Combining Kernel and Model Based Reinforcement Learning for HIV Therapy Selection

Sonali Parbhoo
University of Basel
Basel, Switzerland
sonali.parbhoo@unibas.ch

Jasmina Bogojeska
IBM Research - Zurich
Zurich, Switzerland
JBO@zurich.ibm.com

Volker Roth
University of Basel
Basel, Switzerland
volker.roth@unibas.ch

Finale Doshi-Velez
Harvard University
Cambridge, MA.
finale@seas.harvard.edu

Human Immunodeficiency Virus (HIV-1) currently affects over 36 million people worldwide. To date, the only practical treatment for HIV is life-long administration of combinations of antiretrovirals which target different phases of viral life cycle. Despite the development of new antiretrovirals, every treatment ultimately fails due to the high evolutionary dynamics of the virus, which enable it to escape to resistance by accumulating resistance mutations. This, together with the large number of potential therapy combinations, makes manually searching for an effective therapy challenging, especially for patients with long treatment histories.

The vast majority of computational approaches developed to meet this challenge are based on regression [1]. These map elements of a patient's history directly to some output, such as virological response, and use this to infer appropriate treatments. Recently, [2] present a kernel-based method for predicting treatment outcome based on a patient's treatment history. The premise is that patients with similar treatment histories may respond similarly. Unfortunately, none of these approaches directly address the sequential nature of the therapy selection process — that a choice of combination now might result in drug-resistant viral strains which may be harder to control later. Reinforcement learning approaches, such as [3] and [7], make this sequential nature explicit: they output a treatment policy that chooses therapy combinations not only to optimise virological response in the present, but also in the future. However, these approaches are fragile since they reason about futures on the basis of limited data. Moreover, the heterogeneity in patient data makes it difficult for one particular model to succeed at providing suitable therapy predictions for all patients.

To overcome these problems, we present a mixture-of-experts approach [5] that combines the strengths of both kernel-based regression methods and reinforcement learning for HIV therapy selection. Kernel-based regression methods excel when there are clusters of similar patients: they can model patient-specific aspects in viral response. However, their prediction quality drops when patients are not part of a tight cluster. On the other hand, model-based methods first build a model to reason about how well a series of therapy selections will perform. These approaches tend to find simpler, more robust patterns of response — a better alternative for patients outside of clusters. The mixture-of-experts approach automatically selects between these two options, depending on a patient's particular situation.

Our contributions are as follows: we show that optimising for an immediate reduction in viral load does not control mutations or viral loads in the future. We demonstrate that the therapy combinations proposed by our treatment policy outperform previous methods. Finally, we support our claim that the kernel-based approach is used when a patient lies in a cluster, while the model-based approach is used for patients with few neighbours. In summary, this suggests that more nuanced approaches are required to make optimal treatment recommendations for patients with HIV.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

1 Model

We propose a mixture-of-experts for HIV therapy selection. Our first expert modifies the history alignment model due to [2] to optimise long-term outcomes rather than immediate outcomes. Our second expert is a Bayesian Partially Observable Markov Decision Process (POMDP). Combining these approaches allows us to learn reasonable choices of therapy for individual patients, specific to their situations.

Cohorts We make use of a subset of the EuResist database consisting of HIV genotype and treatment response data for 32 960 patients, together with their corresponding CD4$^+$ and viral load measurements, gender, age, risk group, and number of past treatments recorded. We limit ourselves to the 312 most common drug combinations that occur in the cohort. The database has previously been used to build models such as the therapy alignment model, to predict the outcome of a particular therapy [13, 9]. We are however, specifically interested in optimising the therapy choice for a particular patient.

Long-term Reward Criterion Following [3, 7], we propose the following immediate reward function:

$$r_t = \begin{cases} -0.7 \log V_t + 0.6 \log T_t - 0.2|M_t|, & \text{if } V_t \text{ is above detection limits} \\ 5 + 0.6 \log T_t - 0.2|M_t|, & \text{if } V_t \text{ is below detection limits,} \end{cases}$$

where V_t is the viral load (in copies/mL), T_t is the CD4$^+$ count (in cells/mL), and $|M_t|$ is the number of mutations at time t respectively. This function penalises instances where a patient's viral load increases and rewards instances where a patient's CD4$^+$ count increases (more weight is placed on the viral load, as it is an earlier indicator of whether a therapy is working). We also penalise on the basis of the number of mutations a patient has at a particular time, as these may ultimately contribute to resistance and therapy failure. There is also a bonus for if the viral load is below detectable limits as this is something we would like to sustain over time. The long-term reward criterion sums these immediate rewards over the patient's history.

