning problems consider learning from smaller datasets, the size of those datasets are still considerably larger than what would be able to be provided for most physical experimentation. The second mismatch is the issue of erroneous observations. Considerations in active learning have been made for noisy oracles, however their usage in problems is not that wide spread. In physical experimentation, errors occur regularly enough to warrant the consideration of erroneous observations to be a more mainstream problem when considering active learning problems for data acquisition.

Additionally we have proposed an adjustment to the multi-armed bandit problem, to provide a theoretical problem that is more similar to experiment selection within physical experimentation than existing problems provide. We have de-parameterised the problem of discovery within high dimensional parameter spaces, to address the problems of having very few resources, very large numbers of possible experiments, and limited number of experiments that will return a reward, or useful information in a physical experiment. The issue of erroneous observations are captured within the update on the reward, where in some instances multiple repetitions of an experiment will see the reward drop to zero very quickly, as would be seen by repeating an erroneous experiment and not obtaining the error again. The proposed problem has demonstrated that a translation of the technique used within a previous autonomous discovery system to successfully characterise biological systems [7], outperforms a range of baseline techniques. However, we would expect that new techniques will be able to outperform this in the theoretical problem. Additionally we do not consider this to be a problem that captures all aspects of discovery, for example hypothesis proposal under high uncertainty, but rather a problem that addresses some of the important issues within the data acquisition aspect of active learning.

Future work should be careful to address the issues of the problem most related to autonomous learning or discovery. This is largely the issue of exploration–exploitation within a problem where there are extremely limited resources in

comparison to the number of experiments to choose from. Of less importance within the multi-armed bandit based problem is how to predict the expected reward, due to the abstract nature of reward within this problem. In a real implementation of an autonomous learner, there would be a mechanism for determining expected information gain, for example through measuring the discrepancy between hypothesis predictions within an ensemble of hypotheses. Here we have used the last reward obtained as the prediction for a particular experiment, however alternatively we could have assigned the true reward that will be obtained on repeating an experiment to the predictions for experiments that have been performed previously. This alternate method for setting the predicted reward would only provide the true next reward to those experiments that had been performed, which is analogous to physical experimentation where the predicted reward could be reasonably estimated by comparing the hypotheses under consideration. Whilst experiments that had not been previously performed would have a predicted reward of zero. This alteration would remove the need for considering reward estimation within potential solutions, however may be less acceptable in peer review.

ACKNOWLEDGMENT

The authors would like to acknowledge Gareth Jones for the image of a lab-on-chip experimentation platform in Fig. 1.

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

REFERENCES

[1] C. J. Lovell, G. Jones, S. R. Gunn, and K.-P. Zauner, "An artificial experimenter for enzymatic response characterisation," in *13th International Conference on Discovery Science*, Canberra, Australia, 2010, pp. 42–56.

[2] J. M. Żytkow, J. Zhu, and A. Hussam, "Automated discovery in a chemistry laboratory," in *Proceedings of the 8th National Conference on Artificial Intelligence*. Boston, MA: AAAI Press / MIT Press, 1990, pp. 889–894.

[3] T. Matsumoto, H. Du, and J. S. Lindsey, "A parallel simplex search method for use with an automated chemistry workstation," *Chemometrics and Intelligent Laboratory Systems*, vol. 62, pp. 129–147, 2002.

[4] N. Matsumaru, S. Colombano, and K.-P. Zauner, "Scouting enzyme behavior," in *2002 World Congress on Computational Intelligence, May 12-17*, D. B. Fogel, M. A. El-Sharkawi, X. Yao, G. Greenwood, H. Iba, P. Marrow, and M. Shackleton, Eds. Honolulu, Hawaii: IEEE, Piscataway, NJ, 2002, pp. CEC 19–24.

[5] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, and S. G. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, pp. 247–252, 2004.

[6] G. Jones, C. J. Lovell, H. Morgan, and K.-P. Zauner, "Organising chemical reaction networks in space and time with microfluidics," *International Journal of Nanotechnology and Molecular Computation (IJNMC)*, vol. 3, no. 1, pp. 35–56, 2011.

[7] C. J. Lovell, G. Jones, S. R. Gunn, and K.-P. Zauner, "Autonomous experimentation: Active learning for enzyme response characterisation," *JMLR: Workshop and Conference Proceedings*, vol. 16, pp. 141–154, 2011.

[8] P. Langley, H. A. Simon, G. L. Bradshaw, and J. M. Żytkow, *Scientific Discovery: computational explorations of the creative processes*. MIT Press, 1987.

[9] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin on the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[10] P. Auer, "Using confidence bounds for exploitation-exploration tradeoffs," *Journal of Machine Learning Research*, vol. 3, pp. 397–422, 2002.

[11] R. D. King, "Robot scientist: an autonomous platform for systems biology discovery," Caliper LifeSciences, Tech. Rep. SC-POS-04 09/06, 2006.

[12] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization error," *Journal of Machine Learning Research*, vol. 7, pp. 141–166, 2006.

[13] M. Kuss, "Gaussian process models for robust regression, classification, and reinforcement learning," Ph.D. dissertation, Technische Universität Darmstadt, 2006.

[14] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the ACM Workshop on Computational Learning Theory*, 1992, pp. 287–294.

[15] T. C. Chamberlin, "The method of multiple working hypotheses," *Science (old series)*, vol. 15, pp. 92–96, 1890, Reprinted in: Science, v. 148, p. 754–759, May 1965.

[16] K. Popper, *The Logic of Scientific Discovery*, 2nd ed. New York: Harper & Rowe, 1968.

[17] B. Settles, "Active learning literature survey," University of Wisconsin-Madison, Tech. Rep., 2009.

[18] M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, and C. Lemmen, "Active learning with support vector machines in the drug discovery process," *J. Chem. Inf. Comput. Sci*, vol. 43, pp. 667–673, 2003.

[19] D. J. C. MacKay, "Information–based objective functions for active data selection," *Neural Computation*, vol. 4, pp. 589–603, 1992.

[20] A. C. Atkinson and V. V. Fedorov, "The design of experiments for discriminating between several models," *Biometrika*, vol. 62, no. 2, pp. 289–303, 1975.

[21] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," in *Proceedings of 24th International Conference on Machine Learning*, 2007, pp. 721–728.

[22] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, pp. 1295–1306, 2009.

[23] D. Kulkarni and H. A. Simon, "Experimentation in machine discovery," in *Computational Models of Scientific Discovery and Theory Formation*, J. Shrager and P. Langley, Eds. San Mateo, CA: Morgan Kaufmann Publishers, 1990, pp. 255–273.

[24] J. O. Pfaffmann and K.-P. Zauner, "Scouting context-sensitive components," in *The Third NASA/DoD Workshop on Evolvable Hardware—EH-2001*, D. Keymeulen, A. Stoica, J. Lohn, and R. S. Zebulum, Eds. Long Beach: IEEE Computer Society, Los Alamitos, July 2001, pp. 14–20.