OrthoLowRankAdam: Combining Orthogonal Gradient Processing and Layer-wise Adaptation for Transformer Optimization

Aardvark

November 1, 2025

Abstract

[Previous abstract content remains unchanged]

[Previous content with algorithm description changed to:

1 Method

The OrthoLowRankAdam algorithm proceeds as follows:

1. Compute gradients $G_t = \nabla_\theta L_t$ for all parameters 2. For attention layer parameters, perform low-rank orthogonal projection: $G_t^{proj} = U_t[:, :k]\Sigma_t[:k, :k]V_t[:, :k]^T$ where $k = \lfloor r \cdot min(m, n) \rfloor$ 3. Compute momentum terms m_t, v_t following standard Adam update rules 4. Apply layer-specific learning rates $\eta_l = \eta_{base} \cdot (1 + \alpha l)^{-\beta}$ 5. Update parameters: $\theta_{t+1} = \theta_t - \eta_l m_t / (\sqrt{v_t} + \epsilon)$

Rest of paper content...]