Kernel-based History Alignment It is well-established that a patient's prior history is a key factor for predicting the efficacy of HIV treatment [11, 10]. [2] use this fact in a history alignment model that measures the similarity between two therapy sequences[1]. Two therapy sequences are considered to be similar if they consist of similar drug combinations, are administered in a similar order, and produce similar genomic fingerprints in the viral population. If two patients have similar histories, [2] demonstrate that they often respond similarly to treatment.

The history-alignment model first uses a resistance mutations kernel to quantify the pairwise similarities between different therapy combinations. The kernel assumes that similarity between the different drug groups is additive, since drugs belonging to different groups have different therapeutic targets and could thus be assumed to act independently. The pairwise similarity between two drug combinations, z and z' is given by averaging the similarities of their corresponding drug groups g:

$$k(z, z') = \sum_{g \in G} \frac{\texttt{sim}_g(z, z')}{|G|}; \quad \texttt{sim}_g(z, z') = \frac{u_{zg}^\top u_{z'g}}{\max(||u_{zg}||^2, ||u_{z'g}||^2)} \tag{1}$$

where u_{zg} and $u_{z'g}$ are binary vectors of resistance relevant mutations occurring in drug group g of z and z' respectively.

With this similarity measure between drug combinations, [2] then compute a similarity score between therapy sequences via the Needleman-Wunsch sequence alignment algorithm [6]. The alphabet for the sequence comprises the distinct drug combinations in the data set, while the mutations kernel in Equation 1 determines the pairwise similarities between the characters of the alphabet. The score from the alignment similarity kernel, together with the viral genotype and drug history information, can then be used to train a regression model for predicting the outcome of a therapy in terms of success or failure. [2] define a therapy as successful if that patient's viral load falls below 400 copies/mL after 21 days of treatment under the therapy. We replace this success criteria with the

[1]The combinations of drugs a patient takes at a particular time is defined as a therapy. A therapy sequence refers to a sequence of such combinations over time.

potential long-term value of a therapy choice. First, we convert the binary problem of 'Will this drug combination succeed?' into a multi-class problem of 'Which drug combination should I choose?'. Second, and more importantly, this prediction is made by summing the long-term reward criterion over *all* the time-steps in the patient's future history. We call this treatment policy the 'Long-Term History Alignment' model.

Bayesian POMDP The HIV therapy selection problem involves making a sequence of decisions with long-term consequences. Reinforcement learning formalises this process as a series of exchanges between an agent and its environment. At each time step, the agent selects an action a (such as a drug combination) and the environment returns some observations o (e.g. CD4 counts, viral loads, mutations) as well as an immediate reward r (e.g. whether the viral load drops below a desired threshold). Given a history of length t, $h = \{a_1, o_1, r_1 \ldots, a_t, o_t, r_t\}$, the agent's goal is to choose the subsequent action such that it maximises discounted sum of its expected rewards, $\mathbb{E}[\sum_t \gamma_t r_t]$, where $\gamma \in [0, 1)$ trades off between current and future rewards.

A POMDP m is defined by a finite set of hidden states $\mathcal{S}$, actions $\mathcal{A}$ and observations $\mathcal{O}$. A transition function $T(s'|s, a)$ specifies the probability of transitioning from state s to s' when taking an action a. Similarly, an observation function $\Omega(o|s, a)$ specifies the probability of observing o from state s when taking action a. The reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the immediate reward that an agent receives upon performing an action from a particular state. Evidently, the true state of a patient corresponding to their underlying health status may not be directly observable. We limit our actions to the 312 most common drug combinations in the cohort and learn a model with 7 hidden physiological states. Our observation space consists of (a) binning the values of the viral load using a log scale of [0.0, 1.0, 10.0, 1000, 10000, 10M, 100M] copies/ML, (b) 70 resistance mutations that may occur as a result of a particular therapy together with a patient's CD4$^+$ count, gender, risk group. We treat the parameters for the transitions and observations in a Bayesian fashion by placing Dirichlet priors on them. Inference here, entails sampling multiple models of the parameters and updating the agent's beliefs about the patient state accordingly.

