Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis

Audrey Durand AUDREY.DURAND@MCGILL.CA
School of Compute Science, McGill University

Charis Achilleos ACHILLEOS.CHARIS@UCY.AC.CY
Department of Biological Sciences, University of Cyprus

Demetris Iacovides IACOVIDES.DEMETRIS@UCY.AC.CY
Department of Biological Sciences, University of Cyprus

Katerina Strati STRATI@UCY.AC.CY
Department of Biological Sciences, University of Cyprus

Georgios D. Mitsis GEORGIOS.MITSIS@MCGILL.CA
Department of Bioengineering, McGill University

Joelle Pineau JPINEAU@CS.MCGILL.CA
School of Compute Science, McGill University

Abstract

In this work, we present a specific case study where we aim to design effective treatment allocation strategies and validate these using a mouse model of skin cancer. Collecting data for modelling treatments effectiveness on animal models is an expensive and time consuming process. Moreover, acquiring this information during the full range of disease stages is hard to achieve with a conventional random treatment allocation procedure, as poor treatments cause deterioration of subject health. We therefore aim to design an adaptive allocation strategy to improve the efficiency of data collection by allocating more samples for exploring promising treatments. We cast this application as a contextual bandit problem and introduce a simple and practical algorithm for exploration-exploitation in this framework. The work builds on a recent class of approaches for non-contextual bandits that relies on subsampling to compare treatment options using an equivalent amount of information. On the technical side, we extend the subsampling strategy to the case of bandits with context, by applying subsampling within Gaussian Process regression. On the experimental side, preliminary results using 10 mice with skin tumours suggest that the proposed approach extends by more than 50% the subjects life duration compared with baseline strategies: no treatment, random treatment allocation, and constant chemotherapeutic agent. By slowing the tumour growth rate, the adaptive procedure gathers information about treatment effectiveness on a broader range of tumour volumes, which is crucial for eventually deriving sequential pharmacological treatment strategies for cancer.

1. Introduction

Several recent works have investigated the use of Reinforcement Learning (RL) to automatically discover and optimize sequential treatment strategies that adapt in real-time to the evolution of the disease and the patient's response to previous treatments (Ernst et al.,

© 2018 A. Durand, C. Achilleos, D. Iacovides, K. Strati, G.D. Mitsis & J. Pineau.

2006; Zhao et al., 2009; Panuccio et al., 2013; Bothe et al., 2013; Escandell-Montero et al., 2014). These preliminary results indicate that the adaptive treatment strategies obtained by RL may provide better outcomes than traditional non-adaptive strategies. Yet many challenges remain before this approach is widely transferred to clinical practice, one being the lack of access to appropriate data which can be used to optimize such strategies at a pre-clinical level. For example, the typical randomized treatment allocation procedure may fail to collect data on the full range of possible growth patterns volumes due to fast tumour growth given poor treatments. We thus aim to design an adaptive treatment allocation strategy that would improve the efficiency of data collection by allocating more samples for exploring promising treatments, gathering relevant information for learning policies over the full range of cases (i.e. tumour volumes).

The problem of optimizing a treatment allocation strategy can be formally cast as a contextual bandit episodic game. On each round, we receive a context (e.g. information about the individual, disease symptoms, past treatments), we select the action to perform (e.g. treatment to give) given the context, and we observe a noisy feedback (e.g. treatment effect) related to the action reward. The goal is to estimate the reward functions well enough in order to select actions that maximize the cumulative reward. In this paper, we present a specific case study, where data is obtained from a mouse model of chemically-induced carcinogenesis (Balmain et al., 1984). We present a formal model to analyze the data from an initial exploration phase, including design of the state space representation, policy class, reward function, and try to elucidate what is the best optimized policy that can be recommended from the available data.

Following the work of (Srinivas et al., 2010) and (Krause and Ong, 2011), we adopt a nonparametric approach based on Gaussian Process (GP) regression (Rasmussen and Williams, 2006) to model the reward function. We consider the setting with disjoint actions, which occurs when there is no similarity measure between actions (e.g. independent medical treatments). We develop a new method, called GP BESA, which is an extension of the Best Empirical Sampled Average (BESA) (Baransi et al., 2014) bandit algorithm to the non-linear contextual bandit setting. We present empirical results on three different simulation settings built from data acquired in a previous phase. These early promising results motivated the implementation with real animals in a wet-lab experiment. Preliminary results obtained from 10 mice are encouraging and show that the adaptive policy extends the life spans of animals compared with no treatment, random treatment allocation, and constant chemotherapeutic agent, allowing to gather data with a better coverage of the tumour volume state space.

2. Problem Description

In this work, we address the need for efficient data collection during animal experiments investigating the effectiveness of cancer therapy regimens (Loizides et al., 2015). More specifically, we aim at designing a treatment allocation policy that is personalized and specific to the stage of the disease, hereby represented by the tumour volume x. The effectiveness of a treatment given the tumour volume can be learned by analyzing its impact on the tumour evolution. Here, the evolution of a tumour is represented by triplets (x_i, a_i, x'_i), where x_i denotes the i-th measurement of the tumour volume, a_i is the assigned treatment

on the day of measurement x_i, and x'_i is the measured tumour volume following the treatment administration. The goal is to learn, for each a, the transition function between x_i and x'_i using triplets where $a_i = a$.

