Training Quantization for 16 Bit Factorization Retraining with Maximum Weight Constraint B) Retraining with Quantization Loss 1.0 32 bit Floating Point No Retraining w/Max Weight Constraint 6 bit, No Retraining 8.0 6 bit, with Retraining *p*correct 0.0 + 9.0 0.2 0.0 10⁵ 10⁷ 10⁵ 10^1 10^1 10⁷ 10³ 10³ Samples Samples Weights distribution before Retraining D) Weights distribution after Quantization Retraining C) 10⁴ Frequency 10³ 10¹ 10⁰ -2 4 6 Weight Value Weight Value