INDICIUM Tech & Data Consulting

Technical Portfolio

Expert Consultancy in Machine Learning and Data Science

Ahmed Kriouile & Ilyas Malik
incidiumtechconsulting@gmail.com

Introduction

0.1 Background

We are a team of 2 highly skilled professionals:

Ahmed Kriouile PhD student at **Columbia University** (Operations Research, Engineer graduated from the **École Polytechnique (X)**, MSc in Machine Learning and Computer Vision (MVA[1]) .

Ilyas Malik Engineer graduated from the **École Polytechnique (X)** and MSc in Statistical Science at the **University of Oxford**.

We are childhood friends and colleagues of the École Polytechnique that both specialized in Applied Maths / Computer Science. As demonstrated in that document, we have extensive experience in many tech fields, we are problem solvers and we love challenging topics that require innovation and cutting edge technologies.
During our projects we had the privilege to work with international renown organizations, we are proud to have shined every single time within those organizations, living up to the high standards and committing to the best practices of those organizations. We have worked on numerous highly challenging tech projects and have built various state of the art technologies and algorithms in the domain.
We have considerable contributions in both research and industry, with papers published with the University of Oxford (Oxford) and others forthcoming with Columbia University (New York City), IBM Research Labs (Singapore), and Data61 (Australia). As well as strong contributions exceeding expectations in industry with giant tech leaders like Amazon and IBM.

0.2 Services

We have collaborated with different companies from huge structures (Amazon, IBM, BNP Paribas, EDF, NOMAC) to startups (Farasha Systems) as well as research centers on cutting-edge technologies (Oxford, IBM Research Labs, CSIRO Data61, ENSAE CREST, Inria, CMAP, LIX). See Appendix A.
We are versatile and confident we can approach technical problems in optimal ways and provide state of the art solutions.
We are leveraging our **unique expertise** to provide services in technical solutions especially **Data**

[1] Master of Science: Mathematics, Vision, Learning in the **École Normale Supérieure (ENS)** Paris-Saclay

Science/Data Visualization/Machine Learning/Artificial Intelligence related problems. The world is changing with those new technologies and we are confident we are going to be a part of that change by greatly contributing to developing those technologies and helping companies incorporate them to excel in the competition.
We can study a problem, efficiently approach it, formalize it and convert it to specific technical guidelines for an optimal result, we then execute the guidelines using state of the art technologies to deliver **high quality results**.
Our deep belief is that we are able to understand your technical needs and leverage our unique expertise to efficiently solve your problem in a **reasonable time frame**.

Contents

List of Figures

Chapter 1

Advanced Data Science, Machine Learning, AI

1.1 Oxford Research – Deep Adaptive Design: Amortizing Sequential Bayesian Experimental Design

1.1.1 Context

During a Dissertation Thesis at the university of Oxford, we worked on Bayesian Optimal Experimental Designs (BOED) and were able to contribute to a key and innovative development of that technology, which has tremendous consequences in real world applications in numerous fields: Psychological trials, pharmaceutical trials, any kind of survey, physics experiments . . .

1.1.2 Challenges and contribution

We were able to publish a paper in the International Conference on Machine Learning (ICML[1]). In this work, we introduce Deep Adaptive Design (DAD), a general method for amortizing the cost of performing sequential adaptive experiments using the framework of Bayesian optimal experimental design (BOED). Traditional sequential BOED approaches require substantial computational time at each stage of the experiment. This makes them unsuitable for most real-world applications, where decisions must typically be made quickly. DAD addresses this restriction by learning an amortized design network upfront and then using this to rapidly run (multiple) adaptive experiments at deployment time. This network takes as input the data from previous steps, and outputs the next design using a single forward pass; these design decisions can be made in milliseconds during the live experiment. To train the network, we introduce contrastive information bounds that are suitable objectives for the sequential setting, and propose a customized network architecture that exploits key symmetries. We demonstrate that DAD successfully amortizes the process of experimental design, outperforming alternative strategies on a number of problems.

[1]ICML is the leading international academic conference in machine learning. Along with NeurIPS and ICLR, it is one of the three primary conferences of high impact in machine learning and artificial intelligence research.

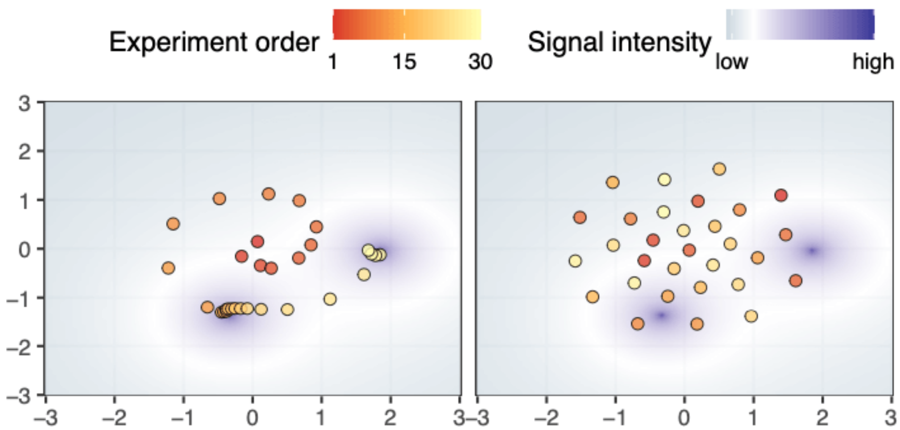


Figure 1.1: An example of the designs learnt by DAD [Left] and the fixed baseline [Right] for a given θ sampled from the prior.

1.1.3 Results

The algorithm resulting from that paper is given in Figure 1.2.
In all experiments DAD performed significantly better than baselines with a comparable deployment time. Further, DAD showed competitive performance against conventional BOED approaches that do not use amortization, but make costly computations at each stage of the experiment. An example of such performance is given in a physics experiment in inspired by the acoustic energy attenuation model, we consider the problem of finding the locations of multiple hidden sources which each emits a signal whose intensity attenuates according to the inverse-square law. The total intensity is a superposition of these signals. The design problem is to choose where to make observations of the total signal to learn the locations of the sources. Results are displayed in Figure 1.1.

1.1.4 Technical details

We implemented DAD by extending PyTorch and Pyro[2] to provide an implementation that is abstracted from the specific problem.

[2]Pyro is a universal probabilistic programming language (PPL) written in Python and supported by PyTorch on the backend. Pyro enables flexible and expressive deep probabilistic modeling, unifying the best of modern deep learning and Bayesian modeling.

Algorithm 1 Deep Adaptive Design (DAD)

Input: Prior $p(\theta)$, likelihood $p(y|\theta, \xi)$, number of steps T
Output: Design network π_ϕ
while Training compute budget not exceeded **do**
  Sample $\theta_0 \sim p(\theta)$ and set $h_0 = \varnothing$
  for $t = 1, ..., T$ **do**
    Compute $\xi_t = \pi_\phi(h_{t-1})$
    Sample $y_t \sim p(y|\theta_0, \xi_t)$
    Set $h_t = \{(\xi_1, y_1), ..., (\xi_t, y_t)\}$
  end
  Compute estimate for $\partial \mathcal{L}_T / \partial \phi$ as per § 4.2
  Update ϕ using stochastic gradient ascent scheme
end
At deployment time, π_ϕ is fixed and each y_t is obtained in turn by running experiment with design ξ_t.

Figure 1.2: Amortization technology

1.2 IBM Research – Safe Deep Q-Learning: Designing Risk-sensitive Reinforcement Learning Agents

1.2.1 Context

We have worked in IBM Research Labs (Department of AI) where we designed AI agents that are risk averse on demand. This technology has numerous applications in sensitive real life scenarios, for instance in security or medical applications where the goal is to accomplish the objective but it is also necessary to avoid worst-case scenarios. Forthcoming paper.

1.2.2 Contribution and results

We have used 3 novel approaches for distributional Reinforcement Learning (See RL representation in Figure 1.3, Deep Q-learning representation in Figure 1.4) to be able to recover (at least implicitly) the distribution of the return $Z(s, a)$ which is the reward of the agent where it performs an action a in a state s. The approaches we have used are:

1. GAN Q-learning
2. A Distributional Perspective on Reinforcement Learning: C51 Algorithm
3. Quantile Regression Deep Q-Learning: This is the most successful approach for which we have kept the experiments, see Algorithm 1.5

We have leveraged the above approaches to develop custom policies AI agents: instead of optimizing the averaged return (a.k.a Q function), the AI agent now optimizes a modified quantity, the Conditional Value at Risk which is basically a weighted average of the "extreme" cases in the tail of the distribution of possible returns.
One of the most visual experiments we have tested our code on are available to watch on **YouTube**:

- Extreme risk averse agent (Only optimizing worst-case scenarios): The lander approaches the flags but hangs in the air for the longest time to avoid crashing on an irregular ground, sometimes it just hangs in the air next to flags until the episode ends.
- Moderately risk averse agent (Optimizing worst-case scenarios and moderate scenarios): The lander has similar behaviour as the extreme risk averse case but it hangs less in the air and always lands between the flags.
- Standard agent (Optimizing over all scenarios): The Vanilla DQN version is the quickest of all but has more worst-case scenarios of crashing.

We conducted Monte Carlo experiments to compare the overall rewards of the risk averse and the standard agent, Figure 1.6 clearly shows that the vanilla DQN has a higher average but the safe version has significantly less worst-case or crashing scenarios.
We have written the code on Python using an object oriented logic and Tensorflow library.

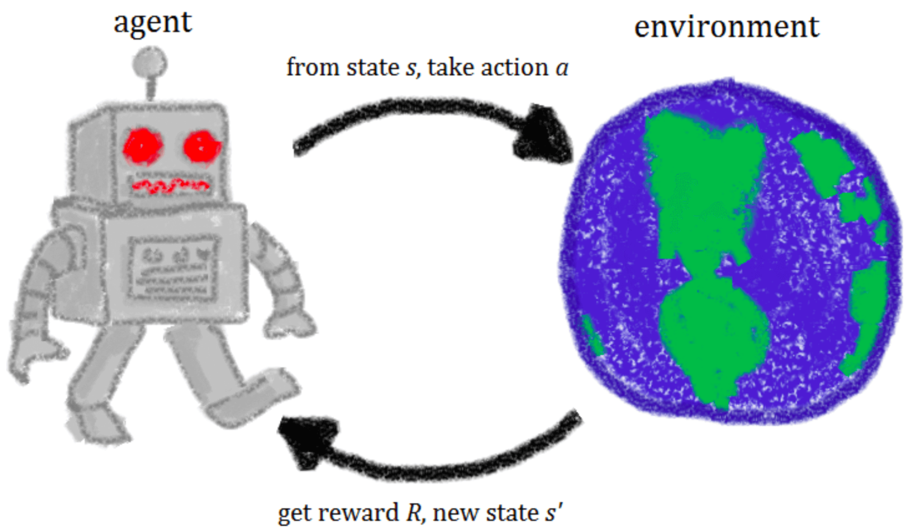


Figure 1.3: Reinforcement Learning representation

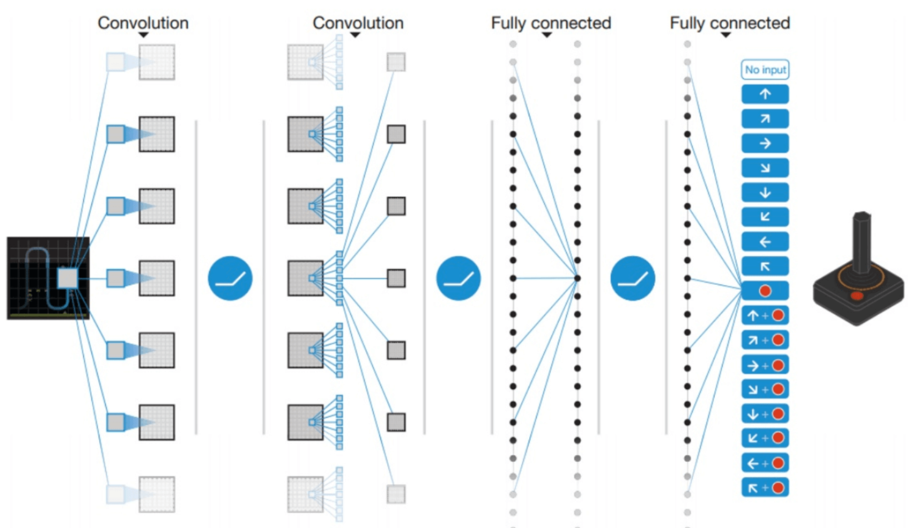


Figure 1.4: Deep Q-learning representation

Require: $N, \kappa\ x, a, r, x', \gamma \in [0,1)$
  # Compute distributional Bellman target
  $Q(x', a') := \sum_j q_j \theta_j(x', a')$
  $a^* \leftarrow_{a'} Q(x, a')$
  $\theta_j \leftarrow r + \gamma \theta_j(x', a^*), \quad \forall j$
  # Compute quantile regression loss (Equation ??) $\sum_{i=1\ j}^{N} \left[\rho^{\kappa}_{\hat{\tau}_i}(\theta_j - \theta_i(x,a))\right]$ =0

Figure 1.5: Algorithm: Quantile Regression Deep Q-learning

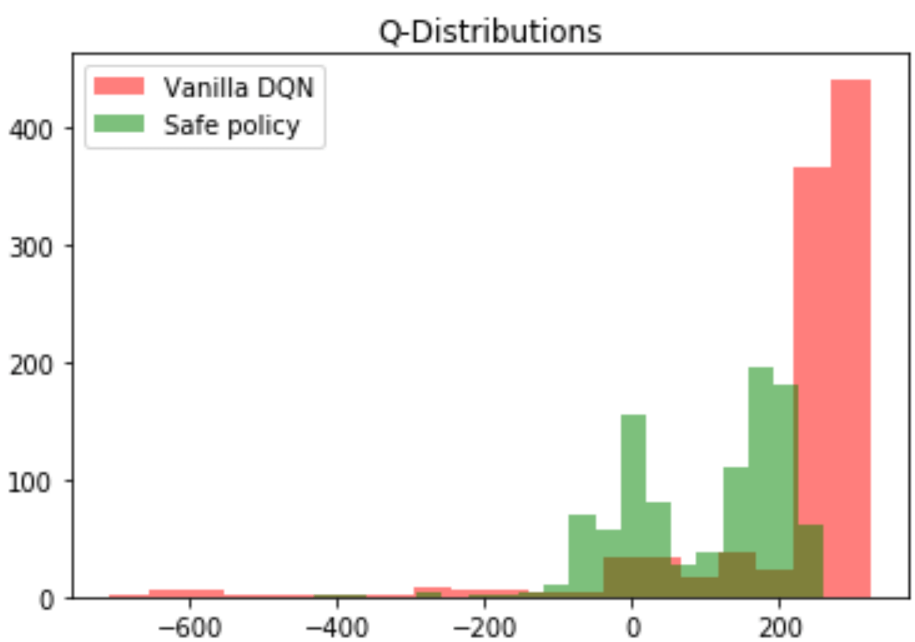


Figure 1.6: Comparison of Vanilla Deep Q-learning and Risk sensitive Q-learning

1.3 Amazon projects – High Confidence Predictions leveraging Monte Carlo Dropout

1.3.1 Context

We have worked with Amazon Luxembourg for 6 months as an Applied Scientist intern in the supply chain Science team. In that time period, we have worked on a Deep Learning project using a multimodal Neural Network (Image branch using Computer Vision and transfer learning, Text branch using NLP, Numerical branch and high cardinality categorical branches as well as numerical branch with categorical and continuous variables) to predict the Amazon products in packaging requirements.

1.3.2 Challenges and results

The existing model was predicting on all the Amazon products in Europe with a reasonable accuracy, but couldn't be deployed on a very large number of products because even a small proportion of errors could lead to a large number of wrongly predicted products.
We have exclusively worked on a pipeline of tasks to return confidence scores[3] along with the predictions in order to have a subset of products for which the accuracy is much higher than the initial one. The method we have developed leverages MonteCarlo Dropout and goes further into setting and proving a rigorous mathematical and statistical framework for computing confidence scores. According to the manager of our team, this cutting edge technology is still not democratised in business and establishes a great improvement and advantage over other work done in different teams.
We have collaborated with the business team and the success of the project led to the deployment of a pilot to automatically switch products packaging.

1.3.3 Technical details

In this context, we have:

- Implemented Hyperparameters tuning using the Python HyperOpt library, which we used to tune the whole architecture of the network and its regularisations, we also introduced Dropout[4] layers in the Network without impairing the model thanks to the hypertuning of the Dropout rates
- Implemented Monte Carlo Dropout and introduced a confidence metric for which we built rigorous statistical framework
- Increased the F1-Score by 26 to 29 percentile points on validation set by restricting the set to high confidence predictions ($\sim$ half of the validation set)

Code was written in Python using the PyTorch library.

[3]See Figure 1.8
[4]See Figure 1.7

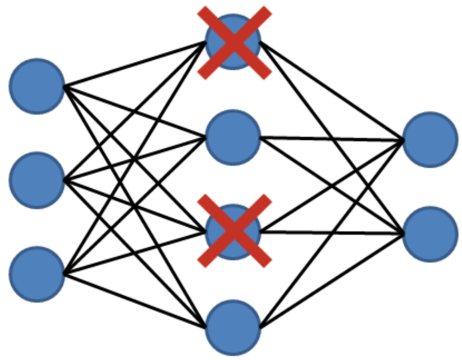


Figure 1.7: Dropout layers

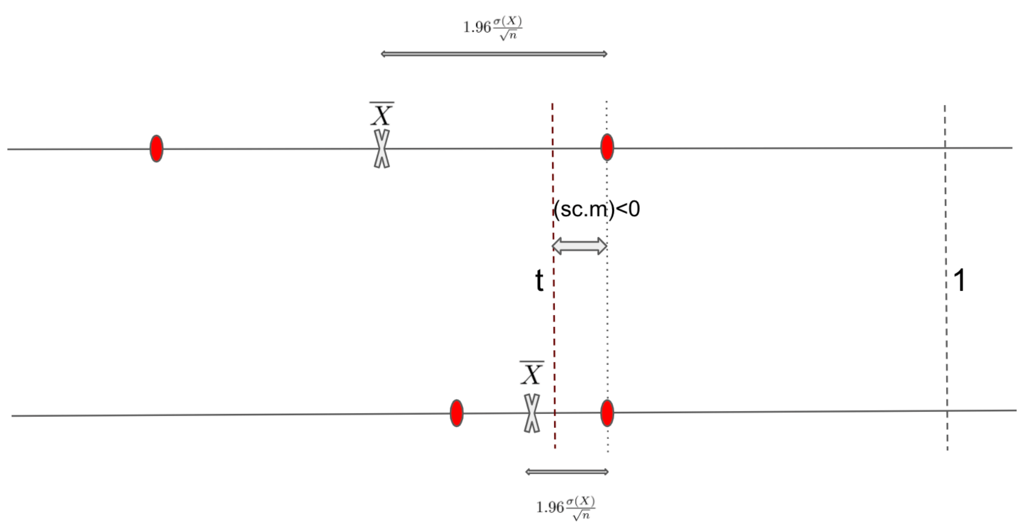


Figure 1.8: Statistical meaning of the predictions' confidence scores

1.4 Data Science Mission with the hospitals of Paris – Causal inference: Assessing the effect of Tranexamic Acid on the mortality of patients with head trauma

1.4.1 Context

We worked in a project in conjunction with the doctors committee of the hospitals of Paris, the goal was to assess the effect of an empirical treatment (Tranexamic Acid) on the mortality of severely damaged patients with head trauma.
We have worked on the whole pipeline of Data Science on a raw extensive Data set with 7000 patients and 350 variables, we have:

1. Preprocessed the data,
2. Performed Missing Values Analysis, imputed missing data using iterative Factorial Analysis for Mixed Data, thus taking into account the interdependance of continuous and categorical covariates
3. Performed distributions analysis to assess the quality of the data imputation
4. Compared the data imputation with the MissForest[5] imputation for further analysis
5. Performed data analysis using multiple methods: Factorial Analysis for Mixed Data[6] (See Figure 1.10), Hierarchical clustering (See Figure 1.11) as well as other custom analyses focusing on the variables of interest
6. Performed advanced causal inference method to assess the intrinsic effect of Tranexamic Acid on the survival of patients: Matching, FAMD Matching, Average Treatment Effect (See Figure 1.12) with Inverse Propensity Weighting (See Figure 1.13) and finally the Double Robust method

1.4.2 Challenges

The project we've worked on was a full stack Data Science project with custom uncommon problematic related to Causal Inference (See Figure 1.9 for illustration), we had to leverage state-of-the-art research to solve the problematic, in addition to the use of multiple Machine Learning models as intermediate steps during the project.

1.4.3 Technical details

The whole project was coded in R, using custom packages for the different steps and methods of the project.
The conclusion of the project (See Figure 1.14) led the doctors committee to pause the use of Tranexamic Acid for head trauma.

[5] Nonparametric missing value imputation for mixed-type data using Random Forests
[6] FAMD: PCA like method for data including continuous and categorical variables

Potential cause(W)	Effect (Y)	Potential bias or real cause (X)
Umbrella sales	Number of car accidents	Weather (rain)
Temperature	Weights of chicks	Gender
Treatment	Survival	Initial state of the patient

Figure 1.9: Causality exemples: Correlated covariate vs. real causes

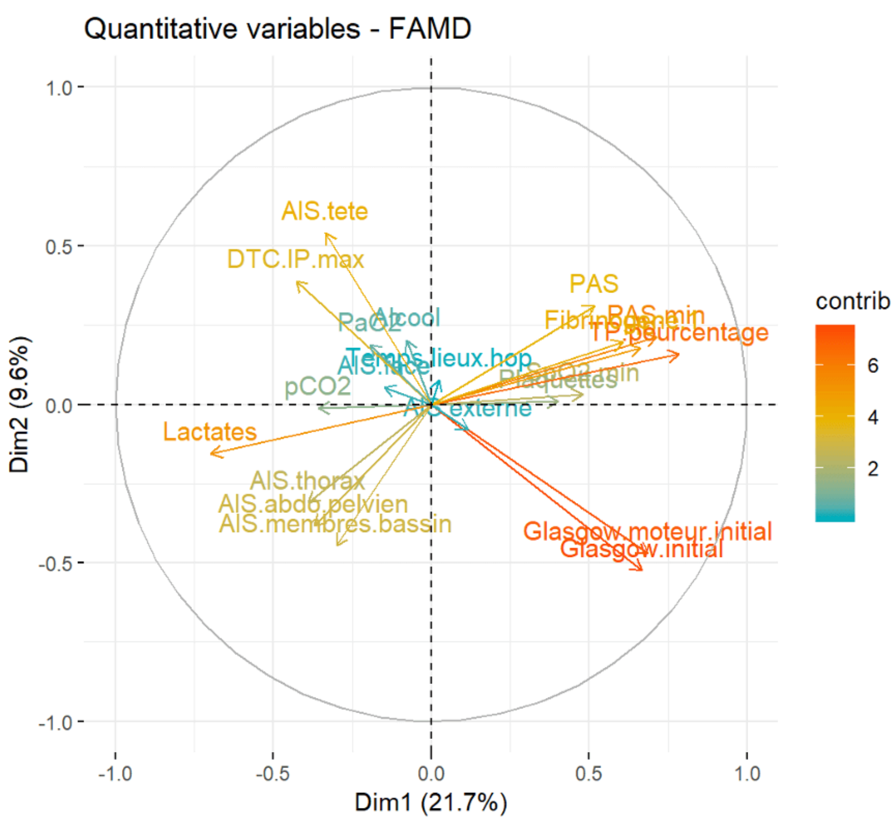


Figure 1.10: FAMD on the data represented on the 2 main components

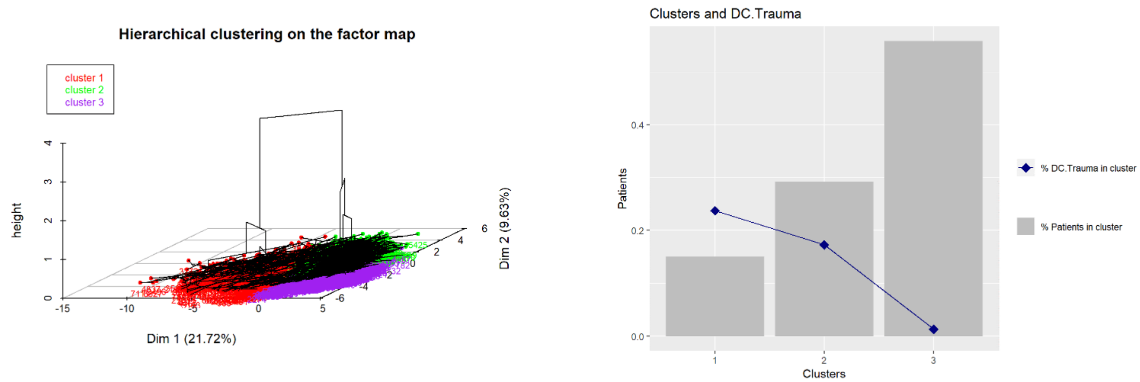
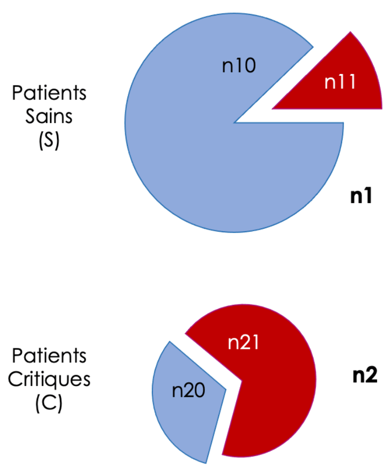


Figure 1.11: Hierarchical Clustering visualised for 3 clusters on the 2 main components returned by the FAMD. The mortality rate is heterogeneous across the clusters



$$\hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{\{X_i = x, W_i = 1\}} Y_i - \frac{1}{n_{x0}} \sum_{\{X_i = x, W_i = 0\}} Y_i.$$

$$\hat{\tau} = \sum_{x \in \mathcal{X}} \frac{n_x}{n} \hat{\tau}(x)$$

Figure 1.12: The Tranexamic Acid is more often administrated to patients in critical states (The red proportion receives the treatment), which explains the need to perform causal inference before concluding its harmfulness

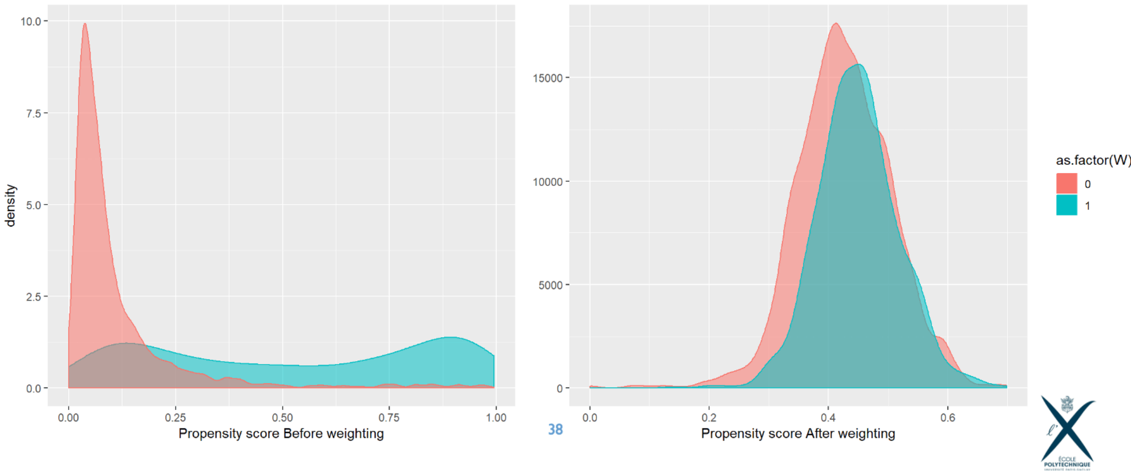


Figure 1.13: Distribution of the Propensity score returned by both regressions without and with the weighting

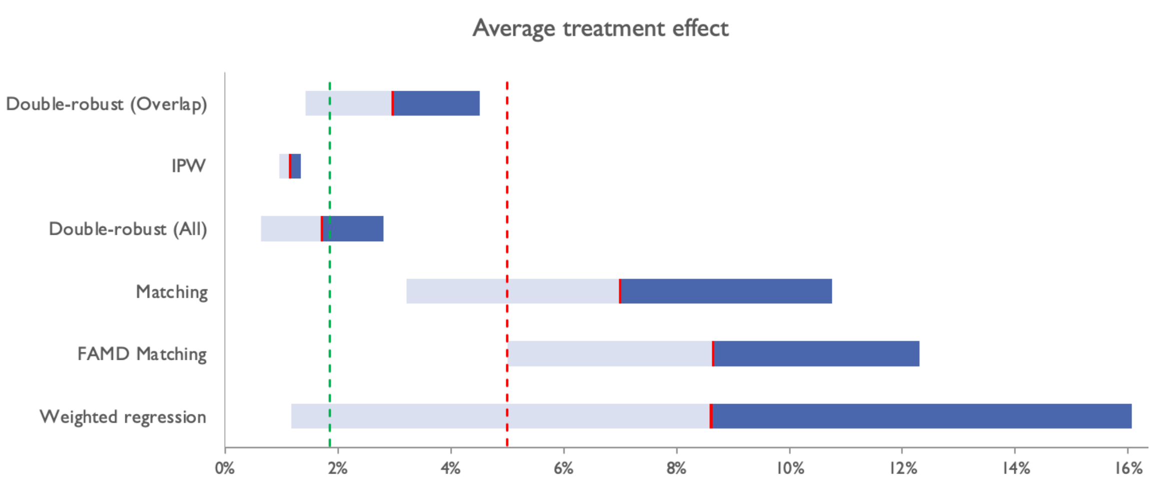


Figure 1.14: Comparison of the multiple computations of the ATE: They all suggest the uselessness of the treatment

1.5 Willow Machine Learning project – Automatic Inpainting with SinGAN

1.5.1 Context

We have worked on a research project with Willow[7], for which we developed an automatic inpainting algorithm on natural images using the SinGAN technology.
Inpainting is the process of editing images or videos in order to remove an object. The industry standard of inpainting is the Photoshop Content Aware software that performs very well with human supervision (semi-automatically), however the results are drastically lower when used without a human guiding and correcting the software (automatically).

1.5.2 Challenges

The goal of this project is to build a fully automatic and realistic photo/video editing software, as it is not scalable to ask graphic designers to edit millions of photos from a large database. The software we build is capable of automatically selecting an element and remove from every data point of the database.
The main challenge of this project is obtaining realistic results that outperfom industry standards.

1.5.3 Technical details

After a thorough investigation of the Machine Learning options to solve this problem, we have opted for a state-of-the art technology: the SinGAN. This technology captures the structure of a single natural image sequentially at different scales. The architecture of the model is basically a multi-layer generative model applied to one natural image only. The dataset of each submodel consists of the sub-images extracted from the original image on a specific scale. See Figure 1.15.
We have managed to implement this technology with an adaptation to fit our specific task.
The plain use of SinGAN was not enough to beat Photoshop Content Aware performances, we ended up introducing a relevant distance in the space of images: Single Image Frechet Inception Distance (SIFID) which is a single image version of the Frechet Inception Distance. This allowed us to compute an intuitive "realism comparator" between the original image and the generated ones.
We developed the following pipeline to removing an object from the image:

1. We train a SinGAN model on the original image capturing its structure on different levels
2. We naively remove the object by replacing it with an average of its neighboring pixels
3. We use this image to generate several images using the SinGAN model with different scales
4. Finally, we use SIFID to decide which one of this images is the most realistic compared to the original one

At the end, we delivered to the Willow Project team an automatic inpainting software based on the SinGAN technology that beats Photoshop Content Aware in the realism survey: 69.2% **score** for the SinGAN software vs. 30.8% **score** for the Photoshop Content Aware software. See Figures 1.16 and 1.17

[7]Computer vision and machine learning research laboratory

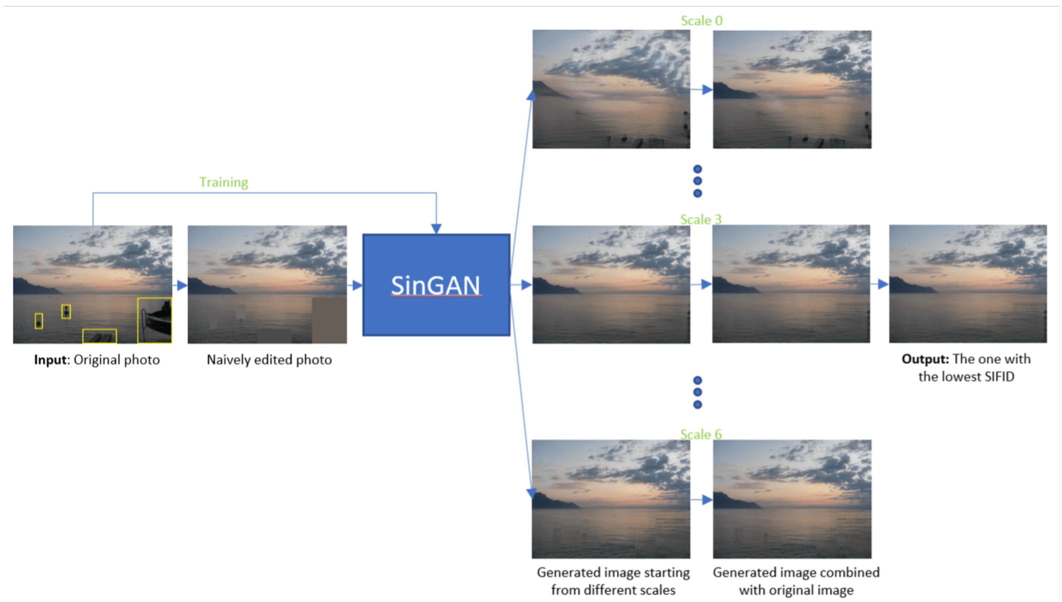


Figure 1.15: Outline of the Inpainting algorithm we developed

Figure 1.16: Comparison between SinGAN and Photoshop Content Aware

Image	Photoshop Content Aware Move	SinGAG Inpaint
Lake	54%	46%
Cows	38%	62%
Wave	69%	31%
Tree	100%	0%
Night Lake	85%	15%

Figure 1.17: Survey results comparing the two technologies

The inpaintings using Photoshop Content Aware and SinGAN are presented and the answer asked is: "Knowing that one is real and one is fake, which one is the fake one?"

1.6 Oxford Machine Learning Kaggle Challenge – Predicting binding molecules

1.6.1 Context

We have participated in a Kaggle challenge where the goal was to predict whether a molecule will bind to proteins or not, without using any mechanical or physical law.
The data consists of different molecules that are either "active" if they can bind to different proteins, or "decoy" if they are assumed not to bind to any protein. Each molecule corresponds to one row of the data set and has 349 associated features, and a label (1 if the molecule is active, 0 if it is decoy).
We used this data to build a model that predicts the behavior of previously unseen molecules. The pipeline of the study starts with an exploratory data analysis, then a discussion of the methods used to preprocess the data, the different models built, and how the final prediction was made.

1.6.2 Challenges and results

The objective of this project was to find a suitable model to predict the behavior of previously unseen molecules. For a given molecule, we wanted to know the probability it had to bind to proteins. Our final model was given by an average of three classifiers: A random forest, a gradient boosting tree, and a neural network (with a specific structure to deal with the class imbalance using a custom initial bias in the penultimate layer).
One of the main challenges faced during this work was to deal with the strong class imbalance in the training data (only 6447 molecules out of 378447 are binding). To do so, we leveraged on various sampling methods, such as oversampling, undersampling or cluster-based sampling (See Figures 1.18 and 1.19). Each one of these methods had some inconveniences, but they allowed us to train our models on balanced sets. Using these strategies and refining them was the main source of improvement of our predictions.

1.6.3 Technical details

On the computational side, we used Python and the models were built with the libraries Scikit-Learn, XGBoost and Tensorflow. The code used for this project can be found at the end of the report.

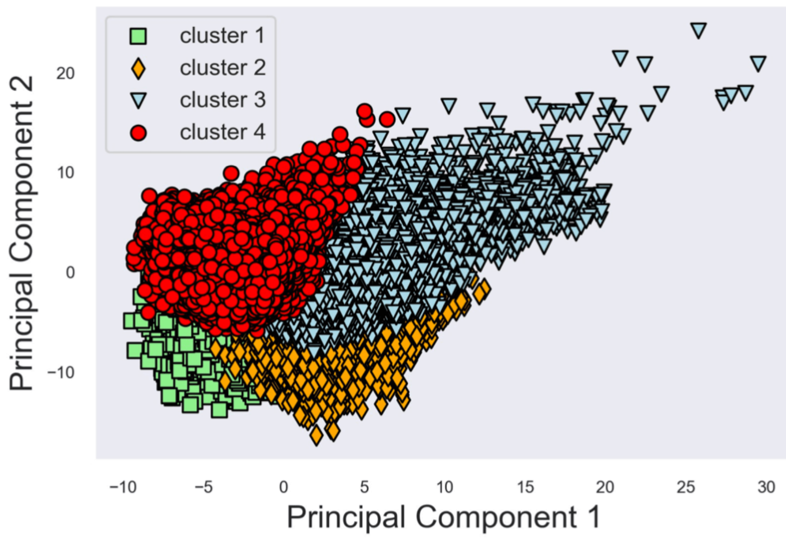


Figure 1.18: Four clusters of majority class against two first PCA components.

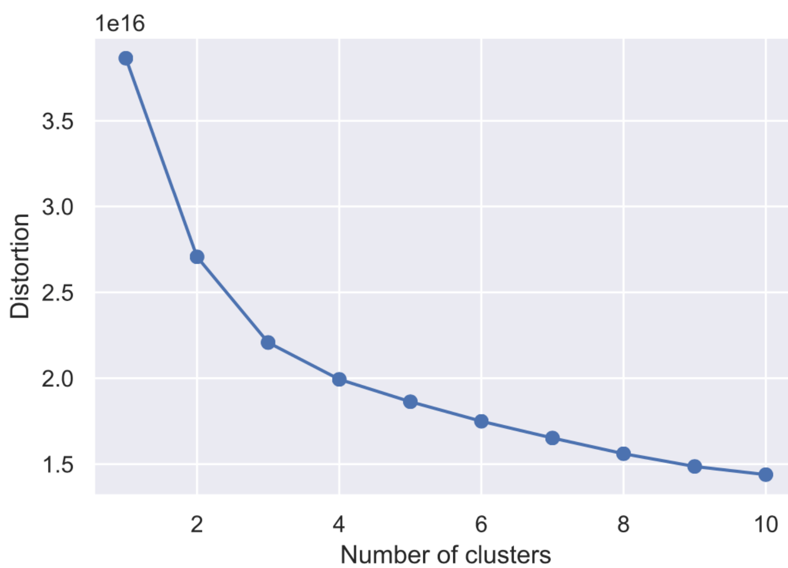


Figure 1.19: Distortion score after K-Means clustering for different values of K

1.7 Inria Machine Learning project – Active Learning in Histology Medical Imagery

1.7.1 Context

We worked with the TAU team from Inria Saclay on an active learning model which goal is to automatically detect the presence or not of certain compounds in microscopique biological tissues in the Atlas of bleast Histology dataset.

1.7.2 Challenges

The main challenge is that although the dataset is large, it is not labeled. Moreover labelling the data points (determining presence of the compounds) is a costly operation: in order for each datapoint to be labeled, it has to be analyzed by an expert which is in this case a specialized doctor. Randomly selecting a subset to be labeled by experts would work but it is costly. We needed to come up with an approach to efficiently select a small number of data points to be labelled and which are sufficient to train the model to a satisfactory accuracy.

1.7.3 Technical details

The technology we used to solve this problem is Active Learning. The optimisation pipeline we used is as follows:

1. At first a small random subset of the data is given to experts to be labelled
2. Then a model is then trained on that subset
3. Its behaviour on the dataset is analyzed using specified metrics (based on the entropy of the model)
4. The previous metric allows us to select the optimal subset to be labeled by the experts to continue the training
5. Finally, the model is trained on that subset and its performance is improved

We used a mixed technique based on the Query by committee and the Entropy sampling to deduce which additional data would be worth labelling. This allows us to improve the quality of the model using a minimal amount of labelled data. This process is iterated to sequentially to subsets of the data to be labelled until the performance is maxed out.
Our algorithm reached a 92% accuracy score on the validation set with less than 800 labelled examples versus a 81% accuracy score with a 2000 randomly labelled examples.

1.8 IBM Research – Deep adversarial training for regularization

1.8.1 Context

We have worked in IBM Research Labs (Department of AI) where we developed a technology that uses Adversarial training to improve the generalization capability of Neural Networks.
The idea is to make use of 2 models instead of 1 model (the original classifier and the discriminator h that's only here to improve the classifier f) with a split of the data to be used in the training. These two models will have conflictual Loss functions whence the denomination Adversarial Training.

1.8.2 Challenges and contribution

The set of models to be trained have conflictual and custom losses that need to be addressed in specific ways (either on Tensorflow or PyTorch). Indeed, unlike in the standard or vanilla neural networks, the gradient of the losses flow backwards and update the parameters of both the networks.
This work is a novel way of regularization that we compared with other regularization techniques and proved to be particularly efficient (and in many experiments outperforms the most common regularisation type: the $\mathcal{L}^2$ regularisation), see Figure 1.23. It is inspired by one of the applications of GANs (See general principle of GANs in Figure 1.20) in ML: Membership Privacy.
We have reduced overfitting by a factor 2 on different datasets ($\sim 1k \times 20$) using this novel technology.

1.8.3 Technical details

We have used a modified version of neural networks training as described in Figure 1.21. The idea is to split the training data into D and D' (See Figure 1.22) and only feed D to f instead of feeding it $D \cup D'$, the whole $D \cup D'$ is given to h to try to determine whether a point is in D or D'. By fooling it, f modifies the distribution of x, y to be independent from D. Thus, the output is less "overfitted" to D and more generalizable.
We have coded in Python and used ScikitLearn for the regressions and the K Nearest Neighbors, and Tensorflow for the Neural Networks and tracking the custom losses.

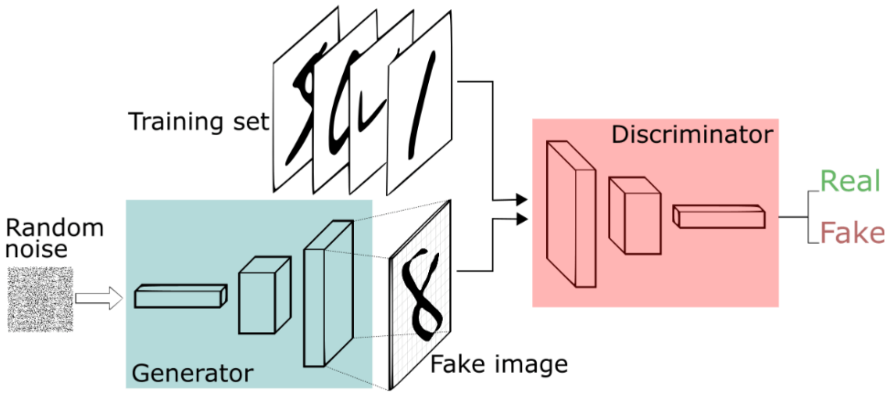


Figure: GANs principle

Figure 1.20: GANs principle

---

for number of the training epochs **do**

  for k steps **do**

    Randomly sample a mini-batch of m training data points $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$ from the training set D.

    Randomly sample a mini-batch of m reference data points $\{(x'_1, y'_1), (x'_2, y'_2), \cdots, (x'_m, y'_m)\}$ from the reference set D'.

    Update the inference model h by ascending its stochastic gradients over its parameters ω:

$$\nabla_\omega \frac{\lambda}{2m}\left(\sum_{i=1}^{m} \log(h(x_i, y_i, f(x_i))) + \sum_{i=1}^{m} (\log(1 - h(x'_i, y'_i, f(x'_i))))\right)$$

  end for

  Randomly sample a fresh mini-batch of m training data points $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$ from D.

  Update the classification model f by descending its stochastic gradients over its parameters θ:

$$\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} ((f(x_i), y_i) + \lambda \log(h(x_i, y_i, f(x_i))))$$

end for=0

Figure 1.21: Custom adversarial training

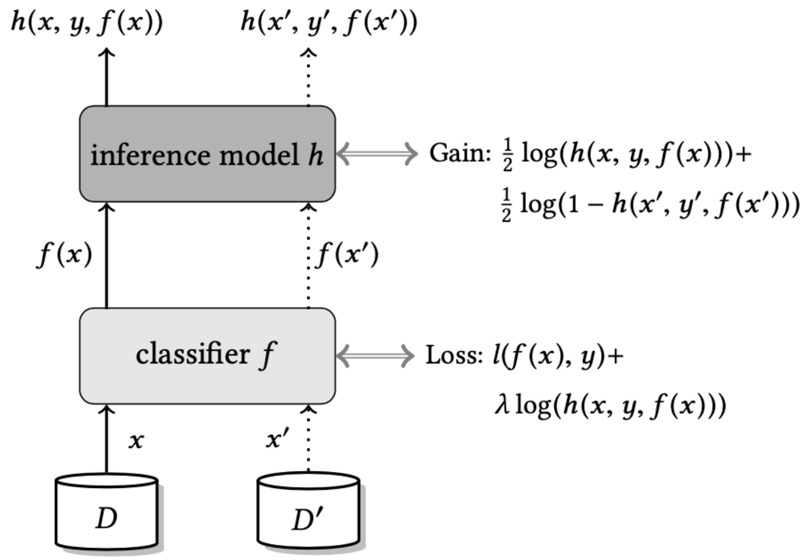


Figure 1.22: Classification loss and inference gain, on the training dataset D and reference dataset D', in our adversarial training. The classification loss is computed over D, but, the inference gain is computed on both sets.

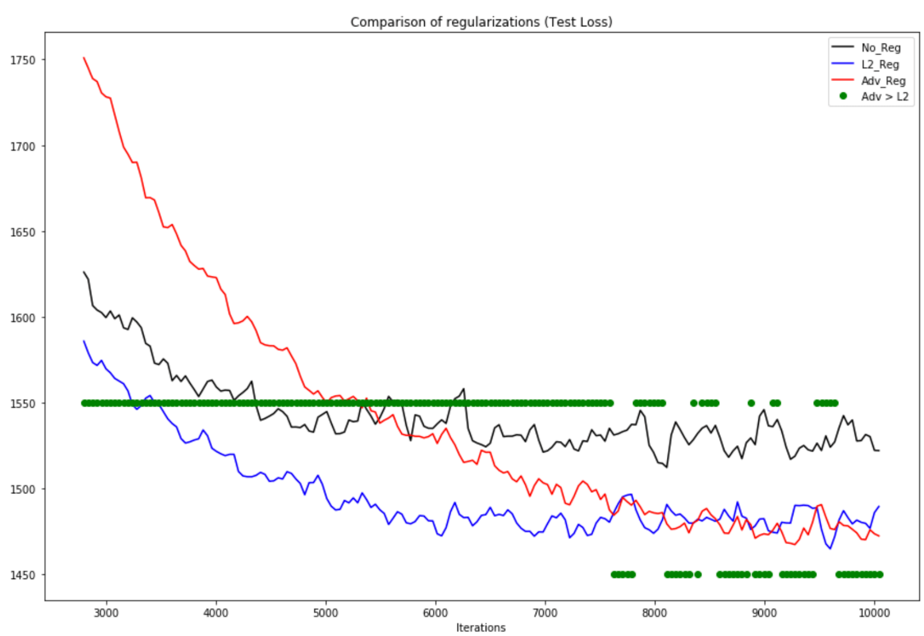


Figure 1.23: Comparison of the losses: Adversarial Regularization vs. $\mathcal{L}^2$ Regularization

1.9 Oxford Research project – Predicting results of a competition using Markov Chain Monte Carlo

1.9.1 Context

We worked on a Bayesian Statistics project where we are given data that represents records of $L = 134$ games played by $n = 24$ players, in each game a small subset of players participate and we keep track of their rankings in each game and their age seniority amongst all the players present in the field of game (not only those participating at that time), the age seniority is a number from 1 to $n_a = 16$.
We performed bayesian analysis of the dataset starting with prior elicitation under the Placket-Luce model before exploratory data analysis, we focused on 2 models that either account for the effect of seniority or not. First, we modelled the data and computed the posterior (See Figures 1.24 and 1.25) using an MCMC approach, then we used that to do parameters estimation, hypothesis testing and model selection (using Bayes Factor).

1.9.2 Challenges and technical details

We have performed Bayesian Data analysis starting from non-informative priors to model the outcomes of the competition taking into account skills of players and seniority ranks as covariates. The main difficulty with such a model is that the problem is high dimensional (40 dimensions). To deal with this issue, we opted for the Gibbs sampler, indeed due to geometric reasons (e.g the volume of a sphere increases exponentially with the dimension) the likelihood in such high dimensions becomes very localised, so to make sure we can go towards the support of the posterior distribution, we made sure the process makes small steps (i.e component by component).
The sequential scan is often better than the random scan Gibbs sampler, in this case this is true. Intuitively, the sequential scan makes sure to go over all the components so the autocorrelation at lag (See Figure 1.26) goes down lower than in the random scan.
We have used the language R for the code and the report.

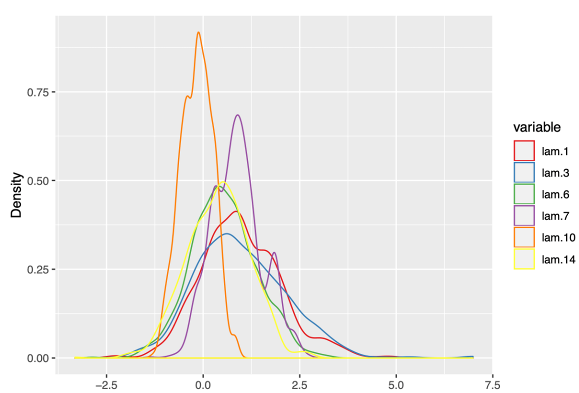


Figure 1.24: Posterior densities of some skills

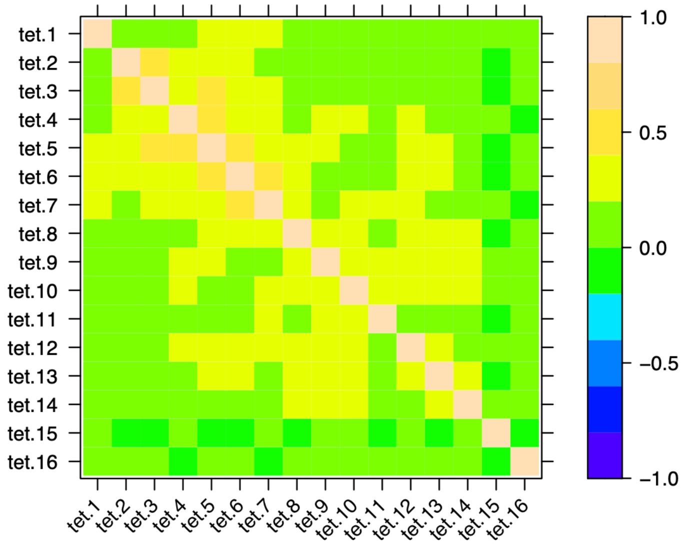


Figure 1.25: Posterior correlation of the theta's

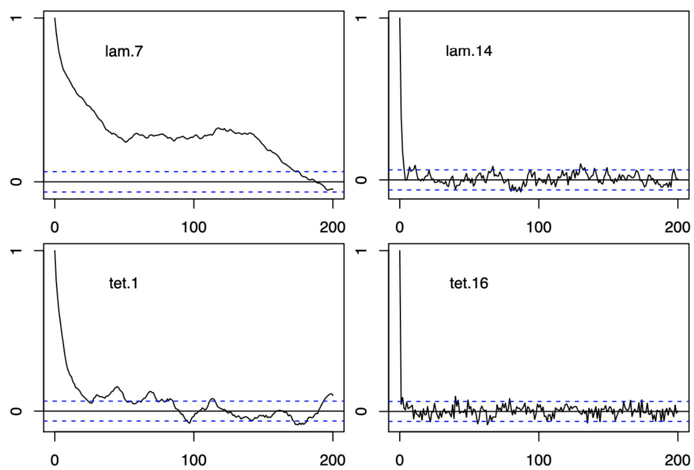


Figure 1.26: ACF plots

Chapter 2

Data Science

2.1 Oxford Data Science project – Predicting coronary heart disease using Generalized Linear Models

We have worked on a Data Science project for which we were given data that consists of records of coronary heart disease status for 4658 individuals. For each individual, heart disease status is recorded as well as some other biological measurements and social characteristics about the individual.
We performed exploratory data analysis where we extracted data correlation and other graphical interpretations (See example in Figure 2.1). Then we built a model with explanatory variables using a generalised linear model with the Bernoulli family (logistic regression), performed outliers analysis (See Figure 2.2) and interpreted effects of different biological measurements and social characteristics.

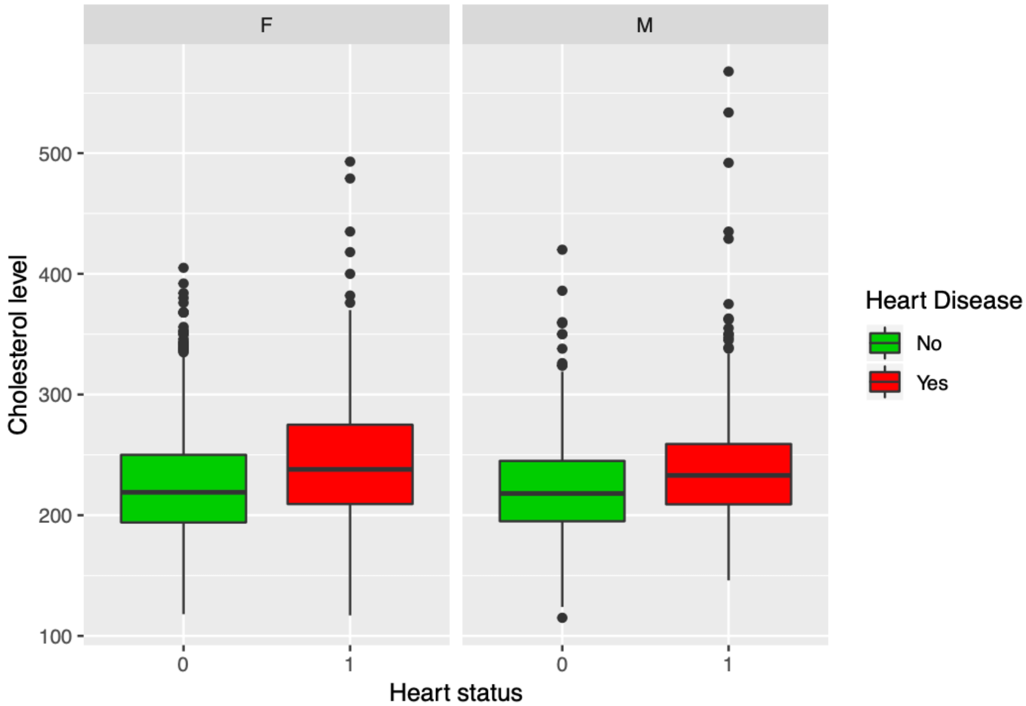


Figure 2.1: Boxplots of Cholesterol level by disease status and sex

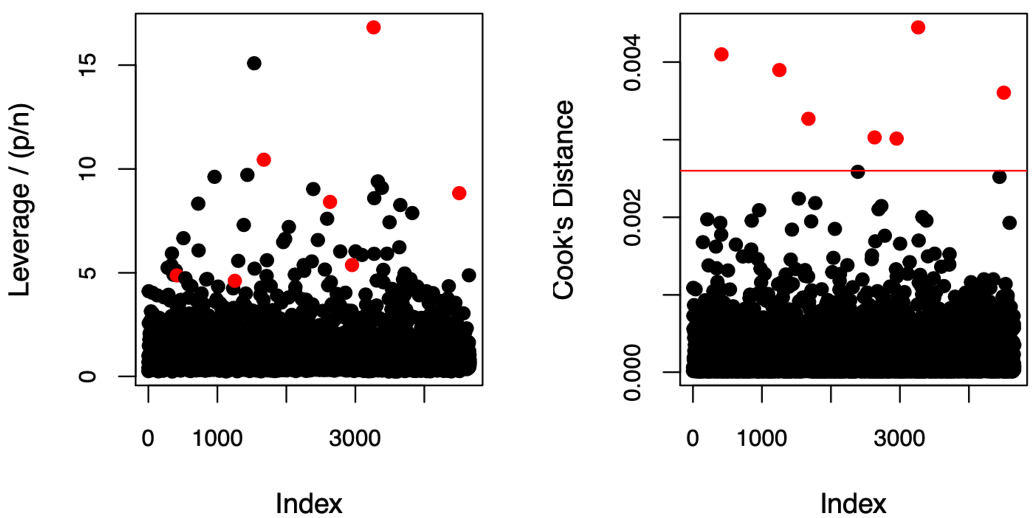


Figure 2.2: Multiplicative model: Outliers analysis

2.2 Oxford Data Science project – Predicting swimming performance records using custom regression models

2.2.1 Context

I've worked on an Oxford project for which we have designed and compared different custom regression models to predict swimming times as a function of other variables as the distance, stroke, sex and length of the course.

2.2.2 Challenges and technical details

We have performed exploratory data analysis using relevant plots (See Figure 2.3) and summary statistics, this step allowed me to notice unusual dispersion of the response (time) as a function of distance, which requires special treatment and could suggest custom models. Then we performed model selection, outliers analysis (See Figure 2.5) and model interpretation.
The common way to deal with the dispersion of the response is the use of weighted regression, we have used weights proportional to the standard deviation of the response, the weighted regression proved better than the standard regression (See Figure 2.4).
To further improve the performance, we have introduced a custom multiplicative model of the form $time = dist \times (e^{\sum other}) \times (randomness)$ where *other* are categorical covariates and thus introduce a multiplication by a constant depending on the state of those covariates. This model accounts for the dispersion seen in *time* which is resulting from the log-normal noise in *randomness* and amplified by the value of *distance*, this model is also a natural consequence of the physical link between time and distance which is velocity.
The final model greatly improves over the observations and also has more physical sense, the mathematical expression reads $time = dist^{1.1} \times \ldots$, which accounts for the tiredness of the swimmers (i.e. the average velocity decreases with the distance), moreover the model is backed up by the Box-Cox model which returns the same exponent.

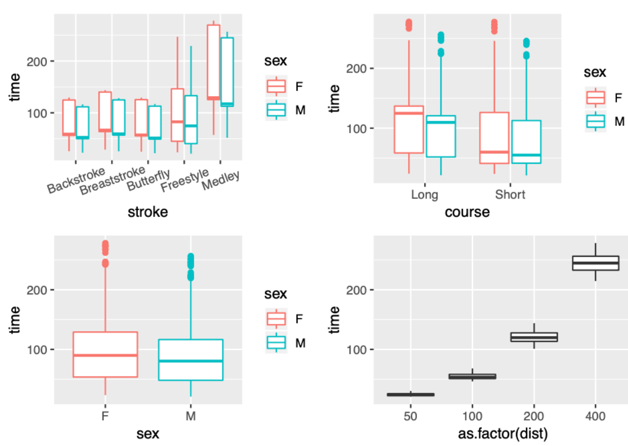


Figure 2.3: ScatterPlots of time against other variables

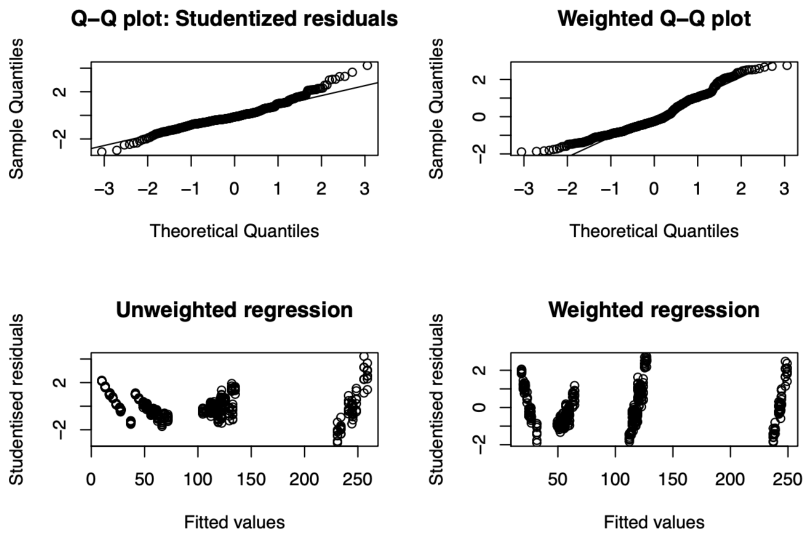


Figure 2.4: Model comparison: Normality and regression quality

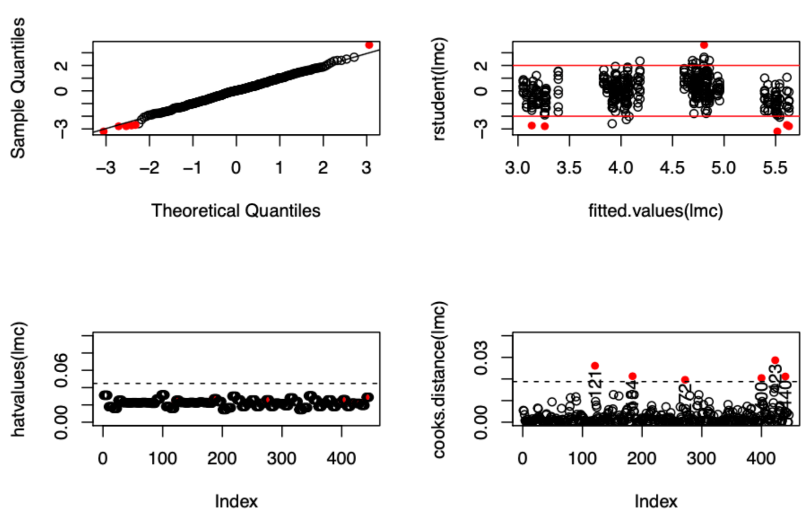


Figure 2.5: Multiplicative model: Outliers analysis

Chapter 3

Research: Operations Reaseach, Theoretical Machine Leaning, Applied Maths

3.1 Machine Learning Research project – Online Learning Pandora's Box

3.1.1 Context

We conceptualized and solved the Online Learning Pandora's Box (OLPB) problem with ENSAE CREST statistic department.
The OLPB problem models sequential decision-making situations where:

- Resources are limited
- Environment specifics are unknown but can be learnt through the process

See figure 3.1 for the mathematical formulation of the problem. To simplify the concept, we will introduce two real-life case studies that can be represented by this model:

1. The first situation models an investor, his limited resource is the capital that he can invest through the year. Opportunities arrive one-by-one to the investor, when an investment opportunity arrives, the investor only has access to some indicators (numbers) based on which he chooses whether to study the opportunity or not. The choice of studying the opportunity costs capital, after the study he can choose to invest or not. The learning dimension is in the studies he does, in this situation, he is new in the market so in the beginning the value of the investment opportunity is poorly judged by the investor but he can learn from the first studies he does how to better judge opportunities.
2. The second real-life situation is a hospital confronted to a new decease. The limited resources in this case are the doctors and the necessary equipment to perform a full examination of the patient, for example only a limited number X of examinations can be made each month.

When patients arrive, they have symptoms based on which we can choose to carry out the examination or not. In the beginning the relationship between the symptoms and the decease are unknown, so we learn it to optimize when to carry out the medical examination.

3.1.2 Challenges

This model is based on the Online Pandora's Box model which does not contain the learning dimension. The learning dimension adds an exploration-exploitation dilemma: when the trader chooses to study the opportunity or not, there is an exploration incentive: learn to judge a promising investment a priori, and there is also an exploitation incentive: find a good investment. In the same way the hospital have an exploration incentive: learn more about the decease and its symptoms, and an exploitation incentive: detect the patients with the decease in order to treat them.
This exploration-exploitation dilemma has to be dealt with in the OLPB problem, it adds a new challenge to the original problem: the Online Pandora's Box.

3.1.3 Technical details

We formalized the problem and tested different approaches to solve it. In each iteration, we analyzed the proposed solution and its limits in order to propose a better approach. In the beginning, we tried to reduce the problem to an equivalent Prophet Inequality problem as there already exist standard solutions to the Prophet Inequality problems. In order for this reduction to work, we needed to solve the equivalent Prophet Inequality with a threshold α-approximation algorithm, however, the new learning dimension makes it sub-optimal.
The following approach was to adapt the Explore-Then-Commit method to our problem because of its easy implementation in spite of its sub-optimality. It has proven to work better than the first approach which made us push more towards Multi-Arms Bandits solutions. The final approach allowed us to attain optimality using what's known as an optimistic approach: it is a UCB-like approach where we predict the value of the opportunity optimistically. Since we are computing the upper confidence bound, a high value means either a high uncertainty about the choice or a confident high predicted value. By being optimistic we align the exploration and the exploitation choices. See figure 3.2 for this algorithm outline.
Our final approach was optimal and we managed to formalize the problem in a well defined model and then solve it to optimality.

We have n boxes, each box has a cost $c_i \in R^+$, a value $v_i \in R^+$, and an indicators vector $t_i \in R^p$. For each box i, we have access to F_i the distribution of (t_i, c_i). We assume that $(F_i)_i$ are independent and we assume that $v_i|t_i \sim \mathcal{N}(\theta^T t_i, \sigma^2)$ with $\theta \in R^p$ a parameter and $\sigma^2 > 0$. When a box is presented, you have access to its indicators vector t_i and you choose to open the box or not. If the box is opened, you get access to the cost c_i and you pay it, you also get access to the value v_i and you choose irrevocably to keep it or not. The set of opened boxes S and the set of kept values R must respect some constraints $\mathcal{F}$ than can depend on the types of the boxes. The reward Pandora is getting is the difference between the kept values and the payed costs:

$$r = \sum_{i \in R} v_i - \sum_{i \in S} c_i$$

The goal is to find an algorithm that maximizes the utility:

$$u_{Alg} = E[\sum_{i \in R} v_i - \sum_{i \in S} c_i]$$

Figure 3.1: Online Learning Pandora's Box setting

Let $\tau_{i,\theta}$ be a threshold strategy for the equivalent prophet inequality assuming θ.

For each step:

- Let CR be a confident region for θ using the linear regression on the past opened boxes data.
- If $\exists \theta \in CR, \sigma_{i,\theta} \geq \tau_{i,\theta}$, then open the box (this condition is verified through a discretization of CR)
- If the box is opened and $\forall \theta \in CR, v_i \geq \tau_{i,\theta}$, then keep the box

Figure 3.2: Outline of the optimistic approach OLPB solver

3.2 Operations Research project for an anonymous energy supplier – Optimizing energy pricing in a two-layered Karush-Kuhn-Tucker model with the possibility of energy stocking

3.2.1 Context

We have worked on a confidential project for an energy supplier company. The project regarded the pricing of energy with respect to customers behaviour to optimize a complex high dimension cost-related quantity.

3.2.2 Contribution and challenges

The project treated a real-life need and the optimization problem that we formulated took into account real-life data of customers behaviours in terms of consumption and pricing decisions. The objective we formulated was a two-layered optimisation problem for which KKT conditions of the lower layer were used to simplify the problem into a one-layer, multi-parameters optimisation problem. Moreover, we have added and modelled the possibility of stocking energy by the consumer into the optimization as many clients now turn towards buying a solar panel.
Our formulation of the problem greatly improved on the existing model.

3.2.3 Technical details

Before solving the formulated problem, we applied the K-means ++ to cluster the clients using the scikit-learn library in order to reduce the dimension of the optimization problem. Then, we ended up with a mixed-integer quadratic problem (MIQP) that we implemented and solved using the IBM optimization libraries CPLEX and DOcplex.

3.3 CMAP Research project – Probability estimation of rare events: Non-extinction of an endangered species

3.3.1 Context

We have worked on a research project with the CMAP[1] for which the goal is to simulate and compute probabilities of rare events ($p \sim 10^{-8}$). The case study we have used for this project is the Galton-Watson model for simulating dynamics of population (See Figure 3.3 for genealogical representation).
At each generation t, each member of the population breeds a random number of offsprings before dying, the total number of offsprings renders the population size at generation $t+1$, this constitutes a Markov Chain for which we study the characteristics. When the average number of offsprings per individual is lower than 1, the population is doomed to the extinction.

3.3.2 Challenges and technical details

The difficulty that rises when estimating rare event probabilities is that standard Monte Carlo estimations no longer work because for an event of probability 10^{-6} for example, using Monte Carlo with a million simulations will only render a very few or no occurrence of the event of interest, thus making the Monte Carlo estimate noisy or even unusable.
We used mathematical properties of the Process of interest to calculate exact values of some rare events for further comparisons with the developed methods (See Figure 3.4 for exact probability calculations). In addition to the naive Monte Carlo estimation (See Figure 3.5), we have used 2 advanced simulation methods to estimate very low probabilities with reasonable computational resources:

1. Importance Sampling: We modify the probability distribution of the whole process and use Monte Carlo with the modified distribution multiplied by a likelihood term to render the original expectation. The new distribution is supposed to allow the rare event more often. See Figure 3.7.

2. Interacting Particle Systems[2]: We used a Markov Chain method that leverages a similar process inspired from Genetic Algorithms. The idea is to have M trajectories of the tree history of the population and at each generation 2 steps are applied: the selection step for which trajectories are sampled with respect to a potential that we define according the trajectories we want to favor for the occurrence of the rare event, the second step is the mutation for which the next generation population is computed. The estimate is then computed using an average corrected by an expression derived from the potential of the trajectories at different generations. See Figure 3.6.

Furthermore, we show an interesting result: if we combine 2 unviable habitats, one with good conditions but with occasional disasters and the second with constant slightly bad conditions. Then we can mathematically prove and empirically verify that there is a specific optimal proportion[3] of the population to keep in the first habitat for the population to thrive exponentially. See Figures 3.8, 3.9 and 3.10.

[1] Centre de Mathématiques Appliquées de l'École polytechnique
[2] This is similar to Sequential Importance Sampling with Resampling in Particle Filters
[3] See paper

Figure 3.3: Genealogical representation of a population evolution

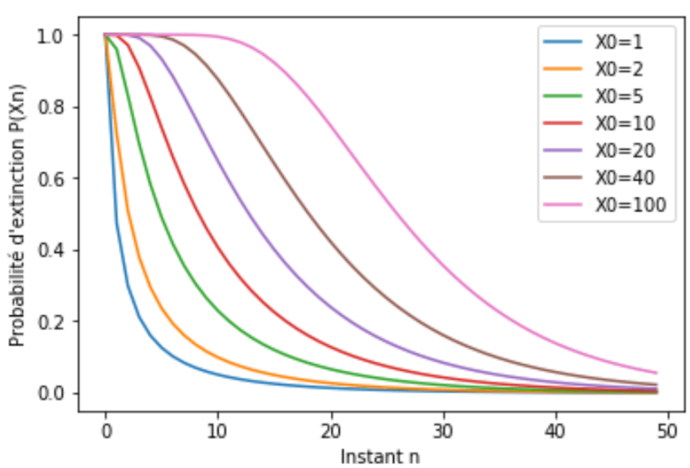


Figure 3.4: Probability of non extinction at generation n: $P(X_n > 0|X_0 = X0)$ for different initial population values $X0$. The average offsprings number per individual is $m = 0.9$

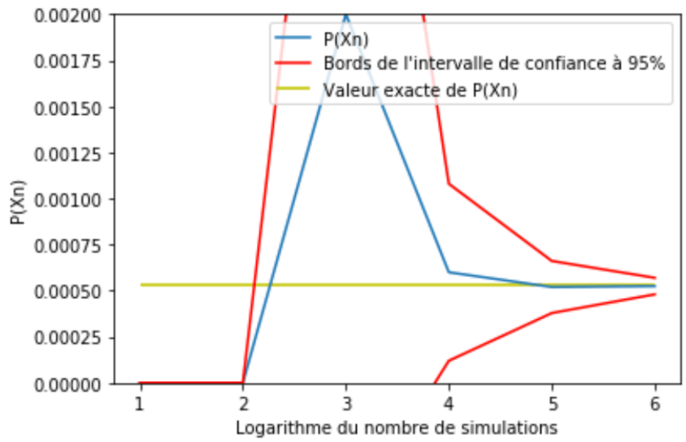


Figure 3.5: Naive Monte Carlo estimate: $P(X_{40} > 0|X_0 = 20)$ for $m = 0.8$ and 10^x simulations

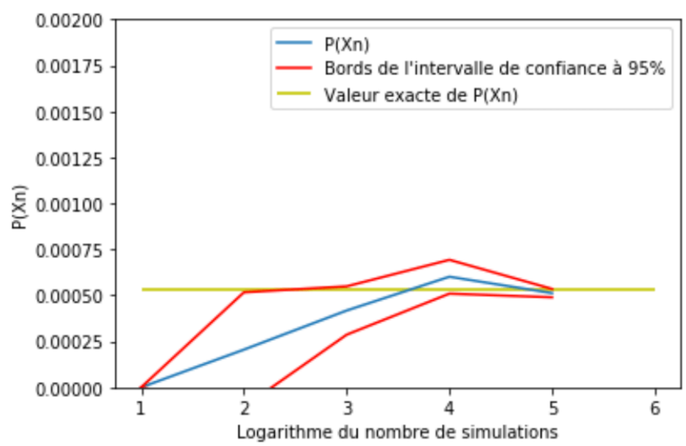


Figure 3.6: Interacting Particle Systems estimate: $P(X_{40} > 0|X_0 = 20)$ for $m = 0.8, \lambda = 0.01$ and 10^x simulations

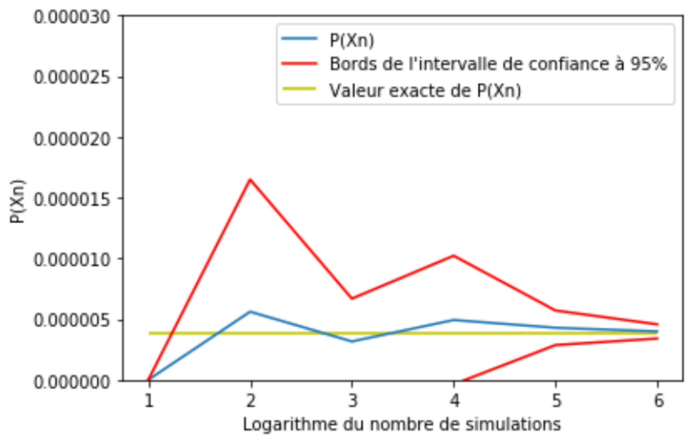


Figure 3.7: Importance Sampling estimate: $P(X_{40} > 0|X_0 = 20)$ for $m = 0.7, m' = 0.97$ and 10^x simulations

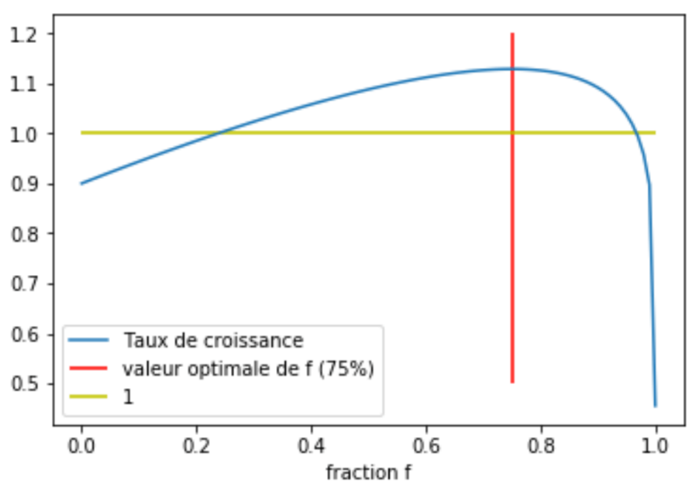


Figure 3.8: Optimal fraction f to put in the first habitat to maximise the growth rate

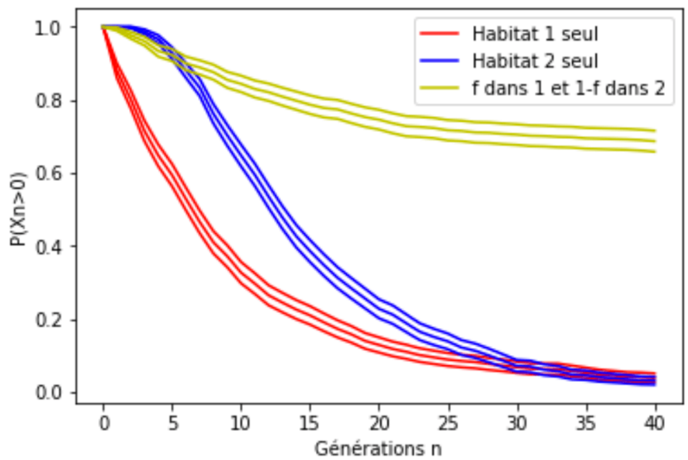


Figure 3.9: Optimal combination of 2 sink habitats where the population can persist

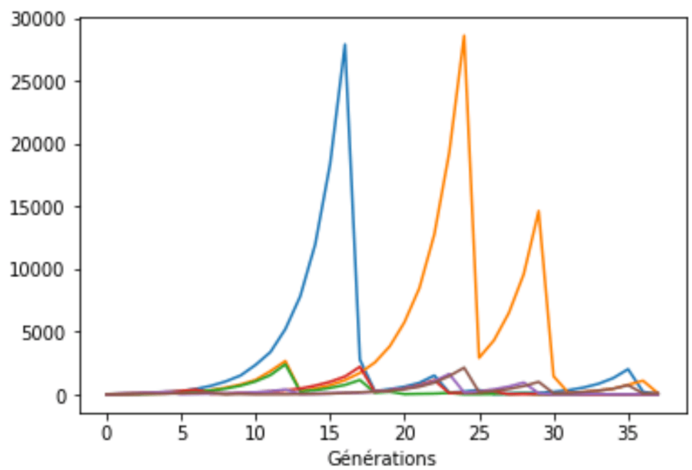


Figure 3.10: Examples of populations that have thrived in the combined habitats, the population stocks in the second habitat to avoid extinction when disasters occur in the first habitat

3.4 LIX Research project – Solving Rush Hour Game efficiently

3.4.1 Context

We have worked on a project with LIX[4] for which we developed an efficient solver for the Rush Hour Game (See Figure 3.11). The game environment consists of a bi-dimensional grid with multiple cars and the goal is to move them in order to get a preselected car out of the grid from a preselected exit. At first glance, the game seems easy, but it is not trivial to develop a universal efficient solver given that the number of outcomes is exponential with respect to the number of moves. The goal of the solver is to find optimal solutions (solutions with the minimum number of moves).
We have developed heuristics that allow to find optimal solutions using reasonable computational resources, the program we delivered has a 100% success rate outputting the optimal solution to the game. On a hard Rush Hour game, the brute-force algorithm took 15.7 hours to deliver the optimal solution while our heuristic based algorithm took only 2 minutes 39 seconds to deliver the same result.

3.4.2 Callenges and technical details

The main challenge of this project is the infinite number of possibilities of playing the game for a particular setup.
A random player solution works to solve the game but it typically takes a huge number of steps and has close to zero probability of rendering the optimal solution. Another solution is the brute-force algorithm which explores all the possibilities in an intelligent way (without re-visiting previously explored paths), this solver is not efficient and has an exponential solving time, however it works and improves on the random player solution.
We have started by implementing the brute-force algorithm from scratch as a best-practice for this type of algorithmic problems as it allows to figure out the overall architecture of the code, its representation and the data structure to use. As anticipated, the brute-force algorithm worked especially on small and easy games but it quickly became too slow to be usable in game setups with big lengths[5].
The following approach we used is based on a heuristic, its role is to prune the exploration tree for paths that we know are sub-optimal to avoid wasting computational resources exploring them further. The idea we used is to compute a function that gives a lower bound on the number of moves needed to finish the game starting from a given position. The challenge is searching a good heuristic is to be easily computable as it will be used on each sequence of moves we explore.
The heuristic we used is the number of cars that are between the main car[6] and the exit plus one for each car that cannot be immediately removed. We mathematically prove that this heuristic is correct then we implement it.
We have used Object Oriented Programming in Java where we built the whole game environment along with the graphical interface.

[4]Laboratoire d'informatique de l'École polytechnique
[5]See Curse of Dimensionality
[6]The main car is the one that needs to get out for the game to be solved

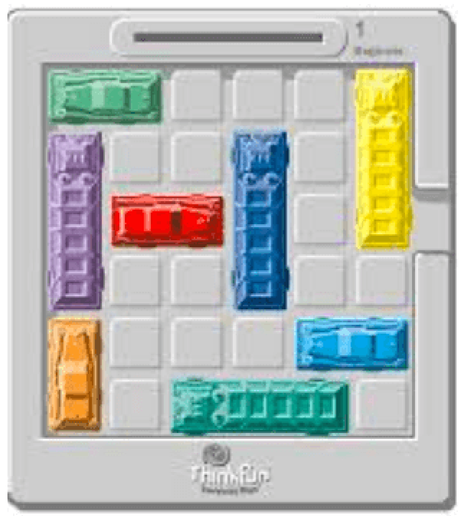


Figure 3.11: Rush Hour Game, playable in this link. The goal is to get the red car out

3.5 Machine Learning Research – Bregman Divergence

3.5.1 Context

We worked with Data61 Artificial Intelligence and Machine Learning Research Group on theorems regarding generalized Bregman Divergences (BD).

3.5.2 Challenges

BD is a family of divergences used in different domains like Information Theory and Convex Optimisation. Indeed, BD are the natural divergence family to use when it comes to parameters of probability distributions, they are used in several statistical estimation and Machine Learning approaches (online learning, clustering, class probability estimation, matrix approximation, boosting, variational inference...). The main issue is that they are usually poorly defined (e.g. they are poorly generalized to tensors and even matrices). In order to correct that, Data61 team chose us to work on a formal rigorous framework to build a reference that can be democratically used in Machine Learning research.

3.5.3 Technical details

After studying the literature, we properly generalized BD on two directions:

1. First, we properly generalized them and their three fundamental properties on tensors (See figure 3.12 for the definition)
2. Then we relaxed the differentiability assumption by defining a BD associated to convex non-differentiable functions

Furthermore, we proved a fourth fundamental property for the BD: the inversion property of a BD with a generalized perspective function. See figure 3.13 for the statement of the theorem.
As a result, our work was used by the the Machine Learning Research Group to propose applications related to a reduction from multi-class density ratio to class-probability estimation, a new adaptive projection free yet norm-enforcing dual norm mirror descent algorithm, and a reduction from clustering on flat manifolds to clustering on curved manifolds. This work is leading to a scientific paper to be published in cooperation with Google Brain research team in IEEE Transactions on Information Theory.

Let p be a natural integer, T_p the set of p-order tensors, $\mathcal{X}$ an open convex subset of T_p, and $\varphi : \mathcal{X} \to R$ a differentiable function.

Definition 3.5.1. *The* ***Bregman distortion*** *associated with φ is defined as:*

$$\forall A, B \in T_p, \quad D_\varphi(A||B) = \varphi(A) - \varphi(B) - \langle \nabla\varphi(B), (A - B)\rangle$$

With $\langle ., . \rangle$ being the canonical inner product.
We call φ the ***generator*** *of this Bregman distortion.*
If φ is convex, it is a divergence, we call it in this cas the ***Bregman divergence*** *associated with φ.*

Figure 3.12: Definition of Bregman divergence on tensors

Let $g : \mathcal{X} \to R_*$ be a differentiable function such as $\breve{\mathcal{X}} \subset \mathcal{X}$ where $\breve{\mathcal{X}} = \{(1/g(T)) \cdot T | T \in \mathcal{X}\}$. We denote for $A \in \mathcal{X}$: $\breve{A} = (1/g(A)) \cdot A$ and $\breve{\varphi}(A) = g(A) \cdot \varphi((1/g(A)) \cdot A)$.

Theorem 3.5.1. *The Scaled Bregman Theorem (SBT):*

$$\forall A, B \in \mathcal{X}, \quad g(A) \cdot D_\varphi(\breve{A}||\breve{B}) = D_{\breve{\varphi}}(A||B) + R_{\varphi,g}(A||B) \tag{3.1}$$

where $R_{\varphi,g}(A||B) = \varphi^*(\nabla\varphi(\breve{B})) \cdot D_g(A||B)$

Figure 3.13: Statement of the Scaled Bregman Theorem, the fourth fundamental theorem of BD

3.6 Télécom Paris Image Processing project – Implementing a video background change detection algorithm

We have worked on a Signal Processing with a teammate, we developed mathematical methods on Matlab to detect background changes in a video. For that, we have computed consecutive image similarities using: images correlation, Grey scale and HSV histograms distance and Fourier transforms distance.
See summary of the project in the next page.

Détection de changement de plan

Projet d'étudiants de 1ère année
Réalisé par Ilyas Malik et Grégoire Dupont.
Encadré par Michel Roux
Traitement des images, Programmation Matlab, Mathématiques Appliquées

TELECOM ParisTech

Problème Général : Comment implémenter un algorithme qui permettrait de trouver les changements de plans d'une vidéo ?

Idée Générale : Comparaison de deux frames successives d'une vidéo et déclaration d'un changement de plan si la différence est trop grande.

*Quelles méthodes implémenter pour rendre compte d'une **distance** ou d'une **ressemblance** entre deux images **?***

La corrélation

Cette méthode compare la « ressemblance » entre deux images pixel par pixel.

i) Opération de plusieurs transformations à l'image

Passage de RGB à Gris

0,3*Red + 0,6*Green + 0,1*Blue

Soustraction de la composante moyenne
Division par l'écart type

ii) Calcul du coefficient de corrélation grâce à la formule ci-dessous :

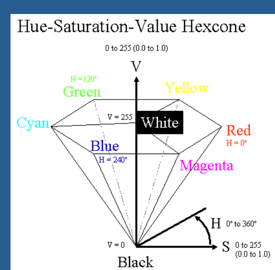
$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)}}$$

Les histogrammes de couleur

Cette méthode repose sur la comparaison des histogrammes d'images successives.

i) Passage en niveau de gris de la même manière qu'en corrélation. Construction d'un histogramme sur cette matrice (à un canal donc).

ii) Passage en coordonnées RGB à HSV.



iii) Construction d'un histogramme en HSV en prenant garde à ne compter la teinte que si la saturation est assez grande pour que la teinte soit stable.

iv) Pour chaque image on dispose donc de deux vecteurs (l'histogramme gris et l'histogramme HSV). Calcul pour chaque couples de frames successives, de la distance euclidienne entre les histogrammes de gris et les histogrammes en HSV.

L'amplitude de Fourier (fft2)

Cette méthode calcule simplement le module de la transformée de Fourier de deux images successives, puis établit les distances euclidiennes entre ces transformées de Fourier.

i) Passage en gris pour pourvoir calculer la fft2 d'un unique canal (économie de calculs).

ii) Calcul de la fft2 de l'image grise

$$TF[f(x,y)] = F(\nu_x, \nu_y) = \iint_{-\infty}^{\infty} f(x,y)\, e^{-i2\pi(\nu_x x + \nu_y y)}\, d\nu_x\, d\nu_y$$

iii) Calcul de l'amplitude de la transformée de Fourier ce qui nous renseigne sur la **présence des éléments de l'image.**

iv) Calcul de la distance euclidienne entre deux amplitudes de transformée de Fourier successives.

Cette méthode permet notamment de ne pas rendre compte des *traveling*.

Fusion des méthodes

La fusion des trois méthodes se fait **après** leur exécution. On dit que l'on **fusionne les décisions.**

Le principe est de considérer que si deux méthodes donnent le même résultat alors il est à peu près sur que c'est un bon résultat.

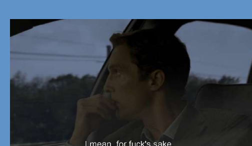
Idées d'applications

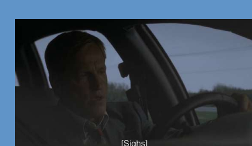
i) Fréquence des échanges pendant les débats télévisés

ii) Afficher le nom des intervenants à la TV, grâce en plus à algorithme de reconnaissance de visages

iii) Étude du rythme d'une série pour proposer des choix à l'utilisateur.

Résultats des Tests





Résultats fournis par les 3 méthodes :
- Corrélation Correlation(Im1,Im2)=0,49 Correlation(Im2,Im3)=0,94
- Amplitude de Fourier d(Im1,Im2)=19.9488 d(Im2,Im3)=8,7468

3.7 CMAP Research project – A conjecture on Random Polynomials Roots

3.7.1 Context

We worked with the department of applied mathematics of l'X on a simulation approach to analyze the roots of random polynomials depending on its degree.

3.7.2 Challenges

Real polynomials in small degrees (0,1,2, or 3) can be easily analyzed, their roots can be analytically described. However when the degree of the polynomial tends to infinity, it is not always possible to analytically describe its roots. We have to approach these kind of problems numerically. In our case, we analyze the behaviour of roots of random polynomials using advanced simulation techniques with numerical roots calculation.

3.7.3 Technical details

We started with second and third degree polynomials in order to isolate the simulation problems while computing the roots is done analytically in those cases. We used a Monte-Carlo approach for different families of random polynomials.
Then we generalized the simulations for every degree. Since we don't have an exact formula to compute the roots, we used the Kac-Rice formula. We used that method to simulate different probability laws of these polynomials: Gaussian polynomials, Kostlan-Shub-Smale polynomials, and Kac polynomials.
We were able to make an interesting observation on the roots: polynomials with degrees tending towards infinity have no or few roots outside the unit ball, with a fixed uniform finite density of probability inside the ball and an infinite density in the unit sphere. See figures 3.14 and 3.15 for Kac polynomials.
We delivered to the department an analysis report with conclusions about the surmised properties of random polynomials roots.

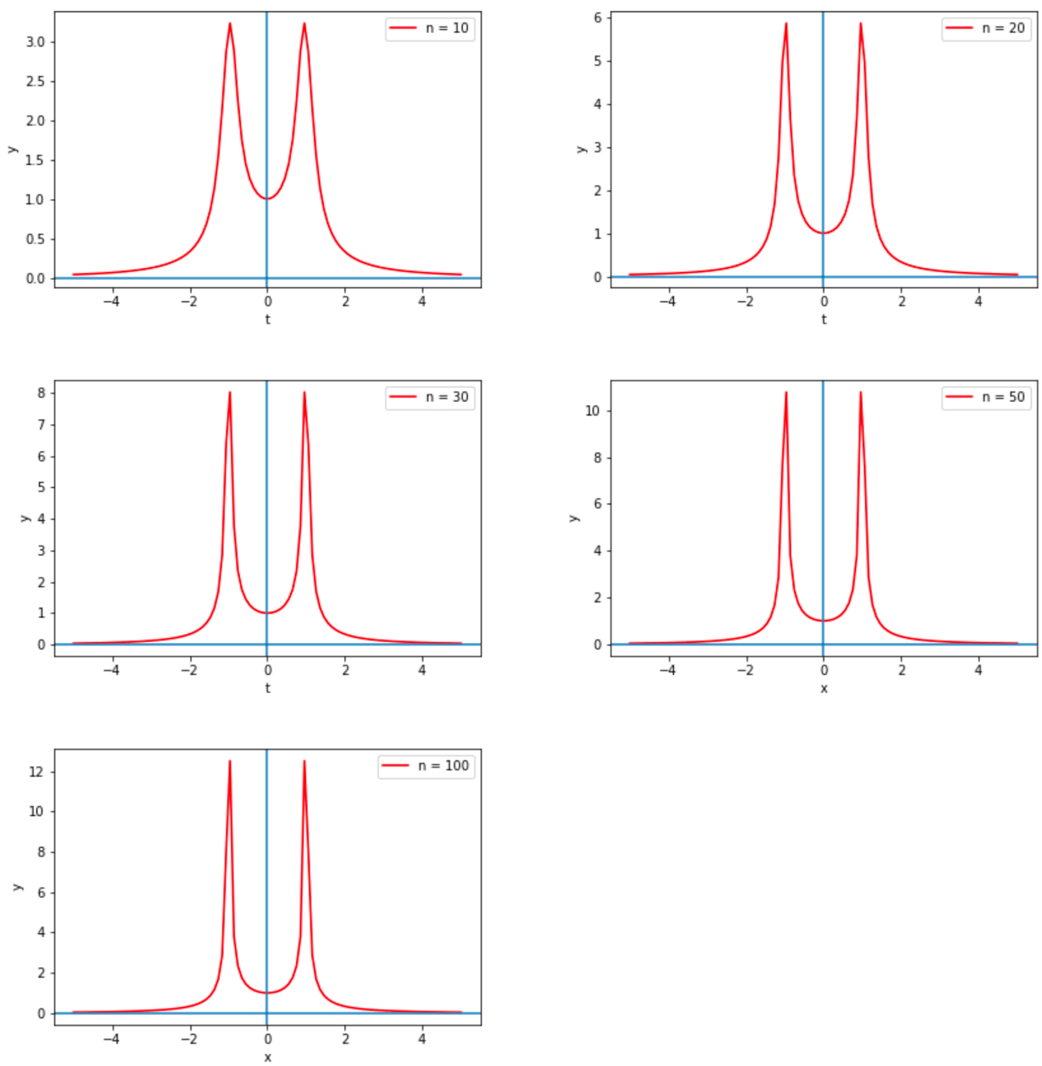


Figure 3.14: Plots of real roots of simulated Kac Polynoms

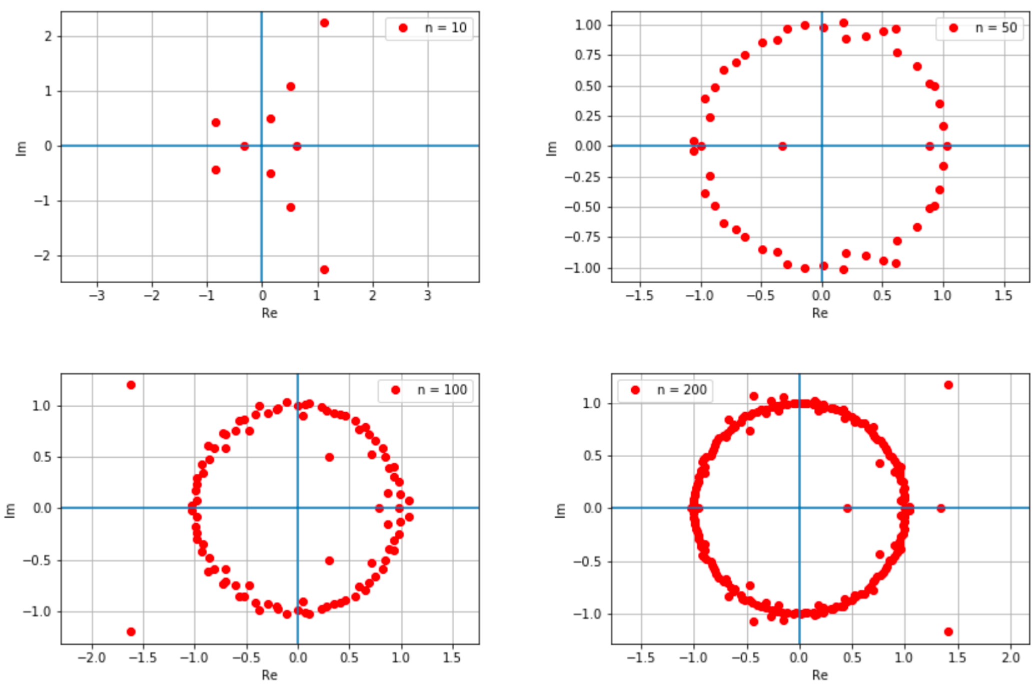


Figure 3.15: Plots of complex roots of simulated Kac Polynoms

Chapter 4

Other Freelance projects

4.1 Noor 1 Freelance project – Building a prioritization software for the CSP plants Mirrors cleaning

4.1.1 Context

We have worked with the NOMAC (the operations and maintenance subsidiary of ACWA Power International) on a project related to their Concentrated Solar Power (CSP) plants. We have built a prioritization software for the mirrors cleaning. The software is based on the drone inspection scan of the plant, this scan gives visible and thermal images of the plant to detect problems and organize operations. Our role was to use the visible images to detect the level of dirt in the mirrors in order to prioritize the cleaning operations in the Noor1 CSP plant of Ouarzazate, Morocco which have Parabolic trough mirrors.

4.1.2 Challenges and Technical details

In this kind of problematic, the standard methods are the gradient-based algorithms. Gradient-based techniques could not work in this case because the images quality wasn't high enough to allow for a robust gradient-based dirt detection algorithm. We had to be more innovative so we developed a different approach designed for this specific solar plant.
We took an optical approach developing a dirt index based on the RGB proportions of each region: the drone takes images of the mirrors when they are directly reflecting the sky (techniques for checking the images are discussed below), a clean mirror has a steady and good proportion of blue while a dirty mirror has a relatively high proportion of red because the dirt comes from the Ourzazat desert soil which color is mainly composed of red. See figure 4.1 for examples.
Hence, we developed an algorithm that automatically computes a dirt index based on this principle. The index was fine tuned based on the data we got from the drone inspections. The next step was to make the index more robust, the first amelioration was to take into account the "blueness" of the sky. The proportion of blue on the photo was compared to the proportion of the blue in the direct image of the sky during the same inspection fly using the pictures of the direct sky in order to have an index robust against the versatility of the weather. See figure 4.2 for examples of index computed for different part of mirrors.

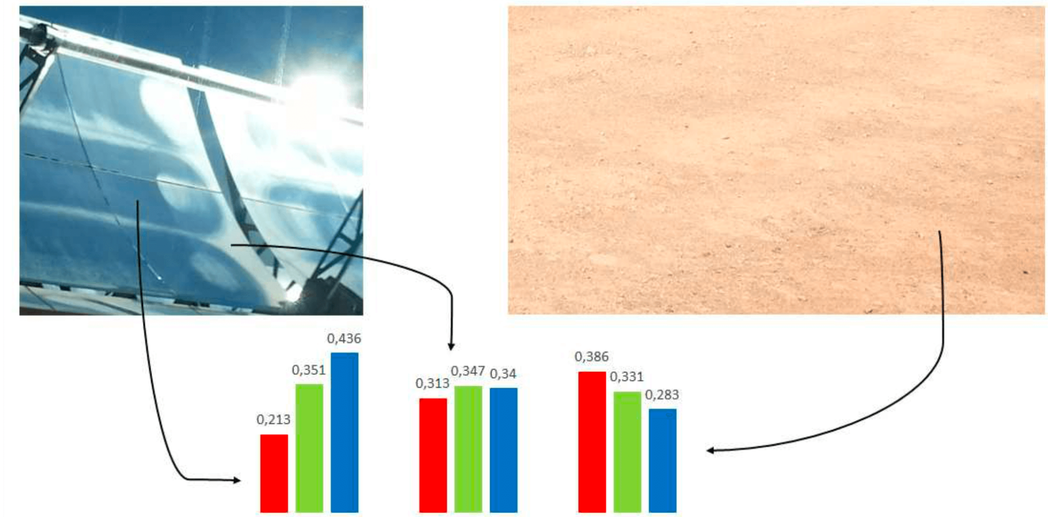


Figure 4.1: Illustration of the difference in RGB proportions between clean mirror, dirty mirror and the source of dirt: the soil

The second important amelioration was to compute with precision the region of space where the drone must be for an efficient comparison of the mirror color and the sky color. Our geometrical analysis showed that there is indeed a sweet spot where the index is the most robust. See figure 4.3. Therefore the software chooses the images corresponding to the sweet spot for each mirror for each mirror, the dirt index is then computed on those images using the image of the direct sky. Then, we end up with a global index for each group of mirrors giving the cleaning team a prioritization for their operations and thus greatly reducing the cost for those operations.
The last step was to check the robustness of our algorithm by comparing the index of groups to the temperature of the fluid the mirrors were heating: the dirt decreases the performance of the mirrors, making them heating the fluid less. So if our index captured well the dirt level of mirror groups, we should find a negative correlation between the dirt level of each group and the difference of temperature of the fluid between the input and the output of each group.
To that end we conducted experiments for which we were able to compute a major negative correlation and thus proved the robustness of the algorithm.

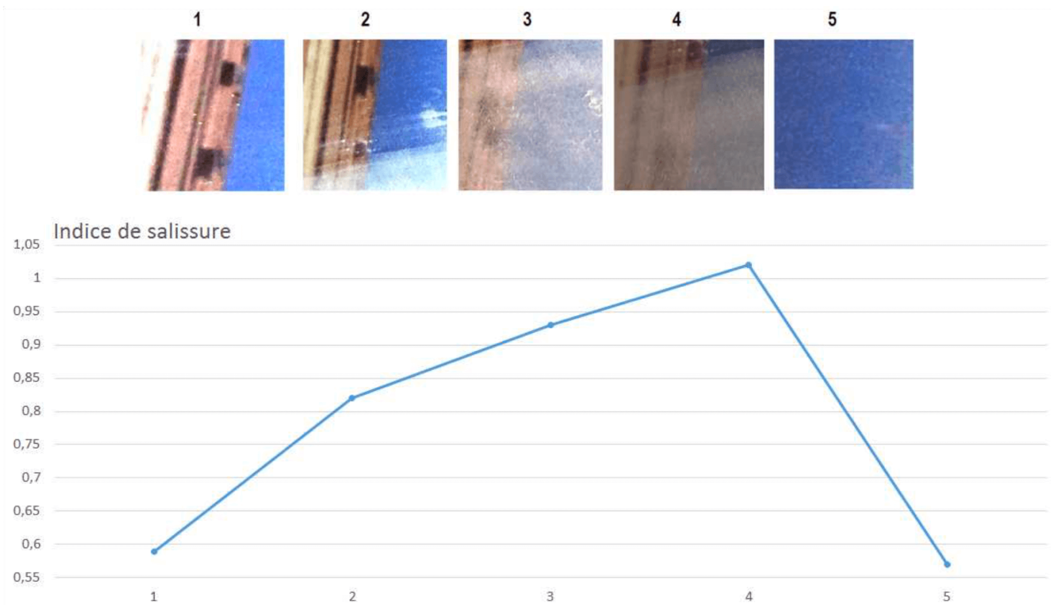


Figure 4.2: Illustration of the dirt index computed on different parts of the mirror with different levels of dirtiness

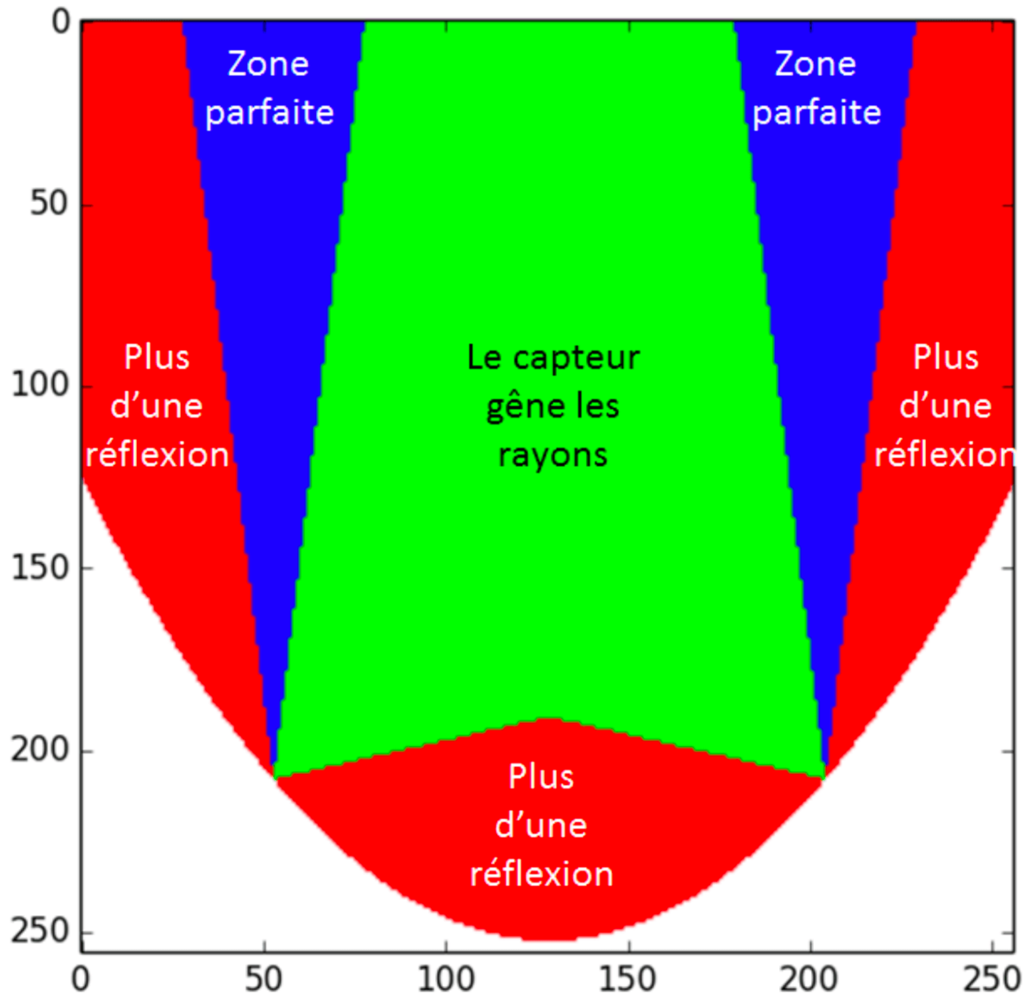


Figure 4.3: Geometrical analyses on the position of the photo

In the red region, the light detected is reflected more than once which influences the index. In the green region, the drone interferes with the light captured. The blue region is the sweet spot where the index is robust.

4.2 Télécom Paris project – Building an Android App for learning Chinese

We have worked on a team project for building an Android App that teaches Chinese from scratch.

- We built a Speech recognition algorithm based on signal processing: Dynamic Time Warping for aligning the input and reference sounds then computing the Mel-Frequency Cepstral Coefficients for their similarity
- We built a character recognition algorithm based on preprocessing the input and computing a similarity score with the reference character
- We implemented the Leitner method for a pedagogic heuristic of presenting the Chinese words

See summary of the project in the next page.

TELECOM ParisTech

Didactibic

Pact 24

DUBARRY-COUCOU Sofyan
LAVIE Emilien
MALIK Ilyas
MATHIEU Axel
MEN Hongling
ROBERT Vincent

Apprendre les idéogrammes chinois efficacement

Apprendre le tracé et la prononciation de nouveaux caractères

S'exercer à son rythme sur les caractères déjà rencontrés

Progresser

Reconnaissance de parole

Vérifier la connaissance de la prononciation grâce à la MFCC et DTW

Thèmes personnalisés

L'utilisateur personnalise son apprentissage en complément d'un cours classique

Une application développeur annexe permet de développer la base de données

Reconnaissance d'écriture

子

Vérifier la connaissance du tracé des caractères grâce à un algorithme de comparaison à une référence

L'ordre et le sens du tracé sont analysés pour valider chaque caractère

Entretiens Semi-Directifs

Cerner le besoin de l'utilisateur

Lui demander son avis sur différentes solutions

Comprendre les méthodes utilisées pour les réexploiter

Application Android

Conçue pour Tablette et Stylet

Création de widgets personnalisés

Simulation de bus d'événements

Programmation selon le modèle MVC

Pédagogie adaptée

La répétition espacée permet d'apprendre plus efficacement

La méthode s'adapte à l'utilisateur

Le système Leitner est appliqué aux évaluations de tracé et de prononciation

4.3 BNP project – Adapting the SWAP pricing model

4.3.1 Context

We worked with BNP Paribas Securities Japan Global Market Quantitative team on adapting the SWAP pricing model in a evolving environment since the LIBOR -which was the major benchmark used in pricing these securities- was going through some major changes and is no more the central benchmark to price with these securities.
The environment of global market was changing, the LIBOR -which was a major benchmark central in a lot of securities pricing model- was evolving and could no more be used as the cental interest-rate in the pricing systems. All the pricing systems depending on the LIBOR had to be adapted to this change. We worked on the SWAP pricing system in the Japanese branch of BNP Paribas.

4.3.2 Challenges and Technical details

The SWAP is a contract of exchange of two cashflows typically a fixed interest rate one against a floating one. The floating rate can typically be the LIBOR which shows why it plays a central role in these securities.
Our first task was to replace the LIBOR with a group of local and global benchmarks. It was a challenging task because the ongoing system was hard coded for the LIBOR benchmark, so we had to re-code it to make it easily configurable before replacing the LIBOR. This change increased the dimension of the system since one benchmark was replaced by multiple ones, this made the models more complex and slower. After discussing this matter with the quantitative team, we optimized it using a Principle Component Analysis (PCA) it order to reduce its dimension.
At the end, we delivered the new SWAP pricer on time as specified and it was integrated to the new version of the global market pricing system.

Appendices

Appendix A

Organizations we collaborate with

During our projects we had the privilege to work with international renown organizations, we are proud to have shined every single time within those organizations, living up to the high standards and committing to the best practices of those organizations, we are grateful for those experiences as they allowed us to learn a lot and build a robust professional expertise.
We were able to bond with many professional parties and we have left our print by the contribution we made each time either in research or from the business perspective, and often in both. We are thankful for the resources and the network those organizations offer us with open arms.
See Figure A.1.

Figure A.1: List of collaborators

Appendix B

Contact us

Don't hesitate to contact us to have an initial discussion of your problematic or your project.

E-mail 1 incidiumtechconsulting@gmail.com

E-mail 2 ilyas.malik@polytechnique.org

E-mail 3 ahmed.kriouile@polytechnique.org

Phone 1 00 212 681 899434

Phone 2 00 212 659 744729