Reframing Long-Tailed Learning via Loss Landscape Geometry

Supplementary Material

7. More Experiment Protocols

7.1. Implementation Details

Our implementation follows [75]. For both CIFAR-10-LT and CIFAR-100-LT, we adopt ResNet-32 as the backbone. Our model is trained for 200 epochs with a batch size of 256 based on an SGD optimizer. The momentum is set to 0.9 and the weight decay is 5×10^{-4}. The learning rate warms up to 0.15 in the first 5 epochs and decays by 0.1 at the 160 and 180 epochs.

We adopt the ResNet-50 [13] architecture as the model backbone for both ImageNet-LT and iNaturalist 2018. The model is optimized based on SGD with a fixed momentum of 0.9 and a batch size of 256. For ImageNet-LT, we train the model for 90 epochs using an initial learning rate of 0.1 and a weight decay of 5×10^{-4}. For iNaturalist 2018, the training is extended to 100 epochs, with an initial learning rate of 0.2 and a weight decay of 1×10^{-4}. A cosine scheduler is employed for learning rate adjustments across all experiments.

7.2. Definition of Feature Quality

As stated in the main paper (Eq.1), the memory bank $\mathcal{M}$ is populated by storing the encoder parameters θ_{enc}^{c} that yield the highest feature quality Q. We provide the mathematical formulation for Q, which is based on the SCoRe framework [38].

First, the inter-class separation for class c is computed as the minimum distance to any other class centroid:

$$\text{Dis}(\theta_{enc}, c) = \min_{c' \neq c} ||\mu_c - \mu_{c'}||_2, \tag{17}$$

Second, the intra-class variance is computed as:

$$\text{Var}(\theta_{enc}, c) = \frac{1}{|\mathcal{A}_c|} \sum_{a \in \mathcal{A}_c} ||a - \mu_c||_2^2, \tag{18}$$

where $\mathcal{A}_c$ is the set of feature vectors a (where $a = f(\theta_{enc}, x)$, Sec. 4.1) for class c, and μ_c is the feature centroid. $\mathcal{A}_c$ and μ_c are computed using the encoder state θ_{enc}.

The final quality score $Q(\theta_{enc}, c)$ is then calculated as:

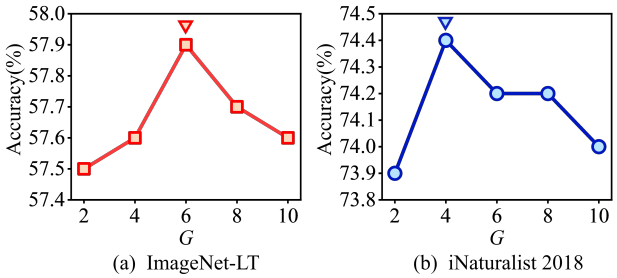
$$Q(\theta_{enc}, c) = \text{Dis}(\theta_{enc}, c) - \beta \log(\text{Var}(\theta_{enc}, c)), \tag{19}$$

where β is a hyperparameter balancing the two components, which is set to 0.5 in our experiments.

7.3. Definition of Adaptive parameter

The adaptive parameter scheduled α (Eq. 16) at epoch t is updated according to the cosine annealing schedule:

$$\alpha = \alpha_{end} + \frac{1}{2}(\alpha_{start} - \alpha_{end}) \left(1 + \cos\left(\frac{t\pi}{T}\right)\right), \tag{20}$$



(a) ImageNet-LT (b) iNaturalist 2018

Figure 7. Ablation study about group numbers on ImageNet-LT and iNaturalist 2018.

where $\alpha_{start} = 0.95$ and $\alpha_{end} = 0.6$ denote the initial and final values respectively, and T represents the total number of training epochs.

8. More Experiment Results

8.1. Ablation Study About Group Numbers

We conduct a detailed ablation study on the impact of the group number G on the large-scale datasets, ImageNet-LT and iNaturalist 2018.We report the Top-1 accuracy for various G values in Fig. 7. The performance peaks at $G = 4$ for iNaturalist 2018, whereas ImageNet-LT achieves its best results at $G = 6$. Similar to our other experiments, employing an excessive number of groups does not yield further gains.

Table 6. Hyperparameter ablation analysis for the knowledge preservation strength λ on CIFAR100-LT ($r = 100$).

Hyperparameter (λ)	**Overall Acc.**
10	52.6
50	52.7
100	**53.2**
500	52.4
1000	52.1

8.2. Ablation Study About Knowledge Preservation Strength

We analyze the model's sensitivity to the knowledge preservation strength, controlled by the hyperparameter λ. We tested a range of λ values ($\lambda \in \{10, 50, 100, 500, 1000\}$) on CIFAR100-LT ($r = 100$). As shown in Table 6, a small λ (*e.g.*, 0.1) is insufficient to prevent catastrophic forgetting, resulting in low overall accuracy. Conversely, a large λ (*e.g.*, 1000) severely hinders the acquisition of new knowledge for

the current group. Our model achieves the optimal balance at $\lambda = 100$, resulting in an accuracy of 53.2%.

8.3. Ablation Study About Perturbation Scale

To further validate our derived characteristic grouped radius, we empirically study the influence of varying the perturbation scale.

To investigate the optimal perturbation scale, we multiply Eq. 13 by a scaling factor $Z = \rho_g / \rho_g^*$, varying the norm from 10^{-1} to 10^{-7}. As shown in Figure 1, the performance of GSA peaks when Z is set to 10^{-2}, confirming the existence of an optimal scale. Nevertheless, the coefficient for the characteristic grouped radius ρ_g^* (Eq. 13) is not a precise value and requires empirical tuning (as further elaborated in the Remarks of Sec. 9.2).

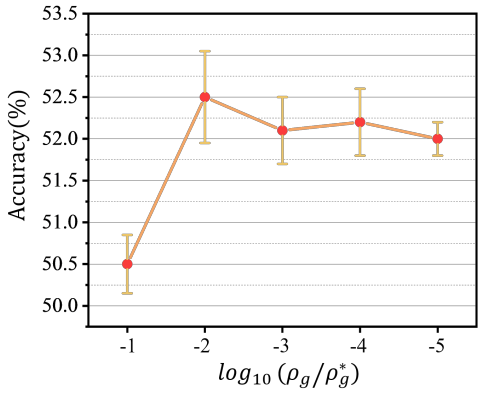


Figure 8. The impact of the perturbation scale.

8.4. Analysis About Loss Landscape

This section presents additional results of the spectral density of hessian for ResNet models trained with CE (the baseline method with the naive Cross Entropy loss), CE+SAM [10] and CE+Ours. We analyze models trained on CIFAR-10 LT datasets using VS and CE loss functions.

Experimental results demonstrate that, compared to SAM, our method achieves lower values for both the largest eigenvalue ($\lambda_{\max}$) and the trace ($\mathrm{tr}(ht)$) of the Hessian matrix across both head and tail classes. This indicates that our approach successfully yields a flatter loss surface.

8.5. Analysis About Training Time Cost

The wall-clock training time of our method ($G = 4$) was assessed against the CE and BCL baselines (Table 7). Our framework required 1.67 hours to train, resulting in an additional 0.62 hours compared to the BCL baseline.

Table 7. Analysis about training time cost.

CE	BCL	Ours
0.52h	1.05h	1.67h

9. Theoretical Proofs

9.1. Theoretical Proofs About Convergence

We present the convergence analysis for our framework (Eq. 16), proving the algorithm converges under standard non-convex stochastic optimization assumptions.

Our proof builds upon the convergence analysis of F-SAM [27], which established convergence for a min-max sharpness objective. We demonstrate that the inclusion of our Grouped Knowledge Preservation (GKP) module, which acts as a convex regularizer, preserves and stabilizes this convergence.

