Modèles de calcul, Complexité, Approximation et Heuristiques

Modèle probabiliste: Algorithmes et Complexité

Jean-Louis Roch

Master-2 Mathématique – Informatique

Grenoble-INP – UJF Grenoble University, France

Techniques d'analyse

Plan du cours :

- Retour sur borne inférieure. Principe Min-Max.
- Projection homomorphique
  - sur les polynômes : Schwartz-Zippel (produit de matrices, primalité)
  - sur les entiers : Word-count
- Marche aléatoire. Lien avec chaîne de Markov. 2-SAT.
- Inégalité de Markov. Application à MAX-SAT
- Bornes de Chernoff. Applications : tri, médiane.

Projection homomorphique

- projection dans un espace plus petit, par hachage
- prob(erreur) = probabilité de collisions
- Schémas génériques de projection

- Evaluation d'un polynôme en une abscisse aléatoire (Schwartz-Zippel)
- Calcul modulo un entier [premier] aléatoire

Hachage modulo un nombre premier

Lemme de projection par hachage

Soit $1 \leq N \leq 2^B$ un entier non nul. Soit α un entier. Soit p un nombre premier choisi uniformément parmi les $\pi(\alpha)$ nombre premiers inférieurs à α.

$$\Pr(N \mod p = 0) \leq \frac{B}{\pi(\alpha)} \leq \frac{B \cdot \log_e \alpha}{\alpha}.$$

Applications

- recherche de chaine de caractères dans un fichier
- WORD-COUNT : recherche du nombre d'ocurence de k mots de m bits dans un fichier de taille n en temps $O(n)$, indépendant de k.

Lemme de Schwartz-Zippel et exemple

Lemme de Schwartz-Zippel

Soit $P \in K[X_1, \ldots, X_n]$ un polynôme non nul à coefficients dans un corps K, à n indéterminées et de degré d.
Soit I un sous-ensemble de K de cardinal $\#I$, soit $(u_1, \ldots, u_n)$ tiré uniformément dans I^n. Alors :

$$\Pr\left(P(u_1, \ldots, u_n) = 0\right) \leq \frac{d}{\#I}.$$

Applications : Preuve probabiliste d'identités algébriques.

- Egalité de deux polynômes.
- Vérification du produit de deux matrices.
- Test de primalité : n est premier ssi $(1+X)^n = 1 + X^n \pmod{n}$.
- Couplage parfait : ensemble d'arcs tel que chaque sommet est atteint une et une seule fois.

Lemme de Schwartz-Zippel et exemple

Test de primalité

Propriété : n est premier ssi $(1+X)^n = 1 + X^n \pmod{n}$.

- Soit $P_n = (1+X)^n - 1 - X^n$. Alors P est nul ssi n est premier.
- Tirer un polynôme Q de degré petit et calculer $R = P_n \bmod Q$. Comment ?
- Vérifier que $R = 0$ en utilisant Schwartz-Zippel.

Couplage parfait

Ensemble d'arcs tel que chaque sommet est atteint une et une seule fois.

- Un couplage parfait existe ssi la matrice antisymétrique de Tutte associé au graphe est inversible ($\det(G) \neq 0$).
  si $(i,j) \notin E$: $T_{i,j} = 0$; sinon si $i < j$: $T_{i,j} = X_{i,j}$ et $T_{j,i} = -X_{i,j}$.

Parcours de graphe

- Soit $G = (V, E)$ un graphe non orienté, connexe, à n sommets et m arcs.
- Question : les 2 sommets s et d sont-ils connectés ?
- Un parcours BFS/DFS à partir de s fait cela en temps linéaire $O(n + m)$ et en espace mémoire $O(n)$. Peut-on faire mieux ?

Algorithme probabiliste dans un graphe

Soit $G = (V, E)$ un graphe non orienté, connexe, à n sommets et m arcs.

```
for( r=s; (r ≠ t) ; r = voisin aléatoire de r ) ;
```


Marche aléatoire

Temps d'atteinte

$H(i, j)$ = temps moyen pour atteindre j à partir de i pour la première fois.

Exemples

- graphe complet : $H(i, j) = n - 1$ $(i \neq j)$
- chaine de 0 à $n - 1$: $H(i, j) = j^2 - i^2$ $(i < j)$. Donc $H(0, n - 1) = (n - 1)^2$.
- graphe général : $H(i, j) = O(n^3)$

Marche aléatoire

Matrice de transition

Associé à une chaîne de Markov $M_G = (X_1, X_2, \ldots)$ de matrice de transition P, avec $P_{i,j} = \frac{1}{d(i)}$.

- M_G est *irréductible* car on peut aller de tout état à tout autre (G est connexe).
- M_G est *apériodique* ssi chacun de ses états v est non-périodique ;
  NB un état v est périodique si
  $\exists \Delta > 1 : \Pr[X_{t+s} = v | X_t = v] = 0$ sauf si s multiple de Δ
  (i.e. partant de v on ne peut revenir à v qu'après un nombre de pas multiples de Δ).

Remarque : si le graphe n'est pas connecté, ou périodique (par exemple bi-partite), il y a problème de convergence

Chaine de Markov

Propriétés

Si M_G est finie, irréductible, apériodique, elle admet une unique distribution stationnaire

- $\Pi = (\Pi_1, \ldots, \Pi_n)$.
  i.e. $\sum_i \Pi_i = 1$ et $\Pi = \Pi \cdot P$.
- Soit $r_{i,j}^k = \Pr[$ atteindre j en k pas à partir de $i]$. Le temps moyen pour aller de i à j est $h_{i,j} = \sum_{t \geq 1} t \cdot r_{i,j}^t$.
- Le temps pour aller de i à i est $h_{i,i} = \frac{1}{\Pi_i}$.

Analyse du temps moyen dans une chaine de Markov

Théorème

Soit M_G de distribution stationnaire unique $\Pi = (\Pi_1, \ldots, \Pi_n)$:

- $\Pi_i = \dfrac{\deg(i)}{2m}$.

  Preuve : il suffit de vérifier que $\Pi \cdot P = \Pi$ et $\sum_i \Pi_i = 1$.
- Si $(i,j) \in E$: $h_{j,i} < 2m$.

  Preuve : on a $h_{i,i} = \frac{2m}{\deg(i)}$ et $h_{i,i} = \sum_{j:(i,j)\in E} \frac{1}{\deg(i)}[1 + h_{j,i}]$. D'où

  $1 + h_{j,i} \leq \sum_{j:(i,j)\in E}[1 + h_{j,i}] = 2m$. Donc $h_{j,i} \leq 2m - 1$.
- Soit C_i = temps moyen pour atteindre tous les sommets à partir de i (*temps de couverture*). Alors $C_i < 4nm$.

  Preuve : Comme G est connexe, il y a un arbre couvrant de racine i, qui possède n arcs (non orientés). En parcourant cet arbre avec retour à la racine (donc 2 fois chaque arc, soit $2n$ arcs), on parcourt tous les sommets. Donc $C_i \leq$ temps moyen de ce parcours $= 2n.2m = 4nm$.

Application : en faisant $K = 8nm = O(n^3)$ pas à partir du sommet s, la probabilité de ne pas atteindre la destination t est (inégalité de Markov) $< \frac{4nm}{8nm} < \frac{1}{2}$

Exemple : 2-SAT

Cas particulier : graphe = chaine

- état à t : $S_t \in \{0, \ldots, n\}$. Transitions :
  $\Pr(S_{t+1} = j + 1 | S_t = j) = \Pr(S_{t+1} = j - 1 | S_t = j) = 2^{-1}$.
- Soit X_j= va qui dénombre le nombre de pas pour aller de j à n :
  $$E[X_j] = \frac{1}{2}(1 + E[X_{j+1}]) + \frac{1}{2}(1 + E[X_{j-1}]).$$
- Posons $h_j = E[X_j]$; on a $h_n = 0$, $h_0 = h_1 + 1$ et
  $h_j = 1 + \frac{1}{2}(h_{j-1} + h_{j+1})$
- Par récurrence, on a : $h_j = h_{j+1} + 2j + 1$
  $h_0 = h_1 + 1$; $h_{j+1} = 1 + \frac{1}{2}(h_j + h_{j+2}) = 1 + \frac{1}{2}(h_{j+1} + 2j + 1 + h_{j+2})$
  d'où $h_{j+1} = h_{j+2} + 2(j + 1) + 1$.
- D'où $h_0 = h_1 + 1 = \sum_{j=1}^{n}(2j - 1) = n(n + 1) - n = n^2$
- Inégalité de Markov :
  $\Pr(X_0 > 2n^2) = 1 - \Pr(X_0 \leq 2n^2) \geq 1 - \frac{1}{2} = \frac{1}{2}$.

