Framework for Optimal Budget Allocation of HIV Intervention Policies

Ali El Moselhy

July 14, 2023

1 Personal Experience

I began my journey into research in 10th grade (2020). Covid-19 was the style du jour, masks were all the rage, and I was enrolling in a class at my school called Science Research. This class was meant to guide students, helping them reach out to experienced mentors in academia and industry. Initially, I was tasked with finding a broad area of study, for example, cancer, in which I could later specify and find a niche that fit me precisely. Even at this early stage, however, I was torn. I had always loved mathematics, but I also adored biology and any topics which had to do with the human body, and I was struggling with finding a topic that neatly tied the two together. It was on a random winter day during the weekend, when I went to a diner with my father for breakfast. There, they had a TV set playing the news, where the latest projected Covid-19 numbers were being displayed. It was at that point my father suggested to me that I should research disease models. There was heavy mathematics used in the creation and analysis of these models, while a strong understanding of diseases and disease epidemiology was required to accurately understand and implement disease models. After some basic investigation, I was smitten.

For the rest of the year, I became engrossed in epidemiology, learning about compartmental and agent-based models, disease routes of infection and interventions, and quite excitingly, following the most up-to-date research on Covid-19. When the summer rolled around, I eagerly sent emails to research groups working with disease models. A group at NYU Langone's Translational Research Department wrote back, and I eagerly began interning at their lab. They worked more on HIV than Covid-19, however, so for the first summer, I investigated their HIV model, EMOD, and learned how to interface with it. I also began learning about HIV itself and applied that knowledge to understanding EMOD. My mentor had also mentioned to me that EMOD, their model, was unable to allocate money to different disease treatment programs. Including this capability was their next goal, and so I also began studying optimization algorithms, specifically those used for more complex problems. I had always loved mathematics, a trait I got from my father, who was trained as an engineer and had been trying to teach me similarly since I was 3. For my research, I had to learn advanced linear algebra to understand the optimization techniques used. To this day, there are still a few finer details I struggle to verbalize. It was also incredibly satisfying to fully grasp and just understand what was going on after a full day of work. That moment, of comprehension after immense struggle, was what made the difficulty and challenge worth it.

By the time my 11th grade summer rolled around, I had absorbed enough knowledge from my mentor, various online sources, as well as some other professors I had reached out to. I felt ready

to begin a project of my own: given an amount of money, a list of interventions (things to do to reduce the effect of HIV), and a disease model which could simulate different investment strategies, I needed to find the most efficient investment strategy (which reduced deaths the most). What I did is detailed in the next section. Research was a fantastic experience for me. It felt more freeing than the slog of work typical of school, and while it was more difficult, it also felt significantly more rewarding to work out a kink in an algorithm than to score well on a test. For those of you starting research as high schoolers, I have a few pointers:

- Do something you love. There are infinite topics on which one can work and if you're spending most of each day on research, you'd better love it.
- Research is not like schoolwork; you should expect to fail multiple times before you find any measure of success. If your code works the first time around, something probably went wrong.
- Your mentor is one of the best sources of information, help, and support you could ever ask for. I was lucky with my mentor, she was incredibly kind and gave me full access to her team and resources, and I am eternally thankful for how she treated me and the time and consideration she gave me. While mentors may seem intimidating, they often want to help and are more than willing to explain complex topics or techniques.
- Packaging and presentation are almost as important as the work itself. The first time one of my relatives asked me about my work, I stumbled into the densest explanation of disease models, tensors, and HIV interventions that I could have possibly given. She probably thought I had no idea what I was talking about. If no one understands your work, it's almost as if you haven't done it.

2 Introduction

Around the world, there exist over 37.9 million people living with HIV (PLHIV), and HIV has been responsible for over 35 million deaths to-date (UNAIDS, 2021). The most affected region of Africa is Eastern and Southern Africa, with 20.6 million PLHIV, accounting for around 54% of all PLHIV (Kharsany, 2016).

Interventions are public health measures taken in order to limit the spread of HIV (Meyer-Rath, 2019). This can range from treatments, such as Anti-retroviral therapy, to preventatives, such as Post- and Pre-Exposure Prophylaxis, to additional programs meant to increase awareness and access to care, like Antenatal Care and Home Counseling and Testing programs. Intervention programs are expensive; yearly costs for medications can be over $100 USD, and non-medical interventions, although cheaper, are still not free (Meyer-Rath, 2019). Since 2021, annual resources for HIV treatment available to Eastern and Southern Africa have decreased by 5%. (Kharsany, 2016). There exists a need for efficient allocation of resources to maximise the impact of every dollar spent in the region (Kharsany, 2016).

The Optima group has previously used their Optima HIV model in allocative efficiency studies (Kerr, 2015). The Optima group has published numerous reports on budget optimization, typically focusing on a single country (Kerr, 2015). In their analyses, the Optima group only seem to focus on around 7 intervention strategies, while it may be of value to analyze larger intervention portfolios (Lukoba, 2020). For our work, we will be using the Epidemiological MODeling software (EMOD) model. Please note that, as opposed to the deterministic Optima model, EMOD is a stochastic model, which will motivate many of our decisions concerning optimization procedures.

