UNIVERSITE PARIS-SUD
ÉCOLE DOCTORALE : Laboratoire de Recherche en Informatique.
DISCIPLINE Graphes Combinatoires.

THÈSE DE DOCTORAT

soutenue le 27/09/2013

par

Marc LETOURNEL

APPROCHES DUALES DANS LA RESOLUTION DE PROBLEMES STOCHASTIQUES

Directeur de thèse : Abdel LISSER Professeur, Laboratoire de Recherche en Informatique, Orsay.

Président du jury : Marc BABOULIN, Professeur Université Paris Sud.

Rapporteurs :

Jean-Baptiste HIRIART-URRUTY, Professeur, Université de Toulouse.

Alexei GAIVORONSKI, Professeur, Université de Trondheim (Norvège).

Examinateur :

Marc BABOULIN, Professeur Université Paris Sud.

Patrice PERNY, Professeur Université Paris 6.

Chapter 1

Résumé en français du rapport de thèse

Le travail général de cette thèse consiste à étendre les outils analytiques et algébriques usuellement employés dans la résolution de problèmes combinatoires algorithmiques déterministes à un cadre combinatoire stochastique. Deux cadres distincts sont abordés : les problèmes combinatoires stochastiques discrets et les problèmes stochastiques continus. Le cadre discret est abordé à travers le problème de la forêt couvrante de poids maximal dans une formulation Two-Stage à multi scénarios. La version déterministe très connue de ce problème établit des liens entre la fonction de rang dans un matroïde et la formulation duale via l'algorithme glouton. La clé de voûte de la preuve mathématique du cas déterministe réside d'une part dans la fomulation duale du problème et l'absence de saut de dualité pour le problème linéaire, et d'autre part dans une transformation d'Abel appliquée sur la différence de coût des arêtes. La formulation stochastique discrète du problème de la forêt maximale couvrante est transformée en un problème déterministe équivalent, mais du fait de la multiplicité des scénarios, le dual associé est en quelque sorte incomplet. Le travail réalisé ici consiste à comprendre en quelles circonstances la fomulation duale atteint néanmoins un minimum égal au problème primal intégral. D'ordinaire, une approche combinatoire classique des problèmes de matroïdes consiste à rechercher des configurations particulières au sein des graphes, comme les circuits, et à explorer d'éventuelles recombinaisons. Le problème classique de l'intersection de deux matroïdes est par exemple résolu par ce type d'approche algorithmique, où la partie analytique est finalement absente. Les preuves combinatoires prennent en compte les éléments de reconfiguration d'un graphe pondéré en inventoriant une liste de reconfigurations possibles . Pour donner une interprétation prosaïque, si on change d'une manière infinitésimale les valeurs de poids des arêtes d'un graphe, il est possible que la forêt couvrante se réorganise complètement. Ceci est vu comme un obstacle dans une approche purement combinatoire. Pourtant, certaines grandeurs analytiques vont varier de manière continue en fonc-

