Pacific Northwest
NATIONAL LABORATORY

PNNL-32925

Scalable Second Order Optimization for Machine Learning

May 2022

Andrew Lumsdaine

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-32925

Scalable Second Order Optimization for Machine Learning

May 2022

Andrew Lumsdaine

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Project Aims

Many machine learning (ML) training tasks are essentially optimization processes that would at first glance appear eminently parallelizable and scalable. However, effective acceleration of these tasks with scalable parallel hardware has proven to be elusive. While standard methods for machine learning, e.g., stochastic gradient descent (SGD) for DNNs, tend to be resource efficient, they appear to be fundamentally sequential in nature.

The increasing availability of scalable computing platforms (including accelerators) presents an opportunity for more sophisticated approaches to be developed. The idea behind such approaches is to apply second-order optimization approaches that, while potentially requiring more memory and computation than a first-order approach, would enable significantly faster convergence while amortizing the resource requirements across a scalable computing platform. The ultimate goal of our work in higher-order solvers is to enable much more rapid time to solution than is currently achieved. The work reported here was part of a larger research agenda aimed at making ML training scalable and significantly improving their performance.

The specific focus of this project was to continue the development of a software library of advanced second order optimization for accelerating ML training. The project designed and prototyped selected second order optimization algorithms in PyTorch [1] and evaluated their convergence behavior and performance with the CIFAR10 dataset [2], using medium to large models (e.g., ResNet18 and ResNet50 [3]). The results obtained were promising. Second-order methods were shown to be competitive with highly-tuned first-order methods such as SGD/Adam [4], suggesting the need for continued research in this area.

The software developed as part of this LDRD will be released publicly as an open-source package to the community. It includes a wide variety of second-order algorithms and supporting functionality, including:

- Newton-Krylov optimizer [5] using a matrix-free conjugate-residual algorithm [6],
- Quasi-Newton optimizers, including limited-memory versions of Broyden, Davidon-Fletcher-Powell (DFP), and Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithms [7]–[9],
- Nonlinear conjugate-direction optimizers, including Fletcher-Reeves [10], Polak-Ribiere [11], Hestenes-Stiefel [12], and Dai-Yuan[13] algorithms,
- Line-search methods, including Armijo [14] and Wolfe [15],
- Trust-region methods, including Levenberg and Levenberg-Marquardt [9], and
- Homotopy-continuation methods [16].

By making this comprehensive software library of second-order methods available in PyTorch, we hope to enable the larger ML community to experiment with them and to develop highly-optimized and scalable approaches based on them.

Key Project Accomplishments

Open Source Repository: https://github.com/pnnl/pytorch_soo

Acknowledgment

Portions of this work were done in collaboration with Eric Silk (University of Washington, Schweitzer Engineering Laboratories) and Tony Chiang (PNNL).

The research described in this report was conducted under the Laboratory Directed Research and Development Program at the U.S. Department of Energy's Pacific Northwest National Laboratory (PNNL), a multiprogram national laboratory operated by the Battelle Memorial Institute under Contract DE-AC06-76RL01830.

Bibliography

[1] A. Paszke *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[2] A. Krizhevsky, *CIFAR10*. 2009.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition." arXiv, 2015. doi: 10.48550/ARXIV.1512.03385.

[4] E. H. Chang *et al.*, "The mixed effects of online diversity training," *Proc. Natl. Acad. Sci.*, vol. 116, no. 16, pp. 7778–7783, Apr. 2019, doi: 10.1073/pnas.1816076116.

[5] P. N. Brown and Y. Saad, "Hybrid Krylov Methods for Nonlinear Systems of Equations," *SIAM J. Sci. Stat. Comput.*, vol. 11, no. 3, pp. 450–481, May 1990, doi: 10.1137/0911026.

[6] Y. Saad, *Iterative methods for sparse linear systems*. SIAM, 2003. doi: 10.1137/1.9780898718003.

[7] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Society for Industrial and Applied Mathematics, 1996. doi: 10.1137/1.9781611971200.

[8] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. Elsevier, 1970. doi: 10.1016/C2013-0-11263-9.

[9] J. Nocedal and S. J. Wright, *Numerical optimization*, 2nd ed. New York: Springer, 2006.

[10] R. Fletcher, "Function minimization by conjugate gradients," *Comput. J.*, vol. 7, no. 2, pp. 149–154, Feb. 1964, doi: 10.1093/comjnl/7.2.149.

[11] E. Polak and G. Ribiere, "Note sur la convergence de méthodes de directions conjuguées," *ESAIM Math. Model. Numer. Anal. - Modélisation Mathématique Anal. Numér.*, vol. 3, no. R1, pp. 35–43, 1969.

[12] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *J. Res. Natl. Bur. Stand.*, vol. 49, no. 6, p. 409, Dec. 1952, doi: 10.6028/jres.049.044.

[13] Y. H. Dai and Y. Yuan, "A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property," *SIAM J. Optim.*, vol. 10, no. 1, pp. 177–182, Jan. 1999, doi: 10.1137/S1052623497318992.

[14] L. Armijo, "Minimization of Functions Having Lipschitz Continuous First Partial Derivatives," *Pac. J. Math.*, vol. 16, no. 1, 1966.

[15] P. Wolfe, "Convergence Conditions for Ascent Methods," *SIAM Rev.*, vol. 11, no. 2, 1969.

[16] H. B. Keller, "Global Homotopies and Newton Methods," in *Recent Advances in Numerical Analysis*, Elsevier, 1978, pp. 73–94. doi: 10.1016/B978-0-12-208360-0.50009-7.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov