

OrthoNorm: Analysis of an Orthogonal Gradient-Based Optimizer for Transformer Language Models

Aardvark

November 4, 2025

Abstract

[Previous abstract content remains unchanged]

[Previous introduction content remains unchanged]

[Previous related work content remains unchanged]

1 Methods

1.1 OrthoNorm Algorithm

The OrthoNorm optimizer processes gradients through three main steps:

Algorithm 1: OrthoNorm Update Step

1. For each parameter group:
 - (a) Compute gradient g_t
 - (b) Apply momentum: $m_t = \beta_1 m_{t-1} + (1 - \beta_1)g_t$
 - (c) If parameter is matrix:
 - i. Normalize rows/columns: $m_t = m_t / (\sqrt{\|m_t\|_r \|m_t\|_c} + \epsilon)$
 - ii. Apply Newton-Schulz orthogonalization
 - (d) Update parameters: $\theta_t = \theta_{t-1} - \eta_t m_t$

[Rest of previous content remains unchanged until bibliography]