GPU Parallelization of Massive Sample-size Survival Analysis

Jianxiao Yang[1], Marc A. Suchard[1,2,3]
1. Department of Computational Medicine, David Geffen School of Medicine at UCLA 2. Department of Biostatistics, UCLA Fielding School of Public Health.
3. Department of Human Genetics, David Geffen School of Medicine at UCLA

Introduction

Large-scale Observational Data

- Observational databases have millions of individuals [1] with thousands of patient characteristics and up to 10 years of data per life [2].
- Resource for comparative effectiveness and safety study.
- Survival analysis is a main statistical method in comparative effectiveness and safety study.
- **Problem**: computational burden.

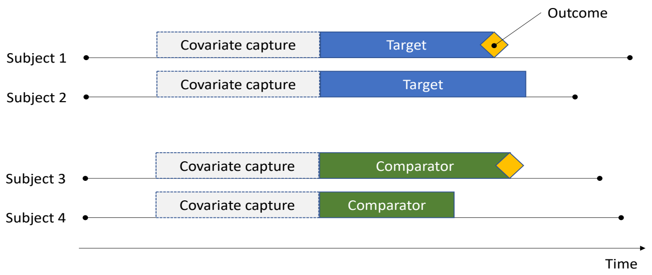

Figure 1: Comparative effectiveness study with cohort design.

GPU Parallelization

- Graphics processing units (GPUs) contain thousands of processor cores that can apply the same numerical operations simultaneously to elements of large data arrays under a “Single Instruction, Multiple Threads” (SIMT) programming paradigm.
- GPUs are relatively inexpensive, easy-to-use hardware that offers impressive potential for speeding up computations.

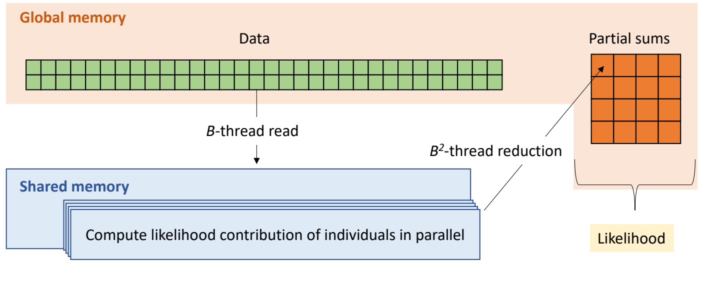

Figure 2: Massive parallelization strategy for computing the likelihood.

Methods

Cox Proportional Hazards Model

- The hazard model formula depends on a baseline survival function and a set of explanatory variables.
- Parameter estimation of the Cox proportional hazards model follows from the log-partial likelihood.

$$h(t|\boldsymbol{\beta}) = h_0(t|\boldsymbol{\beta}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}\right)$$

$$l_{partial}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ \boldsymbol{\beta}^T \mathbf{x}_i - \log \left[\sum_{k \in R(Y_i)} \exp\left(\boldsymbol{\beta}^T \mathbf{x}_k\right) \right] \right\}$$

Fine-Gray Sub-distribution Proportional Hazards Model

- The Fine-Gray model generalizes the Cox proportional hazards model to competing risks time-to-event data that consists of more than one type of events.
- Competing risks arise when individuals can experience more than one type of event and the occurrence of one type of event will prevent the occurrence of the others.

Massive Parallelization for Parameter Estimation with Prefix Sums and Reduction

- We identify prefix sums [3] and reductions in log-partial likelihood of Cox model and Fine-Gray model due to the cumulative structure of the risk set.
- We avoid unnecessary memory transactions by fusing prefix sums and reductions operations in likelihood calculations into a single kernel.
- We minimize data movements by exploiting the sparsity of the design matrix.

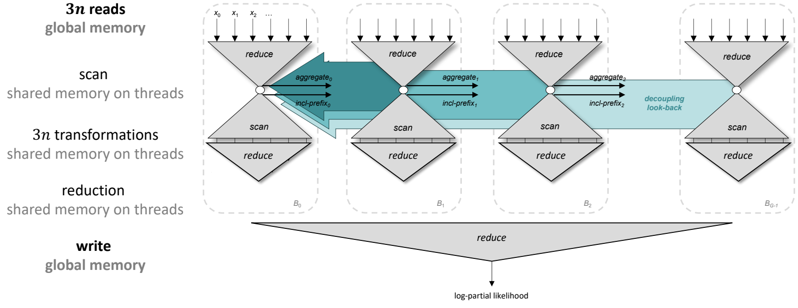

Figure 3: Fused kernel for maximum likelihood estimation.

Results

Simulation experiments

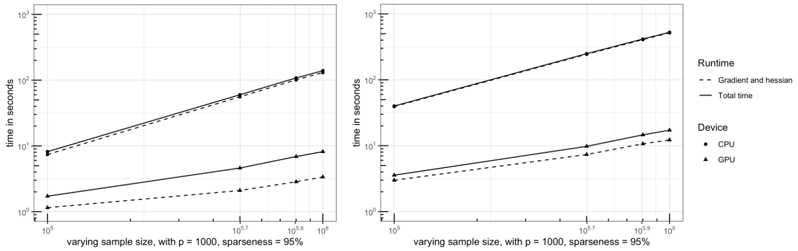

Figure 4: GPU vs CPU runtimes for Cox model (left) and Fine-Gray model (right).

Antihypertensive drug classes study

- Dataset: thiazide or thiazide-like diuretics (THZ) and angiotensin-converting enzyme inhibitors (ACEi) cohort study, reproduction of [4] in CCAE dataset.
- Cohort size: 1,065,745 hypertension patients and 7891 covariates.
- Outcome of interest: major cardiovascular events.
- GPU parallelization reduce the time of parameter estimation from 16 hours on multi-core CPU to less than one hour.

Conclusions

- We implement the massive parallelization of the Cox proportional hazards model and Fine-Gray model by using NVIDIA’s CUB library for parallel computing of prefix sums and exploiting the sparsity of data.
- By saving data movement and clever manipulation of likelihood structure, our parallelization significantly reduces the runtime of large-scale survival analysis.

References

[1] G. Hripcsak, P. B. Ryan, J. D. Duke, N. H. Shah, R. W. Park, V. Huser, M. A. Suchard, M. J. Schuemie, F. J. DeFalco, A. Perotte, et al., "Characterizing treatment pathways at scale using the ohdsi network," Proceedings of the National Academy of Sciences, vol. 113, no. 27, pp. 7329-7336, 2016.

[2] M. A. Suchard, S. E. Simpson, I. Zorych, P. Ryan, and D. Madigan, "Massive parallelization of serial inference algorithms for a complex generalized linear model," ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 23, no. 1, p. 10, 2013.

[3] D. Merrill and M. Garland, "Single-pass parallel prefix scan with decoupled look-back," NVIDIA, Tech. Rep. NVR-2016-002, 2016.

[4] M. A. Suchard, M. J. Schuemie, H. M. Krumholz, S. C. You, R. Chen, N. Pratt, C. G. Reich, J. Duke, D. Madigan, G. Hripcsak, et al., "Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis," The Lancet, vol. 394, no. 10211, pp. 1816-1826, 2019.