

Projet 1

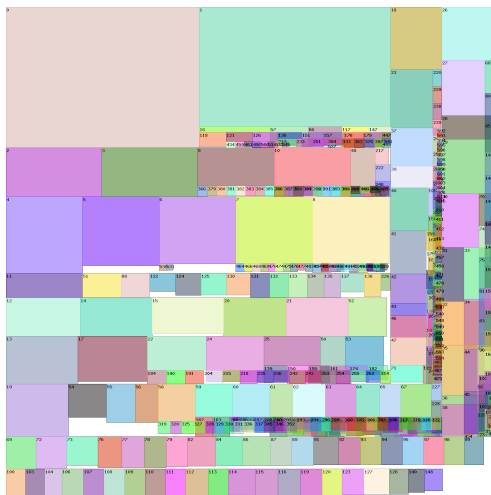
Bin packing avec paquets aléatoires

sujet proposé par Lucas Gerin

lucas.gerin@polytechnique.edu

Mots-clés : Algorithmes probabilistes, Calculs de densités, Convergence en loi, Probabilités conditionnelles.

Le problème du *Bin packing* est un problème célèbre en Algorithmique et en Optimisation combinatoire. Il s'agit de ranger des paquets en utilisant un nombre minimum de boîtes. Nous allons étudier certains aspects du *Bin packing* en dimension 1 lorsque les tailles de paquets sont aléatoires.



La solution approchée d'un problème de Bin Packing en dimension 2. Source : <https://stackoverflow.com/questions/19673588/>.

1.1 Bin packing : Définitions et notations

On considère une suite de boîtes B_1, B_2, \dots de taille 1 et une suite de paquets aléatoires de tailles X_1, X_2, \dots dans $[0, 1]$. Le problème du *Bin packing* (en dimension 1) consiste à affecter chaque paquet à une boîte, de la façon la plus compacte possible.

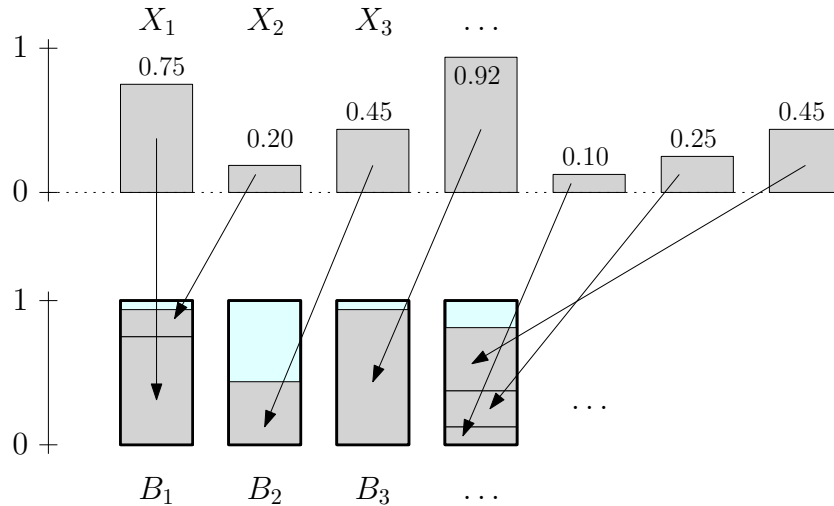


Figure 1.1 – Un exemple d'affectation selon la stratégie *Prochaine-Qui-Convient* avec $X_1 = 0.75$; $X_2 = 0.20$; $X_3 = 0.45$; ... On a $\phi(1) = 1, \phi(2) = 1, \phi(3) = 2, \dots$ et $L_1 = X_1 + X_2 = 0.95, L_2 = X_3 = 0.45$. On voit sur cet exemple que cette stratégie n'est pas du tout optimale : on pourrait affecter $X_5 = 0.10$ à la boîte B_2 pour gagner de la place.