Application : PageRank

- A chaque page web, Google associe un nombre caractérisant son importance.
- Le résultat d'une recherche trie les pages par importance décroissante.
- PageRank : calcule l'importance d'une page à partir des liens entre pages :
  - une page A qui pointe vers B : A donne de l'importance à B
  - une page B pointée par beaucoup de pages est importante
  - plus A a de pointeurs, moins chacun a de l'importance
- Equation :

$$R(B) = \sum_{A:A\to B} R(A)/d(A)$$

avec $d(A)$=#pages pointées par A. Attention aux puits !

Les puits accumulent de l'importance sans en donner : l'équation a des solutions avec des poids qu'aux puits, ce qui n'est pas la motivation initiale !

PageRank et marche aléatoire

- Equation corrigée : une faction du poids de chaque page va vers toutes les pages.

$$R(B) = (1-\alpha)/N + \alpha \sum_{A:A\to B} R(A)/d(A)$$

En pratique : $\alpha \simeq 0,85$.

- C'est une marche aléatoire où à chaque page B :
  - avec probabilité α, on suit un lien (aléatoire)
  - avec probabilité $(1-\alpha)$, on va a une page (aléatoire)
- Comment résoudre avec 100 milliards de pages ?
- La marche aléatoire converge vers la distribution stationnaire !
  - Partir d'une distribution initiale arbitraire.
  - Faire quelques itérations : le résultat est assez proche de la distribution stationnaire
    (pour Google, 50 à 100 itérations en quelques jours)

Dénominations communes...

Définition

Un évènement E se produit

- *surely* (ou *est vrai*) ssi $\complement_E = \emptyset$;
- *almost surely* (a.s.) ssi $\Pr(E) = 1$;
- *with overwhelming probability* (w.o.p.) ssi $\forall A > 0, \exists c_A > 0$:

$$\Pr(E) \geq 1 - \frac{c_A}{n^A}$$

  – par exemple : $\Pr(E) \geq 1 - e^{cn}$ avec c constante – ;
- *with high probability* (w.h.p.) ssi $\exists c > 0 : \Pr(E) \geq 1 - n^{-c}$;
- *asymptotically almost surely* (a.a.s.) ssi $\lim_{n\to\infty} \Pr(E) = 1$.

Formules et encadrement utiles

Formules utiles

Stirling : $n! \simeq \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$.
Coefficients du binôme : $\left(\frac{n}{k}\right)^k \leq C_n^k = \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$.
Approximation de Poisson : $\left(1 - \frac{a}{n}\right) n \simeq e^{-a}$.

Inégalités de Markov, Hoeffding, Bienaymé-Tchebychev

- Markov : soit Z une v.a. réelle presque sûrement positive ou nulle :

$$\forall a > 0 : \Pr(Z \leq a) \leq \frac{E[Z]}{a}.$$

- Majoration de l'écart à la moyenne d'une somme de va Bernoulli X_k indépendantes avec $\Pr(X_k = 1) = 1 - \Pr(X_k = 0) = p$. Soit $S_n = \sum_{k=1}^n X_k$, alors :
  $\Pr\left(|S_n - E[S_n]| \geq x\sqrt{n}\right) \leq 2\exp\left(-2x^2\right)$ (par inégalité de Hoeffding)
  $\Pr\left(|S_n - E[S_n]| \geq x\sqrt{n}\right) \leq \frac{p(1-p)}{x^2}$ (par Bienaymé-Tchebychev)

Bornes de Chernoff

Variable binomiale et Bornes de Chernoff

Soit $X_1, \dots, X_n$ n variables de Bernouilly, et soit $X = \sum_{i=1}^{n} X_i$.
Soit p la probabilité de succès et $q = 1 - p$ celle d'échec.

- $\Pr\{X \geq m\} = \sum_{j=m}^{n} C_n^j p^j q^{n-j}$ (queue de la distribution)
- $\forall 0 < \epsilon < 1 : \begin{cases} \Pr\{X \leq (1-\epsilon)pn\} & \leq exp\left(-\frac{1}{2}\epsilon^2 np\right) = e^{-\epsilon^2 np/2} \\ \Pr\{X \geq (1+\epsilon)pn\} & \leq exp\left(-\frac{1}{3}\epsilon^2 np\right) = e^{-\epsilon^2 np/3} \end{cases}$

Autre formulation

- Soit $X_1, \dots, X_n$ des v.a. indepéndantes discrètes telles que $\forall i : E[X_i] = 0$ et $|X_i| \leq 1$. Soit $X = \sum_i X_i$ et σ^2 la variance de X. Alors, $\forall \lambda, 0 \leq \lambda \leq 2\sigma$: $\Pr[|X| \geq \lambda\sigma] \leq 2e^{-\lambda^2/4}$.

Illustration : estimation de la moyenne. Dans une population de N éléments, $p \cdot N$ vérifient une propriété (p est inconnu). On fait n tirages indépendants ; soit X le nombre d'éléments tirés vérifiant la propriété. Alors l'estimation $\frac{X}{n}$ de p vérifie : $\Pr\left\{\frac{X}{n} = p(1 \pm \epsilon)\right\} \geq 1 - 2e^{-\epsilon^2 np/2}$.

MAX-3-SAT

MAX-3-SAT

- Entrée : conjonction de m clauses, chacune avec 3 variables $\neq$.
- Sortie : le nombre maximal $k \leq m$ de clauses satisfaisables.

Algorithme probabiliste : choisir une affectation aléatoire !

Soit la va $X_i = 1$ si la clause i est satisfaite, 0 sinon.
$\Pr(X_i = 0) = 1/8$, donc
$E[X_i] = O \cdot \Pr(X_i = 0) + 1 \cdot \Pr(X_i = 1) = 7/8$.
L'algorithme a un ratio d'approximation à 7/8 de l'optimal !
Corollaire $\exists$ une affectation qui satisfait au moins $7m/8$ clauses.

Exercice : on tire une affectation tant que moins de 7m/8 clauses sont satisfaites. Quel est le temps moyen ? Estimer la probabilité d'être proche de cette moyenne (Indication : Chernoff).

Briser la symétrie aléatoirement

Eléments indépendants dans un cycle

- Entrée : une liste circulaire (cycle) $L_0, \ldots, L_{n-1}$ avec Succ(L_i) $= L_{i+1 \mod n}$
- Sortie : un ensemble X d'éléments indépendants de grande taille.

Trivial en séquentiel. Mais en parallèle ?

Algorithme parallèle de temps O(1) par tirage aléatoire

- (en parallèle) associer à chaque L_i une couleur col(L_i) $\in \{0, 1\}$ aléatoire ;
- si (col(L_i) = 1) et col(Succ(L_i)) = 0 alors mettre L_i dans X.

Analyse

$\forall 0 < \alpha < \frac{1}{8}$ et soit $\beta = \frac{(1-8\alpha)^2}{16}$. On a : $\Pr\{|X| \leq \alpha n\} \leq e^{-\beta n}$.
[Chernoff avec $|X| > n/2$ tirages de Bernouilly de probabilité de succès 1/4]

Preuve à faire au tableau :

