Approche par encadrement pour l'analyse de modèles complexes

Farah AIT SALAHT

Laboratoire SAMOVAR, Télécom SudParis

09 février 2015

Sommaire

Sommaire

Contexte et problématique étudiés en thèse

- Analyse de performance de systèmes réels sur lesquels nous avons des mesures partielles
- Facteur d'incertitude
- Analyse exacte très difficile voire impossible à effectuer
- Aspects de l'incertitude :
  - Imprécision dans les probabilités de transition
  - Incertitude de mesure et bruit de quantification
- Notre objectif :
  - Apporter des solutions de modélisation pour les systèmes partiellement connus et résolution numérique
  - Fournir des encadrements sur les performances des systèmes étudiés

Sommaire

Sommaire

a) Probabilités de transition imprécises dans les chaînes de Markov

- Difficulté de déterminer tous les paramètres du modèle
- Impossibilité/coût élevé de déterminer précisément les probabilités de transition
- Description par intervalles des probabilités de transition
- Ensemble défini par deux matrices $\mathbf{L}$ et $\mathbf{U}$, tel que $\mathbf{L} \leq \mathbf{P} \leq \mathbf{U}$ ($\leq$: par élément)

Exemple

On considère la chaîne de Markov imprécise à 4 états dont les probabilités de transitions sont exprimées comme suit :

$$\begin{pmatrix} [0.2744, 0.3456] & [0.2765, 0.3477] & [0.3779, 0.4490] & 0 \\ [0.1024, 0.3067] & 0 & [0.2651, 0.7282] & [0.1430, 0.4282] \\ [0.3844, 0.5085] & [0.3844, 0.5152] & 0 & [0.0488, 0.1004] \\ [0.0248, 0.0544] & 0.5571 & [0.2229, 0.4885] & [0, 0.0003] \end{pmatrix}$$

a) Probabilités de transition imprécises dans les chaînes de Markov

- Difficulté de déterminer tous les paramètres du modèle
- Impossibilité/coût élevé de déterminer précisément les probabilités de transition
- Description par intervalles des probabilités de transition
- Ensemble défini par deux matrices $\mathbf{L}$ et $\mathbf{U}$, tel que $\mathbf{L} \leq \mathbf{P} \leq \mathbf{U}$ ($\leq$: par élément)

Exemple

On considère la chaîne de Markov imprécise à 4 états dont les probabilités de transitions sont exprimées comme suit :

$$\begin{pmatrix} [\mathbf{0.2744}, 0.3456] & [0.2765, 0.3477] & [0.3779, 0.4490] & 0 \\ [\mathbf{0.1024}, 0.3067] & 0 & [0.2651, 0.7282] & [0.1430, 0.4282] \\ [\mathbf{0.3844}, 0.5085] & [0.3844, 0.5152] & 0 & [0.0488, 0.1004] \\ [\mathbf{0.0248}, 0.0544] & 0.5571 & [0.2229, 0.4885] & [0, 0.0003] \end{pmatrix}$$

a) Probabilités de transition imprécises dans les chaînes de Markov

- Difficulté de déterminer tous les paramètres du modèle
- Impossibilité/coût élevé de déterminer précisément les probabilités de transition
- Description par intervalles des probabilités de transition
- Ensemble défini par deux matrices $\mathbf{L}$ et $\mathbf{U}$, tel que $\mathbf{L} \leq \mathbf{P} \leq \mathbf{U}$ ($\leq$: par élément)

Exemple

On considère la chaîne de Markov imprécise à 4 états dont les probabilités de transitions sont exprimées comme suit :

$$\begin{pmatrix} [0.2744, \mathbf{0.3456}] & [0.2765, 0.3477] & [0.3779, 0.4490] & 0 \\ [0.1024, \mathbf{0.3067}] & 0 & [0.2651, 0.7282] & [0.1430, 0.4282] \\ [0.3844, \mathbf{0.5085}] & [0.3844, 0.5152] & 0 & [0.0488, 0.1004] \\ [0.0248, \mathbf{0.0544}] & 0.5571 & [0.2229, 0.4885] & [0, 0.0003] \end{pmatrix}$$

a) Probabilités de transition imprécises dans les chaînes de Markov

- Difficulté de déterminer tous les paramètres du modèle
- Impossibilité/coût élevé de déterminer précisément les probabilités de transition
- Description par intervalles des probabilités de transition
- Ensemble défini par deux matrices $\mathbf{L}$ et $\mathbf{U}$, tel que $\mathbf{L} \leq \mathbf{P} \leq \mathbf{U}$ ($\leq$: par élément)

