Approximation efficace de mélanges bootstrap d'arbres de Markov pour l'estimation de densité

F. Schnitzler[1] S. Ammar[2] P. Leray[2] P. Geurts[1] L. Wehenkel[1]

fschnitzler@ulg.ac.be

[1]Université de Liège

[2]Université de Nantes

25 mai 2012

But : estimation d'une densité de probabilité conjointe sur un grand nombre de variables.

But à long terme :

- Bioinformatique
- Réseaux électriques (16 000 noeuds de transmission en Europe)

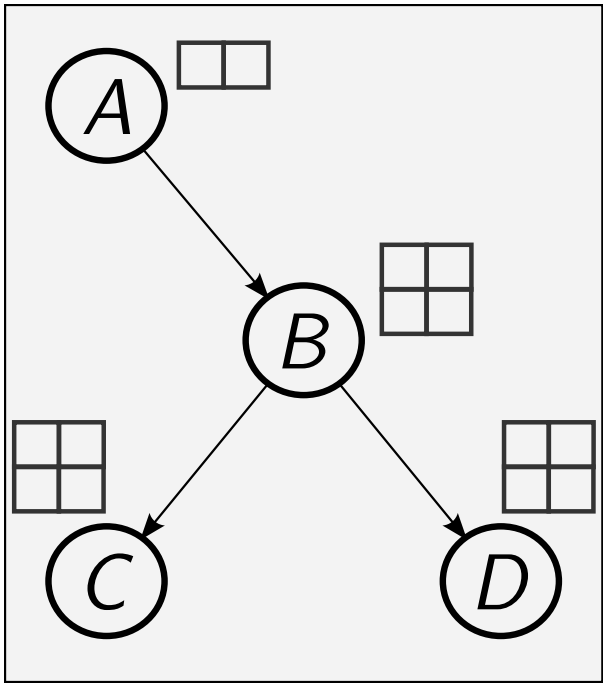
Deux problèmes principaux :

- Complexité algorithmique $\rightarrow$ modèles simples (arbres de Markov)
- Peu d'échantillons : grande variance $\rightarrow$ méthodes d'ensemble

Dans cette présentation :

- Accélération du bagging d'arbres de Markov : 10 fois plus rapide
- Résultats empiriques

Un arbre de Markov encode une distribution de probabilités sur n variables $\mathcal{X}$.



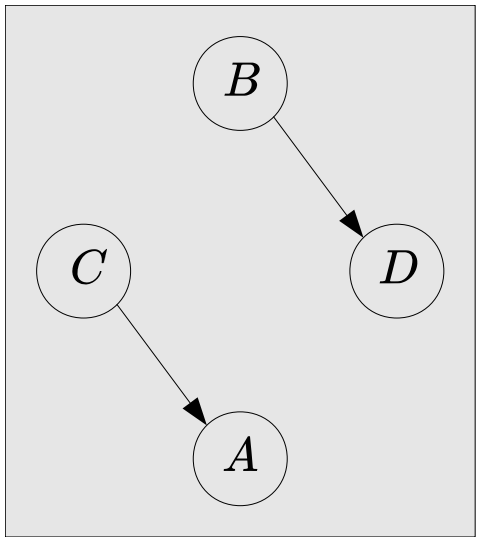
- Pas de cycle, chaque variable a un seul parent (racine exceptée)
- Inférence : $\mathcal{O}(n)$
- Apprentissage au maximum de vraisemblance : $\mathcal{O}(n^2 \log n)$

$$P_T(\mathcal{X}) = P(A)P(B|A)P(C|B)P(D|B)$$

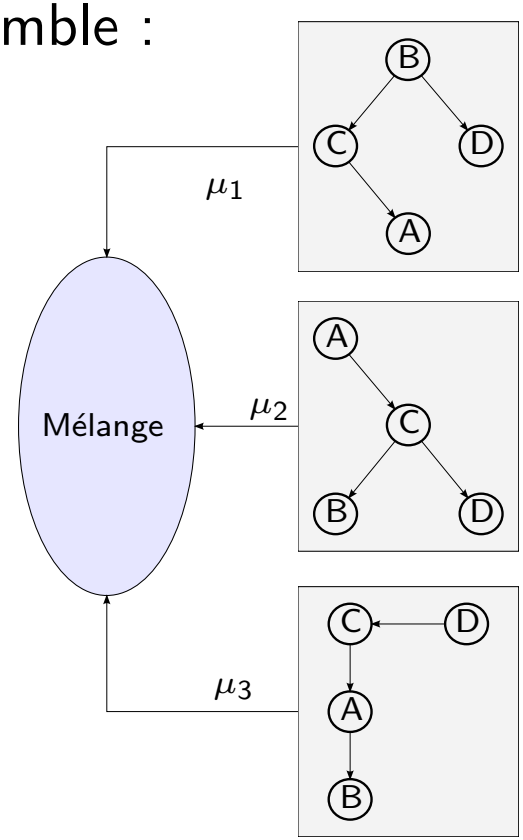
Factorisation : produit des densités marginales de chaque variable, conditionnellement à ses parents dans le graphe.

Un mélange d'arbres possède certaines propriétés intéressantes des arbres.