- Un élément x est sélectionné ssi il est colorié avec 1 et son successeur avec 0 : cet évènement est de probabilité 1/4.
- Pour se ramener à des Bernouilly indépendantes, on se limite à une borne inf sur les éléments x_{2i} d'indice pair.
  Soit X_{2i} la v.a. de Bernouilly de probabilité de succès 1/4 et qui vaut 1 ssi x_{2i} est sélectionné. Il y a n/2 va X_{2i} indépendantes.
  Soit $Y = \sum_{i=0}^{n/2} X_{2i}$ et soit C la va qui dénombre le nombre d'éléments sélectionnés : on a $C \geq Y$.
- Chernoff : $\Pr(Y < (1-\epsilon)n/8) < exp(-\epsilon^2 n/16)$.
  On choisit ϵ en posant $\alpha = (1-\epsilon)/8$ soit $\epsilon = 1 - 8\alpha$: pour $\alpha \in]0, 1/8[$, on a bien $\epsilon > 0$ et aussi $\beta = (1-8\alpha)^2/16 > 0$;
  donc $\forall \alpha \in]0, 1/8[$: $\Pr(C < \alpha n) \leq \Pr(Y < \alpha n) < exp(-\beta n)$.
- Par exemple, avec $\alpha = 1/16$:
  $\Pr(C < n/16) < e^{-64n} < (2.10^{-28})^n$.
  Donc $C > n/16$ w.h.p. (et même très grande probabilité !)

D&C probabiliste : Médiane

Médiane

- entrée : n éléments $a_0, \dots, a_n - 1$ d'un ensemble ordonné.
- sortie : l'élément de rang k (par exemple, $k = n/2$)

Algorithme RandomMedian(i, j, k)

1. si $(j - i == 1)$ retourner a_i ;
2. choisir p au hasard dans $i, \dots, j - 1$
3. $r :=$ segmenter(i, j, p) ; // r est la position de a_p
4. si $r = k - 1$ retourner a_p ;
5. si $rs \geq > k$ retourner RandomMedian(i, r, k) ;
6. sinon retourner $(r + 1, j, k - r - 1)$;

Analyse : Soit X_j le nombre d'éléments à l'étape j.
$\Pr(X_{j+1} \leq 3/4X_j) = 1/2$: donc le nombre moyen d'appels pour diviser par 4/3 est 2. D'où $E[X] = \sum_j E[X_j] \leq \sum_j 2(3/4)^j n = 8n$.

Quicksort probabiliste - Analyse

Borner la probabilité $\text{Prob}^{(e)}_{\text{échec}}(t)$ qu'un élément arbitraire e du tableau à trier soit mal placé après t partitions sur cet élémént.

- Soit n_j la taille du sous-tableau contenant e après j étapes de partition. Alors : $\Pr\left(n_{j+1} \leq \frac{3n_j}{4}\right) \geq \frac{1}{2}$.
  Une telle partition est dite réussie.
- Après $\log_{\frac{4}{3}} n$ partitions réussies, e est correctement placé.
- Soit la v.a. X_t = #partitions réussies : $\Pr^{(e)}_{\text{échec}}(t) = \Pr(X_t \leq \log_{\frac{4}{3}} n)$.
- [Chernoff] choisir ϵ tel que $(1 - \epsilon)pt = \log_{\frac{4}{3}} n$: $\epsilon = 1 - \frac{2}{t} \log_{\frac{4}{3}} n \in]0, 1[$ pour $t > 2 \log_{\frac{4}{3}} n$.
- Finalement, en prenant $t = 5 \log_{\frac{4}{3}} n$, on a :

$$\text{Prob}_{\text{échec}} = n\text{Prob}^{(e)}_{\text{échec}}(5 \log_{\frac{4}{3}} n) < n^{1 - \frac{9}{5} \times 0.86)} < n^{-0.54} < \frac{1}{\sqrt{n}}$$

- Exercice : Reprendre le calcul avec $\Pr\left(n_{j+1} \leq \frac{7n_j}{8}\right) \geq \frac{3}{4}$.

Retour sur le calcul de l'élément médian

Trouver x_n l'élément médian de l'ensemble ordonné $\{x_1, \ldots, x_{2n-1}\}$.

Algorithme

- Choisir Y de taille $2n^{2/3}$;
- Trier Y et prendre $a = Y_{n^{2/3} - n^{1/3} \log n}$ et $b = Y_{n^{2/3} + n^{1/3} \log n}$
- Calculer $Z := X \cap [a, b]$, en comptant $n_a = \#$éléments $< a$.
  Avec une grande probabilité, $x_n \in Z$ et $Z < 2n^{2/3} \log n$ [Chernoff].
- Trier Z et retourner $x_n = z_{n - n_a}$.

Cout moyen : $\frac{3}{2} n + o(n)$ comparaisons !