The typical approach for collecting those triplets is to run a standard Randomized Clinical Trial (RCT), in which each treatment is randomly assigned to tumours of different volumes. However, ineffective treatments may lead to exponential tumour growths, In turn, this causes a rapid deterioration of subject, limiting the amount of data that can be collected, and also restricts the space of states (tumour volumes) visited. This could be problematic in the subsequent use of data to establish a sequential treatment strategy. This motivates the design of an Adaptive Clinical Trial (ACT) phase, in which accumulated data on prior treatment responses is used to help allocate *better* treatments, thus reducing exposure to less effective treatments. This gives rise to the famous trade-off between *exploration* (searching for optimal treatments) and *exploitation* (treating patients as efficiently as possible), making it an application framework of choice for bandits algorithms. The ACT bears similarities with response adaptive trials (Zhou et al., 2008; Saville and Berry, 2016; Wen et al., 2017), where patients are clustered into groups and the goal is to provide patients with the optimal treatment given their group and prior observed responses. Here, the goal is rather to choose a sequence of actions (treatments) in order to obtain a given result, typically reduce or prevent tumour growth. This can be formulated as a sequential decision making problem.

This is often tackled using Reinforcement Learning (RL) approaches, under the Markov Decision Process (MDP) setting (Villar et al., 2015) – also known as the Bayesian Bernoulli multi-armed bandit problem. RL techniques make it possible to manage not only the immediate response to a treatment, but also the impact of the treatment sequence on the final result, that is the treatment of the patient. However, solving an MDP based on Bellman (1952) equations or the Gittins (1974) indexes requires a good coverage of the state space by previously acquired data, which is not the case in the currently limited amount of available data. Moreover, for logistical reasons, the ACT in the study considered here is implemented on groups of animals and the history of prior contexts, treatments, and observations used by the allocation algorithm is only updated after the completion of a given group. We are therefore facing a problem of delayed feedbacks, where current deterministic ACT algorithms based on MDPs are inefficient (Williamson et al., 2017). Indeed, randomized algorithms are known to be more robust at obtaining delayed feedbacks (Chapelle and Li, 2011). For these reasons, we tackle this problem under the contextual bandits setting, where we propose a randomized treatment allocation technique.

3. Contextual Bandits

A contextual bandit problem is described by a set of contexts $\mathcal{X}$ and a set of actions $\mathcal{A}$. Here the space of contexts corresponds to the space of tumour volumes and the actions set corresponds to the considered treatment options. Each episode t is described by the measurement of a tumour volume x_t and the assignment of treatment a_t. We describe the outcome of a treatment as the tumour volume following the treatment, x'_t. The goal of a bandits algorithm is to select treatment a_t in order to minimize the next tumour volume or, inversely, to maximize $y_t = -x'_t$. This should make it possible to collect samples in the

longer term by stretching the lifetime of animal subjects, while visiting more diverse states by delaying uncontrolled tumour growth.

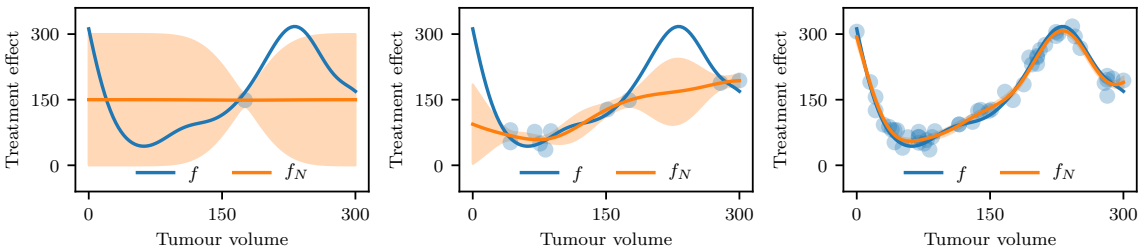
In the standard, non-delayed feedback, setting, a bandit algorithm has access to the complete history of prior contexts, actions, and observations, $(x_s, a_s, y_s)_{s=1}^{t-1}$, in order to make a decision at time t. Here we rather consider an ACT where the history available to the algorithm is only updated *after the completion of a group of subjects*. This is known as delayed feedback and is addressed more effectively by randomized, rather than deterministic, approaches (Chapelle and Li, 2011). Alg. 1 shows the resulting adaptive treatment allocation routine.

Algorithm 1 ACT procedure per groups of animals.

