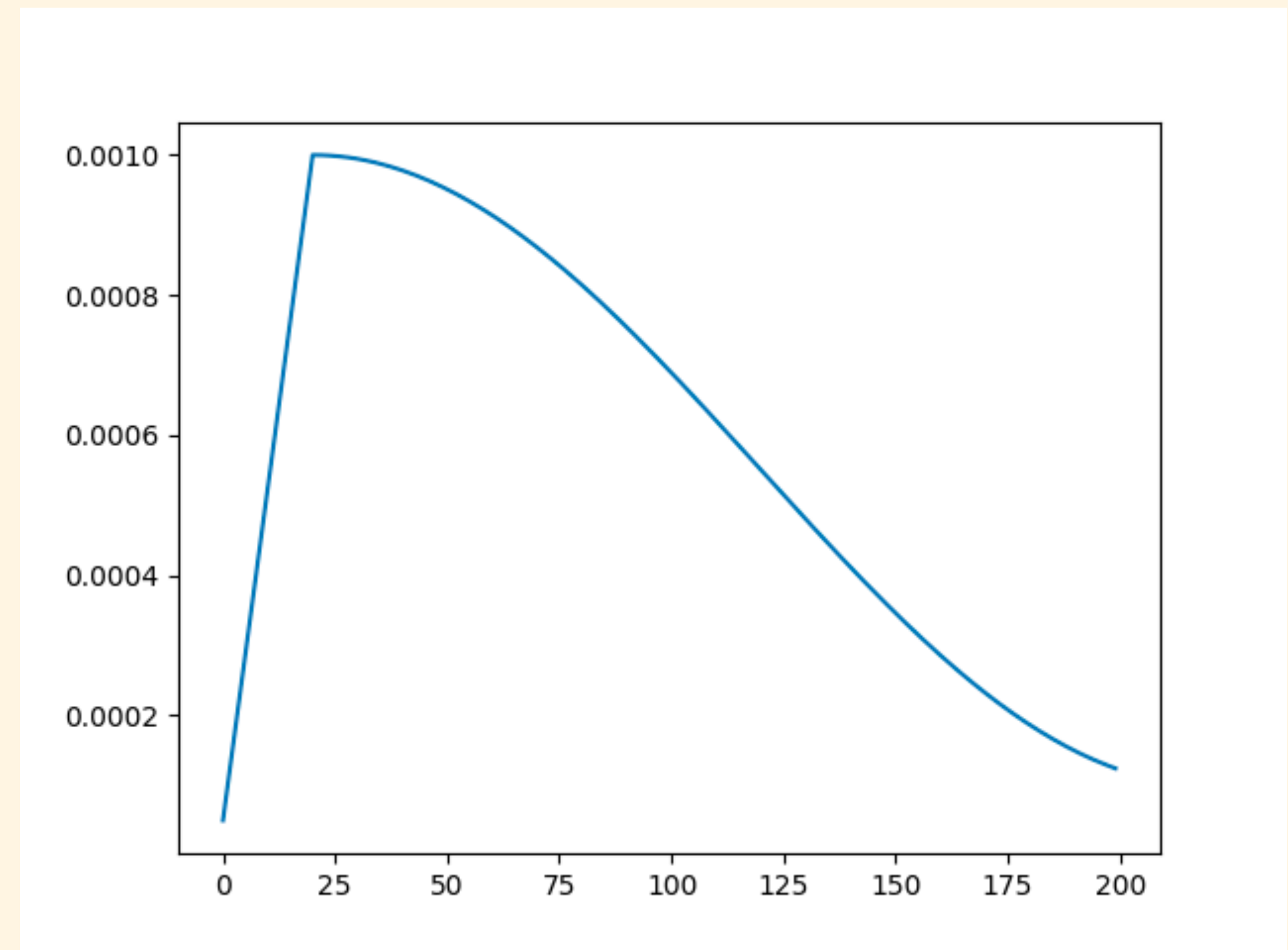


最適化設定：学習速度

- 学習初期には最大値にいたるまで learning rate を線形で徐々に増加させる (Warmup)
- その後、コサイン関数に従って徐々に学習率を低下させていく (Cosine Decay)



スケーラブルなトレーニング

- 効率的に学習を行うには、スループットを増やし、より大きなモデルをGPUメモリにロードする必要がある
- 3次元並列処理
 - ❖ データ並列：学習コーパスを分散させる
 - ❖ パイプライン並列：LLMの異なるレイヤーを複数のGPUに分割する
 - ❖ テンソル並列：LLMのパラメータ行列を分解する
- ZeRO：GPUのメモリ冗長性を排除する
- 混合精度トレーニング：パラメータのビット数を削減する