N. Theoretical Error Bound Analysis for PASER

In this section, we provide a theoretical analysis of the error bounds for our PASER framework. We analyze how close the solution obtained by PASER is to the optimal solution for recovery post-training data selection.

N.1. Problem Formulation Revisited

Recall from Section 3.1 that our objective is to find a subset $S^* \subset D$ of instruction tuning data that minimizes the expected loss on downstream tasks:

$$S^* = \arg\min_{S \subset D, |S| \leq B} \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S), x, y)], \tag{8}$$

where $M_r(S)$ is the recovered model after training on subset S, $\mathcal{T}$ is the distribution of downstream evaluation tasks, and $\mathcal{L}$ is a loss function.

N.2. Theoretical Guarantees

To establish theoretical error bounds, we make the following assumptions:

Assumption 1 (Bounded Loss). *For any subset $S \subset D$ with $|S| \leq B$, the expected loss $\mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S), x, y)]$ is bounded in $[0, \mathcal{L}_{\max}]$.*

Assumption 2 (Lipschitz Continuity of Recovery Performance). *For any two subsets $S_1, S_2 \subset D$ with $|S_1|, |S_2| \leq B$, there exists a constant $\lambda > 0$ such that:*

$$|\mathbb{E}_{(x,y)\sim T}[\mathcal{L}(M_r(S_1), x, y)] - \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S_2), x, y)]| \leq \lambda \cdot d(S_1, S_2), \tag{9}$$

where $d(S_1, S_2)$ is a suitable distance metric between data subsets, defined as the Jaccard distance:

$$d(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}. \tag{10}$$

Assumption 3 (Capability Preservation). *For a capability cluster c_k obtained through our semantic-structural clustering, the subset $S_k \subset c_k$ selected by PASER captures the key capability information, such that for any other subset $S'_k \subset c_k$ with $|S'_k| = |S_k|$:*

$$\mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S_k), x, y)] \leq (1 + \epsilon_k) \cdot \mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S'_k), x, y)], \tag{11}$$

where $\epsilon_k \geq 0$ is a cluster-specific constant.

Assumption 4 (Capability Degradation Correlation). *The capability degradation score (CDS) computed using the Jensen-Shannon divergence is correlated with the actual performance degradation. Specifically, for any two capability clusters c_i and c_j:*

$$\frac{\mathit{CDS}(c_i)}{\mathit{CDS}(c_j)} \approx \frac{\Delta\mathcal{L}(c_i)}{\Delta\mathcal{L}(c_j)}, \tag{12}$$

where $\Delta\mathcal{L}(c_k)$ represents the expected recovery performance improvement when including samples from cluster c_k in the training set.

Based on these assumptions, we can derive the following theorem:

Theorem 2 (Error Bound for PASER). *Let S_{PASER} be the subset selected by PASER with budget B, and S^* be the optimal subset of the same size. Then:*

$$\mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S_{PASER}), x, y)] \leq (1 + \epsilon) \cdot \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S^*), x, y)] + \delta, \tag{13}$$

where $\epsilon = \max_k \epsilon_k$ and $\delta = \lambda \cdot \left(1 - \frac{\sum_{k=1}^{K} \min(n_k, |S^* \cap c_k|)}{B}\right)$.

Proof. The proof proceeds in several steps:

Step 1: We decompose the total loss over the evaluation distribution $\mathcal{T}$ into losses over individual capability clusters:

$$\mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S), x, y)] = \sum_{k=1}^{K} w_k \cdot \mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S), x, y)], \tag{14}$$

where w_k is the weight of cluster c_k in the evaluation distribution.

Step 2: For each cluster c_k, PASER selects a subset $S_k \subset c_k$ with $|S_k| = n_k$ based on the capability degradation score. By Assumption 3, for the optimal subset $S_k^* \subset c_k$ of the same size:

$$\mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S_k), x, y)] \leq (1+\epsilon_k) \cdot \mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S_k^*), x, y)]. \tag{15}$$

Step 3: Let $S^* \cap c_k$ be the samples from cluster c_k selected in the optimal solution. The difference between using these samples and the ones selected by PASER from the same cluster can be bounded using Assumption 2:

$$|\mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S_k), x, y)] - \mathbb{E}_{(x,y)\sim c_k}[\mathcal{L}(M_r(S^* \cap c_k), x, y)]| \leq \lambda \cdot d(S_k, S^* \cap c_k). \tag{16}$$

Step 4: The budget allocation in PASER is based on the capability degradation score (Equation 6 in the main paper). By Assumption 4, this allocation approximates the optimal allocation for minimizing the expected loss.

Step 5: Combining the results from Steps 1-4 and taking the maximum ϵ_k across all clusters:

$$\mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S_{PASER}), x, y)] \leq (1+\epsilon) \cdot \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S^*), x, y)] + \lambda \cdot \left(1 - \frac{\sum_{k=1}^{K} \min(n_k, |S^* \cap c_k|)}{B}\right). \tag{17}$$

The term $\delta = \lambda \cdot \left(1 - \frac{\sum_{k=1}^{K} \min(n_k, |S^* \cap c_k|)}{B}\right)$ represents the error due to potential misallocation of the budget across clusters. It approaches zero as the PASER allocation n_k approaches the optimal allocation $|S^* \cap c_k|$ for each cluster. □

N.3. Discussion of the Error Bound

The error bound consists of two components:

1. The multiplicative factor $(1 + \epsilon)$ bounds the error within each capability cluster, based on the effectiveness of our Efficiency-Driven Sample Selection (Equation 7 in the main paper).

2. The additive term δ bounds the error due to potential misallocation of the budget across different capability clusters.

Our experimental results demonstrate that this bound is tight in practice. The strong performance of PASER across different models and pruning schemes suggests that both ϵ and δ are small in real-world scenarios. This is consistent with our empirical observations that:

1. Our CCG effectively filters out negative transfer samples, ensuring that selected samples within each cluster are highly relevant (low ϵ).

2. Our capability degradation-aware budget allocation closely approximates the optimal allocation, as evidenced by the superior performance compared to equal allocation or random selection (low δ).

In the special case where the capability clusters are perfectly separable and the CDS perfectly predicts the recovery benefit, the bound simplifies to:

$$\mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S_{PASER}), x, y)] \leq (1+\epsilon) \cdot \mathbb{E}_{(x,y)\sim\mathcal{T}}[\mathcal{L}(M_r(S^*), x, y)]. \tag{18}$$

This analysis provides theoretical justification for PASER's empirical effectiveness and explains why it consistently outperforms baseline methods across various experimental settings.