Parameters: contextual bandits algorithm φ

Initialize time $t \leftarrow 1$
for all group g **do**
  Initialize the history $\mathcal{D}_g$ of the group
  for all mouse in group g **do**
    repeat
      Observe tumour volume x_t
      Select (using φ) and apply treatment a_t
      Observe the treatment effect x'_t
      Add the tuple (x_t, a_t, y_t) to the history $\mathcal{D}_g$
      Update time $t \leftarrow t + 1$
    until end of animal protocol
  end for
  $\mathcal{D}_g$ becomes accessible to φ
end for

Such problems are typically addressed through different assumptions on similarity between actions and contexts. Formally, it is supposed that there exists, for each action $a \in \mathcal{A}$, a function $f_a : \mathcal{X} \mapsto \mathbb{R}$ describing the expected observation when selecting action a given the context. Therefore the observation $y_t = f_{a_t}(x_t) + \xi_t$ is a noisy observation of this function for action a_t at the current context x_t. It is common to assume a zero-mean Gaussian noise ξ_t. Many previous works (Auer et al., 2002; Li et al., 2010; Chu et al., 2011; Agrawal and Goyal, 2013) assume linear functions f_a on $\mathcal{X}$ Others tackle the situation where there exists a similarity measure over $\mathcal{A}$. Hence they consider a function $f : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ such that $f_a(x) = f(x, a)$. Among those, some assume that f is a Lipschitz function (Slivkins, 2014), while others assume that f is sampled from a Gaussian Process (GP) distribution (Krause and Ong, 2011; Valko et al., 2013). One can see the GP distribution as a generalization of a Gaussian probability distribution: it is a distribution over functions. The specific case of *disjoint actions* describes the situation where there is no similarity between actions, which is also known as *bandits with covariates* (Rigollet and Zeevi, 2010; Perchet and Rigollet, 2013).



(a) $N = 1$ (b) $N = 10$ (c) $N = 50$

Figure 1: GP posterior (mean and standard deviation) on a function f for different numbers (N) of observations.

4. Proposed Approach

Here we address the ACT problem under the contextual bandit setting with disjoint actions, such that each function f_a is assumed to be sampled from a different GP distribution. This covers the case where functions f_a have different regularities, which is realistic given that the effects of different treatments may vary more or less abruptly according to the stage of the disease. It is thus natural to rely on GP regression for maintaining a posterior distribution on each f_a (Srinivas et al., 2010; Krause and Ong, 2011; Valko et al., 2013).

4.1. Gaussian Process Regression

GP (Rasmussen and Williams, 2006) regression is a non-parametric regression technique that uses a *kernel* to project data into a new space, where linear regression can be performed to recover the target function. For two points $x, x' \in \mathcal{X}$, let us define a kernel function $k(x, x') : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ encoding the amount of information shared between points, along with priors over the regularity of the target function. Consider the $N \times 1$ vector $\mathbf{y}_N = (y_1, \dots, y_N)$ of N previous observations obtained using action a in contexts $x_1, \dots, x_N$. Given the $N \times N$ kernel matrix and $N \times 1$ vector,

$$\mathbf{K}_N = \big[k(x_s, x_{s'})\big]_{s,s' \leqslant N} \quad \text{and} \quad \mathbf{k}_N(x) = \big(k(x, x_s)\big)_{s \leqslant N},$$

and assuming Gaussian noise $\xi_t \sim \mathcal{N}(0, \sigma^2)$, the posterior distribution on the underlying target function f given prior observations is a multivariate Gaussian distribution,

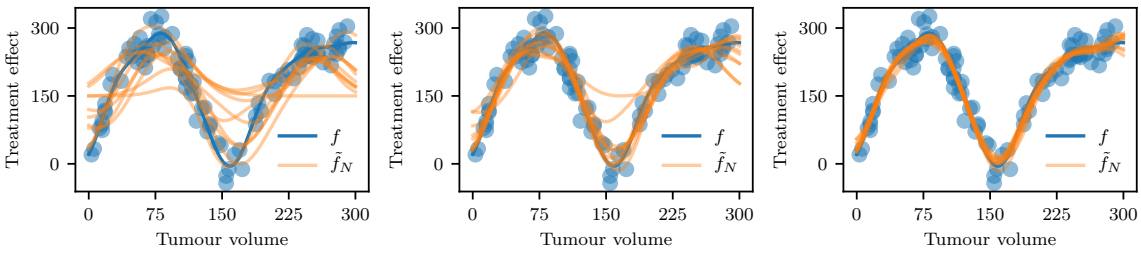
$$\mathbb{P}[f | x_1, \dots, x_N, y_1, \dots, y_N] \sim \mathcal{N}\Big(\big(f_N(x)\big)_{x \in \mathcal{X}}, \big[k_N(x, x')\big]_{x,x' \in \mathcal{X}}\Big),$$

where

$$f_N(x) = \mathbf{k}_N(x)^\top (\mathbf{K}_N + \sigma^2 \boldsymbol{I}_N)^{-1} \mathbf{y}_N \quad \text{and} \tag{1}$$

$$k_N(x, x') = k(x, x') - \mathbf{k}_N(x)^\top (\mathbf{K}_N + \sigma^2 \boldsymbol{I}_N)^{-1} \mathbf{k}_N(x'). \tag{2}$$

Fig. 1 shows examples of posterior distributions that could be obtained with GP regression



(a) $N = 5$ (b) $N = 20$ (c) $N = 50$

Figure 2: Examples of posterior mean obtained by GP regression for 10 different subsamples of observations of size N from the 100 displayed observations.

for different number of observations from a synthetic function. We observe that the more observations available, the closer is the posterior mean to the true underlying function. In practice, computing the posterior mean requires the inversion of the $N \times N$ matrix $(\mathbf{K}_N + \sigma^2 \boldsymbol{I}_N)$, typically performed using a Cholesky decomposition of complexity $\mathcal{O}(N^3)$ (Krishnamoorthy and Menon, 2013), making this approach primarily suitable for small size samples.

4.2. Action Selection Strategy

We now introduce a new contextual bandits strategy to select the next action based on GP regression. To this aim, we rely on the BESA (Baransi et al., 2014) principle, that is on a *fair* comparison between empirical estimators. A comparison between the estimators of two actions is said to be fair if both actions had the same opportunities of showing their potential. Concretely, this can be achieved by computing the estimators using the same amount of observations for both actions. This results in estimators with comparable confidence intervals. Here, each function f_a is estimated by GP regression. The proposed approach, GP BESA, considers subsets of observations such that posterior models are computed using the same amount of observations for each action. Fig. 2 shows various posterior means obtained by GP regression over 10 subsamples of observations obtained from a synthetic function, of different sizes. We observe that the posterior mean is more dependent upon to the sampled subset for contexts where less data is available. Relying on the posterior mean on subsamples (rather than on all observations) for selecting the next treatment therefore allows some exploration.

Let $N_{a,t}$ denote the amount of observations obtained by playing action a up to episode t (inclusively). Let $\tilde{f}_{a,N}$ denote the posterior mean on f_a obtained by GP regression conditioned on a random subsample (without replacement) of $N \leqslant N_{a,t-1}$ previous observations acquired with action a (Eq. 1). Alg. 2 shows the resulting GP BESA action selection algorithm for two actions. It defines a contextual bandits algorithm (φ) that can be used to select the next action on line 7 of Alg.1. For more than two actions, a tournament can be set up (Baransi et al., 2014). By aiming for treatment a_t maximizing $\tilde{f}_{a_t,N}$ in context x_t, GP BESA aims for the treatment with the lowest expected next tumour volume.

Algorithm 2 GP BESA for selecting among two actions.

Parameters: current episode t, context x_t, actions a and b

1: $N = \min(N_{a,t-1}, N_{b,t-1})$
2: $a_t = \text{argmax}_{i \in \{a,b\}} \tilde{f}_{i,N}(x_t)$, where $\tilde{f}_{i,N}()$ is defined as per Eq. 1 on N subsampled observations of action i
3: **return** a_t

5. Experimental Setting

In our study, we considered mice with induced cancer tumours (Balmain et al., 1984), treated using combinations of 5-FU, a chemotherapeutic agent, and imiquimod, a synthetic compound modifying the immune response. We considered the following, fixed-dose, options: no treatment, 5-FU (100mg/kg), imiquimod (8mg/kg), and simultaneous combination of imiquimod and 5-FU. More precisely, we aimed to learn treatment policies that adapt to the stage of the disease.

Currently, the only information about cancer progression that is collected during the trial is through tumour measurements. We therefore characterized the disease progression using the approximation (Tomayko and Reynolds, 1989) of the tumour volume given by $x = \frac{\pi}{6}(\ell w)^{\frac{3}{2}}$, assuming an ellipsoid shape for the tumour, where ℓ and w respectively denote the ellipsoid tumour length and width measured using calipers (for detailed measurement procedures see Loizides et al. (2015)).

5.1. Animal Model

Skin tumours were induced in mice using the DMBA/TPA (Tumor Promoting Agent) model of chemical carcinogenesis. A mouse was treated with TPA for the whole duration of the experiment, which is the standard practice in this type of skin carcinogenesis model (Balmain et al., 1984). Specifically, data was gathered following this experimental procedure:

- Data collection commenced for a mouse when its largest tumour reached 3mm.
- A mouse was sacrificed when its largest tumour reached 10mm or if its general health was extensively deteriorating, according to animal handling guidelines (Simonson et al., 2005).
- Twice weekly, all tumours were measured and a treatment was assigned.
- Each administered treatment was recorded, including the tumour volume preceding and following the treatment.

5.2. Initial Data Collection

During an initial RCT phase, the goal was to randomly assign one of the four treatment (none, 5-FU, imiquimod, or 5-FU combined with imiquimod) twice weekly to six mice with three tumours each (at least one of 3mm diameter), which would result in 18 tumours. Note that this randomization was independent of tumour size. Due to unpredictable biological events, a small portion of the data was removed from further analysis. Specifically, data from one mouse in which tumours took too long to grow, two tumours that disappeared because they were exposed to treatment in an early stage, and a tumour that merged with a

Table 1: Number of samples per treatement in the RCT dataset.

	None	5-FU	Imiquimod	5-FU + Imiquimod	Total
N	42	66	24	31	163

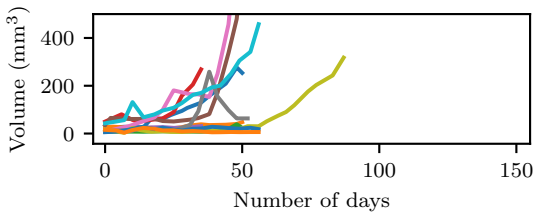


Figure 3: Evolution of tumours during the RCT phase. Each line corresponds to a different tumour. Time scale is set to facilitate comparison with further results (Fig. 5(a)).

new one were removed from the dataset. The resulting dataset contained 12 tumours from five mice, which corresponds to a total of 163 (x_i, a_t, x_i') triplets. Table 1 shows how samples are distributed across treatment options. Fig. 3 shows the observed tumour growths.

