# Risk versus Return: Betting optimally on the Soccer World Cup using Robust Optimization 

Virgile Galle and Ludovica Rizzo


#### Abstract

The sports betting industry is estimated to be worth between $\$ 700$ billion and $\$ 1$ trillion a year internationally. Approximately $70 \%$ of this business comes from soccer betting. It is important to note that the Soccer World Cup is one of the most lucrative events of the sports betting industry. For every single game, it is possible to bet on a multitude of events: the final score (as one would expect) as well as the time of the first goal, the number of yellow cards, whether a certain player scores, etc. Current state-of-the art betting approaches, currently rely on stochastic optimization techniques. These approaches are criticized for unnecessary volatility and inability to handle a large number of simultaneous events that currently take place during the World Cup. In this work, we propose a novel risk management approach to sports betting that handles risk based on Robust Optimization. A key advantage is that this approach can be scaled to the large number of events that are present in the World Cup while reducing the risk exposure.

Our effective betting uses three components: 1. A machine learning model that predicts the probability of an outcome 2. The bookmaker's odds for that outcome 3. An optimization algorithm that compares the two numbers above and decides how much to bet

We first create a data-driven predictive model for the outcome of World Cup games. Our model is built based on logistic regression and data from the last 5 World Cups such as the number of consecutive participations to the tournament, the qualifying zone and whether a team is playing in its home continent. On the 2014 World Cup, the model has a notable accuracy of $68 \%$ in predicting the group winners and $50 \%$ on the outcome of group stage games. These numbers clearly outperform the corresponding accuracies of $62 \%$ and $33 \%$ of Goldman Sachs' prediction team.

Subsequently, we develop a Robust Optimization model for the betting problem. Robust Optimization is a modern decision-making framework that has been widely used in a variety of settings including portfolio allocations. With a large number of simultaneous bets, we use the overall accuracy of the predictive model to optimize the worst case gain. This provides a conservative approach that guarantees a minimum gain with high probability. Applied to the 2014 World Cup, our betting strategy generated a return of $51 \%$. Simulation experiments demonstrate that this strategy reduces risk exposure by a factor of 7 compared to a widely used stochastic optimization approach. Our analysis leads us to believe that Robust Optimization is a novel and valuable approach to modeling betting problems on a large number of simultaneous events. Combined with a good prediction model, this betting strategy can guarantee significant gains while reducing the risk.


## 1 Introduction and Literature Review

This paper considers the problem of betting on a large number of events with fixed odds, as during the Soccer World Cup. A typical setting is the following : a player identifies a set of $N$ events (soccer games here) on which he wants to bet. He has a fixed budget $B$, and knows the bookmakers' odds for every event (represented as the return on a unit stake and denoted $c_{i}$ ). Let us also assume that the player has a statistical model that gives good estimates on the probability of each event $\left(p_{i}\right)$. Comparing these probabilities $p_{i}$ with the odds $c_{i}$, the player needs to decide his "betting strategy", ie how to allocate his budget among the different games.

Betting has been widely studied in the Stochastic Optimization literature ( [1], [2], [3] and [4]). It has widespread applications such as portfolio optimization, gambling games or investment strategies. However most of the classical approaches do not apply to the problem of betting on a large number of simultaneous events. A very common assertion in the literature is that the customer should bet only on events where his expected gain is positive meaning when he has an edge over the bookmakers (denoted later $p_{i} c_{i}>1$ ). Efficient predictive models should be able to create this edge. However, this is not always possible depending on the event that the player bets on. We show that theoretically, this condition might not be necessary.

Betting on multiple simultaneous events has only been studied recently. The first step (as explained in [1]) is to observe that maximizing the expected revenue is an extremely risky approach. In fact, in this case, the optimal strategy is to allocate the entire budget to the event with the highest expected gain. Using this approach the player doesn't diversify his risk and, if a repeated betting game is considered, will be ruined almost surely. After this step, it is common (see [1], [2], [3]) to consider a risk-averse utility function and to look for a strategy that maximizes this expected utility. The most common utility function is the logarithm leading to the Kelly criterion ( [3], [4]) that is broadly used in blackjack for example. However, the Kelly criterion can only be used in the case of sequential events. In the case of simultaneous events, [1] maximizes the expected gain using a stochastic gradient-descent approach. This approach is computationally expensive and cannot be applied to a large number of events. Furthermore, as noticed in [1], assuming a utility function can be considered as unrealistic (there is no 'right shape' for the utility, it is hard to quantify 'riskaversion' in a betting setting and different utility functions can lead to very different strategies). To tackle these problems, we introduce here a new approach to define a risk-averse betting strategy using Robust Optimization.

Robust Optimization (RO) has become more and more popular in the last 10 years and has been applied to many fields. It tackles the problem of optimization under uncertain data. Let's consider a player that faces some uncertainty (a set of events) and wants to define a betting strategy that maximizes his worst case revenue. This 'worst case situation' can be modeled by a game between the player and an opponent 'Nature' who decides the outcome of every event: if the player chooses the betting strategy $x$ then 'Nature' will choose the outcomes that minimize the revenue of strategy $x$. The player can anticipate Nature's reaction and thus choose the strategy $x$ that maximizes his worst case revenue. The main assumption of Robust Optimization is that Nature's decisions are constrained (for instance the player's predictions cannot always be wrong), they belong to what is called an 'uncertainty set' ([7]). The underlying idea is that the uncertainty (in our application, being right or wrong on the outcome of every event) can be controlled. This is represented by a 'budget of uncertainty' accorded to Nature. Optimizing the worst case scenario guarantees a lower bound for the players' earnings for every realization of Nature in the uncertainty set. To the authors' knowledge, only convex uncertainty sets have been studied, and in this problem discrete
uncertainty sets emerge. The authors propose techniques from Integer Programming and a cutting plane algorithm to tackle this point.

The aim of this new approach is twofold: it provides an alternative to the choice of an arbitrary utility function while building a risk averse betting strategy. Indeed, it is observed in [1] and [3] that most of the solutions proposed so far have high volatility. We propose an approach that guarantees a certain gain (or a maximum loss) with a controlled risk as the uncertainty sets parameters, hence the risk, are chosen by the player.

Section 2 formulates the betting problem as a robust linear optimization problem. Section 3 introduces two types of integral uncertainty sets and studies their structure, assets and drawbacks. Section 4 derives simulations to understand the behavior of the proposed strategies and compare it with previous existing solutions. Section 5 gives betting predictions for the 2014 FIFA World Cup in Brazil using this approach. Finally Section 6 gives some concluding remarks.

## 2 Robust Formulation

In this section, we define the variables and parameters of our problem and formulate the robust betting problem. We have a set of $N$ games on which a player wants to bet simultaneously. The player has a budget $B$ and he has to allocate it between the different events. We denote $\left(x_{1}, \ldots, x_{N}\right) \geq 0$ the decisions variables: $x_{i}$ is the amount bet on game $i$. The budget constraint can be written as $\sum_{i=1}^{N} x_{i} \leq B$. Hence we define the polyhedron $P$ of feasible strategies as

$$
\begin{equation*}
P=\left\{x \in \mathbb{R}_{+}^{N}: \sum_{i=1}^{N} x_{i} \leq B\right\} \tag{1}
\end{equation*}
$$

In previous papers, a risk threshold is sometimes added in the constraints. In this case, the following constraint is added to the definition of $P: x_{i} \leq Q, \forall i \in 1, \ldots, N$ The risk threshold $Q$ is an upper bound on every coordinate of $x$, meaning that the player cannot bet more than a certain amount on a every game, in order not to be too risk seeking. However we will see that the robust approach is conservative enough that this kind of constraint is not required.

