BiLoRA: Almost-orthogonal Parameter Spaces for Continual Learning

Supplementary Material

Authors:
Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz.

Appendices: Proofs of Theoretical Results

A.1 Proof of Theorem 1 (Frequency Separation)

Theorem 1 (Frequency Separation). For tasks i and j, let $\mathbf{S}_i$ and $\mathbf{S}_j$ be independently and uniformly sampled frequency selection masks with k components each from a total of d^2 components. For $k \ll d$, the probability of perfect separation (no frequency collision) is at least:

$$p\big(\text{supp}(\mathbf{S}_i) \cap \text{supp}(\mathbf{S}_j) = \emptyset\big) \geq 1 - \frac{k^2}{d^2}.$$

Proof. We approach this proof using the probabilistic method. Let $\mathbf{S}_i$ and $\mathbf{S}_j$ be binary masks where exactly k elements are set to 1 (representing the selected frequency components) and the remaining elements are 0.

1) The probability of a collision (*i.e.*, $\text{supp}(\mathbf{S}_i) \cap \text{supp}(\mathbf{S}_j) \neq \emptyset$) can be analyzed using the complementary counting principle. Let us consider the probability that there is at least one frequency component that is selected by both tasks.

2) First, we select k frequency components for task i uniformly at random from the d^2 available components. Then, we select k components for task j similarly. The probability that a specific component in $\mathbf{S}_j$ collides with any component in $\mathbf{S}_i$ is $\frac{k}{d^2}$, since exactly k out of d^2 components are selected for $\mathbf{S}_i$.

3) Using the union bound, the probability of at least one collision among the k components of $\mathbf{S}_j$ is bounded by:

$$p(\text{collision}) \leq k \cdot \frac{k}{d^2} = \frac{k^2}{d^2}.$$

4) Therefore, the probability of perfect separation (no collision) is:

$$p(\text{supp}(\mathbf{S}_i) \cap \text{supp}(\mathbf{S}_j) = \emptyset) = 1 - P(\text{collision}) \geq 1 - \frac{k^2}{d^2}.$$

Note that this bound is derived using the union bound, which can be loose. The exact probability of "no collision" can be computed using the hypergeometric distribution as:

$$p\big(\text{supp}(\mathbf{S}_i) \cap \text{supp}(\mathbf{S}_j) = \emptyset\big) = \frac{\binom{d^2-k}{k}}{\binom{d^2}{k}}.$$

However, for $k \ll d^2$, our bound $1 - \frac{k^2}{d^2}$ provides a good approximation and conveys the key insight that the collision probability scales quadratically with the ratio $\frac{k}{d^2}$.

A.2 Proof of Theorem 2 (Task Interference Bound)

Theorem 2 (Task Interference Bound). For tasks i and j with updates $\Delta\mathbf{W}_i$ and $\Delta\mathbf{W}_j$, with probability at least $1 - \delta$:

$$\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\|_F \leq \varepsilon,$$

when the number of frequency components k satisfies:

$$k \leq c\frac{d^2}{T}\log\Big(\frac{1}{\delta}\Big),$$

where c is a universal constant and T is the total number of tasks.

Proof. 1) We begin by expressing the weight updates in the frequency domain using our bilinear formulation:

$$\Delta\mathbf{W}_i = \mathbf{F}\mathbf{B}_i\mathbf{F}^H,$$
$$\Delta\mathbf{W}_j = \mathbf{F}\mathbf{B}_j\mathbf{F}^H,$$

where $\mathbf{B}_i$ and $\mathbf{B}_j$ are sparse matrices with at most k non-zero components each, and $\mathbf{F}$ is the DFT matrix.

2) The interference between tasks can be quantified by the Frobenius norm of the product of their weight updates:

$$\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\|_F = \|(\mathbf{F}\mathbf{B}_i\mathbf{F}^H)^T(\mathbf{F}\mathbf{B}_j\mathbf{F}^H)\|_F.$$

3) Since $\mathbf{F}$ is unitary ($\mathbf{F}^H\mathbf{F} = \mathbf{I}$), we can simplify it:

$$\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\|_F = \|(\mathbf{F}^H)^T\mathbf{B}_i^T\mathbf{F}^T\mathbf{F}\mathbf{B}_j\mathbf{F}^H\|_F = \|\mathbf{F}^*\mathbf{B}_i^T\mathbf{B}_j\mathbf{F}^H\|_F.$$

4) Using the unitary invariance of the Frobenius norm (i.e., $\|\mathbf{U}\mathbf{A}\mathbf{V}\|_F = \|\mathbf{A}\|_F$ for unitary matrices $\mathbf{U}$ and $\mathbf{V}$), we have:

$$\|\Delta\mathbf{W}_i^T \Delta\mathbf{W}_j\|_F = \|\mathbf{B}_i^T\mathbf{B}_j\|_F.$$

5) Now, the interference depends entirely on the overlap between the frequency components selected for each task. When $\mathbf{B}_i$ and $\mathbf{B}_j$ have non-overlapping support (i.e., $\text{supp}(\mathbf{B}_i) \cap \text{supp}(\mathbf{B}_j) = \emptyset$), we have $\mathbf{B}_i^T\mathbf{B}_j = \mathbf{0}$, resulting in zero interference.

6) From Theorem 1, the probability of perfect separation between any two tasks is at least $1 - \frac{k^2}{d^2}$. For T tasks, there are $\binom{T}{2} < \frac{T^2}{2}$ pairs of tasks that could potentially interfere. Using the union bound, the probability that at least one pair of tasks has a collision is bounded by:

$$p(\text{any collision}) \leq \binom{T}{2} \cdot \frac{k^2}{d^2} < \frac{T^2}{2} \cdot \frac{k^2}{d^2} = \frac{T^2k^2}{2d^2}.$$

7) Setting this probability to be at most δ, we get:

$$\frac{T^2k^2}{2d^2} \leq \delta.$$

8) Solving for k, we get:

$$k \leq \frac{d\sqrt{2\delta}}{T}.$$

9) For small δ, we can express this as:

$$k \leq c\frac{d^2}{T}\log\left(\frac{1}{\delta}\right),$$

where c is a constant that absorbs the factor $\sqrt{2}$ and the approximation from the square root to the logarithm for small δ.

10) When this condition is satisfied, with probability at least $1-\delta$, all pairs of tasks will have non-overlapping frequency components, resulting in $\|\mathbf{B}_i^T\mathbf{B}_j\|_F = 0$ for all $i \neq j$. Even in cases where there is some overlap, the interference remains bounded by the magnitude of the components and the extent of the overlap, which we can bound by a small constant ε with high probability.

This proves that by appropriately setting the number of frequency components k based on the total number of tasks T and the desired confidence level $1-\delta$, one can ensure that the interference between any pair of tasks remains below a threshold ε with high probability.