Expressions booléennes aléatoires Probabilité et complexité de fonctions booléennes

Danièle GARDY

PRiSM, Univ. Versailles-Saint Quentin

avec B. Chauvin, P. Flajolet, H. Fournier, A. Genitrini,
B. Gittenberger, J. Kozik, M. Kuba, A. Woods, M. Zaionc...

GDR-IM

25 Janvier 2008

Un exemple pour commencer...

$((x \vee \bar{x}) \wedge x) \wedge (\bar{x} \vee (x \vee \bar{x}))$

$(x \vee (y \wedge \bar{x})) \vee (((z \wedge \bar{y}) \vee (x \vee \bar{u})) \wedge (x \vee y))$

Probabilité qu'une formule écrite "au hasard", sur n variables booléennes, soit une tautologie (toujours vraie)?

- $n = 1$: 4 fonctions booléennes; $Proba(Vrai) = 0.2886$
- $n = 2$: 16 fonctions booléennes; $Proba(Vrai) = 0.209$
- $n = 3$: 256 fonctions booléennes; $Proba(Vrai) = 0.165$
- $n \to +\infty$: 2^{2^n} fonctions booléennes

$$Proba(Vrai) \sim ?$$

De quoi parlons-nous?

Formule logique, construite à partir de $\vee$, $\wedge$, et des littéraux positifs ou négatifs

Une formule $\sim$ un arbre

Arbre binaire complet, étiqueté:
- Noeuds internes: $\vee$, $\wedge$
- Feuilles: littéraux

Formule/Arbre $\Rightarrow$ **fonction booléenne**

Taille d'une formule: nombre de littéraux/feuilles

Expression écrite “au hasard” $\Rightarrow$ fonction booléenne aléatoire

- Que veut précisément dire “au hasard”?
- Quelle est la **probabilité** d’une **tautologie**? d’un littéral? d’une fonction donnée?
- **Complexité** d’une fonction: taille minimale d’un arbre la représentant.
  Complexité d’une fonction aléatoire?
- Les fonctions les plus probables sont-elles celles de “petite” complexité?
- Et si l’arbre a toutes ses feuilles au même niveau?

Cadre général

Système pour la logique propositionnelle, défini par des règles pour construire les formules:

- Ensemble de connecteurs logiques
- Chaque connecteur a une arité (fixe ou variable), peut être commutatif ou associatif
- On autorise, ou non, les littéraux négatifs.
- On peut demander que tous les littéraux soient à la même profondeur

Cadre général

Formules booléennes: sur un nombre fixe n de variables

B_n: ensemble des fonctions booléennes à n variables

$Card(B_n) = 2^{2^n}$

- Une formule sur n variables détermine une unique fonction de B_n
- Une fonction de B_n est représentée par une infinité de formules booléennes sur (au moins) n variables

Cadre général

Une formule = un arbre

Différents types d'arbres modélisent différentes contraintes

- Noeuds internes étiquetés par les opérateurs logiques
- Noeuds externes étiquetés par les littéraux
- Opérateurs binaires $\sim$ arbres binaires (Catalan)
- Opérateurs commutatifs $\sim$ arbres non planaires
- Opérateurs associatifs $\sim$ arbres d'arité variable
- Feuilles toutes au même niveau $\sim$ arbres équilibrés

Cadre général

Modèles probabilistes sur les arbres/formules:

1. **Modèle combinatoire:**
   - Tous les arbres de même taille m sont équiprobables
   - La taille m tend vers l'infini
2. **Processus de branchement:** arbres de Galton-Watson, processus de Boltzmann...
   - On construit un arbre, de taille aléatoire
   - On étiquette l'arbre obtenu suivant les règles du système logique

Loi image sur l'ensemble des fonctions booléennes?

Comment évaluer les probabilités sur les arbres et sur les fonctions?

Enumérer des arbres = calculer des fonctions génératrices

1. Construction de l'arbre $\sim$ établir un système d'equations (algébriques, implicites,...)

2. Résoudre le système?

   - Explicitement? mais 2^{2^n} fonctions booléennes $\Rightarrow$ système de **très** grande taille!

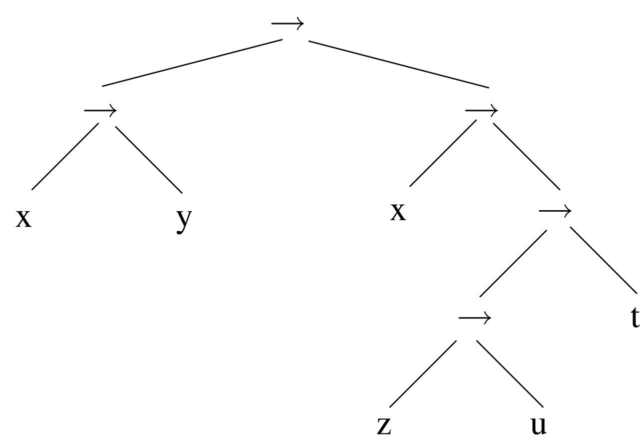
   - Définir des classes de fonctions équivalentes?
     Il y en a encore trop! [Polyà 40, Harrison 60]

   - Asymptotiquement? $\Rightarrow$ outils de combinatoire analytique
     Ex: système algébrique $\Rightarrow$ th. de Drmota-Lalley-Woods

Un système simple: l'implication

- Connecteur: $\rightarrow$
- Pas de littéraux négatifs

$$(x \rightarrow y) \rightarrow (x \rightarrow (z \rightarrow u) \rightarrow t)$$



Un système simple: l'implication

Pourquoi ce modèle?

- Simplicité: un seul connecteur, pas de négation, pas toutes les fonctions
  $\Rightarrow$ on peut espérer calculer la probabilité $P(f)$ d'une fonction f et étudier le lien avec sa complexité $C(f)$
- Logique intuitionniste
  Une tautologie $\sim$ une démonstration du but à partir des prémisses

Un système simple: l'implication

Quelles fonctions obtient-on?

Classe de Post $S_0 = \{f \in \mathcal{B}_n, f = x \vee g\}$

Combien de telles fonctions?

- Pour $n = 1$, 2 fonctions, $Vrai$ et x
- Pour $n = 2$, 6 fonctions, $Vrai$, x, y, $x \to y$, $y \to x$, $x \vee y$
- Pour $n = 3$, 38 fonctions; pour $n = 4$, 942 fonctions
- Pour un alphabet de n variables,
  $Card(S_0) = \sum_{i=1}^{n} \binom{n}{i} (-1)^{i+1} 2^{2^{n-1}}$
- E.I.S.: suite $A005530$

Un système simple: l'implication

Densité d'un sous-ensemble de formules

$\mathcal{I}_m$ ensemble des formules/arbres de taille m construites dans le système de l'implication

$\mathcal{I} = \cup_m \mathcal{I}_m$ ensemble de toutes les formules

Soit $E \subset \mathcal{I}$ et $E_m = E \cap \mathcal{I}_m$

$$\lim_{m \to +\infty} \frac{Card(E_m)}{Card(\mathcal{I}_m)} \ ?$$

Si cette limite exite, c'est la *densité* $\delta(E)$ de E dans $\mathcal{I}$

Théorème: Soit $f \in B_n$, et soit $\mathcal{I}(f)$ l'ensemble des arbres qui calculent f. Alors la limite $\delta(\mathcal{I}(f))$ existe pour tout f; ceci définit une loi de probabilité P sur B_n

Un système simple: l'implication

Quelques valeurs pour n petit

- $n = 1$: $P(1) = 0.72$, $P(x) = 0.28$.
- $n = 2$: $P(1) = 0.52$, $P(x) = 0.11$, $P(x \to y) = 0.10$, $P(x \vee y) = 0.06$.
- $n = 3$: $P(1) = 0.396$, $P(x) = 0.057$, $P(x \to y) = 0.033$, $P(x \vee y) = 0.013$, ...
- $n = 4$: $P(1) = 0.3$, $P(x) = 0.034$, $P(x \to y) = 0.014$, $P(x \vee y) = 0.004$, ...

Logique intuitionniste (simplifiée)

Règles de calcul:

- Initial

$$\frac{}{G, A \vdash A}$$

- Introduction de $\rightarrow$

$$\frac{G, A \vdash B}{G \vdash (A \rightarrow B)}$$

- Elimination de $\rightarrow$ (Modus Ponens)

$$\frac{G \vdash A \qquad G \vdash (A \rightarrow B)}{G \vdash B}$$

Tautologies intuitionnistes

Une formule T est une tautologie intuitionniste
$\Leftrightarrow$ on peut trouver une preuve de T avec ces trois règles.

Tautologies intuitionnistes

- $A \rightarrow A$ est une tautologie intuitionniste
- $A \rightarrow (B \rightarrow A)$ est une tautologie intuitionniste
- $A_1 \rightarrow (A_2 \rightarrow (.... \rightarrow (A_p \rightarrow B)...))$ est une tautologie dès que $B \in \{A_1, ..., A_p\}$: tautologie *simple*
- $((A \rightarrow B) \rightarrow A) \rightarrow A$ est-elle une tautologie intuitionniste?

Tautologies intuitionnistes

- $\{$Taut. intuitionnistes (de $\mathcal{I}$)$\} \subset \{$Taut. classiques (de $\mathcal{I}$)$\}$
- Proportion des tautologies intuitionnistes qui sont "simples"?

  Conjecture: [Zaionc et al. 00]

  > *Asymptotiquement, lorsque le nombre n de variables booléennes tend vers l'infini, toute tautologie intuitionniste est simple*

- Formule de Peirce: $((A \rightarrow B) \rightarrow A) \rightarrow A$
  - formule de $\mathcal{I}$ toujours vraie: tautologie
  - non démontrable en logique intuitionniste
- Proportion des taut. de $\mathcal{I}$, aussi taut. intuitionnistes?

  $\Leftrightarrow$ "densité" de la logique intuitionniste dans la logique classique?

  $\Leftrightarrow$ densité des formules "de Peirce"?

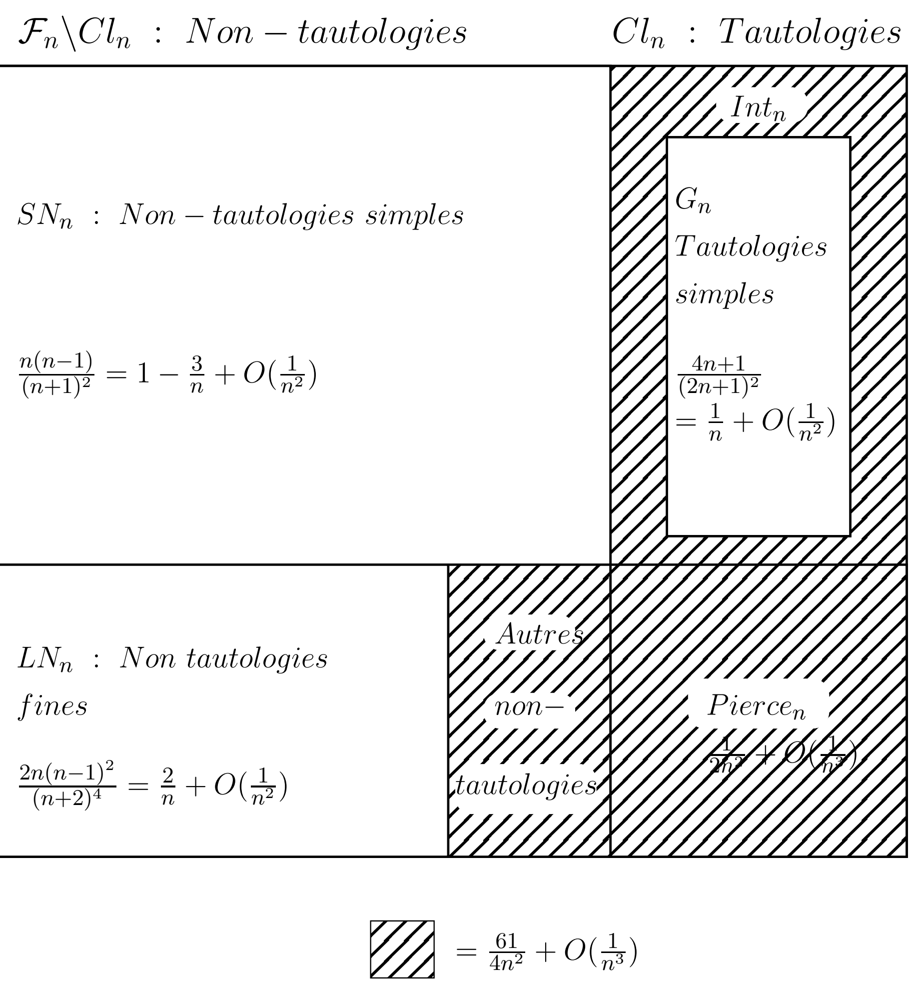
Tautologies intuitionnistes

On définit des classes de formules, incluses dans $\mathcal{I}$

- Tautologies simples
- Non-tautologies simples: les buts des prémisses diffèrent du but global
- Non-tautologies "fines"
- Tautologies "de Peirce"

Les densités des ensembles correspondants existent, et peuvent être calculées explicitement

[Fournier et al. 07, Genitrini et al. 08]


$\mathcal{F}_n \backslash Cl_n$: $Non-tautologies$
Cl_n : $Tautologies$
Int_n
SN_n : $Non-tautologies\ simples$
$\frac{n(n-1)}{(n+1)^2} = 1 - \frac{3}{n} + O(\frac{1}{n^2})$
G_n
$Tautologies$
$simples$
$\frac{4n+1}{(2n+1)^2}$
$= \frac{1}{n} + O(\frac{1}{n^2})$
LN_n : $Non\ tautologies$
$fines$
$\frac{2n(n-1)^2}{(n+2)^4} = \frac{2}{n} + O(\frac{1}{n^2})$
$Autres$
$non-$
$tautologies$
$Pierce_n$
$\frac{1}{2n^2} + O(\frac{1}{n^3})$
$= \frac{61}{4n^2} + O(\frac{1}{n^3})$

Logique intuitionniste vs. logique classique

Densité des tautologies intuitionnistes par rapport aux tautologies classiques?

$$\lambda := \frac{\delta(Taut.\ intuit.)}{\delta(Taut.)}$$

- Connecteur $\rightarrow$ seul: $\lambda = 1$
- On ajoute $\wedge$ ou 0: $\lambda = 1$
- On ajoute $\vee$: $\lambda = 5/8 < 1$

[Fournier et al. 08]

Un système simple: l'implication

Calcul de $P(f)$ pour $f \neq 1$

- Peut-on calculer la probabilité $P(f)$ de toute fonction f représentable par implication et (au plus) n variables?
- Peut-on relier $P(f)$ et $C(f)$?

Un système simple: l'implication

- Littéral x

$$\frac{1}{2n^2} + O\left(\frac{1}{n^3}\right)$$

- Fonction $x \to y$

$$\frac{9}{16n^3} + O\left(\frac{1}{n^4}\right)$$

- Pour $f \in S_0 \setminus \{1\}$:

$$P(f) = \frac{\alpha(f)}{4^n C(f)^{n+1}} + O\left(\frac{1}{C(f)^{n+2}}\right)$$

  $\alpha(f)$ lié aux arbres minimaux de f
  arbres de $\mathcal{A}(f)$ "simples" (obtenus p.s. par une expansion d'un arbre minimal)

[Fournier et al. 08]

Les arbres Et/Ou

Connecteurs $\vee$ et $\wedge$ binaires et non commutatifs, équiprobables
Littéraux (équiprobables) aux feuilles

Modèle d'arbre sous-jacent: les arbres de Catalan

- Système complet: on obtient toutes les fonctions booléennes
- Importance "historique": le plus étudié

Les arbres Et/Ou: résultats

- Définition de la loi de probabilité $P(f)$ par élagage d'arbres infinis; bornes reliant $P(f)$ et $C(f)$ [Lefmann-Savickỳ 97]
- Evaluation du nombre de fonctions calculées par des formules de taille L sur n variables [Savickỳ-Woods 98]
- Définition alternative de P, définition d'une autre loi π par processus de branchement, amélioration de la borne supérieure: $P(f) \leq e^{-\alpha C(f)/n^2}$ [Chauvin et al. 04]
- Améliorations de bornes reliant $P(f)$ et $C(f)$, comparaison de P et π pour des fonctions "read-once" [Gardy-Woods 05]
- $P(1) \sim 3/4n$ pour $n \to +\infty$; tautologies asymptotiquement "simples" $(x \vee \bar{x} \vee ...)$ [Woods 05, Kozik 08]
- Proba d'une fonction f fixée d'ordre $1/n^{C(f)+1}$ [Kozik 08]

Les tautologies dans différents systèmes logiques

Etude systématique initiée par Zaionc et al. vers 2000: probabilité asymptotique d'une tautologie dans divers systèmes logiques?

- Loi uniforme sur B_n: $1/2^{2^n}$
- Connecteur $\leftrightarrow$: $1/2^n$ (taille de l'arbre paire) [Matecki 03]
- Arbres Et/Ou: $1/n$; tautologies simples [Woods 05, Kozik 08]
- Implication: $1/n$; tautologies simples [Fournier et al. 07]
- Système logique t.q. les tautologies sont de probabilité $1/n$ $\Leftrightarrow$ il existe une tautologie avec *exactement une* répétition, et *au moins une* variable a *exactement une* occurrence

  [Fournier et al. 08]

Les outils pour étudier un système logique

- Représentation arborescente des formules
- Définition d'une classe d'arbres et énumération (exacte ou asymptotique)
- Fonction génératrice ϕ_f énumérant les arbres calculant $f \in B_n$
- Relations reliant les fonctions booléennes entre elles (ex. $f = g \wedge h$) $\Rightarrow$ système d'équations sur les ϕ_f
- Théorème de Drmota-Lalley-Woods: si ce système est algébrique, alors les ϕ_f ont la même singularité dominante et on peut calculer l'asymptotique

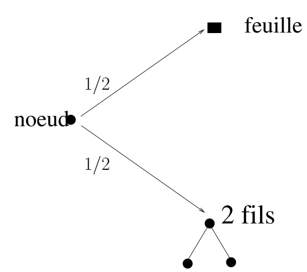
Distributions en arbre sur B_n

On choisit un système logique (connecteurs, littéraux).

Deux lois de probabilité sur $\mathcal{A} = \{$arbres/formules booléennes$\}$

$\Rightarrow$ deux lois image sur B_n: P et π

1. Tous les arbres de même taille m sont équiprobables, puis $m \to +\infty$; $P(f) =$ densité de $\mathcal{A}(f)$
2. Les arbres sont construits suivant un processus de branchement, puis étiquetés; la taille d'un arbre est aléatoire



$\pi(f)$ est la probabilité cumulée des arbres qui calculent f.

Distributions en arbre sur B_n: calcul explicite

Fonctions génératrices:

- $\Phi(z)$ énumère tous les arbres
- ϕ_f énumère les arbres représentant f

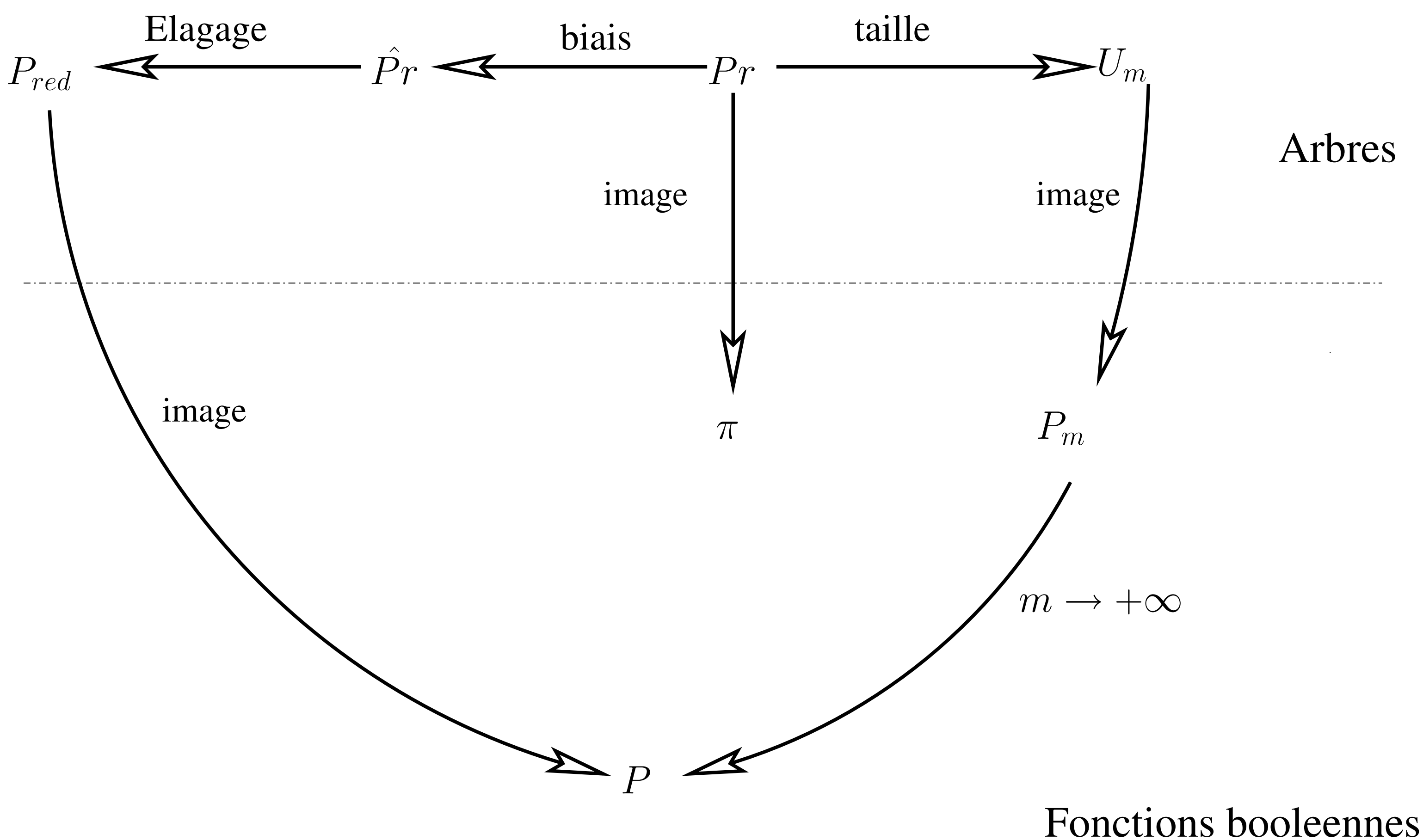
(Souvent) système algébrique $\Rightarrow$ singularité algébrique commune ρ

$$\begin{aligned}\Phi(z) &= \alpha - \beta\sqrt{1 - z/\rho} + O(1 - z/\rho) \\ \phi_f(z) &= \alpha_f - \beta_f\sqrt{1 - z/\rho} + O(1 - z/\rho)\end{aligned}$$

Alors

$$\pi(f) = \frac{\alpha_f}{\alpha}; \qquad P(f) = \frac{\beta_f}{\beta}$$

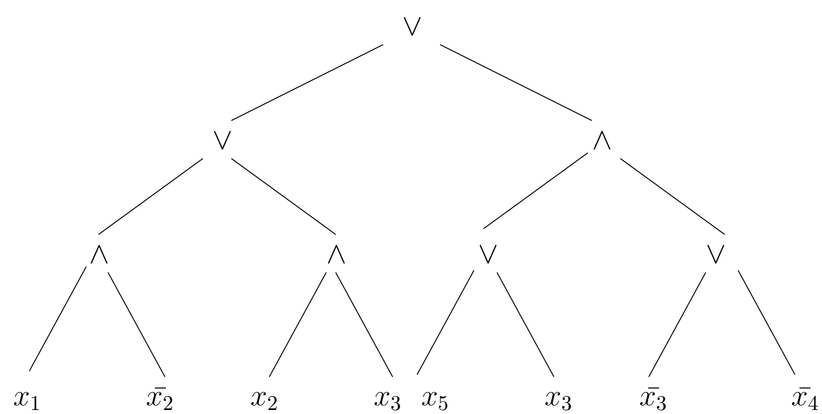
Les différentes lois de probabilité



Arbres équilibrés

Toutes les feuilles sont au même niveau

Exemple: étiquettes $\vee$, $\wedge$ et littéraux



Arbres équilibrés

Pour les définir:

- choisir un processus de croissance $\Leftrightarrow$ connecteur logique
- choisir un ensemble H_0 d'étiquettes sur les feuilles (variables booléennes, littéraux, constantes) et une distribution de probabilité sur H_0

On itère le processus de croissance $\Rightarrow$ ensemble H_h d'arbres de hauteur h

Loi uniforme sur H_h: induit une loi p_h sur B_n

- Construction d'une (classe de) fonction(s) donnée?
- Etude de la probabilité limite sur B_n?

Arbres équilibrés

- Valiant 84: construction *Majorité* par itération du connecteur $\alpha(x_1, \ldots, x_4) = (x_1 \vee_2) \wedge (x_3 \vee x_4)$; taille de l'expression $O(k^{5.3})$
- Boppana 85: généralisation pour les fonctions ℓ-seuil (*Majorité:* fonction $k/2$-seuil); taille des expressions $O(\ell^{4.3} k \log k)$
- Gupta et Mahajan 97, Servedio 99: améliorations/extensions pour *Majorité* ou les fonctions seuil
- Savicky 90: construction d'une distribution uniforme sur B_n; connecteur non linéaire équilibré
- Pippenger et Brosky 05: étude systématique suivant la nature du connecteur
- Fournier et al. 08: connecteur aléatoire $\vee$ ou $\wedge$

Arbres équilibrés: la classification de Brosky et Pippenger

- Connecteur α non linéaire, équilibré; $H_0 = \{$littéraux, 0, 1$\}$
  $\Rightarrow$ loi uniforme sur B_n
- Connecteur α non linéaire, auto-dual; $H_0 = \{$littéraux$\}$
  $\Rightarrow$ loi uniforme sur les fonctions auto-duales
  f est auto-duale $\Leftrightarrow f(\bar{x}_1, ..., \bar{x}_n) = \overline{f(x_1, ..., x_n)}$
- α non linéaire, non équilibré et monotone; pas de négations
  $\Rightarrow$ fonction seuil
- α non linéaire, équilibré et monotone; pas de négations
  $\Rightarrow$ majorité (n impair) ou loi uniforme sur $\{T_{n/2,n}\}$ (n pair)

Arbres équilibrés: connecteur aléatoire

Arbres équilibrés; chaque connecteur = $\vee$ ou $\wedge$ avec proba. $1/2$
$H_0 \subset \{0, 1, x_1, ..., x_n\}$; distribution π_0 sur H_0

- Existence d'une distribution limite π sur B_n
- Support $\subset \{0, 1, \phi_i, 0 \leq i \leq n+1\}$
  $\phi_i(x_1, ..., x_n) = 1$ ssi $\sum_j \pi_0(x_j)x_j > i/(n+2)$
- On peut caractériser π
- Vitesse de convergence accessible, et évaluation de la taille d'une expression calculant une des fonctions limites
- Extension possible pour admettre les littéraux négatifs

Arbres équilibrés: connecteur aléatoire

Exemples

- $H_0 = \{x_1, ..., x_n\}$; $\pi_0 = \mathcal{U}(H_0)$
  $\Rightarrow \pi$ uniforme sur les n fonctions seuil
- $H_0 = \{0, 1, x_1, ..., x_n\}$; $\pi_0 = \mathcal{U}(H_0)$
  $\Rightarrow \pi$ uniforme sur les $n + 2$ fonctions seuil ou constantes
- $H_0 = \{x_1, x_2, x_3\}$; $\pi_0(x_i) = i/6$ $(1 \leq i \leq 3)$
  $\Rightarrow \pi$ uniforme sur 6 fonctions: 1-seuil, $x_2 \vee x_3$, $x_3 \vee (x_1 \wedge x_2)$, $x_3 \wedge (x_1 \vee x_2)$, $x_2 \wedge x_3)$, 3-seuil

Valeur moyenne d'une expression booléenne

- Modèle d'expression/arbre: choix des connecteurs
- Feuilles étiquetées par des littéraux tous distincts
- Loi de Bernoulli sur les feuilles: $Proba(1) = p$, $Proba(0) = 1 - p$
- Chaque assignation aux feuilles associe une valeur 0 ou 1 à l'arbre

Valeur moyenne de l'expression booléenne?

