Analyse d'algorithmes

Nicolas Pouyanne

Université d'été de St Flour, 26 août 2009

Notes de conférence

1 Introduction : analyse d'algorithmes

Objet : en termes très généraux, l'analyse d'un algorithme est l' évaluation de son coût, sous des formes diverses comme par exemple le temps de calcul ou le nombre d'opérations (vision très grossière).

Exemple du tri de clefs ordonnées dans un arbre binaire de recherche, base du fameux algorithme Quicksort et de ses variantes. Insertion d'une nouvelle clef dans l'arbre ou recherche d'une clef déjà insérée. Le modèle pour l'exemple : on insère une suite $(k_n)_{n\geq 1}$ de clefs qui sont des nombres réels de l'intervalle $[0,1]$. Développer la croissance de l'arbre sur l'exemple $0,3$, $0,1$, $0,4$, $0,15$, $0,9$, $0,2$, $0,6$, $0,5$, *etc.* Le coût d'insertion de la $n^{\text{ième}}$ clef est ici le nombre de comparaisons nécessaires à son insertion, *i.e.* la hauteur de k_n dans l'arbre (on dit plutôt *profondeur*).

Dans le **cas le pire**, les clefs se présentent triées dans l'ordre ($(k_n)_n$ est monotone). Dans ce cas, la profondeur de k_n est n ; on ne peut pas rêver plus cher ! En général, l'analyse dans le cas le pire n'est pas très compliquée mais est bien peu réaliste.

Le stade suivant est l'**analyse en moyenne**. On considère toutes les suites possibles et on calcule la moyenne de leur coût. Quelle pondération adopter ? Introduction, dans le modèle, d'un aléa sur les clefs. Dans le cas de l'arbre binaire de recherche, la suite $(k_n)_n$ est une suite de variables aléatoires indépendantes identiquement distribuées. Autre formulation équivalente (l'équivalence est un peu subtile) : pour tout n, la permutation du groupe symétrique $\mathfrak{S}_n$ qui décrit l'ordre relatif des n premières clefs est tirée avec probabilité $1/n!$. Par des méthodes probabilistes, on démontre le théorème suivant, démontré dans les années 80, qui majore le coût maximal en moyenne.

Théorème 1 *La hauteur moyenne d'un arbre binaire de recherche après insertion de la $n^{\text{ième}}$ clef est équivalente, lorsque n tend vers l'infini, à $c \ln n$ où c est la plus grande des deux solutions de l'équation transcendante*

$$x \ln \frac{x}{2} = x - 1.$$

Le stade suivant est celui de l'**analyse en distribution** : quelle est la probabilité pour que la $n^{\text{ième}}$ clef soit insérée à la profondeur p, pour $p = 0, 1, 2, \ldots$? Par des méthodes de probabilités plus avancées encore, on démontre qu'avec probabilité 1, l'arbre croît sous la forme d'un radis (dessin, développer un peu).

D'une façon générale, trois grandes approches méthodiques se dégagent pour l'analyse d'algorithme : **la combinatoire énumérative et analytique, les probabilités, les méthodes dynamiques**. La problématique dépasse le cadre de l'algorithmique en informatique (par exemple, physique théorique, biologie du génôme ou des populations, mathématiques, *etc*).
Evocation de l'algorithme d'Euclide (et de ses variantes classiques caractérisées par le choix du reste) : système dynamique $T(x) = \{1/x\}$. Méthodes de l'analyse dynamique, intervention de l'opérateur transformateur de densité, propriétés spectrales sur des espaces de fonctions *ad hoc*.

Le présent exposé n'a pas la prétention de dresser un tableau exhaustif des problématiques et méthodes de l'analyse d'algorithme (!). Il consiste au mieux en la présentation d'un aperçu, sous la forme d'un petit échantillon et tend à témoigner des interactions organiques et réciproques entre mathématiques et informatique. En tant qu'objet de recherche en informatique théorique ou en mathématiques, c'est un domaine très riche et en expansion constante. Les aspects numériques (au sens de l'analyse numérique) y interviennent de façon marginale, éventuellement comme outils (approcher ou caractériser des constantes).

Les aspects abordés le seront ici sous le seul l'angle de leur fondement scientifique. Il s'agit ensuite, si l'on veut, de les situer dans une perspective d'une prise en compte de ces interactions pour l'enseignement des mathématiques dans le secondaire, dans les premiers cycles du supérieur (licence, CPGE) et dans la formation des maîtres (amorce d'une réflexion).

Pour la suite de l'exposé, après cette introduction extrêmement générale, le choix a été fait de ne parler que de combinatoire analytique.

2 Combinatoire analytique

Dans une démarche qui s'apparente une modélisation, les algorithmes de l'informatique s'analysent le plus souvent après qu'on les a décrits (représentés) en termes d'objets qui relèvent des mathématiques discrètes. On tombe ainsi, dans le champ mathématique, sur l'étude de structures combinatoires aléatoires. Une référence historique incontournable : le travail de D. Knuth.

Les objets de prédilection de l'aléa discret sont les mots dans un alphabet donné, les arbres graphes et cartes, les permutations, les partitions d'entiers, *etc.*
L'objet a une taille (par exemple, le nombre de sommets d'un graphe). On lui associe le paramètre étudié (par exemple, le nombre de composantes connexes). La question est celle de l'asymptotique du paramètre lorsque la taille tend vers l'infini (équivalent ou développement asymptotique).
Voir le nouveau livre de P. Flajolet et R. Sedgewick, *Analytic combinatorics*.
Exemple : lorsque n tend vers l'infini, évaluer la probabilité qu'une permutation uniforme de $\mathfrak{S}_n$ soit un carré, un cube, *etc.*

2.1 Méthode symbolique

Un exemple un peu brutal : les arbres binaires (planaires, enracinés ; formes d'arbres binaires). Dessin des arbres à 1, 2, 3 nœuds internes. On note

$$C(z) = \sum_{n\geq 0} \#\{\text{arbres binaires à } n \text{ nœuds internes}\} z^n = \sum_{n\geq 0} C_n z^n,$$

série (formelle) génératrice ordinaire. Les premiers dessins fournissent $C(z) = 1 + z + 2z^2 + 5z^3 + \dots$ Un arbre est un arbre vide ou une racine et deux sous-arbres (un à gauche, un à droite). Cette décomposition de la classe combinatoire $\mathcal{C} = \square \cup (\mathcal{C}, \bullet, \mathcal{C})$ mène à l'équation $C = 1 + zC^2$. La quadrature impose

$$C(z) = \frac{1 - \sqrt{1-4z}}{2z}$$

(la racine conjuguée contient un terme en z^{-1} et des coefficients négatifs). Ainsi,

$$C_n = \frac{1}{n+1}\binom{2n}{n}$$

est le fameux nombre de Catalan.

Ce genre de raisonnement est rigoureusement validé et se généralise en la *méthode symbolique*. Les outils principaux des preuves sont mathématiquement peu exigeants : ce sont pour l'essentiel les règles de calcul dans l'anneau des séries formelles et la formule du multinôme. Aperçu rapide de la méthode symbolique. Deux types de séries génératrices : les ordinaires ($\sum_n A_n z^n$) qui comptent les objets non étiquetés et les exponentielles ($\sum_n A_n z^n/n!$) qui comptent les objets étiquetés (le $n!$ correspond aux permutations des n étiquettes possibles). A une algèbre d'opérations symboliques du côté des classes combinatoires (union disjointe $\mathcal{A}+\mathcal{B}$, produit cartésien $\mathcal{A}\times\mathcal{B}$, séquences $SEQ(\mathcal{A}) = \{\emptyset\}+\mathcal{A}+\mathcal{A}^2+\mathcal{A}^3+\dots$,

cycles $CYC(\mathcal{A}) = (SEQ(\mathcal{A}) \setminus \{\emptyset\})/permutations\ cycliques$, multi-ensembles $MSET(\mathcal{A}) = SEQ(\mathcal{A})/permutations$, ensemble de parties, $PSET(\mathcal{A})$ *etc*) correspond une algèbre d'opérations sur les séries génératrices ordinaires ou exponentielles ($A+B$, AB, $1/(1-A)$, $\sum_{k\geq 1} \frac{\varphi(k)}{k} \log \frac{1}{1-A(z^k)}$ (cas non étiqueté, φ est la fonction d'Euler) ou $\log \frac{1}{1-A}$ (cas étiqueté), $\exp \sum_{k\geq 1} \frac{(-1)^{k-1}}{k} A(z^k)$ (cas non étiqueté) ou $\exp(A)$ (cas étiqueté), *etc*). En transposant des considérations combinatoires – le plus souvent d'ordre récursif – qui caractérisent la classe combinatoire étudiée, on obtient des propriétés des séries génératrices : formes explicites, équations implicites, équations différentielles ou intégrales, équations fonctionnelles, *etc.*

Exemple des permutations carrées ou puissances m (cadre étiqueté, forme explicite, produit infini). Une permutation $\sigma \in \mathfrak{S}_n$ est un carré ssi pour tout entier naturel l, dans la décomposition de σ en cycles disjoints, le nombre de cycles de longueur $2l$ est pair (c'esztélémentaire). Ainsi, forme d'un carré : des cycles impairs, un nombre pair de 2-cycles, un nombre pair de 4-cycles, *etc*). Le cadre est celui d'une classe combinatoire étiquetée (car $(12)(34) \neq (13)(24)$). Sur les fonctions génératrices exponentielles, la méthode symbolique fournit

$$P_2(z) = e^z \operatorname{ch} \frac{z^2}{2} e^{z^3/3} \operatorname{ch} \frac{z^4}{4} e^{z^5/5} \cdots = \sqrt{\frac{1+z}{1-z}} \prod_{k\geq 1} \operatorname{ch} \frac{z^{2k}}{2k}.$$

Cas des puissances m : note

$$l^\infty \wedge m := \lim_{k\to\infty} l^k \wedge m$$

et

$$e_d(z) = \sum_{k\geq 0} \frac{z^{kd}}{(kd)!} = \frac{1}{d} \sum_{\zeta,\ \zeta^d=1} \exp(\zeta z)\ ;$$

on trouve dans la même veine la forme explicite

$$P_m(z) = \prod_{k|m} (1-z^k)^{-\mu(k)/k} \prod_{k\geq 1,\ k\wedge m\neq 1} e_{k^\infty \wedge m}\left(\frac{z^k}{k}\right)$$

où μ est la fonction de Möbius.

Exemple des permutations zig-zag (équation différentielle, pour interprétation combinatoire et analyse des singularités des nombres de tangente). Une permutation $\sigma \in \mathfrak{S}_{2n+1}$ est dite *zig-zag* lorsqu'elle vérifie $\sigma(1) < \sigma(2) > \sigma(3) < \sigma(4) \cdots > \sigma(2n+1)$. La construction récursive de ces permutations peut se décrire à partir de leur maximum :

$$\mathcal{T} = (e) \cup (\mathcal{T}, \max, \mathcal{T})$$

(ces permutations se décrivent simplement à l'aide d'une bijection devenue usuelle avec les arbres décroissants). Cette relations sur les classes combinatoires se traduit sur les fonctions génératrices exponentielles par l'équation intégrale

$$T(z) = z + \int_0^z T^2(x)dx,$$

équivalente à

$$T'(z) = 1 + T^2(z) \text{ et } T(0) = 0.$$

On résout cette équation différentielle et on trouve

$$T(z) = \sum_{n\geq 0} T_n \frac{z^n}{n!} = \tan z = 1\frac{z}{1!} + 2\frac{z^3}{3!} + 16\frac{z^5}{5!} + 272\frac{z^7}{7!} + \cdots .$$

Interprétation combinatoire des nombres de tangente.

2.2 Analyse des singularités

Une fois l'algorithme modélisé, l'analyse proprement dite du paramètre A_n d'un objet de taille n consiste à en calculer une asymptotique lorsque n tend vers l'infini (équivalent, ou développement asymptotique). Il s'agit de tirer cette asymptotique des propriétés de la série génératrice $A(z) = \sum_n A_n z^n$ que l'on a elles-mêmes déduites des propriétés récursives de la classe combinatoire.
L'*analyse des singularités* est une méthode récemment systématisée (Flajolet *et al.*), qui consiste à déduire les propriétés asymptotiques de A_n du comportement analytique de la fonction A au voisinage de ses singularités dominantes (les plus proches de l'origine).
Description succincte du principe général : camembert, formule de Cauchy le long de contours édentés, échelle log-puissance. Si camembert et si $A(z) \in O_1(1-z)^\alpha$, alors $A_n \in O_n(1/n^{\alpha+1})$. Il existe des version o et $\sim$ ainsi qu'une version étendue à la comparaison avec les fonctions log-puissance.