tion de ces variations infinitésimales, comme la somme des poids des arêtes choisies. Il apparaît également que les choix de telle ou telle arête est une fonction de son poids, mais également du poids des autres arêtes. Ainsi, il est naturel d'essayer de formuler ces sauts décisionels comme autant de fonctions implicites (je serais tenté d'écrire fonctions implicites les unes des autres si cela n'était par essence même le rôle des fonctions implicites). Après un premier chapitre d'introduction des concepts de base, la section 2 décrit la formulation déterministe et la formulation stochastique du problème de la forêt couvrante de poids maximal. Signalons dès à présent que j'ai choisi de répartir les références bibliographiques dans chacun des chapitres séparément, dans la mesure où la lecture des chapitres peut se faire séparément elle aussi. Dans le chapitre 3, la formulation stochastique de la forêt couvrante dans le cas de deux scénarios seulement est abordée avec une preuve de la conservation du caractère intégral du dual. Le chapitre 4 présente le cas de trois scénarios ou plus et donne les situations où le système dual perd son caractère intégral. Le chapitre 5 propose une réduction du problème considéré et aborde un algorithme d'approximation dans le cas d'un dual non intégral. Dans le cas où le dual n'est pas intégral, on peut explorer les forêts couvrantes après relaxation du problème. Une autre difficulté surgit liée au fait que le nombre d'inégalités du système est exponentiel. En effet, pour chaque sous ensemble de sommets, une contrainte apparait dans le fait que le nombre d'arêtes internes doit être strictement inférieur au cardinal de l'ensemble de sommets. Le chapitre 6 propose un modèle polynomial de contraintes par rapport au cardinal de l'ensemble de sommets en introduisant une orientation arbitraire des arêtes, des résultats numériques sont présentés dans une mise en oeuvre du modèle. Les problèmes stochastiques continus sont abordés au cours du chapitre 7 dans le cadre du problème de sac à dos avec contrainte stochastique. La formulation est de type "chance constraint", et la dualisation par variable lagrangienne est adaptée à une situation où la probabilité de respecter la contrainte doit rester proche de 1. Le modèle étudié est celui d'un sac à dos où les objets ont une valeur et un poids déterminés par des distributions normales. Cette situation présente un certain nombre d'avantages calculatoires. En premier lieu, la contrainte étant linéaire, son expression devient une espérance d'une loi normale. Cette formulation permet de s'affranchir de problèmes de convexité, voire de connexité de l'espace admissible des solutions. De plus, la loi normale étant déterminée par sa moyenne et son écart-type, il est possible de géométriser complètement le problème. C'est cette particularité qui est exploitée par dans la littérature pour affirmer que le problème est convexe pour $p > \frac{1}{2}$, mais c'est également la même particularité qui permet de mettre en oeuvre une résolution par la méthode du "second order cone programming". Dans notre approche, nous nous attachons à appliquer des méthodes de gradient directement sur la formulation en espérance de la fonction objectif et de la contrainte. Nous délaissons donc une possible reformulation du problème sous forme géométrique pour détailler les conditions de convergence de la méthode du gradient stochastique. Cette partie est illustrée par des tests numériques de comparaison avec la méthode SOCP

sur des instances combinatoires avec méthode de Branch and Bound, et sur des instances relaxées.

1.1 chapitre 1

Le premier chapitre introduit les outils mathématiques nécessaires pour modéliser et traiter les problèmes combinatoires stochastiques discrets. Les concepts introduits sont les matroïdes :

Definition 1 *Soit $N = \{1, \ldots, n\}$ un ensemble fini, et une collection $\mathscr{F}$ de sous ensembles. $(N, \mathscr{F})$ est un système d'indépendants si*

$$\forall F_1 \in \mathscr{F}, \forall F_2 \subset N, F_2 \subset F_1 \Rightarrow F_2 \in \mathscr{F}.$$

Les éléments de $\mathscr{F}$ sont appelés les ensembles indépendants, les autres ensembles sont appelés les ensembles dépendants.

Definition 2 *Considérons un système d'indépendants $(N, \mathscr{F})$, un sous-ensemble $F \in \mathscr{F}$ est appelé indépendant maximal si $F \cup \{j\} \notin \mathscr{F}$ pour tout $j \notin F$.*

Definition 3 *Un indépendant T est maximum si $|S| \leq |T|$ pour tout $S \in \mathscr{F}$.*

On introduit la notation $m(T) = \{\max_{S \subset T} |S| : S \in \mathscr{F}\}$.

Definition 4 *$M = (N, \mathscr{F})$ est un matroïde si M est un système d'indépendants tel que pour tout sous-ensemble $T \subset N$, tout indépendant contenu dans T qui est maximal dans T a le même cardinal $m(T)$.*

Le second concept introduit dans ce chapitre est celui de formulation duale. Considérons le problème de maximisation suivant : $c \in \mathbb{R}^n$, $A \in \mathscr{M}_{pn}(\mathbb{R})$, $x \in \mathbb{R}^n$ et $b \in (\mathbb{R}^+)^p$

$$\mathrm{Z}_{LP} = \begin{cases} max \sum_{j=1}^{n} c_j x_j \\ Ax \leq b \end{cases} \quad (1.1)$$

Le problème dual associé est :

$$\mathrm{Z}_{LD} = \begin{cases} min \sum_{i=1}^{p} b_i y_i \\ A^T y \geq c \end{cases} \quad (1.2)$$

Le troisième concept introduit est celui de système totalement dual intégral (TDI). Un tel système est caractéristique d'un polyèdre de contraintes dont les sommets sont à coordonnées entières.

Definition 5 *Un système d'inégalités linéaires $Ax \leq b$ est TDI si, pour tout vecteur entier c tel que $Z_{LP} = max\{cx : Ax \leq b\}$ admet une solution finie, le dual $Z_{LD} = min\{yb : A^t y \geq c, y \in (\mathbb{R}^+)^p\}$ a une solution optimale entière.*

Proposition 6 *Si $Ax \leq b$ est TDI et b est entier, alors $P = \{x \in \mathbb{R}^n : Ax \leq b\}$ est entier. Par conséquent, pour tout c tel que Z_{LP} est fini, Z_{LP} est entier.*

1.2 Chapitre 2

Ce chapitre introduit le problème de la forêt couvrante de poids maximal. Considérons un graphe $G = (V, E)$ et une fonction de poids c associée aux arêtes. Pour un graphe donné, les sous-ensembles d'arêtes ne comportant pas de cycles forment un matroïde. Le problème de la forêt couvrante de poids maximal consiste à sélectionner une forêt sans cycle maximisant la somme des poids sélectionnés. $r(S)$ désigne le cardinal maximum d'un ensemble indépendant contenu dans l'ensemble S. la La formulation est la suivante :

$$\mathrm{Z}_{IP} = \begin{cases} max \sum_{j=1}^{n} c_j x_j \\ x \in \{0,1\}^n \\ \sum_{j \in S} x_j \leq r(S) \quad \forall S \subseteq E \end{cases} \quad (1.3)$$

et la formulation duale associée :

$$\mathrm{Z}_{LD} = \begin{cases} min \sum_{S \subseteq E} r(S) y_S \\ \sum_{S : j \in S} y_S \geq c_j \quad \forall j \in E \\ y_S \geq 0 \quad \forall S \subset E \end{cases} \quad (1.4)$$

Ce type de problème est TDI grâce à l'algorithme glouton, qui est présenté dans ce chapitre. Ce chapitre présente les premiers résultats relatifs à la continuité des solutions en fonction des valeurs c de la fonction de poids. Le chapitre introduit égalemetn la formulation stochastique de ce problème. On considère que certains poids suivent des variables aléatoires discrètes tandis que d'autres ont un coût déterminé par avance. Celles de coût fixe sont dites de premier niveau, tandis que celles de coût variable son dites de second niveau. La maximisation du coût des arêtes choisies se transforme en une maximisation de l'espérance de coût. Les arêtes fixes peuvent être choisies par avance, tandis que les arêtes de poids variable sont choisies après réalisation d'un tirage de variable aléatoire. La loi de probabilité est discrète, de sorte qu'on parle de scénarios en nombre fini. Le problème se formule ainsi :

$$z_{LP} = \begin{cases} max \sum_{j \in X} c_j x_j + \sum_{k=1}^{k=K} \pi_k \sum_{j \in Y} c_{jk} z_{jk} : \\ \sum_{j \in S \cap X} x_j + \sum_{j \in S \cap Y} z_{jk} \leq r(S), \quad k \in \{1, \ldots, K\}, \quad \forall S \subseteq E \\ (x, z_k) \in [0,1]^n \times [0,1]^q, k \in \{1, \ldots, K\}. \end{cases} \quad (1.5)$$

et son dual :

$$z_{LD} = \begin{cases} min \sum_{k=1}^{K} \sum_{S \subseteq E} r(S) y_{S,k} : \\ \sum_{k=1}^{K} \sum_{S \subseteq E : i \in X \cap S} y_{S,k} \geq c_i, \quad i \in X \\ \sum_{S \subseteq E : j \in Y \cap S} y_{S,k} \geq \pi_k c_{jk}, \quad j \in Y, k \in \{1, \ldots, K\} \\ y_{S,k} \geq 0, \quad k \in \{1, \ldots, K\}, \quad S \subseteq E. \end{cases} \quad (1.6)$$

Le chapitre se termine par une introduction générale à la littérature sur le sujet.

1.3 chapitre 3

Le chapitre 3 traite le cas où il n'existe que deux réalisations possibles pour le poids des arêtes de second niveau. On dit que le problème a deux scénarios. Dans ce cas, le système associé conserve le caractère TDI du cas déterministe. La preuve passe par un découpage formel du poids des arêtes de coût fixe en deux coûts partiels. Le système d'inégalités est ensuite scindé en deux systèmes ne concernant que des poids relatifs à l'un ou l'autre des scénarios. Les deux systèmes sont maximisés séparément, puis la somme des deux optimums est comparée au dual du système initial. Le chapitre donne des conditions nécessaires et suffisantes pour que les solutions optimales des différents systèmes coïncident. Le résultat majeur est que pour toute formulation avec seulement deux scénarios, le problème de la forêt de poids couvrant maximal conserve son caractère TDI.

