# Low-bit quantization for training neural networks

Jonathan Ashbrock, Alexander M. Powell*
Department of Mathematics, Vanderbilt University, Nashville, USA
*Email:* alexander.m.powell@vanderbilt.edu

---

We investigate the problem of training neural networks with low-bit weights. This is motivated by settings where neural networks are trained on memory-constrained platforms. We study a modified version of stochastic gradient descent that only utilizes low-bit weight vectors at every stage of the training process. We show that this approach performs well numerically, and we also show mathematical error bounds for the associated quantization algorithm.