SAND2024-08427C

Sandia National Laboratories

Exploiting Parallelism in Optimization with Evolutionary Constraints

Eric C. Cyr (PI, SNL), D. Ridzal, R. Vuchkov (SNL), J. Schroder, D. Vargas (UNM), N. Moore (West Texas A&M), M. Heinkenschloss (Rice), S. Lin (Mathworks), B. Southworth (LANL), J. Hahne (Bergische Universitat)

Overview

Data-driven modeling requires solutions of constrained optimization problems

Current optimization approaches for time dependent constraints serialize gradient-descent and second order methods, impacting:

- Deep neural network (DNN) training
- PDE-constrained optimization (PDEOpt)

$$\min_{z} \quad \text{Loss}(u(T)) + \text{Reg}(z)$$
$$\text{s.t.} \quad \partial_t u(t) = -F(u(t), z(t)) \text{ for } t \in [0, T]$$

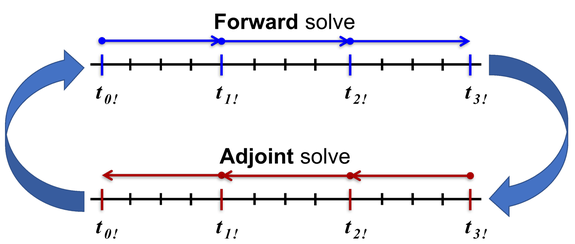
We develop parallel techniques that avoid the serialization using multigrid and domain decomposition methods

Deep Neural Networks

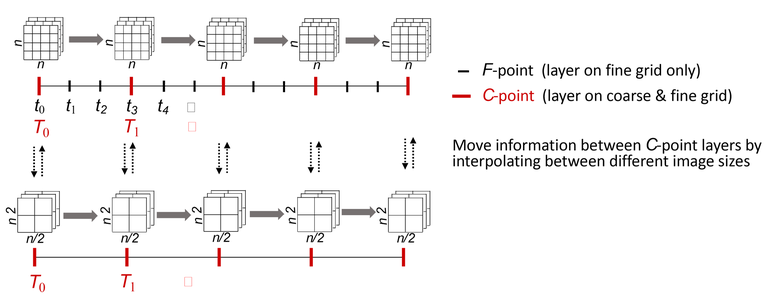
Neural ODEs replace DNN layers with time dim.

$$\underbrace{u_{l+1} = u_l - F(u_l, z_l)}_{\text{ResNet}} \Rightarrow \underbrace{\partial_t u = -F(u, z)}_{\text{NeuralODE}}$$

Computing gradients uses a forward ODE solve followed by a backward adjoint, this serializes

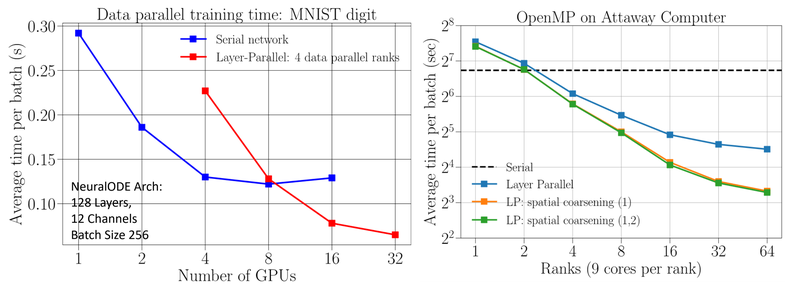


Our **Layer-Parallel**[1,2] algorithm uses MGRIT to accelerate forward and adjoint solves:

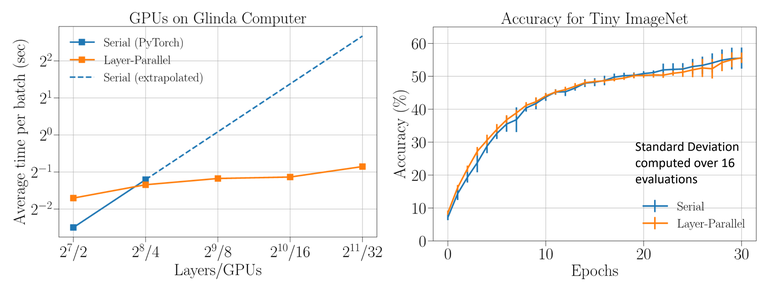


Recently we've added **spatial coarsening**[5]

Training Neural ODEs: with data-parallel (left), spatial coarsening (right)

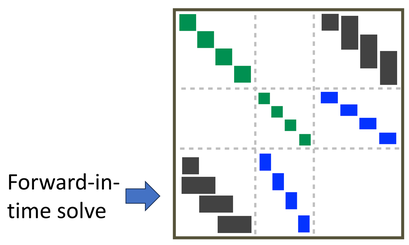


Layer-Parallel on Tiny ImageNet scales well with good accuracy :