First, notice that we consider outcomes of events that are binary. In portfolio allocations problems, for example, the return of an asset can take values on a continuous interval. Here, the player can just be "right" or "wrong" on his predictions, the return can take only two values. We assume that the player has access to a priori probabilities for every event (given by a predictive model or his own beliefs). Let us denote these probabilities by ( $p_{1}, \ldots, p_{N}$ ) and the bookmaker's odds by $\left(c_{1}, \ldots, c_{N}\right)$. In this paper we use the European odds system. It is defined as follows: if the player bets $x_{i}$ on game $i$, then there are two possible outcomes: if the player is correct, then he earns $\left(c_{i}-1\right) x_{i}$, otherwise the players loses his money and his gain is $-x_{i}$. Consequently, we have $Z_{i}$ the random variable corresponding to the profit in game i.

$$
Z_{i}=\left\{\begin{array}{l}
\left(c_{i}-1\right) x_{i} \text { if the player guessed right }  \tag{2}\\
-x_{i} \text { if not }
\end{array}\right.
$$

and let $Z$ be the total gain over the N games: $Z=\sum_{i=1}^{N} Z_{i}$. Usually, the goal is to maximize the expected value of a certain utility function U of the profit on a set of $N$ games i.e the problem is

$$
\begin{equation*}
\underset{x \in P}{\operatorname{maximize}}(\mathbb{E}[U(Z)]) \tag{3}
\end{equation*}
$$

Our approach is the following. To the player's point of view, the outcome are obviously unknown. We model this unknown information by binary variables $y_{1}, \ldots, y_{N}$ defined as follows:

$$
y_{i}=\left\{\begin{array}{l}
1 \text { if the player guessed right for game } i  \tag{4}\\
0 \text { otherwise }
\end{array}\right.
$$

These variables represent nature's decisions. Let us call $\Xi$ the uncertainty budget for nature $(y \in \Xi)$. The robust formulation of our problem is

$$
\begin{equation*}
z_{\text {rob }}=\underset{x \in P}{\operatorname{maximize}}\left(\underset{y \in \Xi}{\operatorname{minimize}}\left(\sum_{i=1}^{N} c_{i} x_{i} y_{i}-x_{i}\right)\right) \tag{5}
\end{equation*}
$$

The intuition behind this formulation is the following: the player wants to maximize his gain using the decision variable $x$, while nature tries to minimize the player's gain using the variable $y$ (worst case realization in the uncertainty set). Let us make a key observation: $z_{\text {rob }}>0$ means that there exists a betting strategy $x^{*} \neq 0$ such that, for every realization of $y$ in the uncertainty set, the gain will be positive. Thus, if $z_{\text {rob }}>0$ and the player is confident that the realization of $y$ will be in the uncertainty set, he should bet.

In the following section we will propose two approaches to build uncertainty sets, we will solve the corresponding robust formulations and provide probabilistic guarantees for the uncertainty set (lower bounds for $\mathbb{P}(y \in \Xi)$ ). We will point out the trade off between optimization ('high' value of $z_{\text {rob }}$ ) and robustness ('high' lower bound for $\mathbb{P}(y \in \Xi)$ ).

## 3 Two approaches

### 3.1 The 'Probabilistic Confident' uncertainty set

The first type of uncertainty set we consider is the 'Probabilistic Confident' case where the player is confident in the probabilities he uses (for example if the player predicts two different games where $p=85 \%$ and another where $p=90 \%$, then he should be right significantly more on the second one). The Probabilistic Confident uncertainty set (PC) is defined in the following way:

$$
\begin{equation*}
\Xi_{P C}=\left\{y \in\{0,1\}^{N}: \sum_{i=1}^{N} y_{i} \geq \Delta, \sum_{i=1}^{N} p_{i} y_{i} \geq \Omega\right\} \tag{6}
\end{equation*}
$$

The intuition behind this uncertainty is the following: To penalize the player the most, nature should try to have $y_{i}=0, \forall i \in 1, \ldots, N$. However, the model has a certain 'worst case' accuracy on the test sets, hence nature should not be able to make the player fail more than this accuracy which gives us the first constraint in $\Xi_{P C}$. In other words, if $\alpha$ is the 'worst case' accuracy of our model take $\Delta=\lfloor\alpha N\rfloor$, one can interpret $N-\Delta$ as a budget for nature to make the player fail. Notice that the first budget constraint is independent of $\left(p_{i}\right)$. We incorporate $\left(p_{i}\right)$ in the second inequality. The weighted sum of $p_{i} y_{i}$ gives different importance to the events according to their probability. Intuitively it translates the fact that it is hard for nature to make a player fail on a event where its predicted probability is very high. If the parameters $\Delta$ and $\Omega$ are well chosen, the nature won't be allowed to reverse the player's predictions on all the events where he has a high probability of success.

The major challenge here is the discrete character of the uncertainty sets that we are considering. Most of the cases treated so far in Robust Optimization have convex sets. Two approaches are possible: we can solve the robust formulation with integrality constraints using integer programming or relax the integrality constraints and solve a linear program to get a lower bound on the optimal cost.

Since $p$ is not an integer vector, here these two approaches are not equivalent, the relaxation provides a lower bound to the robust integral problem. This is illustrated in the following example.

Example For $\mathrm{N}=2$, take $\Delta=1, c_{1} x_{1}>c_{2} x_{2}$ and $p_{1}>p_{2}$ such that $p_{1}+p_{2}=1$ and $\Omega=\frac{p_{1}+p_{2}}{2}$. Hence we can notice that the optimal solution is clearly $y_{1}=1$ and $y_{2}=0$, hence $z_{I P}=c_{1} x_{1}$ but if we relax the integrality and consider the LP, one can prove that $y_{1}=y_{2}=0.5$ is a feasible solution with objective function $z=\frac{1}{2}\left(c_{1} x_{1}+c_{2} x_{2}\right)<z_{I P}$. Hence relaxing integrality constraint has a cost in getting the optimal solution.

The cutting plane algorithm (described in Appendix A and presented in [8]) solves efficiently large scale IP Robust by adding cuts as we generate solutions. It appears to be efficient and give the exact solution to our problem. However there is no guarantee on the number of steps needed to find the optimal solution. Knowing that at each step we solve a binary IP, it might be hard to solve this problem for very large N .