Exemple

On considère la chaîne de Markov imprécise à 4 états dont les probabilités de transitions sont exprimées comme suit :

$$\begin{pmatrix} [0.2744, 0.3456] & [0.2765, 0.3477] & [0.3779, 0.4490] & 0 \\ [0.1024, 0.3067] & 0 & [0.2651, 0.7282] & [0.1430, 0.4282] \\ [0.3844, 0.5085] & [0.3844, 0.5152] & 0 & [0.0488, 0.1004] \\ [0.0248, 0.0544] & 0.5571 & [0.2229, 0.4885] & [0, 0.0003] \end{pmatrix}$$

Question : *Est-ce qu'on peut dire quelque chose sur les probabilités stationnaires ?*

a) Probabilités de transition imprécises dans les chaînes de Markov

- Difficulté de déterminer tous les paramètres du modèle
- Impossibilité/coût élevé de déterminer précisément les probabilités de transition
- Description par intervalles des probabilités de transition
- Ensemble défini par deux matrices $\mathbf{L}$ et $\mathbf{U}$, tel que $\mathbf{L} \leq \mathbf{P} \leq \mathbf{U}$ ($\leq$: par élément)

Exemple

On considère la chaîne de Markov imprécise à 4 états dont les probabilités de transitions sont exprimées comme suit :

$$\begin{pmatrix} [0.2744, 0.3456] & [0.2765, 0.3477] & [0.3779, 0.4490] & 0 \\ [0.1024, 0.3067] & 0 & [0.2651, 0.7282] & [0.1430, 0.4282] \\ [0.3844, 0.5085] & [0.3844, 0.5152] & 0 & [0.0488, 0.1004] \\ [0.0248, 0.0544] & 0.5571 & [0.2229, 0.4885] & [0, 0.0003] \end{pmatrix}$$

Question : *Est-ce qu'on peut dire quelque chose sur les probabilités stationnaires ?* ***OUI***

Sommaire

b) Incertitude de mesure et bruit de quantification

- Traces de trafic réelles
- Mesures de données trop importantes
- Précision limitée des mesures
- Bruit de quantification

Concrètement

- Distributions empiriques avec un grand nombre de bins (plusieurs milliers)
- Équation de structure du modèle parfaitement connue (Loynes, Max-Plus, . . .)
- Effectuer des calculs numériques sur les distributions de probabilités

Inconvénients :

- Croissance très rapide du nombre de bins suite aux opérations numériques
- Calcul numérique très difficile voire impossible

Sommaire

Sommaire

Une méthodologie : les méthodes de comparaison pour le calcul de bornes

a) Probabilités de transition imprécises dans les chaînes de Markov

- Calculer des bornes par éléments sur la distribution de probabilité stationnaire
- Utiliser la comparaison par élément des matrices et des vecteurs ($\leq$)
- Faire un calcul de faible complexité
- Développer un algorithme itératif avec une amélioration des bornes à chaque itération
  - Fondé sur la théorie polyhédrale de Courtois et un résultat de Muntz
  - Et sur un algorithme prouvé par J.M. Fourneau et A. Busic (NLA 2011)
  - Systèmes dynamiques (c.-à-d. non stochastique) reposant sur les séquences (max, +) ou (min, +)

a) Probabilités de transition imprécises dans les chaînes de Markov

- Précision : nous devons avoir un résultat très précis pour les distributions stationnaires.
- Complexité : on suppose que la taille de la chaîne est de n.
- n résolutions numériques de l'état stationnaire.
- Si nous utilisons GTH (en raison de la précision), nous avons une complexité cubique pour la résolution d'une chaîne.
- Ainsi, une complexité en $\theta(n^4)$ pour le calcul global.
- **Solution** : remplacer la GTH par un algorithme fondé sur ∇.
- Par itération : le calcul de ∇ nécessite : $\theta(nz)$, avec nz le nombre d'éléments non nuls
- Nouveau schéma de calcul pour la distribution stationnaire d'un ensemble de DTMC
- Construction de nouvelles bornes par éléments prouvées sur la MTTF d'une chaîne de Markov imprécise et absorbante

Sommaire

Une méthodologie : les méthodes de comparaison et de bornes

b) Incertitude de mesure et bruit de quantification

- Simplification du modèle initial
- Définition de nouvelles bornes sur les distributions fondées sur l'ordre stochastique
- Réduire la taille des distributions
- Offrir un compromis entre la qualité des bornes et la complexité du calcul

Exemples étudiés en thèse :

- Distribution de la durée d'exécution dans un graphe de tâches stochastique
- Distribution des délais d'attente dans une file FIFO
- Mesures de performance dans une file d'attente simple et dans les réseaux

Extension envisagée : Dériver des bornes sur la date à laquelle une expression booléenne devient vraie.

Brève introduction sur l'ordre stochastique

- $\mathcal{G} = \{1, 2, \ldots, n\}$ un espace d'état fini
- X, Y : distributions discrètes sur $\mathcal{G}$
- $p_X(i) = prob(X = i)$ et $p_Y(i) = prob(Y = i)$ pour $i \in \mathcal{G}$

Propriétés sur l'ordre stochastique $\leq_{st}$

- **Définition de l'ordre** $\leq_{st}$: $X \leq_{st} Y \; ssi \; \sum_{k=i}^{n} p_X(k) \leq \sum_{k=i}^{n} p_Y(k), \quad \forall i.$
- **Comparaison de fonctions non décroissantes** :

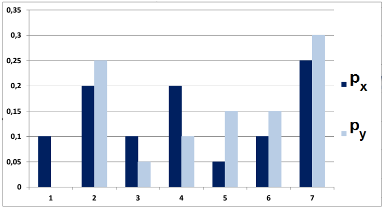
$$X \leq_{st} Y \iff \mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$$

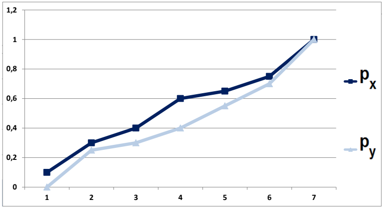
pour toute fonction non décroissante $f : \mathcal{G} \to \mathbb{R}^+$ à condition que les espérances existent.

- Soient F_X et F_Y leurs probabilités cumulées. Alors,

$$X \leq_{st} Y \Leftrightarrow F_X(a) \geq F_Y(a), \; \forall a \in \mathcal{G}$$

Brève introduction sur l'ordre stochastique





-pmfs des distributions X et Y-

-Leurs fonctions de répartition-

FIGURE : $\mathscr{G} = \{1, 2, \ldots, 7\}$, $\boldsymbol{p}_X = [0.1, 0.2, 0.1, 0.2, 0.05, 0.1, 0.25]$ et $\boldsymbol{p}_Y = [0, 0.25, 0.05, 0.1, 0.15, 0.15, 0.3]$.

Bornes stochastiques sur les histogrammes

- Hypothèse : ordre total sur l'espace $\mathscr{H}$, de taille N
- Nous avons une distribution $\boldsymbol{d}$ et $\boldsymbol{r}$: fonction de récompense positive croissante, $R[\boldsymbol{d}] = \sum \boldsymbol{r}(i)\boldsymbol{d}(i)$
- **Déterminer *d1* et *d2* tel que :**
  1. $\boldsymbol{d}2 \leq_{st} \boldsymbol{d} \leq_{st} \boldsymbol{d}1$,
  2. $\boldsymbol{d}1$ et $\boldsymbol{d}2$ ont exactement K bins (pas nécessairement les mêmes) ; ***d1*** a comme support $\mathscr{H}^u$ et ***d2*** a comme support $\mathscr{H}^l$,
  3. $\sum_{i\in\mathscr{H}} \boldsymbol{r}(i)\boldsymbol{d}(i) - \sum_{i\in\mathscr{H}^l} \boldsymbol{r}(i)\boldsymbol{d}2(i)$ est minimal pour les distributions bornes inférieures $\boldsymbol{d}$ avec K bins,
  4. $\sum_{i\in\mathscr{H}^u} \boldsymbol{r}(i)\boldsymbol{d}1(i) - \sum_{i\in\mathscr{H}} \boldsymbol{r}(i)\boldsymbol{d}(i)$ est minimal pour les distributions bornes supérieures $\boldsymbol{d}$ avec K bins.

