Final Project: GPU Accelerated Support Vector Machines via Quadratic Programming
Armaan Kohli - ECE453 Advanced Computer Architecture
Spring 2020

Remarks

We implemented a fast support vector machine (SVM) classifier that leverages GPU architecture to achieve performance gains for large datasets. Specifically, our project formulates a SVM classifier as a quadratic program, which we can solve using the alternating direction method of multipliers (ADMM) with either LDL factorisation on the CPU or preconditioned conjugate gradient (PCG) on the GPU. Our project is an extension of the the `osqp` API [1], having added additional tools for SVM problem generation and linear system solving using PCG and CUDA to target an Nvidia Jetson Nano. We compare the results of our GPU implementation to the benchmarks provided by `osqp`.

Support Vector Machines

The support vector machine (SVM) is an example of a convex optimization problem w/ quadratic criteria. SVM is a technique for classifying data that may not be linearly separable. This is accomplished by performing a linear classification problem in a higher dimensional feature space where the data is linearly separable.

First, let's assume we have N tuples, (x_i, y_i), where $x_i \in R^m$ represent feature vectors, and y_i represents the true class, $y_i \in \{-1, 1\}$. For instance, in the Fig 1, the blue dots correspond to x_i from $y_i = -1$, and the red dots correspond to features from $y_i = 1$. Let's also assume we've already done some kind of projection into a space where the classes are linearly separable.

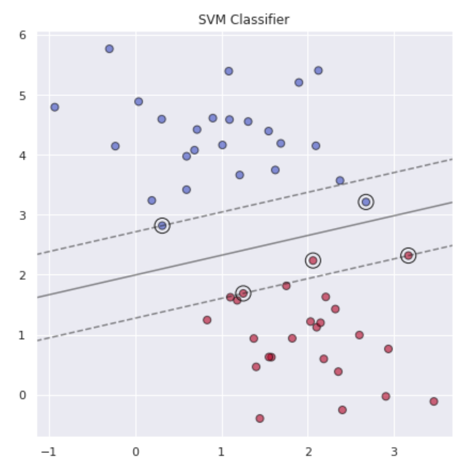

Figure 1: An illustrative example of an SVM classifier. Most of the construction of the SVM section is based on *The Elements of Statistical Learning* [2].

Now, we only have access to each value of x_i, the class they belong to, the corresponding y_i, is unknown. Let's define a hyperplane by:

$$\{x : f(x) = x^T\beta + 1 = 0\} \tag{1}$$

A classification rule induced by this hyperplane $f(x)$ is:

$$G(x) = sign[x^T\beta + 1] \tag{2}$$

Since we are in a space where the classes are linearly separable, we can find a function $f(x) = x^T\beta + 1$ with $y_i f(x_i) > 0 \quad \forall\ i$. Hence, we are able to find a hyperplane that creates the largest margin between the two classes. This is what the SVM accomplishes.
The circled points on the dotted lines are the so-called support vectors. The points between the dashed line and the decision boundary are within the margin. The SVM seeks to find the hyperplane that creates the largest margin, subject to the constraint to minimize the total distance of points on the wrong side of the margin.

We can formulate an SVM as a quadratic program using the method described in [2]. In this form, we can use an algorithm called alternating direction method of multipliers (ADMM).

Implementation Details

The `osqp` paper presents an two new ways of solving ADMM, a direct method an indirect method. The algorithm in full-form is presented below:

Algorithm 1: ADMM algorithm as presented in [1]

given: x^0, z^0, y^0 *and parameters* $\rho > 0,\ \sigma > 0,\ \alpha \in [0, 2]$

while *not terminated* **do**

$$(\tilde{x}^{k+1}, v^{k+1}) \leftarrow \begin{bmatrix} P + \sigma I & A^T \\ A & -\rho^{-1} I \end{bmatrix} \begin{bmatrix} \tilde{x}^{k+1} \\ v^{k+1} \end{bmatrix} = \begin{bmatrix} \sigma x^k - q \\ z^k - \rho^{-1} y^k \end{bmatrix}$$

$$\tilde{z}^{k+1} \leftarrow z^k + \rho^{-1}(v^{k+1} - y^k)$$

$$x^{k+1} \leftarrow \alpha \tilde{x}^{k+1} + (1 - \alpha) x^k$$

$$z^{k+1} \leftarrow \prod \left(\alpha \tilde{z}^{k+1} + (1 - \alpha) z^k + \rho^{-1} y^k \right)$$

