Appendix

A. Technical Details

In this section, we provide the detailed proofs for our results. Let us denote

$$R_{emp} = \frac{1}{m}\sum_{j=1}^{m} l(\mathcal{A}_S, z_j),\ R_{emp}^{-i} = \frac{1}{m}\sum_{j=1}^{m} l(\mathcal{A}_{S^{-i}}, z_j)$$

To prove Theorem 3.1, we require the following key lemma.

Lemma A.1. *Suppose an algorithm $\mathscr{A}$ satisfies locally elastic stability with $\beta_m(\cdot,\cdot)$ for loss function l. For any $\eta > 0$, let $M = 2(M_\beta + \sup_{z\in\mathcal{Z}} \mathbb{E}_{z_j}\beta(z, z_j) + M_l)$ and $\tilde{M} = 2(2\sup_{z\in\mathcal{Z}} \mathbb{E}_{z_j}\beta(z, z_j) + \eta + M_l)$. There exists a constant $C' > 0$ depending on the Lipchitz constant L and dimension d of z, if m is large enough and ε is small enough, such that*

$$\frac{\eta^2}{32M_\beta^2} - \frac{\log C'm}{m} \geqslant \frac{\varepsilon}{2\tilde{M}^2}(-\varepsilon + \frac{4\varepsilon M^2}{\tilde{M}^2} + 4M),$$

we have

$$\mathbb{P}\big(\mathbb{E}_z[l(\mathcal{A}_S, z)] \geqslant \frac{1}{m}\sum_{j=1}^{m} l(\mathcal{A}_S, z_j) + \frac{2\sup_{z\in\mathcal{Z}} \mathbb{E}_{z_j}\beta(z, z_j)}{m} + \varepsilon\big) \leqslant 2\exp\big(-\frac{m\varepsilon^2}{2\tilde{M}^2}\big).$$

Theorem A.1 (Restatement of Theorem 3.1)**.** *Let $\mathscr{A}$ be an algorithm that has locally elastic stability $\beta_m(\cdot,\cdot)$ with respect to the loss function l. Fixing $0 < \delta < 1$ and $\eta > 0$, for large enough m, with probability at least $1-\delta$, we have*

$$\Delta(\mathcal{A}_S) \leq \frac{2\sup_{z'\in\mathcal{Z}} \mathbb{E}_z\beta(z', z)}{m} + 2\left(2\sup_{z'\in\mathcal{Z}} \mathbb{E}_z\beta(z', z) + \eta + M_l\right)\sqrt{\frac{2\log(2/\delta)}{m}}.$$

Proof. Let $\delta = 2\exp(-m\varepsilon^2/(2\tilde{M}^2))$, which gives us

$$\varepsilon = \tilde{M}\sqrt{\frac{2\log(2/\delta)}{m}}.$$

Plugging the value of ε into the inequality

$$\frac{\eta^2}{32M_\beta^2} - \frac{\log C'm}{m} \geqslant \frac{\varepsilon}{2\tilde{M}^2}(-\varepsilon + \frac{4\varepsilon M^2}{\tilde{M}^2} + 4M),$$

we obtain that

$$\frac{\eta^2}{32M_\beta^2} - \frac{\log C'm}{m} \geqslant \frac{1}{2\tilde{M}}\sqrt{\frac{2\log(2/\delta)}{m}}(-\tilde{M}\sqrt{\frac{2\log(2/\delta)}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}\frac{4M^2}{\tilde{M}} + 4M).$$

It is sufficient if we have

$$\frac{\eta^2}{32M_\beta^2} - \frac{\log C'm}{m} \geqslant \frac{2M^2}{\tilde{M}^2}\frac{\log(2/\delta)}{m} + \frac{2M}{\tilde{M}}\sqrt{\frac{2\log(2/\delta)}{m}}.$$

For simplicity, we let m large enough such that

$$\frac{\log C'm}{m} \leqslant \frac{\eta^2}{64M_\beta^2}, \frac{2M^2\log(2/\delta)}{\tilde{M}^2 m} \leqslant \frac{\eta^2}{128M_\beta^2}, \frac{2M}{\tilde{M}}\sqrt{\frac{2\log(2/\delta)}{m}} \leqslant \frac{\eta^2}{128M_\beta^2},$$

which could be achieved once we notice that

$$\lim_{m\to\infty}\frac{\log C'm}{m} \to 0.$$

Then, we obtain the desirable results by applying Lemma A.1. □

To finish the proof of Theorem 3.1, the only thing left is to prove Lemma A.1. We prove it in the following subsection.

A.1. Proof of Lemma A.1

By locally elastic stability,

$$|R_{emp} - R_{emp}^{-i}| \leqslant \frac{1}{m}\sum_{j\neq i}\frac{\beta(z_i, z_j)}{m} + \frac{M_l}{m}.$$

Recall $S = \{z_1, z_2, \cdots, z_m\}$ and we denote

$$L(S) = \sum_{j=1}^{m} l(\mathcal{A}_S, z_j),\ L(S^{-i}) = \sum_{j\neq i}^{m} l(\mathcal{A}_{S^{-i}}, z_j).$$

Let $\mathcal{F}_k$ be the σ-field generated by $z_1, \cdots, z_k$. We construct Doob's martingale and consider the associated martingale difference sequence

$$D_k = \mathbb{E}[L(S)|\mathcal{F}_k] - \mathbb{E}[L(S)|\mathcal{F}_{k-1}]. \tag{4}$$

Consider event

$$E_{-k} = \Big\{S\Big|\ \sup_{z'\in\mathcal{Z}}\Big|\sum_{j\neq k}\frac{\beta(z', z_j)}{m} - \mathbb{E}_z\beta(z', z)\Big| \leqslant \eta\Big\},$$

where z is drawn from the same distribution as the training examples $\{z_i\}_{i=1}^m$. Let us decompose D_k as $D_k^{(1)} + D_k^{(2)}$, where

$$D_k^{(1)} = \mathbb{E}[L(S)I_{E_{-k}}|\mathcal{F}_k] - \mathbb{E}[L(S)I_{E_{-k}}|\mathcal{F}_{k-1}],\ \ D_k^{(2)} = \mathbb{E}[L(S)I_{E_{-k}^c}|\mathcal{F}_k] - \mathbb{E}[L(S)I_{E_{-k}^c}|\mathcal{F}_{k-1}].$$

By Jensen's inequality,

$$\mathbb{E}[e^{\lambda(\sum_{k=1}^m D_k)}] \leqslant \frac{1}{2}\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(1)})}] + \frac{1}{2}\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}]$$

Now, let us bound the two terms $\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(1)})}]$ and $\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}]$ separately in the following paragraphs, so as to further apply Chernoff bound to obtain a concentration bound for $\sum_{k=1}^m D_k$.

Bounding $\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}]$. First, we consider bounding $\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}]$. Let us further define

$$A_k^{(2)} = \inf_x \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}],$$

$$B_k^{(2)} = \sup_x \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}].$$

Apparently,

$$A_k^{(2)} \leqslant D_k^{(2)} \leqslant B_k^{(2)}.$$

Next, we provide an upper bound for $B_k^{(2)} - A_k^{(2)}$. Consider

$$\begin{aligned}
B_k^{(2)} - A_k^{(2)} &= \sup_x \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \inf_x \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] \\
&\leqslant \sup_{x,y} \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = y] \\
&= \sup_{x,y} \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S^{-k})I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] \\
&+ \mathbb{E}[L(S^{-k})I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S^{-k})I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = y] \\
&+ \mathbb{E}[L(S^{-k})I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = y] - \mathbb{E}[L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = y].
\end{aligned}$$

By the boundedness conditions that $|\beta(\cdot,\cdot)| \leqslant M_\beta$, $0 \leqslant l(\cdot,\cdot) \leqslant M_l$

$$\begin{aligned}
&\mathbb{E}[L(S)I_{E_{-k}^c} - L(S^{-k})I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = x] + \mathbb{E}[L(S^{-k})I_{E_{-k}^c} - L(S)I_{E_{-k}^c}|z_1, \cdots, z_{k-1}, z_k = y] \\
&\leqslant (2M_\beta + M_l)\mathbb{P}(E_{-k}^c|z_1, \cdots, z_{k-1}).
\end{aligned}$$