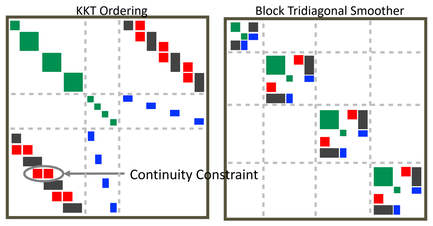
PDE-Constrained Optimization

PDEOpt: precondition augmented KKT matrices



First approach: **Multigrid-in-time**[6]

- Introduce continuity constraint
- Reorder to a block diagonal smoother relaxing continuity
- Geometric coarse grid correct
- Using inexact SQP method
- Scaling results on Burgers equation control



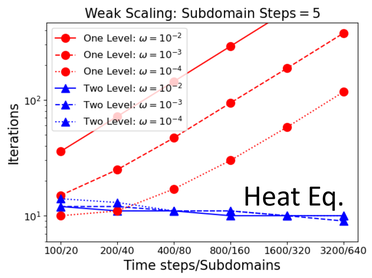
	$\nu = 10^{-1}$			$\nu = 10^{-2}$		
Ns	CG	Ls	Opt	CG	Ls	Opt
32	22	3.94	3	26	5.46	4
64	47	4.84	5	25	7.16	4
128	15	5.92	3	29	9.67	4
256	15	6.96	3	27	12.43	4
512	11	8.10	3	33	14.42	6
1024	14	9.75	3	17	15.97	5
2048	8	10.56	3	22	17.06	4

Second approach: **Domain-Decomposition**[4]

- Schur-complement (SC) is elliptic in time

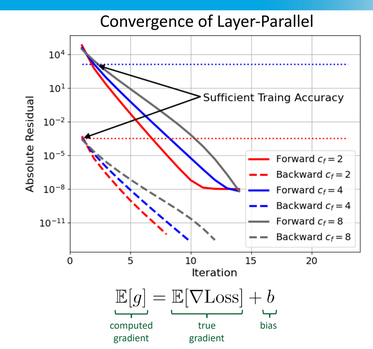
$$\sum_{s=1}^{N_D} R_s^T D_s \left(R_s Q_s^T \hat{J}_s^{-T} M_s \hat{J}_s^{-1} Q_s R_s^T \right) R_s$$

- Applies parallel domain decomposition (DD) to SC



Biased Gradients in SGD

Layer-Parallel derives parallelism by inexactly computing the batch gradient.



$$\underbrace{\mathbb{E}[g]}_{\text{computed gradient}} = \underbrace{\mathbb{E}[\nabla \text{Loss}]}_{\text{true gradient}} + \underbrace{b}_{\text{bias}}$$

Our theory[3,7] extends stochastic gradient descent (SGD) for statistically biased gradients

Theorem (Lin, 2022): For a nonconvex objective, with inexact gradients, and learning rate $\alpha_k < 1/L$, then

$$\sum_{k=1}^{K} (\alpha_k - L\alpha_k^2)\, \mathbb{E}[\|\nabla F(\Theta_k)\|^2] \leq 2(F(\Theta_1) - F^*) + \Delta B + L\nu^2 \sum_{k=1}^{K} \alpha_k^2$$

ΔB: Controls Inexactness

Next steps will look to exploiting this theory to yield new parallel training algorithms through the SEA-CROGS project

Acknowledgement

We are grateful for funding from

- DOE Early Career Research Program
- MMICCs project SEA-CROGS
- Sandia LDRD Program

Our References:

1. S. Guenther, L. Ruthotto, J. B. Schroder, E. C. Cyr, N. R. Gauger, Layer-Parallel Training of Deep Residual Neural Networks, SIMODs, Vol. 2 (1), 2020.
2. E. C. Cyr, G. Moon, Parallel Training of GRU Networks with a Multi-Grid Solver for Long Sequences, ICLR 2022.
3. S. Lin, Multilevel-in-Time Methods for Optimal Control of PDEs and Training of Recurrent Neural Networks 2022 (PhD thesis, Rice University).
4. E. C. Cyr, A 2-Level Domain Decomposition Preconditioner for KKT Systems with Heat-Equation Constraints, DD Methods in Science and Engineering XXVII, Series: Lecture Notes in Computational Science and Engineering, Springer, 2023.
5. E. C. Cyr, J. Hahne, N. S. Moore, J. B. Schroder, B. S. Southworth, D. A. Vargas, TorchBraid: High-Performance Layer-Parallel Training of Deep Neural Networks with MPI and GPU Acceleration, Submitted to ACM TOMS, 2023.
6. E. C. Cyr, D. Ridzal, R. Vuchov, Multigrid-in-time preconditioner for KKT systems, Submitted to Comp. Optimization and Applications, 2024.
7. S. Lin, M. Heinkenschloss, E. C. Cyr, Inexact Stochastic Gradient Methods and Training of Recurrent Neural Networks, In Preparation.

U.S. DEPARTMENT OF ENERGY
NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories