

## Supplemental Material: Efficient Representativeness-Aware Coreset Selection

### A Proof of Theorem 1

**Note:** Due to the non-convexity of deep learning models, the complexity of gradient distributions, and the dynamic nature of training, our theoretical analysis here is limited to simplified conditions that help us understand the behavior of the proposed algorithm under idealized settings. These results are intended to offer theoretical insight rather than serve as the main contribution of this work. The practical effectiveness of the method is ultimately demonstrated through empirical results.

**Theorem 1** (Relative Gradient Approximation Concentration Bound). *Consider a supervised learning task with parameter vector  $w_t \in \mathbb{R}^d$  at training step  $t$ . Let:*

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$  denote the full training dataset
- $\mathcal{S} \subset \mathcal{D}$  be a coreset with cardinality  $|\mathcal{S}| = m$
- $g_i := \nabla \mathcal{L}(w_t; x_i, y_i) \in \mathbb{R}^d$  represent per-sample gradients
- $\mu_t := \frac{1}{N} \sum_{i=1}^N g_i$  be the full-data mean gradient (signal)
- $\hat{\mu}_t := \frac{1}{m} \sum_{j \in \mathcal{S}} g_j$  denote the coreset gradient mean
- $\hat{\sigma}_t^2 := \|\text{Cov}_{\mathcal{S}}(g_j)\|$  be the spectral norm of coreset gradient covariance
- $\text{SNR}_t := \|\hat{\mu}_t\|/\hat{\sigma}_t$  define the signal-to-noise ratio

Under the following assumptions:

- (A1) **Independent Sampling:** Coreset indices are sampled i.i.d. from  $\mathcal{D}$  (or equivalently,  $N \gg m$  allowing approximation by i.i.d. sampling)
- (A2) **Sub-Gaussian Gradients:** Centered gradients satisfy the sub-Gaussian condition

$$\forall u \in \mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\| = 1\}, \mathbb{E} \left[ \exp \left( \frac{\langle u, g_j - \mu_t \rangle^2}{\hat{\sigma}_t^2} \right) \right] \leq 2$$

Then for any confidence parameter  $\delta \in (0, 1)$ , the relative gradient approximation error

$$\epsilon_t(\mathcal{S}) := \frac{1}{\|\mu_t\|} \|\mu_t - \hat{\mu}_t\|$$

satisfies the probabilistic guarantee:

$$\mathbb{P} \left( \epsilon_t(\mathcal{S}) \leq \frac{2}{\text{SNR}_t} \sqrt{\frac{2}{m} \log \left( \frac{2 \cdot 5^d}{\delta} \right)} \approx \frac{C \sqrt{\log(1/\delta)}}{\sqrt{m} \cdot \text{SNR}_t} \right) \geq 1 - \delta \quad (12)$$

where  $C > 0$  is a constant.

**Proof.** Let  $g_j := \nabla \mathcal{L}(w_t; x_j, y_j)$  denote the per-sample gradient, and define the full-data mean gradient as:

$$\mu_t := \frac{1}{N} \sum_{i=1}^N g_i,$$

and the empirical mean of the coreset gradients as:

$$\hat{\mu}_t := \frac{1}{m} \sum_{j \in \mathcal{S}} g_j.$$

758 We aim to bound the relative gradient approximation error:

$$\epsilon_t(\mathcal{S}) := \frac{1}{\|\hat{\mu}_t\|} \|\mu_t - \hat{\mu}_t\|.$$

759 **Step 1: Centered gradient representation.**

760 Let  $Z_j := g_j - \mu_t$  be the centered gradients. Then:

$$\hat{\mu}_t - \mu_t = \frac{1}{m} \sum_{j=1}^m Z_j,$$

761 so that:

$$\epsilon_t(\mathcal{S}) = \frac{1}{\|\hat{\mu}_t\|} \left\| \frac{1}{m} \sum_{j=1}^m Z_j \right\|.$$

762 **Step 2: One-dimensional projection concentration.**

763 By assumption (A2), for all unit vectors  $u \in \mathbb{S}^{d-1}$ :

$$\mathbb{E} \left[ \exp \left( \frac{\langle u, Z_j \rangle^2}{\hat{\sigma}_t^2} \right) \right] \leq 2.$$

764 This implies that each projection  $\langle u, Z_j \rangle$  is a sub-Gaussian random variable with parameter at most  
765  $\hat{\sigma}_t$ . Define the empirical projection mean:

$$X_u := \left\langle u, \frac{1}{m} \sum_{j=1}^m Z_j \right\rangle.$$

766 By standard sub-Gaussian concentration, for fixed  $u \in \mathbb{S}^{d-1}$  and any  $t > 0$ :

$$\mathbb{P}(|X_u| \geq t) \leq 2 \exp \left( -\frac{mt^2}{2\hat{\sigma}_t^2} \right).$$

