Assignment 1
SPM course a.a. 24/25

March 7, 2025

Softmax vectorization

The *softmax* function is a fundamental algorithm in machine learning, widely used in classification tasks and neural network output layers. It converts raw scores (logits) into a probability distribution, ensuring that the sum of the outputs equals one. Given its frequent use in large-scale models, optimizing the *softmax* operation is critical to improving overall performance in real-world applications. Its mathematical formulation is as follows:

$$\sigma : \mathbb{R}^K \to \left\{ z \in \mathbb{R}^K \middle| z_i > 0,\ \sum_{i=1}^{K} z_i = 1 \right\}$$

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad j = 1, \dots, K$$

Starting from a scalar implementation of the *softmax* function in C++ using FP32 arithmetic (provided by the teacher), optimize the *softmax* function by manually vectorizing the code using AVX intrinsics and FMA. Then modify the baseline code (if necessary) and apply appropriate compiler flags and pragmas to enable auto-vectorization. Compare the resulting performance with your manually vectorized version. In the code provided by the teacher (softmax.zip) you can find the AVX implementation of the exponential function (*exp256_ps*) that you should use in your AVX version.

Write a brief report (max 3 pages) summarizing your findings, including:

- A description of your implementation choices
- Performance evaluation and comparisons.
- Discussion of potential trade-offs between manual and auto-vectorization.
- Any challenges encountered and possible improvements for future work

Send the teacher your code and report (both in a zip file with the name softmax_NameSurname.zip) by the deadline.

Deadline: March 14 EOB.

Content of the softmax.zip file:

- Makefile
- softmax_plain.cpp : full scalar implementation
- softmax_auto.cpp : partial implementation, the file contains the softmax_auto function you should implement for auto-vectorization
- Softmax_avx.cpp : partial implementation, the file contains the softmax_avx function you should implement using AVX intrinsics
- Include folder
  - avx_mathfun.h : files containing some mathematical functions including *exp256_ps*
  - hpc_helpers.hpp : helper functions for getting time measurements
  - README : a file with some notes related to the avx_mathfun software

Your code should execute on the spmcluster.unipi.it machines.