Formellement, on cherche une fonction $\phi : \mathbb{N}^* \rightarrow \mathbb{N}^*$ qui vérifie, pour toute boîte B_j ,

$$L_j := \sum_{i \geq 1, \phi(i)=j} X_i \leq 1.$$

Si $\phi(i) = j$ on considère que la paquet X_i est affecté à la boîte B_j . La quantité L_j représente le remplissage de la boîte B_j . Enfin, on pose \mathcal{N}_j le nombre de paquets dans B_j :

$$\mathcal{N}_j = \sum_{i \geq 1} \mathbf{1}_{\phi(i)=j}.$$

En général il est très difficile de trouver la meilleure fonction ϕ (celui qui permet un rangement le plus compact possible). Ici on va considérer une stratégie simple (mais pas optimale) appelée stratégie *Prochaine-Qui-Convient*. On affecte les paquets dans l'ordre où ils arrivent, et on remplit également les boîtes dans l'ordre. L'idée est de remplir chaque boîte au maximum, puis considérer la boîte suivante. Formellement :

1. On affecte X_1 à B_1 : $\phi(1) = 1$. On pose $j = 1$.
2. Pour chaque $i \geq 2$ il y a deux cas selon que l'on peut ranger X_i dans B_j ou pas :
 - Si $X_i + \sum_{i' < i; \phi(i')=j} X_{i'} \leq 1$ on affecte X_i à B_j : $\phi(i) = j$.
 - Si $X_i + \sum_{i' < i; \phi(i')=j} X_{i'} > 1$ on affecte X_i à B_{j+1} : $\phi(i) = j + 1$. On pose $j = j + 1$.

Dans tout ce projet on va considérer un modèle aléatoire dans lequel les (X_i) sont des variables aléatoires indépendantes et de même loi. Les questions que l'on se pose sont par exemple :

- Que peut-on dire du nombre de paquets dans la boîte B_1 ? dans la boîte B_j ?
- Que peut-on dire du taux de remplissage L_1 ? de L_j ?
- Combien faut-il de boîtes pour ranger n paquets, pour n grand?

1.2 Paquets avec loi continue uniforme.

Dans cette section, les X_i sont supposées suivre la loi uniforme sur l'intervalle $[0, 1]$, on ne considère pour l'instant que la stratégie *Prochaine-Qui-Convient*.

Remplissage de la boîte B_1

S1. Simuler $S = 100000$ fois la variable \mathcal{N}_1 et afficher la moyenne des résultats.

S2. Simuler $S = 100000$ fois la variable L_1 , et afficher les résultats dans un histogramme normalisé¹ avec 50 bâtons. Afficher sur le même graphique la courbe $x \mapsto xe^x$.

(Si vos simulations sont correctes l'histogramme devrait être très proche de la courbe.)

On va maintenant confirmer par la théorie ces deux résultats expérimentaux. Les preuves vont reposer sur une formule assez simple à propos de la densité de $S_n = X_1 + X_2 + \dots + X_n$ pour tout n . Observons tout d'abord que pour tout $n \geq 1$ la variable aléatoire S_n , qui est à valeurs dans $[0, n]$, admet une densité f_n . En effet c'est la somme de variables aléatoires indépendantes à densité.

T1. [Preliminaire] En utilisant la formule de convolution (voir le photocopié de MAP361), démontrer par récurrence que pour tout $n \geq 1$ et tout $x \in [0, 1]$,

$$f_n(x) = \frac{x^{n-1}}{(n-1)!}. \quad (\star)$$

(Pour $x > 1$ l'expression de $f_n(x)$ est bien plus compliquée et n'est pas utile pour ce projet.)

T2. Justifier que pour tout $k \geq 1$,

$$\mathbb{P}(\mathcal{N}_1 \geq k) = \mathbb{P}(S_k \leq 1).$$

En utilisant la formule (\star) , en déduire la loi de \mathcal{N}_1 , puis son espérance. Vérifier que votre calcul est cohérent avec votre résultat expérimental obtenu en S1.