EMOD is a stochastic agent-based model developed by the Institute for Disease Modeling (IDM) (Bershteyn, 2018). EMOD is designed to be highly modular, allowing for many different diseases and regions to be simulated, based upon the input files provided to the model (Bershteyn, 2018). For this work, we will be using two sets of input files, the first modeling HIV in South Africa and the second modeling HIV in Kenya. These input files were calibrated by Prof. Kim in the Bershteyn Lab group (Bershteyn, 2018).

EMOD typically takes 15 minutes to complete a simulation, prohibitively long for most optimization algorithms. Case in point, evaluating 100 simulations would take 25 hours to run. In order to reduce evaluation time, we will use surrogate models to approximate the value of EMOD. Specifically, we will use a form of surrogate model based on tensor trains.

Tensor trains were originally conceived in a seminal paper by Prof. Ivan Oseledets, as high-dimensional extensions of the SVD process (Oseledets, 2011). They were further developed in works by Prof. Alex Gorodetsky and Prof. Youssef M. Marzouk, who used them in approximations of high-dimensional functions (Bigoni, 2016) (Gorodetsky, 2018). Tensor Trains were also used in work by Prof. Zheng Zhang in uncertainty quantification (Zhang, 2015). We use tensor trains because they are highly differentiable, quick to fit and evaluate, and they do not require continuous, online data gathering for the fitting process (Oseledets, 2011). They can be calibrated with previously gathered data, which simplifies the fitting process.

In this paper, we propose a framework for intervention optimization based around the EMOD epidemic model and utilizing surrogate models for fast model evaluations. By using tensor trains, our framework escapes the curse of dimensionality, and the high differentiability of tensor trains allows us to use a gradient-based optimization algorithm. Our framework does not use online data gathering, and can be fit to already-available data.

This paper is organized as follows: In section 3, we describe EMOD in greater detail. In section

4, we define the objective functions and metrics we will be using for the remainder of the paper. In section 5, we describe the proposed algorithm, and in section 6, we report the results of our application of the proposed algorithm to two budget optimization problems. We conclude this paper in section 7, with a discussion on the implication of our results and directions for future work.

3 Model Description

Epidemic MODel (EMOD) is a stochastic, agent-based disease model (Bershteyn, 2018). EMOD simulates the propagation of an endemic disease within a population (Bershteyn, 2018). EMOD has enough degrees of freedom in order to allow the user to accurately fit a variety of populations and diseases (Bershteyn, 2018). In this paper, we will be using config files that have been calibrated by Prof. Kim in the Bershteyn Lab group to model the HIV epidemics in both South Africa and Kenya (Bershteyn, 2018).

Upon specifying a parameter set, EMOD generates a sample population consistent with the input parameters, and starts the process of modelling the disease propagation among that population across discrete time steps for a period of 70 years (Bershteyn, 2020). At each time step, there is a probability that any one agent will transition from one state to the other. For example, a person who is HIV negative may become HIV positive (Bershteyn, 2020). A single sample results in one potential path for disease propagation. To minimize the effect of randomness, EMOD runs 50 samples for each set of input parameters. The results are then averaged in order to get a single final result for a given simulation (Bershteyn, 2018).

Within EMOD, there are certain interventions which can be used to limit the spread and effect of HIV. In this paper, we will deal with 5 interventions: Anti-Retroviral Therapy (ART), Pre-Exposure Prophylaxis (PrEP), Voluntary Male Medical Circumcision (VMMC), Ante-Natal Care (ANC) and Home Counseling and Testing (HCT). AntiRetroviral Therapy (ART) is a medication taken by PLHIV, which, after being taken for a short period of time, lowers the amount of HIV found in their bodies, to undetectable levels, meaning that PLHIV on ART can live as if they do not have HIV (Meyer-Rath, 2019). Pre-Exposure Prophylaxis (PrEP) is a medication that is taken by those who do not have HIV and greatly reduces their chances of infection upon exposure to HIV (Meyer-Rath, 2019). People on PrEP are typically high-risk individuals (Meyer-Rath, 2019). Voluntary Male Medical Circumcision (VMMC) is a form of circumcision that lowers the chance for a man to both receive and transmit HIV (Meyer-Rath, 2019). Ante-Natal Care (ANC) is a form of testing and medication given to pregnant women to reduce their probability of passing HIV to their child (Meyer-Rath, 2019). Home Counseling and Testing (HCT) is a form of outreach which reaches people about the symptoms of HIV and encourages them to get tested (Meyer-Rath, 2019).

4 Fitness Functions

Money is a limiting factor for the interventions that can be implemented within any population, so figuring out the combination of interventions that best reduces HIV-attributed death for any given amount of money is a matter of life or death (Sharma, 2021).

Cost is the total amount of money spent on interventions past a threshold year, t_0

$$\text{Total Cost} = \sum_t \sum_i (N_{ti} * C_i * df^{t-t_0})$$

N_{ij} is the number of times intervention i is given in year t. We can obtain this from the simulation results, C_i is the per-unit cost for intervention i (Meyer-Rath, 2019). *df* is the decay factor, typically 0.97. This factor is applied to the product to mitigate uncertainty caused by randomness.

To compute HIV-attributed death we use a quantity called disability adjusted life years (DALYs). Given a PLHIV who has died from HIV, DALYs are calculated by taking their life expectancy and subtracting their lifespan. In short, DALYs are the measure of how many years were taken from them due to HIV.

We begin counting DALYs after t_0. For each person who died due to HIV:

$$\text{DALYs} = \sum (L_{\text{Expected}} - L_{\text{Realized}}) * df^{t-t_0}$$

t is the time of death, in years. L_{Expected} is the expected lifespan, which we obtain from UN average lifespan measures (United Nations, 2019). L_{Realized} is the actual lifespan, which we obtain from the simulation results *df* is defined as before.

Our optimization problem is as follows:

$$\begin{aligned} \min_{x_0, x_1 \cdots x_{n-1}} \quad & \text{DALY}(x_1, x_2, \cdots x_{n-1}) \\ \text{s.t.} \quad & \text{Cost}(x_1, x_2, \cdots x_{n-1}) < \text{Budget} \end{aligned} \tag{1}$$

5 Proposed Algorithm

Our goal is to find the optimal investment strategy to minimize DALYs under certain budget constraints. We implement our analysis within the EMOD framework.

Each parameter represents a probability of an agent moving from one state to another. Therefore, all our parameters range between [0-1]. Note that although some parameters accept an range of 0-1, it does not make practical sense for them to take on certain values, so we will limit their input range.

We define the optimal investment strategy as a set of input parameters to EMOD. This set of input parameters controls the usage of interventions. Note that these parameters do not represent population characteristics or disease characteristics. For the purpose of describing our algorithm to solve the above optimization problem, we will focus on the three main interventions, ART, PrEP, and VMMC. In the results section, we will give examples where we have many more interventions.

To solve the optimization problem, we naturally need to run many simulations corresponding to different parameter sets. As previously stated, each such simulation takes approximately 15 minutes to run. This results in an extremely slow optimization problem. In order to reduce the time needed to solve the optimization problem, we propose to use a surrogate model-based approach.

In this approach, we approximate the EMOD simulation using a surrogate model. That model typically less than one second to evaluate. After each optimization cycle (which involves many evaluations of the surrogate model), we reach a hypothesized optima. To iterate, this optima is based on the surrogate model and may or may not hold true under the EMOD simulation. To mitigate uncertainty, we simulate this optima using the EMOD simulation. If the EMOD simulation agrees that the hypothesized optima is indeed optimal, then we stop our algorithm. Otherwise, we refit our surrogate model on the new data, and rerun the optimization algorithm. Therefore, the main algorithm consists of the following steps:

1. An initial sampling of parameter space to obtain 'training' data.
2. While True:
   (a) Construct surrogate models using the training data.
   (b) Find the optimal allocation policy using the surrogate models.
   (c) Simulate this hypothesized optimum.
   (d) If the simulation output satisfies our stopping criteria, BREAK.
   (e) Otherwise, augment training data using new data and loop.

There are two important ingredients to make the above work. First is the optimization algorithm. For that, we use Newton's algorithm (a gradient-based algorithm). This algorithm is explained in section 5.2. The second ingredient is the surrogate model. For that we use Tensor Trains as described in the following section.

5.1 Tensor Train

The surrogate model's aim is to mimic the forward simulation, evaluating in a fraction of the time, yet preserving as much accuracy as possible. For that, we use tensor trains (Oseledets, 2011).

The main reason behind using tensor trains is that it allows us to find a high-dimensional surrogate model while breaking the curse of dimensionality (Oseledets, 2011). The number of degrees of freedom to fit the model grows linearly with the number of input parameters (and not exponentially) (Oseledets, 2011).

A tensor train is an approximation of a complex model in terms of the product of univariate matrix functions. This can be interpreted in the context of separation of variables and/or high-dimensional SVD:

$$\hat{f}(x_0, x_1, \cdots x_{n-1}) = G_0^T(x_0)G_1(x_1)\cdots G_{n-1}(x_{n-1})$$
$$G_i(x_i) = \sum_{j=0}^{k} b_j(x_i)C_{ij}$$

$G(x_i)$ is a matrix function of the variable x_i, and is of dimension $r_i \times r_{i+1}$. Note that $r_0 = r_n = 1$. This means that the end product of all the $G_i(x_i)$'s is a scalar. We chose to approximate these matrix functions in terms of monomial basis functions, where $b_j(x_i) = \frac{x_i^j}{j!}$ and C_{ij} is the coefficient matrix of dimension $r_i \times r_{i+1}$. Fitting the surrogate model involves computing all C_{ij} in order to best approximate $f(x_0, x_1, \cdots x_{n-1})$.

5.1.1 Fitting Procedure

To fit our model, we need data. We use an offline procedure to collect outputs from EMOD corresponding to random samples of the parameter space. Then, we solve the following nonlinear least squares optimization problem:

$$\min_{C_{ij}} \quad \sum_k (f(x_0^m, \cdots, x_{n-1}^m) - \hat{f}(x_0^m, \cdots, x_{n-1}^m))^2 \tag{2}$$

Where $f(x_0^m, \cdots, x_{n-1}^m)$ is output of the EMOD simulation and x^m refers to the m^{th} observation. We will now intentionally drop the m superscript in order to simplify the following explanation, but we will restore it at the end.

This problem is a nonlinear optimization problem, and we use an iterative algorithm to solve this optimization. As stated above, $\hat{f}(x_0, \cdots, x_{n-1}) = G_0^T(x_0)G_1(x_1)\cdots G_{n-1}(x_{n-1})$. Attempting to solve for all C_{ij} simultaneously results in a nonlinear least squares problem. Instead, we will iterate over i, and for each i, we will fix all unknowns before and after $G_i(x_i)$, and we will solve for $G_i(x_i)$ individually. Each $G_i(x_i)$ a linear combination of unknown coefficient matrices, which we can reformulate into a linear least squares solve:

$$G_0^T(x_0)\cdots G_i(x_i)\cdots G_{n-1}(x_{n-1}) = f(x_0, \cdots, x_{n-1})$$

We define $\mathbf{u}^T = G_0^T(x_0)\cdots G_{i-1}(x_{i-1})$ and $\mathbf{v} = G_{i+1}(x_{i+1})\cdots G_{n-1}(x_{n-1})$. We can also expand $G_i(x_i) = \sum_{j=0}^{k} b_j(x_i)C_{ij}$.

$$\mathbf{u}^T(b_0(x_i)C_{i,0} + \cdots + b_{k-1}(x_i)C_{i,k-1})\mathbf{v} = f(x_0, \cdots, x_{n-1})$$
$$(v^T \otimes u^T)\,\text{vec}(b_0(x_i)C_{i,0} + \cdots + b_{k-1}(x_i)C_{i,k-1}) = \text{vec}(f(x_0, \cdots, x_{n-1}))$$

We define $\mathbf{b}^T = [b_0(x_i), \cdots, b_{k-1}(x_i)]^T$.

$$(\mathbf{v}^T \otimes \mathbf{u}^T \otimes \mathbf{b}^T)[\text{vec}(C_{i,0}), \cdots, \text{vec}(C_{i,k-1})] = \text{vec}(f(x_0, \cdots, x_{n-1}))$$

We now have the form of a linear least squares solve. We can repeat the same analysis for ever one of the sample points. We then assemble those equations into a linear system of equations of the form:

$$A[\text{vec}(C_{i,0}), \cdots, \text{vec}(C_{i,k-1})] = \mathbf{y}$$

Where the m^{th} row of A is $\mathbf{v}^T \otimes \mathbf{u}^T \otimes \mathbf{b}^T$ and the m^{th} row of $\mathbf{y}$ is $f(x_0, \cdots, x_{n-1})$.

Solving this least squares problem will yield the desired coefficient matrices for only a single $G_i(x_i)$. We will then iterate over all $G_i(x_i)$ multiple times, fitting each $G_i(x_i)$ to the available data. We then terminate this iteration after reaching desired accuracy, where we define the accuracy based on two measures. The first measure is the L_2 error in predicting the data. The second measure is the change in coefficients from one iteration to the next. It should be mentioned that, before the least squares solve is performed, a preconditioner is applied to A. This preconditioner is a diagonal matrix which reduces the magnitude of the entries of A, mitigating the chance of numerical error when performing the least squares solve. This preconditioner, P, has diagonal entries $p_{ii} = \frac{1}{\max A}$, where $\max A$ denotes the largest entry of A.

Note that our $b_j(x_i)$'s did not necessarily need to be monomials. We could have replaced the monomials with orthogonal polynomials (e.g. Hermite polynomials, Legendre polynomials, etc.), and utilized the same process.

5.1.2 Application

We apply the above fitting procedure to fit two separate functions. The first function is for DALYs as a function of the parameters controlling the intervention policy. We will refer to that function with $\hat{f}_{DALY}(\mathbf{x})$. The second function is for Cost as a function of the parameters controlling the intervention policy. We will refer to that function as $\hat{f}_{Cost}(\mathbf{x})$. As will be seen in section 6, we managed to construct high-dimensional surrogate models, up to 42 dimensions. We believe that the proposed fitting algorithm can scale up to hundreds of parameters.

5.2 Optimization Algorithm

As part of our proposed algorithm, we require an optimization algorithm to find hypothesized optima. We will use Newton's algorithm (a gradient-based approach) for optimization.

