The 8th European Congress on Computational Methods in Applied Sciences and Engineering
ECCOMAS Congress 2022
5-9 June 2022, Oslo, Norway

# Role of rounding in implementing gradient descent with low-precision representation

**Lu Xia[1], Stefano Massei[2], Michiel E. Hochstenbach[3], and Barry Koren[4]**

[1,2,3,4] Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven, The Netherlands
[1]l.xia1@tue.nl, [2]s.massei@tue.nl, [3]m.e.hochstenbach@tue.nl, [4]b.koren@tue.nl

**Keywords**: *Gradient descent method, Convergence analysis, Stochastic rounding, Low-precision computation*

When implementing the gradient descent method (GD) in low precision, the rounding to the nearest (RN) method normally suffers from vanishing gradient problems [1, 2], while the gradient updates that are below the minimum rounding precision are partially captured, with certain probability, bt conventional stochastic rounding (CSR) [2]. In this study, we trade the zero bias property of CSR [3] with a larger probability to preserve small gradients. In particular, we propose a new stochastic rounding method with constant rounding bias that, in the context of GD, is always tuned in a descent direction. We prove that, for convex problems, the proposed rounding method ensures a tighter convergence rate than the one of CSR.

We validate our theoretical analysis by training a multinomial logistic regression model with 8-bit floating-point representation using different rounding methods. The experiments demonstrate that the new rounding method provides higher classification accuracy than those obtained using RN and CSR, with the same number of training epochs. It is shown that the convergence rate obtained by the new rounding method with 8-bit floating-point representation is even higher than the one of RN with 32-bit floating-point representation. Further, we test the approach for training a two-layer NN, that is a non-convex problem, using 8-bit floating-point number representation. Also in this case, the performance of the new rounding method is superior to those of RN and CSR.

## REFERENCES

[1] M. Höhfeld and S.E. Fahlman, Probabilistic rounding in neural network learning with limited precision. *Neurocomputing*, Vol. **4**, pp. 291–299, 1992.

[2] S. Gupta, A. Agrawal, K. Gopalakrishnan and P. Narayanan, Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1737–1746, 2015.

[3] M.P. Connolly, N.J. Higham and T. Mary, Stochastic rounding and its probabilistic backward error analysis. *SIAM J. Sci. Comput.*, Vol. **43**, pp. A566–A585, 2021.