9.1.1. Objective Function and Update Rule

First, We define the complete objective function $\mathcal{L}(\theta)$ as:

$$\mathcal{L}(\theta) = \sum_{g=1}^{G} \left[\alpha \left(\mathcal{L}_{gsa}^{g}(\theta) \right) + (1-\alpha) \mathcal{L}_{gkp}^{g}(\theta) \right], \tag{21}$$

where G is the total number of groups, $\mathcal{L}_{gsa}^{g}(\theta)$ is the loss for group g, and $\mathcal{L}_{gkp}^{g}(\theta)$ is the GKP loss (Eq.6).

Second, we define our framework gradient g_t'. This is the stochastic gradient of our full objective $\mathcal{L}$ (Eq.16), computed on a mini-batch $\mathcal{S}$ at step t:

$$g_t' = \sum_{g=1}^{G} \left[\alpha \nabla_\theta \mathcal{L}_{\mathcal{S},gsa}^{g}(\theta_t) + (1-\alpha) \nabla_\theta \mathcal{L}_{\mathcal{S},gkp}^{g}(\theta_t) \right], \tag{22}$$

expanding $\mathcal{L}_{\mathcal{S},gsa}^{g}$ using its definition from Eq.15:

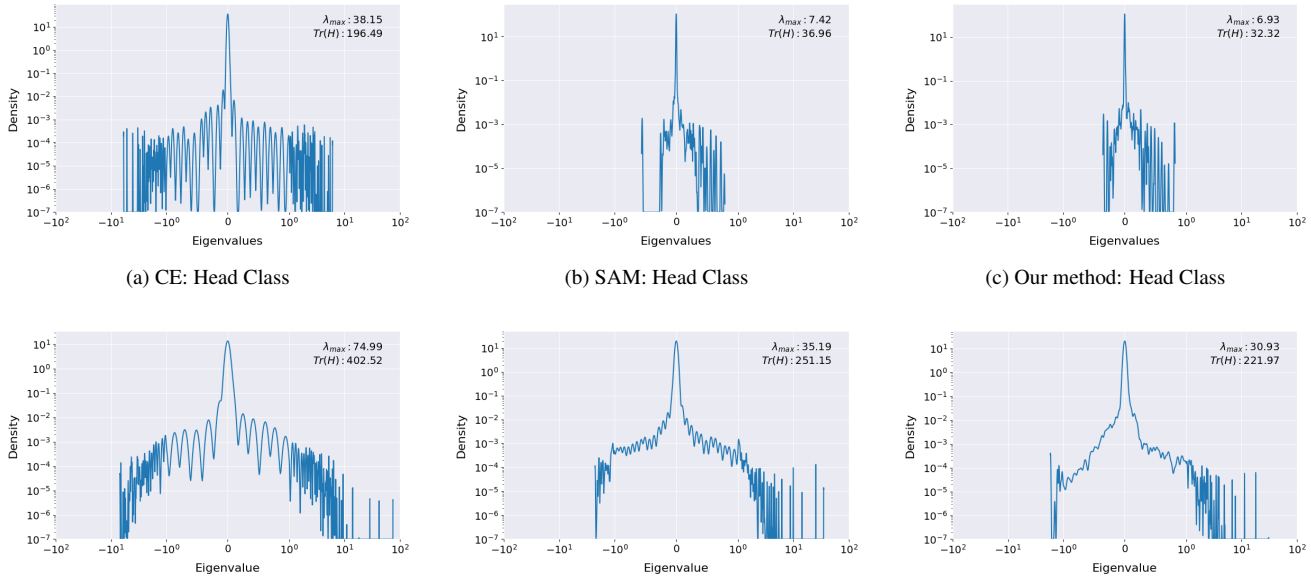
$$g_t' = \sum_{g=1}^{G} \Bigg[\alpha \left(\nabla_\theta \mathcal{L}_{\mathcal{S},gsa}^{g}(\theta_t + \hat{\epsilon}_g^*) + \nabla_\theta \mathcal{L}_{\mathcal{S},reg}^{g}(\theta_t) \right) + (1-\alpha) \nabla_\theta \mathcal{L}_{\mathcal{S},gkp}^{g}(\theta_t) \Bigg], \tag{23}$$

with the update rule is $\theta_{t+1} = \theta_t - \eta_t g_t'$.

9.1.2. Core Assumptions

We adopt standard assumptions from stochastic non-convex optimization [27]:

- β-Smoothness: The objective $\mathcal{L}(\theta)$ is assumed to be β-smooth.
- Bounded Gradients: The stochastic gradients of all components are assumed to be bounded by positive constants M:
  - $\mathbb{E}[\|\nabla \mathcal{L}_{\mathcal{S},gsa}^{g}(\theta)\|_2^2] \leq M_{gsa}^2$



(a) CE: Head Class (b) SAM: Head Class (c) Our method: Head Class

(d) CE: Tail Class (e) SAM: Tail Class (f) Our method: Tail Class

Figure 9. Eigenvalue density distributions for head and tail layers, arranged by CE (the baseline method with the naive Cross Entropy loss), SAM [10], and our method.

  - $\|\nabla_\theta \mathcal{L}^g_{\mathcal{S},reg}(\theta)\|_2^2 \le M^2_{reg}$
  - $\mathbb{E}[\|\nabla_\theta \mathcal{L}^g_{\mathcal{S},gkp}(\theta)\|_2^2] \le M^2_{gkp}$
- Bounded Perturbation: The GSA perturbation vector $\hat{\epsilon}^*_g$ (Eq.14) is bounded by the radius ρ_t: $\|\hat{\epsilon}^*_g\| \le \rho_t$ for all $g \in \{1, ..., G\}$.

9.1.3. Convergence Derivation

We begin with the standard β-smoothness inequality for the objective $\mathcal{L}(\theta)$, taking expectation $\mathbb{E}_t$ conditioned on θ_t:

$$\mathbb{E}_t[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) - \eta_t \mathbb{E}_t[\langle \nabla \mathcal{L}(\theta_t), g'_t \rangle] + \frac{\beta \eta_t^2}{2} \mathbb{E}_t[\|g'_t\|^2], \tag{24}$$

by the bounded gradients assumption and the triangle inequality, we bound the final term:

$$\begin{aligned} \mathbb{E}_t[\|g'_t\|^2] &\le 3G \cdot (\alpha^2 M^2_{gsa} + \alpha^2 M^2_{reg} + (1-\alpha)^2 M^2_{gkp}) \\ &\triangleq M^2_{total}. \end{aligned} \tag{25}$$

Next, we bound the crucial inner product term. We relate g'_t to the gradient $\nabla \mathcal{L}(\theta_t)$ by defining the non-perturbed stochastic gradient $g_t(\theta_t)$ (for which $\mathbb{E}_t[g_t(\theta_t)] = \nabla L(\theta_t)$). This gives:

$$\begin{aligned} g'_t &= g_t(\theta_t) + \alpha \sum_{g=1}^{G} \left(\nabla_\theta \mathcal{L}^g_{\mathcal{S},gsa}(\theta_t + \hat{\epsilon}^*_g) - \nabla_\theta \mathcal{L}^g_{\mathcal{S},gsa}(\theta_t) \right) \\ &\triangleq g_t(\theta_t) + \alpha \sum_{g=1}^{G} \Delta^g_t, \end{aligned} \tag{26}$$

then we yield:

$$\mathbb{E}_t[\langle \nabla \mathcal{L}(\theta_t), g'_t \rangle] = \|\nabla \mathcal{L}(\theta_t)\|^2 + \alpha \cdot \mathbb{E}_t[\langle \nabla \mathcal{L}(\theta_t), \sum_{g=1}^{G} \Delta^g_t \rangle]. \tag{27}$$

Additionally, we bound the magnitude of the second term, $\mathcal{C}_t \triangleq \mathbb{E}_t[\langle \nabla \mathcal{L}(\theta_t), \sum_{g=1}^{G} \Delta^g_t \rangle]$, using β-smoothness, Bounded Perturbation, and Young's inequality ($ab \le a^2/2 + b^2/2$):

$$|\mathcal{C}_t| \le \mathbb{E}_t[\|\nabla \mathcal{L}(\theta_t)\| \| \sum_{g=1}^{G} \Delta^g_t \|] \le \mathbb{E}_t[\|\nabla \mathcal{L}(\theta_t)\| (G \beta_s \rho_t)], \tag{28}$$

$$|\mathcal{C}_t| \le \frac{\|\nabla \mathcal{L}(\theta_t)\|^2}{2} + \frac{G^2 \beta_s^2 \rho_t^2}{2}, \tag{29}$$

Where ρ_t is the perturbation radius for the GSA optimizer at iteration t to find a flat minimum and β_s is the smoothness parameter of the loss $\mathcal{L}(\theta)$, which bounds the Lipschitz constant of its gradient.

This bound (Eq. 29), combined with the standard β-smoothness inequality, gives us:

$$\mathbb{E}_t[\langle \nabla \mathcal{L}(\theta_t), g'_t \rangle] \ge \left(1 - \frac{\alpha}{2}\right) \|\nabla \mathcal{L}(\theta_t)\|^2 - \frac{\alpha G^2 \beta_s^2 \rho_t^2}{2}. \tag{30}$$

.

Now, we combine all bounds into the main smoothness inequality. Let $M_1 = (1 - \frac{\alpha}{2})$, $M_2 = \frac{\alpha G^2 \beta_s^2}{2}$, and $M_3 =$

$\frac{\beta M_{total}^2}{2}$:

$$\mathbb{E}_t[\mathcal{L}(\theta_{t+1})] \leq \mathcal{L}(\theta_t) - M_1\eta_t\|\nabla\mathcal{L}(\theta_t)\|^2 + M_2\eta_t\rho_t^2 + M_3\eta_t^2. \tag{31}$$

Considering the GKP component $\mathcal{L}_{gkp}^g$ (Eq. 6) is a form of Elastic Weight Consolidation (EWC), which acts as a σ-strongly convex regularizer (assuming the Fisher Information Matrix F is positive definite). The inclusion of this term, weighted by $(1-\alpha)$, ensures that our overall objective $\mathcal{L}(\theta)$ satisfies the Polyak-Lojasiewicz (P-L) inequality. This condition states that the gradient norm bounds the sub-optimality:

$$|\nabla\mathcal{L}(\theta_t)|^2 \geq 2\sigma(\mathcal{L}(\theta_t) - \mathcal{L}^*), \tag{32}$$

where $\sigma > 0$ is the P-L constant and $\mathcal{L}^*$ is the optimal loss value. Then, we substitute the P-L condition (Eq. 32) into our main inequality (Eq. 31):

$$\begin{aligned}\mathbb{E}_t[\mathcal{L}(\theta_{t+1})] \leq \mathcal{L}(\theta_t) &- \eta_t M_1 \left(2\sigma(\mathcal{L}(\theta_t) - \mathcal{L}^*)\right) \\ &+ \eta_t M_2\rho_t^2 + \eta_t^2 M_3.\end{aligned} \tag{33}$$

Subtracting $\mathcal{L}^*$ from both sides and taking the full expectation $\mathbb{E}$ over the history:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}^*] \leq &\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}^*] - 2\sigma\eta_t M_1\mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}^*] \\ &+ \eta_t M_2\rho_t^2 + \eta_t^2 M_3.\end{aligned} \tag{34}$$

Let $E_{noise} = (\eta M_2\rho^2 + \eta^2 M_3)$ be the constant error floor introduced by stochastic noise and the GSA perturbation; we unroll this recursion Eq. 34 from $t = 0$ to T:

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_T) - \mathcal{L}^*] \leq &(1 - 2\sigma\eta M_1)^T\mathbb{E}[\mathcal{L}(\theta_0) - \mathcal{L}^*] \\ &+ \sum_{i=0}^{T-1}(1 - 2\sigma\eta M_1)^i E_{noise},\end{aligned} \tag{35}$$

as $T \to \infty$, $\mathbb{E}[\mathcal{L}(\theta_T) - \mathcal{L}^*]$ (Eq. 35) converges to:

$$\mathbb{E}[\mathcal{L}(\theta_T) - \mathcal{L}^*] \to \frac{E_{noise}}{2\sigma\eta M_1} = \frac{\eta M_2\rho^2 + \eta^2 M_3}{2\sigma\eta M_1} = \mathcal{O}(\rho^2 + \eta). \tag{36}$$

This proves that our algorithm converges linearly to a neighborhood of the optimum. If we further employ a decaying schedule where $\eta_t \to 0$ and $\rho_t \to 0$, the error floor E_{noise} vanishes and $\mathbb{E}[\mathcal{L}(\theta_T) - \mathcal{L}^*] \to 0$. This analysis explicitly highlights the crucial role of our GKP module (Eq. 6 in the main paper). By enforcing the P-L condition (strong convexity), the GKP term fundamentally ensures a more stable and efficient linear convergence rate, which is significantly faster than the $\mathcal{O}(1/\sqrt{T})$ rate of general non-convex optimization [26, 27, 74].

9.2. Theoretical Proofs About Group-Specific Radius

Our GSA (Eq.15) derives from the PAC-Bayesian bound (Eq.7), which assumes i.i.d. training and test distributions ($p_s(x, y) = p_t(x, y)$). However, LT inherently violates this due to divergent label distributions ($p_s(y) \neq p_t(y)$). Consequently, the global i.i.d.-based Eq.7 is theoretically invalid in the LT setting.

To construct a theoretically sound generalization bound, we must rely on a weaker, yet more reasonable assumption. We follow the standard practice in LT and adopt the conditional i.i.d. assumption as our axiom: $p_s(x|y) = p_t(x|y), \forall y \in \{1, ..., C\}$.

The CC-SAM [74] was the first to apply this axiom to the PAC-Bayesian framework. They mathematically demonstrated that because the i.i.d. assumption only holds at the class level, the global generalization bound (Eq.7) must be decomposed into C independent class-conditional bounds. By minimizing these C individual bounds, CC-SAM proved that the optimal perturbation radius ρ_c^* must be class-specific and is an explicit function of the class sample size n_c, specifically $\rho_c^* \propto (n_c - 1)^{-1/4}$. While, this class-conditional approach is computationally infeasible for datasets with a large number of classes C, as it requires computing C separate gradients and perturbations in each step.

Our framework builds directly upon this insight. We propose to aggregate the C classes into $G \ll C$ groups using our GKP module (Section 4.2.1). Specifically, the classes are clustered based on parameter space similarity (θ_{enc}^c), serving as a valid and efficient proxy for groups that also share a coherent underlying data distribution. This allows us to posit a group-conditional i.i.d. assumption: $p_s(x|g) \approx p_t(x|g), \forall g \in \{1, ..., G\}$.

By applying the same mathematical derivation established by CC-SAM [74] to our G groups instead of their C classes, we can directly adopt their conclusion. This means we replace the class-specific sample count n_c with our group-specific sample count $n_g = |\mathcal{G}^g|$. This directly yields our Eq.13, which defines a group-specific optimal radius ρ_g^* that is both theoretically justified by the conditional i.i.d. principle and computationally tractable:

$$\rho_g^* \approx \left(\frac{||\theta||_2}{2||\tilde{\nabla}_\theta\mathcal{L}_{\mathcal{D}_g}(\theta)||_2}\right)^{\frac{1}{2}} d^{-\frac{1}{2}}(|\mathcal{G}^g| - 1)^{-\frac{1}{4}} \tag{37}$$

This demonstrates that our GSA module is a principled and efficient extension of the theoretical groundwork laid by CC-SAM.

Remarks: The approximations made during this derivation (*e.g.*, first-order Taylor expansion and omitted $\mathcal{O}(1)$ terms) indicate that the radius ρ_g^* (Eq.13) is not an exact number. Thus, it should be empirically tuned to realize the full potential of our GSA module.