In that case, we might not have access to the actual robust solution but giving bounds on the optimal cost is also valuable. So let's consider the relaxation of our subproblem, given a vector $x \in P, \underset{y \in R_{P C}}{\operatorname{minimize}}\left(\sum_{i=1}^{N} c_{i} x_{i} y_{i}\right)$ where $R_{P C}$ is the polyhedra obtained by relaxing the integrality constraints on $\Xi_{P C}$. We have the following:

Lemma 3.1. All the extreme points of $R_{P C}$ have at most two non integral components
Let us consider the relaxation problem

$$
\begin{equation*}
z_{\text {relax }}=\underset{x \in P}{\operatorname{maximize}}\left(\operatorname{minimize}_{y \in R_{P C}}\left(\sum_{i=1}^{N} c_{i} x_{i} y_{i}-x_{i}\right)\right) \tag{7}
\end{equation*}
$$

Proposition 3.1. We have

$$
\begin{equation*}
z_{\text {relax }} \leq z_{\text {rob }} \leq z_{\text {relax }}+M \tag{8}
\end{equation*}
$$

where $M=\max _{i, j=1, \ldots, N,}\left(c_{i} x_{i}^{*}+c_{j} x_{j}^{*}\right)$, and $x^{*}$ is the solution of the relaxed problem.
Hence solving the relaxation gives an approximation on the robust optimal cost and moreover if $z_{\text {relax }} \geq 0$ then $z_{\text {rob }} \geq 0$ which insures positive gain within the uncertainty set. To solve the relaxation problem, see Appendix B which uses standard robust techniques to reformulate the problem.

### 3.2 The 'Buckets' uncertainty set

In the previous robust formulation we considered the following budget of uncertainty constraint: $\sum_{i=1}^{N} p_{i} y_{i} \geq \Omega$. This constraint means that it costs more to nature to set $y_{i}=0$ for a game with
a really high $p_{i}$. In other words, if the player is confident in the outcome of a game, it is more costly for nature to reverse the player's prediction. This constraint assumes that the player is really confident about his a priori probabilities. But in most of the cases the a priori probabilities are learned using a model and thus come with a confidence interval, or even worst in sport betting they can come from the player's own 'believes'. In this cases, the previous formulation can be too precise and can lead to errors. The authors propose here another approach well suited in these cases.

Let $K \in \mathbb{N}$ and let $I_{1}, \ldots, I_{K}$ be interval subsets of $[0,1]$ not necessarily disjoint. Let $G_{k}=\{i \in$ $\left.[1, N] \mid p_{i} \in I_{k}\right\}$. We can define an uncertainty set in the following way:

$$
\begin{equation*}
\Xi_{b}=\left\{y \in\{0,1\}^{N} \mid \sum_{i=1}^{N} y_{i} \geq \Delta, \sum_{i \in G_{k}} y_{i} \geq \Gamma_{k} \forall k\right\} \tag{9}
\end{equation*}
$$

Intuitively, the probabilities are split into 'buckets' of similar probability and each bucket has a separate budget of uncertainty (above which we add a general budget as in the previous approach). For example, we can set $I_{1}=[0.5,0.75]$ and $I_{2}=[0.75,1]$ and every game where the player's estimated $p_{i}$ is in $I_{1}$ will be in $G_{1}$, every game where $p_{i}>0.75$ will be in $G_{2}$. As for games in $G_{2}$ the player is more confident in the outcome, then $\frac{\Gamma_{1}}{\left|G_{1}\right|}<\frac{\Gamma_{2}}{\left|G_{2}\right|}$ where $\left|G_{i}\right|$ is the cardinal of the set $G_{i}$. This approach has two main advantages. First of all, it doesn't ask the player to give exact values for the a priori probabilities but it is flexible to confidence intervals. Second of all, from a computational point of view, we show, due to the next property, that it is easier to solve optimally than the PC case.

Proposition 3.2. If $\Delta$ and $\left(\Gamma_{k}\right)_{k=1 \ldots K}$ are integers and $I_{k}$ are intervals, then all extreme points of $\Xi_{b}$ are integer.

The first assumption can always be verified as replacing $\Delta$ (or $\Gamma_{k}$ ) by $\lfloor\Delta\rfloor$ (or $\left\lfloor\Gamma_{k}\right\rfloor$ ) does not change the uncertainty set. The second assumption comes from the a priori probabilities and their confidence intervals. From now on, both assumptions are supposed to be verified.

This proposition is useful because the integer constraint in $\Xi_{b}$ can be dropped and $z_{\text {relax }}=$ $z_{\text {rob }}$. Let $R_{b}$ be the relaxation of $\Xi_{b}$. Again, using standard robust techniques relying on duality (Appendix D), the robust counterpart can be derived and it is efficiently solvable.

## Insight on special cases

In this paragraph we will consider two simpler cases. This will help us understanding the advantages and drawbacks of a robust betting strategy.

First of all, let us consider the case where there is just one bucket. Then $\sum_{i=1}^{N} y_{i} \geq \Delta$ is the only constraint in the uncertainty set.

Proposition 3.3. Let $\Delta \in \mathbb{N}$, and the Optimal Objective value has the two following properties

$$
\begin{equation*}
z_{\text {single bucket }} \geq B\left(\frac{\Delta}{\sum_{i=1}^{N} 1 / c_{i}}-1\right) \tag{10}
\end{equation*}
$$

This case also occurs in the PC case when $\sum_{i=1}^{\Delta} p_{(i)}>\Omega$ where $p(i)$ is the i-th smallest probability. One can notice that either the player do not bet anything or the robust gain increases at
least linearly in the budget and in the ratio between $\Delta / N$ and $\frac{1}{N} \sum_{i=1}^{N} 1 / c_{i}$. This ratio, called the 'edge ratio' represents the overall edge of the player on the bookmakers, which is different from the game to game edge normally assumed in the literature. The results shows that a positive game by game ratio is better but not a necessary assumption in the study of betting. It is important to understand that $\Delta$ represents two features of the probabilities: $\Delta$ is large not only when the player has an edge but also when he is confident about this edge.

Let us consider a second case where the buckets form a partition of $\left[p_{\min }, p_{\max }\right]$ and there is no global budget $\Delta$. We have a similar result:

## Proposition 3.4.

$$
\begin{equation*}
z_{\text {partition }} \geq B\left(\frac{1}{K} \sum_{k=1}^{K} \frac{\Gamma_{k}}{\sum_{j \in G_{k}}^{N} 1 / c_{j}}-1\right) \tag{11}
\end{equation*}
$$

Notice that, this case also corresponds to take $\Delta$ such as $\sum_{k=1}^{K} \Gamma_{k}>\Delta$. Here the important value is the average of the edge ratios for each bucket which make sense since it represents the overall edge of the player.

## Probabilistic guarantees

In the paragraphs above we proposed two ways to build uncertainty sets based on some parameters ( $\Delta$ and $\Omega$ ) that intuitively represent the accuracy of the predictive model.

Let us assume that our model is correct, i.e that $p_{i}=\mathbb{P}\left(y_{i}=1\right)$ and let us assume that the games are independent (the outcome of a game cannot influence the outcome if another, this is true if the games are simultaneous).

## Lemma 3.2.

$$
\begin{equation*}
\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right) \simeq \Phi\left(\frac{\Delta-\sum_{i=1}^{N} p_{i}}{\sqrt{\sum_{i=1}^{N} p_{i}\left(1-p_{i}\right)}}\right) \tag{12}
\end{equation*}
$$

where $\Phi$ is the cumulative distribution function of a standard normal random variable.
Notice two things: first, we have the same property within each bucket by replacing the whole set of N games by the bucket. Second, we can have the same type of result for the second inequality of $\Xi_{P C}$.

$$
\begin{equation*}
\mathbb{P}\left(\sum_{i=1}^{N} p_{i} y_{i}<\Omega\right) \simeq \Phi\left(\frac{\Omega-\sum_{i=1}^{N} p_{i}^{2}}{\sqrt{\sum_{i=1}^{N} p_{i}^{3}\left(1-p_{i}\right)}}\right) \tag{13}
\end{equation*}
$$

And finally we can apply the following property to get probabilistic guarantees.

## Proposition 3.5.

$$
\begin{equation*}
\mathbb{P}\left(y \in \Xi_{P C}\right) \geq\left(1-\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right)\right) \times\left(1-\mathbb{P}\left(\sum_{i=1}^{N} p_{i} y_{i}<\Omega\right)\right) \tag{14}
\end{equation*}
$$

Similarly, if the buckets don't overlap, we have

$$
\begin{equation*}
\mathbb{P}\left(y \in \Xi_{b}\right) \geq\left(1-\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right)\right) \times \prod_{k=1}^{K}\left(1-\mathbb{P}\left(\sum_{i \in G_{k}} y_{i}<\Gamma_{k}\right)\right) \tag{15}
\end{equation*}
$$

## 4 Simulations

This section has two main goals. First, we want to illustrate the trade off between 'high worst case return' and probabilistic guarantee and thus provide an approach to select the appropriate values for the parameters $(\Delta, \Gamma, \Omega)$. Secondly, we want to compare the performances of robust strategies to other stochastic optimization approaches.

To achieve these goals, we use simulated data from [1]. We have a set of 37 events where the player always has an edge on the bookmaker. The details are in table 1. For our simulations we assume that the player's probabilities are 'right' (ie $p_{i}=\mathbb{P}\left(y_{i}=1\right)$ ).

| Event | Actual <br> Probability | Bookmaker's <br> Probability | Odds | Event | Actual <br> Probability | Bookmaker's <br> Probability |  |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.128 | 0.125 | 8 | 20 | 0.421 | 0.347 | 2.88 |
| 2 | 0.167 | 0.133 | 7.5 | 21 | 0.422 | 0.385 | 2.6 |
| 3 | 0.197 | 0.167 | 6 | 22 | 0.427 | 0.382 | 2.62 |
| 4 | 0.227 | 0.182 | 5.5 | 23 | 0.449 | 0.42 | 2.38 |
| 5 | 0.232 | 0.23 | 4.35 | 24 | 0.49 | 0.476 | 2.1 |
| 6 | 0.245 | 0.192 | 5.2 | 25 | 0.491 | 0.435 | 2.3 |
| 7 | 0.256 | 0.231 | 4.33 | 26 | 0.506 | 0.465 | 2.15 |
| 8 | 0.266 | 0.263 | 3.8 | 27 | 0.514 | 0.513 | 1.95 |
| 9 | 0.267 | 0.23 | 4.35 | 28 | 0.55 | 0.5 | 2 |
| 10 | 0.27 | 0.25 | 4 | 29 | 0.58 | 0.556 | 1.8 |
| 11 | 0.283 | 0.25 | 4 | 30 | 0.585 | 0.55 | 1.82 |
| 12 | 0.286 | 0.263 | 3.8 | 31 | 0.59 | 0.513 | 1.95 |
| 13 | 0.288 | 0.286 | 3.5 | 32 | 0.595 | 0.571 | 1.75 |
| 14 | 0.3 | 0.294 | 3.4 | 33 | 0.62 | 0.571 | 1.75 |
| 15 | 0.31 | 0.267 | 3.75 | 34 | 0.638 | 0.588 | 1.7 |
| 16 | 0.316 | 0.303 | 3.3 | 35 | 0.67 | 0.625 | 1.6 |
| 17 | 0.323 | 0.286 | 3.5 | 36 | 0.7 | 0.667 | 1.5 |
| 18 | 0.325 | 0.3058 | 3.25 | 37 | 0.778 | 0.769 | 1.3 |
| 19 | 0.388 | 0.385 | 2.6 |  |  |  |  |

Table 1: Dataset 2

Selecting robust strategies' parameters We want to provide a method for selecting the uncertainty sets' parameters. For different values of these parameters we compute the worst case gain and the probabilistic guarantee. To select the right range of parameters, we first notice that $\Delta^{0}=\sum_{i=1}^{N} p_{i}$ corresponds to the average accuracy of the model. For the PC case, we define $\Omega^{0}=\sum_{i=1}^{N} p_{i}^{2}$. We then take several values of $\Gamma, \Omega$ around these baseline values: we analyze the robust strategies for $\Delta \in\left[\left\lfloor 0.8 \Delta^{0}\right\rfloor ;\left\lfloor 1.2 \Delta^{0}\right\rfloor\right]$ and $\Omega \in\left[\left\lfloor 0.8 \Omega^{0}\right\rfloor ;\left\lfloor 1.2 \Omega^{0}\right\rfloor\right]$.

Important Remark: The budget constraint is set to equality so that even if $z_{\text {rob }}<0$, we can study the case.


Figure 1: Trade off worst case gain/probabilistic guarantee

In Figure 1, we represent, for different values of the parameters $(\Delta, \Gamma, \Omega)$ the corresponding probabilistic guarantees (probability to be in the uncertainty set) and values of $z_{\text {rob }}$. The shape of these plots makes sense: the highest is the probabilistic guarantee the lowest is the worst case gain. These plots illustrate the trade off between robustness and optimality.

In the authors' opinion, the most interesting points are the one with a worst case gain close to 0 (slightly positive or negative), they have relatively high probabilistic guarantees and a 'not too negative' worst case gain. Every player, according to his risk aversion, can choose the appropriate values of the parameters.

Comparing robust strategies and expected utility approaches We then select four strategies and analyze in detail their performance. We compare them to the expected utility approach used in [1]. [1] assumes that the player has a logarithmic utility function and wants to maximize his expected utility, where the budget is constrained. The authors propose a Monte Carlo method coupled with a gradient ascent algorithm to solve this problem. To compute the expected utility, assume that the player's probabilities are 'correct' (ie $\left.\mathbb{P}\left(y_{i}=1\right)=p_{i}\right)$.

We are going to use the same assumption to simulate 10,000 realizations of the uncertainty and compare the performances. In table 2 and Figure 2, we compare the overall performances of the different strategies. We can notice that the 4 robust strategies have, here, a similar distribution of returns. As expected, the robust strategies have a lower expected return than the gradient descent approach but have lower volatility.

|  | Strategy 1 (b) | Strategy 2 (b) | Strategy 3 (PC) | Strategy 4 (PC) | Gradient descent |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\Delta$ | 14 | 13 | 10 | 12 |  |
| $\Omega$ |  |  | 7 | 5 |  |
| $\Gamma_{1}$ | 4 | 4 |  |  |  |
| $\Gamma_{2}$ | 9 | 9 |  |  |  |
| Worst case return | $0.17 \%$ | $-3.56 \%$ | $4 \%$ | 0.82 |  |
| Probabilistic | 0.32 | 0.38 | 0.54 |  |  |
| guarantee |  |  |  | $7.9 \%$ | $12.5 \%$ |
| Mean return | $7.9 \%$ | $7.3 \%$ | $7.3 \%$ | $20 \%$ | $28 \%$ |
| Standard deviation | $20 \%$ | $22 \%$ | $19 \%$ |  |  |
| return |  |  |  |  |  |

Table 2: Performances of different robust strategies


Figure 2: Comparing the robust strategies to the gradient descent approach

## 5 Application to the FIFA Soccer World Cup

In this section, we apply the approaches developed above to the FIFA Soccer World Cup. The World Cup is a Soccer championship held every 4 years where 32 national teams from every continent qualify for a final tournament. Predicting the outcome of a World Cup edition is a challenge, as it is held every 4 years teams' players (and thus performances) can be quite different from one tournament to the other. Furthermore, and for the same reason, there is a limited amount of data available to train a predictive model.

