Premier League

Un modèle de simulation d’évènements rares

Raphaël Montaud et Gabriel Misrachi

Modal Simulation d’évènements rares MAP474D - Ecole Polytechnique

Table of contents

Introduction

Introduction

- les championnats de football: un monde incertain

Introduction

- les championnats de football: un monde incertain
- Premier League 2015-2016:

Introduction

- les championnats de football: un monde incertain
- Premier League 2015-2016:
  - Leicester gagne

Introduction

- les championnats de football: un monde incertain
- Premier League 2015-2016:
  - Leicester gagne
  - Manchester City et Manchester United hors podium

Introduction

- les championnats de football: un monde incertain
- Premier League 2015-2016:
  - Leicester gagne
  - Manchester City et Manchester United hors podium
  - Liverpool et Chelsea hors places européennes

Introduction

- les championnats de football: un monde incertain
- Premier League 2015-2016:
  - Leicester gagne
  - Manchester City et Manchester United hors podium
  - Liverpool et Chelsea hors places européennes
- une côte à un 1 contre 5000: 30 millions de pertes pour les bookmakers

Le modèle de Bradley-Terry (1952)

Le modèle de Bradley-Terry (1952)

Un vecteur de force ou de mérite V. On a ensuite chaque rencontre qui est le résultat d'une expérience de Bernoulli:

$$\mathrm{P}(X_{i,j} = 1) = \frac{V_i}{V_i + V_j}$$

Premières implémentations

Premières implémentations

- un modèle pour le calcul des forces: (score année précédente)$^\kappa$

Premières implémentations

- un modèle pour le calcul des forces: (score année précédente)$^\kappa$
- matrice des résultats:

$$\begin{pmatrix} N & 0 & 0 & 1 \\ 1 & N & 1 & 0 \\ 1 & 0 & N & 0 \\ 0 & 1 & 1 & N \end{pmatrix}$$

Premières implémentations

- un modèle pour le calcul des forces: (score année précédente)$^{\kappa}$
- matrice des résultats:

$$\begin{pmatrix} N & 0 & 0 & 1 \\ 1 & N & 1 & 0 \\ 1 & 0 & N & 0 \\ 0 & 1 & 1 & N \end{pmatrix}$$

- premier résultat:

Premières implémentations

```
according to our model, we compute the chances of winning for the best teams
(given with a 95% interval)
proba of winning for Chelsea with n = 80000: 37.0625% +- 0.334872092679%
proba of winning for ManCity with n = 80000: 25.5375% +- 0.301060990353%
proba of winning for Arsenal with n = 80000: 20.4425% +- 0.279516471263%
proba of winning for ManU with n = 80000: 15.007499999999999% +- 0.24668839648%
proba of winning for Tottenham with n = 80000: 9.3125% +- 0.202190825793%
We now compare it to the probabilities according to the bookmakers
Chelsea: 38.095238095238095%
ManCity: 28.57142857142857%
Arsenal: 22.22222222222222%
ManU: 16.666666666666664%
Tottenham: 0.9900990099009901%
```

Evènements rares

Evènements rares

- Les limites de la simulation de Monte-Carlo standard
- La méthode de décalage en probabilité
- Théorème ergodique et splitting

Décalage en probabilité: un résultat utile

$Y_1, \ldots, Y_k$ Bernoulli indépendantes $B(p_1), \ldots, B(p_k)$; et $Z_1, \ldots, Z_k$ Bernoulli indépedantes $B(q_1), \ldots, B(q_k)$. Pour toute fonction mesurable bornée g, on a :

$$\mathrm{E}[g(Y_1, \ldots, Y_k)] = \left(\prod_{i=1}^{k} \frac{1-p_i}{1-q_i}\right) \mathrm{E}\left[g(Z_1, \ldots, Z_k) \prod_{i=1}^{k} \left(\frac{p_i(1-p_i)}{q_i(1-q_i)}\right)^{Z_i}\right]$$

On l'utilise en remplaçant g par l'indicatrice de l'événement considéré.

Décalage en probabilité: implémentation et résultats

- forcer l'événement

Décalage en probabilité: implémentation et résultats

- forcer l'événement
- simplification des calculs

Décalage en probabilité: implémentation et résultats

- forcer l'événement
- simplification des calculs
- résultats

Evénement	Probabilité	Intervalle	Simulations
Leicester gagne	0,41%	0,025%	10 000
Evénement complexe	0,0069%	0,0024%	1 000 000

Théorème ergodique et splitting

- On ne resimule qu'une fraction des matchs $\rightarrow$ construction d'une chaîne de Markov avec un nombre d'états finis.

Théorème ergodique et splitting

- On ne resimule qu'une fraction des matchs $\rightarrow$ construction d'une chaîne de Markov avec un nombre d'états finis.
- rejet des scores trop faibles

Théorème ergodique et splitting

- On ne resimule qu'une fraction des matchs $\rightarrow$ construction d'une chaîne de Markov avec un nombre d'états finis.
- rejet des scores trop faibles
- comptage des scores suffisants

Théorème ergodique et splitting

- On ne resimule qu'une fraction des matchs $\rightarrow$ construction d'une chaîne de Markov avec un nombre d'états finis.
- rejet des scores trop faibles
- comptage des scores suffisants
- calcul de la probabilité finale:

$$P(Leicester\,gagne) = P(Leicester\,gagne\,|score > 10)\ldots P(score > 3)$$

Théorème ergodique et splitting

- taux de rejet idéal $\simeq 20\%$

Théorème ergodique et splitting

- taux de rejet idéal $\simeq$ 20%
- taux de rejet élevés (même avec $\rho = 1$ *match*)

Théorème ergodique et splitting

- taux de rejet idéal $\simeq$ 20%
- taux de rejet élevés (même avec $\rho = 1$ *match*)
- $\{$*Leicester gagne*$|$*Leicester deuxieme*$\}$ est un évènement rare

Théorème ergodique et splitting

- taux de rejet idéal $\simeq 20\%$
- taux de rejet élevés (même avec $\rho = 1$ *match*)
- $\{$*Leicester gagne*$|$*Leicester deuxieme*$\}$ est un évènement rare
- $\rightarrow$méthode peu adaptée au problème

Comportement sur de grands championnats

Comportement sur de grands championnats

R. Chetrite, R. Diel, M. Lerasle *The number of potential winners in Bradley-Terry model in random environment*, Ann. Appl. Probab., 2016. lien

$\Rightarrow$ trois théorème que nous allons illustrer

Théorème 1

Thérorème 1:

Q la fonction quantile, U la distribution des forces. Si:

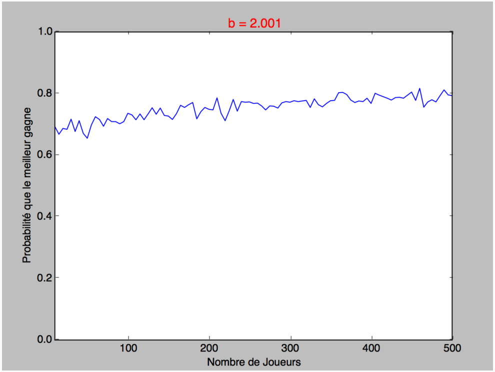
- Il existe β dans $(0, 1/2)$ et $x_0 > 0$ dans l'intérieur de *suppQ* tels que $Q^{1/2-\beta}$ est convexe sur $[x_0, \infty)$.
- $\mathrm{E}\left[U^2\right] < \infty$.

Alors, la probabilité que le joueur le plus fort gagne tend vers 1 avec le nombre N de joueurs:

$$P(\textit{le plus fort gagne}) \xrightarrow[N \to +\infty]{} 1.$$

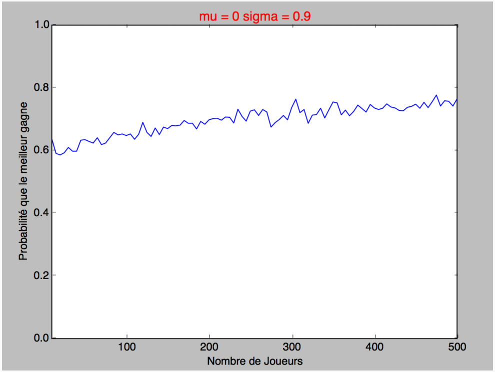
Theorème 1: Résultats

Cas de la distribution de queue en x^{-b}



Theorème 1: Résultats

Cas de la distribution log-normale



Théorème 2

Hypothèse A:

Le maximum de $supp(Q)$ est 1 et il existe $\alpha \in [0, 2)$ tel que:

$$logQ(1-u) = \alpha log(u) + o(logu) \quad quand\ u \to 0 \qquad (A)$$

Théorème 2:

Pour tout $\gamma < 1 - \alpha/2$, on a pour N tendant vers l'infini:

$$\mathrm{P}(aucun\ des\ N^{\gamma}\ meilleurs\ joueurs\ gagne) \to 1$$

Pour tout $\gamma > 1 - \alpha/2$, on a pour N tendant vers l'infini:

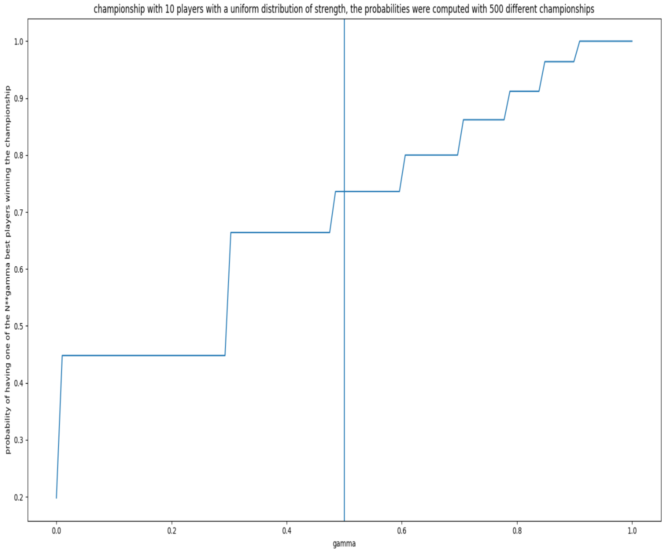
$$\mathrm{P}(un\ des\ N^{\gamma}\ meilleurs\ joueurs\ gagne) \to 1$$

Théorème 2

Lois respectant l'hypothèse A

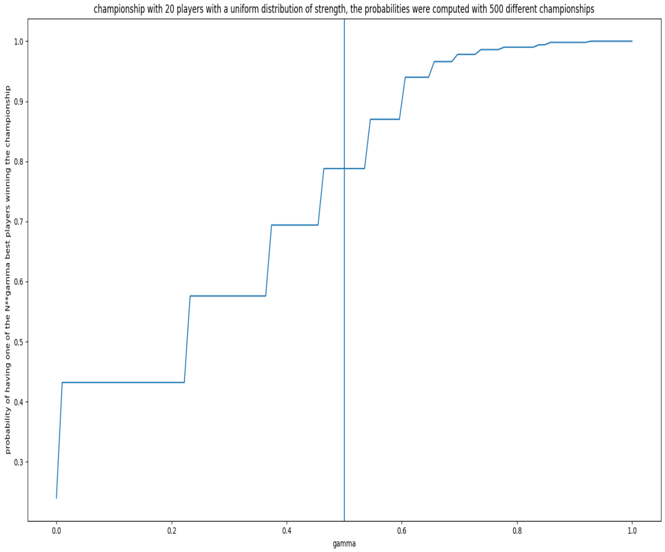
- la distribution uniforme avec $\alpha = 1$
- la distribution *arcsin* avec $\alpha = 1/2$
- les distributions $\beta(a, b)$ avec $\alpha = b$ si $b < 2$

