VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Optimizations for CPUs, GPUs and numerical stability

Georg Zitzlsberger georg.zitzlsberger@vsb.cz

26-03-2021

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Agenda

CPU Optimizations

GPU Optimizations

Numerical Stability

Summary

CPU Optimizations

- Intel offers own version of scikit-learn via Intel Distribution for Python
- It builds on Intel's performance libraries:
  - Intel Data Analytics Acceleration Library (Intel DAAL):
    Replaces some algorithms/tools from scikit-learn
  - Intel Math Kernel Library (Intel MKL):
    Backend to NumPy and SciPy
- Intel DAAL substitutions are turned off by default, enable with:

```
import daal4py.sklearn
daal4py.sklearn.patch_sklearn()
```

. . . or

```
$ USE_DAAL4PY_SKLEARN=YES python ...
```

- Don't use toy sets to test speedups; there is some overhead involved

GPU Optimizations

- From the scikit-learn ▸ FAQ:
  Q: Will you add GPU support?
  A: No, or at least not in the near future. ...
- Some tools/algorithms are available by ▸ H2O4GPU:

```
#from sklearn.cluster import KMeans
from h2o4gpu.solvers import KMeans
```

or

```
import h2o4gpu as sklearn
```

$\Rightarrow$ It can be used as drop-in replacement

- *H2O4GPU*:
  - Inherits *scikit-learn* tools/algorithms
  - Adds GPU support to some; falls back to CPU if not available
  - Builds on existing GPU solvers

H2O4GPU

Advantages:

- Selected algorithms can significantly benefit from higher GPU throughput:
  - GLM: Lasso, Ridge Regression, Logistic Regression, Elastic Net Regulariation
  - KMeans
  - Gradient Boosting Machine (GBM) via XGBoost
  - Singular Value Decomposition(SVD) + Truncated Singular Value Decomposition
  - Principal Components Analysis(PCA)
- Transparently falls back to scikit-learn implementation (Intel DAAL)
- Supports multiple GPUs
- Offers more tools than scikit-learn

Disadvantages:

- Data transfers between host and GPU are limiting the speedup (or even slow down)
- Lacks behind scikit-learn and has smaller community

H2O4GPU Example

```
import matplotlib
from h2o4gpu import DAAL_SUPPORTED
from sklearn import datasets, linear_model
import matplotlib.pyplot as plt
import numpy as np

diabetes = datasets.load_diabetes()
# Define diabetes_X_train, diabetes_X_test, diabetes_y_train, diabetes_y_test
...

if DAAL_SUPPORTED:
    from h2o4gpu.solvers.daal_solver.daal_data import getNumpyShape
    import h2o4gpu

    lin_solver_daal = h2o4gpu.LinearRegression(fit_intercept=True,
                                               verbose=True,
                                               backend='daal',
                                               method=h2o4gpu.LinearMethod.normal_equation)

    rows, cols = getNumpyShape(diabetes_y_train)
    y = diabetes_y_train.reshape(cols, rows)
    lin_solver_daal.fit(diabetes_X_train, y)
    daal_predicted = lin_solver_daal.predict(diabetes_X_test)
else:
    from sklearn.metrics import mean_squared_error, r2_score

    # Create linear regression object
    regr = linear_model.LinearRegression()

    # Train the model using the training sets
    regr.fit(diabetes_X_train, diabetes_y_train)

    # Make predictions using the testing set
    diabetes_y_pred = regr.predict(diabetes_X_test)
```

Numerical Stability

- Optimizations can cause changes in numerical results:
  - Change of FP operation order
  - Different operations (e.g. use of FMA)
  - Rounding modes
  - ...

  Example: $(a + b) + c \neq a + (b + c)$

  $2^{-63} + 1 + -1 = 2^{-63}$: mathematical result
  $(2^{-63} + 1) + -1 \approx 0$: correct IEEE result
  $2^{-63} + (1 + -1) \approx 2^{-63}$: correct IEEE result

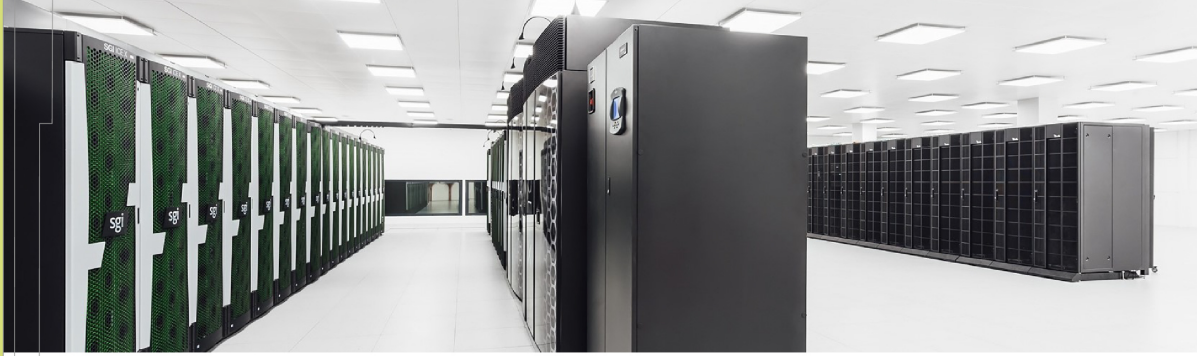
- The scikit-learn community did not pick Intel's version as default
- If you like (Intel) optimizations, you need to enable them explicitly
- Selectively use H2O4GPU:
  Uses Intel DAAL as CPU fallback before switching to scikit-learn!
- Even if results are (bitwise) the same on one system, another system might execute different code paths!

Summary

- Some algorithms/tools from scikit-learn exist in optimized versions for CPUs and GPUs
- They require sufficient workload/minimized data transfer to benefit
- H2O4GPU and Intel DAAL offer even more algorithms but with their own API

- **Be aware:**
  They are different implementations and can lead to different results[1]!

[1] If bitwise reproducibility is required.



VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

IT4Innovations National Supercomputing Center

VŠB – Technical University of Ostrava
Studentská 6231/1B
708 00 Ostrava-Poruba, Czech Republic
www.it4i.cz

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS