Supplementary Material

Throughout this discussion, we will make frequently use of the following standard results concerning the exponential concentration of random variables:

Lemma 4 (Hoeffding's inequality for independent RVs (Hoeffding, 1994))**.** *Let $Z_1, Z_2, \ldots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}\left[Z_i\right]) \geqslant t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}\left[Z_i\right]) \leqslant -t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

for all $t \geqslant 0$.

Lemma 5 (Hoeffding's inequality for sampling with replacement (Hoeffding, 1994))**.** *Let $\mathcal{Z} = (Z_1, Z_2, \ldots, Z_N)$ be a finite population of N points with $Z_i \in [a.b]$ for all i. Let $X_1, X_2, \ldots X_n$ be a random sample drawn without replacement from $\mathcal{Z}$. Then for all $t \geqslant 0$, we have*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \geqslant t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu) \leqslant -t\right) \leqslant \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} Z_i$.

We now discuss one condition that generalizes the exponential concentration to dependent random variables.

Condition 2 (Bounded difference inequality)**.** *Let $\mathcal{Z}$ be some set and $\phi : \mathcal{Z}^n \to \mathbb{R}$. We say that ϕ satisfies the bounded difference assumption if there exists $c_1, c_2, \ldots c_n \geqslant 0$ s.t. for all i, we have*

$$\sup_{Z_1, Z_2, \ldots, Z_n, Z_i' \in \mathcal{Z}^{n+1}} \left|\phi(Z_1, \ldots, Z_i, \ldots, Z_n) - \phi(Z_1, \ldots, Z_i', \ldots, Z_n)\right| \leqslant c_i \,.$$

Lemma 6 (McDiarmid's inequality (McDiarmid, 1989))**.** *Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables on set $\mathcal{Z}$ and $\phi : \mathcal{Z}^n \to \mathbb{R}$ satisfy bounded difference inequality (Condition 2). Then for all $t > 0$, we have*

$$\mathbb{P}\left(\phi(Z_1, Z_2, \ldots, Z_n) - \mathbb{E}\left[\phi(Z_1, Z_2, \ldots, Z_n)\right] \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

and

$$\mathbb{P}\left(\phi(Z_1, Z_2, \ldots, Z_n) - \mathbb{E}\left[\phi(Z_1, Z_2, \ldots, Z_n)\right] \leqslant -t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

A. Proofs from Sec. 3

Additional notation Let m_1 be the number of mislabeled points ($\widetilde{S}_M$) and m_2 be the number of correctly labeled points ($\widetilde{S}_C$). Note $m_1 + m_2 = m$.

A.1. Proof of Theorem 1

Proof of Lemma 1. The main idea of our proof is to regard the clean portion of the data ($S \cup \widetilde{S}_C$) as fixed. Then, there exists an (unknown) classifier f^* that minimizes the expected risk calculated on the (fixed) clean data and (random draws of) the mislabeled data $\widetilde{S}_M$. Formally,

$$f^* := \arg\min_{f \in \mathcal{F}} \mathcal{E}_{\breve{\mathcal{D}}}(f) \,, \tag{10}$$

where

$$\breve{\mathcal{D}} = \frac{n}{m+n}\mathcal{S} + \frac{m_2}{m+n}\widetilde{\mathcal{S}}_C + \frac{m_1}{m+n}\mathcal{D}' \,.$$

Note here that $\breve{\mathcal{D}}$ is a combination of the *empirical distribution* over correctly labeled data $S \cup \widetilde{S}_C$ and the (population) distribution over mislabeled data $\mathcal{D}'$. Recall that

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(f) \,. \tag{11}$$

Since, $\widehat{f}$ minimizes 0-1 error on $S \cup \widetilde{S}$, using ERM optimality on (11), we have

$$\mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(f^*) \,. \tag{12}$$

Moreover, since f^* is independent of $\widetilde{S}_M$, using Hoeffding's bound, we have with probability at least $1-\delta$ that

$$\mathcal{E}_{\widetilde{\mathcal{S}}_M}(f^*) \leqslant \mathcal{E}_{\mathcal{D}'}(f^*) + \sqrt{\frac{\log(1/\delta)}{2m_1}} \,. \tag{13}$$

Finally, since f^* is the optimal classifier on $\breve{\mathcal{D}}$, we have

$$\mathcal{E}_{\breve{\mathcal{D}}}(f^*) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(\widehat{f}) \,. \tag{14}$$

Now to relate (12) and (14), we multiply (13) by $\frac{m_1}{m+n}$ and add $\frac{n}{m+n}\mathcal{E}_{\mathcal{S}}(f) + \frac{m_2}{m+n}\mathcal{E}_{\widetilde{\mathcal{S}}_C}(f)$ both the sides. Hence, we can rewrite (13) as follows:

$$\mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(f^*) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(f^*) + \frac{m_1}{m+n}\sqrt{\frac{\log(1/\delta)}{2m_1}} \,. \tag{15}$$

Now we combine equations (12), (15), and (14), to get

$$\mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(\widehat{f}) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(\widehat{f}) + \frac{m_1}{m+n}\sqrt{\frac{\log(1/\delta)}{2m_1}} \,, \tag{16}$$

which implies

$$\mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{\log(1/\delta)}{2m_1}} \,. \tag{17}$$

Since $\widetilde{S}$ is obtained by randomly labeling an unlabeled dataset, we assume $2m_1 \approx m$ [3]. Moreover, using $\mathcal{E}_{\mathcal{D}'} = 1 - \mathcal{E}_{\mathcal{D}}$ we obtain the desired result. □

Proof of Lemma 2. Recall $\mathcal{E}_{\widetilde{\mathcal{S}}}(f) = \frac{m_1}{m}\mathcal{E}_{\widetilde{\mathcal{S}}_M}(f) + \frac{m_2}{m}\mathcal{E}_{\widetilde{\mathcal{S}}_C}(f)$. Hence, we have

$$\begin{aligned} 2\mathcal{E}_{\widetilde{\mathcal{S}}}(f) - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(f) - \mathcal{E}_{\widetilde{\mathcal{S}}_C}(f) &= \left(\frac{2m_1}{m}\mathcal{E}_{\widetilde{\mathcal{S}}_M}(f) - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(f)\right) + \left(\frac{2m_2}{m}\mathcal{E}_{\widetilde{\mathcal{S}}_C}(f) - \mathcal{E}_{\widetilde{\mathcal{S}}_C}(f)\right) &(18)\\ &= \left(\frac{2m_1}{m} - 1\right)\mathcal{E}_{\widetilde{\mathcal{S}}_M}(f) + \left(\frac{2m_2}{m} - 1\right)\mathcal{E}_{\widetilde{\mathcal{S}}_C}(f) \,. &(19)\end{aligned}$$

[3] Formally, with probability at least $1-\delta$, we have $(m - 2m_1) \leqslant \sqrt{m\log(1/\delta)/2}$.

Since the dataset is labeled uniformly at random, with probability at least $1-\delta$, we have $\left(\frac{2m_1}{m}-1\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m}}$. Similarly, we have with probability at least $1-\delta$, $\left(\frac{2m_2}{m}-1\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m}}$. Using union bound, with probability at least $1-\delta$, we have

$$2\mathcal{E}_{\widetilde{S}} - \mathcal{E}_{\widetilde{S}_M}(f) - \mathcal{E}_{\widetilde{S}_C}(f) \leqslant \sqrt{\frac{\log(2/\delta)}{2m}} \left(\mathcal{E}_{\widetilde{S}_M}(f) + \mathcal{E}_{\widetilde{S}_C}(f)\right) . \tag{20}$$

With re-arranging $\mathcal{E}_{\widetilde{S}_M}(f) + \mathcal{E}_{\widetilde{S}_C}(f)$ and using the inequality $1-a \leqslant \frac{1}{1+a}$, we have

$$2\mathcal{E}_{\widetilde{S}} - \mathcal{E}_{\widetilde{S}_M}(f) - \mathcal{E}_{\widetilde{S}_C}(f) \leqslant 2\mathcal{E}_{\widetilde{S}}\sqrt{\frac{\log(2/\delta)}{2m}} . \tag{21}$$

□

Proof of Lemma 3. In the set of correctly labeled points $S \cup \widetilde{S}_C$, we have S as a random subset of $S \cup \widetilde{S}_C$. Hence, using Hoeffding's inequality for sampling without replacement (Lemma 5), we have with probability at least $1-\delta$

$$\mathcal{E}_{\widetilde{S}_C}(\widehat{f}) - \mathcal{E}_{\mathcal{S}\cup\widetilde{S}_C}(\widehat{f}) \leqslant \sqrt{\frac{\log(1/\delta)}{2m_2}} . \tag{22}$$

Re-writing $\mathcal{E}_{\mathcal{S}\cup\widetilde{S}_C}(\widehat{f})$ as $\frac{m_2}{m_2+n}\mathcal{E}_{\widetilde{S}_C}(\widehat{f}) + \frac{n}{m_2+n}\mathcal{E}_{\mathcal{S}}(\widehat{f})$, we have with probability at least $1-\delta$

$$\left(\frac{n}{n+m_2}\right)\left(\mathcal{E}_{\widetilde{S}_C}(\widehat{f}) - \mathcal{E}_{\mathcal{S}}(\widehat{f})\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m_2}} . \tag{23}$$

As before, assuming $2m_2 \approx m$, we have with probability at least $1-\delta$

$$\mathcal{E}_{\widetilde{S}_C}(\widehat{f}) - \mathcal{E}_{\mathcal{S}}(\widehat{f}) \leqslant \left(1+\frac{m_2}{n}\right)\sqrt{\frac{\log(1/\delta)}{m}} \leqslant \left(1+\frac{m}{2n}\right)\sqrt{\frac{\log(1/\delta)}{m}} . \tag{24}$$