Théorème 2

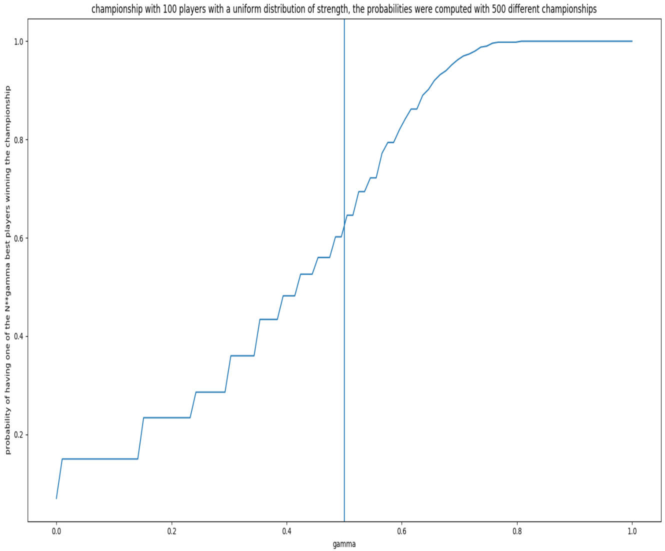


championship with 10 players with a uniform distribution of strength, the probabilities were computed with 500 different championships
probability of having one of the N**gamma best players winning the championship
gamma
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Théorème 2

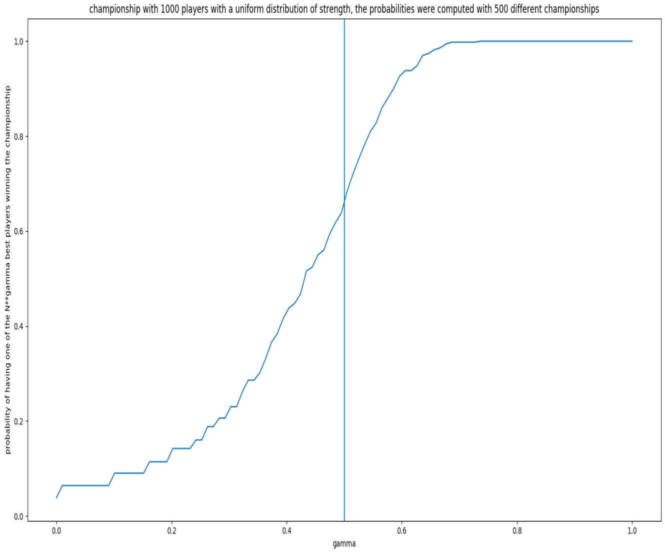

championship with 20 players with a uniform distribution of strength, the probabilities were computed with 500 different championships
probability of having one of the N**gamma best players winning the championship
gamma
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Théorème 2



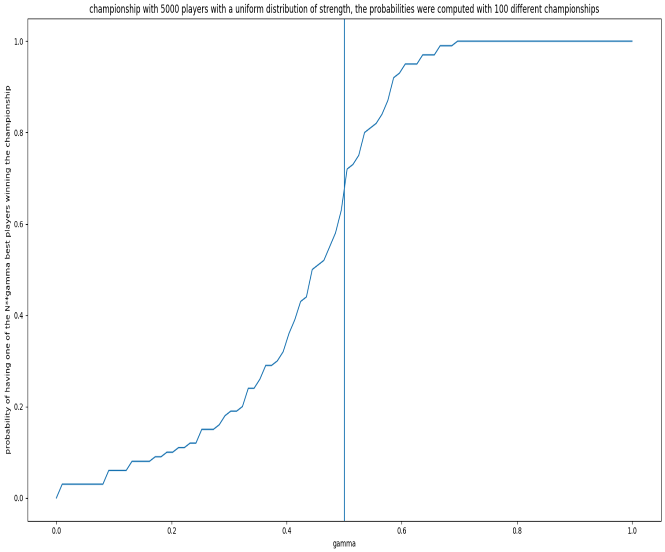
championship with 100 players with a uniform distribution of strength, the probabilities were computed with 500 different championships
probability of having one of the N**gamma best players winning the championship
gamma
0.0
0.2
0.4
0.6
0.8
1.0

Théorème 2



championship with 1000 players with a uniform distribution of strength, the probabilities were computed with 500 different championships
probability of having one of the N**gamma best players winning the championship
gamma
0.0
0.2
0.4
0.6
0.8
1.0

Théorème 2



championship with 5000 players with a uniform distribution of strength, the probabilities were computed with 100 different championships
probability of having one of the N**gamma best players winning the championship
gamma
0.0
0.2
0.4
0.6
0.8
1.0

Encore sous l'hypothèse A:

Théorème 3

$$Soit\ V_U = \mathrm{E}[\frac{U}{(U+1)^2}]\ et\ \epsilon_N = \left(\frac{(2-\alpha)(logN)}{NV_U}\right)^{1/2}$$

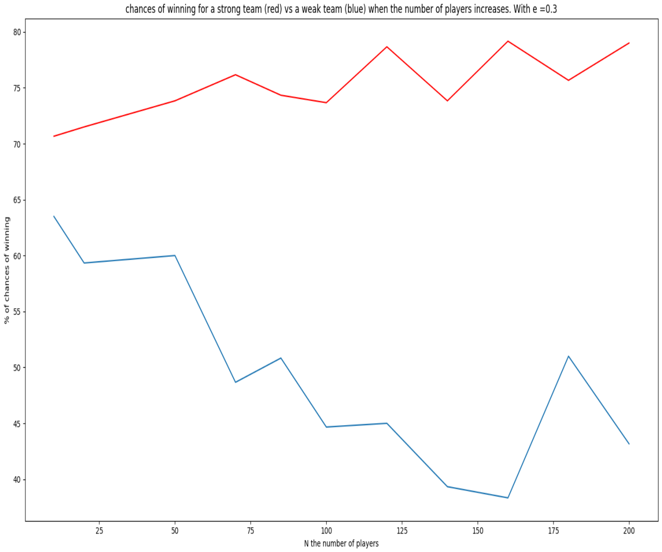
Si x_N représente la force d'un $N+1e$ joueur alors: Si $lim\ inf_{N\to\infty}\frac{x_N-1}{\epsilon_N} > 1$ alors P presque sûrement:

$$\mathrm{P}(le\ N+1e\ joueur\ gagne) \to 1$$

Si $lim\ sup_{N\to\infty}\frac{x_N-1}{\epsilon_N} < 1$ alors P presque sûrement:

$$\mathrm{P}(le\ N+1e\ joueur\ ne\ gagne\ pas) \to 1$$

Théorème 3



chances of winning for a strong team (red) vs a weak team (blue) when the number of players increases. With e =0.3
% of chances of winning
80
75
70
65
60
55
50
45
40
25
50
75
100
125
150
175
200
N the number of players

Théorème 3

Un cas particulier:

- Une équipe possède une force de 2 tandis que toutes les autres possèdent une force de 1
- Une équipe possède une force de 2, une autre une force de 1 et tout le reste a une force de 0.5.