T3. On admet que la variable aléatoire L_1 admet une densité sur l'intervalle $[0, 1]$, que l'on va noter f_L . On admet également que f_L est continue. Soit $x \in [0, 1[$ et $k \geq 1$, démontrer que si $\varepsilon > 0$ est suffisamment petit alors

$$\mathbb{P}(S_k \in [x, x+\varepsilon], X_{k+1} > 1-x) \leq \mathbb{P}(L_1 \in [x, x+\varepsilon], \mathcal{N}_1 = k) \leq \mathbb{P}(S_k \in [x, x+\varepsilon], X_{k+1} > 1-x-\varepsilon). \quad (\#)$$

T4. Démontrer que pour $x \in [0, 1[$ et $\varepsilon > 0$ suffisamment petit

$$x(e^{x+\varepsilon} - e^x) \leq \int_x^{x+\varepsilon} f_L(u) du \leq (x+\varepsilon)(e^{x+\varepsilon} - e^x).$$

En déduire que $f_L(x) = xe^x$.

¹Cela signifie que l'aire totale des bâtons vaut 1. Pour afficher la liste L dans un histogramme normalisé avec 50 bâtons on peut utiliser la commande `plt.hist(L, bins=50, density=True)`.

1.2.1 Dépendance des boîtes

On cherche à illustrer le fait que le remplissage des deux premières boîtes n'est pas indépendant, d'abord théoriquement puis expérimentalement.

T5. Démontrer mathématiquement que les variables L_1 et L_2 ne sont pas indépendantes.

S3. Pour chercher à quantifier cette dépendance, simuler $S = 100000$ couples de variables aléatoires $(L_1^{(1)}, L_2^{(1)}), \dots, (L_1^{(S)}, L_2^{(S)})$, où chaque couple est indépendant et a même loi que (L_1, L_2) . Afficher la quantité

$$C_S = \overline{L_1 L_2} - \overline{L_1} \times \overline{L_2} := \frac{1}{S} \sum_{s=1}^S L_1^{(s)} L_2^{(s)} - \frac{1}{S} \sum_{s=1}^S L_1^{(s)} \frac{1}{S} \sum_{s=1}^S L_2^{(s)}.$$

(Cette quantité est appelée covariance empirique.) Comment peut-on interpréter le signe de C_S ?

1.2.2 Remplissage de la boîte B_j , pour j grand.

Les boîtes ne sont pas remplies de façon homogène : on peut se convaincre que L_1, L_2, \dots n'ont pas la même loi. Il est difficile d'étudier théoriquement la variable aléatoire L_j pour j quelconque mais il a toutefois été démontré² que lorsque $j \rightarrow +\infty$,

$$L_j \xrightarrow{\text{(loi)}} \mathcal{L}, \quad (\$)$$

où \mathcal{L} est une variable aléatoire sur $[0, 1]$, de densité $x \mapsto 3x^2 \mathbf{1}_{0 < x < 1}$. Nous allons chercher à illustrer expérimentalement ce résultat. (Il se trouve que la convergence dans (\$) est très rapide donc on peut illustrer ce résultat avec j assez petit, par exemple $j = 20$.)

S4. Simuler $N = 100000$ fois la variable L_j pour $j = 20$, et afficher les résultats dans un histogramme normalisé avec 50 bâtons. Afficher sur le même graphique la courbe $x \mapsto 3x^2$. (Si vos simulations sont correctes l'histogramme devrait être très proche de la courbe.)

1.2.3 Comparaison expérimentale avec deux autres stratégies

Nous allons étudier expérimentalement deux autres stratégies pour classer N paquets :

Stratégie Prochaine-Qui-Convient-Décroissant

On trie d'abord X_1, \dots, X_N dans l'ordre décroissant. Ensuite on applique la stratégie Prochaine-Qui-Convient. Voir l'exemple en Fig.1.2.

Stratégie Gloutonne

On affecte chaque paquet X_i à la boîte la plus à gauche parmi celles qui ont assez de place libre. Voir l'exemple en Fig.1.3.

²E. G. Coffman Jr., M. Hofri, Kimming, So., AC. Yao. A Stochastic Model of Bin Packing. *Inf. and Control*, n.44,105-115, (1980).

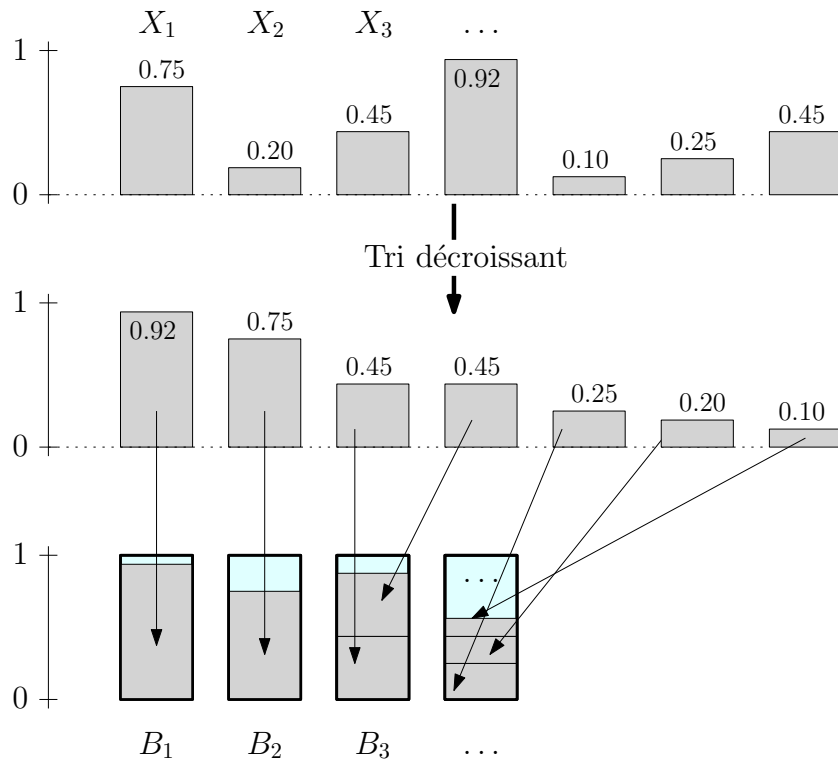


Figure 1.2 – L'affectation selon la stratégie *Prochaine-Qui-Convient-Décroissant*, avec les mêmes données que précédemment.

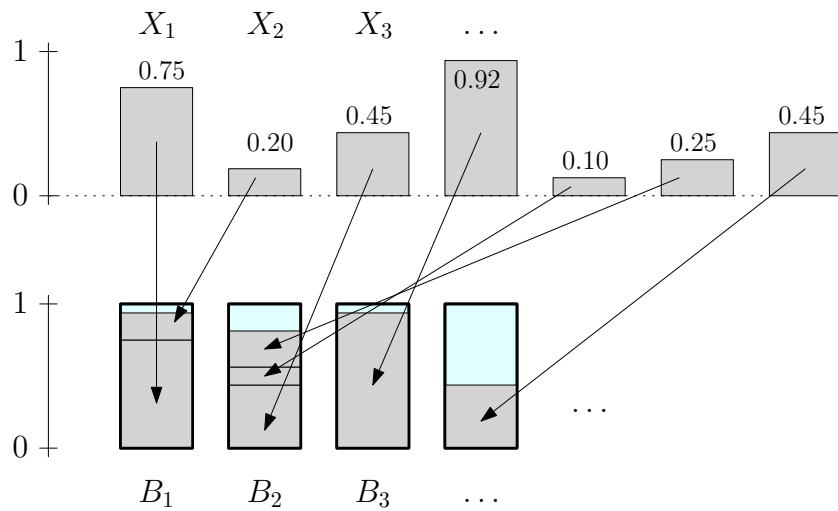


Figure 1.3 – L'affectation selon la stratégie *Gloutonne*, avec les mêmes données que précédemment.

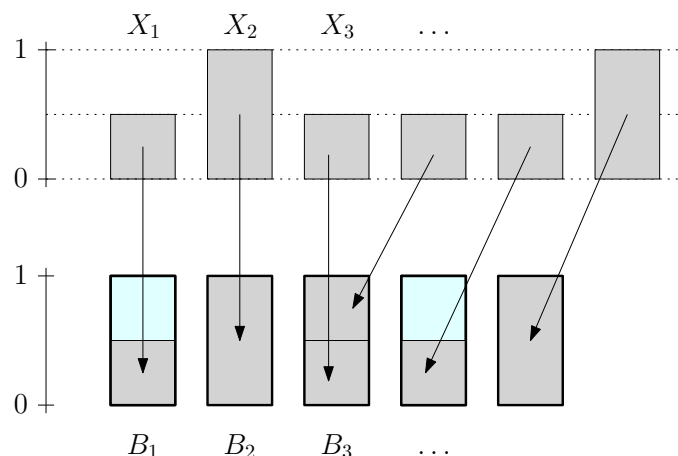


Figure 1.4 – Un exemple d'affectation selon la stratégie *Prochaine-Qui-Convient* avec des paquets discrets.

S5. Pour chaque $N \in \{10, 20, 30, 40\}$, simuler $S = 1000$ fois les variables aléatoires $\phi(N)$ (c'est donc le nombre de boîtes requises pour ranger les N paquets) pour les trois stratégies *Prochaine-Qui-Convient*, *Prochaine-Qui-Convient-Décroissant* et *Gloutonne*. Pour chaque stratégie et chaque N , calculer la moyenne empirique des S simulations.

Afficher les 12 points obtenus sur un même graphique (en mettant N en abscisse). Conclure sur quelle est la meilleure stratégie.

(Pour trier une liste L dans l'ordre croissant on peut utiliser `np.sort(L)`. Ensuite on peut parcourir la liste dans le sens inverse avec `L[::-1]` pour obtenir la suite dans l'ordre décroissant.)

1.3 Paquets discrets

Dans cette partie du projet on ne considère que la stratégie *Prochaine-Qui-Convient*.

À partir de maintenant on suppose que les (X_i) sont discrets et à valeurs dans $\{1/2, 1\}$ (la suite est toujours i.i.d.). Les remplissages des boîtes L_1, L_2, \dots sont ainsi à valeurs dans $\{1/2, 1\}$. Soit $p \in [0, 1]$ tel que

$$\mathbb{P}(X_1 = 1) = p = 1 - \mathbb{P}(X_1 = 1/2).$$

(Voir un exemple en Fig.1.4.) Pour cette loi très simple il est possible de démontrer des formules exactes et de les comparer aux simulations.

S6. [Remplissage moyen] Pour $S = 10000$, $N = 100$ et chaque $p \in \{0; 0.1; 0.2; 0.3; \dots 0.9; 1\}$, simuler S fois la variable aléatoire $(L_1 + \dots + L_N)/N$. Calculer pour chaque p la moyenne empirique obtenue et afficher les 11 points sur un même graphique (en mettant p en abscisse).

T6. Démontrer que pour tout j ,

$$\begin{aligned} \mathbb{P}(L_{j+1} = 1/2 \mid L_j = 1/2) &= 0, \\ \mathbb{P}(L_{j+1} = 1/2 \mid L_j = 1) &= p(1 - p). \end{aligned}$$

T7. En déduire l'expression de $\mathbb{P}(L_j = 1)$ pour tout p et tout j . Calculer le remplissage moyen asymptotique

$$g(p) := \lim_{N \rightarrow +\infty} \frac{1}{N} \mathbb{E}[L_1 + \cdots + L_N]. \quad (\S)$$

Pour vérifier votre calcul, vous pouvez tracer sur la même courbe que la question S6 la courbe $p \mapsto g(p)$.

S7. Il se trouve que non seulement on a la convergence en espérance dans (§) mais on a également la convergence presque-sûre $(L_1 + \cdots + L_N)/N \rightarrow g(p)$. Confirmer par une simulation le fait que la convergence est presque-sûre.

(On se limitera à un seul p fixé. À vous de choisir p , N , le nombre de simulations et le type de simulation appropriés pour illustrer la convergence.)

T8. (Bonus) En utilisant les questions précédentes, faites une conjecture (argumentée) pour la limite $\lim_N \frac{1}{N} \phi(N)$ en fonction de p . (On ne demande pas une preuve rigoureuse, vous pouvez vérifier votre formule à l'aide de simulations.)