1.4 chapitre 4

Le cas de trois scénarios ou plus est abordé dans ce chapitre. Dans ce cas, un premier contre-exemple de configuration non TDI est donné. Ce premier contre-exemple est prolongé par une analyse d'une classe de graphes ne conduisant pas à un système TDI. Dans de tels graphes, il existe au moins trois scénarios et trois arêtes de premier niveau associées respectivement à chacun de ces trois scénarios, telles qu'une seule de ces arêtes de premier niveau n'appartiennent à aucun cycle pour les scénarios respectifs, mais que dans les deux autres scénarios, elles soient engagées dans des cycles où elles sont les arêtes de poids le plus faible. Ce type de configuration entraîne qu'il n'est pas possible de produire une répartition formelle des poids des arêtes de premier niveau de manière à leur conférer un statut identique dans tous les scénarios.

La seconde partie du chapitre propose une réduction du problème multi-scénarios à un problème connu de classe NP : le problème de la famille couvrante de cardinal minimal. Considérons $\mathscr{S} = \{S_1, S_2, \ldots, S_m\}$ une famille couvrante de parties de V. Le problème de la famille couvrante de cardinal minimal consiste à déterminer une sous famille de $\mathscr{S}$ minimale qui couvre tous les éléments de V. Nous construisons une instance de ce problème où sont ajoutés m points correspondant à chaque sous ensemble $\mathscr{S}_i$, et un point supplémentaire source r. Les arêtes de premier niveau correspondent aux arêtes entre la source r et les m nouveaux points créés. Le choix de ces arêtes est équivalent à la sélection en premier niveau d'une sous famille d'ensembles de $\mathscr{S}$. Nous construisons des scénarios tels qu'il est nécessaire de couvrir chaque sommet de V mais avec un choix minimal d'arêtes de premier niveau.

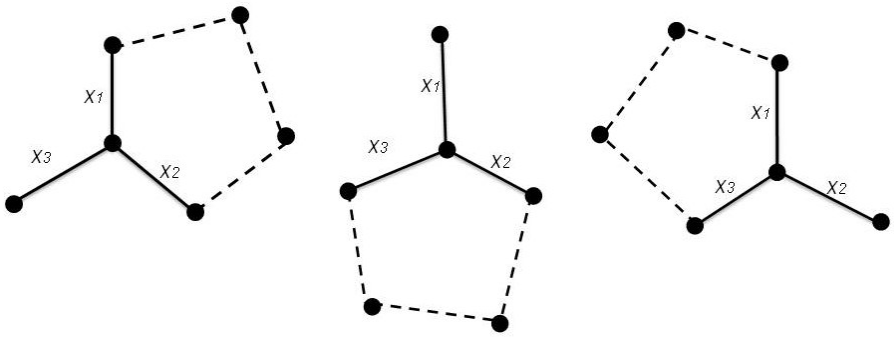

Figure 1.1: Cas général de graphe non TDI : 3 scenarios avec 3 cycles comportant deux arêtes de premier niveau et une arête de premier niveau isolée dans chaque scénario.

1.5 chapitre 5

Dans le cas de trois scénarios ou plus, le système n'étant pas TDI, nous abordons la question de l'approximation. De nouveaux résultats sont présentés sur le caractère fractionnaire des solutions optimales au problème Z_{LD}. En particulier, l'analyse montre que les solutions entières ne peuvent comporter toutes les arêtes fractionnaires de la solution Z_{LP}. Nous proposons donc d'exclure certaines arêtes de premier niveau en réduisant leur poids, de soirte qu'elles ne fassent plus partie de la solution gloutonne dans aucun scénario. Nous construisons une solution optimale entière à un problème voisin, puis nous réappliquons une variation au poids des arêtes écartées restant compatibles aevc une configuration gloutonne inchangée. Ce Chapitre s'appuie sur des illustrations numériques et nous fournissons également une évaluation de la borne d'approximation.

1.6 chapitre 6

Une autre difficulté du problème étudié est le caractère exponentiel du nombre de contraintes, donc de leur formulation. Chaque sous-ensemble de k sommets donne lieu à $(k-1)!$ contraintes. Pour contourner cette difficulté, nous proposons d'orienter aléatoirement les arêtes du graphe. Nous étudions alors les configurations possibles d'un sous graphe présentant un cycle. Certaines configurations sont caractéristiques d'un sous graphe présentant un cycle; EN particulier, si un sous graphe présente un cycle, il existe un sous ensemble d'arêtes tel que deux arêtes pointent sur le même sommet, ou alors il est possible de trouver indéfiniment un successeur à tout sommet en parcourant les arêtes dans le sens de leur orientation. La formulation de ces situations en termes de variables d'orientation nécessite un

nombre polynomial d'expression de contraintes. Ces formulations sont testées sur des arbres générés aléatoirement.

1.7 chapitre 7

Ce dernier chapitre est en réalité une seconde partie indépendante des six premiers chapitres de la thèse. Le chapitre porte sur le problème du sac à dos stochastique. Soit n objets ayant une valeur r_i pour $i \in \{1, \ldots, n\}$ et un poids p_i pour $i \in \{1, \ldots, n\}$ égal à leurs valeurs respectives. On suppose que les poids et les valeurs des objets sont des variables aléatoires continues qui suivent des lois de distribution normale notées χ_i. Un sac à dos ayant une contenance c, on souhaite choisir des objets afin de maximiser l'espérance de valeur des objets placés dans le sac. Les objets ayant un poids aléatoire, la contrainte de capacité est elle même une variable aléatoire. Le respect de la contrainte de capacité est exprimé selon une formulation dite de "chance constraint" : le poids total des objets choisis n'excède pas la capacité selon une probabilité p déterminée par avance et généralement proche de 1.

Ce problème se formule ainsi :

Chance Constrained Knapsack Problem ($CCKP$)

$$\max_{x \in \{0,1\}^n} \mathbb{E}\left[\sum_{i=1}^{n} r_i \chi_i x_i\right] \tag{1.7}$$

$$\text{s.t.} \qquad P\{g(x, \chi) \leq c\} \geq p \tag{1.8}$$

Où $\mathbb{E}[\cdot]$ est l'espérance de valeur des objets,$g(x, \chi) = \sum_{i=1}^{n} \chi_i x_i$ est le poids total des objets sélectionnés et $p \in (0.5, 1]$ est un seuil de probabilité prescrit par avance.

La contrainte probabiliste est reformulée en une contrainte en espérance, mais l'espérance est alors calculée sur une fonction non régulière (fonction de Heaviside), de plus, la contrainte est intégrée à la fonction objectif par une technique de multiplicateur de Lagrange :

$$\mathcal{L}(x, \lambda) \quad = \mathbb{E}\left[\textstyle\sum_{i=1}^{n} r_i \chi_i x_i\right] - \lambda \left(p - \mathbb{E}\left[\mathbb{H}_{\mathbb{R}^+}(c - g(x, \chi))\right]\right),$$

où $\mathbb{H}_{\mathbb{R}^+}$ désigne la fonction indicatrice de l'ensemble des réels positifs et λ est le multiplicateur de Lagrange associé à la contrainte.

Pour traiter ce problème, nous employons une méthode de gradient stochastique appelée Arrow-Hurwicz. Cette méthode emploie une double descente en gradient spatial et selon le multiplicateur de Lagrange. Le calcul des gradients porte donc sur les grandeurs intégrées. D'ordinaire, des techniques de convolution sont utilisées pour substituer une fonction infiniment dérivable à la grandeur dont on calcule l'espérance. Mais ces techniques sont déjà un processus d'approximation de la grandeur intégrée. Nous utilisons donc une technique d'intégration par parties qui substitue une nouvelle fonction régulière à la contrainte dont est calculée

l'espérance.

$$\mathcal{L}(x,\lambda) = \mathbb{E}[\sum_{i=1}^{n} r_i \chi_i x_i] - \lambda \left(p + \mathbb{E}[\mathbb{Y}_{\mathbb{R}^+}(c - g(x,\chi)) \frac{(\chi_\kappa - \mu_\kappa)}{x_\kappa \sigma_\kappa^2}] \right). \tag{1.9}$$

où $\mathbb{Y}(x) = \int_0^x \mathbb{H}_{\mathbb{R}^+}$ et κ est un indice d'un objet particulier choisi pour opérer l'intégration par parties sur la loi normale χ_κ de moyenne μ_k et de variance σ_k. L'ensemble du travail théorique exposé dans ce chapitre examine les conditions de convergence de l'algorithme stochastique appliqué sur la reformulation du problème, ainsi que les choix techniques opératoires pour assurer une convergence efficace (en particulier le choix de la variable d'intégration par parties κ).
Une seconde partie du chapitre expose les résultats obtenus en application à un problème combinatoire de type sac à dos. Une comparaison des performances est établie avec une méthode de type SOCP.