A Tale of Two Problems:

Real-World Experimental Data and Benchmarking Synthetic Data

Group 3

*ChBE 6746/4746 Spring 2023*

*Jessica Bonsu, Lisa Brown, Omar Jimenez, Jennifer Najm, Lucas Timmerman*

*March 31, 2023*

# Introduction

Here, we present two investigations which we will call Problem 1 and Problem 2. Problem 1 involves real-world experimental data for chitin-cellulose films, and Problem 2 is a benchmark optimization problem comparing different gradient-free solver algorithms.

For Problem 1, our work began with an experimental dataset collected in Dr. Carson Meredith bio-renewable materials lab was used. After some difficulties with data management of the dataset for an optimization approach, we concluded that using the films dataset to solve an optimization problem was not the optimal approach for our group to take and sought an alternative optimization problem.

For Problem 2, we approached a gradient-free benchmark optimization problem. In gradient-free optimization (also known as derivative-free or blackbox optimization), the derivative information of a deterministic function is not used to find the global optimum. A Python package called Gradient-Free-Optimizers by SimonBlanke9 provides a series of gradient-free solver algorithms. Our goal is to assess the performance of these solvers by applying them to a series of test problems. We will model our benchmarking approach after a review paper by Rios and Sahinidis8.

# Problem 1: Real-World Experimental Data

In the food packaging industry, the utilization of chitin and cellulose materials for packaging films poses as a sustainable method due to the biodegradability and functionality of these materials. Two important properties of such films are Oxygen Permeability (OP) and Water Vapor Transmission Rate (WVTR). To effectively extend food shelf life and mimic the effectiveness of current environmentally harmful packing materials, these properties need to be optimized through minimization.

# Optimization Problem

The objective of Problem 1 was to determine the optimum film composition and preparation parameters to minimize the transmission of oxygen and water vapor through cellulose-chitin films. This first required determining the relationships between the controlled variables and the target outputs. Therefore, we planned to approach this Problem in two parts: (1) use data analytics methods to identify relationships between input and output variables, and (2) use optimization methods to determine the ideal parameters based on the variable relationships we determined in part 1. Unfortunately, our results from part 1 indicated that the dataset is not amenable to implementing a regression model.

# Dataset Description

The dataset we used for Problem 1 was experimentally collected by a postdoctoral researcher of Dr. Carson Meredith bio-renewable materials lab. Both the postdoctoral researcher and Dr. Meredith gave permission to use and analyze this dataset. This is a static and complete dataset consisting of 104 rows representing 78 unique film compositions, 3 feature columns representing preparation parameters, and 3 output columns. Each film composition includes the concentration of primary component(s) and any additional chemical treatments. Preparation parameters include the mold substrate from which the film was prepared, the drying temperature, and whether the film was covered during drying. Once the film material is prepared, characterization is done which yields the following outputs: film thickness, oxygen transmission rate, and water vapor transmission rate. For our optimization problem, we planned to use the derived outputs Oxygen Permeability (OP) and Normalized Water Vapor Transmission Rate (N-WVTR) by normalizing the oxygen transmission rate and water vapor transmission rate, respectively, to the film thickness.

# Preliminary ideas on data analysis, optimization method(s), and solver(s)

For the first part of Problem 1, we planned to use regression techniques to explore the relationships between inputs and output variables. We probed the feature and target space for correlations and attempted to identify especially important features and relationships to target values. We attempted to expand the feature space by using one hot encoding to quantify various compositional formulations and also extracted solution composition from the “Name” column of the dataset. This resulted in a large and sparse feature space.

In the second part of Problem 1, we planned to derive the mathematical formulation for the optimization problem based on minimizing transport through the films. Due to the nonlinear nature of equations in transport phenomena, we realized this may result in a Non-Linear Programming (NLP) or potentially a Mixed Integer Non-Linear Programming (MINLP) problem. To this end, we planned to use Pyomo and an appropriate method(s) (e.g., IPOPT if NLP, Baron if MINLP) to solve the formulated optimization problem and determine the optimum film composition and preparation parameters.

# Contingency Plan

Our group acknowledged that the size of the dataset may hinder us from receiving feasible and sufficient optimization results. In this instance, we planned to explore established mechanistic models1–5 and realized we would likely need to simplify these models as our dataset was not constructed with the goal of modeling it.

We considered supplementing our project with machine-learning image analysis of film material images, also collected from Dr. Meredith’s lab. We also considered generating data using molecular dynamics simulations should our plan completely fall apart. We could extract important chemical parameters from such simulations (Diffusion coefficients, gas permeability) to develop a model of Oxygen Permeability/Water transfer rate based on a mechanistic explanation of the phenomena, such as Fickian Diffusion or directly via simulation6.

# Data Analytics & Feasibility

Data analytics calculations for Problem 1 are available in Preprocessing-LR.ipynb. We utilized the built- in pandas correlation function to visualize the correlation matrix for the feature and target space. We also utilized principal component analysis to evaluate the separability of our data. Histogram features were also constructed for each of the features. Lastly, we used sklearn’s built in linear regression modules to probe the feasibility of developing regression models to relate the features to our targets.

As can be seen from Figure 1, our results strongly suggest that data-driven modeling techniques will not be useful for this dataset. The correlation matrix shows that the strongest correlations in the dataset are between the target properties themselves. From the feature histograms, it is obvious that our dataset is incredibly sparsely sampled. Indeed, the one-hot encoding techniques that were used to expand the feature space will by definition yield a discontinuous feature space. However, the fact that almost all of these features are strongly skewed to 0 as opposed to a more reasonable split between 0 and 1 shows that the dataset is sparsely sampled. Figure 1 c.) and d.) show the results of PCA, with the scree plot in c.) demonstrating that nearly all of the variance in the feature set can be captured in a single principal component. This again speaks to the sparsity of our dataset and the lack of variation in input variables. Finally, d.) shows the results of simple linear regression for our dataset. We experimented with several approaches to partition our dataset (e.g., full dataset, pairs of features). We note that in a previous deliverable report we slightly misinterpreted our results since some of the lower error values were binned into larger value groups giving the false impression that *all* errors were catastrophic (see Fig. 1.d below). Upon a more careful analysis we found that many errors were reasonable, but a few of the relevant ones to our project’s predictive tasks were still fairly large – some with errors of *O*(106). This analysis has led our group to conclude that additional attempts would likely be fruitless as this dataset was not generated with the intent of informing data driven modeling.

|  |  |
| --- | --- |
| a.) | b.) |

|  |  |
| --- | --- |
| c.) | d.) |

*Figure 1: a.) Correlation matrix for extended features b.) Histogram of all features c.) Scree plot for PCA analysis d.) Target properties plotted vs the first two principal components. Color corresponds to numerical value of target property*

# Alternative Approaches

In an effort to construct a more feasible optimization problem, we explored using a different dataset from the literature on a similar topic, and we found one generated from molecular dynamics (MD) simulations7. However, we concluded that the applicability of this dataset would be limited since it is unclear how MD trajectories could be useful within the context of our project – our problem formulation is based on transport phenomena. We note that it might be possible to extract data from simulations that could prove useful for our project, but this slightly falls outside the scope of this class. In addition, the downloadable dataset included all data in a single 8 GB file, which we were not able to download due to limitations with the network, server, and/or our PCs. The raw dataset also clearly exceeds the 10 MB requirement by a significant amount.

We also explored calculating new metrics for a more universal comparison of film types: oxygen & water vapor transport coefficients and a score. To prepare for these calculations, we obtained more information from Dr. Meredith’s lab about the experimental conditions for the existing dataset, and we considered finding similar metrics in the literature for films currently used in the food industry (e.g., PET). As we continued thinking about this approach, we realized the resulting optimization problem would primarily be binary and therefore simpler than the desired type of optimization for this project. Also, most of the necessary work would likely be outside the scope of the course material, i.e., more data mining and less optimization.

# Conclusions for Problem 1: Real-World Experimental Data

While our efforts for Problem 1 were a great learning experience, our group decided to pivot to a different project topic that was more conducive for utilizing more of the concepts covered in this class. We discussed this plan with Dr. Boukouvala and obtained the new topic information from her group.

# Problem 2: Benchmarking Synthetic Data

* 1. **Optimization Problem**

The objective of Problem 2 is to compare several gradient-free optimization algorithms for global optimization of a set of benchmark optimization problems. We will utilize the implementations developed by Simon Blanke in his Gradient-Free-Optimizers software package9. This is an open source package which can be installed via pip on any machine. We will evaluate the performance of the different optimizers in accordance with procedures laid out by Rios and Sahinidis8. Namely, we will evaluate the different algorithms’ ability to

* + - Find the global optimum with no starting point
    - Improve a starting point far from the global optimum
    - Refine a starting point close to the global optimum

# Dataset Description

The test problems consist of convex and non-convex continuous test functions with varying dimensionality and pre-defined upper and lower bounds. They originated from the bound-constrained problem set (bcp) publicly available from The Optimization Firm LLC [cite URL]. Teaching assistant Suryateja Ravutla provided these problems to our group in numpy file format (.npy) along with a Jupyter notebook with code to load the problems and visualize two-dimensional problems. The provided data includes the name, number of dimensions, and function form for each problem.

Note that since we have functional forms, we don’t actually have an a-priori dataset. Rather, we have the ability to generate data on command. The functions themselves are continuous on the defined domains and could in theory be analyzed with derivative-based optimizers. For this project, we are treating each of the functions as black-box models.

The dimensionality of the test problems is summarized in Figure 2. We analyzed a total of 263 problems, ranging between 1 and 1000 dimensions. Most problems (89) were two-dimensional.

We also attempted to calculate the convexity of each problem by finding the eigenvalues of the Hessian matrix. However, this approach was not successful for problems in which no solutions could be found for one or more partial derivatives. Given the limited time constraints for the project, we decided to focus our benchmark comparisons on dimensionality only.

Chart, bar chart

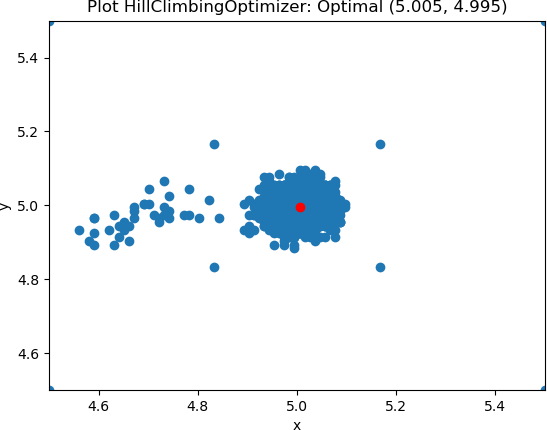
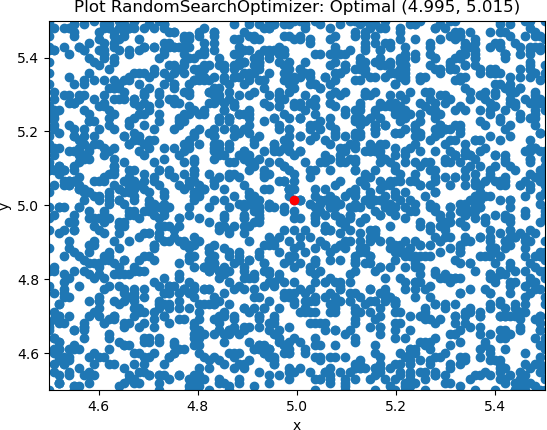
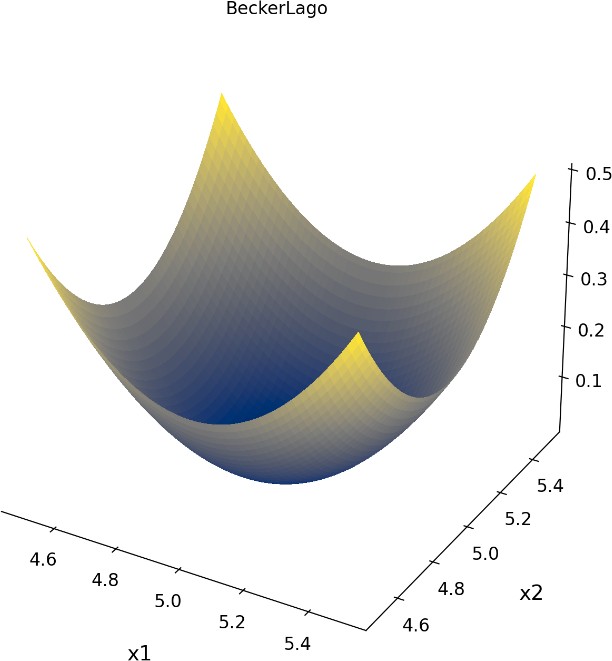
Description automatically generated

*Figure 2: Summary of dimensionality for the 263 test problems analyzed in this project.*

# Preliminary Implementation

Our implementation is available in sample.ipynb. This notebook and associated files are based on the original repository provided by Dr. Boukouvala, which contains a tutorial on how to run the package using a few helper functions. So far, we have experimented with the 2-dimensional BeckerLago problem (Figure 2) using two solvers as a proof-of-concept: Random Search and Hill Climbing. One thing we were initially unsure of was whether changes to the source code would be necessary to achieve the project goals. We found that our analysis could be achieved directly with the package API, such that all solvers can be implemented with the same code structure, and backend changes would not be necessary. Therefore, experimentation with additional solvers and test problems should be trivial.

Some exploratory analysis is presented in Figure 2. Figure 2a.) is a visualization of the 2 dimensional Becker Lago problem. It should be apparent from the figure that this is a convex problem. This contributed in part to our decision to utilize the Becker Lago problem for testing our implementation as we could easily visualize the results of our trials. Figures 2b.) and 2c.) show the final state of the Random Search and Hill Climbing algorithms respectively. Each algorithm was run for 2500 iterations in accordance with the work of Rios and Sahinidis. This visualization, along with the supporting gifs submitted, provide insight into how the solvers conduct their search. The random search algorithm is clearly a global search as the entire space is sampled. Figure 2c.) indicates that the hill climbing algorithm is a local search algorithm. It is also apparent that the geometry of the search space corresponds to a multivariate gaussian. This is easier to see from the gif. A peak inside the source code for this optimizer confirms this.



a.)

b.)

c.)

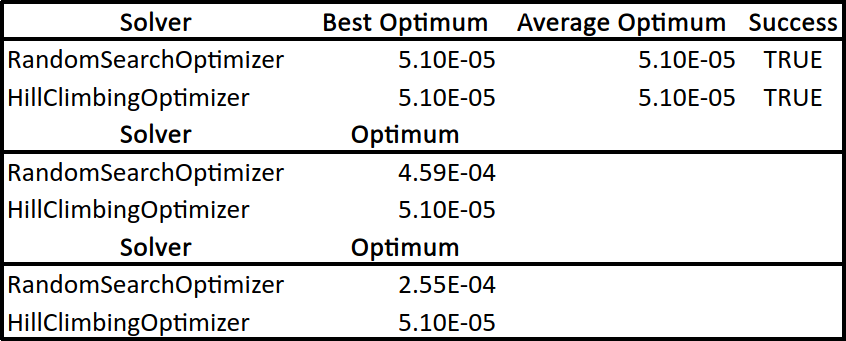
*Figure 2: a.) Visualization of the 2 dimensional Becker Lago problem plotted from (-4.5,5.5). b.) Visualization of the search space for the*

*Random Search Optimizer and c.) the Hill Climbing Optimizer.*

The optimization results are given in Table 1. Both the Random Search and Hill Climbing algorithm are able to find the global optimum within the tolerance set forth by Rios and Sahinidis. The global optimum for this problem is exactly 0. We start to see discrepancies between the optimizers for the other two tasks. The climbing hill optimizer finds the same optimum for all three tasks. The random search optimizer returns different results when refining specified initial guesses and near optimal solutions. Both of the solutions returned would still qualify as near optimal solutions; however, they are clearly not quite as good as the global optimum that was returned for the initial search.

These results are indicative of the kind of analysis that we will run on a larger scale for the rest of the problems and optimizers in the set we were given.

*Table 1: Summary of results for 2 optimizers on the Becker Lago problem. The three metrics evaluated are ability to find an optimal solution, ability to improve an arbitrary guess, and ability to refine near optimal solutions.*



# Evaluation of Performance Criteria

[overview of how and what we evaluated]

[we decided not to evaluate sequential solvers because they timed out/had memory errors]

# 2.4.1. Find Global Optimum with No Initial Guess

[results and discussion]

# 2.4.2. Improve Initial Guess Far From Global Optimum

[results and discussion]

# 2.4.3. Refine Initial Guess Close to Global Optimum

[results and discussion]

# 2.5. Conclusions for Problem 2: Benchmarking Synthetic Data

[results and discussion]

1. **Contributions**

**Deliverable 1: Proposal**

|  |  |
| --- | --- |
| Jessica | Proposed topic, provided data, and wrote parts of *Dataset Description, Data Analytic & Feasibility,* and *Contingency Plan* sections of report. |
| Lisa | Wrote report title, *Optimization Problem* section, and overall edits. |
| Omar | Wrote *Optimization Methods/Solvers* section of report. |
| Jennifer | Wrote *Introduction* section of report. |
| Lucas | Proposed ideas for data analysis, wrote parts of *Dataset Description* and *Contingency Plan*  sections of report, and conducted literature review via ChatGPT6. |

**Deliverable 2: Data Update**

|  |  |
| --- | --- |
| Jessica | Conducted literature search7, provided molecular dynamics dataset, and obtained additional information about existing dataset. |
| Lisa | Proposed ideas for alternative datasets and prepared updated report draft. |
| Omar | Wrote Linear Regression (LR) code in Jupyter notebook and participated in discussions. |
| Jennifer | Participated in discussions and conducted some research. |
| Lucas | Performed data analysis of existed dataset, created GitHub repository and Jupyter notebook, and proposed ideas for universal metric calculation. |

**Deliverable 3: Project Pivot**

|  |  |
| --- | --- |
| Jessica | Contributed with introduction and working with one of the derivative free solvers. |
| Lisa | Prepared report draft, proposed ideas for next steps, and edited several sections of the report. |
| Omar | Edited Section 1.5 of the report. Wrote drafts for Sections 2.3-2.4. |
| Jennifer | Wrote Sections 2.1 and 2.2 of the report. |
| Lucas | Contributed to Section 2. Wrote the optimization code. Generated preliminary results |

**Deliverable 4: Final Report & Presentation**

|  |  |
| --- | --- |
| Jessica |  |
| Lisa |  |
| Omar |  |
| Jennifer |  |
| Lucas |  |

**Literature**

1. Viana Batista, R. *et al.* Food packaging and its oxygen transfer models in active multilayer structures: A theoretical review. *Journal of Plastic Film and Sheeting* **38**, 458–488 (2022).
2. Seong, D. M., Lee, H., Kim, J. & Chang, J. H. High oxygen and water-vapor transmission rate and in vitro cytotoxicity assessment with illite-polyethylene packaging films. *Materials* **13**, (2020).
3. Zhang, K. *et al.* The preparations and water vapor barrier properties of polyimide films containing amide moieties. *Polymers (Basel)* **9**, (2017).
4. Mastromatteo, M. & del Nobile, M. A. A simple model to predict the oxygen transport properties of multilayer films. *J Food Eng* **102**, 170–176 (2011).
5. Bhunia, K., Dhawan, S. & Sablani, S. S. Modeling the Oxygen Diffusion of Nanocomposite-based Food Packaging Films. *J Food Sci* **77**, (2012).
6. Open AI. “ChatGPT: Optimizing Language Models for Dialogue.” November 30, 2022. <https://openai.com/blog/chatgpt/>
7. Hudek, M., Kubiak-Ossowska, K., Johnston, K., Ferro, V. A., & Mulheran, P. A. (2023). Chitin and Chitosan Binding to the α-Chitin Crystal: A Molecular Dynamics Study. *ACS Omega.*, *8*(3), 3470–3477. https://doi.org/10.1021/acsomega.2c07495
8. Rios, Luis Miguel, and Nikolaos V. Sahinidis. “Derivative-Free Optimization: A Review of Algorithms and Comparison of Software Implementations.” Journal of Global Optimization, vol. 56, no. 3, 2012, pp. 1247–1293., https://doi.org/10.1007/s10898-012-9951-y.
9. SimonBlanke. “Gradient-Free-Optimizers”. March 24, 2023. https://github.com/SimonBlanke/Gradient- Free-Optimizers