$$y^{k+1} \leftarrow y^k + \rho(\alpha \tilde{z}^{k+1} + (1 - \alpha) z^k - z^{k+1}$$

end

Using this algorithm, we can very quickly compute solutions to SVM problems. However, there is one big issue with the algorithm, namely the matrix inversion step, which must be done every iterations until convergence. This is what we'll exploit to get performance gains. The CPU-based algorithm that `osqp` presents uses LDL matrix factorization to solve this linear system, called the KKT matrix. LDL factorization can be costly as the size of the problem increases, so for large datasets, we can leverage parallelism, and implement an indirect solver for this KKT matrix using PCG. As opposed to a direct method such as KKT, PCG can be efficiently parallelized.

Preconditioned Conjugate Gradient

In cases where the number of datapoints or the dimensionality of the space is too large for LDL factorization to work effectively, we can instead use an indirect method for solving this KKT linear system. PCG was suggested as a solution in [3]. This is accomplished by re-writing the KKT constraints as in the following form:

$$(P + \sigma I + A^T R A) \tilde{x}^{k+1} = \sigma x^k - q + A^T (R z^k - y^k) \tag{3}$$

Which can be solved using the PCG algorithm below.

GPU Optimizations

In order to write an efficient GPU implementation of PCG, we used the following CUDA Toolkit libraries: `Thrust`, `cuBLAS` and `cuSPARSE`. `Thrust` provides a high-level interface for essential data parallel primitives, such as scan and sort operations. `cuBLAS` is a CUDA implementation of BLAS (Basic Linear Algebra Subprograms). We use only level-1 cuBLAS API functions that implement the inner product, axpyoperation, scalar-vector multiplication, and computation of norms. `cuSPARSE` is a CUDA library that contains a set of linear algebra subroutines for handling sparse matrices. Throughout the GPU codebase, we use a CSR format of sparse

Algorithm 2: PCG algorithm as presented in [3]

initialise: $r^0 = Kx^0 - b,\ y^0 = M^{-1}r^0,\ p^0 = -y^0,\ k = 0$
while $||r^k|| > \epsilon||b||$ **do**

$$\alpha^k \leftarrow -\frac{(r^k)^T y^k}{(p^k)^T K p^k}$$
$$x^{k+1} \leftarrow x^k + \alpha^k p^k$$
$$r^{k+1} \leftarrow r^k + \alpha^k K p^k$$
$$y^{k+1} \leftarrow M^{-1} r^{k+1}$$
$$\beta^{k+1} \leftarrow -\frac{(r^{k+1})^T y^{k+1}}{(r^k)^T y^k}$$
$$p^{k+1} \leftarrow -y^{k+1} + \beta^{k+1} p^k$$
$$k \leftarrow k + 1$$

end

matricies, since that is the format that `cuSPARSE` uses. The rest of the `osqp` api uses CSC representation, so additional code was needed to perform conversion.

For the kernel functions, we used 64 elements per thread and 1024 threads per block. We found that this gave the best performance empirically, through a more thorough analysis or parameter search would be required to test for optimal settings.

Results & Discussion

In order to test the performance of our implementation, we generated random SVM problems of various dimensions and computed the run time. To create the CPU benchmarks, we use the CPU code provided by `osqp`. To test the GPU code, we used our own PCG implementation that we then interfaced with the `osqp` api. We generated the SVM problems using a simple `python` script and used the provided code generation tools provided by `osqp` to generate header files containing the data. Below in Fig. 2 is the performance of the SVM classifier according to the results from [3].

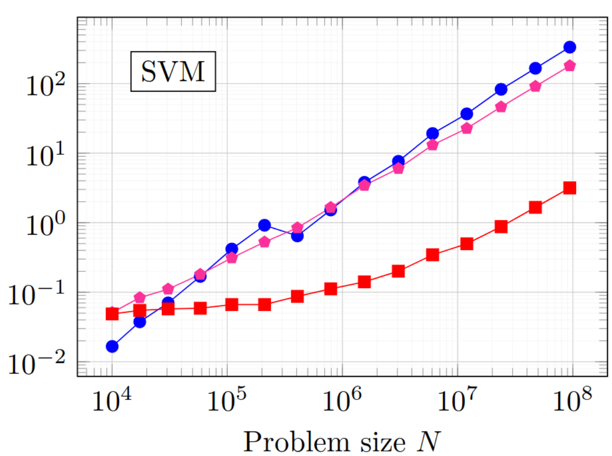

Figure 2: This a plot replicated from [3]. The red line on the right-hand graph shows the performance of the algorithm from [3] in terms of runtime measured in seconds versus problem size. The blue line is the CPU implementation from [1].

They found that using a GPU implementation for SVM solving didn't give massive improvements. However the CPU implementation scales much faster than the GPU implementation as the problem size increases. Furthermore, latency between the host and target devices meant that for smaller problems, the CPU algorithm was superior to the GPU algorithm. But as the problem size increased, the overhead was overcome and the GPU achieved better performance.

For our implementation, we were unable to generate problems greater than 10^4. However, we did see similar performance improvements, our results are pictured in Fig. 3. As was the case for the official implementation, we found that for small problem size, the overhead made the CPU more efficient than the GPU. But, as the problem size increased, the GPU gave superior performance. And, as the problem size increased, the GPU runtime increased at a much slower rate than the CPU runtime.

Looking at absolute performance however, and not trends, tells us that our GPU implementation is actually superior, since it surpasses CPU performance for smaller problems faster than the official version. GPU performance beats CPU performance around for $N = 3 \times 10^4$ in Fig. 2, whereas this cross over happens much earlier for our implementation, at around $N = 5 \times 10^3$. This could be due to the choice of GPU specific parameters, such as the number of threads per block or elements per thread, or even the physical GPU itself.

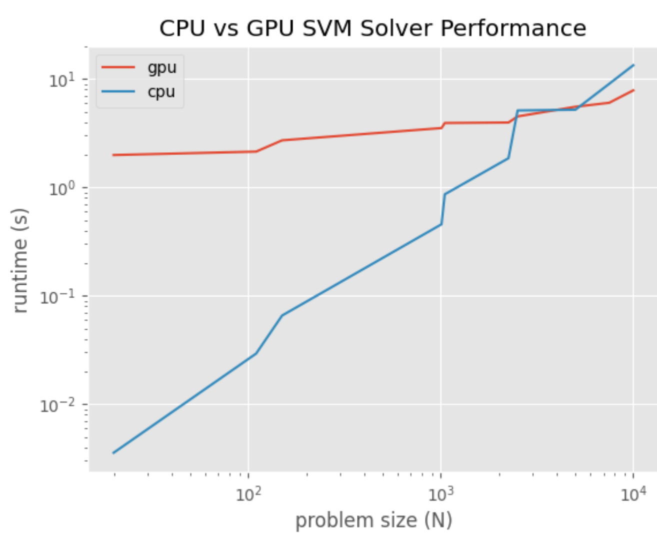

Figure 3: This a plot of our SVM solver results. Unlike the results presented by [3], we were not able to generate problems larger than 10^4. However, the trends in the graphs are the same. For larger problem sizes, GPU outperforms CPU, and GPU performance scales better for larger problems. These metrics were compiled by generating random SVM problems of different sizes, then averaging over several problems. All experiments were run on a Jetson Nano.

Conclusion

In conclusion, we were able to develop our own version of a GPU implementation of ADMM for SVM classification using the `osqp` api. We also successfully verified the results from [3] a regime of relatively scale problems. Furthermore, our GPU implementation achieves higher performance than reported in the paper in the regime of small scale problems.

References

[1] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "OSQP: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, 2020. [Online]. Available: https://doi.org/10.1007/s12532-020-00179-2

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[3] M. Schubiger, G. Banjac, and J. Lygeros, "GPU acceleration of ADMM for large-scale quadratic programming," *arXiv:1912.04263*, 2019.

Appendix A: Code

The code below is `generate_problem.py`. It is used to generate SVM classification problems

```
"""
generate random svm problems
"""

import numpy as np
import scipy as sp

from scipy import sparse
import utils.codegen_utils as cu

sp.random.seed(1234)
n = 50
m = 100
N = int(m / 2)
gamma = 1.0
b = np.hstack([np.ones(N), -np.ones(N)])
A_upp = sparse.random(N, n, density=0.5)
A_low = sparse.random(N, n, density=0.5)
Ad = sparse.vstack(
    [
        A_upp / np.sqrt(n) + (A_upp != 0.0).astype(float) / n,
        A_low / np.sqrt(n) - (A_low != 0.0).astype(float) / n,
    ],
    format="csc",
)

# osqp data
Im = sparse.eye(m)
P = sparse.block_diag([sparse.eye(n), sparse.csc_matrix((m, m))], format="csc")
q = np.hstack([np.zeros(n), gamma * np.ones(m)])
A = sparse.vstack(
    [
        sparse.hstack([sparse.diags(b).dot(Ad), -Im]),
        sparse.hstack([sparse.csc_matrix((m, n)), Im]),
    ],
    format="csc",
)
l = np.hstack([-np.inf * np.ones(m), np.zeros(m)])
u = np.hstack([-np.ones(m), np.inf * np.ones(m)])

cu.generate_problem_data(P, q, A, l, u, "svm")
```

The code below is cuda_pcg.cu. It contains all of the main CUDA routines for the PGC algorithm.

```
/* cuda pcg algorithm */

#include "cuda_pcg.h"
#include "csr_type.h"
#include "cuda_handler.h"
#include "cuda_malloc.h"
#include "cuda_lin_alg.h"
#include "cuda_wrapper.h"
#include "helper_cuda.h"

#ifdef __cplusplus
extern "C" {
extern CUDA_Handle_t *CUDA_handle;
}
#endif

__global__ void scalar_division_kernel(c_float *res, const c_float *num, const c_float *den)
{
    *res = (*num) / (*den);
}

/* computes:  d_y = (P + sigma*I + A'*R*A) * d_x */
static void mat_vec_prod(cudapcg_solver *s, c_float *d_y, const c_float  *d_x, c_int device)
{

    c_float *sigma;
    c_float H_0 = 0.0;
    c_float H_1  = 1.0;
    c_int n = s->n;
    c_int m = s->m;
    csr *P  = s->P;
    csr *A  = s->A;
    csr *At = s->At;

    if (device)
    {
        sigma = s->d_sigma;
    }
    else
    {
        sigma = s->h_sigma;
    }

    /* d_y = d_x */
    checkCudaErrors(cudaMemcpy(d_y, d_x, n * sizeof(c_float), cudaMemcpyDeviceToDevice));

    /* d_y *= sigma */
    checkCudaErrors(cublasTscal(CUDA_handle->cublasHandle, n, sigma, d_y, 1));

    /* d_y += P * d_x */
```

```
    checkCudaErrors(cusparseCsrmv(CUDA_handle->cusparseHandle, P->alg, P->m, P->n, P->nnz, &H_1, P->
     MatDescription, P->val, P->row_ptr, P->col_ind, d_x, &H_1, d_y, P->buffer));

    if (m == 0) return;

    if (!s->d_rho_vec)
    {
        /* d_z = rho * A * d_x */
        checkCudaErrors(cusparseCsrmv(CUDA_handle->cusparseHandle, A->alg, A->m, A->n, A->nnz, s->h_rho,
     A->MatDescription, A->val, A->row_ptr, A->col_ind, d_x, &H_0, s->d_z, A->buffer));
    }
    else
    {
        /* d_z = A * d_x */
        checkCudaErrors(cusparseCsrmv(CUDA_handle->cusparseHandle, A->alg, A->m, A->n, A->nnz, &H_1, A->
     MatDescription, A->val, A->row_ptr, A->col_ind, d_x, &H_0, s->d_z, A->buffer));

        /* d_z = diag(d_rho_vec) * dz */
        cuda_vec_ew_prod(s->d_z, s->d_z, s->d_rho_vec, m);
    }

    /* d_y += A' * d_z */
    checkCudaErrors(cusparseCsrmv(CUDA_handle->cusparseHandle, At->alg, At->m, At->n, At->nnz, &H_1, At
     ->MatDescription, At->val, At->row_ptr, At->col_ind, s->d_z, &H_1, d_y, A->buffer));
}

/* pcg algorithm */
c_int cuda_pcg(cudapcg_solver *s, c_float eps, c_int max_niter)
{

    c_float *ptr_tmp;
    c_int niter = 0;
    c_int n = s->n;
    c_float H_m_1 = -1.0;

    /* set up problem */

    if (!s->warm_start)
    {
        /* d_x = 0 */
        checkCudaErrors(cudaMemset(s->d_x, 0, n * sizeof(c_float)));
    }

    /* d_p = 0 */
    checkCudaErrors(cudaMemset(s->d_p, 0, n * sizeof(c_float)));

    /* d_r = K * d_x */
    mat_vec_prod(s, s->d_r, s->d_x, 0);

    /* d_r -= d_rhs */
    checkCudaErrors(cublasTaxpy(CUDA_handle->cublasHandle, n, &H_m_1, s->d_rhs, 1, s->d_r, 1));
```

```

    /* h_r_norm = |d_r| */
    s->vector_norm(s->d_r, n, s->h_r_norm);

    /* need to change CUBLAS mode */
    cublasSetPointerMode(CUDA_handle->cublasHandle, CUBLAS_POINTER_MODE_DEVICE);

    if (s->precondition)
    {
        /* d_y = M \ d_r */
        cuda_vec_ew_prod(s->d_y, s->d_diag_precond_inv, s->d_r, n);
    }

    /* d_p = -d_y */
    checkCudaErrors(cublasTaxpy(CUDA_handle->cublasHandle, n, s->D_MINUS_ONE, s->d_y, 1, s->d_p, 1));

    /* rTy = d_r' * d_y */
    checkCudaErrors(cublasTdot(CUDA_handle->cublasHandle, n, s->d_y, 1, s->d_r, 1, s->rTy));

    /* synchronize for timing */
    cudaDeviceSynchronize();

    /* Run the PCG algorithm */
    while ( *(s->h_r_norm) > eps && niter < max_niter )
    {

        /* d_Kp = K * d_p */
        mat_vec_prod(s, s->d_Kp, s->d_p, 1);

        /* pKp = d_p' * d_Kp */
        checkCudaErrors(cublasTdot(CUDA_handle->cublasHandle, n, s->d_p, 1, s->d_Kp, 1, s->pKp));

        /* alpha = rTy / pKp */
        scalar_division_kernel<<<1,1>>>(s->alpha, s->rTy, s->pKp);

        /* d_x += alpha * d_p */
        checkCudaErrors(cublasTaxpy(CUDA_handle->cublasHandle, n, s->alpha, s->d_p, 1, s->d_x, 1));

        /* d_r += alpha * d_Kp */
        checkCudaErrors(cublasTaxpy(CUDA_handle->cublasHandle, n, s->alpha, s->d_Kp, 1, s->d_r, 1));

        if (s->precondition)
        {
            /* d_y = M \ d_r */
            cuda_vec_ew_prod(s->d_y, s->d_diag_precond_inv, s->d_r, n);
        }

        /* Swap pointers to rTy and rTy_prev */
        ptr_tmp = s->rTy_prev;
        s->rTy_prev = s->rTy;
        s->rTy = ptr_tmp;
```

```

        /* rTy = d_r' * d_y */
        checkCudaErrors(cublasTdot(CUDA_handle->cublasHandle, n, s->d_y, 1, s->d_r, 1, s->rTy));

        /* Update residual norm */
        s->vector_norm(s->d_r, n, s->d_r_norm);
        checkCudaErrors(cudaMemcpyAsync(s->h_r_norm, s->d_r_norm, sizeof(c_float), cudaMemcpyDeviceToHost));

        /* beta = rTy / rTy_prev */
        scalar_division_kernel<<<1,1>>>(s->beta, s->rTy, s->rTy_prev);

        /* d_p *= beta */
        checkCudaErrors(cublasTscal(CUDA_handle->cublasHandle, n, s->beta, s->d_p, 1));

        /* d_p -= d_y */
        checkCudaErrors(cublasTaxpy(CUDA_handle->cublasHandle, n, s->D_MINUS_ONE, s->d_y, 1, s->d_p, 1));

        cudaDeviceSynchronize();
        niter++;

    } /* End of the PCG algorithm */

    /* change CUBLAS pointer mode back */
    cublasSetPointerMode(CUDA_handle->cublasHandle, CUBLAS_POINTER_MODE_HOST);

    return niter;
}

/* update preconditioning  */
void cuda_pcg_update_precond(cudapcg_solver *s, c_int P_updated, c_int A_updated, c_int R_updated)
{

    void    *buffer;
    c_float *mem_tmp;
    c_int    n  = s->n;
    csr     *At = s->At;

    size_t buff_size = n * (sizeof(c_float) + sizeof(c_int));

    if (!P_updated && !A_updated && !R_updated) return;

    if (P_updated)
    {
        /* Update d_P_diag_val */
        checkCudaErrors(cusparseTgthr(CUDA_handle->cusparseHandle, n, s->P->val, s->d_P_diag_val, s->d_P_diag_ind, CUSPARSE_INDEX_BASE_ZERO));
    }

    if (A_updated || R_updated)
```

```
    {
        /* Allocate memory */
        cuda_malloc((void **) &mem_tmp, At->nnz * sizeof(c_float));
        cuda_malloc((void **) &buffer, buff_size);

        /* Update d_AtRA_diag_val */
        if (!s->d_rho_vec)
        {   /* R = rho*I  -->  A'*R*A = rho * A'*A */

            if (A_updated)
            {
                /* Update d_AtA_diag_val */
                cuda_vec_ew_prod(mem_tmp, At->val, At->val, At->nnz);
                cuda_vec_segmented_sum(mem_tmp, At->row_ind, s->d_AtA_diag_val, buffer, n, At->nnz);
            }

            /* d_AtRA_diag_val = rho * d_AtA_diag_val */
            cuda_vec_add_scaled(s->d_AtRA_diag_val, s->d_AtA_diag_val, NULL, *s->h_rho, 0.0, n);
        }
        else
        {   /* R = diag(d_rho_vec)  -->  A'*R*A = A' * diag(d_rho_vec) * A */
            cuda_mat_rmult_diag_new(At, mem_tmp, s->d_rho_vec);   /* mem_tmp = A' * R */
            cuda_vec_ew_prod(mem_tmp, mem_tmp, At->val, At->nnz);     /* mem_tmp = mem_tmp * A */
            cuda_vec_segmented_sum(mem_tmp, At->row_ind, s->d_AtRA_diag_val, buffer, n, At->nnz);
        }

        cuda_free((void **) &mem_tmp);
        cuda_free((void **) &buffer);
    }

    /* d_diag_precond = sigma */
    cuda_vec_set_sc(s->d_diag_precond, *s->h_sigma, n);

    /* d_diag_precond += d_P_diag_val + d_AtRA_diag_val */
    cuda_vec_add_scaled3(s->d_diag_precond, s->d_diag_precond, s->d_P_diag_val, s->d_AtRA_diag_val, 1.0,
     1.0, 1.0, n);

    /* d_diag_precond_inv = 1 / d_diag_precond */
    cuda_vec_reciprocal(s->d_diag_precond_inv, s->d_diag_precond, n);
}
```

