Factorization-Based Data Modeling
Practical Work 3

Umut Şimşekli
Télécom ParisTech, Université Paris-Saclay, Paris, France
umut.simsekli@telecom-paristech.fr

Instructions: (please read carefully)

1. This homework can be done in groups of **maximum 2** people.
2. Prepare your report as a pdf file in English by using LaTeX or a similar software (Word etc). Do not submit scanned papers.
3. Put all your files (code and/or report) in a zip file: *surname_name_tp3.zip* and submit it to `https://www.dropbox.com/request/KldORL8rXkdz1DA6jahv`. The deadline is **January 12, 2017**. Late submissions will not be accepted.
4. One submission per group is sufficient.

1 Distributed Stochastic Gradient Descent

In this practical work, the aim is to implement the Distributed Stochastic Gradient Descent (DSGD)[1] algorithm, which we covered earlier. You will implement the algorithm in C/C++ by using the OpenMPI and GSL libraries.

Throughout this practical work, we will only consider the usual matrix factorization problem, given as follows:

$$(Z_1^\star, Z_2^\star) = \underset{Z_1, Z_2}{\arg\min} \frac{1}{2} \|M \odot (X - Z_1 Z_2)\|_F^2 \tag{1}$$

where we have changed the notation from the earlier notes.

2 Exercises

In the following questions, we will work on the MovieLens 1 Million dataset. We will assume we have 4 processors, therefore the observed matrix will be partitioned into a 4×4 matrix.

1. Complete the file `dsgd_mf_template.cpp`.
2. Set the rank to 10 and the step size to 0.00001. Run the code for MovieLens 1 Million Dataset.
3. Compute the RMSE by using the code `compute_rmse.cpp` and plot the RMSE in Matlab by using `plot_rmse.m`.
4. Play with the rank and the step-size. What do you observe?

[1] Gemulla, Rainer, et al. "Large-scale matrix factorization with distributed stochastic gradient descent.", Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011.