Predicting the outcome of soccer games is a problem relatively well studied in the literature. Different approaches have been proposed to predict the outcome of every game in a national championship or in an international tournament (see [9] and [10]). Poisson processes have been used to predict the number of goals of each team and logit and probit models have been used to predict the outcome of a game.

| Entire Data Set |  | Restricted Data Set |  |
| :--- | :---: | :---: | :---: |
|  | Accuracy |  | Accuracy |
| Train Baseline | $33 \%$ | Train Baseline | $33 \%$ |
| Train Model | $38 \%$ | Train Model | $57 \%$ |
| Test Baseline | $33 \%$ | Test Baseline | $33 \%$ |
| Test Model | $50 \%$ | Test Model | $52 \%$ |

Table 3: Model Global Accuracy

We will present here a relatively simple approach that we built to forecast the outcome of the 2014 World Cup edition.

## Data and model

Data We collected data from the last 5 editions of the World Cup (1994 to 2010). We collected the outcome of every game and features of every team: FIFA Ranking before the competition, number of consecutive participation to the tournament, performance during the qualifying phase (average number of goals scored and conceded), performance of the Under20 team two and four years before the current edition, qualifying zone and a binary variable (called 'Continent') that indicates whether the world cup edition is held in the team's continent. We built a model to predict the outcome of every game and a model that predicts which teams will qualify to the bracket phase.

Predicting the outcome of a game We decided to focus on the group stage where the list of games is known in advance (as required by the robust betting strategy). In the group stage, every game can have three possible outcomes: one team may win or there can be a draw. To account for this soccer-specific characteristic of draws, we used an ordered logistic model (generalization of a logit model that can handle multiple and ordered outcomes). We trained our model with the four oldest editions (1994 to 2006) data and validated it with 2010 data. The final test set is the 2014 edition. Among all the features we selected, only three of them turned out to have a significant predictive power:

- the FIFA Ranking
- the number of consecutive participations to the competition
- the 'Continent' binary variable

This model gives, for every game, the probability for every outcome (Team 1 wins, Team 2 wins or draw). The outcome with the highest probability is the predicted outcome. In figure 4 is the results in R of the ordered logistic regression and in table 3 we compare the accuracy of our model to a random baseline.

The first observation to make is that it is difficult to correctly guess that a game is going to be a draw. Our model predicts a draw when the two teams have very similar features, but few of this 'close games' turn out to be draws. Furthermore, when the model predicts a draw, the associated probability is always low (below $40 \%$ ) which means that it is never confident about this

| Train |  | Test |  |
| :---: | :---: | :---: | :---: |
|  | Accuracy |  | Accuracy |
| Baseline | $50 \%$ | Baseline | $50 \%$ |
| Smart Baseline | $72.9 \%$ | Smart Baseline | $65.6 \%$ |
| Model | $75.7 \%$ | Test Model | $68.7 \%$ |

Table 4: OutGroup Model Accuracy
outcome. For our betting purposes, we decide to bet only on games where our model does not predict a draw and where the outcome probability is higher than 0.5 On this new subset of games (restricted subset), our model performs significantly better as reported in table 3 .

Predicting which teams qualify after the group phase As predicting the outcome of every single game can be hard, we might also want to bet on the event of team advancing from group stage to round of 16 . Therefore for each team in the World Cup, we consider another set of variables:

- Team (str): Name of the team
- FIFApts (num): The FIFA points normalized
- PartInaRow (int): The consecutive number of participation in the World Cup
- Team. 1 (str): Name of a team in Team's group
- FIFApts. 1 (num): The FIFA points also rescaled of Team. 1
- Qualifzone. 1 (int): The Qualifying zone of Team.1. In Part II, it was introduced as a string. Here we model it by integers; 'SA' (South America) is replaced by 1, 'AS' (Asia) by -1 and all the other ones by 0 (this choice will be explained further down)
- Same 3 features for Team. 2 and Team. 3
- OutGroup (int): Binary variable. 1 if Team qualified for the final bracket, 0 if not.

In this case, the training set is going to be every team participating in the 1994, 1998 and 2002 World Cups (88 observations). The validation set is the 2006 and 2010 World Cups (64 observations) and our test set 2014. We use a logistic regression. The coefficients we obtain are in Table 8 and the results are available in 4. It appears that group outcomes are more easy to predict as we take into account different games (and thus less random effects). This averages the error of the model by compensation. Therefore, we have access for each team to the probability of this team advancing to the knock-out stage. We select the event we bet on the following way: For each group, we consider the two teams the most likely to qualify and we bet on the fact that these teams qualify.

Robust betting strategy for 2014 Applying those two models to the 2014 World Cup which took place this summer in Brazil, we have access to a list of $\mathrm{N}=14$ events of two types: games and teams advancing to the next stage. Each game considered has an edge ( $p_{i} c_{i}>1$ ). With the same methods presented in section 4, we select our parameters for both policies. After analysis, we selected the following strategy:

- The PC Approach using $\Delta=9$ and $\Omega=6$ with $z_{\text {rob }}=9.78$ and $\mathbb{P}\left(y \in \Xi_{P C}\right) \geq 80 \%$

Let us compare our strategy with the one given by the gradient method. Both strategies are given in table 7. First of all, let us compare qualitatively the two approaches. The gradient approach trusts the model in the sense that, the more the player has an edge, the more it bets. This seems very risky, as more than $75 \%$ of the budget is distributed on only 3 events (where the edge is important). On the opposite, the robust strategy still bets on those events as the reward is important but also considers the other games to diversify its risk. On the other hand, the expected gain seems higher in the gradient strategy than in the robust one. We can verify it quantitatively.

Assuming that our probabilities are right, we simulate $50^{\prime} 000$ instances of the World Cup and gather the results in Table 5 and Figure 3. As seen in the interpretation, the mean is significantly greater for the gradient case. However, the variance dramatically decreases with the robust strategy. A good way to evaluate the trade-off between the two is the Sharpe Ratio (defined as the ratio between the mean return and its standard deviation) which is almost ten times higher for the robust model.

Until now, we have always assumed that the 'probabilities were right' but what happen if it is not the case? As our probabilities come from a model they are just an estimation of the real probabilities. Furthermore, even if we assume that our estimator is unbiased it is likely that the player's probabilities are an overestimation for games where he has an edge (ie the player's probability is higher than the bookmaker's). As an example, let us consider the case where the real probabilities are the average between the player's and the bookmaker's predictions. We keep the strategies developed above and recompute the distribution of return. The results are in Table 6. We can see that the mean return is lower and that the variance is almost unchanged for the two strategies. Nevertheless, we can see that, in this setting the $25 \%$ quantile of the gradient descent approach is negative, while the robust strategy's stays next to 0 . This example illustrates the fact that, when dealing with probability estimations, a stochastic approach can be risky in cases of overestimation, while a robust approach that doesn't assume any probability distribution can be safer.

In conclusion, in this paper we developed a risk-averse betting strategy. It guarantees a minimum gain with a certain probability. Compared to standard stochastic optimization approaches, the mean return is lower but the volatility is dramatically decreased. Thus, we developed a "safe" risk betting approach that, combined with a good prediction model, can guarantee a minimum gain and prevent from significant losses.

|  | Robust Betting | Gradient Betting |
| ---: | :---: | :---: |
| Mean Return | $29 \%$ | $175 \%$ |
| Standard Deviation Return | $4,1 \%$ | $203 \%$ |
| Sharpe Ratio | $\mathbf{6 . 9 1}$ | $\mathbf{0 . 8 6}$ |

Table 5: Comparison of the Betting Strategies for the 2014 FIFA World Cup

|  | Robust Betting | Gradient Betting |
| ---: | :---: | :---: |
| Mean Return | $14 \%$ | $86 \%$ |
| Standard Deviation Return | $4,6 \%$ | $189 \%$ |
| Sharpe Ratio | $\mathbf{3 . 1 2}$ | $\mathbf{0 . 4 6}$ |

Table 6: Comparison of the Betting Strategies for the 2014 FIFA World Cup (wrong probabilities)

|  | Event | Model's <br> Probability | Bookmaker's <br> Probability | Odds | Robust Betting <br> Strategy | Gradient Betting <br> Strategy |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
| 1 | Iran adv. | 0.405 | 0.118 | 8.5 | 1.61 | 21.66 |
| 2 | Italy bt. England | 0.537 | 0.407 | 2.46 | 5.58 | 1.81 |
| 3 | Mexico adv. | 0.571 | 0.465 | 2.15 | 6.38 | 0.83 |
| 4 | Greece adv. | 0.604 | 0.333 | 3.00 | 4.57 | 12.02 |
| 5 | Algeria adv. | 0.610 | 0.182 | 5.50 | 2.49 | 39.78 |
| 6 | USA adv. | 0.642 | 0.333 | 3.00 | 4.57 | 16.30 |
| 7 | Netherlands adv. | 0.656 | 0.637 | 1.57 | 8.74 | 0.00 |
| 8 | Argentina bt. Nigeria | 0.806 | 0.735 | 1.36 | 10.09 | 0.15 |
| 9 | Brazil bt. Croatia | 0.823 | 0.787 | 1.27 | 10.81 | 0.01 |
| 10 | Italy adv. | 0.909 | 0.694 | 1.44 | 9.53 | 7.08 |
| 11 | Brazil bt. Cameroon | 0.916 | 0.833 | 1.20 | 11.44 | 0.20 |
| 12 | Spain adv. | 0.930 | 0.877 | 1.14 | 12.04 | 0.02 |
| 13 | Germany adv. | 0.966 | 0.885 | 1.13 | 12.14 | 0.14 |
| 14 | Brazil adv. | 0.973 | 0.971 | 1.03 | 0.00 | 0.00 |

Table 7: Betting Strategy for the 2014 FIFA World Cup


Figure 3: Boxplots of both Strategies (1:Robust, 2:Gradient) for the World Cup

## 6 Conclusion

We considered a player that wants to bet on a set of $N$ games ( $N$ large). He has a model that predicts the probability of every game. He has tested his model on past data and knows its performances (accuracy). The player wants to bet robustly, i.e wants to follow a betting strategy that will maximize the worst case gain knowing the accuracy of his model.

We presented two different approaches to model the uncertainty set. The first one uses explicitly the probabilities given by the predictive model, the second one is more aggregate and splits the games into 'buckets'. On a mathematical point of view, these two ways of modeling the uncertainty set are not convex but integer. We explained how we can use the relaxation to get a lower bound (or the exact value in the bucket case) or how we can use the cutting plane algorithm to solve the problem exactly. We explained how to choose the robust parameters using a simulated dataset and we compared this new approach to a classical stochastic optimization approach. Finally, we proposed two predicting models of the Soccer World Cups and tested our approaches on the 2014 FIFA World Cup data.

Our analysis leads us to believe that Robust Optimization is a novel and valuable approach to modeling betting problems on a large number of simultaneous events. Combined with a good prediction model, this betting strategy can guarantee significant gains while reducing the risk.

## References

[1] Whitrow, Algorithms for optimal allocation of bets on many simultaneous events. Applied Statistics, Vol. 56, Part 5, pp. 607/623, 2007.
[2] Browne \& Ward Whitt, Portfolio Choice and the Bayesian Kelly Criterion. Advances in Applied Probability, Vol. 28, No. 4, pp. 1145/1176, 1996.
[3] Haigh, The Kelly Criterion and Bet Comparisons in Spread Betting. Journal of the Royal Statistical Society, Series D, The Statistician, Vol.49, No. 4, pp. 531/539, 2000.
[4] Thorp, The Kelly criterion in blackjack, sports betting and the stock market. Edward O. Thorp \& Associates, Newport Beach, 1997.
[5] Bertsimas \& Sim, The Price of Robustness. Operations Research, Vol. 52, pp. 35/53, 2004.
[6] Bertsimas, Pachamanova \& Sim, Robust linear optimization under general norms. Operations Research Letters, Vol. 32, pp. 510/516, 2004.
[7] Bertsimas, Brown \& Caramanis, Theory and Applications of Robust Optimization. Society for Industrial and Applied Mathematics, Vol. 53, No. 3, pp. 464/501, 2011.
[8] Bertsimas, Dunning \& Lubin, Reformulations versus cutting planes for robust optimization, A computational and machine learning perspective. Working paper.
[9] Dyte \& Clark, A Ratings Based Poisson Model for World Cup Soccer Simulation. The Journal of the Operational Research Society, Vol. 51, No. 8, pp. 993/998,Aug. 2000.
[10] Goddard, Regression models for forecasting goals and match results in association football. International Journal of Forecasting 21, pp.331/340, 2005.

## Appendices

## A The Robust Cutting Plane algorithm

The algorithm is basically the cutting plane algorithm as it is used in Integer programming. It was developed by Bertsimas \& al in [8]. The problem can be written as follows

$$
\begin{array}{ll}
\underset{t \in \mathbb{R}, x \in P}{\operatorname{maximize}} & \left(t-\sum_{i=1}^{N} x_{i}\right) \\
\text { subject to } & 0 \leq \sum_{i=1}^{N} c_{i} x_{i} y_{i}-t u, \forall y \in \Xi_{P C}, u \in\{1\} \tag{16}
\end{array}
$$

In order not to have an unbounded initial problem and knowing that $(1, \ldots, 1,1) \in \Xi_{P C} \times\{1\}$, the algorithm starts by solving the following problem called the master problem

$$
\begin{array}{ll}
\underset{t \in \mathbb{R}, x \in P}{\operatorname{maximize}} & \left(t-\sum_{i=1}^{N} x_{i}\right) \\
\text { subject to } & 0 \leq \sum_{i=1}^{N} c_{i} x_{i}-t \tag{17}
\end{array}
$$

It gives a solution $\left(\mathrm{t}^{*}, \mathrm{x}^{*}\right)$. Then it solves the separation problem

$$
\underset{(y, u) \in \Xi_{P C} \times\{1\}}{\operatorname{minimize}}\left(\sum_{i=1}^{N} c_{i} x_{i}^{*} y_{i}-t^{*} u\right)
$$

This problem is a binary IP that looks very similar to the knapsack problem. It gives an optimal cost $\mathrm{v}^{*}$ corresponding to a certain $y^{*} \in \Xi_{P C}$. If $v^{*} \geq 0$ then the robust solution is ( $\mathrm{x}^{*}, \mathrm{t}^{*}$ ). If not the algorithm add the constraint

$$
0 \leq \sum_{i=1}^{N} c_{i} x_{i} y_{i}^{*}-t
$$

to the master problem. The number of cuts that the algorithm adds is finite since it adds at most one cut per element of $\Xi_{P C}$ and $\left|\Xi_{P C}\right| \leq 2^{N}$. Hence the algorithm terminates and gives an optimal solution. However, the number of iterations might be might be exponential.

## B Linear Robust Formulations

The Relaxation formulation First, we write the dual of the subproblem:

$$
\begin{align*}
& \underset{y}{\operatorname{minimize}} \sum_{i=1}^{N} c_{i} x_{i} y_{i} \\
& \text { subject to } \sum_{i=1}^{N} y_{i} \geq \Delta \quad \underset{a, b, f}{\operatorname{maximize}} \Delta a+\Omega b-\sum_{i=1}^{N} f_{i} \\
& \begin{array}{ll}
\sum_{i=1}^{N} p_{i} y_{i} \geq \Omega & \text { subject to } \begin{array}{l}
a+b-f_{i} \leq c_{i} x_{i} \forall i \\
\\
\\
a, b, f \geq 0
\end{array}
\end{array}  \tag{18}\\
& y \leq 1 \\
& y \geq 0
\end{align*}
$$

By merging the two maximization problem, we get a simple LP called the Robust counterpart. This is indeed very efficiently solvable and tractable for large N .

$$
\begin{array}{cl}
\underset{a, b, f, x}{\operatorname{maximize}} & \Delta a+\Omega b-\sum_{i=1}^{N} f_{i}-\sum_{i=1}^{N} x_{i} \\
\text { subject to } & a+b-f_{i} \leq c_{i} x_{i} \forall i \\
& \sum_{i=1}^{N} x_{i} \leq B  \tag{19}\\
& a, b, f, x \geq 0
\end{array}
$$

The Buckets formulation By strong duality

$$
\begin{array}{cl}
\underset{y}{\operatorname{minimize}} & \sum_{i=1}^{N} c_{i} x_{i} y_{i} \\
\text { subject to } & \sum_{i=1}^{N} y_{i} \geq \Delta \\
& \sum_{i \in G_{k}} y_{i} \geq \Gamma_{k} \forall k \\
& y \leq 1 \\
& y \geq 0
\end{array}
$$

The robust formulation becomes

$$
\begin{array}{ll}
\underset{a, b, f, x}{\operatorname{maximize}} & \sum_{k=1}^{K} \Gamma_{k} b_{k}+\Delta a-\sum_{i=1}^{N} f_{i}-\sum_{i=1}^{N} x_{i} \\
\text { subject to } & a+\sum_{k \mid i \in G_{k}} b_{k}-f_{i} \leq c_{i} x_{i} \forall i  \tag{21}\\
& \sum_{i=1}^{N} x_{i} \leq B \\
& a, b, f, x \geq 0
\end{array}
$$

## C Proofs

Lemma C.1. All the extreme points of $R_{P C}$ have at most two non integral components
Proof. Every extreme point has N binding inequalities. In our problem we have $N$ inequalities $0 \leq y_{i} \leq 1$ and 2 other inequalities. Thus at least $N-2 y$ s have binary values.

Proposition C.1. We have

$$
\begin{equation*}
z_{\text {relax }} \leq z_{\text {rob }} \leq z_{\text {relax }}+M \tag{22}
\end{equation*}
$$

where $M=\max _{i, j=1, \ldots, N, i \neq j}\left(c_{i} x_{i}^{*}+c_{j} x_{j}^{*}\right)$, and $x^{*}$ is the solution of the relaxed problem.

Proof. The first part of the inequality is direct using the fact that $\Xi_{P C} \subset R_{P C}$. Let x* be the robust solution. By solving the subproblem for $x=x^{*}$, the optimal solution is an extreme point of $R_{P C}$ (let us call it $u$ ). By lemma 3.1, u has at most two non integral components. Let's consider $\mathrm{y}^{*}$ such as $y_{i}^{*}=\left\{\begin{array}{c}y_{i} \text { if } u_{i} \in\{0,1\} \\ 1 \text { if } 0<u_{i}<1\end{array}\right.$ Notice that $y^{*} \in \Xi_{P C}$, and moreover by taking $x=x^{*}$ and $y=y^{*}$, let z be the objective value for this solution. We have $z \geq z_{\text {rob }}$ and $z-z_{\text {relax }} \leq \sum_{i=1}^{N} c_{i} x_{i}^{*}\left(y_{i}^{*}-u_{i}\right) \leq M$.

Proposition C.2. If $\Delta$ and $\left(\Gamma_{k}\right)_{k=1 \ldots} K$ are integers and $I_{k}$ are intervals, then $\Xi_{b}$ has integer extreme points.

Proof. We can then write $\Xi_{b}$ using a matrix notation:

$$
\Xi_{b}=\left\{y \in\{0,1\}^{N} \mid A y \geq \boldsymbol{\Gamma}\right\}
$$

where

$$
A=\left(\begin{array}{ccccccc}
1 & 1 & 1 & 1 & \cdots & 1 & 1 \\
1 & 1 & 1 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & 1 & 0 & \cdots & 0 \\
0 & & & & \cdots & & \\
0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}\right)
$$

and

$$
\boldsymbol{\Gamma}=\left(\begin{array}{c}
\Delta \\
\Gamma_{1} \\
\ldots \\
\Gamma_{K}
\end{array}\right)
$$

The matrix $A$ has the 'consecutive ones property' (in every row the ones are consecutive) thus it is totally unimodular. As we assumed that $\boldsymbol{\Gamma}$ has integer values then the extreme points of

$$
\left\{y \in[0,1]^{N} \mid A y \geq \boldsymbol{\Gamma}\right\}
$$

are integer.
Proposition C.3. Let $\Delta \in \mathbb{N}$, and the Optimal Objective value has the two following properties

$$
\begin{equation*}
z_{\text {single bucket }} \geq B\left(\frac{\Delta}{\sum_{i=1}^{N} 1 / c_{i}}-1\right) \tag{23}
\end{equation*}
$$

Proof. A feasible solution of single bucket case is such as $\forall i \in\{1, \ldots, N\}, x_{i}=\frac{B}{c_{i} \times \sum_{j=1}^{N} 1 / c_{j}}$ Indeed, $\sum_{i=1}^{N} x_{i}=B$ and by taking $a=\frac{B}{\sum_{j=1}^{N} 1 / c_{j}}$ and $f=0$, one can easily see that x is feasible. Computing the cost for this solution gives the lower bound wanted on the optimal cost.

## Proposition C.4.

$$
\begin{equation*}
z_{\text {partition }} \geq B\left(\frac{1}{K} \sum_{k=1}^{K} \frac{\Gamma_{k}}{\sum_{j \in G_{k}}^{N} 1 / c_{j}}-1\right) \tag{24}
\end{equation*}
$$

Proof. First of all, let us notice that since $\left(I_{k}\right)$ form a partition of $\left[p_{\min }, p_{\max }\right]$, for all $i \in\{1, \ldots, N\}$, there exists a unique $k$ such that $i \in I_{k}$, let us note it $k(i)$. Here the solution is $x_{i}=\frac{B / K}{c_{i} \times \sum_{j \in G_{k(i)}} 1 / c_{j}}$ $\forall i \in\{1, \ldots, N\}$ and $b_{k}=\frac{B / K}{\sum_{j \in G_{k(i)}}^{1 / c_{j}}}, \forall k \in 1, \ldots, K$ and $f=0$.

$$
\sum_{i=1}^{N} x_{i}=\sum_{k=1}^{K} \sum_{i \in G_{k}} x_{i}=\sum_{k=1}^{K} \sum_{i \in G_{k}} \frac{B / K}{c_{i} \times \sum_{j / i n G_{k(i)}} 1 / c_{j}}=\frac{B}{K} \sum_{k=1}^{K} \frac{\sum_{i \in G_{k}} 1 / c_{i}}{\sum_{j \in G_{k}} 1 / c_{j}}=B
$$

and let $i \in\{1, \ldots, N\}, \sum_{k \mid i \in G_{k}} b_{k}=b_{k}(i)=c_{i} x_{i}$ so the solution given is feasible. To get the bound we compute its cost which gives the wanted lower bound.

$$
\sum_{k=1}^{K} \Gamma_{k} b_{k}-\sum_{i=1}^{N} x_{i}=B\left(\frac{1}{K} \sum_{k=1}^{K} \frac{\Gamma_{k}}{\sum_{j \in G_{k}}^{N} 1 / c_{j}}-1\right)
$$

## Lemma C.2.

$$
\begin{equation*}
\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right) \simeq \Phi\left(\frac{\Delta-\sum p_{i}}{\sqrt{\sum p_{i}\left(1-p_{i}\right)}}\right) \tag{25}
\end{equation*}
$$

where $\Phi$ is the cumulative distribution function of a standard normal.

Proof. The result comes from the application of the central limit theorem for independent but non identically distributed variables (Lindeberg-Feller thereom). Let us consider $X_{i}=y_{i}-p_{i}$ a sequence of independent random variables such that $\mathbb{E}\left(X_{i}\right)=0$ and $\mathbb{E}\left(X_{i}^{2}\right)=p_{i}\left(1-p_{i}\right)$. Let $s_{n}^{2}=\sum_{i=1}^{N} p_{i}\left(1-p_{i}\right)$. Let us also assume that $\exists p_{\max }<1, p_{\min }>0$ s.t. $p_{i} \in\left[p_{\min }, p_{\max }\right] \forall i$. This assumption is realistic as our data is bounded, the probabilities of our model are not too close to 0 and 1. Then using Lyapunov condition and Lindeberg-Feller theorem we get

$$
\frac{\sum_{i=1}^{N} X_{i}}{s_{n}^{2}} \longrightarrow \mathcal{N}(0,1) \text { in distribution }
$$

and thus gives us the result.

## Proposition C.5.

$$
\begin{equation*}
\mathbb{P}\left(y \in \Xi_{P C}\right) \geq\left(1-\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right)\right) \times\left(1-\mathbb{P}\left(\sum_{i=1}^{N} p_{i} y_{i}<\Omega\right)\right) \tag{26}
\end{equation*}
$$

Similarly, if the buckets don't overlap, we have

$$
\begin{equation*}
\mathbb{P}\left(y \in \Xi_{b}\right) \geq\left(1-\mathbb{P}\left(\sum_{i=1}^{N} y_{i}<\Delta\right)\right) \times \prod_{k=1}^{K}\left(1-\mathbb{P}\left(\sum_{i \in G_{k}} y_{i}<\Gamma_{k}\right)\right) \tag{27}
\end{equation*}
$$

Proof. Let $\delta$ be the event $\left\{\sum_{i=1}^{N} y_{i}<\Delta\right\}, \omega=\left\{\sum_{i=1}^{N} p_{i} y_{i}<\Omega\right\}$ and $\forall k=\{1, \ldots, K\}, \gamma_{k}=$ $\left\{\sum_{i \in G_{k}} y_{i}<\Gamma_{k}\right\}$ We have

$$
\mathbb{P}\left(y \in \Xi_{P C}\right)=\mathbb{P}(\bar{\delta}, \bar{\omega})=\mathbb{P}(\bar{\omega} \mid \bar{\delta}) \mathbb{P}(\bar{\delta})
$$

Now notice that $\mathbb{P}(\bar{\omega} \mid \bar{\delta}) \geq \mathbb{P}(\bar{\omega})$ because if $\bar{\delta}$ is true, then $\bar{\omega}$ has more chance to be true than if we didn't have any information on $\bar{\delta}$. So

$$
\mathbb{P}\left(y \in \Xi_{P C}\right) \geq \mathbb{P}(\bar{\omega}) \mathbb{P}(\bar{\delta})=(1-\mathbb{P}(\omega)) \times(1-\mathbb{P}(\delta))
$$

For the bucket case, notice that since the bucket are not overlapping then $\gamma_{k}$ are independent. Using the same reasoning:

$$
\mathbb{P}\left(y \in \Xi_{b}\right)=\mathbb{P}\left(\bar{\delta}, \forall k \overline{\gamma_{k}}\right)=\mathbb{P}\left(\bar{\delta} \mid \forall k \overline{\gamma_{k}}\right) \mathbb{P}\left(\forall k \overline{\gamma_{k}}\right)=\mathbb{P}\left(\bar{\delta} \mid \forall k \overline{\gamma_{k}}\right) \prod_{i=1}^{K} \mathbb{P}\left(\overline{\gamma_{k}}\right)
$$

Again for a similar reason as before $\mathbb{P}\left(\bar{\delta} \mid \forall k \overline{\gamma_{k}}\right) \geq \mathbb{P}(\bar{\delta})$ So

$$
\mathbb{P}\left(y \in \Xi_{b}\right) \geq \mathbb{P}(\bar{\delta}) \prod_{i=1}^{K} \mathbb{P}\left(\overline{\gamma_{k}}\right)=(1-\mathbb{P}(\delta)) \times \prod_{i=1}^{K}\left(1-\mathbb{P}\left(\gamma_{k}\right)\right)
$$

## D Models' Coefficients

```
Coefficients:
    Estimate Std. Error z value Pr(>|z|)
FIFARanking -0.023388 0.007877 -2.969 0.00299 **
PartInaRow 0.097836 0.030984 3.158 0.00159 **
Continent 0.733048 0.225714 3.248 0.00116 **
---
Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ' ' 1
Threshold coefficients:
    Estimate Std. Error z value
2|X -0.8606 0.1847 -4.659
X|1 1.1304 0.1913 5.910
```

Figure 4: Ordered logistic: significant variables

| Variables | Coefficient |
| :---: | :---: |
| FIFApts | 0.578148 |
| PartInaRow | 0.171173 |
| FIFApts. 1 | -0.07393 |
| Qualifzone. 1 | 0.39428 |
| FIFApts.2 | -0.07393 |
| Qualifzone. 2 | 0.39428 |
| FIFApts.3 | -0.07393 |
| Qualifzone.3 | 0.39428 |
| Intercept | -0.44935 |

Table 8: Coefficients of the Logistic Regressions