The code below is `cuda_pcg_interface.c`. It wraps `cuda_pcg.cu` functions into a `lin_sys` solver module specified by the `osqp` api so that we can actually solve a linear system via PCG.

```
/* interface for pcg_solver with main osqp api */

#include "cuda_pcg_interface.h"
#include "cuda_pcg.h"

#include "cuda_lin_alg.h"
#include "cuda_malloc.h"

#include "glob_opts.h"

static c_float compute_tolerance(cudapcg_solver *s, c_int admm_iter)
{
    c_float eps;
    c_float rhs_norm;

    /* compute the norm of RHS of the linear system */
    s->vector_norm(s->d_rhs, s->n, &rhs_norm);

    if (s->polish) return c_max(rhs_norm * CUDA_PCG_POLISH_ACCURACY, CUDA_PCG_EPS_MIN);

    switch (s->eps_strategy)
    {

    /* SCS strategy */
    case SCS_STRATEGY:
        eps = rhs_norm * s->start_tol  / pow((admm_iter + 1), s->dec_rate);
        eps = c_max(eps, CUDA_PCG_EPS_MIN);
        break;

    /* residual strategy */
    case RESIDUAL_STRATEGY:
        if (admm_iter == 1)
        {
            /* In case rhs = 0.0 we don't want to set eps_prev to 0.0 */
            if (rhs_norm < CUDA_PCG_EPS_MIN)
                s->eps_prev = 1.0;
            else
                s->eps_prev = rhs_norm * s->reduction_factor;
            /* Return early since scaled_pri_res and scaled_dua_res are meaningless before the first
     ADMM iteration */
            return s->eps_prev;
        }

        if (s->zero_pcg_iters >= s->reduction_threshold) {
            s->reduction_factor /= 2;
            s->zero_pcg_iters = 0;
        }

```

```
        eps = s->reduction_factor * sqrt((*s->scaled_pri_res) * (*s->scaled_dua_res));
        eps = c_max(c_min(eps, s->eps_prev), CUDA_PCG_EPS_MIN);
        s->eps_prev = eps;
        break;
    }
    return eps;
}

/* d_rhs = d_b1 + A' * rho * d_b2 */
static void compute_rhs(cudapcg_solver *s, c_float *d_b)
{

    c_int n = s->n;
    c_int m = s->m;

    /* d_rhs = d_b1 */
    cuda_vec_copy_d2d(s->d_rhs, d_b, n);

    if (m == 0)
        return;

    /* d_z = d_b2 */
    cuda_vec_copy_d2d(s->d_z, d_b + n, m);

    if (!s->d_rho_vec)
    {
        /* d_z *= rho */
        cuda_vec_mult_sc(s->d_z, *s->h_rho, m);
    }
    else
    {
        /* d_z = diag(d_rho_vec) * d_z */
        cuda_vec_ew_prod(s->d_z, s->d_z, s->d_rho_vec, m);
    }

    /* d_rhs += A' * d_z */
    cuda_mat_Axpy(s->At, s->d_z, s->d_rhs, 1.0, 1.0);
}

/* api fcns as perscribed by the osqp documentation */
c_int init_linsys_solver_cudapcg(cudapcg_solver **sp,
                                 const OSQPMatrix *P,
                                 const OSQPMatrix *A,
                                 const OSQPVectorf *rho_vec,
                                 OSQPSettings *settings,
                                 c_float *scaled_pri_res,
                                 c_float *scaled_dua_res,
                                 c_int polish) {

    c_int n, m;
    c_float H_m_1 = -1.0;
```

```

    /* allocate linsys solver structure */
    cudapcg_solver *s = c_calloc(1, sizeof(cudapcg_solver));
    *sp = s;

    /* assign type and the number of threads */
    s->type = CUDA_PCG_SOLVER;
    s->nthreads = 1; // dummy, this changes for target device

    /* dimensions */
    n = OSQPMatrix_get_n(P);
    m = OSQPMatrix_get_m(A);
    s->n = n;
    s->m = m;

    /* pcg states */
    s->polish = polish;
    s->zero_pcg_iters = 0;

    /* default norm and tolerance strategy */
    s->eps_strategy = RESIDUAL_STRATEGY;
    s->norm = CUDA_PCG_NORM;
    s->precondition = CUDA_PCG_PRECONDITION;
    s->warm_start_pcg = CUDA_PCG_WARM_START;
    s->max_iter = (polish) ? CUDA_PCG_POLISH_MAX_ITER : CUDA_PCG_MAX_ITER;

    /* tolerance strategy parameters */
    s->start_tol = CUDA_PCG_START_TOL;
    s->dec_rate = CUDA_PCG_DECAY_RATE;
    s->reduction_threshold = CUDA_PCG_REDUCTION_THRESHOLD;
    s->reduction_factor = CUDA_PCG_REDUCTION_FACTOR;
    s->scaled_pri_res = scaled_pri_res;
    s->scaled_dua_res = scaled_dua_res;

    /* set pointers settings and data */
    s->A = A->S;
    s->At = A->At;
    s->P = P->S;
    s->d_P_diag_ind = P->d_P_diag_ind;
    if (rho_vec)
        s->d_rho_vec = rho_vec->d_val;
    if (!polish)
    {
        s->h_sigma = &settings->sigma;
        s->h_rho = &settings->rho;
    }
    else
    {
        s->h_sigma = &settings->delta;
        s->h_rho = (c_float*) c_malloc(sizeof(c_float));
        *s->h_rho = 1. / settings->delta;
```

```
    }

    /* allocate pcg iterates */
    cuda_calloc((void **) &s->d_x, n * sizeof(c_float));
    cuda_malloc((void **) &s->d_p, n * sizeof(c_float));
    cuda_malloc((void **) &s->d_Kp, n * sizeof(c_float));
    cuda_malloc((void **) &s->d_y, n * sizeof(c_float));
    cuda_malloc((void **) &s->d_r, n * sizeof(c_float));
    cuda_malloc((void **) &s->d_rhs, n * sizeof(c_float));

    if (m != 0) cuda_malloc((void **) &s->d_z, m * sizeof(c_float));

    /* allocate scalar in host memory that is page-locked and accessible to target */
    cuda_malloc_host((void **) &s->h_r_norm, sizeof(c_float));

    /* allocate target-side scalar values. This way scalars are packed in target memory */
    cuda_malloc((void **) &s->d_r_norm, 8 * sizeof(c_float));
    s->rTy = s->d_r_norm + 1;
    s->rTy_prev = s->d_r_norm + 2;
    s->alpha = s->d_r_norm + 3;
    s->beta = s->d_r_norm + 4;
    s->pKp = s->d_r_norm + 5;
    s->D_MINUS_ONE = s->d_r_norm + 6;
    s->d_sigma = s->d_r_norm + 7;
    cuda_vec_copy_h2d(s->D_MINUS_ONE, &H_m_1, 1);
    cuda_vec_copy_h2d(s->d_sigma, s->h_sigma, 1);

    /* allocate memory for PCG preconditioning */
    if (s->precondition)
    {
        cuda_malloc((void **) &s->d_P_diag_val, n * sizeof(c_float));
        cuda_malloc((void **) &s->d_AtRA_diag_val, n * sizeof(c_float));
        cuda_malloc((void **) &s->d_diag_precond, n * sizeof(c_float));
        cuda_malloc((void **) &s->d_diag_precond_inv, n * sizeof(c_float));
        if (!s->d_rho_vec) cuda_malloc((void **) &s->d_AtA_diag_val, n * sizeof(c_float));
    }

    /* Set the vector norm */
    switch (s->norm) {
    case 0:
        s->vector_norm = &cuda_vec_norm_inf;
        break;

    case 2:
        s->vector_norm = &cuda_vec_norm_2;
        break;
    }

    s->solve = &solve_linsys_cudapcg;
    s->warm_start = &warm_start_linsys_solver_cudapcg;
    s->free = &free_linsys_solver_cudapcg;
```

```
    s->update_matrices = &update_linsys_solver_matrices_cudapcg;
    s->update_rho_vec  = &update_linsys_solver_rho_vec_cudapcg;

    /* init PCG preconditioner */
    if (s->precondition) cuda_pcg_update_precond(s, 1, 1, 1);

    return 0;
}

/* main driver function for pcg */
c_int solve_linsys_cudapcg(cudapcg_solver *s, OSQPVectorf *b,c_int admm_iter)
{

    c_int pcg_iter;
    c_float eps;

    /* compute the RHS of the reduced KKT system */
    compute_rhs(s, b->d_val);

    /* compute the required solution precision */
    eps = compute_tolerance(s, admm_iter);

    /* solve the linear system with PCG */
    pcg_iter = cuda_pcg(s, eps, s->max_iter);

    /* copy the first part of the solution to b->d_val */
    cuda_vec_copy_d2d(b->d_val, s->d_x, s->n);

    /* solution polishing */
    if (!s->polish)
    {
        /* Compute d_z = A * d_x */
        if (s->m) cuda_mat_Axpy(s->A, s->d_x, b->d_val + s->n, 1.0, 0.0);
    }
    else
    {
        /* Compute yred = (A * d_x - b) / delta */
        cuda_mat_Axpy(s->A, s->d_x, b->d_val + s->n, 1.0, -1.0);
        cuda_vec_mult_sc(b->d_val + s->n, *s->h_rho, s->m);
    }

    if (pcg_iter == 0)
        s->zero_pcg_iters++;

    return 0;
}

void warm_start_linsys_solver_cudapcg(cudapcg_solver *s, const OSQPVectorf *x)
{
    cuda_vec_copy_d2d(s->d_x, x->d_val, x->length);
}
```

```

void free_linsys_solver_cudapcg(cudapcg_solver *s)
{

    if (s)
    {
        if (s->polish)
            c_free(s->h_rho);

        /* PCG iterates */
        cuda_free((void **) &s->d_x);
        cuda_free((void **) &s->d_p);
        cuda_free((void **) &s->d_Kp);
        cuda_free((void **) &s->d_y);
        cuda_free((void **) &s->d_r);
        cuda_free((void **) &s->d_rhs);
        cuda_free((void **) &s->d_z);

        /* free host memory */
        cuda_free_host((void **) &s->h_r_norm);

        /* target-side scalar values */
        cuda_free((void **) &s->d_r_norm);

        /* pcg preconditioner */
        cuda_free((void **) &s->d_P_diag_val);
        cuda_free((void **) &s->d_AtA_diag_val);
        cuda_free((void **) &s->d_AtRA_diag_val);
        cuda_free((void **) &s->d_diag_precond);
        cuda_free((void **) &s->d_diag_precond_inv);

        c_free(s);
    }
}

c_int update_linsys_solver_matrices_cudapcg(cudapcg_solver *s, const OSQPMatrix *P, const OSQPMatrix *A)
{

    if (s->precondition) cuda_pcg_update_precond(s, 1, 1, 0);
    return 0;
}

c_int update_linsys_solver_rho_vec_cudapcg(cudapcg_solver *s, const OSQPVectorf *rho_vec, c_float rho_sc
    )
{

    if (s->precondition) cuda_pcg_update_precond(s, 0, 0, 1);
    return 0;
}
```