Supplement to: ECOC-IL: Robust and Efficient Label LDP for Imbalanced Learning

Mengyang Li
Tianjin Normal University
Tianjin, China, 300387
limengyang@tjnu.edu.cn

Ou Wu
HIAS, University of Chinese Academy of Sciences
Hangzhou, China, 310024
wuou@ucas.ac.cn *

A. Proofs of Theoretical Conclusions

A.1 Proof for Lemma 1

Let $\mathcal{M}(y)$ denote the ECOC-RAPPOR mechanism for label y: first map y to $c_y = \mathbf{C}_{\text{opt}}[y, :]$, then perturb each bit of c_y via RAPPOR (Eq. 1) to get $\mathbf{z}$. For any $y_1 \neq y_2$ and $\mathbf{z} \in \{0,1\}^M$, we need to show $\mathbb{P}(\mathcal{M}(y_1) = \mathbf{z}) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y_2) = \mathbf{z})$.

Let $c_1 = \mathbf{C}_{\text{opt}}[y_1, :]$ and $c_2 = \mathbf{C}_{\text{opt}}[y_2, :]$. Since RAPPOR perturbation is bit-wise independent:

$$\mathbb{P}(\mathcal{M}(y_i) = \mathbf{z}) = \prod_{m=1}^{M} \mathbb{P}(z_m | c_{i,m}). \tag{1}$$

where $\mathbb{P}(z_m|c_{i,m})$ follows Eq. 1 with bit-wise budget ϵ/M:

$$\mathbb{P}(z_m = 1|c_{i,m}) = \begin{cases} \frac{e^{\epsilon/(2M)}}{1+e^{\epsilon/(2M)}} & \text{if } c_{i,m} = 1, \\ \frac{1}{1+e^{\epsilon/(2M)}} & \text{if } c_{i,m} = 0. \end{cases} \tag{2}$$

Let $S = \{m \mid c_{1,m} \neq c_{2,m}\}$ and $T = \{m \mid c_{1,m} = c_{2,m}\}$. For $m \in T$, $\mathbb{P}(z_m|c_{1,m}) = \mathbb{P}(z_m|c_{2,m})$, so these terms cancel out in the ratio $\frac{\mathbb{P}(\mathcal{M}(y_1)=\mathbf{z})}{\mathbb{P}(\mathcal{M}(y_2)=\mathbf{z})}$.

For $m \in S$, assume $c_{1,m} = 1$ and $c_{2,m} = 0$. Then:

$$\frac{\mathbb{P}(z_m|c_{1,m} = 1)}{\mathbb{P}(z_m|c_{2,m} = 0)} = \begin{cases} e^{\epsilon/(2M)} & \text{if } z_m = 1, \\ e^{-\epsilon/(2M)} & \text{if } z_m = 0. \end{cases} \tag{3}$$

The maximum ratio occurs when $z_m = 1$ for all $m \in S$:

$$\frac{\mathbb{P}(\mathcal{M}(y_1) = z)}{\mathbb{P}(\mathcal{M}(y_2) = z)} \leq \prod_{m \in S} e^{\epsilon/(2M)} = e^{\epsilon \cdot |S|/(2M)} \tag{4}$$

Since $|S| \leq M$ (as S is a subset of M bits):

$$e^{\epsilon \cdot |S|/(2M)} \leq e^{\epsilon \cdot M/(2M)} = e^{\epsilon/2} \leq e^{\epsilon} \tag{5}$$

Thus, $\mathbb{P}(\mathcal{M}(y_1) = \mathbf{z}) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y_2) = \mathbf{z})$, completing the proof.

*Corresponding author: Ou Wu.

A.2 Proof of Proposition 1 and extension

A.2.1 Proof of Proposition 1

The one-hot+RAPPOR baseline requires K perturbation events, while ECOC-RAPPOR uses M events. Since $M = O(\log K)$, the number of perturbations is reduced by $K/M \sim O(K/\log K)$. Each bit perturbation has variance $p_{\text{err}}(1 - p_{\text{err}})$, so the total noise variance is:

$$\text{One-hot:} \quad K \cdot p_{\text{err}}(1 - p_{\text{err}}) \tag{6}$$

$$\text{ECOC-RAPPOR:} \quad M \cdot p_{\text{err}}(1 - p_{\text{err}}) \tag{7}$$

The variance reduction ratio is $K/M \sim O(K/\log K)$.

A.2.2 Error Correction Analysis (Supplementary)

For RAPPOR with $\epsilon \in [0.5, 5]$, $p_{\text{err}} = \frac{1}{e^{\epsilon}+1} \in [0.12, 0.4]$. The single-bit to multi-bit error ratio is:

$$R(M, p_{\text{err}}) = \frac{M p_{\text{err}}(1 - p_{\text{err}})^{M-1}}{\sum_{t=2}^{M} \binom{M}{t} p_{\text{err}}^{t}(1 - p_{\text{err}})^{M-t}}. \tag{8}$$

Key insight: $R \geq 5$ only when $M \leq \frac{1-p_{\text{err}}}{5p_{\text{err}}}$. For typical ECOC parameters ($M = \Theta(\log K) \geq 5$), this condition fails (e.g., $M_{\text{crit}} \approx 1.47$ at $\epsilon = 0.5$), meaning multi-bit errors dominate. Despite this, error correction remains beneficial because It corrects the *most probable single error pattern* (single-bit flips). Error correction delivers **marginal gains** compared to dimensionality reduction (e.g., 30% error reduction vs. $100\times$ variance reduction when $K = 1000$).

A.3 Proof of Proposition 2 and Supplementary Properties

A.3.1 Proof of Proposition 2

Proof of Proposition 2 (Revised)

The estimation error decomposes as:

$$\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2 \leq \|\hat{\mathbf{f}} - \mathbb{E}[\hat{\mathbf{f}}]\|_2 + \|\mathbb{E}[\hat{\mathbf{f}}] - \mathbf{f}^*\|_2. \tag{9}$$

Bias term analysis: From the regularized solution, we have:

$$\mathbb{E}[\hat{\mathbf{f}}] = (\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T\mathbf{f}^*. \tag{10}$$

Therefore,

$$\mathbb{E}[\hat{\mathbf{f}}] - \mathbf{f}^* = -\lambda_{\text{reg}}(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\mathbf{f}^*. \tag{11}$$

Taking norms:

$$\|\mathbb{E}[\hat{\mathbf{f}}] - \mathbf{f}^*\|_2 \leq \lambda_{\text{reg}}\|(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\| \cdot \|\mathbf{f}^*\|_2. \tag{12}$$

The matrix $(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}$ has norm bounded by $1/(\sigma_{\min}^2 + \lambda_{\text{reg}})$, where $\sigma_{\min}$ is the smallest singular value of $\mathbf{C}_{\text{opt}}$. Under the ECOC design with $d_{\min} \geq 3$, we have $\sigma_{\min} = \Omega(M)$. Thus:

$$\|(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\| = O\left(\frac{1}{M^2}\right), \tag{13}$$

and:

$$\|\mathbb{E}[\hat{\mathbf{f}}] - \mathbf{f}^*\|_2 \leq \lambda_{\text{reg}} \cdot O\left(\frac{1}{M^2}\right) \cdot \|\mathbf{f}^*\|_2 = \mathcal{O}\left(\frac{\lambda_{\text{reg}}}{M^2} \cdot \|\mathbf{f}^*\|_2\right). \tag{14}$$

Variance term analysis: The variance term $\|\hat{\mathbf{f}} - \mathbb{E}[\hat{\mathbf{f}}]\|_2$ arises from the noise $\mathbf{w} = \mathbf{b} - \mathbb{E}[\mathbf{b}]$, where $\mathbf{b} = \frac{1}{p}(\hat{\pi} - q\mathbf{1})$. Since $\hat{\pi}$ is the average of N independent Bernoulli vectors:

$$\mathbb{E}\|\hat{\pi} - \mathbb{E}[\hat{\pi}]\|_2^2 = \sum_{m=1}^{M} \text{Var}(\hat{\pi}_m) \leq \frac{M}{4N}. \tag{15}$$

Using $p \approx \frac{\epsilon}{4M}$ for small ϵ/M:

$$\mathbb{E}\|\mathbf{w}\|_2^2 \leq \frac{M}{4Np^2} \approx \frac{4M^3}{N\epsilon^2}. \tag{16}$$

In regularized least squares, the estimator variance satisfies:

$$\mathbb{E}\|\hat{\mathbf{f}} - \mathbb{E}[\hat{\mathbf{f}}]\|_2^2 \leq \|(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\mathbf{C}_{\text{opt}}\|^2 \cdot \mathbb{E}\|\mathbf{w}\|_2^2. \tag{17}$$

Assuming $\sigma_{\min} = \Omega(M)$:

$$\|(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^T + \lambda_{\text{reg}} I)^{-1}\mathbf{C}_{\text{opt}}\| = O(1/M), \tag{18}$$

and thus:

$$\mathbb{E}\|\hat{\mathbf{f}} - \mathbb{E}[\hat{\mathbf{f}}]\|_2^2 = O\left(\frac{1}{M^2} \cdot \frac{M^3}{N\epsilon^2}\right) = O\left(\frac{M}{N\epsilon^2}\right). \tag{19}$$

Taking square roots:

$$\|\hat{\mathbf{f}} - \mathbb{E}[\hat{\mathbf{f}}]\|_2 = O\left(\sqrt{\frac{M}{N\epsilon^2}}\right). \tag{20}$$

Combining terms: Putting together the variance and bias terms:

$$\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2 \leq O\left(\sqrt{\frac{M}{N\epsilon^2}}\right) + \mathcal{O}\left(\frac{\lambda_{\text{reg}}}{M^2} \cdot \|\mathbf{f}^*\|_2\right). \tag{21}$$

A.3.2 Supplementary Properties

Unique Solution The QP in Eq. 5 has a unique global minimum because:

- Objective is strictly convex (Hessian $2(\mathbf{C}_{\text{opt}}\mathbf{C}_{\text{opt}}^\top + \lambda_{\text{reg}} I) \succ 0$)
- Feasible set (probability simplex) is convex
- $\mathbf{C}_{\text{opt}}$ has full column rank (by ECOC's "unique codeword per class" design)

A.4 Proof of Proposition 3 and Privacy Analysis

A.4.1 Proof of Noise Suppression Properties

Step 1: Error dynamics

Define $\mathbf{e}(t) = \mathbf{f}(t) - \mathbf{f}^*$. From the EMA update and noise model $\mathbf{f}_{\text{batch}}(t) = \mathbf{f}^* + \boldsymbol{\varepsilon}(t)$ (Proposition 1):

$$\mathbf{e}(t) = \alpha\mathbf{e}(t-1) + (1-\alpha)\boldsymbol{\varepsilon}(t), \tag{22}$$

where $\mathbb{E}[\boldsymbol{\varepsilon}(t)] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{\varepsilon}(t)\|_2^2] \leq \sigma^2 = \Theta(M/(n\epsilon^2))$.

Step 2: Time-averaged error bound

Taking expectation of $\|\mathbf{e}(t)\|_2^2$ and using $\mathbb{E}[\langle \mathbf{e}(t-1), \boldsymbol{\varepsilon}(t)\rangle] = 0$:

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}(t)\|_2^2] &= \alpha^2\mathbb{E}[\|\mathbf{e}(t-1)\|_2^2] + (1-\alpha)^2\mathbb{E}[\|\boldsymbol{\varepsilon}(t)\|_2^2] \\ &\leq \alpha^2\mathbb{E}[\|\mathbf{e}(t-1)\|_2^2] + (1-\alpha)^2\sigma^2. \end{aligned} \tag{23}$$

Solving this recurrence with $\mathbb{E}[\|\mathbf{e}(0)\|_2^2] \leq D^2$ ($D \leq \sqrt{2}$):

$$\begin{aligned}\mathbb{E}[\|\mathbf{e}(t)\|_2^2] &\leq \alpha^{2t}D^2 + (1-\alpha^2)\sigma^2\sum_{k=0}^{t-1}\alpha^{2k} \\ &= \alpha^{2t}D^2 + \frac{(1-\alpha^{2t})(1-\alpha^2)}{1+\alpha}\sigma^2. \end{aligned} \tag{24}$$

Time-averaging and taking $\limsup_{T\to\infty}$ yields:

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{e}(t)\|_2^2] \leq \frac{1-\alpha}{1+\alpha}\sigma^2. \tag{25}$$

A.4.2 Privacy Safety Analysis

By post-processing invariance of LDP [1], since $\mathbf{f}_{\text{batch}}$ is a deterministic function of privatized outputs $\{\mathbf{z}^{(i)}\}$ (satisfying ϵ-label LDP), and $\mathbf{f}(t)$ is a deterministic function of $\mathbf{f}_{\text{batch}}$, both quantities retain ϵ-label LDP with no additional privacy loss.

Proof: For any $\mathbf{y} \in \Delta^{K-1}$ and neighboring datasets $\mathcal{D} \sim \mathcal{D}'$:

$$\begin{aligned}\mathbb{P}[\mathbf{f}_{\text{batch}}(\mathcal{D}) = \mathbf{y}] &= \mathbb{P}[\mathcal{M}(\mathcal{D}) \in g^{-1}(\mathbf{y})] \\ &\leq e^{\epsilon}\mathbb{P}[\mathcal{M}(\mathcal{D}') \in g^{-1}(\mathbf{y})] \\ &= e^{\epsilon}\mathbb{P}[\mathbf{f}_{\text{batch}}(\mathcal{D}') = \mathbf{y}], \end{aligned} \tag{26}$$

where $\mathcal{M}$ is the ECOC-RAPPOR mechanism and g is the aggregation function. The same argument applies to $\mathbf{f}(t)$ as a function of $\mathbf{f}_{\text{batch}}$.

B. Experimental

B.1 Final Codeword Generation Algorithm and Constraint Validation

This section details the definitive algorithm used to generate ECOC codewords in all our experiments and provides a rigorous validation for its core design constraint. The goal is to create the most computationally efficient code (i.e., the shortest code length M) that meets our required level of noise resilience.

The Search-Based Generation Algorithm. Our method for selecting a codebook is a systematic search for the smallest integer M that can support at least K classes while satisfying our primary noise resilience constraint. The algorithm proceeds as follows:

1. **Initialize Search.** The search for the code length M commences at its theoretical minimum, $M_{start} = \lceil \log_2 K \rceil$. 2. **Iterative Search.** For each integer $M \geq M_{start}$, we perform an existence check to determine if a valid codebook can be constructed. 3. **Existence Check.** The check for a given (K, M) configuration against a target $d_{\min}$ is performed using a two-tiered approach: (a) First, we query for a standard construction (e.g., BCH codes via the `galois` library). (b) If none is found, we employ a heuristic search. 4. **Termination.** The algorithm terminates and returns the first value of M for which a valid codebook is found.

Application of the Algorithm. This algorithm, when run with our chosen constraint of $d_{\min} \geq 3$, yields the code lengths used in our paper: $M = 6$ for CIFAR-10, $M = 12$ for CIFAR-100, and $M = 15$ for ImageNet-LT.

Validation of the $d_{\min} \geq 3$ Constraint. The choice of the $d_{\min}$ constraint is critical as it represents a trade-off between noise resilience and code efficiency. A stricter constraint may improve error correction but requires a longer M, which increases computational cost and the total amount of LDP noise. To validate that $d_{\min} \geq 3$ is a well-justified choice, we conduct a sensitivity analysis on CIFAR-100-LT ($\epsilon = 2$), treating the $d_{\min}$ constraint as a hyperparameter. For each constraint, we run our generation algorithm to find the corresponding shortest M and evaluate the final model performance.

The results are presented in Table 1. Increasing the constraint from $d_{\min} = 3$ to $d_{\min} = 4$ necessitates a longer code ($M = 14$) but yields a marginal performance improvement. However, further increasing the constraint to $d_{\min} = 5$ requires a substantially longer code ($M = 18$), at which point performance begins to decline. This is likely due to the diminishing returns of error correction being outweighed by the negative effect of perturbing a larger number of bits. No feasible solution was found for $d_{\min} \geq 6$ within a reasonable search space.

Table 1. Sensitivity analysis of the $d_{\min}$ constraint on CIFAR-100-LT ($\epsilon = 2$). For each constraint, we report the shortest code length M found and the resulting model performance.

$d_{\min}$ Constraint	Resulting M	Feasible?	Macro-F1	Acc (Few)
≥ 3	12	✓	42.1	36.5
≥ 4	14	✓	**42.3**	**36.6**
≥ 5	18	✓	41.2	35.5
≥ 6	> 20	No	–	–

This analysis reveals that the optimal performance lies in the region of $d_{\min} = 3$ or $d_{\min} = 4$. Given that $d_{\min} = 3$ provides nearly identical performance to $d_{\min} = 4$ while using a significantly more compact and efficient code ($M = 12$ vs. $M = 14$), we conclude that it represents the most effective trade-off. Therefore, we adopt the $d_{\min} \geq 3$ constraint for all experiments, confident that it is a robust and empirically-supported choice.

B.2 Sensitivity Analysis on EMA Coefficient α

Our imbalanced learning module utilizes an Exponential Moving Average (EMA) to dynamically update class proportion estimates during training. The coefficient α controls the update speed, balancing stability (high α) and responsiveness to the model's evolving predictions (low α).

To analyze the impact of this hyperparameter, we evaluate the performance of ECOC-IL on CIFAR-100-LT ($\epsilon = 2$) across a range of α values from 0.5 to 0.99. The results are shown in Figure 1. The performance is suboptimal for lower values of α (e.g., 0.5 to 0.8), as the proportion estimates become too noisy and unstable, being overly influenced by individual batches. As α increases, the performance rises sharply and enters a stable, high-performance plateau in the range of $[0.9, 0.99]$. This indicates that our method is not sensitive to the precise choice of α within this range. The results validate our selection of $\alpha = 0.95$ for all main experiments, as it resides comfortably within this optimal region.

B.3 Rationale for a Unified Codeword Strategy under Varying Privacy Levels

A Principled Approach for Fair Comparison. Our theoretical analysis in Section 4 suggests that for stricter privacy regimes (e.g., $\epsilon \leq 0.5$), a more robust code with a higher minimum Hamming distance (e.g., $d_{\min} \geq 5$) could be employed to counteract the stronger noise. This would, in turn, necessitate a longer code length M than our standard $d_{\min} \geq 3$ constraint requires.

However, to ensure a fair and rigorous comparison across all privacy levels, we adopted a **unified approach** in our experiments. By using the exact same codebook generated by our standard algorithm (with the $d_{\min} \geq 3$ constraint)

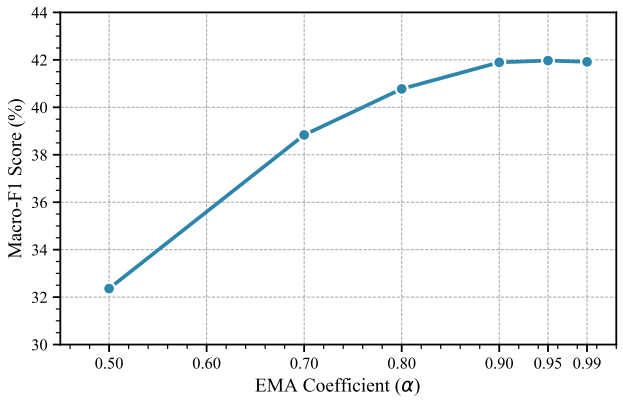

Figure 1. Sensitivity of ECOC-IL to the EMA coefficient α on CIFAR-100-LT ($\epsilon = 2$). The model achieves robust, high performance for $\alpha \in [0.9, 0.99]$.

for all values of ϵ, we eliminate potential confounding variables, such as changes in the model's output dimensionality, that would arise from switching codebooks. This principled choice ensures that observed performance changes are attributable solely to the variation in the privacy budget ϵ. Nevertheless, it raises a critical question: does this unified practice compromise performance at low ϵ values where theory suggests a more complex code?

Empirical Validation and Trade-off Analysis. To answer this, we conduct a direct comparison on CIFAR-100-LT at a strong privacy setting of $\epsilon = 1.0$ (the lowest in our main experiments). We compare the performance of our standard ECOC-IL (using the code with $M = 12$ derived from the $d_{\min} \geq 3$ constraint) against a "theory-compliant" variant (ECOC-IL-Robust) that uses a code with $M = 18$, derived from the stricter $d_{\min} \geq 5$ constraint.

The results, presented in Table 2, show that the variant with the more robust code ($M = 18$) achieves nearly identical performance to our standard, more efficient version ($M = 12$). This outcome reveals a crucial trade-off inherent in the label LDP setting. While the $d_{\min} \geq 5$ code offers superior error-correction capabilities, its required longer code length forces a finer distribution of the total privacy budget ϵ. Consequently, the per-bit privacy budget (ϵ/M) is significantly smaller, leading to a higher level of noise being introduced to each bit during the initial RAPPOR perturbation step. In our tested regime, the theoretical advantage of enhanced error correction is effectively nullified by the practical disadvantage of this increased initial noise.

Conclusion. This analysis confirms that our decision to use a single, unified code generation algorithm across all ϵ values is an empirically well-justified and principled choice. By effectively navigating the intrinsic trade-off between error-correction power and initial perturbation noise, our approach provides a consistent, efficient, and high-performing solution without the need for privacy-level-dependent hyperparameter tuning of the code structure.

Table 2. Performance comparison on CIFAR-100-LT ($\rho = 100, \epsilon = 1.0$) between our standard ECOC-IL and a variant using a theoretically more robust code. The performance is nearly identical, justifying our unified approach.

Method Variant	$d_{\min}$ Constraint	Resulting M	Macro-F1	Acc (Few)
ECOC-IL-Robust	≥ 5	18	**35.9**	32.2
ECOC-IL (Standard)	≥ 3	12	35.8	**32.3**

B.4 Full Experimental Results

This section provides the comprehensive results for all experiments, supplementing the summarized analysis in the main paper.

On the inclusion of the VA-IL baseline. To provide a more fine-grained ablation of our contributions, we introduce an additional strong baseline, **VA-IL**. This baseline is constructed by applying our full imbalanced learning (IL) module directly onto the original Vector Approximation (VA) framework. By comparing VA-IL to the standard VA, we can isolate the "pure" contribution of the IL module. Subsequently, by comparing our final ECOC-IL model to VA-IL, we can precisely quantify the additional benefits brought by our ECOC encoding scheme.

Organization of Results. To provide maximum clarity, the performance on long-tailed benchmarks is organized across three dedicated tables. Table 3 details the results for the severe imbalance setting ($\rho = 100$), while Table 4 presents the results for the moderate $\rho = 50$ setting. The results for the large-scale ImageNet-LT are provided separately in Table 5 to highlight scalability.

B.4.1 Full Results for Severe Imbalance Setting ($\rho = 100$)

This table presents the detailed experimental results for the severe imbalance setting ($\rho = 100$) on both CIFAR-10-LT and CIFAR-100-LT. The data covers all baseline methods across all tested privacy budgets, providing a comprehensive view of their performance.

The results in Table 3 demonstrate that our method, ECOC-IL, consistently outperforms all baseline methods across both datasets and all tested privacy budgets. The performance advantage is particularly substantial under stricter privacy constraints (e.g., $\epsilon = 1$). For instance, on CIFAR-100-LT, the Macro-F1 score of ECOC-IL is 9.8 points higher than the VA baseline at $\epsilon = 1$, while this gap is 7.2 points at $\epsilon = 8$. This trend highlights the superior noise resilience of our ECOC-based encoding in high-privacy (i.e.,

Table 3. Full experimental results on long-tailed datasets (CIFAR-10-LT and CIFAR-100-LT) for the severe imbalance setting ($\rho = 100$). The data covers all methods across all tested privacy budgets. All metrics are in percent (%).

Dataset	Method	ϵ	Macro-F1	Overall Acc	Many	Medium	Few
CIFAR-10-LT	RR	1	25.1	34.0	49.8	28.1	11.5
		2	31.5	40.2	55.1	33.8	15.2
		4	38.2	46.5	60.9	41.0	20.8
		8	43.0	51.1	65.2	46.5	25.1
	LP-2ST	1	38.5	48.1	63.5	42.0	23.4
		2	45.1	53.8	68.2	48.5	28.6
		4	51.0	59.2	72.8	55.1	34.9
		8	54.5	62.9	75.9	59.8	39.5
	ALIBI	1	46.8	56.0	70.1	52.5	33.1
		2	53.6	61.5	75.0	58.1	39.3
		4	59.1	66.8	79.2	64.0	45.5
		8	62.8	70.1	82.0	68.3	49.9
	VA	1	48.2	58.5	75.1	55.3	37.0
		2	55.4	64.1	79.2	61.3	43.2
		4	61.3	69.5	83.0	67.8	50.1
		8	65.0	72.8	85.5	71.9	54.8
	VA-IL	1	54.5	60.1	75.5	58.0	45.2
		2	60.8	65.9	79.6	64.2	50.5
		4	66.0	70.8	83.3	69.5	56.1
		8	69.1	74.0	85.8	73.5	60.2
	ECOC-IL	1	**58.8**	**63.2**	**76.5**	**62.0**	**50.4**
		2	**65.1**	**68.5**	**80.8**	**67.1**	**55.9**
		4	**70.5**	**73.0**	**84.1**	**72.5**	**61.2**
		8	**73.8**	**76.1**	**86.6**	**76.0**	**64.9**
CIFAR-100-LT	RR	1	8.0	20.2	29.5	18.1	6.4
		2	11.5	24.3	35.1	21.8	8.2
		4	16.8	28.1	40.2	25.5	10.1
		8	17.2	31.0	43.8	28.3	11.7
	LP-2ST	1	15.8	29.5	42.1	27.0	12.3
		2	20.1	33.8	48.2	30.5	14.6
		4	26.0	39.1	55.0	35.8	18.1
		8	28.8	42.5	59.1	39.0	20.4
	ALIBI	1	22.5	36.8	50.1	35.4	18.9
		2	28.6	41.5	55.0	40.1	21.3
		4	36.5	45.8	60.1	44.5	24.2
		8	37.5	49.2	64.0	48.1	27.0
	VA	1	26.0	43.5	59.5	41.2	22.1
		2	32.9	48.1	63.2	45.3	25.2
		4	38.5	52.3	67.8	49.9	29.1
		8	42.0	55.4	71.0	53.1	32.5
	VA-IL	1	31.8	44.2	**59.9**	44.1	28.3
		2	38.8	49.0	**63.5**	48.8	32.1
		4	44.1	53.1	**68.0**	53.2	36.5
		8	47.5	56.7	**71.2**	56.4	39.7
	ECOC-IL	1	**35.8**	**45.1**	59.8	**48.0**	**32.3**
		2	42.1	**49.5**	62.8	**51.1**	**36.5**
		4	**46.5**	**53.8**	67.5	**55.0**	**40.1**
		8	**49.2**	**56.9**	70.8	**58.2**	**43.3**

high-noise) regimes. This superiority is primarily driven by our framework's ability to protect minority class signals, as evidenced by the consistently higher scores in the Few-shot accuracy column. These findings validate the robustness and effectiveness of our proposed framework in challenging, privacy-constrained, and severely imbalanced learning scenarios.

Table 4. Full experimental results on long-tailed datasets (CIFAR-10-LT and CIFAR-100-LT) for the moderate imbalance setting ($\rho = 50$). All metrics are in percent (%).

Dataset	Method	ϵ	Macro-F1	Overall Acc	Many	Medium	Few
CIFAR-10-LT	RR	1	35.1	44.0	56.8	38.5	21.3
		2	40.2	49.5	61.3	44.1	25.8
		4	46.0	55.1	66.0	50.3	31.9
		8	50.5	59.2	69.8	55.4	36.8
	LP-2ST	1	50.2	57.8	68.5	54.0	36.5
		2	55.8	62.2	73.0	59.5	41.6
		4	61.1	67.0	77.1	65.2	48.0
		8	64.9	70.5	80.2	69.3	52.8
	ALIBI	1	59.0	66.5	76.2	64.8	48.1
		2	64.3	70.9	80.1	69.3	53.4
		4	69.0	75.1	83.8	73.9	59.0
		8	72.1	78.0	86.1	77.2	62.9
	VA	1	61.5	68.9	79.8	67.5	51.3
		2	66.8	73.1	82.5	71.8	56.0
		4	71.2	77.0	85.9	76.1	61.5
		8	74.5	79.9	88.0	79.3	65.4
	VA-IL	1	65.2	70.5	80.1	69.8	56.2
		2	70.1	74.9	**82.8**	74.0	61.3
		4	74.0	78.2	86.1	78.0	65.8
		8	77.1	81.0	88.2	81.1	69.0
	ECOC-IL	1	**69.8**	**73.0**	**80.5**	**72.1**	**62.5**
		2	**74.9**	**77.5**	82.1	**76.2**	**65.1**
		4	**78.5**	**81.0**	**87.2**	**80.1**	**69.3**
		8	**81.2**	**83.5**	**89.5**	**82.9**	**72.0**
CIFAR-100-LT	RR	1	13.1	27.0	37.5	25.0	11.4
		2	16.2	31.5	42.3	29.1	14.5
		4	20.0	36.1	47.2	33.8	18.3
		8	22.8	39.5	51.0	37.1	21.0
	LP-2ST	1	22.0	36.5	49.2	34.0	17.5
		2	26.8	41.2	54.0	38.5	21.1
		4	32.1	46.5	59.1	44.0	26.0
		8	35.9	50.1	62.9	47.9	29.8
	ALIBI	1	30.1	44.0	56.5	42.8	25.1
		2	35.1	48.9	61.1	47.3	28.4
		4	40.0	53.5	65.8	52.0	32.9
		8	43.8	56.9	69.1	55.8	36.5
	VA	1	33.5	50.1	63.8	48.0	29.2
		2	38.5	54.6	67.5	51.8	34.0
		4	43.1	58.8	71.9	56.5	39.0
		8	46.9	62.0	75.0	60.1	42.8
	VA-IL	1	37.8	52.0	**64.1**	50.5	35.1
		2	42.5	56.0	**67.8**	54.9	40.0
		4	47.0	60.1	**72.0**	58.9	44.5
		8	50.1	63.0	**75.1**	62.0	47.7
	ECOC-IL	1	**41.8**	**52.1**	64.0	**52.3**	**39.5**
		2	**47.0**	**56.1**	67.1	**57.2**	**44.0**
		4	**51.5**	**60.3**	71.0	**61.5**	**48.8**
		8	**55.0**	**63.5**	74.2	**64.9**	**52.5**

B.4.2 Full Results for Moderate Imbalance Setting ($\rho = 50$)

This table presents the detailed experimental results for the moderate imbalance setting ($\rho = 50$) on both CIFAR-10-LT and CIFAR-100-LT. As expected, the performance of all methods improves compared to the more severe $\rho = 100$ setting, as the learning task becomes inherently easier with more samples available for minority classes.

Despite the simplified setting, the results in Table 4 confirm that ECOC-IL maintains a consistent and significant

Table 5. Full experimental results on the large-scale ImageNet-LT dataset. The data covers all methods across all tested privacy budgets. All metrics are in percent (%).

Dataset	Method	ϵ	Macro-F1	Overall Acc	Many	Medium	Few
ImageNet-LT	RR	1	3.9	14.1	22.0	11.5	2.2
		2	5.8	16.5	25.4	13.1	3.0
		4	7.9	19.1	29.0	15.8	4.2
		8	9.8	21.5	32.1	18.0	5.5
	LP-2ST	1	8.5	21.0	32.5	17.1	5.3
		2	11.2	24.1	36.8	20.4	7.1
		4	14.3	27.8	41.0	24.1	9.5
		8	16.9	30.9	44.5	27.3	11.8
	ALIBI	1	12.8	26.2	38.0	23.9	8.8
		2	16.5	29.8	42.1	27.5	11.4
		4	20.1	33.9	46.8	31.8	14.5
		8	23.0	37.1	50.5	35.0	17.1
	VA	1	13.5	31.8	48.1	28.5	10.1
		2	20.3	36.2	51.5	33.1	14.8
		4	26.1	40.5	55.8	38.0	18.9
		8	30.5	44.0	59.1	42.1	22.4
	VA-IL	1	18.2	33.5	**48.4**	31.0	14.1
		2	25.1	37.8	**51.8**	36.3	18.2
		4	30.2	41.9	**56.0**	40.5	22.0
		8	34.0	45.1	**59.3**	44.2	25.5
	ECOC-IL	1	**22.1**	**34.0**	48.2	**34.5**	**18.5**
		2	**28.7**	**38.1**	51.1	**39.2**	**24.1**
		4	**33.5**	**42.0**	55.0	**43.8**	**28.0**
		8	**37.1**	**45.2**	58.2	**47.1**	**31.5**

performance lead. On CIFAR-100-LT at $\epsilon = 2$, for example, ECOC-IL achieves a Macro-F1 score of 47.0%, which is 8.5 points higher than the VA baseline. While the absolute performance of all methods is higher, the sustained advantage of our method demonstrates its robust superiority, which is not limited to only the most extreme imbalance scenarios. The consistent lead across all privacy budgets and on both datasets reinforces the conclusions drawn in the main paper.

B.4.3 Full Results for Large-Scale ImageNet-LT

This table provides the detailed experimental results on the large-scale, long-tailed ImageNet-LT dataset. This benchmark serves as a crucial test for the scalability and effectiveness of the compared methods in a real-world, high-complexity scenario with 1000 classes.

The results in Table 5 confirm that the advantages of ECOC-IL, demonstrated on the CIFAR-based datasets, scale effectively to this much more challenging task. ECOC-IL consistently achieves the highest performance across all metrics and privacy budgets. The Macro-F1 score, which is particularly indicative on such a highly imbalanced dataset, shows a substantial lead. At the strictest privacy setting of $\epsilon = 1$, ECOC-IL outperforms the VA baseline by 8.6 percentage points. This significant gap underscores the critical importance of our method's noise resilience and imbalanced learning capabilities when both the number of classes and the degree of imbalance are large. As with other datasets, the primary driver of this improvement is

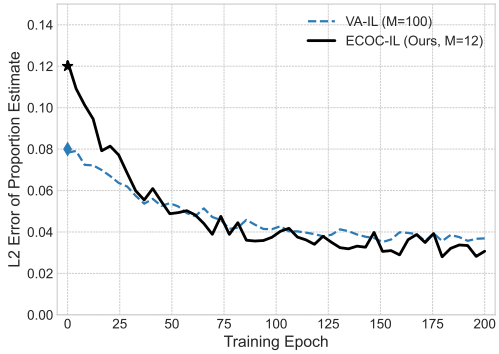

Figure 2. L2 error of the dynamic proportion estimate $f^{(t)}$ over training epochs on CIFAR-100-LT ($\epsilon = 2$). Despite starting with a higher initial error, ECOC-IL's dynamic update process converges to a highly accurate estimate, matching or surpassing the VA-IL baseline.

the significantly better performance on minority (Few-shot) classes, validating ECOC-IL as a robust and scalable solution for practical, large-scale applications under label privacy constraints.

B.5 Analysis of Training Dynamics

A central claim of our paper is that the final classification superiority of ECOC-IL is not due to a more accurate initial static proportion estimate, but rather stems from the improved trainability and noise resilience conferred by the structured ECOC representation. This section provides direct empirical evidence for this claim by visualizing the training dynamics. We analyze two key aspects: the convergence of the dynamic proportion estimates and the learning progress on challenging few-shot classes.

B.5.1 Convergence of Dynamic Proportion Estimates

Our imbalanced learning module relies on dynamic EMA updates to refine the initial static proportion estimate, $f^{(0)}$. To verify the effectiveness of this process, we track the L2 error ($||f^{(t)} - f^*||_2$) between the estimated proportion vector at epoch t, $f^{(t)}$, and the ground-truth vector f^*.

Figure 2 plots this error curve for ECOC-IL and the VA-IL baseline on CIFAR-100-LT ($\rho = 100, \epsilon = 2$). As established in the main paper, ECOC-IL's initial static estimate ($t = 0$) has a higher error than VA-IL's due to information compression. However, as training progresses, the EMA update allows ECOC-IL to effectively leverage the model's increasingly accurate predictions. The error curve for ECOC-IL descends rapidly and, crucially, converges to a level comparable to, or even slightly better than, the VA-IL baseline in the later stages of training. This demonstrates that our dynamic update mechanism successfully compensates for the initial estimation loss, providing the model with a highly accurate proportion estimate during the criti-

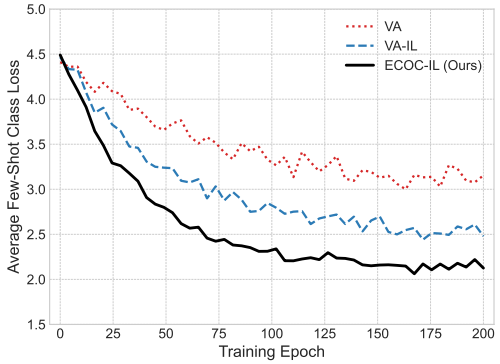

Figure 3. Average training loss on few-shot classes over epochs on CIFAR-100-LT ($\epsilon = 2$). ECOC-IL's loss curve is consistently lower and more stable, demonstrating its superior ability to learn from minority classes.

cal final training phases.

B.5.2 Learning Dynamics of Few-Shot Classes

To directly visualize the impact of our framework on the most challenging classes, we plot the average training loss on the set of few-shot classes over epochs. A lower, more stable loss curve indicates more effective learning for these minority classes.

Figure 3 compares the few-shot training loss for three key methods on CIFAR-100-LT ($\rho = 100, \epsilon = 2$): the standard VA, the VA-IL baseline, and our full ECOC-IL model. The standard VA model struggles significantly, with its loss curve remaining high and noisy, indicating difficulty in learning from the scarce and noisy signals. The VA-IL baseline shows a clear improvement, as the IL module forces the model to pay more attention to these classes.

Most importantly, the ECOC-IL curve is substantially lower and more stable than both baselines throughout the entire training process. It descends faster and converges to the lowest final loss value. This provides direct, compelling evidence for our central claim: the structured, noise-resilient ECOC representation provides a superior inductive bias and a more stable learning signal, enabling the model to learn minority classes far more effectively than is possible with sparse, high-dimensional one-hot representations.

References

[1] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014. 2