Une **forêt** est un arbre, moins quelques arcs :



Un **mélange** d'arbres est une méthode d'ensemble :



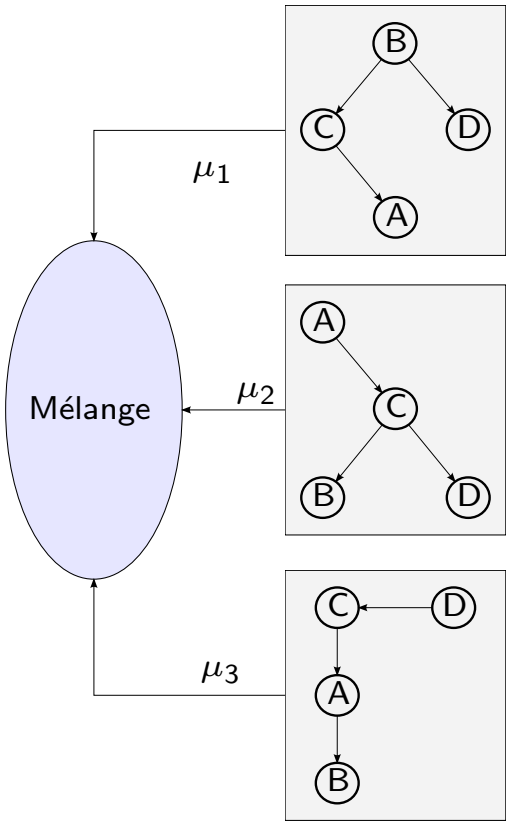
$$\mathbb{P}_{\hat{\mathcal{T}}}(X) = \sum_{i=1}^{m} \mu_i \mathbb{P}_{\mathcal{T}_i}(X)$$

Un mélange d'arbres possède certaines propriétés intéressantes des arbres.

- Plusieurs modèles $\rightarrow$ modélisation améliorée
- Modèles simples $\rightarrow$ faible complexité :
  - inférence : linéaire,
  - apprentissage : logquadratique.

Il y a deux types de mélanges :

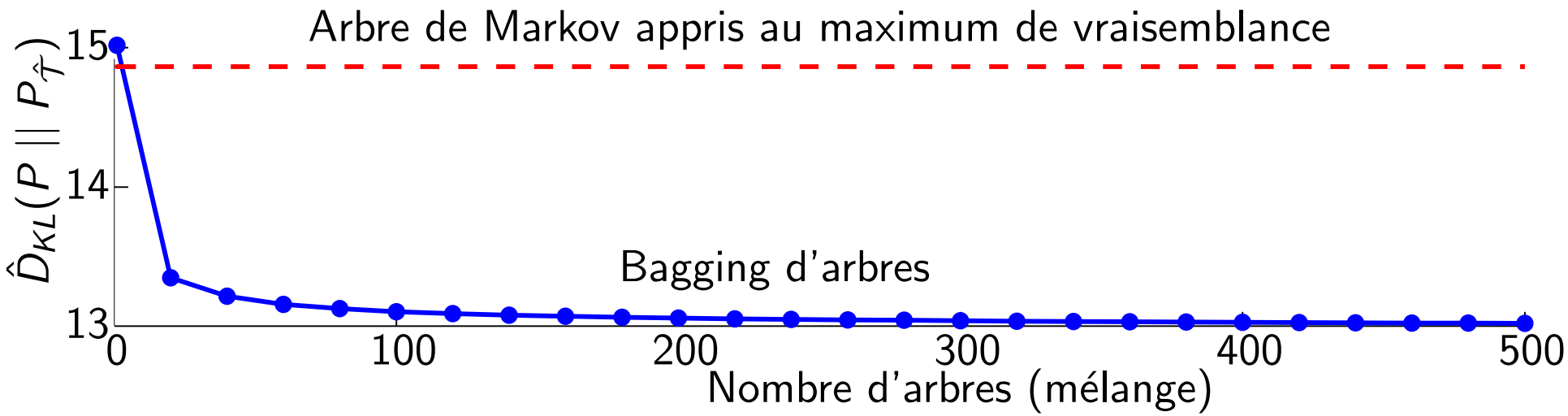
- Réduction du biais
  - ex : mélanges de gaussiennes
- Réduction de variance
  - ex : random forests (classification)



Le bagging réduit la variance.

- moyenne sur m arbres de Markov appris chacun sur un réplicat bootstrap :
  - $\rightarrow$ présente typiquement une variance plus faible,
  - $\rightarrow$ moins de surapprentissage.
- Un réplicat bootstrap $\mathbf{D}'$ d'un ensemble d'apprentissage $\mathbf{D}$ est échantillonné avec remise dans $\mathbf{D}'$.
- Complexité : $\mathcal{O}(mn^2 \log n)$

Exemple : 200 variables et 200 observations (problème synthétique)



Nous développons une approximation pour accélérer le bagging.

Complexité : $\mathcal{O}(mn^2 \log n)$

- But : accélérer l'apprentissage sans sacrifier la précision.
- Motivation : il faut beaucoup d'arbres : le mélange est d'autant meilleur.
- Le terme quadratique vient du nombre d'arcs considérés pour chaque arbre.

$$T_i(\mathbf{D}') = \arg\max_T \sum_{(X,Y)\in\mathcal{E}(T)} I_{\mathbf{D}'}(X;Y) \ ,$$

Réplicat

A	B	C	D
0	1	0	1
1	1	0	1
0	0	1	1
1	1	1	0

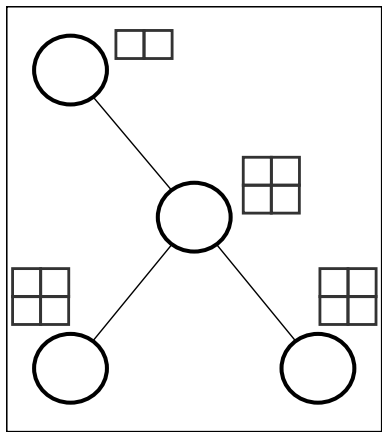
$I(X,Y)$

Poids des arcs

	A	B	C	D
A		*	*	*
B	*		*	*
C	*	*		*
D	*	*	*	

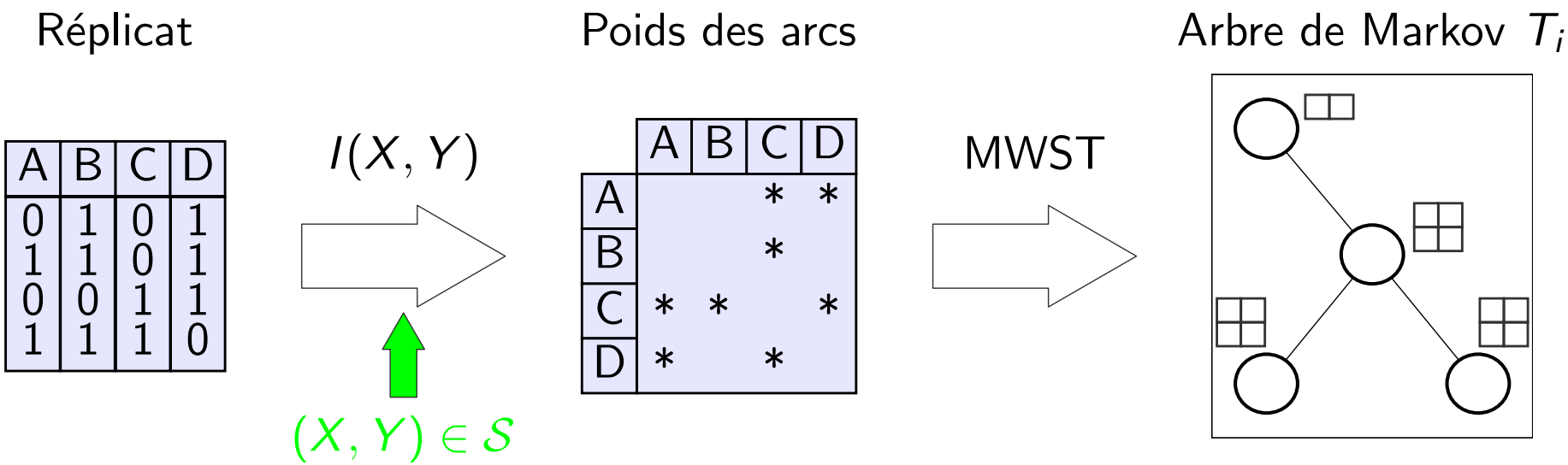
MWST

Arbre de Markov T_i



Approximation : ne considérer qu'un sous-ensemble d'arcs.

- Idées :
  - premier arbre appris au maximum de vraisemblance sur les données.
  - exploiter ce calcul pour obtenir un bon ensemble $\mathcal{S}$ d'arcs candidats, utilisés pour les arbres suivants.
    - $\rightarrow$ les termes du mélange ne sont plus indépendants.
    - $\star$ Seul le sous-ensemble d'arbres (ou forêts) inclus à $\mathcal{S}$ est exploré.
- Résultat :
  - complexité : $\mathcal{O}(mn^2 \log n) \rightarrow \mathcal{O}(n^2 \log n + m|\mathcal{S}| \log |\mathcal{S}|)$
  - temps de calcul : un ordre de grandeur plus rapide.

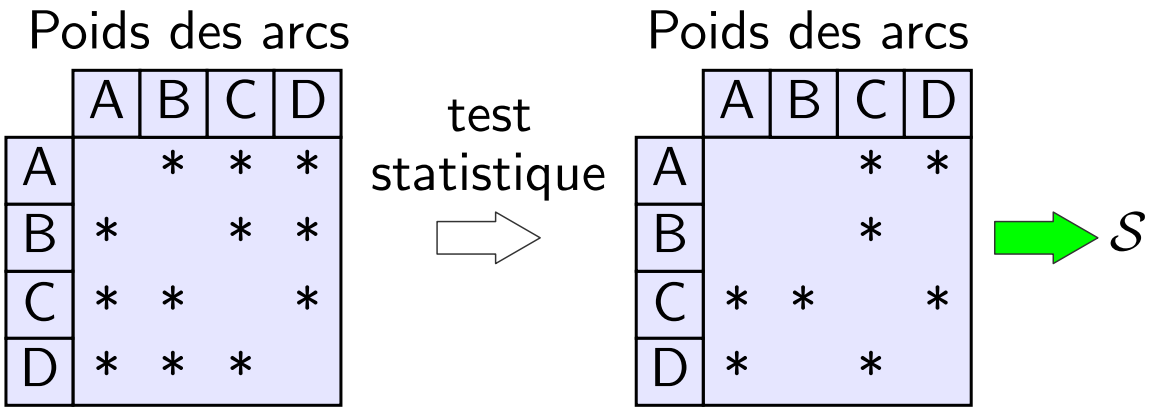


$\mathcal{S}$ doit contenir les arcs dont l'information mutuelle est élevée.

- Des arcs dont l'information mutuelle est faible
  - ont peu de chance de faire partie d'un arbre (même si les poids sont perturbés),
  - sont probablement peu significatifs (bruit, ou relation indirecte).

  $\rightarrow$ Ils peuvent sans doute être ignorés.

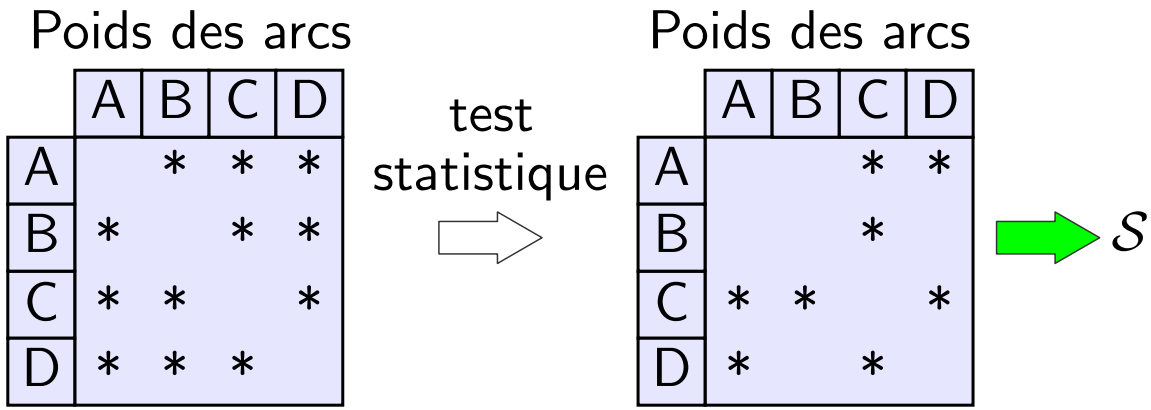
Premier arbre :



Un arc est inclu dans $\mathcal{S}$ en fonction d'un test d'indépendance.

- $I_{\mathbf{D}}(X;Y)$ est comparé à une valeur limite dépendant d'une risque de première espèce choisi, par exemple $\alpha = 0.05$.
- Similaire à une régularisation :
  $T_{CL}^{\lambda}(\mathbf{D}) = \arg\max_T \sum_{(X,Y)\in\mathcal{E}(T)} I_{\mathbf{D}}(X;Y) - \lambda|T|$
- $\mathcal{S}$ contient les paires de variables dont l'information mutuelle (sur les données initiales) est supérieure à la valeur seuil.

Premier arbre :



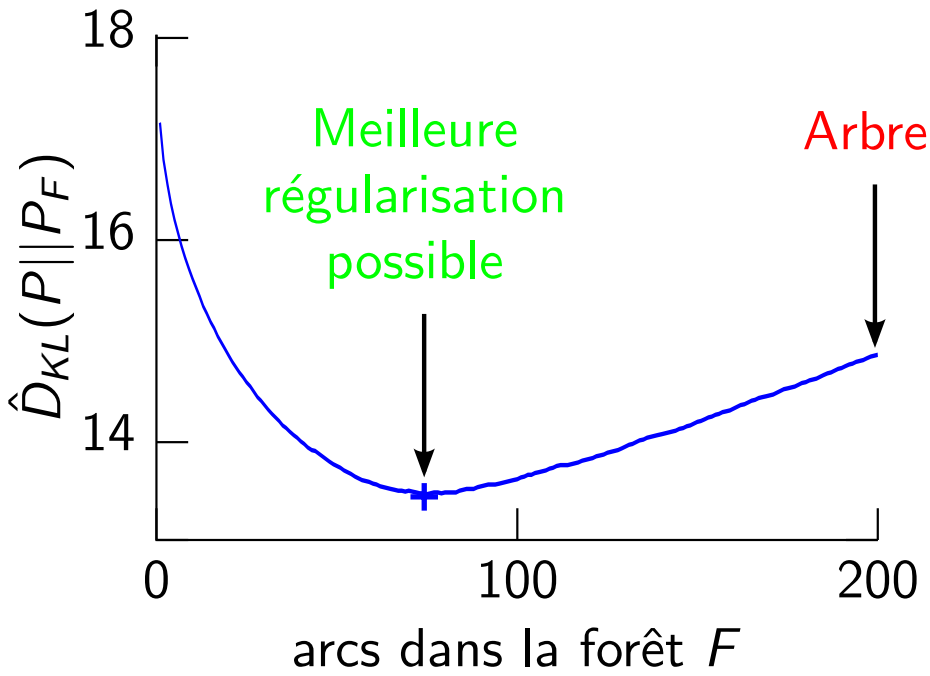
Régulariser est un autre moyen de réduire la variance.

Arbre régularisé :

$$T_{CL}^{\lambda}(\mathbf{D}) = \arg\max_{T} \sum_{(X,Y)\in\mathcal{E}(T)} I_{\mathbf{D}}(X;Y) - \lambda |T|$$

λ est ici optimisé sur l'ensemble de test.

$\rightarrow$ Meilleure régularisation possible (pour comparaison)



Ces algorithmes sont évalués sur des problèmes synthétiques et plus réalistes.

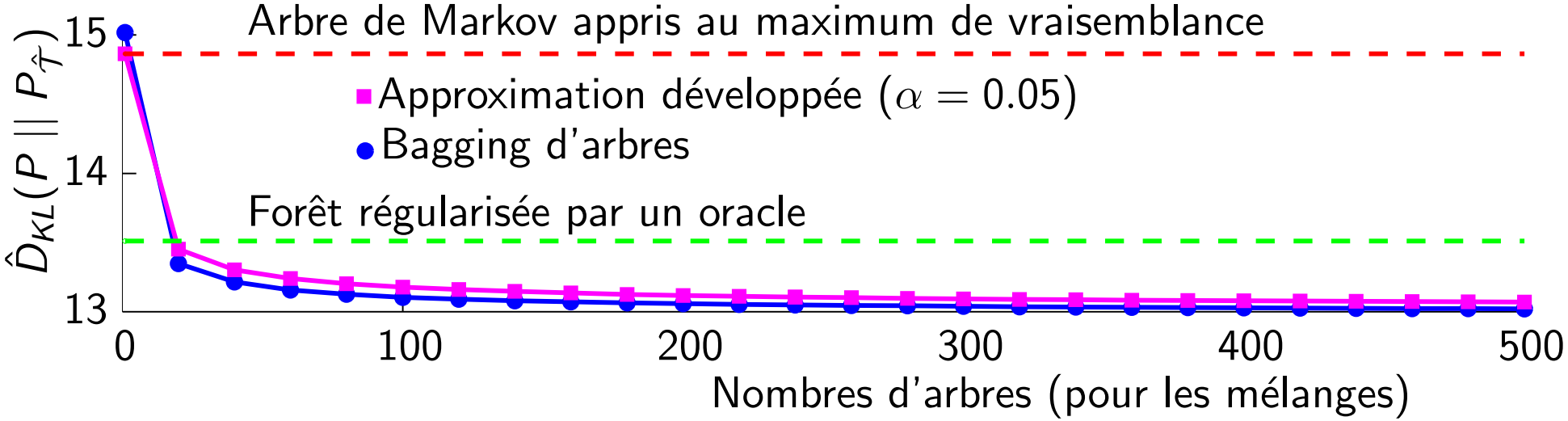
Réseaux synthétiques :

- Pour chaque X_i
  - nombre de parents aléatoirement choisis dans $[0, max(5, i-1)]$
  - parents choisis aléatoirement dans $\{X_1, ..., X_{i-1}\}$.

- 200 et 1000 variables ; 200, 600 et 1000 observations.
- Validation par estimation Monte-Carlo de la divergence de Kullback-Leibler entre le modèle réel et le mélange.

Les deux mélanges ont souvent une précision similaire.

200 échantillons, 200 variables :

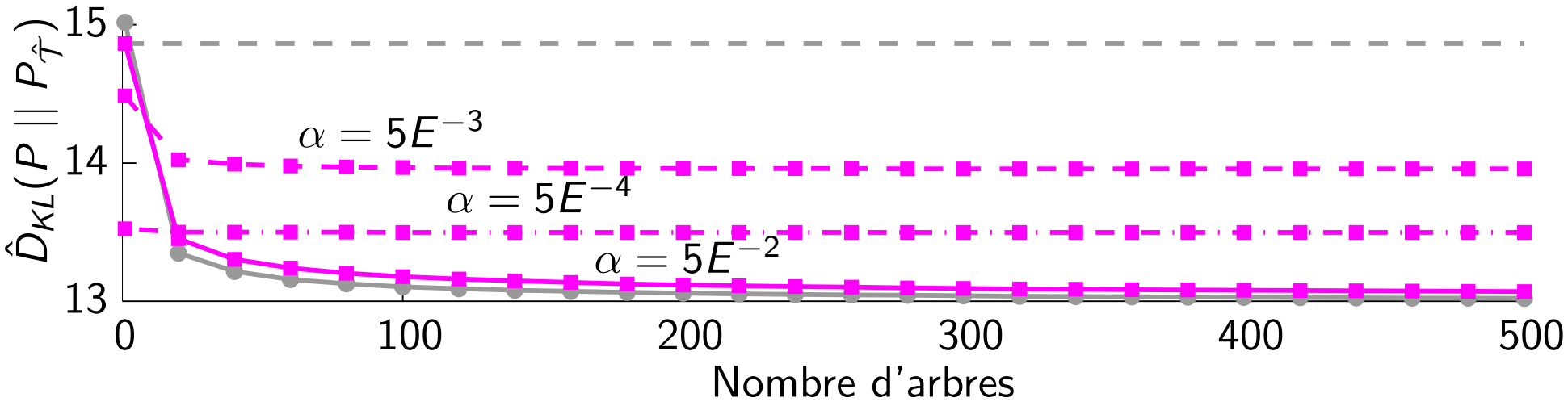


Temps de calcul relatif pour un mélange de 500 arbres (premier arbre : 1) :

- Bagging d'arbres : 532
- Approximation : 21

Influence du risque de première espèce α :

200 variables, 200 observations :



- Plus α est petit, plus faible est la variance du premier arbre.
  - $\rightarrow$ Ici, amélioration de la précision.
  - Augmentation du biais.
- Plus α est grand, plus la convergence est lente, mais meilleure est la précision.
  - $\rightarrow$ Plus grande diversité dans les arbres.
  - $\rightarrow$ Meilleure réduction de la variance due au mélange.
  - Le biais des arbres est également meilleur.

Problèmes plus réalistes.

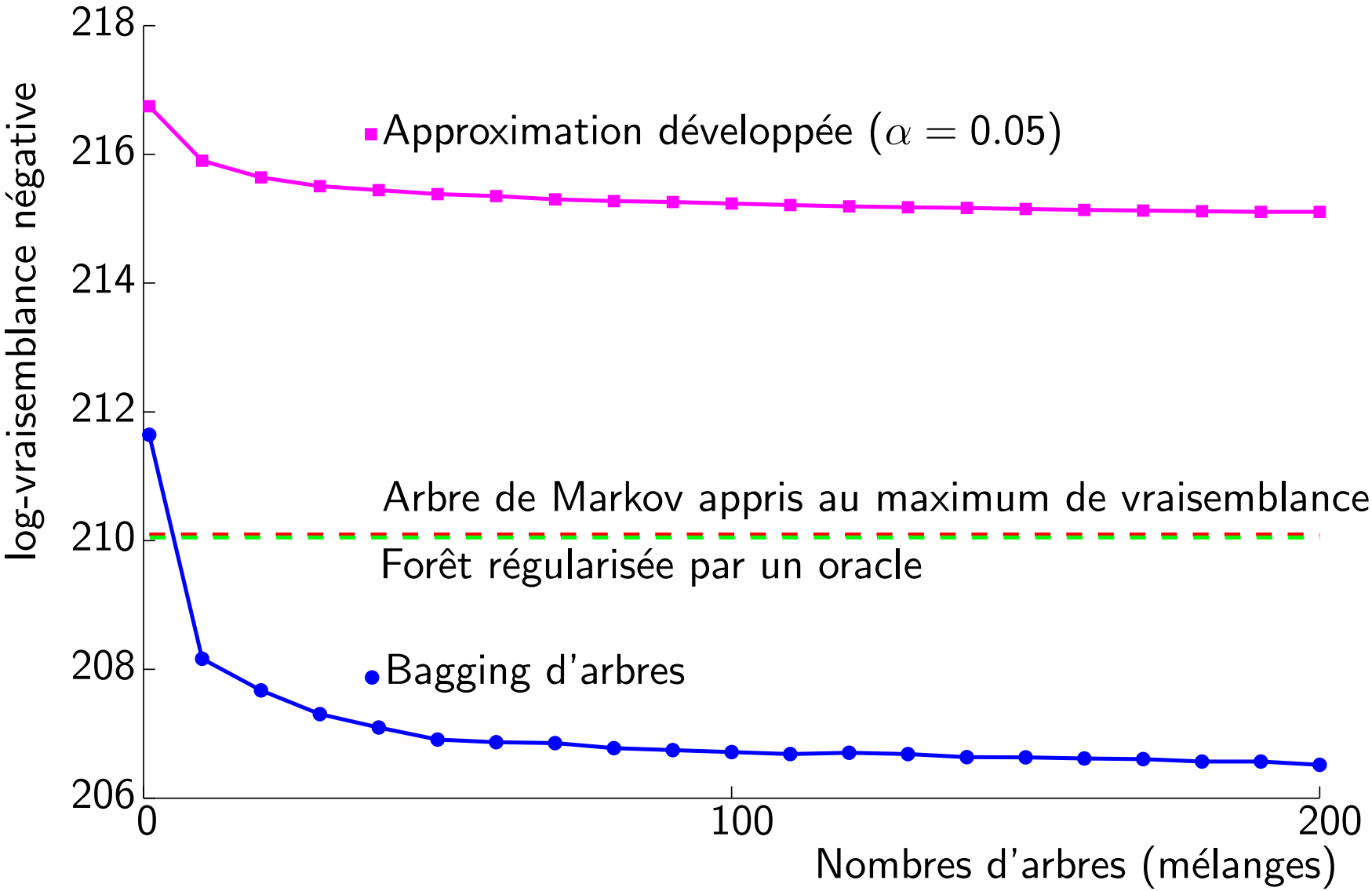
- 8 modèles comptant entre 200 et 801 variables ; 200 et 500 observations :
  - 4 distributions classiques (Child10, Insurance10, Alarm10, Hailfinder10)
  - 2 modèles ressimulés à partir de données d'expression génétique (Gene, Lung Cancer)
  - 2 systèmes experts (Munin,Pigs)
- Score : log-vraisemblance négative de 5000 observations indépendantes.

$\alpha = 0.05$ pour l'approximation.
Résumé de la précision de l'approximation par rapport au bagging :

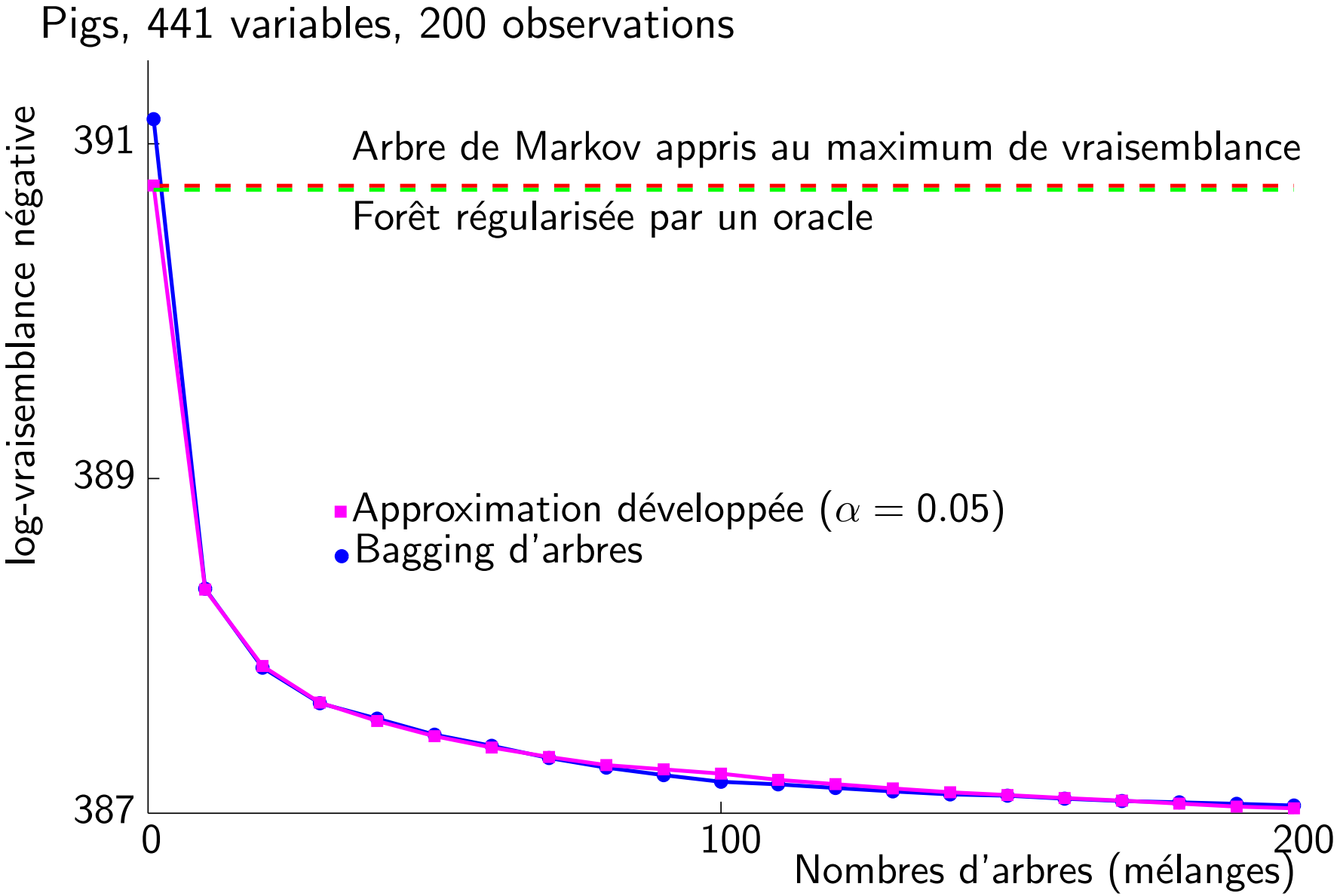
- Approximation moins bonne que Bagging : 3 configurations sur 16
- Approximation $\approx$ Bagging : 9/16
- Approximation meilleure que Bagging : 4/16

Exemple où le bagging est meilleur que l'approximation.

Insurance10, 270 variables, 200 observations

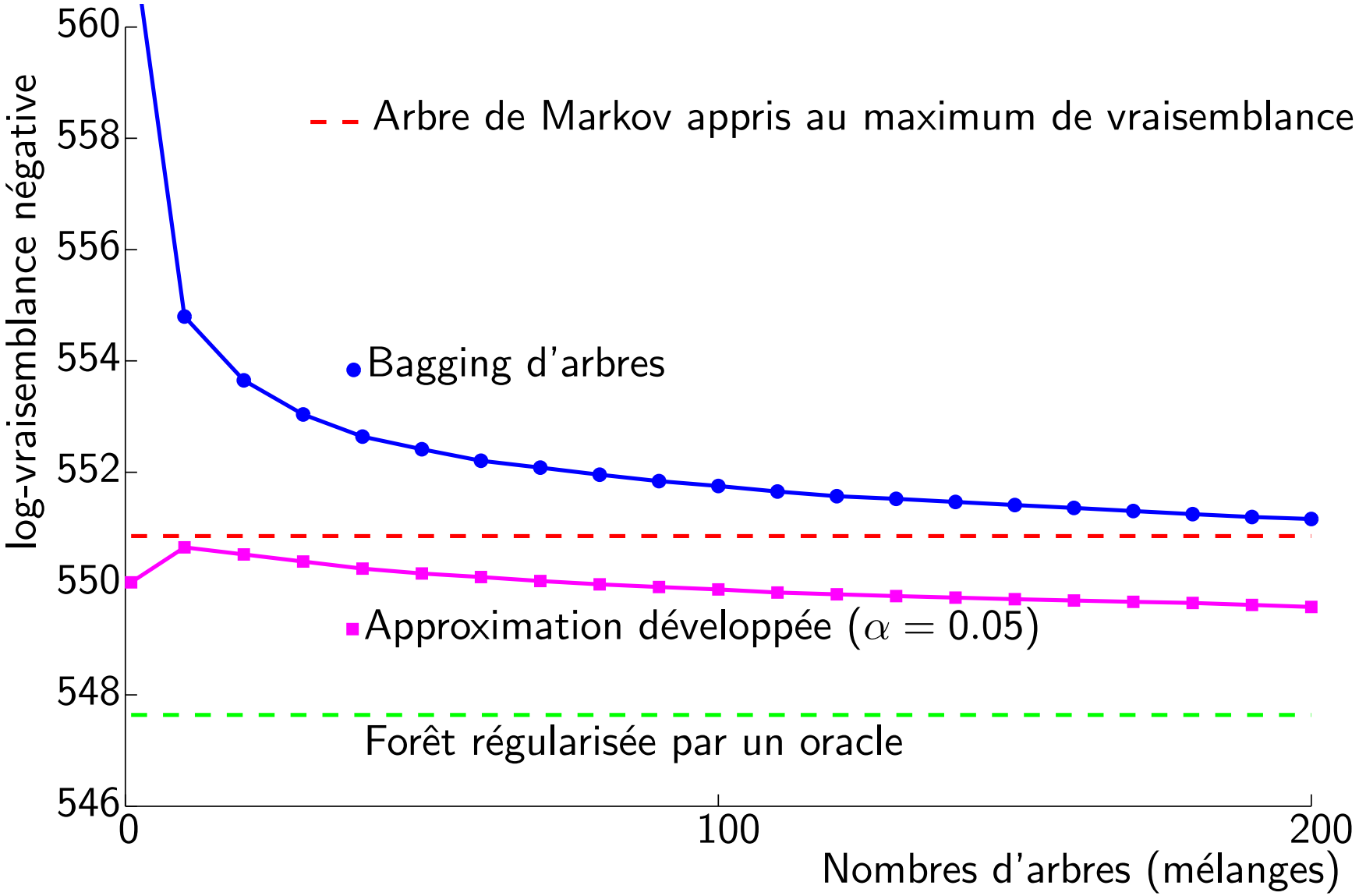


Exemple où la précision des deux mélanges est similaire.



Exemple où l'approximation est meilleure que le bagging.

Hailfinder10, 560 variables, 200 observations



Conclusions

- Nous développons un algorithme pour apprendre un mélange d'arbres de Markov, avec une précision généralement similaire au bagging d'arbres, et un temps d'apprentissage plus court.
- Il exploite le calcul du premier arbre du mélange pour réduire le nombre d'arcs évalués pour les arbres suivants.
- Les arcs sont sélectionnés sur base d'un test d'indépendance.
- Choisir un bon risque de première espèce instance pour le test est nécessaire.

TABLE: Influence d'α sur le nombres d'arcs, moyenne sur 5 densités fois 6 ensembles de données, pour $n = 1000$ variables et $p = 200$ observations

Edges	Nombres d'arcs (% du total) for $\alpha =$			
	$1E^{-1}$	$5E^{-2}$	$5E^{-3}$	$5E^{-4}$
T_1	998	997.9	993.2	626.8
$\mathcal{S}$	52278(10.5%)	26821(5.36%)	3311(0.66%)	683 (0.13%)

TABLE: Temps de calcul série minimum

Méthode	temps relatif minimum pour l'apprentissage					
	n=200, m=500 - sauf CL			n=1000, m=100 - sauf CL		
	p=200	p=600	p=1000	p=200	p=600	p=1000
CL	1	3.07	5.3	37	98	174
Bagg.	532	1531	2674	5037	11662	19431
Approx.	21	82	191	139	612	1005

TABLE: Log-vraisemblance négative (moyenne sur 5 ensembles d'apprentissage)

Distribution	n	N	CL	regCL	Bagg.	Approx.
Alarm10	370	200	166.65	166.65	**163.59**	166.80
Alarm10	370	500	136.37	136.28	**135.31**	135.61
Child10	200	200	135.29	135.08	**133.94**	134
Child10	200	500	131.71	131.71	**131.01**	131.02
Gene	801	200	485.21	483.6	482.80	**482.66**
Gene	801	500	477.48	476.75	**473.75**	473.79
Hailfinder10	560	200	550.85	**547.64**	551.75	549.89
Hailfinder10	560	500	523.81	**523.26**	523.61	523.31
Insurance10	270	200	210.1	210.1	**206.77**	215.23
Insurance10	270	500	198.87	198.87	**195.47**	202.01
Lung Cancer	800	200	435.72	**435.46**	437.41	436.01
Lung Cancer	800	500	424.69	424.44	418.31	**418.30**
Munin	189	200	42.614	36.987	41.799	**35.566**
Munin	189	500	37.66	35.414	37.656	**35.140**
Pigs	441	200	390.75	390.75	**387.19**	387.24
Pigs	441	500	385.59	385.59	**382.22**	382.26