First, we transform our constrained optimization problem using Lagrange multipliers.

$$\text{Lower Bound Penalty} = LBP(\mathbf{x}) = \sum_{i=0}^{n-1} \exp\left(\frac{1}{\lambda_0}(lb - x_i)\right)$$

$$\text{Upper Bound Penalty} = UBP(\mathbf{x}) = \sum_{i=0}^{n-1} \exp\left(\frac{1}{\lambda_1}(x_i - ub)\right)$$

$$\text{Budget Penalty} = BP(\mathbf{x}) \quad = \exp\left(\frac{1}{\lambda_2}(\hat{f}_{\text{Cost}}(\mathbf{x}) - B)\right)$$

Where λ_0, λ_1, and λ_2 are Lagrange multipliers, lb and ub are lower and upper bounds on the parameters, and B is the budget constaint. Our optimization algorithm will therefore aim to solve the following problem:

$$\min_{\mathbf{x}} \quad U(\mathbf{x}) = \hat{f}_{\text{DALY}}(\mathbf{x}) + LBP(\mathbf{x}) + UBP(\mathbf{x}) + BP(\mathbf{x}) \tag{3}$$

Recall that Newton's algorithm is a gradient-based algorithm where the solution at every step is incrementally updated using the following update function:

$$\mathbf{x_{\ell+1}} = \mathbf{x}_\ell - H^{-1}\mathbf{j}$$

Where ℓ is the index of the iteration, $\mathbf{x}_\ell$ is the current estimate of the solution, $\mathbf{x}_{\ell+1}$ is the updated guess to the solution, H^{-1} is the inverse of the Hessian matrix, and $\mathbf{j}$ is the Jacobian. The Jacobian is a vector of length n and the Hessian is a matrix of size $n \times n$. The elements of the Jacobian are:

$$\mathbf{j}_i = \frac{\partial U}{\partial \mathbf{x}_i}(\mathbf{x}_\ell)$$

The elements of the Hessian are:

$$H_{ij} = \frac{\partial^2 U}{\partial \mathbf{x}_i \partial \mathbf{x}_j}(\mathbf{x}_\ell)$$

We stop once our update of $\mathbf{x}$ becomes smaller than a certain threshold. Please note that one of the main advantages of using a tensor train-based surrogate model is that it is very easy to compute the derivatives.

6 Results

In this section, we will discuss the application of our algorithm to two separate problems. First, on identifying optimal intervention policies to for HIV in South Africa. Second, on identifying optimal intervention policies to for HIV in Kenya. The first problem will be referred to as South Africa, and the second problem will be referred to as Kenya. The South Africa problem is parameterized using 11 parameters, and the Kenya problem is parameterized using 42 parameters. For each of the two problems, we will explain the processes of data gathering, tensor train fitting, and the identified optimal allocation policies.

To run each simulation, we use a config file calibrated by Prof. Kim in the Bershteyn Lab group (Bershteyn, 2018). As mentioned above, the config files allow us to accurately simulate HIV spread within their respective countries. The parameters that we change only control the intervention policies, but do not affect population demographics or HIV behavior. Please note that some of the changes we do to the parameters may be hard to implement in real life. That does not undermine the validity of the final results, but slightly increases the difficulty of a real-world implementation. With that being said, our algorithm is flexible enough to allow tighter constrains on the parameters in order to better mimic real-life constraints. We leave this refinement up to the public health experts.

6.1 HIV in South Africa

6.1.1 Parameter Descriptions

We identified 11 parameters which controlled the coverage of intervention policies: 6 for ART, 1 for VMMC, 1 for PrEP, and 3 for HCT.

These parameters are listed in table 6.1.1:

Table 1: List of Parameters for South Africa Model Healthcare System

Parameter Name	Parameter Description	Default Value	Range
PresProb	Prob. a symptomatic agent gets HIV test	0.94	[0,1]
Child_6w	Prob. 6-week-old child gets HIV test	0.337	[0,1]
TestUptake	Increased rate of testing after 2009	0.9	[0,1]
Staging	Prob. agent begins ART Staging	0.837	[0,1]
PreART	Prob. agent continues PreART	0.75	[0,1]
FastART	Prob. agent begins ART after 2016	0.95	[0,1]
On_ART	Prob. agent who re-enrolls for ART	0.9	[0,1]
KeepART	Prob. agent keeps using ART after 2016	0.9	[0,1]
ARTInterrupted	Prob. ART is interrupted	0.2	[0,0.2]
PrEP	Prob. HIV negative agent begins to use PrEP	0.15	[0,1]
VMMC	VMMC coverage over the male population	0.8	[0,1]

6.1.2 Surrogate Model Fitting

We will start by fitting a surrogate model. To do that, we start our process by running simulations on 1000 randomly sampled points in the parameter space, and generating the corresponding Cost and DALY outputs. Then we fit two surrogate models; one for Cost and one for DALYs.

We choose the same specifications for both surrogate models, namely, rank $r_i = 5$ and order $k = 4$, yielding a third-order polynomial approximation.

Our choices of 1000 samples and the rank and order of approximation has been determined based on both available computational resources and sensitivity analyses to ensure robustness of our results. A rank of 5 and third-order polynomials produced the most accurate model while avoiding overfitting to the available data.

We defined the surrogate model error as the L_2 norm of the difference of the sampled points from the training data divided by the L_2 norm of the training data:

$$\text{Error} = \frac{||\mathbf{f}(\mathbf{x}) - \hat{\mathbf{f}}(\mathbf{x})||}{||\mathbf{f}(\mathbf{x})||}$$

The model achieved an error of under 1% for both Cost and DALY. The error was 0.29% for the Cost model, and 0.91% for the DALY model. The surrogate typically took 2 minutes to train on 50 iterations of 1000 training points, and evaluated a single data point in under 100 ms. This represents a run time close to 4 orders of magnitude faster than the base EMOD model.

6.1.3 Optimal Allocation

We implemented the proposed algorithm as described in the previous section, with a cost constraint of $1M. The Cost and DALY combinations generated, compared to those generated randomly, are found in 1.

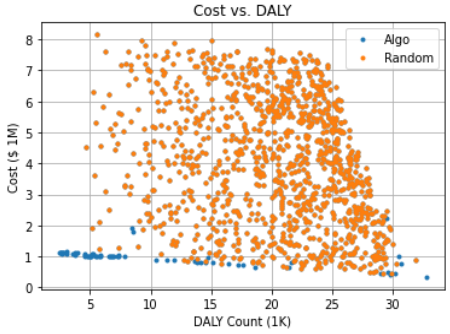

Figure 1: Graph of Cost vs DALYs for both Random and Chosen Points

The above graph shows 1000 randomized points in orange, comparing the cost for each simulation with the DALY count for that simulation. In blue are the points computed with our algorithm. The most notable feature is the 'tail' of blue points on the bottom left of the graph, which represents the convergent behavior of the algorithm: all these points carry a low DALY measure and are very close to the cost constraint of $1M. The scattered points across the bottom and right represent points where our surrogate models made certain inaccurate predictions. In those cases, our algorithm

was able to take advantage of these inaccuracies and use them to reinforce the accuracy of our surrogates.

We wanted to evaluate the effect of increased budget on DALY Count and optimal parameters of the intervention policies. To achieve that, we changed the allotted budget from $1M to $3M in steps of $500K. The results of these runs are shown in Figure 2.

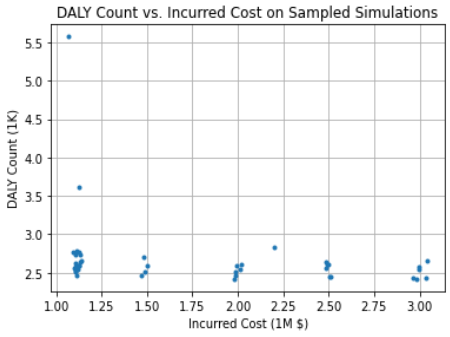

Figure 2: Graph of Cost vs DALYs for Varied Budgets

We have multiple observations. The DALY count was relatively stable, ranging from 2416 to 2470, a small variance relative to the possible range of DALYs. For small budgets, ART is by far the most cost-effective policy, while PrEP remains at zero. For larger budgets, PrEP becomes more cost-effective. One surprising observation is that our algorithm sometimes terminates at optimal points that do not use all the allowed budget. For example, even though we set a budget of $3M, the algorithm has terminated at both $1.1M and $2M. We studied this observation and found that there are 5 different local minima at which the algorithm terminates. In Table 6.1.3, we give the parameter set yielding the best point within each cluster and the corresponding DALY count and Cost.

Table 2: Parameter and Output Values for South Africa Model Clusters

Parameter Name	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
KeepART	1	1	1	1	1
TestUptake	1	1	1	1	1
FastART	0.423	0.567	0.623	0.003	0.304
PresProb	1	1	1	1	1
Child_6w	1	0.743	0.31	0.802	0.998
Staging	1	0.979	1	1	1
PreART	1	1	1	1	1
On_ART	1	1	1	1	1
VMMC	0.016	1	0.64	1	0.003
PrEP	0	0.05	0.12	0.186	0.247
ARTInterrupted	0.2	0.129	0.071	0.035	0
Cost	$1,113,055.47	$1,465,936.00	$1,978,418.95	$2,513,331.67	$2,985,608.23
DALY	2,470.75	2,461.45	2,417.12	2,442.63	2,416.71

It seems that ART is the most heavily emphasized intervention, with 3 different parameters set to 1 between all 5 clusters. HCT is a close second, with 2 parameters set to 1 between all 5 clusters. This may be due to the way they work within the simulation. ART is a medication taken by those with HIV which extends their lifespan while preventing them from passing HIV on to other people. This has a dual-effect in reducing DALYs, limiting the DALYs created from their own deaths while limiting the DALYs created from the deaths of others. Testing works by taking those who may not know they have HIV, and putting them in a position where they have the option to get treated. It is also relatively inexpensive compared to the other interventions, meaning that, even with a lower reduction in DALYs than other medications, it can still be more cost-effective as it is cheaper.

PrEP, on the other hand, begins with 0 investment, and slowly increases as the total amount of money spent increases. PrEP is a preventative, taken by those without HIV, which protects them from catching HIV. This has an impact compared to other medications only if the number of people *catching* HIV is creating more DALYs than those already with HIV. PrEP is also expensive, meaning that it needs to have a very large impact in order to be cost-effective. It seems like this is not the environment of the South Africa model.

6.2 HIV in Kenya

6.2.1 Parameter Descriptions

In the Kenya disease model, there exist 42 parameters which control the available interventions: 6 parameters for ART, 18 parameters for VMMC, 12 parameters for PrEP, 4 parameters for HCT, and 2 parameters for ANC.

6.2.2 Surrogate Model Fitting

As with the South Africa problen, we fit two surrogate models, one for Cost and one for DALYs. The surrogate models were fit with 1000 randomly generated points, which took EMOD 2 days to evaluate. Initially, we chose rank $r_i = 5$ and order $k = 4$, yielding a third-order polynomial approximation with matrix coefficients of shape 5×5. Upon beginning optimization using Newton's method, this tensor train structure had too many unknown coefficients to solve for, as compared to the surrogate models used in the South Africa example. This led to overfitting the surrogate models to the available data, meaning that the surrogate's projections became unreliable. Outputs ranged from -10^{12} to 10^{12} and were not accurate to EMOD. As a result, r_i was reduced to 3, and k was reduced to 2, yielding a linear function for each variable, with coefficients matrices of shape 2×2. This greatly reduced the number of free variables, mitigating the overfitting.

6.2.3 Optimal Allocation

We implemented the proposed algorithm as described in the previous section, with a cost constraint of $3M. The Cost and DALY combinations computed by the framework, compared to those generated randomly, are found in Figure 3.

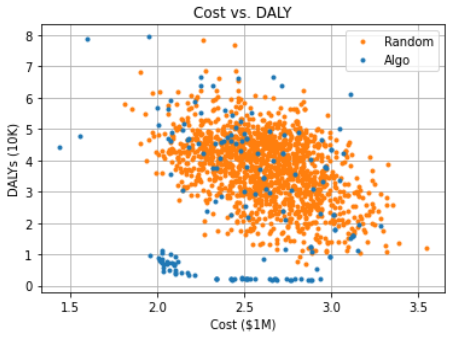

Figure 3: Graph of Cost vs DALYs for Varied Budgets

As before, there are some points dispersed in suboptimal areas with high DALY counts and varied Costs. These are points where the framework has predicted inaccurately, and is able to update its predictions based on the real EMOD value. Notably, there is a thick cluster of points ranging between $2M budget and $3M budget with low DALY counts, ranging from the low thousands up to $10k$ DALYs. This region is highlighted in Figure 4.

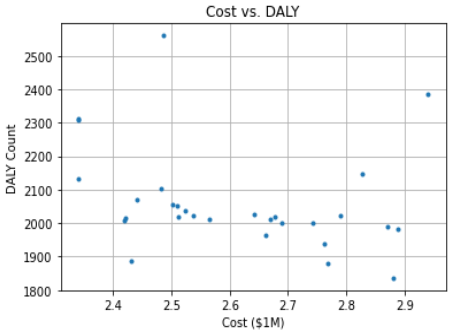

Figure 4: Graph of Cost vs DALYs for Optimal Simulations in Kenya

The most optimal points range in DALY counts from 1800 to 2500, a value significantly lower than the randomly generated simulations. Optimal Cost also ranges between $2.4M to $3M, well within our allowed budget of $3M. Notably, there are no clusters within the located points. There is only a scattering of lower values spread out fairly evenly within the area. Importantly, while all of these points have low DALY counts, the predictions that the surrogate models made were not accurate within this region, as shown in Figure 5.

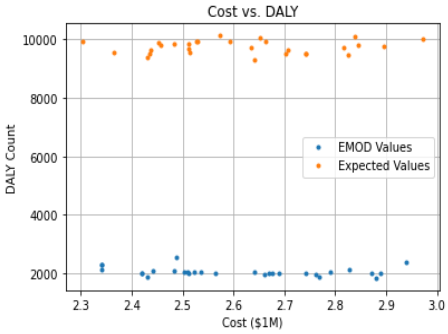

Figure 5: Graph of Cost vs DALYs for Surrogate Projections Compared to EMOD Values in Kenya

The surrogate models predicted DALY values in the 10k range, but EMOD actually returned DALY values near 2k DALYs. This is most likely because we used a linear model for each term in the tensor train. Compared to cubic polynomials, which we were using for the South Africa problem, linear polynomials have less ability to model complex behavior of a single parameter. They would tend to model the bulk of the data while ignoring the fringe behavior. In our case, the DALY values of the points used for fitting range from 10k DALYs to 70K DALYs. It could make sense for a model with relatively few parameters to stick to the bulk of the data, precisely this range. Points with low DALY counts are abnormal, which could explain why the surrogate models drastically overestimate the DALY count of low-DALY points.

It should be remembered that these points, while having high predicted DALYs (as predicted by the surrogate), still had low DALYs when evaluated by EMOD. This is probably due to the fact that a gradient-based optimization algorithm was used. The algorithm was able to accurately traverse the surrogate models to a place of relative minimum, and although the actual value of this relative minimum was overestimated, it was a relative minimum nonetheless, in the surrogate model's estimation and within EMOD as well.

The Cost within this range is predicted accurately, which would align with our explanation. The Cost of the randomly generated points ranges from \$2M to \$3.25M, and our predicted Costs landed within the middle of this range. A model with few degrees of freedom may be able to predict Costs accurately if they fall within the range of the Costs used to fit the model.

7 Conclusions & Future Work

In this project, we were able to construct a framework for HIV intervention optimization which greatly reduced the number of DALYs computed within the EMOD simulation tool. Our framework is able to adapt to various price points and recommend optimal investment strategies for each one. Our framework is also self-improving, in that it was able to either locate an optima with respect to the main EMOD model, and if not, it improved upon its accuracy and predictive ability. This framework is highly customizable in terms of metaparameters and since it is not specific to HIV models, it can be applied to a variety of optimization problems. It is powerful enough to solve high-dimensional problems ($d \geq 40$), and reports a variety of results to the end user. It uses a gradient-based optimization algorithm which takes advantage of the surrogate model's differentiability and low evaluation time. It can be fit to already-existent data.

In terms of optimal investment, it seems that ART and HCT are the most cost effective interventions which should be prioritized on a limited budget. Preventative medications like PrEP, however, should be skipped out when strapped for cash, but as available funding increases, investment should increase.

This work can be extended in a number of ways. First and foremost, it should be taken by a public health researcher and applied to a realistic example of HIV. In this project, the parameters used for each application represented the maximum possible reach over each country's healthcare system. In reality, South Africa may be less able to give ART medication and more able to give HCT kits. This would be represented by bounds on the values of the corresponding parameters easily inputted into the existing software. Additionally, this framework should be tested on other models and optimization problems. It is not specific to HIV models, or disease models as a whole, and is instead general. It can be applied to any form of optimization problem, and should be tested as such.

8 Acknowledgements

I would like to thank my mentor, my parents, and my science research teacher.

References

[1] South Africa. UNAIDS. https://www.unaids.org/en/regionscountries/countries/southafrica. Published April 30, 2022. Accessed November 1, 2022.

[2] Kharsany AB, Karim QA. HIV Infection and AIDS in Sub-Saharan Africa: Current Status, Challenges and Opportunities. Open AIDS J. 2016;10:34-48. Published 2016 Apr 8. doi:10.2174/1874613601610010034

[3] Meyer-Rath, Gesine, et al. "The per-Patient Costs of HIV Services in South Africa: Systematic Review and Application in the South African HIV Investment Case." PLOS ONE, vol. 14, no. 2, 2019, https://doi.org/10.1371/journal.pone.0210497.

[4] Kerr CC, Stuart RM, Gray RT, et al. Optima: A Model for HIV Epidemic Analysis, Program Prioritization, and Resource Optimization. J Acquir Immune Defic Syndr. 2015;69(3):365-376. doi:10.1097/QAI.0000000000000605

[5] Lukoba, Benard; Simiyu, Joseph; Chege, Wendy; Kelly, Sherrie; Minnery, Mark; Sithole, Lonjezo; Shubber, Zara. 2020. Improving Allocative Efficiency of the HIV Response in Kenya : A Country-Level Analysis Using the Optima HIV Model. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/35196 License: CC BY 3.0 IGO.

[6] Sharma M, Mudimu E, Simeon K, et al. Cost-effectiveness of point-of-care testing with Task-shifting for HIV care in South Africa: A modelling study. The Lancet HIV. 2021;8(4):216-224. doi:10.1016/s2352-3018(20)30279-4

[7] Bershteyn, A., Sharma, M., Akullian, A. N., Peebles, K., Sarkar, S., Braithwaite, R. S. and Mudimu, E. Impact along the HIV pre-exposure prophylaxis "cascade of prevention" in western Kenya: a mathematical modelling study. J Int AIDS Soc. 2020; 23(S3):e25527

[8] Korenromp EL, Bershteyn A, Mudimu E, et al. The impact of the program for medical male circumcision on HIV in South Africa: analysis using three epidemiological models. Gates Open Res. 2021;5:15. Published 2021 Jan 25. doi:10.12688/gatesopenres.13220.1

[9] Bershteyn A, Gerardin J, Bridenbecker D, et al. Implementation and applications of EMOD, an individual-based multi-disease modeling platform. Pathog Dis. 2018;76(5):fty059. doi:10.1093/femspd/fty059

[10] Gorodetsky A, Karaman S, Marzouk Y. A continuous analogue of the tensor-train decomposition. Computer Methods in Applied Mechanics and Engineering. 2019;347:59-84. doi:10.1016/j.cma.2018.12.015

[11] Bigoni D, Engsig-Karup AP, Marzouk YM. Spectral tensor-train decomposition. SIAM Journal on Scientific Computing. 2016;38(4). doi:10.1137/15m1036919

[12] Z. Zhang, X. Yang, I. V. Oseledets, G. E. Karniadakis and L. Daniel, "Enabling High-Dimensional Hierarchical Uncertainty Quantification by ANOVA and Tensor-Train Decomposition," in IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 34, no. 1, pp. 63-76, Jan. 2015, doi: 10.1109/TCAD.2014.2369505.

[13] Oseledets, I. V. "Tensor-Train Decomposition." SIAM Journal on Scientific Computing, vol. 33, no. 5, 2011, pp. 2295–2317., https://doi.org/10.1137/090752286.

[14] United Nations, Department of Economic and Social Affairs, Population Division (2019). World Mortality 2019: Data Booklet (ST/ ESA/SER.A/436).