Exemple des arbres binaires et des nombres de Catalan. La singularité dominante de $C(z) =$ est $1/4$ et $C(z) = 2 - 4\sqrt{1/4 - z} + \cdots$ au voisinage de $z = 1/4$. Cela fournit l'équivalent

$$C_n \sim_\infty \frac{4^n}{\sqrt{\pi n^3}}$$

On peut développer plus loin, à n'importe quel ordre. Ce résultat se déduit aussi de la formule de Stirling à partir de la forme explicite de C_n. Le $\sqrt{\pi}$ vient de $1/\Gamma(-1/2) = -1/(2\sqrt{\pi})$ et le 4^n de la singularité en $1/4$. Naturellement, la valeur

$C(0) = 2$ n'apporte aucune contribution à l'asymptotique (une fonction constante est analytique !).

Exemple des permutations zig-zag. La fonction $T(z) = \tan z$ a deux singularités dominantes en $\pm\pi/2$. En chacune d'elles,

$$T(z) \sim_{\pm\frac{\pi}{2}} \frac{8z}{\pi^2 - 4z^2}.$$

Les contributions en les deux singularités s'ajoutent et on obtient par théorème de transfert la proportion asymptotique des permutations zig-zag (et l'asymptotique des nombres de tangente) :

$$\frac{T_n}{n!} \sim_\infty 2.\left(\frac{2}{\pi}\right)^{n+1}.$$

2.3 Autres méthodes d'analyse asymptotique

Classiquement Darboux : si f, analytique dans le disque ouvert, est de classe $\mathcal{C}^s$ sur le cercle, alors $[z^n]f(z) \in o(1/n^s)$. Permet aussi de trouver des développements asymptotiques, y compris lorsque le cercle est une frontière naturelle (exemple : $\sum z^{2^n}/2^{nr}$, où $r \geq 1$). Exemple : la série génératrice exponentielle des graphes 2-réguliers (expliquer) est $e^{-z/2-z^2/4}/\sqrt{1-z}$. D'où une proportion asymptotique $\sim_n e^{-3/4}/\sqrt{\pi n}$.

Tauber. Exemple : permutations aux cycles de longueurs toutes distinctes, série génératrice exponentielle $\prod_n(1 + z^k/k)$, terme général équivalent à $e^{-\gamma}$ où γ est la constante d'Euler.

Hybride. Exemple des permutations puissances m. Ré-écriture pour les carrés, en développant le cosinus hyperbolique sur le disque ouvert et en ré-ordonnant les termes :

$$P_2(z) = \sqrt{\frac{1+z}{1-z}} \prod_{n\geq 1} \exp\left(\frac{(-1)^{n-1}\tau_{n-1}}{n2^{2n+1}} \operatorname{Li}_{2n}(z^{4n})\right)$$

où $\tau_p = T_{2p+1}/(2p+1)!$ est le $p^{\text{ième}}$ nombre de tangente. Cas de frontière naturelle d'analyticité, échec de Darboux et de l'analyse des singularités. Hybride. On obtient pour les carrés la proportion asymptotique $\sim_n \pi_2/\sqrt{n}$ où la constante π_2 est explicite (elle s'exprime avec les nombres de tangentes et la fonction ζ de Riemann en les entiers pairs) Enoncé du résultat général des puissances m (équivalent et existence du développement asymptotique) : la proportion asymptotique est

$$\sim_\infty \frac{\pi_m}{n^{1-\varphi(m)/m}}$$

où π_m est également explicite et se décrit comme la limite d'une série numérique qui converge rapidement (vitesse géométrique).

Méthode du col. Exemple à la proportion des involutions de $\mathfrak{S}_n$. La série génératrice exponentielle des involutions est $e^{z+z^2/2}$, elle est entière (singularité en l'infini). La méthode du col fournit la proportion asymptotique

$$\sim_\infty \frac{e^{-1/4}}{2\sqrt{\pi n}} n^{-n/2} e^{n/2+\sqrt{n}}.$$

3 Conclusion

Cette présentation n'est qu'un survol, à peine ébauche.
Richesse (épistémologique) du sujet et place dans la recherche à l'interface des mathématiques et de l'informatique théorique.
Conséquences sur l'enseignement : réflexion à prévoir. Place de l'algorithmique (quelle algorithmique ?), des structures combinatoires, du langage de l'aléa, du calcul formel, de l'analyse complexe, de la géométrie, *etc* dans les contenus enseignés en mathématiques à tous les niveaux, y compris bien sûr en formation des maîtres.

Nicolas Pouyanne,
Laboratoire de Mathématiques de Versailles,
CNRS, UMR 8100,
Université de Versailles - St-Quentin,
45, avenue des Etats-Unis, 78035 Versailles CEDEX (France)
pouyanne@math.uvsq.fr
http://www.math.uvsq.fr/∼pouyanne/