Novel Exploration Techniques (NETs) for Malaria Policy Interventions

Oliver Bent
University of Oxford
oetbent@robots.ox.ac.uk

Sekou L. Remy
IBM Research Africa
sekou@ke.ibm.com

Stephen Roberts
University of Oxford
sjrob@robots.ox.ac.uk

Aisha Walcott-Bryant
IBM Research Africa
awalcott@ke.ibm.com

arXiv:1712.00428v1 [cs.AI] 1 Dec 2017

***Abstract*—The task of decision-making under uncertainty is daunting, especially for problems which have significant complexity. Healthcare policy makers across the globe are facing problems under challenging constraints, with limited tools to help them make data driven decisions. In this work we frame the process of finding an optimal *malaria* policy as a stochastic multi-armed bandit problem, and implement three agent based strategies to explore the policy space. We apply a Gaussian Process regression to the findings of each agent, both for comparison and to account for stochastic results from simulating the spread of malaria in a fixed population. The generated policy spaces are compared with published results to give a direct reference with human expert decisions for the same simulated population. Our novel approach provides a powerful resource for policy makers, and a platform which can be readily extended to capture future more nuanced policy spaces.**

I. Introduction

Malaria is a mosquito-borne disease which is endemic in sub-Saharan Africa (SSA). There has been a significant progress in the prevention and control of the disease resulting in a reduction of the mortality rate of malaria and the number of new cases. Many countries in SSA rely heavily on external funding for malaria prevention and control which, in recent years, investments have started to level-off [1]. This means that policy makers will have more difficult decisions to ensure continued successes in the management of the disease and of their populations given the available resources. Moreover, it is projected that up to $450M in research and development (R&D) is required each year for malaria up to 2018, with slower growth needed thereafter [2]. It is critical for SSA countries to have access to tools to develop policies that maximize the cost-effectiveness of malaria interventions.

Individual distributed decision makers (e.g., NGOs, governments and charities) must be able to explore the possible set of actions for appropriate malaria interventions within their populations. Such policies include a mix of actions like the distribution of long-lasting insecticide-treated nets (ITNs), indoor residual spraying (IRS), vector larvicide in bodies of water, and malaria vaccinations. The space of possible policies for malaria interventions is daunting and inefficient for human decision makers to explore without adequate decision support tools.

In this work we describe a novel formulation for the systematic exploration of malaria intervention actions. This formulation is well-suited for applying Artificial Intelligence (AI) agents to learn the most effective intervention strategies for a specific environment. To date, the applications of AI in healthcare have focused around prediction of disease spread, diagnosis, and personalized care planning tools (e.g., [3] and [4]). Building on these applications, our work leverages the OpenMalaria Platform [5], which provides a simulation environment for an agent to learn optimal policies for the control of the disease. The OpenMalaria codebase gives access to stochastic transmission models of malaria and can be used by researchers to evaluate the impact of various malaria control interventions. OpenMalaria therefore provides a platform to create a simulation environment from which an agent may explore optimal policies for the control of malaria transmission. Specifically, the work presented will make use of a parameterisation of OpenMalaria models which describes the Rachuonyo South district in Western Kenya [6].

Our approach is to apply multiple agents to determine the optimal malaria policy based on any combination of coverage of ITN and IRS for the target population. The reward function is determined by an application of the cost of disability adjusted life years. A key benefit of this work is analytical search space exploration to converge on an optimal policy. We demonstrate how agent-based exploration techniques and advances in compute infrastructure can be leveraged to determine the optimal policy of malaria interventions for a particular environment, without expert human guidance. Moreover, our work shows the potential for a systematic agent based decision support system for human decision makers exploring cost-effective intervention strategies.

II. Stochastic Multi-Armed Bandit

Finding an optimal malaria policy from OpenMalaria simulations can be posed as a stochastic multi-armed bandit problem. For example this formulation has been used as an approach to develop models which may aid in the design of clinical trials, where actions should be made to balance exploitation (positive patient outcomes) and exploration (searching for actions which may lead to a clinical ‘breakthrough’). In our framing we wish to efficiently determine high performing policies for a simulated population of individuals over a 5 year intervention time frame.

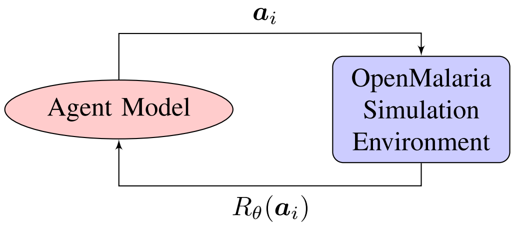

$\boldsymbol{a}_i$
Agent Model
OpenMalaria Simulation Environment
$R_\theta(\boldsymbol{a}_i)$

Fig. 1: Policies $\boldsymbol{a}_i$ are chosen by the Agent Model which receives rewards $R(\boldsymbol{a}_i)$

A. State

Due to the multi-armed bandit framing of the problem there is no state transition between OpenMalaria simulations. Instead we are trying to solve the problem of making one-shot policy recommendations for the simulation intervention period of 5 years. The state is therefore defined by the simulation's initial parameters θ and the policy $\boldsymbol{a}$ simulated.

B. Action

The main control methods used in Rachuonyo South district are: mass-distribution of long-lasting insecticide-treated nets (ITNs); Indoor Residual Spraying (IRS) with pyrethroids; and the prompt and effective treatment of malaria. This work will explore a policy space made up of the first two components (ITNs and IRS) which are direct intervention strategies, while prompt and effective treatment is described by the OpenMalaria simulation parameters and impacts the rewards detailed in the proceeding section. The domain of the first component is the deployment of nets, which defines the coverage of the population ($a_{\text{ITN}} \in (0, 1]$). The domain for the second component is the application of seasonal spraying, which defines the proportion of population coverage for this intervention ($a_{\text{IRS}} \in (0, 1]$). The spraying regimens of IRS are conducted through alternating the intervention between April/June each year [7]. Since the policy decision is framed as how much of the simulated population should be covered by a particular intervention, the policy space A is constructed through $\boldsymbol{a}_i \in A = \{a_{\text{ITN}}, a_{\text{IRS}}\}$.

For the studied scenario the simulation environment handles distribution of the interventions across the simulated population. The agent is not controlling the more complex actions of targeted interventions which have not been previously reported on, though it should be noted that while the action space is finite (there are a finite number of individuals in the simulation model), the size of this space will grow exponentially as more interventions or targeted interventions are added. Simulation compute time also grows linearly with number of simulated individuals. As such, a complete exploration of the entire action space quickly becomes in-feasible as complexity grows toward any real-world equivalent simulation. What is presented here is a first approximation to a real-world scenario.

C. Reward

The reward associated with each policy $R_\theta(\boldsymbol{a}_i)$ is stochastic through the parameterisation of the simulation θ, which generates a randomised distribution of parameters for the OpenMalaria simulation. The magnitude of the reward is determined through an economic cost-effectiveness analysis of the stochastic simulation output. The following sections give an overview of the calculations used, as specified in the health economics literature.

1) DALYs: Disability adjusted life years [8], are a measure defined by the total years of life lost (YLL) due to fatality linked with contraction of the disease, and number of years of life with disability (YLD) as a result of the disease. Upon termination, the OpenMalaria simulation produces outcomes for each individual in the population for the considered scenario. If an individual experienced a malaria episode ($\mathtt{ME_k}$), the simulation results allow YLD to be quantified (See (1)). Additionally if an individual contracted malaria and subsequently died ($\mathtt{D_z}$), either directly or indirectly from malaria, the simulation output allows YLL to be calculated (See (2)). We also use a discount factor $\gamma = 0.97$ to discount the value of future years of life lost, and a life expectancy of 46.6 years [9]. The work of [10] may be referred to for an explicit mapping of OpenMalaria simulation outputs to the calculation of DALYs.

$$\text{YLD} = \sum_{k=0}^{K} \mathtt{Duration(ME_k)} * \mathtt{Weight(Age(ME_k))} \tag{1}$$

$$\text{YLL}_z = \max(0, \mathtt{LifeExpectancy} - \mathtt{Age(D_z)})) \tag{2}$$

$$\text{YLL} = \sum_{z=0}^{Z} \text{YLL}_z \times \gamma^{\text{YLL}_z} \tag{3}$$

$$\text{DALY} = \text{YLL} + \text{YLD} \tag{4}$$

2) Simulated Costs: In this work we simulate two types of costs, the cost to treat and manage malaria episodes, and the cost to implement interventions which minimise malaria prevalence. We call these healthcare system costs (HSC), and intervention costs (IC) respectively.

For each malaria episode that a patient seeks treatment, hospitals incur costs to treat the disease, to manage the patient's recovery process, and also to deal with the patient's death if that were to occur. As such, the HSC can be broken down in terms of total in-hospital treatment costs (TTC), the total in-hospital recovery cost (TRC), and the cost for in-hospital mortality (See (5-7)). We use values from the literature [11] to define the costs implemented in this work.

$$\text{TTC} = \sum_{k=0}^{K} \mathtt{Cost(Treatment(InHospital(ME_k)))} \tag{5}$$

$$\text{TRC} = \sum_{k=0}^{K} \mathtt{Cost(Recovery(InHospital(ME_k)))} \tag{6}$$

$$\text{HSC} = \text{TTC} + \text{TRC} + \sum_{z=0}^{Z} \mathtt{Cost(InHospital(D_z))} \tag{7}$$

The cost of the intervention (C_{int}) is the sum of the numbers of individuals covered by each intervention by the average cost of deploying the intervention to an individual. For the region in Kenya under consideration it costs 8.52USD per net and 0.73USD per person covered by spraying intervention. The average cost to seek hospital treatment per person is 0.60USD, and this value is associated with transportation and consumables. [7].

3) Cost Effectiveness: The agent models proposed will receive rewards based on the cost effectiveness of a policy, a metric often used by researchers evaluating the impact of a policy. We define cost effectiveness as the ratio of the relative cost to perform a policy intervention to the health impact realized from that policy intervention. The health impact is defined as the DALYs averted (DA), the difference between the DALYs realized with the application of the considered policy intervention and the DALYs realized when performing no intervention. So the cost effectiveness will be quantified as the cost per DALY averted (C_{DA}):

$$C_{\text{DA}} = \frac{\text{HSC}_{\text{int}} - \text{HSC}_{\text{noint}} + C_{\text{int}}}{\text{DA}} \tag{8}$$

D. Gaussian Process Regression

The rewards received from the OpenMalaria simulation environment are stochastic as there is noise built in to the underlying models. Despite stochastic simulation results, the rewards received from similar policies should be highly correlated. If we consider that each simulated policy returns a stochastic scalar reward $R(\boldsymbol{a}^1)...R(\boldsymbol{a}^n)$ with mean $\mu(\boldsymbol{a}) = E[R(\boldsymbol{a})]$ and covariance $k(\boldsymbol{a}, \boldsymbol{a}') = E[(R(\boldsymbol{a}) - \mu(\boldsymbol{a}))(R(\boldsymbol{a}') - \mu(\boldsymbol{a}'))]$, a gaussian process can be specified by these mean and covariance functions $GP(\mu(\boldsymbol{a}), k(\boldsymbol{a}, \boldsymbol{a}'))$.

Gaussian Process regression (GPR) is a supervised learning technique, in which the stochastic scalar rewards R, are used to train a Gaussian Process to infer with confidence bounds the performance of actions across the policy space [12]. The learnt parameters describe the posterior distribution over $R(\boldsymbol{a})$:

$$\mu_{i+1}(\boldsymbol{a}) = \boldsymbol{k}_i(\boldsymbol{a})^T(\boldsymbol{K}_i + \sigma^2\boldsymbol{I})^{-1}R_i \tag{9}$$

$$\sigma_{i+1}(\boldsymbol{a}) = k(\boldsymbol{a}, \boldsymbol{a}') - \boldsymbol{k}_i(\boldsymbol{a})^T(\boldsymbol{K}_i + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_i(\boldsymbol{a}) \tag{10}$$

At each location $\boldsymbol{a} \in A$, $\boldsymbol{k}_i(\boldsymbol{a}) = [k(\boldsymbol{a}_i^k, \boldsymbol{a})]_{\boldsymbol{a}_i^k \in A_i}$ and $\boldsymbol{K}_i = [k(\boldsymbol{a}, \boldsymbol{a}')]_{\boldsymbol{a}, \boldsymbol{a}' \in A_i}$, here σ^2 is the likelihood variance of the GP posterior. For this specific problem we have used a 0-mean function $\mu_0 \equiv 0$ and a Matern-5/2 covariance function or kernel $k(\boldsymbol{a}, \boldsymbol{a}')$, of length scale $= l$ and parameter $\nu = 5/2$.

E. Agent Models

Each agent performs sequential batch exploration, towards optimisation of an unknown stochastic reward function R. At each batch (i) we will choose $j = 1, 2, .., B$ policies $\boldsymbol{a}_i^j \in A$. Due to the computational expense of calculating $R(\boldsymbol{a}_i)$ and the size of the entire A, we wish to find solutions of maximal reward in as few batches i as possible. The goal being to approximate $\boldsymbol{a}^* = \text{argmax}_{\boldsymbol{a} \in A} R(\boldsymbol{a})$ without prohibitively expensive computation for all possible policies, therefore using a subset $A_c \in A$ of the policy space.

1) Upper/Lower Confidence Bound (GP-ULCB): We introduce the Gaussian Process Upper/Lower Confidence Bound (GP-ULCB) algorithm, inspired by Gaussian Process regression (GPR) and work on Upper Confidence Bound (UCB) solutions to the multi-armed bandit problem [13] [14]. This is a formulation which combines the natural confidence bounds of Gaussian Processes for stochastic multi-armed bandit problems, and variants have already been proposed in the form of GP-UCB [15] and GP-UCB-PE [16].

```
Result: $\boldsymbol{a}_i$: batch $i$
Input: random discretised actions $\boldsymbol{a} \in A_c$;
GP priors $\mu_0 = 0$, $\sigma_0, l$;
$B$ =batch size, $f_m$ =mixing factor, $f_c$ =masking factor;
for $i$ = 1,2,... do
    reset: $\boldsymbol{a}_{\text{upper}}, \boldsymbol{a}_{\text{lower}}, A_c$;
    for $j$ = 1,2,..,B. do
        if $j < B \times f_m$ then
            $\boldsymbol{a}_i^j = \underset{a \in A_c}{\text{argmax}}\ \mu_{i-1}(\boldsymbol{a}) + \beta\sigma_{i-1}(\boldsymbol{a})$
            mask: $\boldsymbol{a}_{\text{upper}}$, $|\boldsymbol{a^j} - \boldsymbol{a}_{\text{upper}}| < l \times f_c$
            update: $\boldsymbol{a}_{\text{upper}} \notin A_c$
        else
            $\boldsymbol{a}_i^j = \underset{a \in A_c}{\text{argmin}}\ \mu_{i-1}(\boldsymbol{a}) - \beta\sigma_{i-1}(\boldsymbol{a})$
            mask: $\boldsymbol{a}_{\text{lower}}$, $|\boldsymbol{a^j} - \boldsymbol{a}_{\text{lower}}| < l \times f_c$
            update: $\boldsymbol{a}_{\text{lower}} \notin A_c$
        end
    end
    Return: $R_\theta(\boldsymbol{a}_i)$
    Update Posterior: mean $\mu_i(\boldsymbol{a})$, variance $\sigma_i(\boldsymbol{a})$
end
```

Algorithm 1: GP-ULCB

The algorithm is initialised with a random sample of a discrete policy space (A_c). Subsequent policies are chosen to further explore the policy space regressed by GPR on all preceding simulation runs. The choice of using both *upper* and *lower* confidence bounds was made due to the stochastic nature of rewards. Specifically, minima and maxima can readily occur in the policy space necessitating a search for both potentially optimal and bad strategies. Also, by including the sampling of minima, the agent may present a risk adverse exploration of the policy space.

2) Genetic Algorithm: A genetic algorithm (GA) was implemented to provide comparison of another ‘black box’ optimisation technique for the exploration of the policy space. The GA is a biologically inspired, population-based search technique [17], specifically a meta-heuristic inspired by the process of natural selection. We use the reward generated for a policy as the measure of its fitness, and as OpenMalaria allows us to calculate a stochastic reward for each policy, there is noise in the fitness measure.

Given an evaluated population, in this case a set of policies and their stochastic rewards, the GA with then derive the next generation of the population. In this work we begin this process with roulette wheel selection [18] to select candidate policies. This biases selection of good policies to pass their 'genetic material' to the subsequent generation. The probability of selection p^j of the j^{th} policy in a generation (i.e. batch i), is defined in (11), where f^j is the fitness ($-R(\boldsymbol{a}_i)$ normalised $[0, 1]$).

$$p^j = \frac{f^j}{\sum_{k=1}^{B} f^k} \tag{11}$$

Each policy in the subsequent generation is derived from two policies selected via this approach. Mimicking biological crossover of chromosomes, the two selected policies are mixed, and one of the resulting policies selected at random. Finally, a random subset of the components of each derived policy is perturbed by adding noise. This sequence of processes defines how policies from the current generation are used to derive the next generation.

F. Batch Policy Gradient

The final approach used was a modified policy gradient [19], chosen as a method reported to handle continuous or very large action spaces in the case of this problem, while being robust to stochasticity. In this implementation $R_\theta(\boldsymbol{a})$ is approximated by a neural network, with new policies sampled through ϵ-greedy exploration. The negated rewards $-R(\boldsymbol{a_i})$ were normalised $[0, 1]$ from the batch results and the network trained to update it's weights ($\boldsymbol{w}$) associated with each policy $\boldsymbol{a} \in A_c$, using gradient descent on the negative log-loss of the batch normalised rewards.

During training the policy $\boldsymbol{a}^j$ (12) will be chosen with probability ϵ:

$$\boldsymbol{a}^j = \underset{w \sim A_c}{\text{argmax}}(\boldsymbol{w}) \tag{12}$$

While a random policy will be sampled from A_c with probability $1 - \epsilon$. Similarly to the GP-ULCB algorithm each $\boldsymbol{a}^j$ of batch i is sampled sequentially such that $\boldsymbol{a}^{j-1} \notin A_c$.

III. System Implementation and Deployment

For this work we used OpenMalaria commit a50730b. The simulation environment was run on a 4 node cluster of machines, each with 64 hyper threaded cores (2.20GHz Intel Xeon® CPU E5-2660). On these processors, running one instance of an OpenMalaria simulation, for a representative human population size (100,000), returns results in the timeframe of days. Running experiments in batches can thus take advantage of the natural parallelism of the deployment environment. Parallelism was implemented using Python's *multiprocessing* package. This package supports spawning processes using an API, and was used to execute OpenMalaria simulations, passing the scenario as an argument. This results in a natural expression of the batch size B for each agent model equal to the number of available processor cores.

IV. Results

Published studies [7] give a direct reference to human expert decisions made using OpenMalaria as a research tool, specifically to compare the cost-effectiveness of different interventions in Rachuonyo South District. Their findings stated that the current policy of 56% a_{ITN}, 70% a_{IRS} was the most cost-effective with regards to C_{DA}, while they also recommended that increasing this to 80% a_{ITN} and 90% a_{IRS} (including a school-based screen and treat program) would have the greatest health impact for DALYs averted. In this work we use the same stochastic parameterisation θ of OpenMalaria, but instead explore an automated answer to a less constrained problem for the decision maker: *given the current intervention strategy, what policy decisions can be made for the next 5 years to improve cost-effectiveness*?

Our results indicate that all three agent models extract the same top three performing emergent policies with respect to our primary evaluation metric, C_{DA} (See Table I):

- Maintain a_{ITN} and stop a_{IRS},
- Maintain a_{ITN} and reduce a_{IRS},
- Increase a_{ITN} and stop a_{IRS}.

These findings are extracted from the surface maxima of the posterior mean $\mu(\boldsymbol{a})$ through Gaussian Progress regression of rewards $R_\theta(\boldsymbol{a})$ collected by each respective agent. Figure 2 illustrates these surfaces.

V. Evaluation

The methods shown give a comprehensive evaluation, exploring the cost-effectiveness for a policy of the two main malaria interventions. Such fundamental insight is often missing from empirical studies, which are grounded in determining how much of a single intervention may be implemented to maximise a particular performance metric. Interestingly, at the surface our results challenge the current sentiment in the community - that policy makers in Sub-Saharan Africa should maximise ITN coverage *before* looking into other intervention strategies. One 2017 study even states that:

> coverage of ITNs was consistently the most cost-effective intervention across a range of transmission settings and was found to occur early in the cost-effectiveness scale-up pathway. IRS, RTS, S and SMC entered the cost-effective pathway once ITN coverage had been maximised. [20]

Instead, our results suggest that when the cost effectiveness of strategies is considered, after achieving a threshold of nets deployed for the environment, it may be more cost-effective to start spraying a small proportion of households (approx. 20-30%), instead of continuing to scale the deployment of insecticide treated nets.

If additional investment is available, then these resources should be allocated to scale up the coverage of bednets and further maximise health outcomes. With unlimited resources there are health benefits in the maximization of nets first, however as policy makers are facing tougher budget constraints, there are other factors that should influence the decisions they will make and any smaller investment can more effectively used to spray households. Our findings are exciting, however our contributions are limited as the current approach is only preliminary. We have not explored if interventions could be deployed at different times of year, or even if multiple policies

GP-ULCBII-E1				**Genetic Algorithm**				**Batch Policy Gradient**			
Policy	C_{DA}	DA	C_{int}	Policy	C_{DA}	DA	C_{int}	Policy	C_{DA}	DA	C_{int}
$\{\mathbf{60},\mathbf{4}\}$	28.9	11800	514000	$\{\mathbf{55},\mathbf{0}\}$	25.5	8690	458000	$\{\mathbf{55},\mathbf{8}\}$	27.6	11600	477000
$\{58, 33\}$	30.1	12300	519000	$\{55, 21\}$	26.8	11600	477000	$\{55, 28\}$	30.2	11000	489000
$\{69, 0\}$	30.9	13000	579000	$\{76, 0\}$	28.3	11700	632000	$\{68, 7\}$	30.3	13800	589000

TABLE I: Top three policies with respect to C_{DA} evaluated by each agent model. Policy: $\{a_{\mathrm{ITN}}\%, a_{\mathrm{IRS}}\%\}$, C_{DA}: Cost per DALY Averted USD, DA: DALYs Averted, C_{int}: Intervention Costs USD.

could have been concurrently deployed in a population. Further, the current simulation did not permit interventions to be targeted to specific subsets of the population (e.g. households with young children). Finally, this work is specific to one studied location in Western Kenya, and the generalisation of the insights from multiple agents gathering insights across expansive environments e.g. Sub-Saharan Africa is yet to be explored.

VI. Further AI Research

The techniques presented have been selected and designed with the view to deployment on larger policy spaces as detailed in the Evaluation section. More compute time, expansive environments and policies are a requirement for the real-world human decision maker. While other data sources exist outside of the OpenMalaria simulation environment, notably malaria mapping projects which could serve as visual input, requiring function approximation through Deep Learning. There is also the existing opportunity to embed the agent model deeper into the simulation environment, passing control of simulation parameters in order to allow balancing of computational expense with efficient policy space exploration. This work is viewed as an emerging application for deploying further novel exploration techniques.

References

[1] P. Winskill, M. Rowland, G. Mtove, R. C. Malima, and M. J. Kirby, "Malaria risk factors in north-east Tanzania," *Malaria Journal*, vol. 10, no. 1, p. 98, 2011. [Online]. Available: http://malariajournal.biomedcentral.com/articles/10.1186/1475-2875-10-98

[2] M. Moran, *The malaria product pipeline: planning for the future*. Health Policy Division, The George Institute for International Health, 2007.

[3] J. D. Piette, S. L. Krein, D. Striplin, N. Marinec, R. D. Kerns, K. B. Farris, S. Singh, L. An, and A. A. Heapy, "Patient-Centered Pain Care Using Artificial Intelligence and Mobile Health Tools: Protocol for a Randomized Study Funded by the US Department of Veterans Affairs Health Services Research and Development Program," *JMIR Research Protocols*, vol. 5, no. 2, p. e53, apr 2016. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856067/

[4] N. B. Dimitrov and D. P. Morton, "Combinatorial design of a stochastic Markov decision process," *Operations Research/ Computer Science Interfaces Series*, vol. 47, pp. 167–193, 2009.

[5] T. Smith, N. Maire, A. Ross, M. Penny, N. Chitnis, A. Schapira, A. Studer, B. Genton, C. Lengeler, F. Tediosi, D. De Savigny, and M. Tanner, "Towards a comprehensive simulation model of malaria epidemiology and control," *Parasitology*, vol. 135, no. 13, p. 1507, 2008. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0031182008000371

[6] E. M. Stuckey, J. C. Stevenson, M. K. Cooke, C. Owaga, E. Marube, G. Oando, D. Hardy, C. Drakeley, T. A. Smith, J. Cox, and N. Chitnis, "Simulation of malaria epidemiology and control in the highlands of western Kenya," *Malaria Journal*, vol. 11, no. 1, p. 357, oct 2012. [Online]. Available: http://dx.doi.org/10.1186/1475-2875-11-357

[7] E. M. Stuckey, J. Stevenson, K. Galactionova, A. Y. Baidjoe, T. Bousema, W. Odongo, S. Kariuki, C. Drakeley, T. A. Smith, J. Cox, and N. Chitnis, "Modeling the cost effectiveness of malaria control interventions in the highlands of western Kenya," *PLoS ONE*, vol. 9, no. 10, 2014.

[8] C. J. L. Murray and A. D. Lopez, "The global burden of disease: a comprehensive assessment of mortality and disability from deceases, injuries and risk factors in 1990 and projected to 2010," *Harvard University Press*, vol. 1, pp. 1–35, 1996.

[9] I. Network, "INDEPTH model life tables for sub-Saharan Africa," *Aldershot (England): INDEPTH Network*, 2004.

[10] O. J. Briët, M. a. Penny, D. Hardy, T. S. Awolola, W. Van Bortel, V. Corbel, R. K. Dabiré, J. Etang, B. G. Koudou, P. K. Tungu, and N. Chitnis, "Effects of pyrethroid resistance on the cost effectiveness of a mass distribution of long-lasting insecticidal nets: a modelling study." *Malaria journal*, vol. 12, no. 1, p. 77, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23442575

[11] F. Tediosi, N. Maire, M. Penny, A. Studer, and T. a. Smith, "Simulation of the cost-effectiveness of malaria vaccines." *Malaria journal*, p. 127.

[12] C. K. I. Williams and C. E. Rasmussen, "Gaussian processes for regression," *Advances in Neural Information Processing Systems 8*, vol. 8, no. August, pp. 514–520, 1996. [Online]. Available: http://eprints.aston.ac.uk/651/

[13] P. Auer and R. Ortner, "UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem," *Periodica Mathematica Hungarica*, vol. 61, no. 1, pp. 55–65, 2010.

[14] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[15] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design."

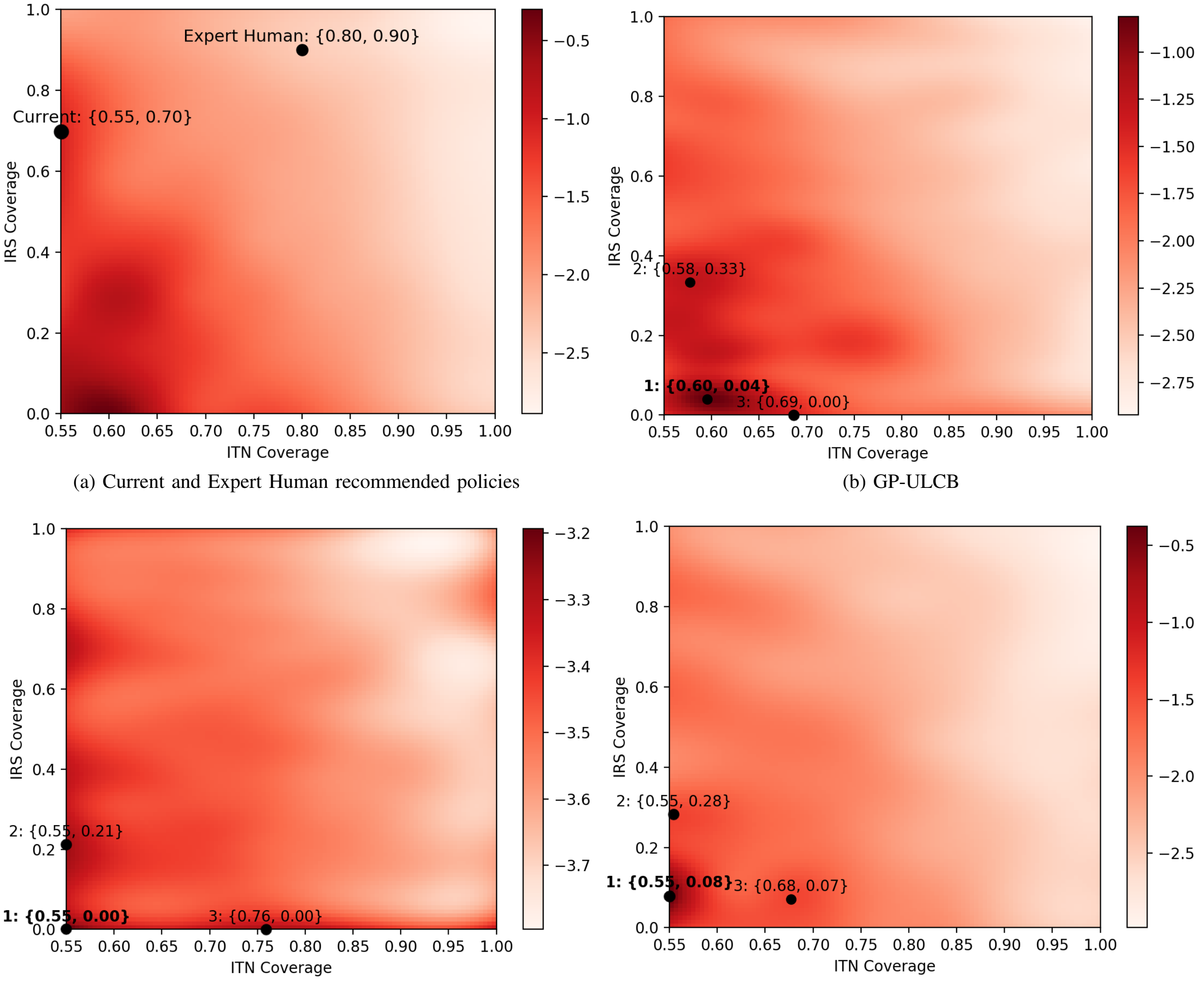
[16] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis, "Parallel Gaussian process optimization with upper confidence bound and pure exploration," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8188 LNAI, no. PART 1, pp. 225–240, 2013.

[17] J. H. Holland, "Genetic Algorithms - Computer programs that ”evolve” in ways that resemble natural selection can solve complex problems even their creators do not fully understand," pp. 66–72, 1992.

[18] D. E. Goldberg and K. Deb, "A Comparative Analysis of Selection Schemes Used in Genetic Algorithms," *Foundations of Genetic Algorithms*, vol. 1, pp. 69–93, 1991.

[19] R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," *In Advances in Neural Information Processing Systems 12*, pp. 1057–1063, 1999.

[20] P. Winskill, P. G. T. Walker, J. T. Griffin, and A. C. Ghani, "Modelling the cost-effectiveness of introducing the RTS,S malaria vaccine relative to scaling up other malaria interventions in sub-Saharan Africa," *BMJ Global Health*, vol. 2, no. 1, p. e000090, 2017. [Online]. Available: http://gh.bmj.com/lookup/doi/10.1136/bmjgh-2016-000090



(a) Current and Expert Human recommended policies

(b) GP-ULCB

(c) Genetic Algorithm

(d) Batch Policy Gradient

Fig. 2: Visualisations of $-\log(R_\theta(\boldsymbol{a}))$ policy surfaces. Generated from each agent, after 8 iterations of a 64 batch size, running a total 512 simulations. All regressed with the same Gaussian Process parameters. The surfaces describe each agent's exploration of the available actions a_{ITN} and a_{IRS} for a human decision maker in Rachuonyo South.