767 **Step 3: Extend to vector norm via  $\varepsilon$ -net.**

768 We have:

$$\left\| \frac{1}{m} \sum_{j=1}^m Z_j \right\| = \sup_{u \in \mathbb{S}^{d-1}} |X_u|.$$

769 Let  $\mathcal{N}_{1/2}$  be a  $1/2$ -net of  $\mathbb{S}^{d-1}$  with cardinality at most  $5^d$ . Then for any vector  $v \in \mathbb{R}^d$ ,

$$\|v\| \leq 2 \max_{u \in \mathcal{N}_{1/2}} \langle u, v \rangle.$$

770 Applying the union bound over the finite net:

$$\mathbb{P}(\exists u \in \mathcal{N}_{1/2}, |X_u| \geq t) \leq 2 \cdot 5^d \cdot \exp \left( -\frac{mt^2}{2\hat{\sigma}_t^2} \right).$$

771 Since

$$\left\| \frac{1}{m} \sum_{j=1}^m Z_j \right\| \leq 2 \max_{u \in \mathcal{N}_{1/2}} |X_u|,$$

772 we obtain:

$$\mathbb{P} \left( \left\| \frac{1}{m} \sum_{j=1}^m Z_j \right\| \geq 2t \right) \leq 2 \cdot 5^d \cdot \exp \left( -\frac{mt^2}{2\hat{\sigma}_t^2} \right).$$

773 **Step 4: Solve for  $t$  in terms of confidence level  $\delta$ .**

774 Let the right-hand side be equal to  $\delta$ :

$$2 \cdot 5^d \cdot \exp \left( -\frac{mt^2}{2\hat{\sigma}_t^2} \right) = \delta.$$

775 Solving yields:

$$t = \hat{\sigma}_t \sqrt{\frac{2}{m} \log \left( \frac{2 \cdot 5^d}{\delta} \right)}.$$

776 Hence with probability at least  $1 - \delta$ ,

$$\|\hat{\mu}_t - \mu_t\| \leq 2\hat{\sigma}_t \sqrt{\frac{2}{m} \log \left( \frac{2 \cdot 5^d}{\delta} \right)}.$$

777 **Step 5: Normalize by  $\|\hat{\mu}_t\|$  and define SNR.**

778 Recall:

$$\epsilon_t(\mathcal{S}) = \frac{\|\mu_t - \hat{\mu}_t\|}{\|\hat{\mu}_t\|}, \quad \text{SNR}_t := \frac{\|\hat{\mu}_t\|}{\hat{\sigma}_t}.$$

779 Thus with probability at least  $1 - \delta$ :

$$\epsilon_t(\mathcal{S}) \leq \frac{2}{\text{SNR}_t} \sqrt{\frac{2}{m} \log \left( \frac{2 \cdot 5^d}{\delta} \right)}.$$

780 **Optional simplification:** For fixed  $d$ , the logarithmic term can be absorbed by a constant into  $\delta$ :

$$\epsilon_t(\mathcal{S}) \lesssim \frac{C \sqrt{\log(1/\delta)}}{\sqrt{m} \cdot \text{SNR}_t}.$$

781

□

## 782 B Boader Impact

783 Coreset selection algorithms aim to significantly reduce the training data volume and computational  
 784 overhead while preserving model performance, making them a key strategy for improving training  
 785 efficiency and reducing energy consumption in deep learning. By identifying a small yet highly  
 786 representative subset from the original training data, such methods can effectively shorten training  
 787 time and lower hardware resource usage, thereby reducing carbon emissions and mitigating the  
 788 environmental impact of AI development. In light of the growing global attention to sustainable  
 789 AI, advancing efficient and eco-friendly training strategies holds strong practical and long-term  
 790 value. Moreover, the reduced resource demand improves accessibility for researchers and developers  
 791 with limited computational capabilities, promoting greater inclusivity and equity in AI research and  
 792 innovation.

793 This study proposes a representativeness-aware coreset selection method based on gradient signal-  
 794 to-noise ratio (SNR), which significantly improves final model performance while maintaining the  
 795 computational efficiency of coreset selection. Compared to conventional approaches, our method  
 796 possesses the ability to aware the representativeness of selected samples during training, effectively  
 797 addressing the issue of coreset degradation in later stages. As a result, it achieves a favorable  
 798 balance between training efficiency and model performance. The method is theoretically grounded,  
 799 simple to implement, and easy to integrate into existing deep learning pipelines. It is well-suited  
 800 for both academic research and industrial deployment, particularly in edge scenarios and developer  
 801 communities with constrained computational resources.