In addition,

$$\mathbb{E}[L(S^{-k})I_{E^c_{-k}}|z_1,\cdots,z_{k-1},z_k=x]-\mathbb{E}[L(S^{-k})I_{E^c_{-k}}|z_1,\cdots,z_{k-1},z_k=y]=0.$$

As a result,

$$B_k^{(2)}-A_k^{(2)}\leqslant(2M_\beta+M_l)\mathbb{P}(E^c_{-k}|z_1,\cdots,z_{k-1})$$

We further use M to denote $2M_\beta+M_l$ and $P_k(z_{1:k-1})$ to denote $\mathbb{P}(E^c_{-k}|z_1,\cdots,z_{k-1})$. Now, by Hoeffding's lemma,

$$\begin{aligned}\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}]&=\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}\mathbb{E}[e^{2\lambda D_m^{(2)}}|\mathcal{F}_{m-1}]\Big]\\&\leqslant\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2P_m^2(z_{1:m-1})}\Big]\end{aligned}$$

Suppose for some constant τ_m (see Lemma A.3 for exact value of τ_m)

$$\sup_k\mathbb{P}(E^c_{-k})\leqslant\tau_m,$$

then for all $k=1,\ldots,m$

$$\mathbb{P}(P_k(z_{1:k-1})\geqslant c)\leqslant\frac{\tau_m}{c}.$$

Then

$$\begin{aligned}\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2P_m(z_{1:m-1})}\Big]&=\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2P_m^2(z_{1:m-1})}I_{\{P_m(z_{1:m-1})\geqslant c\}}\Big]\\&+\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2P_m^2(z_{1:m-1})}I_{\{P_m(z_{1:m-1})<c\}}\Big]\\&\leqslant\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2}I_{\{P_m(z_{1:m-1})\geqslant c\}}\Big]\\&+\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2c^2}I_{\{P_m(z_{1:m-1})<c\}}\Big].\end{aligned}$$

Now we further bound the two terms on the righthand side of the above inequality with the following lemmas.

We first consider bounding $\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2}I_{\{P_m(z_{1:m-1})\geqslant c\}}\Big]$.

Lemma A.2. *For any fixed $\lambda>0$, for any $k=1,\ldots,m$, we have*

$$\mathbb{E}[e^{2\lambda\sum_{i=1}^{k-1} D_i^{(2)}}I_{\{P_k(z_{1:k-1})\geqslant c\}}]\leqslant e^{2M(k-1)\lambda}\mathbb{P}(P_k(z_{1:k-1})\geqslant c)\leqslant e^{2M(k-1)\lambda}\frac{\tau_m}{c}.$$

Proof. Consider

$$\begin{aligned}\mathbb{E}[e^{2\lambda D_{m-1}^{(2)}}I_{\{P_m(z_{1:m-1})\geqslant c\}}|\mathcal{F}_{m-2}]&=\mathbb{E}[e^{2\lambda(D_{m-1}^{(2)}-\mathbb{E}[\tilde{D}_{m-1}^{(2)}|\mathcal{F}_{m-2}])}I_{\{P_m(z_{1:m-1})\geqslant c\}}|\mathcal{F}_{m-2}]\\&\leqslant\mathbb{E}[e^{2\lambda(D_{m-1}^{(2)}-\tilde{D}_{m-2}^{(2)})}I_{\{P_m(z_{1:m-1})\geqslant c\}}|\mathcal{F}_{m-2}]\ \ (\star)\\&\leqslant e^{2\lambda M}\mathbb{E}_{z_{m-1}}[I_{\{P_m(\mathcal{F}_{m-2},z_{m-1})\geqslant c\}}]\end{aligned}$$

In $(\star)$, $\tilde{D}_{m-1}^{(2)}|\mathcal{F}_{m-2}$ is an independent copy of $D_{m-1}^{(2)}|\mathcal{F}_{m-2}$, and the inequality is due to the application of Jensen's inequality. Notice that

$$\mathbb{E}\Big[e^{2\lambda D_{m-2}^{(2)}}\mathbb{E}_{z_{m-1}}[I_{\{P_m(\mathcal{F}_{m-2},z_{m-1})\geqslant c\}}]|\mathcal{F}_{m-3}\Big]\leqslant e^{2\lambda M}\mathbb{E}_{z_{m-2},z_{m-1}}[I_{\{P_m(\mathcal{F}_{m-3},z_{m-2},z_{m-1})\geqslant c\}}].$$

Iteratively, we obtain that

$$\mathbb{E}[e^{2\lambda\sum_{i=1}^{k-1} D_i^{(2)}}I_{\{P_k(z_{1:k-1})\geqslant c\}}]\leqslant e^{2M(k-1)\lambda}\mathbb{P}(P_k(z_{1:k-1})\geqslant c).$$

Similar argument can be obtained for any $k=1,\ldots,m$. □

With the help of Lemma A.2, we can obtain an upper bound for $\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}e^{\frac{1}{2}\lambda^2M^2P_m(z_{1:m-1})}\Big]$.

Lemma A.3. *For any fixed $\lambda > 0$, $c < 1$,*

$$\mathbb{E}\big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}\big] \leqslant e^{2m\lambda^2 M^2 c^2} + m\frac{\tau_m}{c} e^{2m\lambda M \max\{1,\lambda M\}}.$$

Proof. Note that

$$\begin{aligned}\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})} e^{\frac{1}{2}\lambda^2 M^2 P_m(z_{1:m-1})}\Big] \leqslant\ & \mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})} e^{\frac{1}{2}\lambda^2 M^2} I_{\{P_m(z_{1:m-1})\geqslant c\}}\Big] \\ & + \mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})} e^{\frac{1}{2}\lambda^2 M^2 c^2} I_{\{P_m(z_{1:m-1})< c\}}\Big].\end{aligned}$$

We do a decomposition as the following

$$1 = I_{\{P_k(z_{1:k-1})<c\}} + I_{\{P_k(z_{1:k-1})\geqslant c\}}$$

for the second term $\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})} e^{\frac{1}{2}\lambda^2 M^2 c^2} I_{\{P_m(z_{1:m-1})<c\}}\Big]$ sequentially until $I_{\{P_{k+1}(z_{1:k})\geqslant c\}}$ appears for some $k = 1, \ldots, m$, *i.e.*,

$$\begin{aligned} I_{\{P_{k+1}(z_{1:k})<c\}} &= I_{\{P_{k+1}(z_{1:k})<c\}}(I_{\{P_k(z_{1:k-1})\geqslant c\}} + I_{\{P_k(z_{1:k-1})<c\}}) \\ &= I_{\{P_{k+1}(z_{1:k})<c\}} I_{\{P_k(z_{1:k-1})\geqslant c\}} + I_{\{P_{k+1}(z_{1:k})<c\}} I_{\{P_k(z_{1:k-1})<c\}}(I_{\{P_{k-1}(z_{1:k-2})\geqslant c\}} \\ &\quad + I_{\{P_{k-1}(z_{1:k-2})<c\}}) \\ &= \cdots. \end{aligned}$$

Besides the term,

$$\mathbb{E}\Big[e^{2\lambda(\sum_{k=1}^{m} D_k^{(2)})} \Pi_{j=1}^m I_{\{P_j(z_{1:j-1})<c\}}\Big],$$

which is bounded by

$$e^{\frac{1}{2}m\lambda^2 M^2 c^2},$$

doing such a decomposition will provide extra m terms and the sum of them can be bounded by

$$m\frac{\tau_m}{c} e^{2m\lambda M \max\{1,\lambda M\}}$$

by applying Lemma A.2.

Taking the sum of them would yield the results. □

Last, we provide a bound for τ_m.

Lemma A.4. *Recall*

$$E_{-k} = \Big\{S \Big| \sup_{z'\in\mathcal{Z}} \Big|\sum_{j\neq k} \frac{\beta(z', z_j)}{m} - \mathbb{E}_z \beta(z', z)\Big| \leqslant \eta\Big\}.$$

for $\eta > 2M_\beta/m$, we have

$$\sup_k \mathbb{P}(E_{-k}^c) \leqslant \tau_m,$$

where $\tau_m = C\exp(-\frac{m\eta^2}{32M_\beta^2})$ for a constant $C > 0$.

Proof. Notice that if $\eta > 2M_\beta/m$,

$$\cup_k E_{-k}^c \subseteq \Big\{S \Big| \sup_{z'\in\mathcal{Z}} \Big|\sum_{j=1}^{m} \frac{\beta(z, z_j)}{m} - \mathbb{E}_z \beta(z', z)\Big| \geqslant \eta/2\Big\}.$$

Thus,

$$\sup_k \mathbb{P}(E_{-k}^c) \leqslant \mathbb{P}\left(\Big\{S \Big| \sup_{z'\in\mathcal{Z}} \Big|\sum_{j=1}^{m} \frac{\beta(z', z_j)}{m} - \mathbb{E}_z \beta(z', z)\Big| \geqslant \eta/2\Big\}\right).$$

Recall the L-Lipschitz property on the first argument of $\beta(\cdot,\cdot)$, by the standard epsilon net-argument (Wainwright, 2019), and choose $\varepsilon = \eta/(6L)$, we can first obtain via uniform bound such that on a finite $\eta/(6L)$-net of $\mathcal{Z}$, which we define as $\mathcal{Z}_{\mathcal{N}}$, with high probability,

$$\sup_{z' \in \mathcal{Z}_{\mathcal{N}}} \Big| \sum_{j=1}^{m} \frac{\beta(z', z_j)}{m} - \mathbb{E}_z \beta(z', z) \Big| \geqslant \eta/6.$$

Then, by Lipschitz condition, for any z_a, z_b in each cell, *i.e.* $|z_a - z_b| \leq \eta/(6L)$, we have

$$\Big| \sum_{j=1}^{m} \frac{\beta(z_a, z_j)}{m} - \mathbb{E}_z \beta(z_a, z) \Big| - \Big| \sum_{j=1}^{m} \frac{\beta(z_b, z_j)}{m} - \mathbb{E}_z \beta(z_b, z) \Big| \leq \eta/3.$$

By combining the above two steps, we have the uniform bound on $\mathcal{Z}$.

Specifically, we have

$$\mathbb{P}\left(\Big\{ S \Big| \sup_{z' \in \mathcal{Z}} \Big| \sum_{j=1}^{m} \frac{\beta(z', z_j)}{m} - \mathbb{E}_z \beta(z', z) \Big| \geqslant \eta/2 \Big\} \right) = C \exp(-\frac{m\eta^2}{32 M_\beta^2}),$$

where

$$C = \exp(\tilde{C} d \log(Ld/\eta))$$

for a universal constant $\tilde{C}$ and d is the dimension of z, L is the Lipschitz constant of the first variable in $\beta(\cdot,\cdot)$ function in assumption. □

Corollary A.1. *For $\lambda > 0$, $c < 1$, we have*

$$\mathbb{E}\big[e^{2\lambda(\sum_{k=1}^{m-1} D_k^{(2)})}\big] \leqslant e^{\frac{1}{2} m\lambda^2 M^2 c^2} + \frac{m}{c} C \exp(-\frac{m\eta^2}{32 M_\beta^2}) e^{2m\lambda M \max\{1, \lambda M\}}$$

for a constant C that depends on L and d.

Proof. Combining the results of Lemma A.3 and A.4, we obtain the bound. □

Bounding $\mathbb{E}[e^{2\lambda(\sum_{k=1}^{m} D_k^{(1)})}]$. Now, we consider bounding

$$\mathbb{E}[e^{2\lambda(\sum_{k=1}^{m} D_k^{(1)})}].$$

We further define

$$A_k^{(1)} = \inf_x \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}],$$

$$B_k^{(1)} = \sup_x \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}].$$

Again we have,

$$A_k^{(1)} \leqslant D_k^{(1)} \leqslant B_k^{(1)}.$$

Similarly

$$\begin{aligned} B_k^{(1)} - A_k^{(1)} \leqslant & \sup_{x,y} \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S^{-k}) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = x] \\ & + \mathbb{E}[L(S^{-k}) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = x] - \mathbb{E}[L(S^{-k}) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = y] \\ & + \mathbb{E}[L(S^{-k}) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = y] - \mathbb{E}[L(S) I_{E_{-k}} | z_1, \cdots, z_{k-1}, z_k = y]. \end{aligned}$$

By the nature of E_{-k}, and the boundedness conditions that $|\beta(\cdot,\cdot)| \leqslant M_\beta$, $0 \leqslant l(\cdot,\cdot) \leqslant M_l$

$$\begin{aligned}&\mathbb{E}[L(S)I_{E_{-k}} - L(S^{-k})I_{E_{-k}}|z_1,\cdots,z_{k-1},z_k=x] + \mathbb{E}[L(S^{-k})I_{E_{-k}} - L(S)I_{E_{-k}}|z_1,\cdots,z_{k-1},z_k=y]\\ &\leqslant 2\sup_{z'\in\mathcal{Z}}\mathbb{E}_z\beta(z',z) + 2\eta + M_l.\end{aligned}$$

Besides,

$$\mathbb{E}[L(S^{-k})I_{E_{-k}}|z_1,\cdots,z_{k-1},z_k=x] - \mathbb{E}[L(S^{-k})I_{E_{-k}}|z_1,\cdots,z_{k-1},z_k=y] = 0.$$

As a result,

$$B_k^{(1)} - A_k^{(1)} \leqslant 2\sup_{z'\in\mathcal{Z}}\mathbb{E}_z\beta(z',z) + 2\eta + M_l.$$

Then, by the standard argument of concentration of martingale differences, we have the following lemma.

Lemma A.5. *For any $\lambda \in \mathbb{R}$, if we denote $\tilde{M} = 2\sup_{z'\in\mathcal{Z}}\mathbb{E}_z\beta(z',z) + 2\eta + M_l$*

$$\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(1)})}] \leqslant e^{\frac{1}{2}m\lambda^2\tilde{M}^2}.$$

Combined the previous results, we provide the following lemma.

Lemma A.6. *For $\lambda > 0, c > 0$, there exists a constant $C > 0$ depending on η, d and L*

$$\mathbb{E}[e^{\lambda(\sum_{k=1}^m D_k)}] \leqslant \frac{1}{2}e^{\frac{1}{2}m\lambda^2\tilde{M}^2} + \frac{1}{2}\big(e^{\frac{1}{2}m\lambda^2M^2c^2} + Cm\frac{\tau_m}{c}e^{2m\lambda M\max\{1,\lambda M\}}\big).$$

Proof. Combining Lemma A.5 and Corollary A.1, and the fact that

$$\mathbb{E}[e^{\lambda(\sum_{k=1}^m D_k)}] \leqslant \frac{1}{2}\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(1)})}] + \frac{1}{2}\mathbb{E}[e^{2\lambda(\sum_{k=1}^m D_k^{(2)})}],$$

we obtain the above bound. □

A Concentration Bound. If we choose $c = \tilde{M}/M'$, where $\tilde{M} = 2\sup_{z\in\mathcal{Z}}\mathbb{E}_{z_j}\beta(z,z_j) + 2\eta + M_l$ and $M' = 2M_\beta + 2\eta + M_l$. It is easy to see $c < 1$. We denote $\gamma_m = Cm\frac{\tau_m}{2c}e^{2m\lambda M\max\{1,\lambda M\}}$. By Chernoff-bound, for any $\lambda > 0$

$$\mathbb{P}\Big(L(S) - \mathbb{E}[L(S)] \geqslant m\varepsilon\Big) \leqslant \frac{e^{\frac{1}{2}m\lambda^2\tilde{M}^2} + \gamma_m}{e^{\lambda m\varepsilon}}.$$

Let us take $\lambda = \varepsilon/\tilde{M}^2$ for $\varepsilon > 0$, when m is large enough and ε is small enough, we expect $\gamma_m \leqslant e^{\frac{1}{2}m\lambda^2\tilde{M}^2}$. Specifically,

$$\frac{Cm\tau_m}{2c} \leqslant \exp(\frac{1}{2}m\lambda(\tilde{M}^2\lambda - 4M\max\{1,\lambda M\})).$$

Recall $\tau_m = C\exp(-\frac{m\eta^2}{32M_\beta^2})$, plugging in $\lambda = \varepsilon/\tilde{M}^2$ and τ_m, let $C' = C/(2c)$ it is sufficient to let m large enough and ε small enough such that

$$C'm\exp(-\frac{m\eta^2}{32M_\beta^2}) \leqslant \exp\left(\frac{m\varepsilon}{2\tilde{M}^2}(\varepsilon - \frac{4\varepsilon M^2}{\tilde{M}^2} - 4M)\right).$$

which can be further simplified as

$$\frac{\eta^2}{32M_\beta^2} - \frac{\log C'm}{m} \geqslant \frac{\varepsilon}{2\tilde{M}^2}(-\varepsilon + \frac{4\varepsilon M^2}{\tilde{M}^2} + 4M).$$

That will lead to

$$\begin{aligned}\mathbb{P}\Big(L(S) - \mathbb{E}[L(S)] \geqslant m\varepsilon\Big) &\leqslant \frac{e^{\frac{1}{2}m\lambda^2\tilde{M}^2} + \gamma_m}{e^{\lambda m\varepsilon}}\\ &\leqslant 2\exp(-\frac{m\varepsilon^2}{2\tilde{M}^2}).\end{aligned}$$

[Proof of Lemma A.1]

Proof. Now let us consider $\tilde{l}(\mathcal{A}_S, z) = -l(\mathcal{A}_S, z) + \mathbb{E}_z[l(\mathcal{A}_S, z)]$. We remark here as long as $0 \leqslant l \leqslant M_l$ (not $\tilde{l}$).

In addition, if we denote

$$\tilde{L}(S) = \sum_{j=1}^{m} \tilde{l}(\mathcal{A}_S, z_j),\ \tilde{L}(S^{-i}) = \sum_{j\neq i}^{m} \tilde{l}(\mathcal{A}_{S^{-i}}, z_j),$$

we have

$$\begin{aligned}&\mathbb{E}[\tilde{L}(S)I_{E^c_{-k}} - \tilde{L}(S^{-k})I_{E^c_{-k}}|z_1, \cdots, z_{k-1}, z_k = x] + \mathbb{E}[\tilde{L}(S^{-k})I_{E^c_{-k}} - \tilde{L}(S)I_{E^c_{-k}}|z_1, \cdots, z_{k-1}, z_k = y] \\ &\leqslant (2M_\beta + 2\sup_{z\in\mathcal{Z}} \mathbb{E}_{z'}\beta(z, z') + 2M_l)\mathbb{P}(E^c_{-k}|z_1, \cdots, z_{k-1}).\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}[\tilde{L}(S)I_{E_{-k}} - \tilde{L}(S^{-k})I_{E_{-k}}|z_1, \cdots, z_{k-1}, z_k = x] + \mathbb{E}[\tilde{L}(S^{-k})I_{E_{-k}} - \tilde{L}(S)I_{E_{-k}}|z_1, \cdots, z_{k-1}, z_k = y] \\ &\leqslant (4\sup_{z\in\mathcal{Z}} \mathbb{E}_{z_j}\beta(z, z_j) + 2\eta + 2M_l)\mathbb{P}(E_{-k}|z_1, \cdots, z_{k-1}).\end{aligned}$$

Thus, the generalization argument is exactly the same as theory above except the value of M, and $\tilde{M}$.

Specifically, we choose $M = 2(M_\beta + \sup_{z'\in\mathcal{Z}} \mathbb{E}_z\beta(z', z) + M_l)$ and $\tilde{M} = 2(2\sup_{z'\in\mathcal{Z}} \mathbb{E}_z\beta(z', z) + \eta + M_l)$.

Next, by Lemma 7 in Bousquet & Elisseeff (2002), for an independent copy of z_j, which we denote it as z'_j, we have

$$\begin{aligned}\mathbb{E}_S\big[-\frac{1}{m}\sum_{j=1}^{m} l(\mathcal{A}_S, z_j) + \mathbb{E}_z[l(\mathcal{A}_S, z)]\big] &\leqslant \mathbb{E}_{S,z'_j}\big[|l(\mathcal{A}_S, z'_j) - l(\mathcal{A}_{S^j}, z'_j)|\big] \\ &\leqslant \frac{2\sup_{z'\in\mathcal{Z}} \mathbb{E}_z\beta(z', z)}{m}.\end{aligned}$$

Then, the result follows. □

A.2. Proof of Lemma 4.1

For simplicity, let us introduce the following two notations:

$$R_r(g) := \frac{1}{m}\sum_{j=1}^{m} l(g, z_j) + \lambda\|g\|_K^2,$$

$$R_r^{-i}(g) := \frac{1}{m}\sum_{j\neq i}^{m} l(g, z_j) + \lambda\|g\|_K^2.$$

Let us denote f as a minimizer of R_r in $\mathcal{F}$ and f^{-i} as a minimizer of R_r^{-i}. We further denote $\Delta f = f^{-i} - f$.

By Lemma 20 in Bousquet & Elisseeff (2002), we have

$$2\|\Delta f\|_K^2 \leq \frac{\sigma}{\lambda m}|\Delta f(x_i)|.$$

Furthermore since

$$|f(x_i)| \leq \|f\|_K\sqrt{K(x_i, x_i)} \leq \|f\|_K\kappa(x_i),$$

we have

$$\|\Delta f\|_K \leq \frac{\kappa(x_i)\sigma}{2\lambda m}.$$

By the σ-admissibility of l,

$$|l(f, z) - l(f^{-i}, z)| \leq \sigma|f(x) - f^{-i}(x)| = \sigma|\Delta f(x)| \leq \sigma\|\Delta f\|_K\kappa(x) \leqslant \frac{\sigma^2\kappa(x)\kappa(x_i)}{2\lambda m}.$$

A.3. Proof of Locally Elastic Stability of SGD

In this section, we establish our new notion of algorithm stability – *locally elastic stability* for SGD. Specifically, we consider the quantity

$$|\mathbb{E}_{\mathcal{A}}[l(\mathcal{A}_S, z)] - \mathbb{E}_{\mathcal{A}}[l(\mathcal{A}_{S^{-i}}, z)]|.$$

Here, the expectation is taken over the internal randomness of $\mathcal{A}$. The randomness comes from the selection of sample at each step of SGD. Specifically, $\mathcal{A}_S$ returns a parameter θ_T, where T is the number of iterations. And for dataset S with sample size m, SGD is performed in the following way:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta l(\theta_t, z_{i_t}),$$

where η_t is the learning rate at time t, i_t is picked uniformly at random in $\{1, \cdots, m\}$.

We denote

$$L(z) = \sup_{\theta \in \Theta} \|\nabla_\theta l(\theta, z)\|$$

We further denote the gradient update rule

$$G_{l,\eta}(\theta, z) = \theta - \eta \nabla_\theta l(\theta, z).$$

We consider gradient updates $G_1, \cdots, G_T$ and $G'_1, \cdots, G'_T$ induced by running SGD on S and S^{-i}. Most of the proofs are similar to Hardt et al. (2015), the only difference is that for dataset S and S^{-i}, the randomness in $\mathcal{A}$ is different. For S, i_t is randomly picked in $\{1, \cdots, m\}$ but for S^{-i}, i_t is randomly picked in $\{1, \cdots, i-1, i+1, \cdots, m\}$. Thus, we create two coupling sequences for the updates on S and S^{-i}. Notice choosing any coupling sequences will not affect the value of $|\mathbb{E}_{\mathcal{A}}[l(\mathcal{A}_S, z)] - \mathbb{E}_{\mathcal{A}}[l(\mathcal{A}_{S^{-i}}, z)]|$, since the expectations are taken with respect to $l(\mathcal{A}_S, z)$ and $l(\mathcal{A}_{S^{-i}}, z)$ separately.

Convex optimization. We first show SGD satisfies locally elastic stability for convex loss minimization.

Proposition A.1 (Restatement of Proposition 2)**.** *Assume that the loss function $l(\cdot, z)$ is α-smooth and convex for all $z \in \mathcal{Z}$. In addition, $l(\cdot, z)$ is $L(z)$-Lipschitz and $L(z) < \infty$ for all $z \in \mathcal{Z}$: $|l(\theta, z) - l(\theta', z)| \leqslant L(z)\|\theta - \theta'\|$ for all θ, θ'. We further assume $L = \sup_{z \in \mathcal{Z}} L(z) < \infty$. Suppose that we run SGD with step sizes $\eta_t \leqslant 2/\alpha$ for T steps. Then,*

$$|\mathbb{E}[l(\hat{\theta}_T, z)] - \mathbb{E}[l(\hat{\theta}_T^{-i}, z)]| \leqslant \frac{(L + L(z_i))L(z)}{m} \sum_{t=1}^{T} \eta_t.$$

Proof. Notice for i_t which is randomly picked in $\{1, \cdots, i-1, i+1, \cdots, m\}$, we can view it as a two-phase process. Firstly, draw i_t uniformly from a n-element set $\{1, \cdots, i-1, i', i+1, \cdots, m\}$. If any element but i' is drawn, directly output it. Otherwise if i' is drawn, then uniformly draw again from $\{1, \cdots, i-1, i+1, \cdots, m\}$, and output the final index that is drawn. It is not hard to notice that in this way, each index in $\{1, \cdots, i-1, i+1, \cdots, m\}$ has probability

$$\frac{1}{m} + \frac{1}{m(m-1)} = \frac{1}{m-1}$$

to be drawn, which is the same as directly uniformly draw from $\{1, \cdots, i-1, i+1, \cdots, m\}$.

We consider two coupling processes of SGD on S and S^{-i}. The randomness of uniformly drawing from n elements for SGD on S and uniformly drawing from $\{1, \cdots, i-1, i', i+1, \cdots, n\}$ for SGD on S^{-i} share the same random seed ξ_t at each iteration at time t. That will not affect the value of

$$|\mathbb{E}_{\mathcal{A}}[l(A_S, z)] - \mathbb{E}_{\mathcal{A}}[l(A_{S^{-i}}, z)]|.$$

Let $\delta_t = \|\hat{\theta}_t - \hat{\theta'}_t\|$, where $\hat{\theta}_t$ is the parameter obtained by SGD on S at iteration t and $\hat{\theta'}_t$ is the parameter by SGD on S^{-i} obtained at iteration t.

With probability $1/n$ the selected example is different, in that case we use the fact that

$$\delta_{t+1} \leqslant \delta_t + \eta_t L(z_i) + \eta_t L(z_j)$$

for some $j \neq i$, which can further be upper bounded by $\eta_t(L(z_i) + L)$. With probability $1 - 1/n$, the selected example is the same, then we can apply Lemma 3.7 in Hardt et al. (2015) regarding 1-expansivity of the update rule G_t, then we have

$$\mathbb{E}[\delta_{t+1}] \leqslant \left(1 - \frac{1}{m}\right) \mathbb{E}[\delta_t] + \frac{1}{m}\mathbb{E}[\delta_t] + \frac{\eta_t(L(z_i) + L)}{m}.$$

This technique is used repeatedly in the following other theorems, we will not further elaborate it.

Then, unraveling the recursion, by the fact that $L(z)$ continuity of $l(\theta, z)$ for any θ, we obtain

$$|\mathbb{E}[l(\hat{\theta}_T, z)] - \mathbb{E}[l(\hat{\theta}_T^{-i}, z)]| \leqslant \frac{(L + L(z_i))L(z)}{m} \sum_{t=1}^{T} \eta_t.$$

□

Strongly convex optimization. We consider the penalized loss discussed in Hardt et al. (2015):

$$\frac{1}{n}\sum_{i=1}^{n} l(\theta, z_i) + \frac{\mu}{2}\|\theta\|_2^2,$$

where $l(\theta, z)$ is convex with respect to θ for all z. And without loss of generality, we assume Θ is a ball with radius r (this can be obtained by the boundedness of loss l) and apply stochastic projected gradient descent:

$$\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t \nabla \tilde{l}(\theta_t, z_{i_t}))$$

where $\tilde{l}(\theta, z) = l(\theta, z) + \frac{\mu}{2}\|\theta\|_2^2$.

Proposition A.2 (Strongly Convex Optimization)**.** *Assume that the loss function $l(\cdot, z)$ is α-smooth and μ-strongly convex for all $z \in \mathcal{Z}$. In addition, $l(\cdot, z)$ is $L(z)$-Lipschitz and $L(z) < \infty$ for all $z \in \mathcal{Z}$: $|l(\theta, z) - l(\theta', z)| \leqslant L(z)\|\theta - \theta'\|$ for all θ, θ'. We further assume $L = \sup_{z\in\mathcal{Z}} L(z) < \infty$. Suppose that we run SGD with step sizes $\eta \leqslant 1/\alpha$ for T steps. Then,*

$$\mathbb{E}|\tilde{l}(\hat{\theta}_T; z) - \tilde{l}(\hat{\theta'}_T; z)| \leqslant \frac{(L(z_i) + L)L(z)}{m\mu}$$

Proof. By using the same coupling method, we have when the learning rate $\eta\mu \leqslant 1$, By further similarly applying Lemma 3.7 and method in the Theorem 3.9 in Hardt et al. (2015), we have

$$\mathbb{E}[\delta_{t+1}] \leqslant \left(1 - \frac{1}{m}\right)(1 - \eta\mu)\mathbb{E}[\delta_t] + \frac{1}{m}(1 - \eta\mu)\mathbb{E}[\delta_t] + \frac{\eta(L(z_i) + L)}{m}.$$

Unraveling the recursion gives,

$$\mathbb{E}[\delta_T] \leqslant \frac{L(z_i) + L}{m}\sum_{t=0}^{T}(1 - \eta\mu)^t \leqslant \frac{L(z_i) + L}{m\mu}.$$

Plugging the above inequality, we obtain

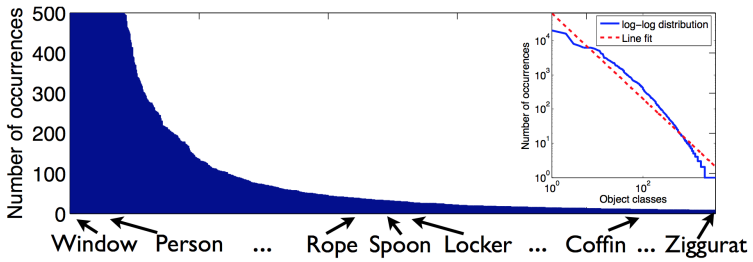
$$\mathbb{E}|\tilde{l}(\hat{\theta}_T; z) - \tilde{l}(\hat{\theta'}_T; z)| \leqslant \frac{(L(z_i) + L)L(z)}{m\mu}$$

□

Non-convex optimization Lastly, we show the case of non-convex optimization.

Proposition A.3 (Restatement of Proposition 3)**.** *Assume that the loss function $l(\cdot, z)$ is non-negative and bounded for all $z \in \mathcal{Z}$. Without loss of generality, we assume $l(\cdot, z) \in [0, 1]$. In addition, we assume $l(\cdot, z)$ is α-smooth and convex for all $z \in \mathcal{Z}$. We further assume $l(\cdot, z)$ is $L(z)$-Lipschitz and $L(z) < \infty$ for all $z \in \mathcal{Z}$ and $L = \sup_{z\in\mathcal{Z}} L(z) < \infty$. Suppose that we run SGD for T steps with monotonically non-increasing step sizes $\eta_t \leqslant c/t$ for some constant $c > 0$. Then,*

$$|\mathbb{E}[l(\hat{\theta}_T, z)] - \mathbb{E}[l(\hat{\theta}_T^{-i}, z)]| \leqslant \frac{1 + 1/(\alpha c)}{m - 1}[c(L(z_i) + L)L(z)]^{\frac{1}{\alpha c+1}} T^{\frac{\alpha c}{\alpha c+1}}.$$



(a) The number of examples by object class in SUN dataset

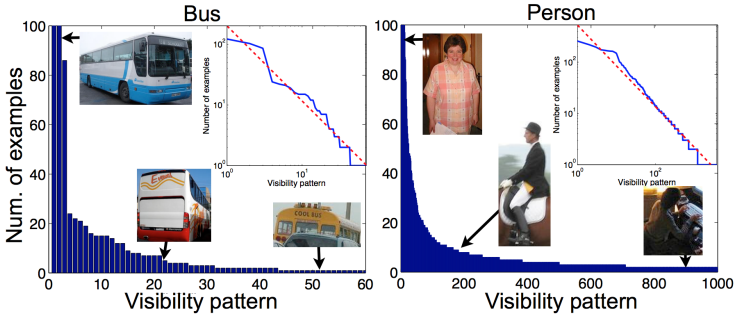


Figure 3. Long tail empirical distribution of classes and subpopulations within classes, taken from (Zhu et al., 2014) with the authors' permission.

Proof. By Lemma 3.11 in Hardt et al. (2015), for every $t_0 \in \{1, \cdots, T\}$ and switch L to $L(z)$, we have

$$|\mathbb{E}[l(\hat{\theta}_T, z)] - \mathbb{E}[l(\hat{\theta}_T^{-i}, z)]| \leqslant \frac{t_0}{m} + L(z)\mathbb{E}[\delta_T | \delta_{t_0} = 0].$$

Let $\Delta_t = \mathbb{E}[\delta_t | \delta_{t_0} = 0]$. By applying Lemma 3.7 and method in the Theorem 3.12 in Hardt et al. (2015), combining the fact regarding boundedness of the gradient – $\|\nabla_\theta l(\hat{\theta}_t, z_i)\| \leqslant L(z_i)$, $\sup_{j \neq i} \|\nabla_\theta l(\hat{\theta}_t, z_j)\| \leqslant L$, we have

$$\begin{aligned}
\Delta_{t+1} &\leqslant \left(1 - \frac{1}{m}\right)(1 + \eta_t \alpha)\Delta_t + \frac{1}{m}\Delta_t + \frac{\eta_t(L + L(z_i))}{m} \\
&\leqslant \left(\frac{1}{m} + (1 - 1/m)(1 + c\alpha/t)\right)\Delta_t + \frac{c(L + L(z_i))}{tm} \\
&= \left(1 + (1 - 1/m)\frac{c\alpha}{t}\right)\Delta_t + \frac{c(L + L(z_i))}{tm} \\
&\leqslant \exp((1 - 1/m)\frac{c\alpha}{t})\Delta_t + \frac{c(L + L(z_i))}{tm}
\end{aligned}$$

By the fact $\Delta_0 = 0$, we can unwind this recurrence relation from T down to $t_0 + 1$, it is easy to obtain

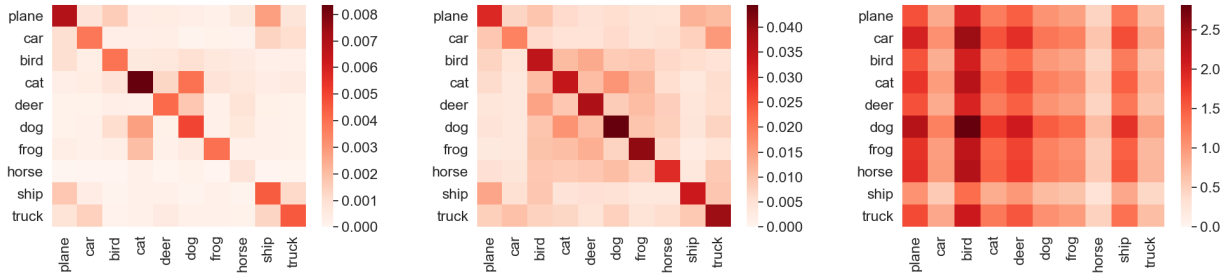
$$|\mathbb{E}[l(\hat{\theta}_T, z)] - \mathbb{E}[l(\hat{\theta}_T^{-i}, z)]| \leqslant \frac{1 + 1/(\alpha c)}{m - 1}[c(L(z_i) + L)L(z)]^{\frac{1}{\alpha c + 1}} T^{\frac{\alpha c}{\alpha c + 1}}.$$

□

B. More about Experiments

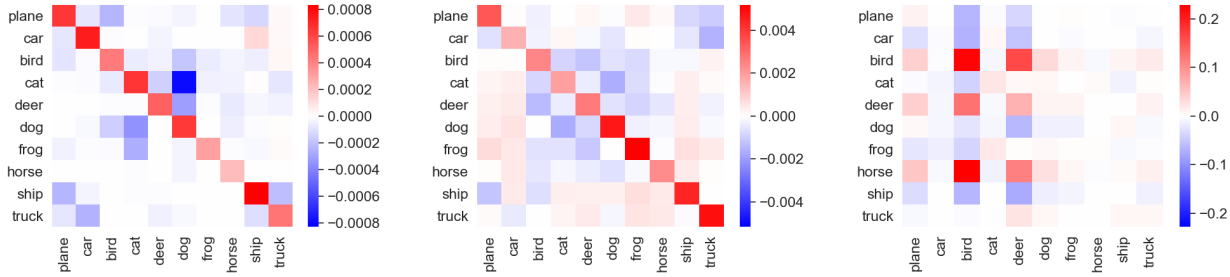
To verify the effectiveness of our proposed locally elastic stability, we conduct experiments on the real-world CIFAR-10 dataset. In our experiments, we randomly choose 100 examples per image class (1000 examples in total) for both training and test data. For the neural network model, we consider an 18-layer ResNet and use its pytorch implementation[3] for

[3]More details are in https://pytorch.org/docs/stable/torchvision/models.html.



(a) Sensitivity of neural networks. (b) Sensitivity of a random feature model. (c) Sensitivity of a linear model.

Figure 4. Class-level sensitivity approximated by influence functions for neural networks (based on a pre-trained 18-layer ResNet), a random feature model (based on a randomly initialized 18-layer ResNet), and a linear model on CIFAR-10. The vertical axis denotes the classes in the test data and the horizontal axis denotes the classes in the training data. The class-level sensitivity from class a in the training data to class b in the test data is defined as $C(c_a, c_b) = \frac{1}{|S_a| \times |\tilde{S}_b|} \sum_{z_i \in S_a} \sum_{z \in \tilde{S}_b} |l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$, where S_a denotes the set of examples from class a in the training data and $\tilde{S}_b$ denotes set of examples from class b in the test data.



(a) Influence for neural networks. (b) Influence for the random feature model. (c) Influence for the linear model.

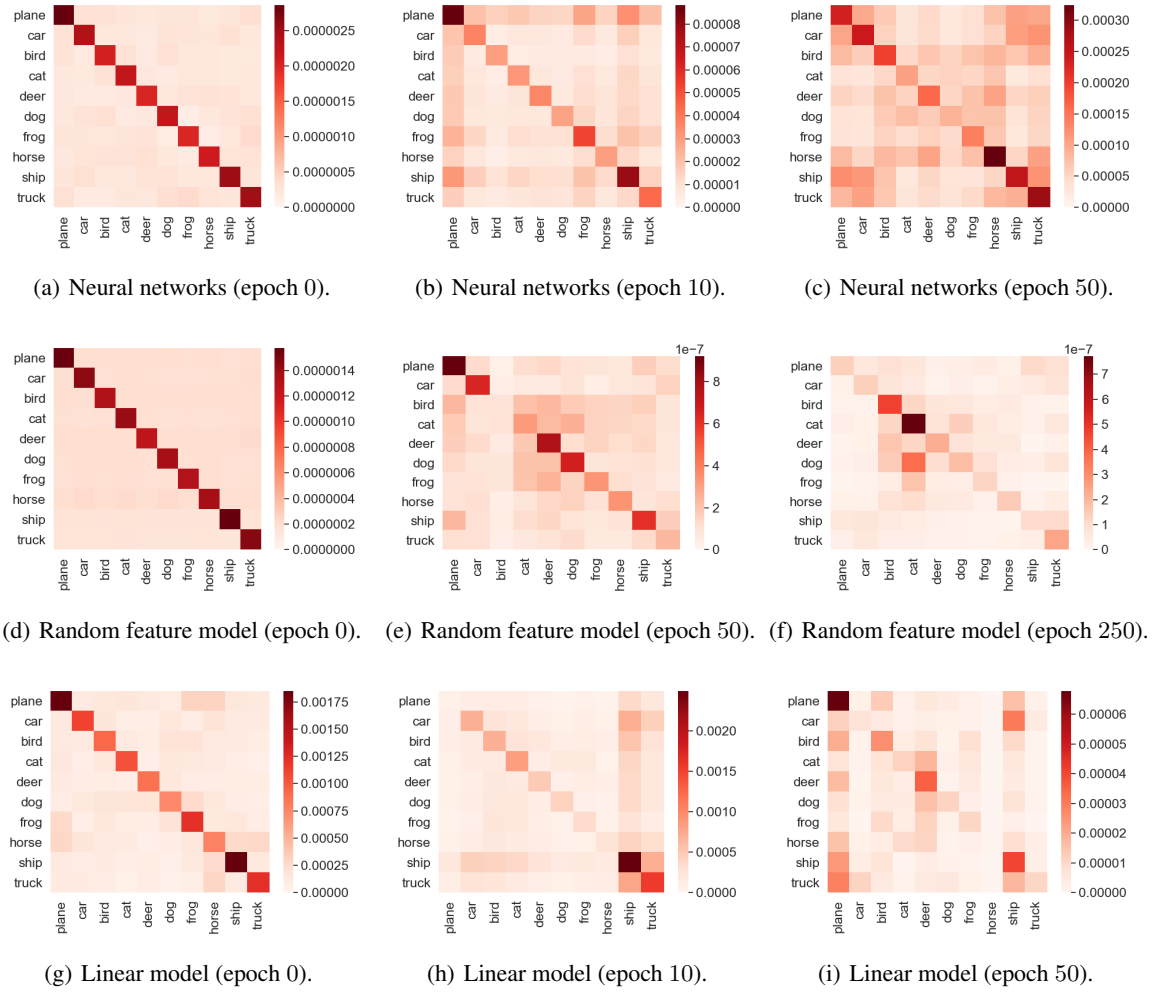
Figure 5. Class-level sensitivity approximated by influence function for neural networks (based on a pre-trained 18-layer ResNet), a random feature model (based on a randomly initialized 18-layer ResNet), and a linear model on CIFAR-10. Note that the sensitivity here is based on sign values ($l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)$) instead of absolute values ($|l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$) as in Eq. (2).

our experiments. For the random feature model, we use the same 18-layer ResNet (with randomly initialized weights) to extract random features and only train the last layer. As for the loss function, we use the cross-entropy loss for linear models, random feature models, and neural networks. Furthermore, we analyze locally elastic stability in two settings: locally elastic stability for the whole algorithm and locally elastic stability for a step-wise update of SGD.

Locally elastic stability via influence functions. As shown in Sec. 2.1, we use influence functions to estimate the quantity $|l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$ for all i's in Eq. (2). Similar to Koh & Liang (2017), we compared the ResNet-18 with all but the top layer frozen[4], and a random feature model based on a randomly initialized ResNet-18 to a linear model in our experiments. In the experiments for locally elastic stability via influence functions, we add the ℓ_2 regularization ($\frac{\lambda \|\theta\|^2}{2}$) with $\lambda = 1e^{-7}$. We train the last layer (randomly initialized) of the ResNet-18, the random feature model, and the linear model using Adam (Kingma & Ba, 2015) with learning rate $3e^{-4}$ for 50 epochs, learning rate 1.0 for 500 epochs[5], and learning rate $3e^{-4}$ for 60 epochs each, and the mini-batch sizes are 50, 20, 50 respectively. The training accuracy for the ResNet-18, the random feature model, and the linear model is 99.3%, 94.7%, and 94.7%, and the test accuracy for them is 93.1%, 29.8%, and 27.3%. The class-level sensitivity approximated by influence function for the neural networks, the random feature model, and the linear model on CIFAR-10 is shown in Fig. 4. Furthermore, we also consider the influence based on sign values,

[4]We pre-train the model on the whole CIFAR-10 dataset first and keep the pre-trained weights.

[5]It is worthwhile to note that it is hard for the random feature model, especially based on large neural networks, to converge. For the random feature model, we also use a widely-used learning rate decay, where the initial learning rate is annealed by a factor of 10 at $1/3$ and $2/3$ during training.



(a) Neural networks (epoch 0). (b) Neural networks (epoch 10). (c) Neural networks (epoch 50).

(d) Random feature model (epoch 0). (e) Random feature model (epoch 50). (f) Random feature model (epoch 250).

(g) Linear model (epoch 0). (h) Linear model (epoch 10). (i) Linear model (epoch 50).

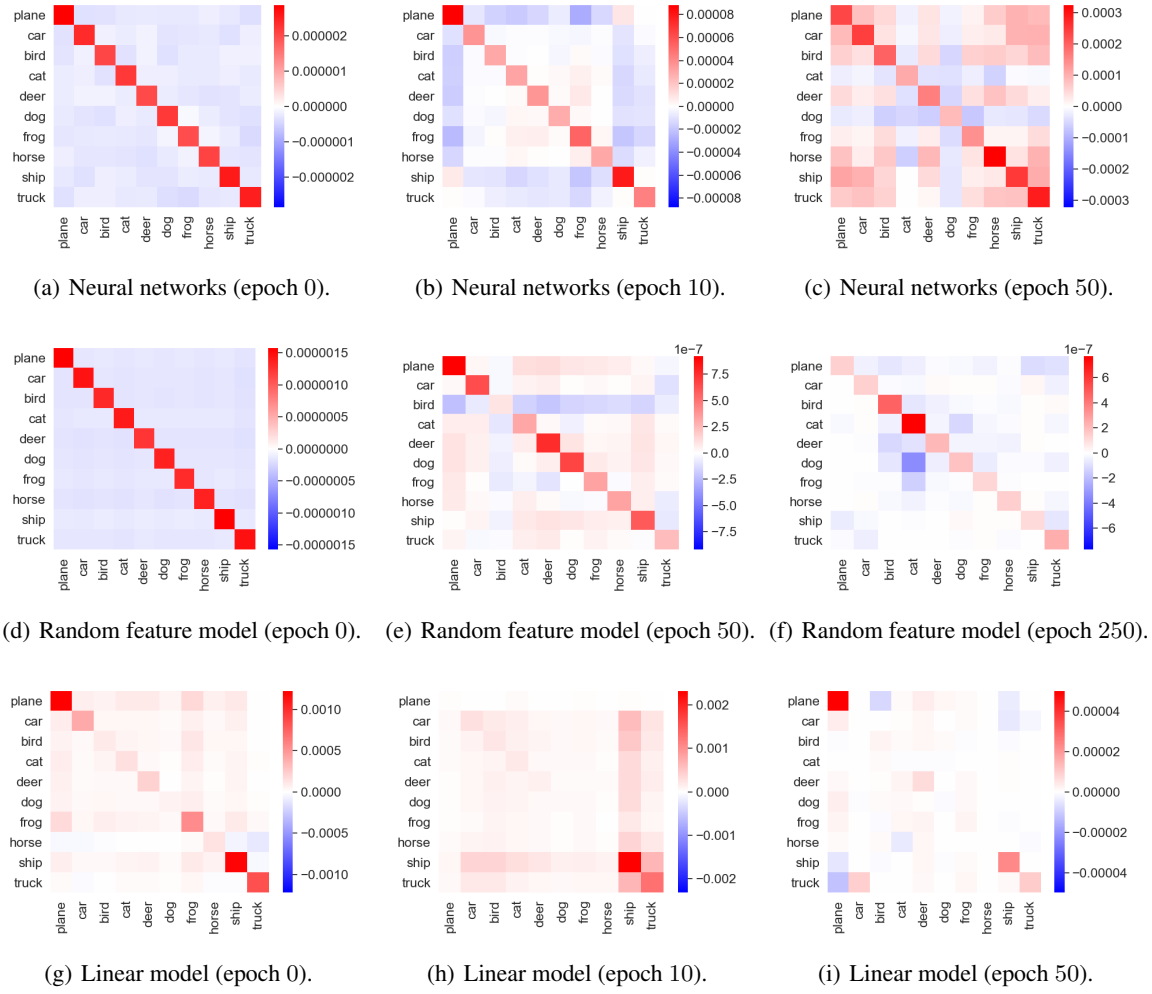
Figure 6. Exact stepwise characterization of class-level sensitivity for neural networks, random feature models, and linear models trained with different numbers of epochs by SGD on CIFAR-10. The class-level sensitivity for a stepwise update of SGD is $C'(c_a, c_b) = \frac{1}{|S_a| \cdot |\tilde{S}_b|} \sum_{z_i \in S_a} \sum_{z \in \tilde{S}_b} |l(\hat{\theta}_t - \eta \nabla_\theta l(\hat{\theta}_t, z_i), z) - l(\hat{\theta}_t, z)|$, where S_a denotes the set of examples with class a in the training data and $\tilde{S}_b$ denotes the set of examples with class b in the test data.

$l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)$ instead of absolute values $|l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$ in Eq. (2), and the corresponding class-level sensitivity is shown in Fig. 5.

Stepwise characterization of locally elastic stability. To provide the stepwise characterization of locally elastic stability, we consider the trained parameters of SGD with different number of training epochs. Note that we didn't make any approximation for the experiments in this part. In the training stage, we train the ResNet-18[6], the random feature model, and the linear model using SGD with learning rate 0.05, 1.0, and 0.3 separately, and the mini-batch sizes are 50, 20, 50 respectively. The training accuracy (test accuracy) for the ResNet-18 at epoch 0, 10, 50, 100 are 10.3% (10.6%), 22% (20.2%), 37.6% (24.3%), and 99.9% (38.9%). Similarly, the training accuracy (test accuracy) for the random feature model[7] at epoch 0, 50, 250 are 10.3% (10.6%), 63.1% (28.0%), and 90.2% (30.1%). Similarly, the training accuracy (test accuracy) for the linear model at epoch 0, 10, 50, 100 are 9.6% (7.6%), 53.4% (20.5%), 98.2% (22.7%), and 100% (23%). To compute the class-level sensitivity $C(c_a, c_b)$, we use the small probing learning rate $1e^{-6}$. The corresponding

[6]Note that we remove the batch normalization for the experiments in step-wise characterization of locally elastic stability for SGD (only in this part).

[7]Because it is hard for the random feature model to converge, we use the Adam optimizer and a widely-used learning rate decay for the random feature model, where the initial learning rate is annealed by a factor of 10 at $1/3$ and $2/3$ during training.



(a) Neural networks (epoch 0). (b) Neural networks (epoch 10). (c) Neural networks (epoch 50).

(d) Random feature model (epoch 0). (e) Random feature model (epoch 50). (f) Random feature model (epoch 250).

(g) Linear model (epoch 0). (h) Linear model (epoch 10). (i) Linear model (epoch 50).

Figure 7. Exact step-wise characterization of class-level sensitivity for neural networks, random feature models, and linear models trained with different numbers of epochs by SGD on CIFAR-10. Note that the sensitivity here is based on sign values ($l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)$) instead of absolute values ($|l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$) as in Eq. (2).

class-level sensitivity based on absolute values($|l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)|$) and sign values ($l(\hat{\theta}, z) - l(\hat{\theta}^{-i}, z)$) are shown in Fig. 6 and Fig. 7.

Comparison among M_β, $\sup_{z' \in \mathcal{Z}} \mathbb{E}_z \beta(z', z)$ and M_l on a 2-layer NNs. We consider a 2-layer NNs in the following format:

$$f(W, a, x) = \frac{1}{k} \sum_{r=1}^{k} a_r \sigma(W_r^T x)$$

where d is the input dimension, k is the dimension of the hidden layer, and σ is the ReLU activation function. As for the loss function, we use the square loss with ℓ_2 regularization as follows:

$$L(W, a, x) = (f(W, a, x) - y)^2 + \frac{\lambda}{2}(\|W\|_2^2 + \|a\|_2^2)$$

In our experiments, the value of each dimension of W, a, x is in $[-1, 1]$, and the value of y is in $\{-1, 1\}$. As for the data distribution, each dimension of x is sampled from a uniform distribution on $[-0.5, 1]$ for positive samples with label $y = 1$. Similarly, each dimension of x is sampled from a uniform distribution on $[-1.0, 0.5]$ for negative samples with label $y = -1$. We randomly sample a total of $m = 10000$ examples equally from positive and negative data distribution

Positive Examples	Cat, Dog	Deer, Horse	Deer, Frog	Car, Cat	Plane, Cat
Negative Examples	Car, Truck	Car, Truck	Ship, Truck	Plane, Bird	Car, Truck
Between Finer Classes (sign)	0.03	0.03	0.02	0.06	0.08
Within Finer Classes (sign)	0.05	0.09	0.08	0.24	0.15
Between Finer Classes (absolute)	1.52	1.90	2.05	4.50	3.19
Within Finer Classes (absolute)	1.53	1.92	2.09	4.66	3.23

Table 2. A fine-grained analysis of the sensitivity within superclasses (within fine-grained classes or between fine-grained classes) for binary classification.

for both training and test data. As for the hyper parameters, we use $d = 10$, $k = 100$, and $\lambda = 1e^{-6}$ in our experiments. We trained the 2-layer NNs 50 epochs with SGD on batches. The corresponding learning rate and batch size are 1.0 and 100.

In this setting, the upper bound of the loss function M_l is 121.00055 and $M_\beta = m\beta_m^U = \sup_{z' \in S, z \in \mathcal{Z}} \beta(z', z)$ estimated by the influence function as shown in Eq. (2) is 3464.97. We can see that M_β is about 29 times of M_l. Similarly, $\sup_{z' \in \mathcal{Z}} \mathbb{E}_z \beta(z', z)$ estimated by the influence function is 22.91. It indicates that $\sup_{z' \in \mathcal{Z}} \mathbb{E}_z \beta(z', z)$ in the locally elastic stability is smaller than M_l and much smaller than $M_\beta = m\beta_m^U = \sup_{z_j \in S, z \in \mathcal{Z}} \beta(z, z_j)$.

Class-level locally elastic stability. In this part, we consider the case where the sensitivity from one class to another class is the maximum sensitivity instead of the mean sensitivity (in Fig. 4) among the 100×100 pairs. In this setting, we have $\sup_{z_j \in S, z \in \mathcal{Z}} \beta_m(z, z_j) = 3.05$ and $\sup_{z \in \mathcal{Z}} \mathbb{E}_{z_j} \beta_m(z, z_j) = 0.73$ for the neural networks, and $\sup_{z_j \in S, z \in \mathcal{Z}} \beta_m(z, z_j) = 314$, $\sup_{z \in \mathcal{Z}} \mathbb{E}_{z_j} \beta_m(z, z_j) = 210$ for the linear model. Furthermore, the maximum and the mean of the diagonal (off-diagonal) elements are 3.05 and 1.02 (1.65 and 0.21) for the neural networks in this setting. Similarly, the maximum and the mean of the diagonal (off-diagonal) elements are 168 and 77 (314 and 82) for the linear model.

Fine-grained analysis. To better understand the locally elastic stability, we also provide some fine-grained analysis on the CIFAR-10 dataset. In this part, we consider binary classification on two superclasses, and each superclass is composed of two fine-grained classes. In the training data, we randomly sample 500 examples for each fine-grained class (2000 examples in total). In the test data, we have 1000 examples for each fine-grained class, so there are 4000 examples in total. We repeat our experiments five times on different compositions of positive and negative examples as shown in Table 2. In this part, we still use ResNet-18 as our model, and the sensitivity is approximated by the influence function. We first train the ResNet-18 from scratch and then froze the weights except for the top layer. After that, we train the last layer (randomly initialized) of the ResNet-18, and the class-level sensitivity is shown in Table 2. *The results show that examples within fine-grained classes (within the same superclass) have stronger sensitivity than examples between fine-grained classes (within the same superclass) for neural networks.*