Sparse Condensed-Space Interior-Point Methods with Inequality Relaxations on GPUs: Will it Work?

Sungho Shin

sshin@anl.gov

Argonne National Laboratory

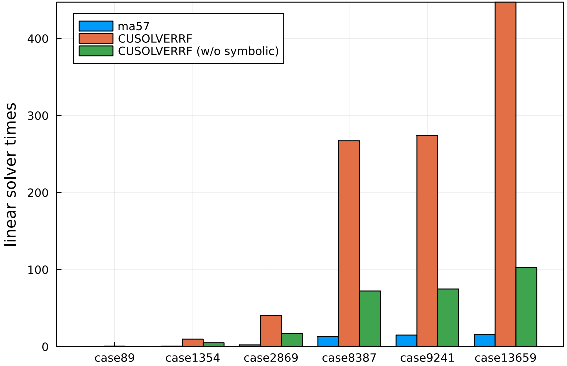
June 27, 2023

Executive Summary

- We exmaine the potential benefit of sparse condensed-space interior-point methods with inequality relaxations, with the goal of solving general large-scale NLPs on GPUs.
- The proposed method relaxes the equality constraints as inequality constraints by replacing $c(x) = 0$ by $c(x) + s = 0$ and $0 \leq s \leq \epsilon_{\mathrm{IR}}$.
- The resulting inequality-constrained NLP is solved with sparse condensed-space interior-point method, which requires solving a sparse positive definite system.
- While the LU solver in CUSOLVERRF cannot handle the sparse indefinite systems wihtin IPM, it can handle the sparse PD systems up to a certain accuracy ($\epsilon_{\mathrm{IR}} = \mathtt{tol} = 10^{-3}$).
- For portability, we will still need a sparse Cholesky solver running on GPU.

A Naive Approach Doesn't Work.

- Can we solve the indefinite KKT systems using sparse LU solver in `CUSOLVERRF`? No.



- Symbolic factorization needed every several iterations (every 3 for above case).
- Numeric factorization is not fast.

Condensed Approach

Consider an NLP of the following form:

$$\min_{x^L \le x \le x^U} \ f(x)$$
$$\text{s.t. } c(x) = 0.$$

The problem is relaxed via *inequality relaxation*:

$$\min_{x^L \le x \le x^U, 0 \le s \le \epsilon_{\text{IR}}} \ f(x)$$
$$\text{s.t. } c(x) + s = 0,$$

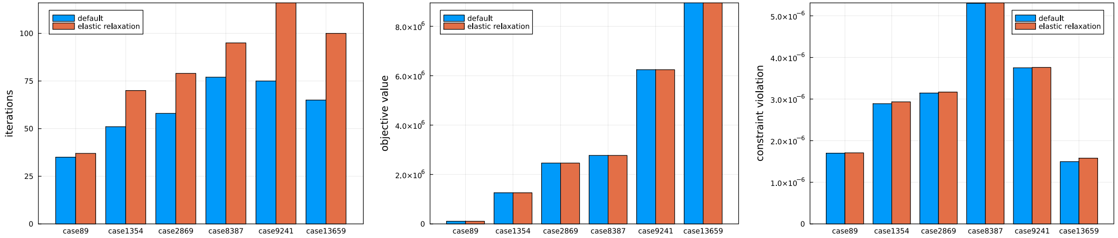
where $\epsilon_{\text{IR}} \approx \texttt{tol}$.

We now observe that the problem has *inequality constriants only*. This allows us to apply the *condensation strategy*:

$$\begin{bmatrix} H + \Sigma_x & & J^\top \\ & \Sigma_s & I \\ J & I & \end{bmatrix} \begin{bmatrix} p^u \\ p^s \\ p^\lambda \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ r_3, \end{bmatrix} \iff \left(H + \Sigma_x + J^\top \Sigma_s J\right) p^u = -r_1 + J r_2 - J^\top \Sigma_s r_3.$$

Of course, $J^\top \Sigma_s J$ can be arbitrarily dense, but our favorite problems (e.g., OPF) will still be sparse.

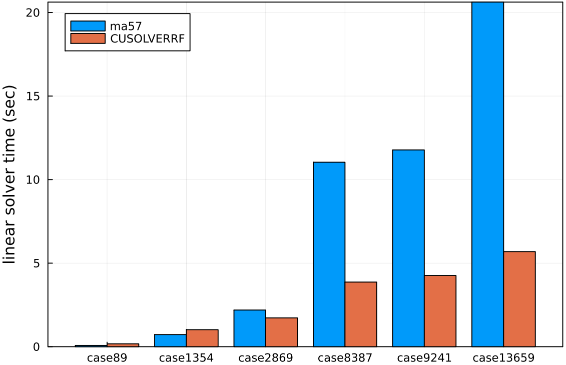
Effects of Inequality Relaxation



- $\epsilon_{\mathrm{IR}} = 10^{-8}$
- Number of iterations increased (5–50%).
- Not much difference in the quality of solution (obj value/constraint violation).

Will it be Efficient?

- We do this experiment to better evaluate the potential benefit of the proposed strategy.
- We take the KKT system from the iterates obtained from MadNLP+Ma57 and solve it with CUSOLVER. We aim to check:
  - Symbolic factorization can be reused between iterations.
  - Numeric factorization/backsolve is fast/precise enough.



- Observation: factorization/backsolve is precise/efficient enough for $\epsilon_{\mathrm{IR}} = \mathtt{tol} = 10^{-3}$.

What do we need for Portabiltiy?

- For now, the sparse LU solver in CUSOLVERRF can do the job.
- For portability, all we need is a sparse Cholesky solver (preferrably, written in `KernelAbstractions.jl`).

Current Status (6/27/2023)

- SparseCondensedKKTSystem works on CPU.
- With inequality relaxation, can solve `case9241` up to $\texttt{tol} = 10^{-7}$, but requires long iterative refinement. Convergence up to 10^{-5} is reliable/efficient.
- Numerical results (w/ linear algebra/AD on GPU) for `case9241`:

```
Total wall-clock secs in solver (w/o fun. eval./lin. alg.)  =  6.892
Total wall-clock secs in linear solver                      =  4.408
Total wall-clock secs in NLP function evaluations           =  0.660
Total wall-clock secs                                       = 11.960
```

  - Linear Sovler: only 0.708 secs out of 4.408 secs are spent on GPU.
  - AD: only 0.104 secs out of 0.660 secs are spent on GPU.
  - MadNLP: operations like $y \leftarrow y + ax$ are $\times 10$ faster on GPUs.
- By running everything on GPUs, we may solve `case9241` in less than 4 secs (JuMP+Ipopt takes 40 secs).
- The group in LLNL may be looking at the same problem (based on their activities on Github).
- Implement the GPU version (2-3 weeks of work) and submit this to for PSCC next year?

Sparse Condensed-Space Interior-Point Methods with Inequality Relaxations on GPUs: Will it Work?

Sungho Shin

sshin@anl.gov

Argonne National Laboratory

June 27, 2023