Valeur moyenne d'une expression booléenne: arbres équilibrés

- Noeuds internes: choix uniforme de $\vee$ ou $\wedge$
- h est la hauteur de l'arbre, qui a donc $r = 2^h$ feuilles
- Pour une feuille: valeurs 0 ou 1 équiprobables
- Un arbre ϕ prend en entrée un r-uplet de $\{0, 1\}^r$, renvoie 0 ou 1
- *Moyenne* X_h de la valeur de sortie, si les 2^r entrées sont équiprobables?

Valeur moyenne d'une expression booléenne: arbres équilibrés

- X_h est une martingale
- $\lim_{h \to +\infty} X_h$ existe, vaut 0 ou 1 p.s.
- $E[X_h] = 1/2$, $E[X_h^2] = 1/2 - 1/h + O(\log h/h^2)$
- Calcul récursif des moments possible
- Soit $Z_h = X_h(1 - X_h)$. Alors
  - $E[Z_h] = 1/(h + O(\log h))$
  - $E[Z_h^2] \sim \alpha/h$
  - $P(Z_h > a) = \Theta(1/h)$

[Pemantle-Wards 05]

Valeur moyenne d'une expression booléenne: famille simple d'arbres

- Ensemble fini $\mathcal{B}$ de connecteurs non commutatifs d'arité bornée
- Arbres non équilibrés $\in$ famille simple d'arbres engendrée par $\mathcal{B}$
- Loi Bernoulli(p) sur les feuilles $\Rightarrow$ valeur moyenne $P_{1,m}(p)$ sur les arbres de taille m

Alors: $\lim_{m\to+\infty} P_{1,m}(p)$ existe, vaut $P_1(p)$ exprimable en fonction des caractéristiques de la famille simple

[Yashunsky 05]

Satisfiabilité

Les problèmes de satisfiabilité peuvent être formulés dans ce cadre

- $3 - SAT$: clauses $C_i = l_{i,1} \vee l_{i,2} \vee l_{i,3}$
- $XOR - SAT$: clauses $C_i = l_{i,1} \oplus l_{i,2} \oplus l_{i,3}$
- Divers CSP $\leftrightarrow$ divers types de clauses
- On veut satisfaire $C_1 \wedge C_2 \wedge ... \wedge C_m$

Une expression est insatisfiable $\Leftrightarrow$ elle calcule FAUX

Peut-on obtenir la probabilité de Faux dans divers systèmes?

[Dershowitz-Harris 03]

Satisfiabilité

Des différences notables avec les systèmes précédents:

- Formule $\sim$ arbre équilibré de hauteur 2 et d'arité à la racine $m \to +\infty \Rightarrow$ Formule $\sim$ suite d'arbres élémentaires (clauses)!

  vs

  Arbre de Catalan (par ex.)

- Système d'equations linéaire; nombre (doublement) exponentiel de pôles; le pole principal tend vers 0

  vs

  Système d'équations algébriques, toutes les fonctions partagent la même singularité principale, fixée

D'où des difficultés pour l'analyse asymptotique....

Satisfiabilité

MAIS... des premiers résultats encourageants sur XOR-SAT

Perspectives

Il reste des questions ouvertes sur les distributions en arbre

- Lien exact entre P et π?
  Ex: pour une fonction "read-once", $\pi(f) > P(f)$
  mais $\pi(Vrai) < P(Vrai)$ [Gardy-Woods 05]
- Influence de la commutativité? de l'associativité?
- Complexité moyenne d'une fonction booléenne?
  - Selon la distribution uniforme, c'est $2^n/\log n$
  - Selon une distribution en arbre?
- Lien entre probabilité et complexité, pour une fonction de grande complexité? Généraliser l'élagage de Lefmann et Savickỳ? l'améliorer?
- ...

Perspectives (suite)

Ce qui a été fait est pour la logique propositionnelle

Et pour la logique du premier ordre? le lambda-calcul?

Peut-on énumérer les formules de taille donnée? Définir la probabilité qu'une formule bien formée soit toujours vraie? équivalente à une formule donnée?