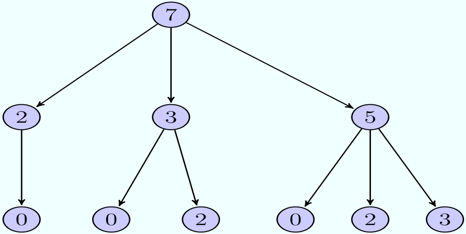
Bornes optimales, programmation dynamique

- Problème de théorie des graphes.
- On considère un graphe pondéré $G = (V, E)$ avec :
  - **Borne inférieure :** $w(e) = \sum_{j \in \mathscr{H}: u < j < v} \boldsymbol{d}(j)(\boldsymbol{r}(j) - \boldsymbol{r}(u))$
  - **Borne supérieure :** $w(e) = \sum_{j \in \mathscr{H}: u < j < v} \boldsymbol{d}(j)(\boldsymbol{r}(v) - \boldsymbol{r}(j))$
- Calcul de la borne optimale $\equiv$ Déterminer le chemin de longueur K ($K << N$) de coût minimum dans le graphe G. Algorithme fondé sur la programmation dynamique avec une complexité : $\theta(N^2 K)$.
- La masse de probabilités des nœuds supprimés est sommée avec
  - **Borne inférieure :** les prédécesseurs immédiats
  - **Borne supérieure :** les successeurs immédiats

Exemple : Borne supérieure optimale

$\mathscr{A}$ une distribution discrète avec support $\{0, 2, 3, 5, 7\}$
et probabilités $[0.05, 0.3, 0.15, 0.2, 0.3]$.
$\boldsymbol{r}$: fonction de récompense, $\boldsymbol{r}(a_i) = a_i$, $R[\mathscr{A}] = \sum_{a_i \in \mathbf{A}} \boldsymbol{r}(a_i)\ p_{\mathbf{A}}(i) = 4.15$.

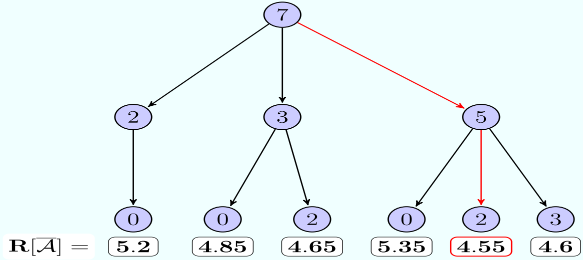
► Calculer la borne optimale supérieure $\overline{\mathscr{A}}$ sur 3 bins tel que $R[\overline{\mathscr{A}}] - R[\mathscr{A}]$ est minimale.



Exemple : Borne supérieure optimale

$\mathscr{A}$ une distribution discrète avec support $\{0, 2, 3, 5, 7\}$
et probabilités $[0.05, 0.3, 0.15, 0.2, 0.3]$.
$\boldsymbol{r}$: fonction de récompense, $\boldsymbol{r}(a_i) = a_i$, $R[\mathscr{A}] = \sum_{a_i \in \mathbf{A}} \boldsymbol{r}(a_i)\; p_{\mathbf{A}}(i) = 4.15$.

▶ Calculer la borne optimale supérieure $\overline{\mathscr{A}}$ sur 3 bins tel que $R[\overline{\mathscr{A}}] - R[\mathscr{A}]$ est minimale.



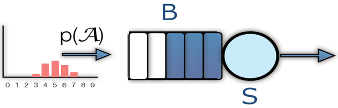
Distribution bornante $[0.35, 0.35, 0.3]$ avec support $\{2, 5, 7\}$ et $R[\overline{\mathscr{A}}] = 4.55$.