We model time in discrete increments of 6 months, and perform a forward search for therapy choices that optimise outcomes over a 5 year horizon (10 total steps). In our results, we evaluate the performance of each of the methods using three off-policy evaluation schemes: importance sampling [8], weighted importance sampling [12], and doubly robust evaluation [4].

2 Results and Discussion

	Doubly Robust	**Importance Sampling**	**Weighted Importance**
Random Policy	–2.31 $\pm$ 1.42	–3.48 $\pm$ 1.36	–2.80 $\pm$ 1.27
Short-term History Alignment	2.17 $\pm$1.47	2.14 $\pm$1.22	2.15 $\pm$1.16
Long-term History Alignment	9.48 $\pm$ 1.90	5.42 $\pm$ 1.93	6.74 $\pm$ 1.89
POMDP	6.34 $\pm$ 2.15	4.36 $\pm$ 2.38	6.76 $\pm$ 2.24
Mixture-of-experts	**11.47 $\pm$ 1.38**	**12.25 $\pm$ 1.41**	**11.23 $\pm$ 1.40**

Table 1: Off-Policy evaluation using importance sampling, weighted importance sampling and doubly robust methods for different therapy selection models.

Table 1 compares the performance of the history alignment method, the POMDP and the mixture-of-experts against a random policy where a completely random therapy choice is made. A higher value indicates a better performing treatment policy.

Optimising for long-term health produces different treatment policies than predicting the most common next therapy Table 1 shows that the short-term history alignment model achieves significantly worse rewards than the long-term history alignment model (and the POMDP). These results suggest that treatments which may initially appear attractive may result in poor patient outcomes at a later stage—unsurprising to many in HIV. Specifically, resistance against a particular drug may lead to cross-resistance against another, leading to long-term dependencies in therapy response.

The mixture-of-experts produces the best treatment policies. The mixture-of-experts approach outperforms the other approaches across all evaluation schemes. While the POMDP performs worse than the long-term alignment kernel in general, the fact that the mixture of experts approach outperforms both the POMDP and alignment kernel suggests that these models are making mistakes in different places—and thus we can do better by choosing between the two. A post-hoc investigation reveals that the mixture-of-experts chooses the POMDP model approximately 26% of the time in comparison to the alignment kernel.

The mixture-of-experts chooses experts based on clustering characteristics. We follow up on our clustering hypothesis: when is each model being chosen? Specifically, we consider role that a patient's history and the lower quantile of their distance to other patients plays here. Figure 1 (a) and (b) provide box plot illustrations of the values of the latter features in relation to the choice of model selected by the mixture-of-experts. As the lower quantile of the distance between a patient and their neighbours increases, the POMDP is more likely to become the model of choice. Moreover, the length of a patient's therapy history seems to play a defining role in the choice of expert. The mixture-of-experts selects the POMDP for patients with longer history lengths and the alignment kernel for the others. This is likely because as a patient's treatment history increases in length, they become more unique and have smaller similarity values relative to other patients in the kernel. Here the POMDP is the model of choice, possibly because it is able to incorporate this rich history into its belief state, while the alignment kernel cannot capture the same level of information.

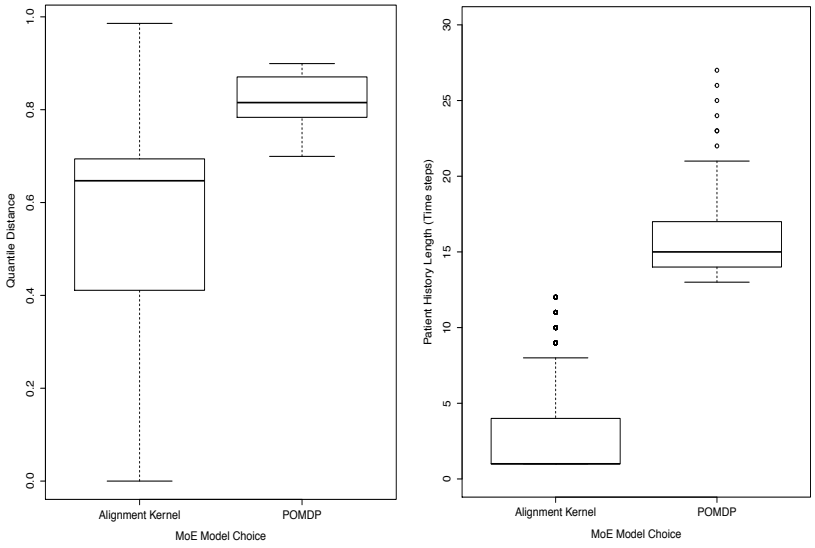

Figure 1: Mixture-of-experts model choice over (a) distances to closest neighbour and, (b) varying history lengths.

3 Conclusion

We showed how kernel methods and model-based RL can be combined for HIV therapy selection using a mixture-of-experts approach. This enabled us to account for heterogeneity in patient data, that typically makes it difficult for a single model to provide reasonable therapy predictions for individuals. Specifically, we showed that the kernel approach is optimal for patients with short treatment histories; the model-based method proved more suited to patients with long treatment histories and rare therapy combinations. These are patients that have been heavily pre-treated. We attribute this difference to the way in which each model uses a patient's history: the POMDP incorporates knowledge about a patient's history implicitly through its beliefs and actions, each influenced by past observations, treatments and mutations, while the history alignment method only uses patient's treatment history from the kernel, and does not account for observations that occur further back in time. By combining

the methods, we are able to automate the task of selecting an appropriate model for a particular patient. In this way, we draw on the strengths of each approach, and eliminate the need to choose between the methods. This ultimately aids in improved decision-making.

References

[1] André Altmann, Niko Beerenwinkel, Tobias Sing, Igor Savenkov, Martin Däumer, Rolf Kaiser, Soo-Yon Rhee, W Jeffrey Fessel, Robert W Shafer, and Thomas Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral therapy*, 12(2):169, 2007.

[2] Jasmina Bogojeska, Daniel Stöckel, Maurizio Zazzi, Rolf Kaiser, Francesca Incardona, Michal Rosen-Zvi, and Thomas Lengauer. History-alignment models for bias-aware prediction of virological response to hiv combination therapy. In *AISTATS*, pages 118–126, 2012.

[3] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

[4] Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

[5] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

[6] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

[7] S. Parbhoo. *A Reinforcement Learning Design for HIV Clinical Trials*. University of the Witwatersrand, Faculty of Science, School of Computer Science, 2014.

[8] Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

[9] Mattia CF Prosperi, Andre Altmann, Michal Rosen-Zvi, Ehud Aharoni, Gabor Borgulya, Fulop Bazso, Anders Sönnerborg, Eugen Schülter, Daniel Struck, Giovanni Ulivi, et al. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antivir Ther*, 14(3):433–42, 2009.

[10] Mattia CF Prosperi, Michal Rosen-Zvi, André Altmann, Maurizio Zazzi, Simona Di Giambenedetto, Rolf Kaiser, Eugen Schülter, Daniel Struck, Peter Sloot, David A Van De Vijver, et al. Antiretroviral therapy optimisation without genotype resistance testing: a perspective on treatment history based models. *PloS one*, 5(10):e13753, 2010.

[11] AD Revell, D Wang, R Harrigan, RL Hamers, AMJ Wensing, F Dewolf, M Nelson, A-M Geretti, and BA Larder. Modelling response to hiv therapy without a genotype: an argument for viral load monitoring in resource-limited settings. *Journal of antimicrobial chemotherapy*, page dkq032, 2010.

[12] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 707. John Wiley & Sons, 2011.

[13] Maurizio Zazzi, R Kaiser, A Sönnerborg, Daniel Struck, Andre Altmann, Mattia Prosperi, M Rosen-Zvi, A Petroczi, Y Peres, E Schülter, et al. Prediction of response to antiretroviral therapy by human experts and by the euresist data-driven expert system (the eve study). *HIV medicine*, 12(4):211–218, 2011.