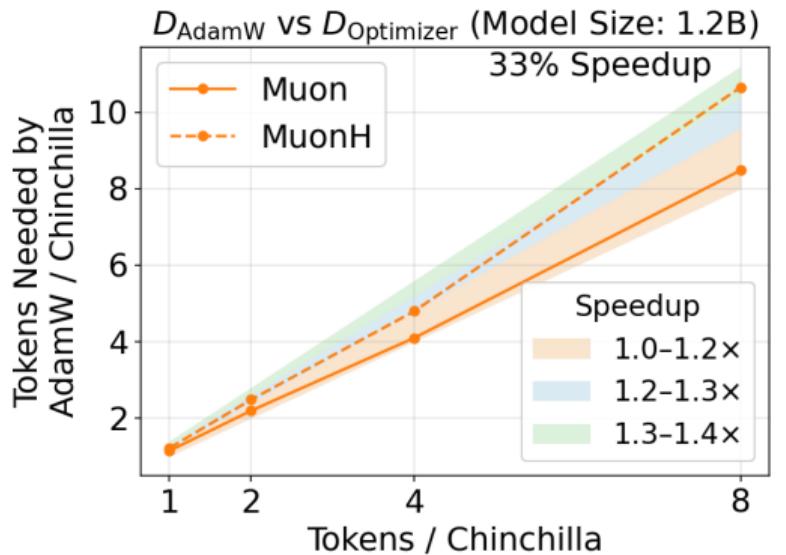
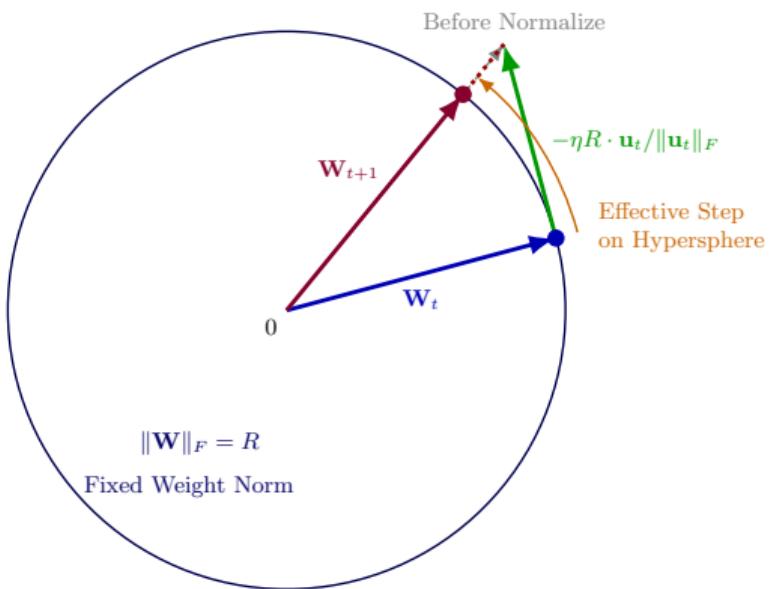


$R = \|W_0\|_F$ (initial norm), $\text{Normalize}(x) = x/\|x\|_F$, u_t : Adam/Muon update.

Hyperball Optimization: $W_{t+1} = R \cdot \text{Normalize}(W_t - \eta R \cdot \text{Normalize}(u_t))$.



Speedup over AdamW on pretraining 1.2B models