□

Proof of Theorem 1. Having established these core intermediate results, we can now combine above three lemmas to prove the main result. In particular, we bound the population error on clean data ($\mathcal{E}_{\mathcal{D}}(\widehat{f})$) as follows:

(i) First, use (17), to obtain an upper bound on the population error on clean data, i.e., with probability at least $1-\delta/4$, we have

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant 1 - \mathcal{E}_{\widetilde{S}_M}(\widehat{f}) + \sqrt{\frac{\log(4/\delta)}{m}} . \tag{25}$$

(ii) Second, use (21), to relate the error on the mislabeled fraction with error on clean portion of randomly labeled data and error on whole randomly labeled dataset, i.e., with probability at least $1-\delta/2$, we have

$$-\mathcal{E}_{\widetilde{S}_M}(f) \leqslant \mathcal{E}_{\widetilde{S}_C}(f) - 2\mathcal{E}_{\widetilde{S}} + 2\mathcal{E}_{\widetilde{S}}\sqrt{\frac{\log(4/\delta)}{2m}} . \tag{26}$$

(iii) Finally, use (24) to relate the error on the clean portion of randomly labeled data and error on clean training data, i.e., with probability $1-\delta/4$, we have

$$\mathcal{E}_{\widetilde{S}_C}(\widehat{f}) \leqslant -\mathcal{E}_{\mathcal{S}}(\widehat{f}) + \left(1+\frac{m}{2n}\right)\sqrt{\frac{\log(4/\delta)}{m}} . \tag{27}$$

Using union bound on the above three steps, we have with probability at least $1-\delta$:

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{S}}(\widehat{f}) + 1 - 2\mathcal{E}_{\widetilde{S}}(\widehat{f}) + \left(\sqrt{2}\mathcal{E}_{\widetilde{S}} + 2 + \frac{m}{2n}\right)\sqrt{\frac{\log(4/\delta)}{m}} . \tag{28}$$

□

A.2. Proof of Proposition 1

Proof of Proposition 1. For a classifier $f : \mathcal{X} \to \{-1, 1\}$, we have $1 - 2\,\mathbb{I}\left[f(x) \neq y\right] = y \cdot f(x)$. Hence, by definition of $\mathcal{E}$, we have

$$1 - 2\mathcal{E}_{\widetilde{\mathcal{S}}}(f) = \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i) \leqslant \sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i)\,. \tag{29}$$

Note that for fixed inputs $(x_1, x_2, \ldots, x_m)$ in $\widetilde{S}$, $(y_1, y_2, \ldots y_m)$ are random labels. Define $\phi_1(y_1, y_2, \ldots, y_m) := \sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i)$. We have the following bounded difference condition on ϕ_1. For all i,

$$\sup_{y_1, \ldots y_m, y_i' \in \{-1,1\}^{m+1}} \left|\phi_1(y_1, \ldots, y_i, \ldots, y_m) - \phi_1(y_1, \ldots, y_i', \ldots, y_m)\right| \leqslant 1/m\,. \tag{30}$$

Similarly, we define $\phi_2(x_1, x_2, \ldots, x_m) := \mathbb{E}_{y_i \sim U\{-1,1\}}\left[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i)\right]$. We have the following bounded difference condition on ϕ_2. For all i,

$$\sup_{x_1, \ldots x_m, x_i' \in \mathcal{X}^{m+1}} \left|\phi_2(x_1, \ldots, x_i, \ldots, x_m) - \phi_1(x_1, \ldots, x_i', \ldots, x_m)\right| \leqslant 1/m\,. \tag{31}$$

Using McDiarmid's inequality (Lemma 6) twice with Condition (30) and (31), with probability at least $1 - \delta$, we have

$$\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i) - \mathbb{E}_{x,y}\left[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} y_i \cdot f(x_i)\right] \leqslant \sqrt{\frac{2\log(2/\delta)}{m}}\,. \tag{32}$$

Combining (29) and (32), we obtain the desired result. □

A.3. Proof of Theorem 2

Proof of Theorem 2 follows similar to the proof of Theorem 1. Note that the same results in Lemma 1, Lemma 2, and Lemma 3 hold in the regularized ERM case. However, the arguments in the proof of Lemma 1 change slightly. Hence, we state the lemma for regularized ERM and prove it here for completeness.

Lemma 7. *Assume the same setup as Theorem 2. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the random draws of mislabeled data $\widetilde{S}_M$, we have*

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant 1 - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f}) + \sqrt{\frac{\log(1/\delta)}{m}}\,. \tag{33}$$

Proof. The main idea of the proof remains the same, i.e. regard the clean portion of the data ($S \cup \widetilde{S}_C$) as fixed. Then, there exists a classifier f^* that is optimal over draws of the mislabeled data $\widetilde{S}_M$.

Formally,

$$f^* := \arg\min_{f \in \mathcal{F}} \mathcal{E}_{\breve{\mathcal{D}}}(f) + \lambda R(f)\,, \tag{34}$$

where

$$\breve{\mathcal{D}} = \frac{n}{m+n}\mathcal{S} + \frac{m_1}{m+n}\widetilde{\mathcal{S}}_C + \frac{m_2}{m+n}\mathcal{D}'\,.$$

That is, $\breve{\mathcal{D}}$ a combination of the *empirical distribution* over correctly labeled data $S \cup \widetilde{S}_C$ and the (population) distribution over mislabeled data $\mathcal{D}'$. Recall that

$$\widehat{f} := \arg\min_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(f) + \lambda R(f)\,. \tag{35}$$

Since, $\widehat{f}$ minimizes 0-1 error on $S \cup \widetilde{S}$, using ERM optimality on (11), we have

$$\mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(\widehat{f}) + \lambda R(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}}(f^*) + \lambda R(f^*)\,. \tag{36}$$

Moreover, since f^* is independent of $\widetilde{S}_M$, using Hoeffding's bound, we have with probability at least $1-\delta$ that

$$\mathcal{E}_{\widetilde{S}_M}(f^*) \leqslant \mathcal{E}_{\mathcal{D}'}(f^*) + \sqrt{\frac{\log(1/\delta)}{2m_1}}\,. \tag{37}$$

Finally, since f^* is the optimal classifier on $\breve{\mathcal{D}}$, we have

$$\mathcal{E}_{\breve{\mathcal{D}}}(f^*) + \lambda R(f^*) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(\widehat{f}) + \lambda R(\widehat{f})\,. \tag{38}$$

Now to relate (36) and (38), we can re-write the (37) as follows:

$$\mathcal{E}_{\mathcal{S}\cup\widetilde{\mathcal{S}}}(f^*) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(f^*) + \frac{m_1}{m+n}\sqrt{\frac{\log(1/\delta)}{2m_1}}\,. \tag{39}$$

After adding $\lambda R(f^*)$ on both sides in (39), we combine equations (36), (39), and (38), to get

$$\mathcal{E}_{\mathcal{S}\cup\widetilde{\mathcal{S}}}(\widehat{f}) \leqslant \mathcal{E}_{\breve{\mathcal{D}}}(\widehat{f}) + \frac{m_1}{m+n}\sqrt{\frac{\log(1/\delta)}{2m_1}}\,, \tag{40}$$

which implies

$$\mathcal{E}_{\widetilde{S}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{\log(1/\delta)}{2m_1}}\,. \tag{41}$$

Similar as before, since $\widetilde{S}$ is obtained by randomly labeling an unlabeled dataset, we assume $2m_1 \approx m$. Moreover, using $\mathcal{E}_{\mathcal{D}'} = 1 - \mathcal{E}_{\mathcal{D}}$ we obtain the desired result. □

A.4. Proof of Theorem 3

To prove our results in the multiclass case, we first state and prove lemmas parallel to those used in the proof of balanced binary case. We then combine these results to obtain the result in Theorem 3.

Before stating the result, we define mislabeled distribution $\mathcal{D}'$ for any $\mathcal{D}$. While $\mathcal{D}'$ and $\mathcal{D}$ share the same marginal distribution over inputs $\mathcal{X}$, the conditional distribution over labels y given an input $x \sim \mathcal{D}_{\mathcal{X}}$ is changed as follows: For any x, the Probability Mass Function (PMF) over y is defined as: $p_{\mathcal{D}'}(\cdot|x) := \frac{1-p_{\mathcal{D}}(\cdot|x)}{k-1}$, where $p_{\mathcal{D}}(\cdot|x)$ is the PMF over y for the distribution $\mathcal{D}$.

Lemma 8. *Assume the same setup as Theorem 3. Then for any $\delta > 0$, with probability at least $1-\delta$ over the random draws of mislabeled data $\widetilde{S}_M$, we have*

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant (k-1)\left(1 - \mathcal{E}_{\widetilde{S}_M}(\widehat{f})\right) + (k-1)\sqrt{\frac{\log(1/\delta)}{m}}\,. \tag{42}$$

Proof. The main idea of the proof remains the same. We begin by regarding the clean portion of the data ($S \cup \widetilde{S}_C$) as fixed. Then, there exists a classifier f^* that is optimal over draws of the mislabeled data $\widetilde{S}_M$.

However, in the multiclass case, we cannot as easily relate the population error on mislabeled data to the population accuracy on clean data. While for binary classification, we could lower bound the population accuracy $1-\mathcal{E}_{\mathcal{D}}$ with the empirical error on mislabeled data $\mathcal{E}_{\widetilde{S}_M}$ (in the proof of Lemma 1), for multiclass classification, error on the mislabeled data and accuracy on the clean data in the population are not so directly related. To establish (42), we break the error on the (unknown) mislabeled data into two parts: one term corresponds to predicting the true label on mislabeled data, and the other corresponds to predicting neither the true label nor the assigned (mis-)label. Finally, we relate these errors to their population counterparts to establish (42).

Formally,

$$f^* := \operatorname*{arg\,min}_{f\in\mathcal{F}} \mathcal{E}_{\breve{\mathcal{D}}}(f) + \lambda R(f)\,, \tag{43}$$

where

$$\breve{\mathcal{D}} = \frac{n}{m+n}\mathcal{S} + \frac{m_1}{m+n}\widetilde{\mathcal{S}}_C + \frac{m_2}{m+n}\mathcal{D}' \,.$$

That is, $\breve{\mathcal{D}}$ is a combination of the *empirical distribution* over correctly labeled data $S \cup \widetilde{S}_C$ and the (population) distribution over mislabeled data $\mathcal{D}'$. Recall that

$$\widehat{f} := \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathcal{E}_{\mathcal{S} \cup \widetilde{S}}(f) + \lambda R(f) \,. \tag{44}$$

Following the exact steps from the proof of Lemma 7, with probability at least $1 - \delta$, we have

$$\mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{\log(1/\delta)}{2m_1}} \,. \tag{45}$$

Similar to before, since $\widetilde{S}$ is obtained by randomly labeling an unlabeled dataset, we assume $\frac{k}{k-1}m_1 \approx m$.

Now we will relate $\mathcal{E}_{\mathcal{D}'}(\widehat{f})$ with $\mathcal{E}_{\mathcal{D}}(\widehat{f})$. Let y^T denote the (unknown) true label for a mislabeled point (x, y) (i.e., label before replacing it with a mislabel).

$$\mathbb{E}_{(x,y) \in \sim \mathcal{D}'}\left[\mathbb{I}\left[\widehat{f}(x) \neq y\right]\right] = \underbrace{\mathbb{E}_{(x,y) \in \sim \mathcal{D}'}\left[\mathbb{I}\left[\widehat{f}(x) \neq y \wedge \widehat{f}(x) \neq y^T\right]\right]}_{\mathrm{I}} + \underbrace{\mathbb{E}_{(x,y) \in \sim \mathcal{D}'}\left[\mathbb{I}\left[\widehat{f}(x) \neq y \wedge \widehat{f}(x) = y^T\right]\right]}_{\mathrm{II}} \,. \tag{46}$$

Clearly, term 2 is one minus the accuracy on the clean unseen data, i.e.,

$$\mathrm{II} = 1 - \mathbb{E}_{x,y \sim \mathcal{D}}\left[\mathbb{I}\left[\widehat{f}(x) \neq y\right]\right] = 1 - \mathcal{E}_{\mathcal{D}}(\widehat{f}) \,. \tag{47}$$

Next, we relate term 1 with the error on the unseen clean data. We show that term 1 is equal to the error on the unseen clean data scaled by $\frac{k-2}{k-1}$, where k is the number of labels. Using the definition of mislabeled distribution $\mathcal{D}'$, we have

$$\mathrm{I} = \frac{1}{k-1}\left(\mathbb{E}_{(x,y) \in \sim \mathcal{D}}\left[\sum_{i \in \mathcal{Y} \wedge i \neq y} \mathbb{I}\left[\widehat{f}(x) \neq i \wedge \widehat{f}(x) \neq y\right]\right]\right) = \frac{k-2}{k-1}\mathcal{E}_{\mathcal{D}}(\widehat{f}) \,. \tag{48}$$

Combining the result in (47), (48) and (46), we have

$$\mathcal{E}_{\mathcal{D}'}(\widehat{f}) = 1 - \frac{1}{k-1}\mathcal{E}_{\mathcal{D}}(\widehat{f}) \,. \tag{49}$$

Finally, combining the result in (49) with equation (45), we have with probability $1 - \delta$,

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant (k-1)\left(1 - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f})\right) + (k-1)\sqrt{\frac{k\log(1/\delta)}{2(k-1)m}} \,. \tag{50}$$

□

Lemma 9. *Assume the same setup as Theorem 3. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the random draws of $\widetilde{S}$, we have*

$$\left|k\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f}) - \mathcal{E}_{\widetilde{\mathcal{S}}_C}(\widehat{f}) - (k-1)\mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f})\right| \leqslant 2k\sqrt{\frac{\log(4/\delta)}{2m}} \,.$$

Proof. Recall $\mathcal{E}_{\widetilde{S}}(f) = \frac{m_1}{m}\mathcal{E}_{\widetilde{S}_M}(f) + \frac{m_2}{m}\mathcal{E}_{\widetilde{S}_C}(f)$. Hence, we have

$$\begin{aligned} k\mathcal{E}_{\widetilde{S}}(f) - (k-1)\mathcal{E}_{\widetilde{S}_M}(f) - \mathcal{E}_{\widetilde{S}_C}(f) &= (k-1)\left(\frac{km_1}{(k-1)m}\mathcal{E}_{\widetilde{S}_M}(f) - \mathcal{E}_{\widetilde{S}_M}(f)\right) \\ &\quad + \left(\frac{km_2}{m}\mathcal{E}_{\widetilde{S}_C}(f) - \mathcal{E}_{\widetilde{S}_C}(f)\right) \\ &= k\left[\left(\frac{m_1}{m} - \frac{k-1}{k}\right)\mathcal{E}_{\widetilde{S}_M}(f) + \left(\frac{m_2}{m} - \frac{1}{k}\right)\mathcal{E}_{\widetilde{S}_C}(f)\right] . \end{aligned}$$

Since the dataset is randomly labeled, we have with probability at least $1-\delta$, $\left(\frac{m_1}{m} - \frac{k-1}{k}\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m}}$. Similarly, we have with probability at least $1-\delta$, $\left(\frac{m_2}{m} - \frac{1}{k}\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m}}$. Using union bound, we have with probability at least $1-\delta$

$$k\mathcal{E}_{\widetilde{S}}(f) - (k-1)\mathcal{E}_{\widetilde{S}_M}(f) - \mathcal{E}_{\widetilde{S}_C}(f) \leqslant k\sqrt{\frac{\log(2/\delta)}{2m}}\left(\mathcal{E}_{\widetilde{S}_M}(f) + \mathcal{E}_{\widetilde{S}_C}(f)\right) . \tag{51}$$

□

Lemma 10. *Assume the same setup as Theorem 3. Then for any $\delta > 0$, with probability at least $1-\delta$ over the random draws of $\widetilde{S}_C$ and S, we have*

$$\left|\mathcal{E}_{\widetilde{\mathcal{S}}_C}(\widehat{f}) - \mathcal{E}_{\mathcal{S}}(\widehat{f})\right| \leqslant 1.5\sqrt{\frac{k\log(2/\delta)}{2m}} .$$

Proof. In the set of correctly labeled points $S \cup \widetilde{S}_C$, we have S as a random subset of $S \cup \widetilde{S}_C$. Hence, using Hoeffding's inequality for sampling without replacement (Lemma 5), we have with probability at least $1-\delta$

$$\mathcal{E}_{\widetilde{\mathcal{S}}_c}(\widehat{f}) - \mathcal{E}_{\mathcal{S}\cup\widetilde{\mathcal{S}}_C}(\widehat{f}) \leqslant \sqrt{\frac{\log(1/\delta)}{2m_2}} . \tag{52}$$

Re-writing $\mathcal{E}_{\mathcal{S}\cup\widetilde{\mathcal{S}}_C}(\widehat{f})$ as $\frac{m_2}{m_2+n}\mathcal{E}_{\widetilde{\mathcal{S}}_C}(\widehat{f}) + \frac{n}{m_2+n}\mathcal{E}_{\mathcal{S}}(\widehat{f})$, we have with probability at least $1-\delta$

$$\left(\frac{n}{n+m_2}\right)\left(\mathcal{E}_{\widetilde{\mathcal{S}}_c}(\widehat{f}) - \mathcal{E}_{\mathcal{S}}(\widehat{f})\right) \leqslant \sqrt{\frac{\log(1/\delta)}{2m_2}} . \tag{53}$$

As before, assuming $km_2 \approx m$, we have with probability at least $1-\delta$

$$\mathcal{E}_{\widetilde{\mathcal{S}}_c}(\widehat{f}) - \mathcal{E}_{\mathcal{S}}(\widehat{f}) \leqslant \left(1 + \frac{m_2}{n}\right)\sqrt{\frac{k\log(1/\delta)}{2m}} \leqslant \left(1 + \frac{1}{k}\right)\sqrt{\frac{k\log(1/\delta)}{2m}} . \tag{54}$$

□

Proof of Theorem 3. Having established these core intermediate results, we can now combine above three lemmas. In particular, we bound the population error on clean data ($\mathcal{E}_{\mathcal{D}}(\widehat{f})$) as follows:

(i) First, use (50), to obtain an upper bound on the population error on clean data, i.e., with probability at least $1-\delta/4$, we have

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant (k-1)\left(1 - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f})\right) + (k-1)\sqrt{\frac{k\log(4/\delta)}{2(k-1)m}} . \tag{55}$$

(ii) Second, use (51) to relate the error on the mislabeled fraction with error on clean portion of randomly labeled data and error on whole randomly labeled dataset, i.e., with probability at least $1-\delta/2$, we have

$$-(k-1)\mathcal{E}_{\widetilde{S}_M}(f) \leqslant \mathcal{E}_{\widetilde{S}_C}(f) - k\mathcal{E}_{\widetilde{S}} + k\sqrt{\frac{\log(4/\delta)}{2m}} . \tag{56}$$

(iii) Finally, use (54) to relate the error on the clean portion of randomly labeled data and error on clean training data, i.e., with probability $1-\delta/4$, we have

$$\mathcal{E}_{\widetilde{\mathcal{S}}_C}(\widehat{f}) \leqslant -\mathcal{E}_{\mathcal{S}}(\widehat{f}) + \left(1 + \frac{m}{kn}\right)\sqrt{\frac{k\log(4/\delta)}{2m}}\,. \tag{57}$$

Using union bound on the above three steps, we have with probability at least $1-\delta$:

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{S}}(\widehat{f}) + (k-1) - k\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f}) + (\sqrt{k(k-1)} + k + \sqrt{k} + \frac{m}{n\sqrt{k}})\sqrt{\frac{\log(4/\delta)}{2m}}\,. \tag{58}$$

Simplifying the term in RHS of (58), we get the desired result. in the final bound. □

B. Proofs from Sec. 4

We suppose that the parameters of the linear function are obtained via gradient descent on the following L_2 regularized problem:

$$\mathcal{L}_S(w;\lambda) := \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \left\| w \right\|_2^2 \,, \tag{59}$$

where $\lambda \geqslant 0$ is a regularization parameter. We assume access to a clean dataset $S = \{(x_i, y_i)\}_{i=1}^{n} \sim \mathcal{D}^n$ and randomly labeled dataset $\widetilde{S} = \{(x_i, y_i)\}_{i=n+1}^{n+m} \sim \widetilde{\mathcal{D}}^m$. Let $\boldsymbol{X} = [x_1, x_2, \cdots, x_{m+n}]$ and $\boldsymbol{y} = [y_1, y_2, \cdots, y_{m+n}]$. Fix a positive learning rate η such that $\eta \leqslant 1/\left(\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_{\text{op}} + \lambda^2\right)$ and an initialization $w_0 = 0$. Consider the following gradient descent iterates to minimize objective (59) on $S \cup \widetilde{S}$:

$$w_t = w_{t-1} - \eta \nabla_w \mathcal{L}_{S\cup\widetilde{S}}(w_{t-1};\lambda) \quad \forall t = 1, 2, \dots \tag{60}$$

Then we have $\{w_t\}$ converge to the limiting solution $\widehat{w} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T\boldsymbol{y}$. Define $\widehat{f}(x) := f(x;\widehat{w})$.

B.1. Proof of Theorem 4

We use a standard result from linear algebra, namely the Shermann-Morrison formula (Sherman & Morrison, 1950) for matrix inversion:

Lemma 11 (Sherman & Morrison (1950))**.** *Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $u, v \in \mathbb{R}^n$ are column vectors. Then $\boldsymbol{A} + uv^T$ is invertible iff $1 + v^T\boldsymbol{A}u \neq 0$ and in particular*

$$(\boldsymbol{A} + uv^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}uv^T\boldsymbol{A}^{-1}}{1 + v^T\boldsymbol{A}^{-1}u} \,. \tag{61}$$

For a given training set $S \cup \widetilde{S}_C$, define leave-one-out error on mislabeled points in the training data as

$$\mathcal{E}_{\text{LOO}(\widetilde{S}_M)} = \frac{\sum_{(x_i,y_i)\in\widetilde{S}_M} \mathcal{E}(f_{(i)}(x_i), y_i)}{\left|\widetilde{S}_M\right|} \,,$$

where $f_{(i)} := f(\mathcal{A}, (S \cup \widetilde{S})_{(i)})$. To relate empirical leave-one-out error and population error with hypothesis stability condition, we use the following lemma:

Lemma 12 (Bousquet & Elisseeff (2002))**.** *For the leave-one-out error, we have*

$$\mathbb{E}\left[\left(\mathcal{E}_{\mathcal{D}'}(\widehat{f}) - \mathcal{E}_{LOO(\widetilde{S}_M)}\right)^2\right] \leqslant \frac{1}{2m_1} + \frac{3\beta}{n+m} \,. \tag{62}$$

Proof of the above lemma is similar to the proof of Lemma 9 in Bousquet & Elisseeff (2002) and can be found in App. D. Before presenting the proof of Theorem 4, we introduce some more notation. Let $\boldsymbol{X}_{(i)}$ denote the matrix of covariates with the i^{th} point removed. Similarly, let $\boldsymbol{y}_{(i)}$ be the array of responses with the i^{th} point removed. Define the corresponding regularized GD solution as $\widehat{w}_{(i)} = \left(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)} + \lambda\boldsymbol{I}\right)^{-1} \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{(i)}$. Define $\widehat{f}_{(i)}(x) := f(x;\widehat{w}_{(i)})$.

Proof of Theorem 4. Because squared loss minimization does not imply 0-1 error minimization, we cannot use arguments from Lemma 1. This is the main technical difficulty. To compare the 0-1 error at a train point with an unseen point, we use the closed-form expression for $\widehat{w}$ and Shermann-Morrison formula to upper bound training error with leave-one-out cross validation error.

The proof is divided into three parts: In part one, we show that 0-1 error on mislabeled points in the training set is lower than the error obtained by leave-one-out error at those points. In part two, we relate this leave-one-out error with the population error on mislabeled distribution using Condition 1. While the empirical leave-one-out error is an unbiased estimator of the average population error of leave-one-out classifiers, we need hypothesis stability to control the variance of empirical leave-one-out error. Finally, in part three, we show that the error on the mislabeled training points can be estimated with just the randomly labeled and clean training data (as in proof of Theorem 1).

Part 1 First we relate training error with leave-one-out error. For any training point (x_i, y_i) in $\widetilde{S} \cup S$, we have

$$\mathcal{E}(\widehat{f}(x_i), y_i) = \mathbb{I}\left[y_i \cdot x_i^T \widehat{w} < 0\right] = \mathbb{I}\left[y_i \cdot x_i^T \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} < 0\right] \tag{63}$$

$$= \mathbb{I}\left[y_i \cdot x_i^T \underbrace{\left(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)} + x_i^T x_i + \lambda \boldsymbol{I}\right)^{-1}}_{\text{I}} (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)} + y_i \cdot x_i) < 0 \right]. \tag{64}$$

Letting $\boldsymbol{A} = \left(\boldsymbol{X}_{(i)}^T \boldsymbol{X}_{(i)} + \lambda \boldsymbol{I}\right)$ and using Lemma 11 on term 1, we have

$$\mathcal{E}(\widehat{f}(x_i), y_i) = \mathbb{I}\left[y_i \cdot x_i^T \left[\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1} x_i x_i^T \boldsymbol{A}^{-1}}{1 + x_i^T \boldsymbol{A}^{-1} x_i}\right] (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)} + y_i \cdot x_i) < 0\right] \tag{65}$$

$$= \mathbb{I}\left[y_i \cdot \left[\frac{x_i^T \boldsymbol{A}^{-1}(1 + x_i^T \boldsymbol{A}^{-1} x_i) - x_i^T \boldsymbol{A}^{-1} x_i x_i^T \boldsymbol{A}^{-1}}{1 + x_i^T \boldsymbol{A}^{-1} x_i}\right] (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)} + y_i \cdot x_i) < 0\right] \tag{66}$$

$$= \mathbb{I}\left[y_i \cdot \left[\frac{x_i^T \boldsymbol{A}^{-1}}{1 + x_i^T \boldsymbol{A}^{-1} x_i}\right] (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)} + y_i \cdot x_i) < 0\right]. \tag{67}$$

Since $1 + x_i^T \boldsymbol{A}^{-1} x_i > 0$, we have

$$\mathcal{E}(\widehat{f}(x_i), y_i) = \mathbb{I}\left[y_i \cdot x_i^T \boldsymbol{A}^{-1} (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)} + y_i \cdot x_i) < 0\right] \tag{68}$$

$$= \mathbb{I}\left[x_i^T \boldsymbol{A}^{-1} x_i + y_i \cdot x_i^T \boldsymbol{A}^{-1} (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)}) < 0\right] \tag{69}$$

$$\leqslant \mathbb{I}\left[y_i \cdot x_i^T \boldsymbol{A}^{-1} (\boldsymbol{X}_{(i)}^T \boldsymbol{y}_{(i)}) < 0\right] = \mathcal{E}(\widehat{f}_{(i)}(x_i), y_i). \tag{70}$$

Using (70), we have

$$\mathcal{E}_{\widetilde{S}_M}(\widehat{f}) \leqslant \mathcal{E}_{\text{LOO}(\widetilde{S}_M)} := \frac{\sum_{(x_i, y_i) \in \widetilde{S}_M} \mathcal{E}(\widehat{f}_{(i)}(x_i), y_i)}{\left|\widetilde{S}_M\right|}. \tag{71}$$

Part 2 We now relate RHS in (71) with the population error on mislabeled distribution. To do this, we leverage Condition 1 and Lemma 12. In particular, we have

$$\mathbb{E}_{\mathcal{S} \cup \widetilde{\mathcal{S}}_M}\left[\left(\mathcal{E}_{\mathcal{D}'}(\widehat{f}) - \mathcal{E}_{\text{LOO}(\widetilde{S}_M)}\right)^2\right] \leqslant \frac{1}{2m_1} + \frac{3\beta}{m+n}. \tag{72}$$

Using Chebyshev's inequality, with probability at least $1 - \delta$, we have

$$\mathcal{E}_{\text{LOO}(\widetilde{S}_M)} \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{1}{\delta}\left(\frac{1}{2m_1} + \frac{3\beta}{m+n}\right)}. \tag{73}$$

Part 3 Combining (73) and (71), we have

$$\mathcal{E}_{\widetilde{S}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{1}{\delta}\left(\frac{1}{2m_1} + \frac{3\beta}{m+n}\right)}. \tag{74}$$

Compare (74) with (17) in the proof of Lemma 1. We obtain a similar relationship between $\mathcal{E}_{\widetilde{\mathcal{S}}_M}$ and $\mathcal{E}_{\mathcal{D}'}$ but with a polynomial concentration instead of exponential concentration. In addition, since we just use concentration arguments to relate mislabeled error to the errors on the clean and unlabeled portions of the randomly labeled data, we can directly use the results in Lemma 2 and Lemma 3. Therefore, combining results in Lemma 2, Lemma 3, and (74) with union bound, we have with probability at least $1-\delta$

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{S}}(\widehat{f}) + 1 - 2\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f}) + \left(\sqrt{2}\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f}) + 1 + \frac{m}{2n}\right)\sqrt{\frac{\log(4/\delta)}{m}} + \sqrt{\frac{4}{\delta}\left(\frac{1}{m} + \frac{3\beta}{m+n}\right)}\,. \tag{75}$$

□

B.2. Extension to multiclass classification

For multiclass problems with squared loss minimization, as standard practice, we consider one-hot encoding for the underlying label, i.e., a class label $c \in [k]$ is treated as $(0, \cdot, 0, 1, 0, \cdot, 0) \in \mathbb{R}^k$ (with c-th coordinate being 1). As before, we suppose that the parameters of the linear function are obtained via gradient descent on the following L_2 regularized problem:

$$\mathcal{L}_S(w;\lambda) := \sum_{i=1}^{n} \left\| w^T x_i - y_i \right\|_2^2 + \lambda \sum_{j=1}^{k} \left\| w_j \right\|_2^2\,, \tag{76}$$

where $\lambda \geqslant 0$ is a regularization parameter. We assume access to a clean dataset $S = \{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}^n$ and randomly labeled dataset $\widetilde{S} = \{(x_i, y_i)\}_{i=n+1}^{n+m} \sim \widetilde{\mathcal{D}}^m$. Let $\boldsymbol{X} = [x_1, x_2, \cdots, x_{m+n}]$ and $\boldsymbol{y} = [e_{y_1}, e_{y_2}, \cdots, e_{y_{m+n}}]$. Fix a positive learning rate η such that $\eta \leqslant 1/\left(\left\|\boldsymbol{X}^T\boldsymbol{X}\right\|_{\text{op}} + \lambda^2\right)$ and an initialization $w_0 = 0$. Consider the following gradient descent iterates to minimize objective (59) on $S \cup \widetilde{S}$:

$$w_j{}^t = w_j{}^{t-1} - \eta\nabla_{w_j}\mathcal{L}_{S\cup\widetilde{S}}(w^{t-1};\lambda) \quad \forall t = 1, 2, \ldots \text{ and } j = 1, 2, \ldots, k\,. \tag{77}$$

Then we have $\{w_j{}^t\}$ for all $j = 1, 2, \cdots, k$ converge to the limiting solution $\widehat{w}_j = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}_j$. Define $\widehat{f}(x) := f(x;\widehat{w})$.

Theorem 5. *Assume that this gradient descent algorithm satisfies Condition 1 with $\beta = \mathcal{O}(1)$. Then for a multiclass classification problem wth k classes, for any $\delta > 0$, with probability at least $1-\delta$, we have:*

$$\begin{aligned}\mathcal{E}_{\mathcal{D}}(\widehat{f}) &\leqslant \mathcal{E}_{\mathcal{S}}(\widehat{f}) + (k-1)\left(1 - \frac{k}{k-1}\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f})\right)\\ &\quad + \left(k + \sqrt{k} + \frac{m}{n\sqrt{k}}\right)\sqrt{\frac{\log(4/\delta)}{2m}} + \sqrt{k(k-1)}\sqrt{\frac{4}{\delta}\left(\frac{1}{m} + \frac{3\beta}{m+n}\right)}\,.\end{aligned} \tag{78}$$

Proof. The proof of this theorem is divided into two parts. In the first part, we relate the error on the mislabeled samples with the population error on the mislabeled data. Similar to the proof of Theorem 4, we use Shermann-Morrison formula to upper bound training error with leave-one-out error on each $\widehat{w}^j$. Second part of the proof follows entirely from the proof of Theorem 3. In essence, the first part derives an equivalent of (45) for GD training with squared loss and then the second part follows from the proof of Theorem 3.

Part-1: Consider a training point (x_i, y_i) in $\widetilde{S} \cup S$. For simplicity, we use c_i to denote the class of i-th point and use y_i as the corresponding one-hot embedding. Recall error in multiclass point is given by $\mathcal{E}(\widehat{f}(x_i), y_i) = \mathbb{I}\left[c_i \notin \arg\max x_i^T\widehat{w}\right]$. Thus, there exists a $j \neq c_i \in [k]$, such that we have

$$\begin{aligned}\mathcal{E}(\widehat{f}(x_i), y_i) &= \mathbb{I}\left[c_i \notin \arg\max x_i^T\widehat{w}\right] = \mathbb{I}\left[x_i^T\widehat{w}_{c_i} < x_i^T\widehat{w}_j\right] &\quad (79)\\ &= \mathbb{I}\left[x_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}_{c_i} < x_i^T\left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}_j\right] &\quad (80)\\ &= \mathbb{I}\left[x_i^T\underbrace{\left(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)} + x_i^Tx_i + \lambda\boldsymbol{I}\right)^{-1}}_{\text{I}}\left(\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} + x_i - \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)}\right) < 0\right]. &\quad (81)\end{aligned}$$

Letting $\boldsymbol{A} = \left(\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)} + \lambda\boldsymbol{I}\right)$ and using Lemma 11 on term 1, we have

$$\begin{aligned}
\mathcal{E}(\widehat{f}(x_i), y_i) &= \mathbb{I}\left[x_i^T\left[\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}x_i x_i^T\boldsymbol{A}^{-1}}{1+x_i^T\boldsymbol{A}^{-1}x_i}\right]\left(\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} + x_i - \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)}\right) < 0\right] && (82)\\
&= \mathbb{I}\left[\left[\frac{x_i^T\boldsymbol{A}^{-1}(1+x_i^T\boldsymbol{A}^{-1}x_i) - x_i^T\boldsymbol{A}^{-1}x_i x_i^T\boldsymbol{A}^{-1}}{1+x_i^T\boldsymbol{A}^{-1}x_i}\right]\left(\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} + x_i - \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)}\right) < 0\right] && (83)\\
&= \mathbb{I}\left[\left[\frac{x_i^T\boldsymbol{A}^{-1}}{1+x_i^T\boldsymbol{A}^{-1}x_i}\right]\left(\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} + x_i - \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)}\right) < 0\right]. && (84)
\end{aligned}$$

Since $1 + x_i^T\boldsymbol{A}^{-1}x_i > 0$, we have

$$\begin{aligned}
\mathcal{E}(\widehat{f}(x_i), y_i) &= \mathbb{I}\left[x_i^T\boldsymbol{A}^{-1}\left(\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} + x_i - \boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)}\right) < 0\right] && (85)\\
&= \mathbb{I}\left[x_i^T\boldsymbol{A}^{-1}x_i + x_i^T\boldsymbol{A}^{-1}\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} - x_i^T\boldsymbol{A}^{-1}\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)} < 0\right] && (86)\\
&\leqslant \mathbb{I}\left[x_i^T\boldsymbol{A}^{-1}\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{c_i\,(i)} - x_i^T\boldsymbol{A}^{-1}\boldsymbol{X}_{(i)}^T\boldsymbol{y}_{j\,(i)} < 0\right] = \mathcal{E}(\widehat{f}_{(i)}(x_i), y_i). && (87)
\end{aligned}$$

Using (87), we have

$$\mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathrm{LOO}(\widetilde{\mathcal{S}}_M)} := \frac{\sum_{(x_i,y_i)\in\widetilde{\mathcal{S}}_M}\mathcal{E}(\widehat{f}_{(i)}(x_i), y_i)}{\left|\widetilde{\mathcal{S}}_M\right|}. \tag{88}$$

We now relate RHS in (71) with the population error on mislabeled distribution. Similar as before, to do this, we leverage Condition 1 and Lemma 12. Using (73) and (88), we have

$$\mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{1}{\delta}\left(\frac{1}{2m_1} + \frac{3\beta}{m+n}\right)}. \tag{89}$$

We have now derived a parallel to (45). Using the same arguments in the proof of Lemma 8, we have

$$\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant (k-1)\left(1 - \mathcal{E}_{\widetilde{\mathcal{S}}_M}(\widehat{f})\right) + (k-1)\sqrt{\frac{k}{\delta(k-1)}\left(\frac{1}{2m_1} + \frac{3\beta}{m+n}\right)}. \tag{90}$$

Part-2: We now combine the results in Lemma 9 and Lemma 10 to obtain the final inequality in terms of quantities that can be computed from just the randomly labeled and clean data. Similar to the binary case, we obtained a polynomial concentration instead of exponential concentration. Combining (90) with Lemma 9 and Lemma 10, we have with probability at least $1-\delta$

$$\begin{aligned}
\mathcal{E}_{\mathcal{D}}(\widehat{f}) \leqslant{}& \mathcal{E}_{\mathcal{S}}(\widehat{f}) + (k-1)\left(1 - \frac{k}{k-1}\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f})\right)\\
&+ \left(k + \sqrt{k} + \frac{m}{n\sqrt{k}}\right)\sqrt{\frac{\log(4/\delta)}{2m}} + \sqrt{k(k-1)}\sqrt{\frac{4}{\delta}\left(\frac{1}{m} + \frac{3\beta}{m+n}\right)}. && (91)
\end{aligned}$$

□

B.3. Discussion on Condition 1

The quantity in LHS of Condition 1 measures how much the function learned by the algorithm (in terms of error on unseen point) will change when one point in the training set is removed. We need hypothesis stability condition to control the variance of the empirical leave-one-out error to show concentration of average leave-one-error with the population error.

Additionally, we note that while the dominating term in the RHS of Theorem 4 matches with the dominating term in ERM bound in Theorem 1, there is a polynomial concentration term (dependence on $1/\delta$ instead of $\log(\sqrt{1/\delta})$) in Theorem 4.

Since with hypothesis stability, we just bound the variance, the polynomial concentration is due to the use of Chebyshev's inequality instead of an exponential tail inequality (as in Lemma 1). Recent works have highlighted that a slightly stronger condition than hypothesis stability can be used to obtain an exponential concentration for leave-one-out error (Abou-Moustafa & Szepesvári, 2019), but we leave this for future work for now.

B.4. Formal statement and proof of Proposition 2

Before formally presenting the result, we will introduce some notation. By $\mathcal{L}_S(w)$, we denote the objective in (59) with $\lambda = 0$. Assume Singular Value Decomposition (SVD) of $\boldsymbol{X}$ as $\sqrt{n}\boldsymbol{U}\boldsymbol{S}^{1/2}\boldsymbol{V}^T$. Hence $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^T$. Consider the GD iterates defined in (60). We now derive closed form expression for the t^{th} iterate of gradient descent:

$$w_t = w_{t-1} + \eta \cdot \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}w_{t-1}) = (\boldsymbol{I} - \eta\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^T)w_{k-1} + \eta\boldsymbol{X}^T\boldsymbol{y}\,. \tag{92}$$

Rotating by $\boldsymbol{V}^T$, we get

$$\widetilde{w}_t = (\boldsymbol{I} - \eta\boldsymbol{S})\widetilde{w}_{k-1} + \eta\widetilde{\boldsymbol{y}}, \tag{93}$$

where $\widetilde{w}_t = \boldsymbol{V}^T w_t$ and $\widetilde{\boldsymbol{y}} = \boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}$. Assuming the initial point $w_0 = 0$ and applying the recursion in (93), we get

$$\widetilde{w}_t = \boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{S})^k)\widetilde{\boldsymbol{y}}\,, \tag{94}$$

Projecting solution back to the original space, we have

$$w_t = \boldsymbol{V}\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{S})^k)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\,. \tag{95}$$

Define $f_t(x) := f(x; w_t)$ as the solution at the t^{th} iterate. Let $\widetilde{w}_\lambda = \arg\min_w \mathcal{L}_S(w;\lambda) = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{V}(\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}$. and define $\widetilde{f}_\lambda(x) := f(x; \widetilde{w}_\lambda)$ as the regularized solution. Assume κ be the condition number of the population covariance matrix and let s_{min} be the minimum positive singular value of the empirical covariance matrix. Our proof idea is inspired from recent work on relating gradient flow solution and regularized solution for regression problems (Ali et al., 2018). We will use the following lemma in the proof:

Lemma 13. *For all $x \in [0,1]$ and for all $k \in \mathbb{N}$, we have (a) $\frac{kx}{1+kx} \leqslant 1 - (1-x)^k$ and (b) $1 - (1-x)^k \leqslant 2 \cdot \frac{kx}{kx+1}$.*

Proof. Using $(1-x)^k \leqslant \frac{1}{1+kx}$, we have part (a). For part (b), we numerically maximize $\frac{(1+kx)(1-(1-x)^k)}{kx}$ for all $k \geqslant 1$ and for all $x \in [0,1]$. □

Proposition 3 (Formal statement of Proposition 2)**.** *Let $\lambda = \frac{1}{t\eta}$. For a training point x, we have*

$$\mathbb{E}_{x\sim\mathcal{S}}\left[(f_t(x) - \widetilde{f}_\lambda(x))^2\right] \leqslant c(t,\eta) \cdot \mathbb{E}_{x\sim\mathcal{S}}\left[f_t(x)^2\right]\,,$$

where $c(t,\eta) := \min(0.25, \frac{1}{s_{min}^2 t^2\eta^2})$. Similarly for a test point, we have

$$\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[(f_t(x) - \widetilde{f}_\lambda(x))^2\right] \leqslant \kappa \cdot c(t,\eta) \cdot \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[f_t(x)^2\right]\,.$$

Proof. We want to analyze the expected squared difference output of regularized linear regression with regularization constant $\lambda = \frac{1}{\eta t}$ and the gradient descent solution at the t^{th} iterate. We separately expand the algebraic expression for squared difference at a training point and a test point. Then the main step is to show that $\left[\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{S})^k) - (\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\right] \preceq c(\eta,t) \cdot \boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{S})^k)$.

Part 1 First, we will analyze the squared difference of the output at a training point (for simplicity, we refer to $S \cup \widetilde{S}$ as S), i.e.,

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{S}}\left[\left(f_t(x)-\widetilde{f}_\lambda(x)\right)^2\right] &= \|\boldsymbol{X}w_t - \boldsymbol{X}\widetilde{w}_\lambda\|_2^2 && (96)\\
&= \left\|\boldsymbol{X}\boldsymbol{V}\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}\boldsymbol{V}(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\right\|_2^2 && (97)\\
&= \left\|\boldsymbol{X}\boldsymbol{V}\left(\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\right)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\right\|_2 && (98)\\
&= \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\underbrace{\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}}_{\text{I}}\right)^2\boldsymbol{S}\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\,. && (99)
\end{aligned}
$$

We now separately consider term 1. Substituting $\lambda = \frac{1}{t\eta}$, we get

$$
\begin{aligned}
\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1} &= \boldsymbol{S}^{-1}\left((\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{I}+\boldsymbol{S}^{-1}\lambda)^{-1}\right) && (100)\\
&= \underbrace{\boldsymbol{S}^{-1}\left((\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{I}+(\boldsymbol{S}t\eta)^{-1})^{-1}\right)}_{\boldsymbol{A}}\,. && (101)
\end{aligned}
$$

We now separately bound the diagonal entries in matrix $\boldsymbol{A}$. With s_i, we denote i^{th} diagonal entry of $\boldsymbol{S}$. Note that since $\eta \leqslant 1/\|S\|_{\text{op}}$, for all i, $\eta s_i \leqslant 1$. Consider i^{th} diagonal term (which is non-zero) of the diagonal matrix $\boldsymbol{A}$, we have

$$
\begin{aligned}
\boldsymbol{A}_{ii} = \frac{1}{s_i}\left(1-(1-s_i\eta)^t - \frac{t\eta s_i}{1+t\eta s_i}\right) &= \frac{1-(1-s_i\eta)^t}{s_i}\left(\underbrace{1-\frac{t\eta s_i}{(1+t\eta s_i)(1-(1-s_i\eta)^t)}}_{\text{II}}\right) && (102)\\
&\leqslant \frac{1}{2}\left[\frac{1-(1-s_i\eta)^t}{s_i}\right]. && \text{(Using Lemma 13 (b))}
\end{aligned}
$$

Additionally, we can also show the following upper bound on term 2:

$$
\begin{aligned}
1-\frac{t\eta s_i}{(1+t\eta s_i)(1-(1-s_i\eta)^t)} &= \frac{(1+t\eta s_i)(1-(1-s_i\eta)^t) - t\eta s_i}{(1+t\eta s_i)(1-(1-s_i\eta)^t)} && (103)\\
&\leqslant \frac{1-(1-s_i\eta)^t - t\eta s_i(1-s_i\eta)^t}{(1+t\eta s_i)(1-(1-s_i\eta)^t)} && (104)\\
&\leqslant \frac{1}{t\eta s_i}\,. && \text{(Using Lemma 13 (a))}
\end{aligned}
$$

Combining both the upper bounds on each diagonal entry $\boldsymbol{A}_{ii}$, we have

$$
\boldsymbol{A} \le c_1(\eta, t)\cdot\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t)\,, \tag{105}
$$

where $c_1(\eta,t) = \min(0.5, \frac{1}{ts_i\eta})$. Plugging this into (99), we have

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{S}}\left[\left(f_t(x)-\widetilde{f}_\lambda(x)\right)^2\right] &\leqslant c(\eta,t)\cdot\boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)^2\boldsymbol{S}\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} && (106)\\
&= c(\eta,t)\cdot\boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)\boldsymbol{S}\left(\boldsymbol{S}^{-1}(\boldsymbol{I}-(\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} && (107)\\
&= c(\eta,t)\cdot\|\boldsymbol{X}w_t\|_2^2 && (108)\\
&= c(\eta,t)\cdot\mathbb{E}_{x\sim\mathcal{S}}\left[(f_t(x))^2\right], && (109)
\end{aligned}
$$

where $c(\eta,t) = \min(0.25, \frac{1}{t^2 s_i^2\eta^2})$.

Part 2 With $\mathbf{\Sigma}$, we denote the underlying true covariance matrix. We now consider the squared difference of output at an unseen point:

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\left(f_t(x) - \widetilde{f}_\lambda(x)\right)^2\right] &= \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\left\|x^T w_t - x^T \widetilde{w}_\lambda\right\|_2\right] && (110)\\
&= \left\|x^T \boldsymbol{V}\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I} - \eta\boldsymbol{S})^t)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} - x^T\boldsymbol{V}(\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\right\|_2 && (111)\\
&= \left\|x^T\boldsymbol{V}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\right)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\right\|_2 && (112)\\
&= \boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\right)\boldsymbol{V}^T\mathbf{\Sigma}\boldsymbol{V} && (113)\\
&\qquad \left((\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}\right)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} && (114)\\
&\leqslant \sigma_{\text{max}}\cdot\boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\underbrace{\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t) - (\boldsymbol{S}+\lambda\boldsymbol{I})^{-1}}_{\text{I}}\right)^2\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\,, && (115)
\end{aligned}
$$

where σ_{max} is the maximum eigenvalue of the underlying covariance matrix $\mathbf{\Sigma}$. Using the upper bound on term 1 in (105), we have

$$
\begin{aligned}
\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\left(f_t(x) - \widetilde{f}_\lambda(x)\right)^2\right] &\leqslant \sigma_{\text{max}}\cdot c(\eta,t)\cdot\boldsymbol{y}^T\boldsymbol{V}\boldsymbol{X}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)^2\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} && (116)\\
&= \kappa\cdot c(\eta,t)\cdot\sigma_{\text{min}}\cdot\left\|\boldsymbol{V}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)\boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y}\right\|_2^2 && (117)\\
&\leqslant \kappa\cdot c(\eta,t)\cdot\left[\boldsymbol{V}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)\boldsymbol{V}^T\boldsymbol{X}^T\right]^T\mathbf{\Sigma} && (118)\\
&\qquad \left[\boldsymbol{V}\left(\boldsymbol{S}^{-1}(\boldsymbol{I} - (\boldsymbol{I}-\eta\boldsymbol{S})^t)\right)\boldsymbol{V}^T\boldsymbol{X}^T\right]\boldsymbol{y} && (119)\\
&= \kappa\cdot c(\eta,t)\cdot\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\left[\left\|x^T w_t\right\|_2\right]\,. && (120)
\end{aligned}
$$

□

B.5. Extension to deep learning

Under Assumption B.6, we present the formal result parallel to Theorem 3.

Theorem 6. *Consider a multiclass classification problem with k classes. Under Assumption 1, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$
\mathcal{E}_\mathcal{D}(\widehat{f}) \leqslant \mathcal{E}_\mathcal{S}(\widehat{f}) + (k-1)\left(1 - \tfrac{k}{k-1}\mathcal{E}_{\widetilde{\mathcal{S}}}(\widehat{f})\right) + c\sqrt{\frac{\log(\frac{4}{\delta})}{2m}}\,, \qquad (121)
$$

for some constant $c \leqslant ((c+1)k + \sqrt{k} + \frac{m}{n\sqrt{k}})$.

The proof follows exactly as in step (i) to (iii) in Theorem 3.

B.6. Justifying Assumption 1

Motivated by the analysis on linear models, we now discuss alternate (and weaker) conditions that imply Assumption 1. We need hypothesis stability (Condition 1) and the following assumption relating training error and leave-one-error:

Assumption 2. *Let $\widehat{f}$ be a model obtained by training with algorithm $\mathcal{A}$ on a mixture of clean S and randomly labeled data $\widetilde{S}$. Then we assume we have*

$$
\mathcal{E}_{\widetilde{S}_M}(\widehat{f}) \leqslant \mathcal{E}_{LOO(\widetilde{S}_M)}\,,
$$

for all $(x_i, y_i) \in \widetilde{S}_M$ where $\widehat{f}_{(i)} := f(\mathcal{A}, S \cup \widetilde{S}_{M(i)})$ and $\mathcal{E}_{LOO(\widetilde{S}_M)} := \frac{\sum_{(x_i,y_i)\in\widetilde{S}_M}\mathcal{E}(\widehat{f}_{(i)}(x_i), y_i)}{|\widetilde{S}_M|}$.

Intuitively, this assumption states that the error on a (mislabeled) datum (x, y) included in the training set is less than the error on that datum (x, y) obtained by a model trained on the training set $S - \{(x, y)\}$. We proved this for linear models trained with GD in the proof of Theorem 5. Condition 1 with $\beta = \mathcal{O}(1)$ and Assumption 2 together with Lemma 12 implies Assumption 1 with a polynomial residual term (instead of logarithmic in $1/\delta$):

$$\mathcal{E}_{\mathcal{S}_M}(\widehat{f}) \leqslant \mathcal{E}_{\mathcal{D}'}(\widehat{f}) + \sqrt{\frac{1}{\delta}\left(\frac{1}{m} + \frac{3\beta}{m+n}\right)}\,. \tag{122}$$

C. Additional experiments and details

C.1. Datasets

Toy Dataset Assume fixed constants μ and σ. For a given label y, we simulate features x in our toy classification setup as follows:

$$x := \texttt{concat}\,[x_1, x_2] \quad \text{where} \quad x_1 \sim \mathcal{N}(y \cdot \mu, \sigma^2 I_{d\times d}) \;\text{ and }\; x_1 \sim \mathcal{N}(0, \sigma^2 I_{d\times d})\,.$$

In experiements throughout the paper, we fix dimention $d = 100$, $\mu = 1.0$, and $\sigma = \sqrt{d}$. Intuitively, x_1 carries the information about the underlying label and x_2 is additional noise independent of the underlying label.

CV datasets We use MNIST (LeCun et al., 1998) and CIFAR10 (Krizhevsky & Hinton, 2009). We produce a binary variant from the multiclass classification problem by mapping classes $\{0, 1, 2, 3, 4\}$ to label 1 and $\{5, 6, 7, 8, 9\}$ to label -1. For CIFAR dataset, we also use the standard data augementation of random crop and horizontal flip. PyTorch code is as follows:

```
(transforms.RandomCrop(32, padding=4),
     transforms.RandomHorizontalFlip())
```

NLP dataset We use IMDb Sentiment analysis (Maas et al., 2011) corpus.

C.2. Architecture Details

All experiments were run on NVIDIA GeForce RTX 2080 Ti GPUs. We used PyTorch (Paszke et al., 2019) and Keras with Tensorflow (Abadi et al., 2016) backend for experiments.

Linear model For the toy dataset, we simulate a linear model with scalar output and the same number of parameters as the number of dimensions.

Wide nets To simulate the NTK regime, we experiment with $2-$layered wide nets. The PyTorch code for 2-layer wide MLP is as follows:

```
nn.Sequential(
    nn.Flatten(),
    nn.Linear(input_dims, 200000, bias=True),
    nn.ReLU(),
    nn.Linear(200000, 1, bias=True)
    )
```

We experiment both (i) with the second layer fixed at random initialization; (ii) and updating both layers' weights.

Deep nets for CV tasks We consider a 4-layered MLP. The PyTorch code for 4-layer MLP is as follows:

```
nn.Sequential(nn.Flatten(),
    nn.Linear(input_dim, 5000, bias=True),
    nn.ReLU(),
    nn.Linear(5000, 5000, bias=True),
    nn.ReLU(),
    nn.Linear(5000, 5000, bias=True),
    nn.ReLU(),
    nn.Linear(1024, num_label, bias=True)
    )
```

For MNIST, we use 1000 nodes instead of 5000 nodes in the hidden layer. We also experiment with convolutional nets. In particular, we use ResNet18 (He et al., 2016). Implementation adapted from: https://github.com/kuangliu/pytorch-cifar.git.

Deep nets for NLP We use a simple LSTM model with embeddings intialized with ELMo embeddings (Peters et al., 2018). Code adapted from: https://github.com/kamujun/elmo_experiments/blob/master/elmo_experiment/notebooks/elmo_text_classification_on_imdb.ipynb

We also evaluate our bounds with a BERT model. In particular, we fine-tune an off-the-shelf uncased BERT model (Devlin et al., 2018). Code adapted from Hugging Face Transformers (Wolf et al., 2020): `https://huggingface.co/transformers/v3.1.0/custom_datasets.html`.

C.3. Additonal experiments

Results with SGD on underparameterized linear models

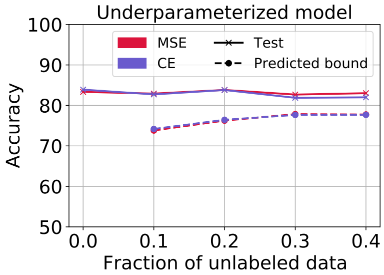
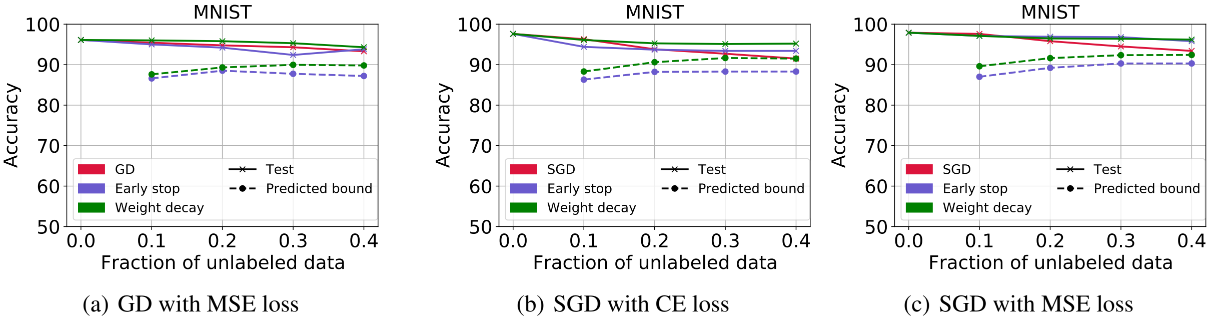


Figure 3. We plot the accuracy and corresponding bound (RHS in (1)) at $\delta = 0.1$ for toy binary classification task. Results aggregated over 3 seeds. Accuracy vs fraction of unlabeled data (w.r.t clean data) in the toy setup with a linear model trained with SGD. Results parallel to Fig. 2(a) with SGD.

Results with wide nets on binary MNIST



(a) GD with MSE loss (b) SGD with CE loss (c) SGD with MSE loss

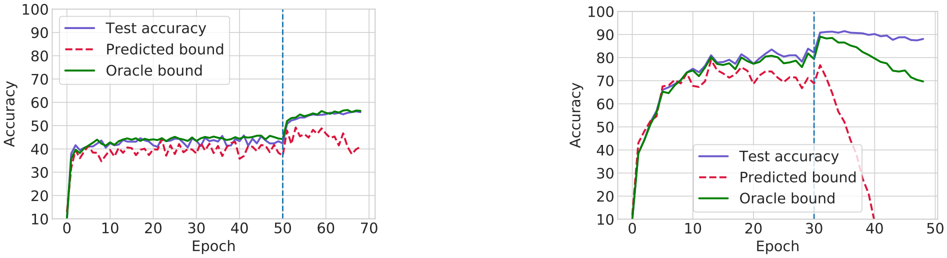
Figure 4. We plot the accuracy and corresponding bound (RHS in (1)) at $\delta = 0.1$ for binary MNIST classification. Results aggregated over 3 seeds. Accuracy vs fraction of unlabeled data for a 2-layer wide network on binary MNIST with both the layers training in (a,b) and only first layer training in (c). Results parallel to Fig. 2(b) .