$\Rightarrow$ On créé un script qui calcule la plus petite force nécessaire pour gagner avec 70% de chances

Théorème 3

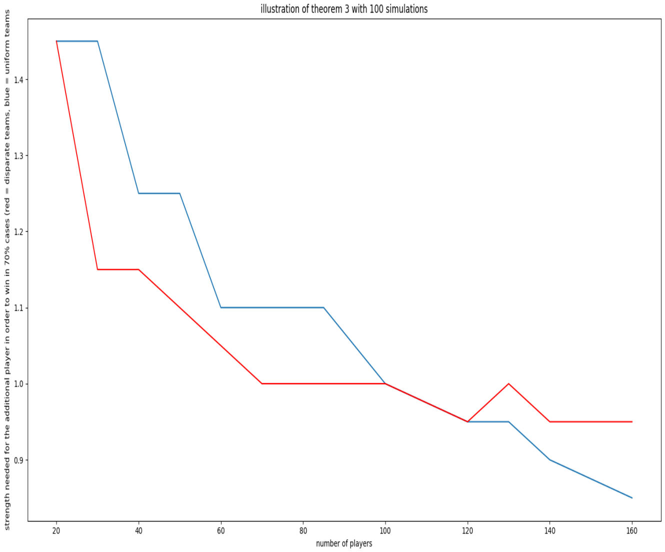


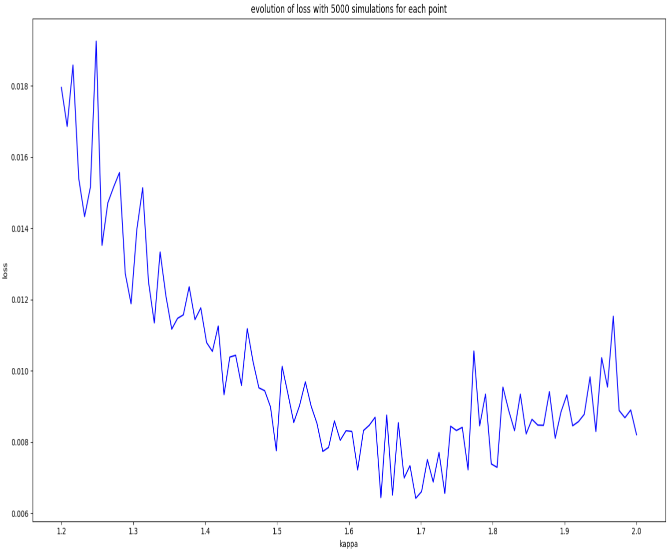
illustration of theorem 3 with 100 simulations
strength needed for the additional player in order to win in 70% cases (red = disparate teams, blue = uniform teams)
number of players
0.9
1.0
1.1
1.2
1.3
1.4
20
40
60
80
100
120
140
160

Recherche de paramètres optimaux

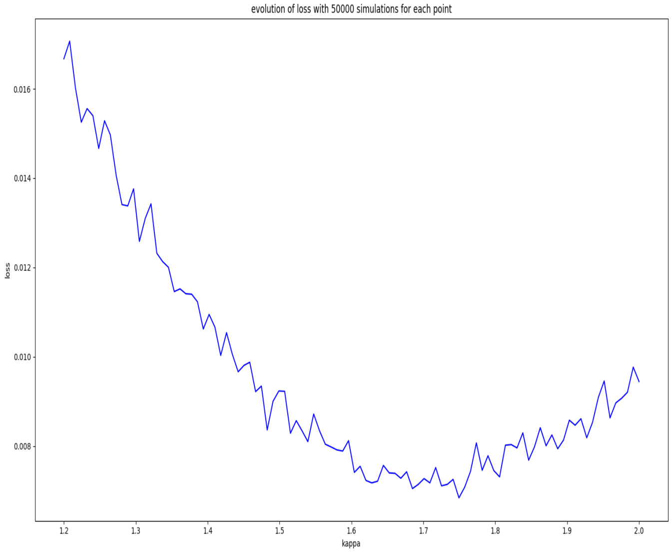
Idée: les bookmakers évaluent bien les probabilités pour les grandes équipes mais pas pour les plus petites.

$\Rightarrow$ on va se fixer nos paramètres sur les valeurs des bookmakers.

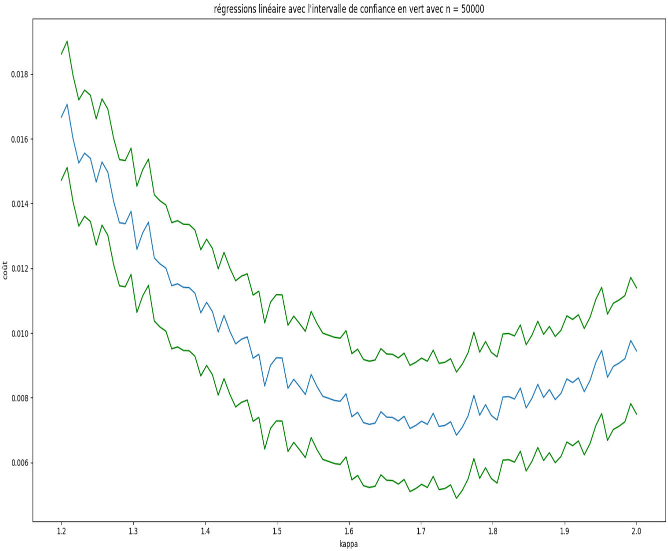
Un premier tracé de la fonction de coût



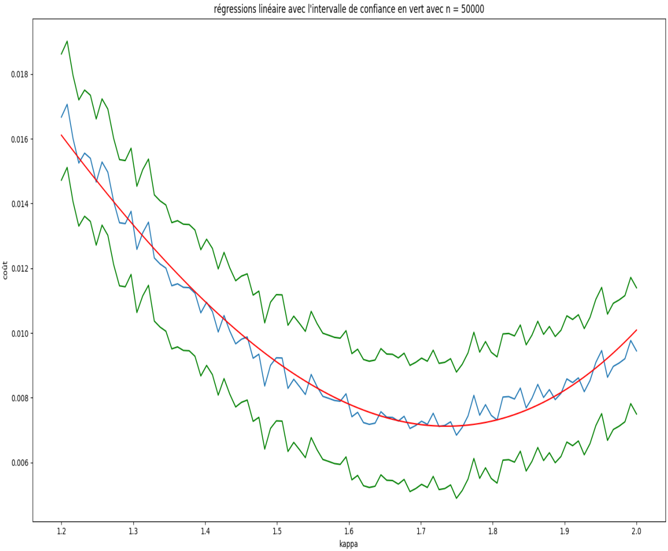
Un deuxième tracé plus précis



Intervalle de confiance



Interpolation



Conclusion partielle

- Minimum de la fonction interpolée $\kappa = 1.690$
- injection dans le modèle. Victoire de Leicester: $0.20\% \pm 0.013\%$
- côte à 5000 $\rightarrow$ une espérance de 90 euros pour 10 euros misés
- 1000 évènements $\rightarrow$ 14% de chances de tout perdre et $\mathrm{E}[gain] = 90\ 000$

Gradient descent ?

- Une solution généralement préférée

Gradient descent ?

- Une solution généralement préférée
- Trop de bruit

Gradient descent ?

- Une solution généralement préférée
- Trop de bruit
- Pas de reproductibilité des résultats

Gradient descent ?

- Une solution généralement préférée
- Trop de bruit
- Pas de reproductibilité des résultats
- $\rightarrow$ on rejette les résultats obtenus

Conclusion

Conclusion

- La nécessité des décalages en probabilité

Conclusion

- La nécessité des décalages en probabilité
- Une théorie intéressante, pas toujours applicable au football

Conclusion

- La nécessité des décalages en probabilité
- Une théorie intéressante, pas toujours applicable au football
- Une optimisation des paramètres très coûteuse en temps de calcul

Conclusion

- La nécessité des décalages en probabilité
- Une théorie intéressante, pas toujours applicable au football
- Une optimisation des paramètres très coûteuse en temps de calcul
- Le football un problème encore ouvert, important pour de nombreux acteurs économiques

Questions?