AutoLoRA with SVD: Adaptive Rank Selection for Efficient LoRA Fine-Tuning

1 Introduction

AutoLoRA with Singular Value Decomposition (SVD) is an approach that dynamically determines the optimal rank for LoRA layers based on the singular value spectrum of the model's weight matrices. The goal is to adaptively allocate LoRA rank, ensuring efficient adaptation while maintaining computational efficiency.

2 Key Idea Behind AutoLoRA with SVD

Instead of setting a fixed rank for all layers in LoRA, this method:

- Uses SVD on weight updates to determine the most informative singular values.
- Adapts the LoRA rank per layer based on the decay of singular values.
- Prunes unimportant components dynamically during training.

3 How AutoLoRA with SVD Works

3.1 Compute SVD on Weight Updates

For a given weight matrix W, LoRA introduces a low-rank decomposition:

$$W' = W + \Delta W \tag{1}$$

where $\Delta W = AB$ (with A and B being low-rank matrices).

To find the optimal rank r:

1. Compute SVD of the weight update matrix:

$$\Delta W = U \Sigma V^T \tag{2}$$

where:

- U and V are orthonormal matrices.

- Σ is a diagonal matrix containing singular values.

2. Analyze the singular value decay:
   - Large singular values contribute more to model adaptation.
   - Small singular values can be pruned without significant loss in expressivity.
3. Define an adaptive threshold:

   Rank selection can be based on energy retention, e.g.:

$$\sum_{i=1}^{r} \sigma_i^2 \geq \alpha \sum_{i=1}^{\text{all}} \sigma_i^2 \tag{3}$$

   where α (e.g., 95%) controls how much of the energy is preserved.
4. Select the minimum rank r that satisfies the condition.

3.2 Apply Adaptive Rank LoRA

- Instead of setting a fixed r, each layer gets its own rank based on the SVD results.
- The rank can be updated per iteration or at predefined intervals.

3.3 Efficient Training with Rank Adaptation

- **Early Training:** LoRA starts with a higher initial rank.
- **Dynamic Pruning:** As training progresses, SVD is periodically applied, and small singular values are removed, reducing rank dynamically.
- **Final Adaptation:** The rank settles into an efficient lower-dimensional subspace that retains key adaptation capacity.

4 Advantages of AutoLoRA with SVD

- **Adaptive Rank Selection** – Different layers get different ranks based on their importance.
- **Efficient Computation** – Reduces memory and computation compared to fixed-rank LoRA.
- **Better Generalization** – Avoids overfitting by pruning unnecessary components.
- **Energy-Based Optimization** – Ensures high information retention with minimal redundancy.