Sommaire

Bornes pour des systèmes dynamiques

$Sortie = f(Entree_1; Entree_2, \ldots)$

Exemple 1 : file d'attente

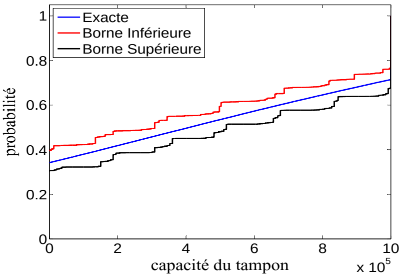


La longueur du tampon à l'instant k :
$Q(k) = \min(B, (Q(k-1) + A(k) - S)^+); \;\; k \in \mathbb{N}$

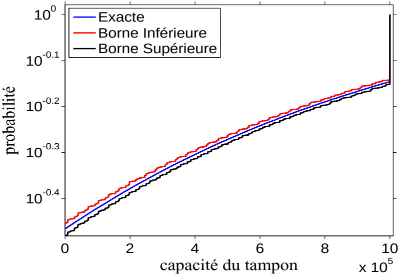
- Opérations sur les v.a. = convolution (+)
- Complexité espace à chaque instant k : au plus $\theta(|A| \times |Q(k-1)|)$, temps : $\theta(|A| \times |Q(k-1)|)$ (algo naif), $\theta((|A| \times |Q(k-1)|) log(|A| \times |Q(k-1)|))$ (FFT)
- Bornes supérieures avec $A^{sup}(k) \geq_{st} A(k)$

Mesures de performance d'une file d'attente simple

Fonction de répartition (cdf) de la longueur du tampon sous la trace MAWI



(a) bins=20



(b) bins=100

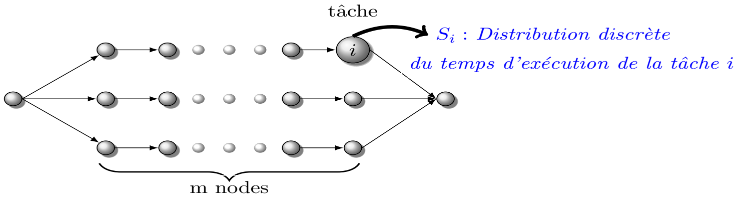
Temps de calcul pour $bins = 100$:

Exacte : 1897 s

Borne Inférieure : 0.35 s **et** **Borne Supérieure :** 0.33 s

Bornes pour des systèmes dynamiques

Exemple 2 : graphe de tâches stochastique (série parallèle)



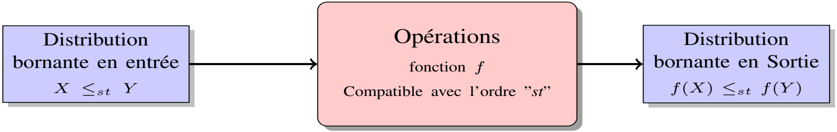
Distribution de la durée d'exécution : $T_i = \max_{j \in \Gamma_i^-} \{T_j\} + S_i$

- Opérations sur les v.a. : Convolution (+), Produit des pmfs (max)
- Complexité convolution : espace, au plus $\theta(\prod_i N_{X_i})$, temps : $\prod_i N_{X_i}$
- Bornes supérieures avec $\forall j \in \Gamma_i^-, T_j^{sup} \geq_{st} T_j$ et $S_i^{sup} \geq_{st} S_i$

Construction des bornes

$Sortie = f(Entree_1, Entree_2, \cdots)$

- "**Monotonie**" :



Si $Entree_1^{inf} \leq_{st} Entree_1 \Rightarrow$
$f(Entree_1^{inf}, Entree_2, \cdots) \leq_{st} f(Entree_1, Entree_2, \cdots)$

$\leq_{st}$: fonctions croissantes : Max, Min, +,... Système Dynamique Max, Plus

- Indépendance des v.a.
  sinon par conditionnement mais complexité augmente exponentiellement
- Analyse de v.a. continues est également possible par discrétisation

Extension pour la Vérification Probabiliste

- $f(t)$: une proposition atomique ou une formule logique à l'instant t : $f(t) = 1(Vrai)$ ou $f(t) = 0(Faux)$,
- T_f : variable aléatoire représentant la plus petite date à laquelle la formule logique $f(t) = 1$ quand $f(0) = 0$,
- H_f : l'histogramme (pmf) associé à T_f :
  - $H_f[i] = Prob(T_f = i)$
  - $Prob(f(t) = 1) = Prob(T_f \leq t) = \sum_{k \leq t} H_f[k]$

Exemple :

- $f(t)(resp.g(t))$: la composante A (resp. B) est en panne à la date t
- $T_f(resp.T_g)$: v.a. représentant la date de panne pour la composante A (resp. B)
  - A dispo. à t $\Rightarrow \forall k \leq t, f(k) = 0$
  - $Prob(A$ dispo. à t$) = Prob(T_f > t) = \sum_{j|j>t} H_f[j]$

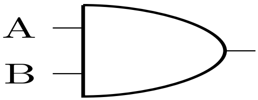
Conjonction

- $z(t) = f(t)\ AND\ g(t)$
- $T_z = \max(T_f, T_g)$
- la taille de H_z, $|H_z| = \theta(|H_f| + |H_g| - 1)$ au plus
  nombre d'opérations $\theta(|H_f| + |H_g|)$

Conjonction

- $z(t) = f(t)\ AND\ g(t)$
- $T_z = \max(T_f, T_g)$
- la taille de H_z, $|H_z| = \theta(|H_f| + |H_g| - 1)$ au plus
  nombre d'opérations $\theta(|H_f| + |H_g|)$

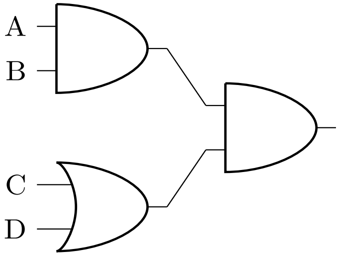
Exemple : Arbre de fautes dynamiques : porte AND



- $z(t) = 1 \Rightarrow A$ et B en panne à la date t
- Histogrammes des dates de panne pour A et B :
  $H_f = \{0.5, 0.2, 0.3\}$ définie sur le support $\{1, 6, 7\}$
  $H_g = \{0.1, 0.7, 0.2\}$ définie sur le support $\{2, 8, 10\}$
- $H_z = \{0.05, 0.02, 0.03, 0.7, 0.2\}$ définie sur le support $\{2, 6, 7, 8, 10\}$.

- $w(t) = f(t)\ OR\ g(t)$
- $T_w = \min(T_f, T_g)$

Exemple :



Avec

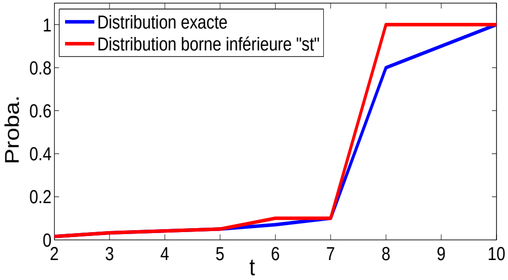
- $x(t)$: C est en panne à la date t,
  $H_x = \{0.5, 0.5\}$ définie sur $\{3, 5\}$.
- $y(t)$: D est en panne à la date t,
  $H_y = \{0.3, 0.7\}$ définie sur $\{1, 9\}$.
- $v(t)$: système est en panne à la date t, T_v la date de panne du système.
- $v(t) = (f(t)\ AND\ g(t))\ AND\ (x(t)\ OR\ y(t))$.

$H_v = \{0.015, 0.0175, 0.0175, 0.02, 0.03, 0.7, 0.2\}$ définie sur $\{2, 3, 5, 6, 7, 8, 10\}$.
Probabilité de la disponibilité à la date $t = \sum_{k>t} H_v[k]$.

Bornes stochastiques

- $H_f \leq_{st} H_f^{sup}, \quad Prob(T_f > t) \leq Prob(T_f^{sup} > t)$
- $H_f \geq_{st} H_f^{inf}, \quad Prob(T_f > t) \geq Prob(T_f^{inf} > t)$

• *Exemple 2*
$H_f^{inf} = \{0.5, 0.5\}$ définie sur $\{1, 6\}$, $H_g^{inf} = \{0.1, 0.9\}$ définie sur $\{2, 8\}$
$H_v^{inf} = \{0.015, 0.0175, 0.0175, 0.05, 0.9\}$ définie sur $\{2, 3, 5, 6, 8\}$.
$H^{inf} \leq_{st} H$



- Proba. que le système dispo. à la date $t = 4$ est : $0.9675 \leq 0.9675$
- dispo. à la date $t = 7$ est $0.9 \leq 0.9$

Sommaire

Conclusion

- Apporter de nouvelles solutions de modélisation et de résolution numérique reposant sur les techniques de bornes
- Nouvelles bornes stochastiques sur les distributions discrètes avec réduction de taille : définition d'un algorithme optimal
- Applications en cours :
  - Fiabilité et sûreté de fonctionnement
  - Vérification probabiliste : DFT
  - Dimensionnement de réseaux dans les cloud computing