Taming the Long Tail: Rebalancing Adversarial Training via Adaptive Perturbation

Appendix

A. Proofs of Sec. 3 (Preliminaries and problem analysis)

A.1. Useful lemmas

Lemma A.1. *Considering two arbitrary distributions Q_1 and Q_2 over instance space $\mathcal{S}$, for any hypothesis $h \in \mathcal{H}$ and loss function $\ell(\cdot,\cdot)$, the following bound holds:*

$$|\mathcal{R}_{nat}(h,Q_1) - \mathcal{R}_{nat}(h,Q_2)| \leq c_1\{\sqrt{\rho^2+1}\,\mathcal{W}_{c_2+q}(Q_1,Q_2)\}^{\frac{c_2+q}{q}}$$

Lemma A.2. *For an arbitrary iteration $t \in [T]$ in the adversarial training process and an arbitrary distributions P over instance space $\mathcal{S}$, the following bound holds:*

$$\mathcal{R}_{rob}(h^{(T)},P) - \mathcal{R}_{nat}(h^{(t)},P_{adv}^{h^{(t-1)}}) \leq \sum_{t'=t}^{T} c_1\{\sqrt{\rho^2+1}\,\mathcal{W}_{c_2+q}(P_{adv}^{h^{(t')}},P_{adv}^{h^{(t'-1)}})\}^{\frac{c_2+q}{q}}.$$

A.2. Proofs of lemmas

Proof of Lemma A.1. According to the Definition 3.1, with $q \geq 1$,

$$\begin{aligned}
&|\mathcal{R}_{\text{nat}}(h,Q_1) - \mathcal{R}_{\text{nat}}(h,Q_2)| \\
=&\left|\mathbb{E}_{(x_1,y_1)\sim Q_1,(x_2,y_2)\sim Q_2}[\ell(h(x_1),y_1) - \ell(h(x_2),y_2)]\right| \\
=&\left|\inf_{\gamma\in\Gamma(Q_1,Q_2)} \mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}[\ell(h(x_1),y_1) - \ell(h(x_2),y_2)]\right| \\
\leq&\inf_{\gamma\in\Gamma(Q_1,Q_2)} \mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[|\ell(h(x_1),y_1) - \ell(h(x_2),y_2)|\right] \\
=&\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[|\ell(h(x_1),y_1) - \ell(h(x_2),y_2)|\right]\right\}^{q\cdot\frac{1}{q}} \\
\leq&\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[|\ell(h(x_1),y_1) - \ell(h(x_2),y_2)|^q\right]\right\}^{\frac{1}{q}},
\end{aligned}$$

where the last line holds because $(\mathbb{E}[|\ell(h(x_1),y_1)-\ell(h(x_2),y_2)|])^q \leq \mathbb{E}[|\ell(h(x_1),y_1)-\ell(h(x_2),y_2)|^q]$ (obtained by applying the Jensen's inequality).
Since the loss function $\ell(\cdot,\cdot)$ is Holder continuous and the hypothesis h is Lipschitz continuous, we can further get that

$$\begin{aligned}
&|\mathcal{R}_{\text{nat}}(h,Q_1) - \mathcal{R}_{\text{nat}}(h,Q_2)| \\
\leq&\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[{c_1}^q(|h(x_1)-h(x_2)| + |y_1-y_2|)^{c_2+q}\right]\right\}^{\frac{1}{q}} \\
\leq&\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[{c_1}^q(\rho\|x_1-x_2\|_1 + |y_1-y_2|)^{c_2+q}\right]\right\}^{\frac{1}{q}} \\
=&c_1\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[\sqrt{\rho^2+1}\left(\frac{\rho}{\sqrt{\rho^2+1}}\|x_1-x_2\|_1 + \frac{1}{\sqrt{\rho^2+1}}|y_1-y_2|\right)\right]^{c_2+q}\right\}^{\frac{1}{q}} \\
\leq&c_1\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[\sqrt{\rho^2+1}(\|x_1-x_2\|_1 + |y_1-y_2|)\right]^{c_2+q}\right\}^{\frac{1}{q}} \\
=&c_1(\sqrt{\rho^2+1})^{\frac{c_2+q}{q}}\inf_{\gamma\in\Gamma(Q_1,Q_2)} \left\{\mathbb{E}_{(x_1,y_1),(x_2,y_2)\sim\gamma}\left[(\|x_1-x_2\|_1 + |y_1-y_2|)^{c_2+q}\right]\right\}^{\frac{1}{c_2+q}\cdot\frac{c_2+q}{q}} \\
=&c_1\left\{\sqrt{\rho^2+1}\,\mathcal{W}_{c_2+q}(Q_1,Q_2)\right\}^{\frac{c_2+q}{q}}
\end{aligned}$$

□

Proof of Lemma A.2. For any $t \in [T]$,

$$
\begin{aligned}
&\mathcal{R}_{\text{rob}}(h^{(T)}, P) - \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) \\
=&\mathcal{R}_{\text{rob}}(h^{(T)}, P) - \mathcal{R}_{\text{rob}}(h^{(t-1)}, P) + \mathcal{R}_{\text{rob}}(h^{(t-1)}, P) - \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) \\
=&\sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{rob}}(h^{(t')}, P) - \mathcal{R}_{\text{rob}}(h^{(t'-1)}, P) \big\} + \mathcal{R}_{\text{rob}}(h^{(t-1)}, P) - \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) \\
=&\sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{rob}}(h^{(t')}, P) - \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) \big\} + \sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) - \mathcal{R}_{\text{rob}}(h^{(t'-1)}, P) \big\} \\
&- \big(\mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) - \mathcal{R}_{\text{rob}}(h^{(t-1)}, P) \big) \\
=&\sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{rob}}(h^{(t')}, P) - \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) \big\} + \sum_{t'=t+1}^{T} \big\{ \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) - \mathcal{R}_{\text{rob}}(h^{(t'-1)}, P) \big\} \\
\leq&\sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{rob}}(h^{(t')}, P) - \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) \big\} \\
=&\sum_{t'=t}^{T} \big\{ \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t')}}) - \mathcal{R}_{\text{nat}}(h^{(t')}, P_{\text{adv}}^{h^{(t'-1)}}) \big\},
\end{aligned}
$$

where the last two lines are obtained by applying Equation (4) and Equation (1).
According to Lemma A.1,

$$
\mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t)}}) - \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) \leq c_1 \big\{ \sqrt{\rho^2 + 1} \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t)}}, P_{\text{adv}}^{h^{(t-1)}}) \big\}^{\frac{c_2+q}{q}},
$$

and thus we can get

$$
\mathcal{R}_{\text{rob}}(h^{(T)}, P) - \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) \leq \sum_{t'=t}^{T} c_1 \big\{ \sqrt{\rho^2 + 1}\, \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t')}}, P_{\text{adv}}^{h^{(t'-1)}}) \big\}^{\frac{c_2+q}{q}}.
$$

□

A.3. Proofs of theorems

Proof of Theorem 3.1. According to Lemma A.2, the following inequality holds for any $t \in [T]$ and $r_t \geq 0$,

$$
r_t\, \mathcal{R}_{\text{rob}}(h^{(T)}, P) \leq r_t \Big\{ \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) + \sum_{t'=t}^{T} c_1 \big\{ \sqrt{\rho^2 + 1}\, \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t')}}, P_{\text{adv}}^{h^{(t'-1)}}) \big\}^{\frac{c_2+q}{q}} \Big\}.
$$

Therefore, for any set $\{r_t\}_{t \in [T]}$ satisfying $r_t \geq 0$ and $\sum_{t=1}^{T} r_t = T$,

$$
\begin{aligned}
\mathcal{R}_{\text{rob}}(h^{(T)}, P) \leq& \frac{1}{T} \sum_{t=1}^{T} r_t \Big\{ \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) + \sum_{t'=t}^{T} c_1 \big\{ \sqrt{\rho^2 + 1}\, \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t')}}, P_{\text{adv}}^{h^{(t'-1)}}) \big\}^{\frac{c_2+q}{q}} \Big\} \\
\leq& \frac{1}{T} \sum_{t=1}^{T} r_t\, \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) + \frac{1}{T} \sum_{t=1}^{T} r_t \sum_{t'=t}^{T} c_1 \big\{ \sqrt{\rho^2 + 1}\, \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t')}}, P_{\text{adv}}^{h^{(t'-1)}}) \big\}^{\frac{c_2+q}{q}} \\
=& \frac{1}{T} \sum_{t=1}^{T} r_t\, \mathcal{R}_{\text{nat}}(h^{(t)}, P_{\text{adv}}^{h^{(t-1)}}) + \frac{1}{T} \sum_{t=1}^{T} R_t c_1 \big\{ \sqrt{\rho^2 + 1}\, \mathcal{W}_{c_2+q}(P_{\text{adv}}^{h^{(t')}}, P_{\text{adv}}^{h^{(t'-1)}}) \big\}^{\frac{c_2+q}{q}},
\end{aligned}
$$

where $R_t = \sum_{t'=1}^{t} r_{t'}$. □

A.4. Derivation of the decomposition

$$
\begin{aligned}
&\mathcal{R}_{\text{rob}}(h^{(t-1)}, \bar{P}) - \mathcal{R}_{\text{rob}}(h^{(t-1)}, P) \\
=&\sum_{i=1}^{|\mathcal{Y}|} \frac{1}{|\mathcal{Y}|} \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_i) - \sum_{i=1}^{|\mathcal{Y}|} P(y_i)\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_i) \\
=&\sum_{i=1}^{|\mathcal{Y}|} \big(\frac{1}{|\mathcal{Y}|} - P(y_i)\big)\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_i) - \sum_{i=1}^{|\mathcal{Y}|} \frac{1}{|\mathcal{Y}|}\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_1) + \sum_{i=1}^{|\mathcal{Y}|} P(y_i)\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_1) \\
=&\sum_{i=1}^{|\mathcal{Y}|} \big(\frac{1}{|\mathcal{Y}|} - P(y_i)\big)\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_i) - \sum_{i=1}^{|\mathcal{Y}|} \big(\frac{1}{|\mathcal{Y}|} - P(y_i)\big)\, \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_1) \\
=&\sum_{i=2}^{|\mathcal{Y}|} \big(\frac{1}{|\mathcal{Y}|} - P(y_i)\big)\big(\mathcal{R}_{\text{rob}}(h^{(t-1)}, y_i) - \mathcal{R}_{\text{rob}}(h^{(t-1)}, y_1)\big)
\end{aligned}
$$

B. Proofs of Sec. 4 (Theoretical insights)

B.1. Useful lemmas

Lemma B.1 (Class-conditional risks)**.** *For a hypothesis $h \in \mathcal{H}$, the class-conditional risks are $\mathcal{R}_{nat}(h, y) = Pr.\{\mathcal{N}(0,1) < \mathcal{Z}_{nat}(h, y)\}$ and $\mathcal{R}_{rob}(h, y) = Pr.\{\mathcal{N}(0,1) < \mathcal{Z}_{rob}(h, y)\}$, where*

$$
\begin{aligned}
\mathcal{Z}_{nat}(h, y) &= \frac{1}{\sigma}\big(-yb - \mu_1\|w_{G_1}\|_1 - \mu_2\|w_{G_2}\|_1\big), \\
\mathcal{Z}_{rob}(h, y) &= \frac{1}{\sigma}\big(-yb - (\mu_1 - \epsilon)\|w_{G_1}\|_1 - (\mu_2 - \epsilon)\|w_{G_2}\|_1\big).
\end{aligned}
$$

B.2. Proofs of lemmas

Proof of Lemma B.1. Focusing on the conditional natural risk $\mathcal{R}_{\text{nat}}(h, y)$ first, we can get that

$$
\begin{aligned}
\mathcal{R}_{\text{nat}}(h, y) &= \text{Pr.}\big\{yh(x) < 0 | y\big\} \\
&= \text{Pr.}\big\{y\langle w, x\rangle + yb < 0 | y\big\} \\
&= \text{Pr.}\Big\{\sum_{k \in G_1 \cup G_2} y w_k \mathcal{N}(y\theta, \sigma^2) + yb < 0\Big\} \\
&= \text{Pr.}\Big\{\sum_{k \in G_1} w_k \mathcal{N}(\mu_1, \sigma^2) + \sum_{k \in G_2} w_k \mathcal{N}(\mu_2, \sigma^2) + yb < 0\Big\} \\
&= \text{Pr.}\big\{\mathcal{N}(0,1) < \frac{yb - \sum_{k \in G_1} w_k \mu_1 - \sum_{k \in G_2} w_k \mu_2}{\sqrt{\sum_{k \in G_1} w_k^2 + \sum_{k \in G_2} w_k^2}\sigma}\big\} \\
&= \text{Pr.}\big\{\mathcal{N}(0,1) < \frac{yb - \mu_1\|w_{G_1}\|_1 - \mu_2\|w_{G_2}\|_1}{\sigma\|w\|_2}\big\} \\
&= \text{Pr.}\big\{\mathcal{N}(0,1) < \frac{1}{\sigma}\big(-yb - \mu_1\|w_{G_1}\|_1 - \mu_2\|w_{G_2}\|_1\big)\big\}.
\end{aligned}
$$

Therefore, $\mathcal{R}_{\text{nat}}(h, y) = \text{Pr.}\{\mathcal{N}(0,1) < \mathcal{Z}_{\text{nat}}(h, y)\}$, where

$$
\mathcal{Z}_{\text{nat}}(h, y) = \frac{1}{\sigma}\big(-yb - \mu_1\|w_{G_1}\|_1 - \mu_2\|w_{G_2}\|_1\big).
$$

Then, we analyze the conditional robust risks. Since

$$\begin{aligned}
\mathcal{R}_{\text{rob}}(h,y) &= \Pr.\big\{\min_{\|\delta\|\le\epsilon} yh(x+\delta)<0|y\big\}\\
&= \Pr.\big\{\max_{\|\delta\|\le\epsilon} y\langle w,x+\delta\rangle + yb<0|y\big\}\\
&= \Pr.\big\{\sum_{k\in G_1\cup G_2}\max_{\|\delta_k\|\le\epsilon} yw_k(\mathcal{N}(y\theta_k,\sigma^2)+\delta_k)+yb<0\big\}\\
&= \Pr.\big\{\sum_{k\in G_1\cup G_2} w_k(\mathcal{N}(\theta_k-\epsilon,\sigma^2))+yb<0\big\}\\
&= \Pr.\big\{\mathcal{N}(0,1)<\frac{yb-\sum_{k\in G_1}w_k(\mu_1-\epsilon)-\sum_{k\in G_2}w_k(\mu_2-\epsilon)}{\sqrt{\sum_{k\in G_1}w_k^2+\sum_{k\in G_2}w_k^2}\sigma}\big\}\\
&= \Pr.\big\{\mathcal{N}(0,1)<\frac{yb-(\mu_1-\epsilon)\|w_{G_1}\|_1-(\mu_2-\epsilon)\|w_{G_2}\|_1}{\sigma\|w\|_2}\big\}\\
&= \Pr.\big\{\mathcal{N}(0,1)<\frac{1}{\sigma}\big(-yb-(\mu_1-\epsilon)\|w_{G_1}\|_1-(\mu_2-\epsilon)\|w_{G_2}\|_1\big)\big\},
\end{aligned}$$

we can come to the conclusion that $\mathcal{R}_{\text{rob}}(h,y)=\Pr.\{\mathcal{N}(0,1)<\mathcal{Z}_{\text{rob}}(h,y)\}$, where

$$\mathcal{Z}_{\text{rob}}(h,y)=\frac{1}{\sigma}\big(-yb-(\mu_1-\epsilon)\|w_{G_1}\|_1-(\mu_2-\epsilon)\|w_{G_2}\|_1\big).$$

□

Proof of Lemma 4.2. According to Lemma B.1, the sign of the difference between the two class-conditional natural risks is

$$\begin{aligned}
\text{sign}\big(\mathcal{R}_{\text{nat}}(h,-1)-\mathcal{R}_{\text{nat}}(h,+1)\big) &= \text{sign}\big(\Pr.\{\mathcal{N}(0,1)<\mathcal{Z}_{\text{nat}}(h,-1)\}-\Pr.\{\mathcal{N}(0,1)<\mathcal{Z}_{\text{nat}}(h,+1)\}\big)\\
&= \text{sign}\big(\mathcal{Z}_{\text{nat}}(h,-1)-\mathcal{Z}_{\text{nat}}(h,+1)\big)\\
&= \text{sign}(b).
\end{aligned}$$

Meanwhile, the sign of the difference between the two class-conditional robust risks is

$$\begin{aligned}
\text{sign}\big(\mathcal{R}_{\text{rob}}(h,-1)-\mathcal{R}_{\text{rob}}(h,+1)\big) &= \text{sign}\big(\Pr.\{\mathcal{N}(0,1)<\mathcal{Z}_{\text{rob}}(h,-1)\}-\Pr.\{\mathcal{N}(0,1)<\mathcal{Z}_{\text{rob}}(h,+1)\}\big)\\
&= \text{sign}\big(\mathcal{Z}_{\text{rob}}(h,-1)-\mathcal{Z}_{\text{rob}}(h,+1)\big)\\
&= \text{sign}(b).
\end{aligned}$$

Therefore, we come to the conclusion that if the bias $b\neq 0$, f will have a disparity between the two class on both natural and robust risks. Meanwhile, $\text{sign}\big(\mathcal{R}_{\text{nat}}(h,-1)-\mathcal{R}_{\text{nat}}(h,+1)\big)=\text{sign}\big(\mathcal{R}_{\text{rob}}(h,-1)-\mathcal{R}_{\text{rob}}(h,+1)\big)=\text{sign}(b)$. □

B.3. Proofs of theorems

Proof of Theorem 4.1. Under the condition that $\epsilon_y\in(\mu_2,\mu_1)$, we first prove that $w_i^{(t)}\ge w_i^{(t-1)}$ for $\forall i\in G_1$ holds by contradiction. Specifically, the contradiction to be proved is that if there exists $w_i<w_i^{(t-1)}$ for $i\in G_1$ where w_i is the i-th dimension of the hypothesis h's weight w, $h\neq h^{(t)}$.
The training objective of $h^{(t)}$ is

$$\begin{aligned}
\mathcal{R}_{\text{rob}}(h^{(t-1)}) &= \Pr.\big\{\min_{\|\delta\|\le\epsilon_y} y(\langle w^{(t-1)},x+\delta\rangle+b^{(t-1)})<0\big\}\\
&= \Pr.\big\{\sum_{k\in G_1\cup G_2}\min_{\|\delta_k\|\le\epsilon_y} yw_k^{(t-1)}(\mathcal{N}(y\theta_k,\sigma^2)+\delta_k)+yb^{(t-1)}<0\big\}\\
&= \Pr.\big\{\sum_{k\in G_1\cup G_2}\min_{\|\delta_k\|\le\epsilon_y} w_k^{(t-1)}(\mathcal{N}(\theta_k+y\delta_k,\sigma^2)+yb^{(t-1)}<0\big\}\\
&= \Pr.\big\{\sum_{k\in G_1\cup G_2} w_k^{(t-1)}\mathcal{N}(\theta_k-\epsilon_y,\sigma^2)+yb^{(t-1)}<0\big\}.
\end{aligned}$$

Then, we can find that if there exists $i \in G_1$ satisfying $w_i < w_i^{(t-1)}$,

$$\begin{aligned}\mathcal{R}_{\text{rob}}(h^{(t-1)}) =& \text{ Pr.}\Big\{\sum_{k\in G_1\cup G_2, k\neq i} w_k^{(t-1)}\mathcal{N}(\theta_k-\epsilon_y,\sigma^2)+w_i^{(t-1)}\mathcal{N}(\mu_1-\epsilon_y,\sigma^2)+yb^{(t-1)}<0\Big\}\\ <& \text{ Pr.}\Big\{\sum_{k\in G_1\cup G_2, k\neq i} w_k^{(t-1)}\mathcal{N}(\theta_k-\epsilon_y,\sigma^2)+w_i\mathcal{N}(\mu_1-\epsilon_y,\sigma^2)+yb^{(t-1)}<0\Big\},\end{aligned}$$

which means replacing $w_i^{(t-1)}$ by w_i will result in a larger robust risk. Since the optimization objective of $h^{(t)}$ is to minimize the robust risk, $h \neq h^{(t)}$. Therefore, we come to the conclusion that $h^{(t)}$ satisfies that $w_i^{(t)} \geq w_i^{(t-1)}$ for $\forall i \in G_1$.
Next, we prove $w_j^{(t)} \leq w_j^{(t-1)}$ for $\forall j \in G_2$ by contradiction, which is that if there exists $w_j > w_j^{(t-1)}$ for any $j \in G_2$ in the weight w of hypothesis h, $h \neq h_t$. Since the minimization objective of $h^{(t)}$ satisfies

$$\begin{aligned}\mathcal{R}_{\text{rob}}(h^{(t-1)}) =& \text{ Pr.}\Big\{\sum_{k\in G_1\cup G_2, k\neq j} w_k^{(t-1)}\mathcal{N}(\theta_k-\epsilon_y,\sigma^2)+yb^{(t-1)}+w_j^{(t-1)}\mathcal{N}(\mu_2-\epsilon_y,\sigma^2)<0\Big\}\\ <& \text{ Pr.}\Big\{\sum_{k\in G_1\cup G_2, k\neq j} w_k^{(t-1)}\mathcal{N}(\theta_k-\epsilon_y,\sigma^2)+yb^{(t-1)}+w_j\mathcal{N}(\mu_2-\epsilon_y,\sigma^2)<0\Big\},\end{aligned}$$

$w_j^{(t-1)}$ gives a lower robust risk than $w_j > w_j^{(t-1)}$ and thus $h \neq h^{(t)}$. Therefore, we have the conclusion that $h^{(t)}$ satisfies that $w_j^{(t)} \leq w_j^{(t-1)}$ for $\forall j \in G_2$.
Then, we prove that $\mathcal{R}_{\text{rob}}(h^{(t)}, y) \leq \mathcal{R}_{\text{rob}}(h^{(t-1)}, y)$ for $y \in \mathcal{Y}$. Since according to Lemma B.1

$$\mathcal{Z}_{\text{rob}}(h^{(t)}, y) - \mathcal{Z}_{\text{rob}}(h^{(t-1)}, y) = \frac{1}{\sigma}\big((\mu_1-\epsilon)(\|w_{G_1}^{(t-1)}\|_1 - \|w_{G_1}^{(t)}\|_1) + (\mu_2-\epsilon)(\|w_{G_2}^{(t-1)}\|_1 - \|w_{G_2}^{(t)}\|_1)\big) \leq 0,$$

where the inequality holds because $w_i^{(t)} \geq w_i^{(t-1)}$ for $\forall i \in G_1$ and $w_j^{(t)} \leq w_j^{(t-1)}$ for $\forall j \in G_2$, we can obtain that

$$\mathcal{R}_{\text{rob}}(h^{(t)}, y) - \mathcal{R}_{\text{rob}}(h^{(t-1)}, y) = \text{ Pr.}\{\mathcal{N}(0,1) < \mathcal{Z}_{\text{rob}}(h^{(t)}, y)\} - \text{Pr.}\{\mathcal{N}(0,1) < \mathcal{Z}_{\text{rob}}(h^{(t-1)}, y)\} \leq 0.$$

□

Proof of Theorem 4.3. According to Lemma 4.2, $b^{(t)} = 0$ if the conditional natural/robust risks are even across different classes. We first get the exact form of optimal $b^{(t)}$, and then show what settings of class-wise perturbation intensity can lead to zero bias in $h^{(t)}$.
We begin with the minimization objective of $h^{(t)}$, which is

$$\begin{aligned}\mathcal{R}_{\text{rob}}(h^{(t-1)}) \propto& \ \mathcal{R}_{\text{rob}}(h^{(t-1)}, -1) + K \cdot \mathcal{R}_{\text{rob}}(h^{(t-1)}, +1)\\ =& \text{ Pr.}\left\{\mathcal{N}(0,1) < \mathcal{Z}_{\text{rob}}(h^{(t-1)}, -1)\right\} + K \cdot \text{Pr.}\left\{\mathcal{N}(0,1) < \mathcal{Z}_{\text{rob}}(h^{(t-1)}, +1)\right\},\end{aligned}$$

where

$$\mathcal{Z}_{\text{rob}}(h^{(t-1)}, y) = \frac{-yb^{(t-1)} - (\mu_1-\epsilon_y)\|w_{G_1}^{(t-1)}\|_1 - (\mu_2-\epsilon_y)\|w_{G_2}^{(t-1)}\|_1}{\sigma}$$

according to Lemma B.1. Since the optimal $b^{(t)}$ will be found by letting $\frac{\partial \mathcal{R}_{\text{rob}}(h^{(t-1)})}{\partial b^{(t-1)}} = 0$ where

$$\begin{aligned}\frac{\partial \mathcal{R}_{\text{rob}}(h^{(t-1)})}{\partial b^{(t-1)}} =& \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, -1)) \cdot \frac{\partial \mathcal{Z}_{\text{rob}}(h^{(t-1)}, -1)}{\partial b^{(t-1)}}\\ &+ K \cdot \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, +1)) \cdot \frac{\partial \mathcal{Z}_{\text{rob}}(h^{(t-1)}, +1)}{\partial b^{(t-1)}}\\ =& \frac{1}{\sqrt{2\pi}\sigma}\Big(\exp(-\frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, -1)) - \exp(\log K - \frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, +1))\Big),\end{aligned}$$

we can obtain the following equation satisfied by optimal $b^{(t)}$

$$\frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, +1) - \frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, -1) - \log K = 0.$$

Letting $A = (\mu_1\|w_{G_1}^{(t-1)}\|_1 + \mu_2\|w_{G_2}^{(t-1)}\|_1)/\|w^{(t-1)}\|_1$, we have

$$\begin{aligned}
&\frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, +1) - \frac{1}{2}\mathcal{Z}_{\text{rob}}^2(h^{(t-1)}, -1) - \log K \\
=&\frac{1}{2\sigma^2}\big((-b^{(t-1)} + (\epsilon_{+1} - A)\|w^{(t-1)}\|_1)^2 - (b^{(t-1)} + (\epsilon_{-1} - A)\|w^{(t-1)}\|_1)^2\big) - \log K \\
=&\frac{1}{2\sigma^2}\big((\epsilon_{+1} - A)^2\|w^{(t-1)}\|_1^2 - (\epsilon_{-1} - A)^2\|w^{(t-1)}\|_1^2 + (-2\epsilon_{+1} - 2\epsilon_{-1} + 4A)\|w^{(t-1)}\|_1 b^{(t-1)}\big) - \log K.
\end{aligned}$$

Therefore, the exact form of optimal $b^{(t)}$ is

$$b^{(t)} = \frac{2\sigma^2 \log K - (\epsilon_{+1} - A)^2\|w^{(t-1)}\|_1^2 + (\epsilon_{-1} - A)^2\|w^{(t-1)}\|_1^2}{(-2\epsilon_{+1} - 2\epsilon_{-1} + 4A)\|w^{(t-1)}\|_1},$$

which indicates that if $(\epsilon_{+1} - A)^2 - (\epsilon_{-1} - A)^2 = \frac{2\sigma^2 \log K}{\|w^{(t-1)}\|_1^2}$ holds, then $b^{(t)} = 0$. Therefore, we can simply find that if

$$\begin{aligned}
\epsilon_{+1} &< A - \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K}, \\
\epsilon_{-1} &= A - \sqrt{(\epsilon_{+1} - A)^2 - 2\sigma^2\|w^{(t-1)}\|_1^{-2}\log K},
\end{aligned}$$

the optimal $b^{(t)} = 0$. Meanwhile, according to Lemma 4.2, $b^{(t)} = 0$ indicates that the conditional risks are even, i.e., $\mathcal{R}_{\text{nat}}(h^{(t)}, -1) - \mathcal{R}_{\text{nat}}(h^{(t)}, +1) = 0$ and $\mathcal{R}_{\text{rob}}(h^{(t)}, -1) - \mathcal{R}_{\text{rob}}(h^{(t)}, +1) = 0$. □

Proof of Theorem 4.4. Since

$$A = \frac{\mu_1\|w_{G_1}^{(t-1)}\|_1 + \mu_2\|w_{G_2}^{(t-1)}\|_1}{\|w^{(t-1)}\|_1} = \frac{\mu_1\|w_{G_1}^{(t-1)}\|_1 + \mu_2\|w_{G_2}^{(t-1)}\|_1}{\|w_{G_1}^{(t-1)}\|_1 + \|w_{G_2}^{(t-1)}\|_1},$$

we can simply find that A can be bounded as $A \in [\mu_2, \mu_1]$. For the perturbation intensity ϵ_{+1}, the intersection

$$\mathcal{F}_{\text{rob}}(\epsilon_{+1}) \cap \mathcal{F}_{\text{bal}}(\epsilon_{+1}) = (\mu_2, \mu_1) \cap (0, A) = (\mu_2, A - \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K}) \neq \emptyset.$$

Next, for the perturbation intensity ϵ_{-1}, since $\mathcal{F}_{\text{bal}}(\epsilon_{-1}) = \{A - \sqrt{(A - \epsilon_{+1})^2 - 2\sigma^2\|w^{(t-1)}\|_1^{-2}\log K}\} \neq \emptyset$ satisfies

$$A - \sqrt{(A - \epsilon_{+1})^2 - 2\sigma^2\|w^{(t-1)}\|_1^{-2}\log K} \leq A \leq \mu_2,$$

and

$$A - \sqrt{(A - \epsilon_{+1})^2 - 2\sigma^2\|w^{(t-1)}\|_1^{-2}\log K} \geq A - \sqrt{(A - \epsilon_{+1})^2} = \epsilon_{+1} \geq \mu_1,$$

$\mathcal{F}_{\text{bal}}(\epsilon_{-1})$ is in the $\mathcal{F}_{\text{rob}}(\epsilon_{-1}) = (\mu_2, \mu_1)$, and thus $\mathcal{F}_{\text{rob}}(\epsilon_{-1}) \cap \mathcal{F}_{\text{bal}}(\epsilon_{-1}) = \mathcal{F}_{\text{bal}}(\epsilon_{-1}) \neq \emptyset$. □

B.4. Derivation of the bounds

For the lower bound of $\epsilon_{-1} = A - \sqrt{(A - \epsilon_{+1})^2 - (\sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K})^2}$, since $(\sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K})^2 \geq 0$, we can get that

$$\epsilon_{-1} \geq A - \sqrt{(\epsilon_{+1} - A)^2} = \epsilon_{+1}.$$

As for the upper bound, we can find that

$$\begin{aligned}
\epsilon_{-1} =& A - \sqrt{(A - \epsilon_{+1} - \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K})(A - \epsilon_{+1} + \sqrt{2}\sigma\frac{\|w^{(t-1)}\|_2}{\|w^{(t-1)}\|_1}\sqrt{\log K})} \\
\leq& A - \sqrt{(A - \epsilon_{+1} - \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K})^2} \\
=& \epsilon_{+1} + \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K}.
\end{aligned}$$

Therefore, $\epsilon_{-1} \in [\epsilon_{+1}, \epsilon_{+1} + \sqrt{2}\sigma\|w^{(t-1)}\|_1^{-1}\sqrt{\log K}]$.

C. Reproduction

We provide a simple PyTorch-style pseudocode in Algorithm 1 for better understand on RobustLT. Meanwhile, our code repository (including the split configurations for long-tail datasets) is available at https://github.com/zhang-lilin/RobustLT.

Algorithm 1 PyTorch-style pseudocode of RobustLT

```
"""
Args:
   alpha, beta: hyper-parameters of RobustLT
   samples_per_class: shape=[num_classes], the number of samples in each class
   base_algorithm: a given base adversarial training algorithm
   eps, step_size, steps: parameters for adversarial example generation
Returns:
   model trained by RobustLT-enhanced base_algorithm
"""

# initial the value of classwise perturbation intensity by CPB
n_max, N = samples_per_class.max(), samples_per_class.sum()
tau = alpha / ((samples_per_class / N) * (n_max / samples_per_class).log().sqrt()).sum()
classwise_eps_max = (1 - alpha) * eps + tau * (n_max / samples_per_class).log().sqrt() * eps

# train for T epochs
for t in range(T):

   # calculate adversarial intensity for current epoch by AIW
   intensity = min((t - 1) / (T * beta), 1)
   classwise_eps = classwise_eps_max * intensity

   for (x, y) in dataloader:

      # generate adversarial examples with RobustLT
      eps_t = classwise_eps[y]
      step_size_t = eps_t / eps * step_size
      x_adv = base_algorithm.get_adversarial_example(x, y, eps_t, step_size_t, steps)

      # model update
      base_algorithm.forward(model, optimizer, x, x_adv, y)

# return the trained model
return model
```

D. Additional experiments

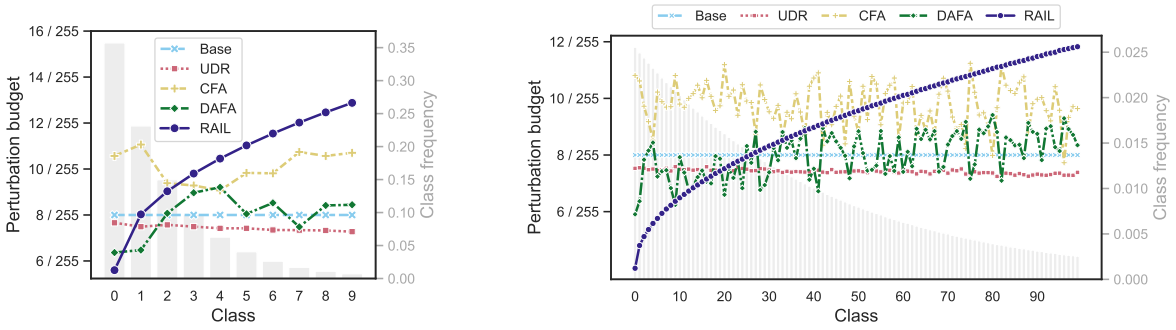
D.1. Detailed configurations

Long-tail dataset generation. We follow the method in [5] to generate the long-tail datasets from balanced datasets (e.g., CIFAR10/100 and TinyImagenet). Specifically, given the imbalance ratio K and the number of available samples per class N, we then set the number of instances of the i-th class $N_i = N \cdot K^{-\frac{i-1}{|\mathcal{Y}|-1}}$, where $i \in \{1, 2, \ldots, |\mathcal{Y}|\}$. The generation are randomly conducted three times for all experiments, and the average with standard deviation are reported.

Other settings. Following [28], all experimental results are obtained with setting the activation function to ReLU, BN-mode to eval, the optimizer to SGD with Nesterov momentum [27], where learning rate, weight decay, and momentum are set to 0.1, 5e-4, and 0.9, respectively. To remain consistent with the settings of the respective base adversarial training algorithms, other hyper-parameters not explicitly mentioned, such as the learning rate schedule, batch size, and parameters for adversarial example generation, are kept the same as in their original implementations. All the experiments are conducted on an NVIDIA RTX 4090 GPU.

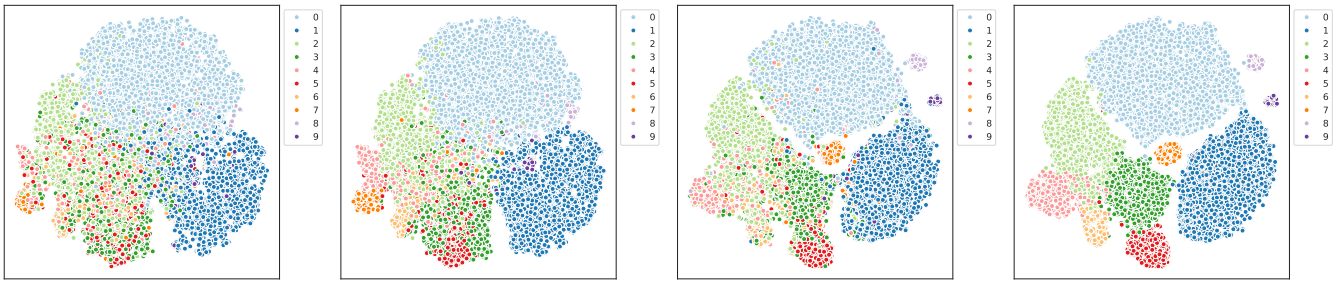
D.2. Study of adversarial distribution

Since RobustLT consists of two components: (i) CPB, which rebalances the skewed training objective on adversarial data, and (ii) AIW, which stabilizes the evolution of adversarial distributions during training, we validate their effectiveness here.



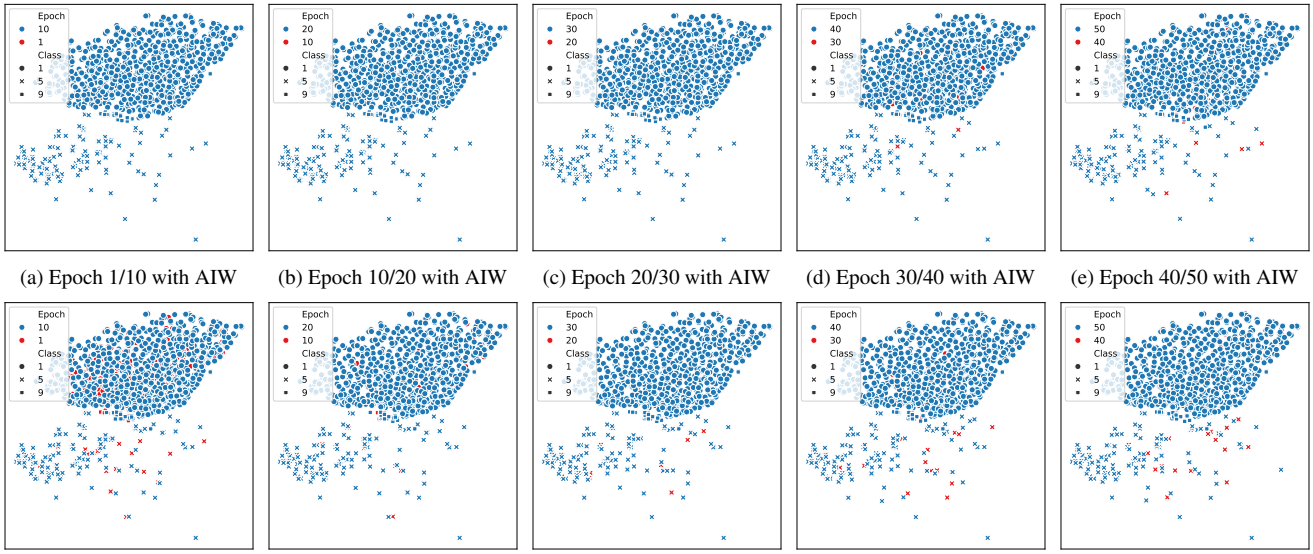
(a) CIFAR10-LF (b) CIFAR100-LF

Figure 3. Adaptive perturbation intensity in final epoch of different enhancement methods, averaged over multiple base algorithms.



(a) AEs against h_1 with CPB (b) AEs against h_1 w/o CPB (c) AEs against h_2 with CPB (d) AEs against h_2 w/o CPB

Figure 4. T-SNE visualizations of the latent space logits of adversarial examples (AEs) generated with and without CPB extracted from h_1 and h_2 on CIFAR10-LT, where h_1 and h_2 are the models trained with AT-BSL and AT, respectively.



(a) Epoch 1/10 with AIW (b) Epoch 10/20 with AIW (c) Epoch 20/30 with AIW (d) Epoch 30/40 with AIW (e) Epoch 40/50 with AIW

(f) Epoch 1/10 w/o AIW (g) Epoch 10/20 w/o AIW (h) Epoch 20/30 w/o AIW (i) Epoch 30/40 w/o AIW (j) Epoch 40/50 w/o AIW

Figure 5. T-SNE visualizations of the latent space logits of adversarial examples (AEs) generated with and without AIW across multiple epochs on CIFAR10-LT. Colors indicate training epochs, while shapes denote ground truth labels. Fewer exposed red samples reflect stronger alignment between successive adversarial distributions.

Rebalanced adversarial distribution. Models suffering from overconfidence issue caused by data imbalance tend to generate biased adversarial examples, which in turn exacerbate the imbalance. CPB helps to rebalance the distribution of

Table 3. Hyper-parameter settings with respect to (α, β).

Dataset	AT	AWP	RoBal	REAT	AT-BSL	TAET
CIFAR10-LT	$(0.3, 0.8)$	$(0.3, 0.8)$	$(0.3, 0.8)$	$(0.3, 0.8)$	$(0.3, 0.8)$	$(0.3, 1.0)$
CIFAR100-LT	$(0.5, 0.6)$	$(0.5, 0.6)$	$(0.5, 0.6)$	$(0.5, 0.6)$	$(0.5, 0.6)$	$(0.5, 0.6)$
TinyImagenet-LT	$(0.3, 0.2)$	$(0.3, 0.2)$	$(0.5, 0.4)$	$(0.5, 0.4)$	$(0.4, 0.6)$	$(0.3, 0.8)$

Table 4. Natural and robust accuracies of various base adversarial training algorithms without and with RobustLT using WRN-28-10 on CIFAR10-LT under different imbalance ratios. Better results are bolded.

Base	Method	Imbalance Ratio = 10				Imbalance Ratio = 100			
		Nat. (all)	Nat. (tail)	Rob. (all)	Rob. (tail)	Nat. (all)	Nat. (tail)	Rob. (all)	Rob. (tail)
AT	origin	$74.31_{\pm 0.25}$	$68.88_{\pm 0.38}$	$35.07_{\pm 0.09}$	$24.45_{\pm 0.19}$	$51.01_{\pm 0.58}$	$40.45_{\pm 0.79}$	$25.18_{\pm 0.15}$	$10.87_{\pm 0.29}$
	RobustLT	$\mathbf{78.07}_{\pm 0.27}$	$\mathbf{73.53}_{\pm 0.26}$	$\mathbf{37.15}_{\pm 0.15}$	$\mathbf{28.14}_{\pm 0.27}$	$\mathbf{53.12}_{\pm 0.02}$	$\mathbf{42.08}_{\pm 0.07}$	$\mathbf{27.23}_{\pm 0.05}$	$\mathbf{13.73}_{\pm 0.08}$
AWP	origin	$76.01_{\pm 0.43}$	$70.78_{\pm 0.55}$	$36.78_{\pm 0.31}$	$26.08_{\pm 0.28}$	$52.24_{\pm 0.61}$	$40.88_{\pm 0.83}$	$26.18_{\pm 0.17}$	$11.74_{\pm 0.30}$
	RobustLT	$\mathbf{81.04}_{\pm 0.26}$	$\mathbf{77.01}_{\pm 0.34}$	$\mathbf{37.22}_{\pm 0.24}$	$\mathbf{27.73}_{\pm 0.25}$	$\mathbf{56.67}_{\pm 0.75}$	$\mathbf{46.32}_{\pm 0.93}$	$\mathbf{26.49}_{\pm 0.21}$	$\mathbf{13.02}_{\pm 0.15}$
RoBal	origin	$77.99_{\pm 0.16}$	$73.33_{\pm 0.11}$	$39.81_{\pm 0.19}$	$30.26_{\pm 0.21}$	$\mathbf{69.42}_{\pm 0.38}$	$\mathbf{63.16}_{\pm 0.36}$	$29.01_{\pm 0.56}$	$18.45_{\pm 0.58}$
	RobustLT	$\mathbf{82.32}_{\pm 0.27}$	$\mathbf{79.73}_{\pm 0.22}$	$\mathbf{42.35}_{\pm 0.12}$	$\mathbf{36.80}_{\pm 0.26}$	$68.74_{\pm 0.46}$	$63.12_{\pm 0.70}$	$\mathbf{34.60}_{\pm 0.55}$	$\mathbf{26.89}_{\pm 0.87}$
REAT	origin	$77.76_{\pm 0.11}$	$73.56_{\pm 0.18}$	$36.57_{\pm 0.25}$	$27.57_{\pm 0.38}$	$64.88_{\pm 0.36}$	$57.38_{\pm 0.34}$	$26.27_{\pm 0.01}$	$14.31_{\pm 0.12}$
	RobustLT	$\mathbf{80.45}_{\pm 0.15}$	$\mathbf{76.87}_{\pm 0.22}$	$\mathbf{39.46}_{\pm 0.21}$	$\mathbf{32.17}_{\pm 0.16}$	$\mathbf{67.71}_{\pm 0.39}$	$\mathbf{61.35}_{\pm 0.44}$	$\mathbf{30.58}_{\pm 0.14}$	$\mathbf{21.17}_{\pm 0.10}$
AT-BSL	origin	$83.47_{\pm 0.21}$	$80.60_{\pm 0.33}$	$46.42_{\pm 0.19}$	$39.86_{\pm 0.34}$	$\mathbf{75.08}_{\pm 0.55}$	$\mathbf{70.62}_{\pm 0.81}$	$34.94_{\pm 0.22}$	$26.39_{\pm 0.38}$
	RobustLT	$\mathbf{83.63}_{\pm 0.07}$	$\mathbf{81.28}_{\pm 0.13}$	$\mathbf{48.86}_{\pm 0.32}$	$\mathbf{44.43}_{\pm 0.58}$	$73.54_{\pm 0.71}$	$68.93_{\pm 1.01}$	$\mathbf{39.25}_{\pm 0.11}$	$\mathbf{32.28}_{\pm 0.50}$
TAET	origin	$77.96_{\pm 0.17}$	$75.37_{\pm 0.21}$	$42.37_{\pm 0.32}$	$37.55_{\pm 0.22}$	$57.89_{\pm 3.23}$	$51.92_{\pm 2.91}$	$27.79_{\pm 0.45}$	$20.28_{\pm 0.44}$
	RobustLT	$\mathbf{79.08}_{\pm 0.92}$	$\mathbf{77.43}_{\pm 0.69}$	$\mathbf{42.60}_{\pm 0.35}$	$\mathbf{39.84}_{\pm 0.48}$	$\mathbf{62.02}_{\pm 1.19}$	$\mathbf{57.09}_{\pm 1.27}$	$\mathbf{30.56}_{\pm 0.07}$	$\mathbf{24.45}_{\pm 0.79}$

these adversarial examples. To validate this, we visualize and compare the distributions of adversarial examples generated with and without CPB in Figure 4. The results indicate that: (i) CPB encourages the generation of more diverse adversarial examples for tail-class regardless the impact of model overconfidence, as shown by the more dispersed adversarial examples of the same color in Figure 4a and Figure 4c, compared to Figure 4b and Figure 4d, respectively. Interpreting adversarial example generation as a form of data augmentation, these findings align with [47], which highlights the importance of data augmentation in increasing sample diversity for enhancing robustness in long-tail settings. (ii) Models affected by severe data imbalance generate more biased adversarial examples, as evidenced by the narrower distribution range of tail-class adversarial examples in Figure 4b. This effect is more pronounced in Figure 4d, since models trained by AT are more sensitive to class-imbalance than that trained by AT-BSL (see Table 1), which further validates our motivation.

Stable evolution over iterations. To validate that AIW stabilizes the evolution of adversarial distributions, we visualize adversarial examples during training with and without AIW. Specifically, we extract checkpoints of adversarial distributions at intervals of 10 epochs and compare the alignment between successive checkpoints, as shown in Figure 5. The results show that: (i) AIW indeed stabilizes the evolution, since distributions with AIW exhibit stronger alignment, evidenced by red samples largely overlapping with blue ones in the first row of Figure 5 compared to the second row. (ii) Adversarial distributions from the early training stage are less aligned than later ones, as seen in Figure 5f, where adversarial examples from epochs 1 and 10 expose more red samples.

D.3. Additional results

Additional experiment results are provided in Tables 4 to 6 and Figure 6, the observations of which are consistent to the analysis in the main text.

E. Discussions

E.1. Limitations

Required assumption. We assume that class frequencies are available in training, which may not always hold in practice. Future works going beyond this assumption are encouraged.

Table 5. Natural and robust accuracies of various base adversarial training algorithms without and with RobustLT using WRN-28-10 on CIFAR100-LT under different imbalance ratios. Better results are bolded.

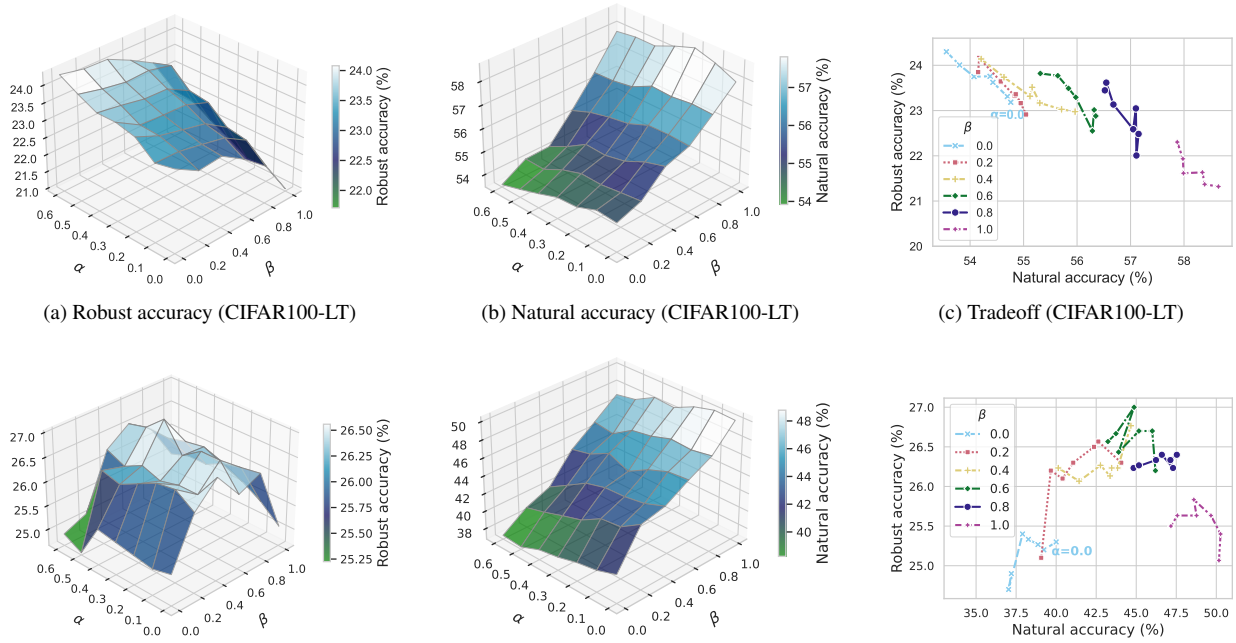
Base	Method	Imbalance Ratio = 5				Imbalance Ratio = 50			
		Nat. (all)	Nat. (tail)	Rob. (all)	Rob. (tail)	Nat. (all)	Nat. (tail)	Rob. (all)	Rob. (tail)
AT	origin	$48.94_{\pm 0.42}$	$46.00_{\pm 0.44}$	$19.34_{\pm 0.10}$	$18.03_{\pm 0.09}$	$33.68_{\pm 0.04}$	$26.53_{\pm 0.05}$	$12.86_{\pm 0.17}$	$9.37_{\pm 0.30}$
	RobustLT	$\mathbf{51.36}_{\pm 0.25}$	$\mathbf{49.63}_{\pm 0.32}$	$\mathbf{20.15}_{\pm 0.36}$	$\mathbf{19.65}_{\pm 0.43}$	$\mathbf{37.06}_{\pm 0.21}$	$\mathbf{30.62}_{\pm 0.17}$	$\mathbf{13.61}_{\pm 0.22}$	$\mathbf{10.75}_{\pm 0.32}$
AWP	origin	$50.55_{\pm 0.11}$	$47.42_{\pm 0.29}$	$20.87_{\pm 0.05}$	$19.39_{\pm 0.06}$	$34.59_{\pm 0.07}$	$27.07_{\pm 0.03}$	$13.65_{\pm 0.27}$	$9.93_{\pm 0.36}$
	RobustLT	$\mathbf{54.30}_{\pm 0.05}$	$\mathbf{52.23}_{\pm 0.16}$	$\mathbf{21.76}_{\pm 0.12}$	$\mathbf{20.96}_{\pm 0.07}$	$\mathbf{38.79}_{\pm 0.22}$	$\mathbf{31.98}_{\pm 0.45}$	$\mathbf{14.39}_{\pm 0.15}$	$\mathbf{11.24}_{\pm 0.16}$
RoBal	origin	$53.74_{\pm 0.03}$	$52.35_{\pm 0.10}$	$21.71_{\pm 0.30}$	$21.14_{\pm 0.27}$	$40.10_{\pm 0.26}$	$35.79_{\pm 0.60}$	$14.62_{\pm 0.04}$	$12.26_{\pm 0.17}$
	RobustLT	$\mathbf{54.04}_{\pm 0.12}$	$\mathbf{52.95}_{\pm 0.17}$	$\mathbf{22.17}_{\pm 0.24}$	$\mathbf{22.28}_{\pm 0.27}$	$\mathbf{41.48}_{\pm 0.43}$	$\mathbf{37.70}_{\pm 0.61}$	$\mathbf{15.61}_{\pm 0.07}$	$\mathbf{14.21}_{\pm 0.18}$
REAT	origin	$51.34_{\pm 0.19}$	$50.11_{\pm 0.18}$	$19.54_{\pm 0.15}$	$18.95_{\pm 0.20}$	$39.44_{\pm 0.54}$	$35.77_{\pm 0.46}$	$13.04_{\pm 0.44}$	$11.18_{\pm 0.54}$
	RobustLT	$\mathbf{51.76}_{\pm 0.31}$	$\mathbf{51.33}_{\pm 0.32}$	$\mathbf{20.29}_{\pm 0.08}$	$\mathbf{20.41}_{\pm 0.13}$	$\mathbf{40.76}_{\pm 0.15}$	$\mathbf{37.97}_{\pm 0.39}$	$\mathbf{14.29}_{\pm 0.48}$	$\mathbf{13.15}_{\pm 0.61}$
AT-BSL	origin	$58.51_{\pm 0.30}$	$57.08_{\pm 0.25}$	$25.72_{\pm 0.21}$	$25.34_{\pm 0.28}$	$46.38_{\pm 0.12}$	$42.36_{\pm 0.20}$	$18.53_{\pm 0.10}$	$16.45_{\pm 0.09}$
	RobustLT	$\mathbf{59.16}_{\pm 0.02}$	$\mathbf{57.92}_{\pm 0.07}$	$\mathbf{25.92}_{\pm 0.31}$	$\mathbf{26.07}_{\pm 0.37}$	$\mathbf{46.81}_{\pm 0.08}$	$\mathbf{42.72}_{\pm 0.06}$	$\mathbf{19.06}_{\pm 0.34}$	$\mathbf{17.54}_{\pm 0.31}$
TAET	origin	$49.73_{\pm 0.64}$	$46.72_{\pm 0.46}$	$21.08_{\pm 0.29}$	$19.54_{\pm 0.46}$	$35.56_{\pm 0.32}$	$28.28_{\pm 0.41}$	$13.96_{\pm 0.18}$	$10.49_{\pm 0.31}$
	RobustLT	$\mathbf{50.00}_{\pm 0.38}$	$\mathbf{47.75}_{\pm 0.23}$	$\mathbf{21.77}_{\pm 0.17}$	$\mathbf{21.16}_{\pm 0.25}$	$\mathbf{35.58}_{\pm 0.09}$	$\mathbf{29.11}_{\pm 0.18}$	$\mathbf{14.79}_{\pm 0.23}$	$\mathbf{11.77}_{\pm 0.26}$

Table 6. Natural and robust accuracies of AT-BSL without and with RAIL across various datasets and model architectures. ImageNet-LT uses the first 20 classes of ImageNet64 [8] with imbalance ratio 50. Adversarial training on DeiT-S follows [26] to use gradient clipping and pretrained initialization. **Better** results are highlighted.