Results on CIFAR 10 and MNIST We plot epoch wise error curve for results in Table 1(Fig. 5 and Fig. 6). We observe the same trend as in Fig. 1. Additionally, we plot an *oracle bound* obtained by tracking the error on mislabeled data which nevertheless were predicted as true label. To obtain an exact emprical value of the oracle bound, we need underlying true labels for the randomly labeled data. While with just access to extra unlabeled data we cannot calculate oracle bound, we note that the oracle bound is very tight and never violated in practice underscoring an importamt aspect of generalization in multiclass problems. This highlight that even a stronger conjecture may hold in multiclass classification, i.e., error on mislabeled data (where nevertheless true label was predicted) lower bounds the population error on the distribution of mislabeled data and hence, the error on (a specific) mislabeled portion predicts the population accuracy on clean data. On the other hand, the dominating term of in Theorem 3 is loose when compared with the oracle bound. The main reason, we believe is the pessimistic upper bound in (45) in the proof of Lemma 8. We leave an investigation on this gap for future.

Results on CIFAR 100 On CIFAR100, our bound in (5) yields vacous bounds. However, the oracle bound as explained above yields tight guarantees in the initial phase of the learning (i.e., when learning rate is less than 0.1) (Fig. 7).

C.4. Hyperparameter Details

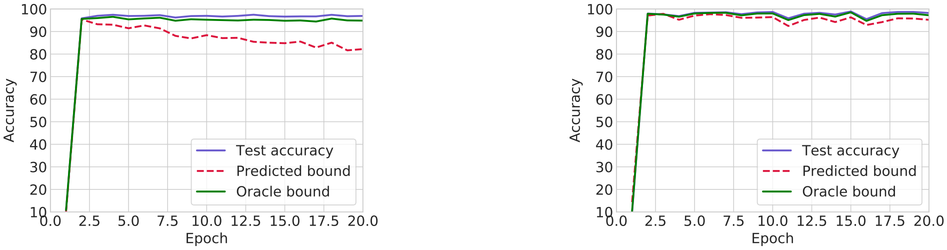
Fig. 1 We use clean training dataset of size $40,000$. We fix the amount of unlabeled data at 20% of the clean size, i.e. we include additional $8,000$ points with randomly assigned labels. We use test set of $10,000$ points. For both MLP and ResNet,



(a) MLP

(b) ResNet

Figure 5. Per epoch curves for CIFAR10 corresponding results in Table 1. As before, we just plot the dominating term in the RHS of (5) as predicted bound. Additionally, we also plot the predicted lower bound by the error on mislabeled data which nevertheless were predicted as true label. We refer to this as "Oracle bound". See text for more details.



(a) MLP

(b) ResNet

Figure 6. Per epoch curves for MNIST corresponding results in Table 1. As before, we just plot the dominating term in the RHS of (5) as predicted bound. Additionally, we also plot the predicted lower bound by the error on mislabeled data which nevertheless were predicted as true label. We refer to this as "Oracle bound". See text for more details.

we use SGD with an initial learning rate of 0.1 and momentum 0.9. We fix the weight decay parameter at 5×10^{-4}. After 100 epochs, we decay the learning rate to 0.01. We use SGD batch size of 100.

Fig. 2 (a) We obtain a toy dataset according to the process described in Sec. C.1. We fix $d = 100$ and create a dataset of $50,000$ points with balanced classes. Moreover, we sample additional covariates with the same procedure to create randomly labeled dataset. For both SGD and GD training, we use a fixed learning rate 0.1.

Fig. 2 (b) Similar to binary CIFAR, we use clean training dataset of size $40,000$ and fix the amount of unlabeled data at 20% of the clean dataset size. To train wide nets, we use a fixed learning of 0.001 with GD and SGD. We decide the weight decay parameter and the early stopping point that maximizes our generalization bound (i.e. without peeking at unseen data). We use SGD batch size of 100.

Fig. 2 (c) With IMDb dataset, we use a clean dataset of size $20,000$ and as before, fix the amount of unlabeled data at 20% of the clean data. To train ELMo model, we use Adam optimizer with a fixed learning rate 0.01 and weight decay 10^{-6} to minimize cross entropy loss. We train with batch size 32 for 3 epochs. To fine-tune BERT model, we use Adam optimizer with learning rate 5×10^{-5} to minimize cross entropy loss. We train with a batch size of 16 for 1 epoch.

Table 1 For multiclass datasets, we train both MLP and ResNet with the same hyperparameters as described before. We sample a clean training dataset of size $40,000$ and fix the amount of unlabeled data at 20% of the clean size. We use SGD with an initial learning rate of 0.1 and momentum 0.9. We fix the weight decay parameter at 5×10^{-4}. After 30 epochs for ResNet and after 50 epochs for MLP, we decay the learning rate to 0.01. We use SGD with batch size 100. For Fig. 7, we use the same hyperparameters as CIFAR10 training, except we now decay learning rate after 100 epochs.

In all experiments, to identify the best possible accuracy on just the clean data, we use the exact same set of hyperparamters except the stopping point. We choose a stopping point that maximizes test performance.

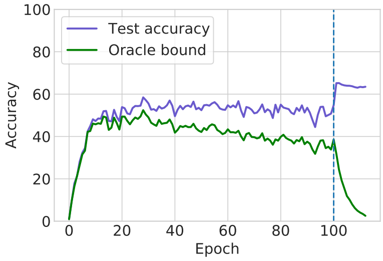


Figure 7. Predicted lower bound by the error on mislabeled data which nevertheless were predicted as true label with ResNet18 on CIFAR100. We refer to this as "Oracle bound". See text for more details. The bound predicted by RATT (RHS in (5)) is vacuous.

C.5. Summary of experiments

Classification type	Model category	Model	Dataset
Binary	Low dimensional	Linear model	Toy Gaussain dataset
	Overparameterized linear nets	2-layer wide net	Binary MNIST
	Deep nets	MLP	Binary MNIST
			Binary CIFAR
		ResNet	Binary MNIST
			Binary CIFAR
		ELMo-LSTM model	IMDb Sentiment Analysis
		BERT pre-trained model	IMDb Sentiment Analysis
Multiclass	Deep nets	MLP	MNIST
			CIFAR10
		ResNet	MNIST
			CIFAR10
			CIFAR100

D. Proof of Lemma 12

Proof of Lemma 12. Recall, we have a training set $S \cup \widetilde{S}_C$. We defined leave-one-out error on mislabeled points as

$$\mathcal{E}_{\text{LOO}(\widetilde{S}_M)} = \frac{\sum_{(x_i, y_i) \in \widetilde{S}_M} \mathcal{E}(f_{(i)}(x_i), y_i)}{\left|\widetilde{S}_M\right|} \,,$$

where $f_{(i)} := f(\mathcal{A}, (S \cup \widetilde{S})_{(i)})$. Define $S' := S \cup \widetilde{S}$. Assume (x, y) and (x', y') as i.i.d. samples from $\mathcal{D}'$. Using Lemma 25 in Bousquet & Elisseeff (2002), we have

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{E}_{\mathcal{D}'}(\widehat{f}) - \mathcal{E}_{\text{LOO}(\widetilde{S}_M)}\right)^2\right] \leqslant & \mathbb{E}_{S', (x,y), (x',y')}\left[\mathcal{E}(\widehat{f}(x), y)\mathcal{E}(\widehat{f}(x'), y')\right] - 2\mathbb{E}_{S', (x,y)}\left[\mathcal{E}(\widehat{f}(x), y)\mathcal{E}(f_{(i)}(x_i), y_i)\right] \\ & + \frac{m_1 - 1}{m_1}\mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i)\mathcal{E}(f_{(j)}(x_j), y_j)\right] + \frac{1}{m_1}\mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i)\right] \,. \end{aligned} \tag{123}$$

We can rewrite the equation above as :

$$\begin{aligned}\mathbb{E}\left[\left(\mathcal{E}_{\mathcal{D}'}(\widehat{f}) - \mathcal{E}_{\text{LOO}(\widetilde{S}_M)}\right)^2\right] \leqslant & \underbrace{\mathbb{E}_{S', (x,y), (x',y')}\left[\mathcal{E}(\widehat{f}(x), y)\mathcal{E}(\widehat{f}(x'), y') - \mathcal{E}(\widehat{f}(x), y)\mathcal{E}(f_{(i)}(x_i), y_i)\right]}_{\text{I}} \\ & + \underbrace{\mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i)\mathcal{E}(f_{(j)}(x_j), y_j) - \mathcal{E}(\widehat{f}(x), y)\mathcal{E}(f_{(i)}(x_i), y_i)\right]}_{\text{II}} \\ & + \underbrace{\frac{1}{m_1}\mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i) - \mathcal{E}(f_{(i)}(x_i), y_i)\mathcal{E}(f_{(j)}(x_j), y_j)\right]}_{\text{III}} \,. \end{aligned} \tag{124}$$

We will now bound term III. Using Cauchy-Schwarz's inequality, we have

$$\begin{aligned}\mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i) - \mathcal{E}(f_{(i)}(x_i), y_i)\mathcal{E}(f_{(j)}(x_j), y_j)\right]^2 &\leqslant \mathbb{E}_{S'}\left[\mathcal{E}(f_{(i)}(x_i), y_i)\right]^2 \mathbb{E}_{S'}\left[1 - \mathcal{E}(f_{(j)}(x_j), y_j)\right]^2 && (125) \\ &\leqslant \frac{1}{4} \,. && (126)\end{aligned}$$

Note that since (x_i, y_i), (x_j, y_j), (x, y), and (x', y') are all from same distribution $\mathcal{D}'$, we directly incorporate the bounds on term I and II from the proof of Lemma 9 in Bousquet & Elisseeff (2002). Combining that with (126) and our definition of hypothesis stability in Condition 1, we have the required claim.

□