ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

modern machines. The results show that the bigger the size of the problem and the batch size are the more using SW-SGD reduces the cache hits misses.

5 Conclusion and future work

In this paper, we introduced SW-SGD which adapts mini batch GD to the access characteristics of modern HPC memory hierarchies to gain extra gradient noise smoothing, hopefully for free. In this way it will combine the epoch efficiency of one point SGD with the lower noise and easier to spot convergence of mini batch SGD. We compare the approach to mini batch SGD using different experimental data sets. We show that SW-SGD can improve over mini batch SGD in terms of convergence, for a given number of loads of training points from the large slow memory level and that is applicable for different variants of SGD.

In subsequent work we will expand the experiments to better understand under what circumstances SW-SGD should be used and how to dimension the cache. Future work should also address how to adapt the algorithm for parallel and distributed settings.

6 Acknowledgements

This work is funded by the European project ExCAPE which received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant no. 671555.

References

[1] Tieleman, Tijmen and Hinton, Geoffrey (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning

[2] Duchi, John; Hazan, Elad; Singer, Yoram (2011). "Adaptive subgradient methods for online learning and stochastic optimization". JMLR 12: pp. 2121–2159.

[3] Zeiler, Matthew D. Adadelta: An adaptive learning rate method. arXiv:1212.5701, 2012.

[4] Schaul, T., Zhang, S., and LeCun, Y. . No more pesky learning rates. arXiv:1206.1106, 2012.

[5] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. Technical report, http://arxiv.org/abs/1012.1367, 2010

[6] S. Shalev-Shwartz and T. Zhang. Accelerated mini-batch stochastic dual coordinate ascent. In Advances in Neural Information Processing Systems, pp. 378–385, 2013.

[7] A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. In NIPS, volume 24, pp. 1647–1655, 2011.

[8] S. Sallinen, N. Satish, M. Smelyanskiy, S. S. Sury and C. RÃ©, "High Performance Parallel Stochastic Gradient Descent in Shared Memory," 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS), Chicago, IL, 2016, pp. 873-882.

[9] P. Xinghao, L. Maximilian, Tu. Stephen, Pa. Dimitris, Zh. Ce, Jo. Michael, R. Kannan, Re. Chris, Re. Benjamin, "CYCLADES: Conflict-free Asynchronous Machine Learning" . eprint arXiv:1605.09721 05/2016.

[10] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86, 2278-2324.

[11] `http://sebastianruder.com/optimizing-gradient-descent/index.html#gradientdescentoptimizationalgorithms`