TP - 2
Parallélisation du prétraitement de la multidimensional scaling method (MDS)

Le Multidimensional Scaling (MDS) est une méthode d'analyse statistique qui permet de représenter visuellement des données complexes dans un espace de faible dimension, tout en préservant autant que possible les relations de distance entre les individus.

Dans cette approche, on utilise comme point de départ la matrice des distances D, qui représente les distances entre les N individus de l'ensemble de données. L'objectif principal du MDS est de construire un ensemble de points X dans un espace euclidien de telle manière que la matrice de Gram associée, définie par $G = XX^T$, soit une approximation aussi fidèle que possible de la matrice des distances D.

Pour obtenir cette matrice de Gram, il est nécessaire d'appliquer un double centrage à la matrice des distances, selon la procédure suivante.

$$G_{ij} = -\frac{1}{2}\left(D_{ij}^2 - \frac{1}{n}D_{i+}^2 - \frac{1}{n}D_{+j}^2 + \frac{1}{n^2}D_{++}^2\right). \tag{1}$$

où $d_{i+}^2 = \sum_j d_{ij}^2$ est la somme des coefficients d'une ligne, $d_j^2 = \sum_i d_{ij}^2$ la somme des coefficients d'une colonne et $d_{++}^2 = \sum_{ij} dij^2$la somme totale des coefficients de la matrice. Comme la matrice des distances est symétrique, on a $D_{i+}^2 = D_{+i}^2$.

L'objectif de ce TP est de paralléliser par une approche à base de tâches les différentes étapes du prétraitement de la méthode MDS à l'aide d'OpenMP.

A Introduction

Dans cette première approche, on considère un découpage en tuile de la matrice des distances et nous utiliserons un parallélisme de tâche.

```
typedef struct block {
 // the shape of the block
 int shape[2];
// the coefficients
 value_type *coeff;
} block;
```

Listing 1: structure de block.

```
typedef struct tiled_matrix {
//  the  block size
 int block_size;
//  the shape of the matrix
 int global_shape[2] ;
//  the shape of the block matrix
 int shape[2];
 //  a pointer on the bloc structure
 struct block *mat_block;
} tiled_matrix;
```

Listing 2: structure de block.

La méthode `pretraitement_MDS_seq` propose une approche séquentielle du prétraitement décrite dansl'algorithme 1.

Algorithm 1: Les différentes étapes du pretraitement

Data: D une matrice de distance par blocs
Result: D la matrice après une double centrage

▷ Etape 1
$D \leftarrow D \ .*D$
▷ Etape 2
$mean_{row} = \frac{1}{N} D * \mathbf{1}$
$d = \frac{1}{N} < mean_{row}, \mathbf{1} >$
▷ Etape 3
$D \leftarrow gram(Dist, mean_{row}, d)$ ▷ Via Équation 1

Le processus de double centrage de la matrice des distances se décompose en trois étapes principales :

1. Élévation au carré de la matrice des distances.
2. Calcul des moyennes :
   — La moyenne des lignes de la matrice D, notée $mean_{\text{row}}$, qui représente la moyenne de chaque ligne de la matrice.
   — La moyenne globale des éléments de D, notée d, qui est simplement la moyenne de tous les éléments de la matrice.
3. Calcul de la matrice de Gram : on applique l'équation 1 pour obtenir la matrice de Gram.

B Parallélisation à base de tâches : approche synchronisée

L'objectif de cette première partie est de paralléliser la fonction `pretraitement_MDS_task` (code `block_mds.c`) en utilisant le pragma `omp task`, selon l'approche suivante :

— On affecte une tâche pour évaluer les différentes opérations sur un bloc.
— Un vecteur (par exemple $mean_{row}$) se découpe en blocs (virtuels) similaires aux découpages en blocs des lignes de la matrice. Une tâche traite un bloc du vecteur.
— Une synchronisation des tâches est effectuée à la fin de chaque étape pour s'assurer qu'elles soient terminées avant de passer à l'étape suivante.

Le calcul de la somme des lignes et de la somme des coefficients est réalisé via une réduction.

C Parallélisation à base de tâches : approche pipeline

L'objectif maintenant est de supprimer les synchronisations explicites afin de pipeliner les étapes du traitement. Pour cela, il est nécessaire d'introduire des dépendances entre les différentes tâches afin de maintenir l'ordre correct de calcul tout en optimisant le parallélisme.

De plus, il faudra reconsidérer la méthode de réduction utilisée. Par exemple, au lieu d'utiliser une réduction classique, vous pourriez reformuler ce calcul comme un produit matrice-vecteur.

Indication : écrire le graphe de tâches pour les deux premières boîtes.