Dataset	Architecture	Method	Nat. (all)	Nat. (tail)	Rob. (all)	Rob. (tail)
CIFAR10-LT	DeiT-S	origin	$58.30_{\pm 1.33}$	$52.74_{\pm 1.62}$	$33.46_{\pm 0.68}$	$27.95_{\pm 0.86}$
		RobustLT	$\mathbf{60.83}_{\pm 0.51}$	$\mathbf{55.34}_{\pm 0.76}$	$\mathbf{34.21}_{\pm 0.16}$	$\mathbf{28.66}_{\pm 0.31}$
	ResNet-18	origin	$72.86_{\pm 0.59}$	$68.49_{\pm 0.59}$	$38.38_{\pm 0.14}$	$31.50_{\pm 0.28}$
		RobustLT	$\mathbf{73.39}_{\pm 0.48}$	$\mathbf{69.08}_{\pm 0.23}$	$\mathbf{38.86}_{\pm 0.20}$	$\mathbf{32.49}_{\pm 0.59}$
	ResNet-50	origin	$72.86_{\pm 0.88}$	$68.88_{\pm 1.06}$	$39.01_{\pm 0.16}$	$33.01_{\pm 0.17}$
		RobustLT	$\mathbf{74.29}_{\pm 1.40}$	$\mathbf{70.41}_{\pm 1.23}$	$\mathbf{39.65}_{\pm 0.24}$	$\mathbf{34.07}_{\pm 0.41}$
	WRN-28-10	origin	$77.09_{\pm 0.41}$	$72.48_{\pm 0.30}$	$37.98_{\pm 0.42}$	$28.60_{\pm 0.85}$
		RobustLT	$\mathbf{77.61}_{\pm 0.54}$	$\mathbf{73.83}_{\pm 0.60}$	$\mathbf{42.11}_{\pm 0.36}$	$\mathbf{35.98}_{\pm 0.79}$
ImageNet-LT	DeiT-S	origin	$25.73_{\pm 1.16}$	$25.67_{\pm 0.16}$	$16.30_{\pm 0.57}$	$17.25_{\pm 0.97}$
		RobustLT	$\mathbf{32.73}_{\pm 0.37}$	$\mathbf{28.46}_{\pm 0.21}$	$\mathbf{18.73}_{\pm 0.25}$	$\mathbf{17.50}_{\pm 0.20}$
	WRN-28-10	origin	$44.03_{\pm 1.17}$	$40.17_{\pm 0.71}$	$22.63_{\pm 0.12}$	$22.67_{\pm 0.36}$
		RobustLT	$\mathbf{45.03}_{\pm 0.74}$	$\mathbf{41.50}_{\pm 0.74}$	$\mathbf{23.43}_{\pm 0.60}$	$\mathbf{23.58}_{\pm 0.60}$

Gap between theory and methodology (from binary classification to multi-class classification). The conclusions in Sec. 4 based on a binary task, by which we can directly define the perturbation intensity for the majority class '+1' and minority class '-1' as: $\epsilon_{+1} = (1-\alpha)\epsilon$ and $\epsilon_{-1} = (1-\alpha)\epsilon + \tau\sqrt{\log K}\epsilon$, where $K = \frac{P(y=+1)}{P(y=-1)}$, α is a hyper-parameter controlling the minimum perturbation intensity, and τ is a hyper-parameter that determines the variation in perturbation intensity between classes. For multi-class case where $P(y_1) \geq P(y_2) \geq \cdots \geq P(y_{|\mathcal{Y}|})$, the bias corresponding to each minority class is dominated by the gap between it and the most frequent class y_1 according to Equation (6). Therefore, we extend it to the more general multi-class case as $\epsilon_y = (1-\alpha)\epsilon + \tau\sqrt{\log K_y}\epsilon$, where $K_y = \frac{P(y_1)}{P(y)}$. This further forms CPB.

Gap between theory and practice. (i) **About the generalization gap.** We use population risk (over distribution) to isolate the effect of class imbalance and provide qualitative insights that motivate RobustLT. In practice, there exists a generalization gap between population risk and empirical risk [22]. (ii) **About the tradeoff between standard generalization and adversarial robustness.** The conclusion in Lemma 4.2 shows that the bias term is an indicator of both overconfidence phenomenons on standard generalization (reflecting by $\mathcal{R}_{\text{nat}}$) and adversarial robustness (reflecting by $\mathcal{R}_{\text{rob}}$), and consequently the offsetting of negative effects caused by data imbalance takes effect simultaneously on both of the two theoretically. The reason this conclusion holds is that there is no conflict between them for the conceptual binary classification task formalized in Sec. 4. While the view that adversarial robustness is not at odd with standard generalization is supported by many works



(a) Robust accuracy (CIFAR100-LT) (b) Natural accuracy (CIFAR100-LT) (c) Tradeoff (CIFAR100-LT)

(d) Robust accuracy (TinyImageNet-LT) (e) Natural accuracy (TinyImageNet-LT) (f) Tradeoff (TinyImageNet-LT)

Figure 6. Robust accuracy, natural accuracy, and the tradeoff between them under varying settings of α and β when applying to AT-BSL on CIFAR100-LT and TinyImageNet-LT. The tradeoff curves correspond to varying values of β, with individual points representing different $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ in order.

theoretically [32, 50], there is always a tradeoff between them in practice due to the limitation of model capacity. Therefore, in our experiment, CPB, the theoretical foundation of which depends on Theorem 4.4, may not always increase both natural and robust accuracies simultaneously as shown in Figure 2. However, RobustLT can do this due to the incorporating of AIW as discussed in Sec. 6.2. (iii) **About the non-universality of the $\sqrt{\log K}$ scaling.** Our analysis intentionally adopts simplified settings (linear classifiers, Gaussian data) to isolate the effect of class imbalance and provide qualitative insights that motivate RobustLT. The $\sqrt{\log K}$ dependence in Theorem 4.4 is not universal and relies on these assumptions. The key takeaway is the monotonic relationship between class imbalance and perturbation intensity, rather than the exact functional form.

E.2. Warmup w.r.t. adversarial intensity in balanced scenario verses imbalanced scenario

Warmup w.r.t. adversarial intensity is not new for adversarial training with balanced data, which often sets the warmup length to a small value (no more than 20% of total training length on CIFAR10) and has limited effect [28]. Different from that, RobustLT adopts a large warmup length (e.g., 80% of total training length on CIFAR10-LT) leading to a better performance.

E.3. Broader impacts

This work addresses the intersection of adversarial robustness and class imbalance, two fundamental challenges in deploying machine learning models in real-world scenarios. We propose RobustLT, a general framework that enhances adversarial training under long-tailed distributions. Positive societal impacts include improved reliability and fairness in applications involving rare but critical classes (e.g., medical anomalies or uncommon traffic signs). As RobustLT is compatible with existing adversarial training methods, it lowers the barrier for broader adoption in practice. Negative societal impacts could arise if adversarial robustness is misused to build more evasive or manipulative AI systems (e.g., in misinformation or surveillance tools). We encourage transparency, auditing, and responsible use to mitigate such risks.