6. Simulations

Before moving to the adaptive data collection on animal subjects, the performance of GP BESA was asserted using a simulated study. Simulation models were built using the initial RCT data (Section 5.2). The context arrival (simulated encounter of tumour volumes) was modeled using the probability density function (PDF) of an exponential distribution:

$$f(x|\gamma, \lambda) = \begin{cases} \lambda e^{-\lambda(x-\gamma)} & \text{for } x \geqslant \gamma \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

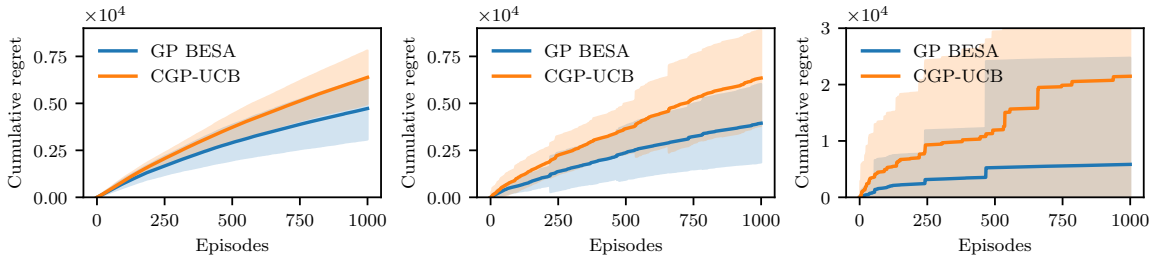
of location $\gamma = 3.42$ mm^3 and scale[1] $\lambda = 66.88$ mm^3 fitted to the volumes encountered during the RCT (Fig. 6(a)). As for the treatment effects, four different models were built for each treatment using linear, cubic, and quartic regression on the RCT data (Table 1).

GP BESA was compared with CGP-UCB (Krause and Ong, 2011), a state-of-the-art contextual bandits algorithm based on GP regression. Let $f_{a,N}(x)$ and $k_{a,N}(x,x)$ respectively denote the predictive mean and variance on f_a at point x obtained by GP regression on the last N observations acquired by trying action a (Eq. 1 and 2). CGP-UCB selects the action

$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left[f_{a,N_{a,t-1}}(x_t) + \beta_t(\delta) k_{a,N_{a,t-1}}(x_t, x_t) \right],$$

where $\beta_t(\delta) = \sqrt{2 \ln \frac{4t^2\pi^2}{6\delta}}$ and $\delta = 0.1$. GP BESA and CPG-UCB used the same GP hyperparameters, which were obtained by maximizing the likelihood in the initial RCT data (Rasmussen and Williams, 2006) and kept fixed afterwards.

1. The scale of an exponential distribution also corresponds to its mean and standard deviation.



(a) Linear (b) Cubic (c) Quartic

Figure 4: Average cumulative regret and standard deviation for each simulation model (lower is better).

Contextual bandit algorithms are typically compared based on their ability to minimize the cumulative *regret*:

$$\mathfrak{R}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \left[\left(\max_{a \in \mathcal{A}} f_a(x_t) \right) - f_{a_t}(x_t) \right]. \tag{4}$$

This quantity measures the expected loss incurred by taking the actions recommended by the algorithm rather than taking the action with the highest expected outcome given the context. In other words, it compares an algorithm with an oracle that would have access to the (unknown) treatment effect functions. An algorithm selecting actions uniformely at random would obtain a linear regret given episodes. Hence, an algorithm that is actually learning should accumulate regret sublinearly. Of course, using the regret as performance measure requires that we have access to the true underlying functions, which is possible only in simulations. For a fair comparison of algorithms, it is required that the sequence of contexts $(x_t)_{t \geqslant 1}$ is the same for both algorithms. This is also possible in simulation studies. Therefore, both algorithms were executed 100 times during 1000 episodes, that means on 100 sequences of 1000 contexts, where the contexts observed during the i-th sequence were the same for both algorithms.

Fig. 4 compares the cumulative regret averaged over the 100 repetitions for both algorithms. We observe that both achieve sublinear regret, and that GP BESA cumulates less regret and shows less variance than CGP-UCB. This is interesting, given that GP BESA does not consider the posterior variance provided by the GP but relies only on the posterior mean. Paired t-tests on the regret accumulated after 1000 episodes rejected the null hypothesis with $p < 10^{-5}$ for all settings. Note that the goal of this experiment is not to claim that GP BESA is *better* overall than CGP-UCB, but rather to evaluate its potential for the current application before deploying the algorithm on real living animals. In this domain at least, GP BESA competes well against the state-of-art, making it a valid and promising candidate for an ACT.

7. Adaptive Data Collection

An ACT was subsequently set up in a laboratory experiment on the same animal model as the previous RCT (Section 5.1), but using GP BESA to assign treatments based on the main (largest) tumour volume. Groups of mice were processed simultaneously (until the death of all members of the group). After group termination, data for all mice in the group were added to the GP BESA history, such that they could be subsampled and used to fit GPs for mice in subsequent groups. A total of ten mice were treated and grouped as follows as they were born:

Group A mice 1 and 2;
Group B mice 3, 4, and 5;
Group C mice 6 and 7;
Group D mice 8, 9, and 10.

The effectiveness of the strategy obtained with GP BESA was compared with the following basic treatment strategies:

None (3 mice) tumour evolution without any treatment;
Random (5 mice) random allocation strategy (RCT);
5-FU (4 mice) fixed dosage of 5-FU.

The last strategy is a standard chemotherapy procedure (Longley et al., 2003). It also corresponds to always selecting 5-FU in the treatment options given to GP BESA.

Comparing different treatment allocation strategies is not trivial since the contexts that are observed (tumour volumes) vary from one experiment to another, and no ground truth regarding the optimal tumour growth is known. However, the lifespan of the animals might be seen as an indicator of how fast tumours grow. By comparing the evolution of tumours observed during the RCT (Fig. 3) and all the animal groups of the ACT (Fig. 5(a)), we observe that tumour volumes exhibit delayed exponential growth over time resulting in longer animal lifetimes. This is highlighted in Fig. 5(b), which shows the distribution of mice lifetime for each basic treatment allocation strategy and GP BESA. We observe on the one hand that GP BESA makes it possible to increase the lifetime of the mice compared to the basic treatment allocation strategies. This is explained in Fig. 5(c), which shows the distribution of mice lifetime for each GP BESA group. We observe that GP BESA improves after each update, that is, after integrating the data collected from the previous group of mice. More specifically, we notice an increase of more than 50% longevity between the best basic strategy (5-FU) and the last group of GP BESA. Note that the significant difference in lifetime for the mice belonging to the different update groups explains the range of GP BESA results in the previous figure. We also note that groups of mice present much less variability in their longevity as the updates are made. This may be due to the use of less variable treatment strategies with the convergence of GP BESA. This effect should, however, be validated in a larger cohort.

Table 2 shows how collected samples are distributed across treatment options, for each group and over the whole ACT dataset. Fig. 6 compares the distribution of tumour volumes observed during the RCT and ACT phases. We observe that the adaptive treatment allocation strategy results in a *better* state space coverage. More specifically, the last five mice

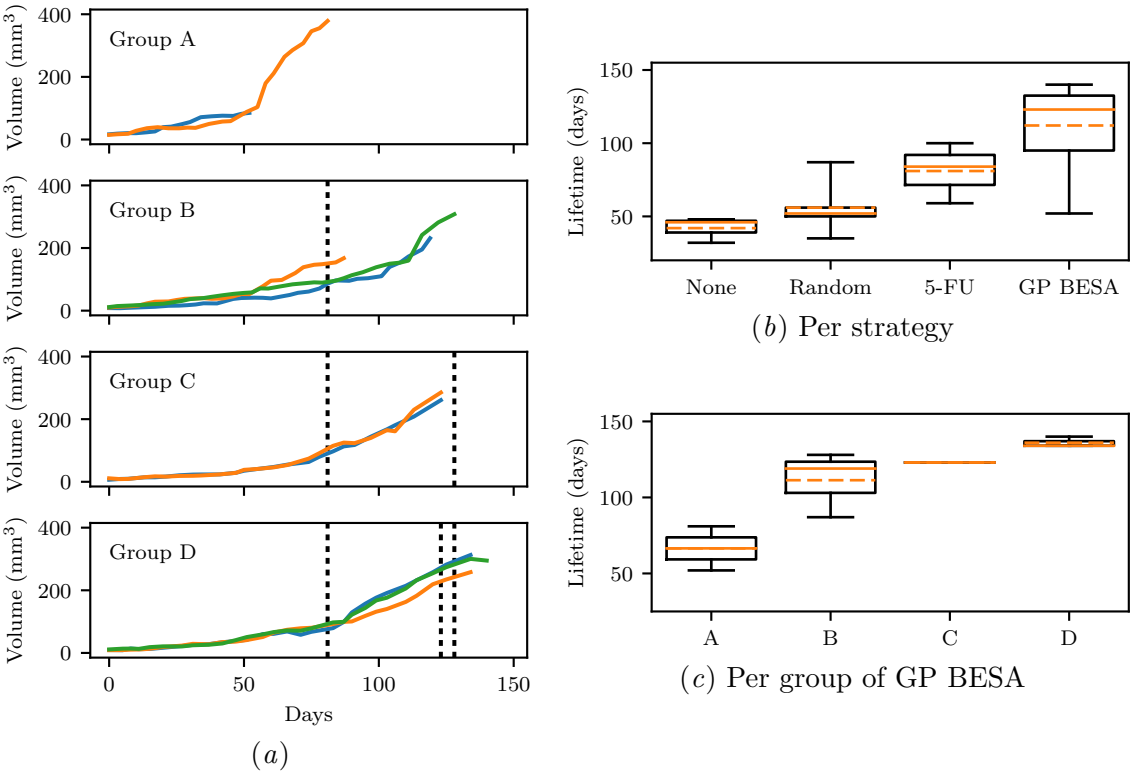


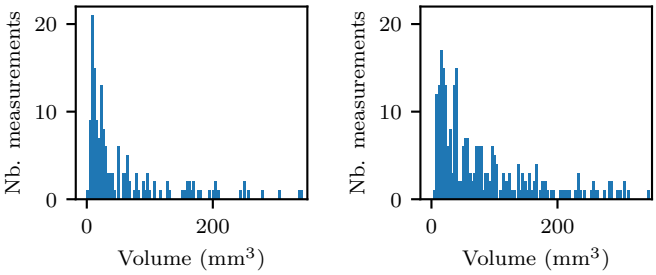
Figure 5: a) Comparative evolution of tumours during the ACT (longer duration in days is better). Dotted lines indicate max duration in previous groups. b), c) Median (line) and mean (dotted line) lifetime of animals (higher is better). Boxes cover 1st to 3rd quartiles and handles show the spread of data.

Table 2: Number of samples per treatement in the ACT dataset.

	None	5-FU	Imiquimod	5-FU + Imiquimod	Total
Group A	3	3	10	24	40
Group B	6	6	24	50	86
Group C	2	3	19	26	50
Group D	1	5	39	40	85
N	12	17	92	140	261

using the ACT strategy allowed to gather 59 data points for volumes larger than 70 mm^3 compared with 42 data points for the five mice from the RCT. This corresponds to an increase of 40%. This additional data is of major importance for exploring better strategies during later stages of the disease.

Fig. 7 shows the treatment allocation policy that was used in each group. We observe that, as updates go on, that the system learns to avoid the "No treatment" (None) option. More specifically, the resulting strategy seems to tend toward alternating between



(*a*) RCT ($N = 163$) (*b*) ACT ($N = 261$)

Figure 6: Distribution of tumour volumes during both phases.

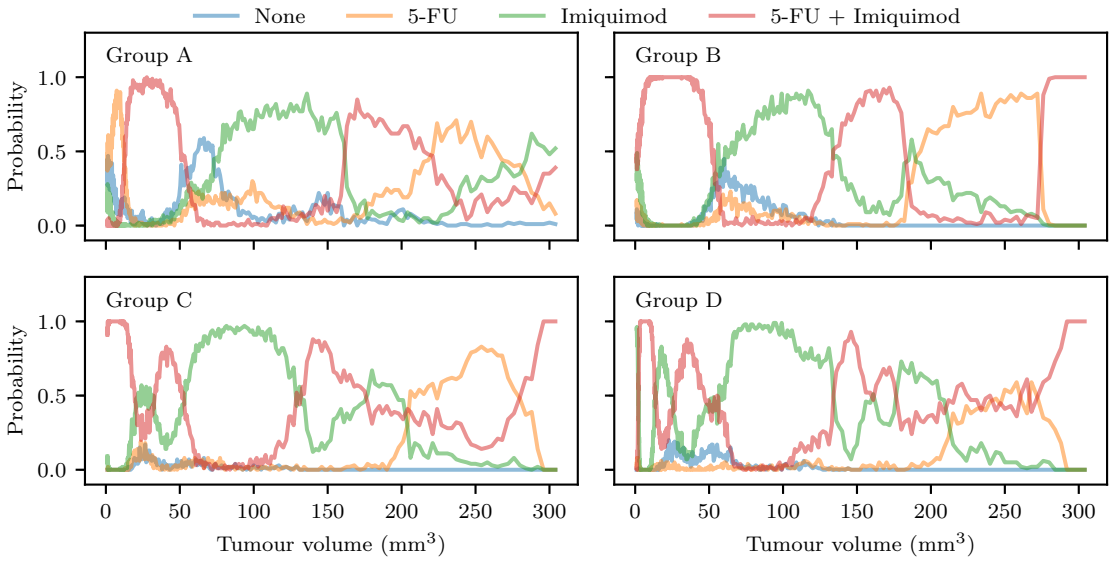


Figure 7: Probability of assigning each treatment given the tumour volume.

Imiquimod and 5-FU + Imiquimod. This could suggest a behaviour trying to avoid the development of drug resistance (Housman et al., 2014).

8. Conclusion

Both simulation and animal experiments results show that GP BESA constitutes a promising randomized alternative to current deterministic contextual bandit strategies. This is interesting, given that GP BESA conducts exploration through subsampling of observations rather than relying on the posterior variance provided by GP regression. These results also support the applicability of GP BESA for contextual learning in a limited data regime, which remains a challenge.

Results on animal experiments suggest that GP regression is able to capture the highly variable features of cancer progression. It can therefore be a useful tool in attempting to model the dynamics of tumour growth, which is a current challenge (Loizides et al., 2015). A better understanding of the evolution of cancerous tumours could help in the discovery

of treatments and promote the implementation of strategies adapted to the disease. Future work will confirm the potential for using the data collected for off-line learning of treatment policies.

These results also support the deployment of ACT strategies in real-world application contexts. Although the bandit setting (Thompson, 1933; Robbins, 1952) takes roots in ACT application, the use of these algorithms in the field still remains limited (Villar et al., 2015). More efforts to provide theoretical guarantees under practically feasible assumptions could help facilitate the adoption of these strategies. In particular, experiments incorporating more variables to better characterize the disease an the subjects should be carried out. Earlier work (Djolonga et al., 2013; Wang et al., 2016; Li et al., 2016) addressed the problem of GP regression on a high dimensional space. Others (Snoek et al., 2015; Springenberg et al., 2016) have considered the use of deep neural networks to approximate posterior distribution. Extensions of BESA to these type of approaches could make it possible to provide randomized contextual bandit to high-dimensional context spaces. This could open the door to ACT for personalized strategies, adapting treatments to patients characteristics.

References

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pages 127–135, 2013.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

A. Balmain, M. Ramsden, G. T. Bowden, and J. Smith. Activation of the mouse cellular harvey-ras gene in chemically induced benign skin papillomas. *Nature*, 307:658–660, 1984.

A. Baransi, O.-A. Maillard, and S. Mannor. Sub-sampling for multi-armed bandits. In *ECML*, pages 115–131, 2014.

R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38(8):716–719, 1952.

M. K. Bothe, L. Dickens, K. Reichel, A. Tellmann, B. Ellger, M. Westphal, and A. A. Faisal. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. *Expert Review of Medical Devices*, 10(5):661–73, 2013.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *NIPS*, pages 2249–2257, 2011.

W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214, 2011.

J. Djolonga, A. Krause, and V. Cevher. High-dimensional gaussian process bandits. In *NIPS*, pages 1025–1033, 2013.

D. Ernst, G.B. Stan, J. Goncalves, and L. Wehenkel. Clinical data based optimal STI strategies for HIV: A reinforcement learning approach. In *45th IEEE Conference on Decision and Control*, pages 667–672, 2006.

P. Escandell-Montero, M. Chermisi, J.M. Martínez-Martinez, J. Gomez-Sanchis, C. Barbieri, E. Soria-Olivas, F. Mari, J. Vila-Frances, A. Stopper, E. Gatti, and J.D. Martín-Guerrero. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1):47–60, 2014.

J. Gittins. A dynamic allocation index for the sequential design of experiments. *Progress in statistics*, pages 241–266, 1974.

G. Housman, S. Byler, S. Heerboth, K. Lapinska, M. Longacre, N. Snyder, and S. Sarkar. Drug resistance in cancer: An overview. *Cancers*, 6(3):1769–1792, 2014.

A. Krause and C. S. Ong. Contextual Gaussian process bandit optimization. In *NIPS*, pages 2447–2455, 2011.

A. Krishnamoorthy and D. Menon. Matrix inversion using cholesky decomposition. In *IEEE Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, pages 70–72, 2013.

C.-L. Li, K. Kandasamy, B. Póczos, and J. Schneider. High dimensional Bayesian optimization via restricted projection pursuit models. In *AISTATS*, pages 884–892, 2016.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, pages 661–0670, 2010.

C. Loizides, D. Iacovides, M. M. Hadjiandreou, G. Rizki, A. Achilleos, K. Strati, and G. D. Mitsis. Model-based tumor growth dynamics and therapy response in a mouse model of de novo carcinogenesis. *PloS One*, 10(12):e0143840, 2015.

D. B. Longley, D. P. Harkin, and P. G. Johnston. 5-fluorouracil: mechanisms of action and clinical strategies. *Nature Reviews Cancer*, 3(5):330, 2003.

G. Panuccio, A. Guez, R. Vincent, M. Avoli, and J. Pineau. Adaptive control of epileptiform excitability in an in vivo model of limbic seizures. *Experimental Neurology*, 241:179–83, 2013.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Annals of Statistics*, 41(2):693–721, 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT press Cambridge, 2006.

P. Rigollet and A. Zeevi. Nonparametric bandits with covariates. In *COLT*, pages 54–66, 2010.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952.

B. R. Saville and S. M. Berry. Efficiencies of platform clinical trials: A vision of the future. *Clinical Trials*, 13(3):358–366, 2016.

S. J. S. Simonson, M. J. Difilippantonio, and P. F. Lambert. Two distinct activities contribute to human papillomavirus 16 e6's oncogenic potential. *Cancer research*, 65(18): 8266–8273, 2005.

A. Slivkins. Contextual bandits with similarity information. *JMLR*, 15:2533–2568, 2014.

J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. A. Patwary, Prabhat, and R. Adams. Scalable Bayesian optimization using deep neural networks. In *ICML*, pages 2171–2180, 2015.

J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian optimization with robust bayesian neural networks. In *NIPS*, pages 4134–4142, 2016.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, 2010.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

M. M. Tomayko and C. P. Reynolds. Determination of subcutaneous tumor size in athymic (nude) mice. *Cancer chemotherapy and pharmacology*, 24(3):148–154, 1989.

M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *UAI*, pages 654–665, 2013.

S. S. Villar, J. Bowden, and J. Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. de Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

S. Wen, J. Ning, S. Collins, and D. Berry. A response-adaptive design of initial therapy for emergency department patients with heart failure. *Contemporary clinical trials*, 52: 46–53, 2017.

S. F. Williamson, P. Jacko, S. S. Villar, and T. Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational Statistics & Data Analysis*, 113:136–153, 2017.

Y. Zhao, M.R. Kosorok, and Zeng D. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–315, 2009.

X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee. Bayesian adaptive design for targeted therapy development in lung cancer—a step toward personalized medicine. *Clinical Trials*, 5(3):181–193, 2008.