A Proof of Lemma 1

Proof. Given the number of clean samples n, for any learner $f \in \mathcal{F}$, the accuracy of f on x_0 is:

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] =& Pr[(x_0, y_0) \in D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \in D_n] \\ &+ Pr[(x_0, y_0) \notin D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n], \end{aligned} \tag{1}$$

where $Pr[(x_0, y_0) \in D_n] = 1 - (1 - 1/k)^n$ and $Pr[(x_0, y_0) \notin D_n] = (1 - 1/k)^n$.

Since the bijection g is unknown to the learner f, when $(x_0, y_0) \notin D_n$, by symmetry the optimal prediction is predicting an arbitrary label that is not in D_n, thus

$$\begin{aligned} & Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n] \\ \leq & Pr[E \mid (x_0, y_0) \notin D_n] \cdot 1 + Pr[\neg E \mid (x_0, y_0) \notin D_n] \cdot \frac{1}{2} \\ = & Pr[E \mid (x_0, y_0) \notin D_n] \cdot \frac{1}{2} + \frac{1}{2} \\ \leq & \left(1 - \left(1 - \frac{1}{k-1}\right)^n\right) \cdot \frac{1}{2} + \frac{1}{2} \\ = & 1 - \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2} \end{aligned}$$

where E denotes the event that all other $k-1$ labels appear in the training set D_n. Above, what we do is to divide the probability into two cases and bound them separately. Case 1 is when E happens, where we simply upper bound the probability that $f_{D_n}(x_0) = y_0$ by 1. Case 2 is when E does not happen, meaning that there is some $y_1 \neq y_0$ that does not appear in D_n. By Definition 1, we have $Pr[f_{D_n}(x_0) = y_0] = Pr[f_{T_{y_0 \leftrightarrow y_1}(D_n)}(x_0) = y_1] = Pr[f_{D_n}(x_0) = y_1]$ thus $Pr[f_{D_n}(x_0) = y_0] \leq \frac{1}{2}$.

With Equation 1, we have

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] \leq & Pr[(x_0, y_0) \in D_n] \cdot 1 + Pr[(x_0, y_0) \notin D_n] \cdot \left(1 - \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2}\right) \\ = & 1 - Pr[(x_0, y_0) \notin D_n] \cdot \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2} \\ = & 1 - \left(1 - \frac{1}{k}\right)^n \cdot \left(1 - \frac{1}{k-1}\right)^n \cdot \frac{1}{2}. \end{aligned}$$

Thus $Pr[f_{D_n}(x_0) = y_0] \geq \tau \Rightarrow \left(1 - \frac{1}{k}\right)^n \cdot \left(1 - \frac{1}{k-1}\right)^n \leq 2 - 2\tau \Rightarrow n \geq \frac{\log(2-2\tau)}{\log(1-2/k)} = \Theta(k)$.

The intuition behind the proof: If the training set contains (x_0, y_0), the learner can obviously predict correctly; Otherwise, the best it can do is to guess a label that is not in the training set. □

B Proof of Lemma 2

Proof. Given x_0, for any N and any learner f, one of following two cases must be true:

Case 1: If $Pr[f_{D_N}(x_0) = y_0] \leq \frac{1}{|Y|}$, using the identity transform $T(D) = D$ for all $D \in \Omega^{\mathbb{N}}$, we have $Pr[f_{T(D_N)}(x_0) = y_0] \leq \frac{1}{|Y|}$ and $\mathbb{E}[|T(D_N) - D_N|] = 0$.

Case 2: If $Pr[f_{D_N}(x_0) = y_0] > \frac{1}{|Y|}$, since $\sum_{y \in Y} Pr[f_{D_N}(x_0) = y] = 1$, there exists $y_1 \neq y_0$ such that $Pr[f_{D_N}(x_0) = y_1] \leq \frac{1}{|Y|}$. Let $T = T_{y_0 \leftrightarrow y_1}$ be a transform swapping labels y_0 and y_1, i.e. $T(D)$ is the same as D except that every $(x, y_0) \in D$ will becomes $(x, y_1) \in T(D)$ and every $(x, y_1) \in D$ will becomes $(x, y_0) \in T(D)$. Since $T = T_{y_0 \leftrightarrow y_1}$ is a transform swapping labels y_0 and y_1 in the training set, $E[|T(D_N) - D_N|]$ is in fact the expected number of samples with a label of y_0 or y_1, which is $\frac{2N}{k}$. Thus we have $Pr[f_{T(D_N)}(x_0) = y_0] = Pr[f_{D_N}(x_0) = y_1] \leq \frac{1}{|Y|}$ and $\mathbb{E}[|T(D_N) - D_N|] = \frac{2N}{k} = \Theta(\frac{1}{k}) \cdot N$.

In both cases, we have a transform T that minimize the accuracy of f while in expectation altering no more than $\Theta(1/k)$ of the training set and therefore the proof completes. Note the underlying assumption used in case 2 is that the Bijection g is unknown to learners in a sense that the output distributions of f change accordingly when labels are permuted. □

C Proof of Lemma 3

Proof. Given the number of clean samples n, for any learner $f \in \mathcal{F}$, the accuracy of f on x_0 is:

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] =& Pr[(x_0, y_0) \in D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \in D_n] \\ &+ Pr[(x_0, y_0) \notin D_n] \cdot Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n], \end{aligned} \tag{2}$$

where $Pr[(x_0, y_0) \in D_n] = 1 - (1 - 1/m)^n$ and $Pr[(x_0, y_0) \notin D_n] = (1 - 1/m)^n$.

Since g is a mapping that assigns labels independently to different inputs and it is unknown to learners, we have $Pr[f_{D_n}(x_0) = y_0 \mid (x_0, y_0) \notin D_n] \leq \frac{1}{k}$ and therefore with Equation 2, we have

$$\begin{aligned} Pr[f_{D_n}(x_0) = y_0] \leq& Pr[(x_0, y_0) \in D_n] \cdot 1 + Pr[(x_0, y_0) \notin D_n] \cdot \frac{1}{k} \\ =& 1 - Pr[(x_0, y_0) \notin D_n] \cdot (1 - \frac{1}{k}) \\ =& 1 - \left(1 - \frac{1}{m}\right)^n \cdot \left(1 - \frac{1}{k}\right). \end{aligned}$$

Thus $Pr[f_{D_n}(x_0) = y_0] \geq \tau \Rightarrow \left(1 - \frac{1}{m}\right)^n \leq \frac{k(1-\tau)}{k-1} \Rightarrow n \geq \frac{\log(1-\tau)+\log(1+1/(k-1))}{\log(1-1/m)} = \Theta(m)$.

The intuition behind the proof: If the training set contains (x_0, y_0), the most data-efficient learner will memorize it to predict correctly; Otherwise it can do nothing but guess an arbitrary label. □

D Proof of Lemma 4

Proof. Given x_0, for any N and any learner f, we let T be a transform that obtain the poisoned training set $T(D)$ by removing all (x_0, y_0) from the clean training set D. By definition of Instance Memorization, there is no information regarding y_0 contained in $T(D_N)$ and therefore we have $Pr[f_{T(D_N)}(x_0) = y_0] =\leq \frac{1}{|Y|}$ given that the mapping g is unknown to learners. Meanwhile, we have $\mathbb{E}[|T(D_N) - D_N|] = \frac{N}{m} = \Theta(\frac{1}{m}) \cdot N$. □

E Proof of Theorem 1

Proof. First we introduce Coupling Lemma as a tool.

Coupling Lemma [19]: *For two distributions U and V over Ω, a coupling W is a distribution over $\Omega \times \Omega$ such that the marginal distributions are the same as U and V, i.e. $U(u) = \int_{v \in \Omega} W(u, v)dv$ and $V(v) = \int_{u \in \Omega} W(u, v)du$. The lemma states that for two distributions U and V, there exists a coupling W such that $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$.*

Intuitively, Coupling Lemma suggests there is a correspondence between the two distributions U and V, such that only the mass within their difference $\delta(U, V)$ will correspond to different elements.

Through coupling lemma, there is a coupling W with $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$.

$$\begin{aligned} \mathbb{E}_{D\sim U^n}[f(D)] - \mathbb{E}_{D\sim V^n}[f(D)] =& \mathbb{E}_{(\forall 1\leq i\leq n)\ (u_i,v_i)\sim W}\left[f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n)\right] \\ \leq& Pr_{(\forall 1\leq i\leq n)\ (u_i,v_i)\sim W}[(\exists i)\ u_i \neq v_i] \\ \leq& \sum_{i=1}^{n} Pr_{(u_i,v_i)\sim W}[u_i \neq v_i] \\ =& n \cdot \delta(U, V). \end{aligned}$$

The second line in the above inequalities is derived as follows: When $u_i = v_i$ for all i, we have $f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n) = 0$; When there exists $u_i \neq v_i$ for some i, we have $f(\{u_i\}_{i=1}^n) - f(\{v_i\}_{i=1}^n) \leq 1$ because the output of f is $\{0, 1\}$.

For the third line, we use the union bound. The probability that for at least one i we have $u_i \neq v_i$ is upper bounded by the sum of probability that $u_i \neq v_i$ for all i.

□

F Proof of Theorem 2

Proof. Through coupling lemma, there is a coupling W with $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$. We define the mapping T as follows: For any $D = (u_1, \ldots, u_N) \in \Omega^N$, the output $T(D) = (v_1, \ldots, v_N)$ is obtained by drawing v_i from $W(v \mid u)$ independently for $i = 1 \ldots N$.

For $i = 1 \ldots N$, we have $P(v_i) = \int_{u_i \in \Omega} W(v_i \mid u_i)U(u_i)du_i = \int_{u_i \in \Omega} W(v_i \mid u_i)W(u_i) = V(v_i)$ and $T(U^N)$ is the same distribution as V^N, meaning that $\mathbb{E}_{D\sim U^N}[f(T(D))] - \mathbb{E}_{D\sim V^N}[f(D)] = 0$; Meanwhile, given $Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V)$, we have $\mathbb{E}_{D\sim U^N}[|T(D) - D|] = \sum_{i=1}^N Pr_{u\sim U, v\sim W(v|u)}[u \neq v] = \sum_{i=1}^N Pr_{(u,v)\sim W}[u \neq v] = \delta(U, V) \cdot N$. □

G Proof of Lemma 6

Proof. We will use Theorem 1. In order to have $Pr[f_{D_n}(x_0) = y_0] \geq \tau$, there must be some $g : \Omega^{\mathbb{N}} \to \{0, 1\}$ discriminating $U = \mathcal{N}(\mu_1, I)$ and $V = \mathcal{N}(\mu_1', I)$ with confidence larger than a constant τ' using in expectation n/k samples, where $\mu_1' = \mu_1 + (d_2 - d_1)/d_1 \cdot (1 + \epsilon)(\mu_1 - x_0)$ for some $\epsilon > 0$. Note that $||\mu_1' - \mu_1|| = (1 + \epsilon)(d_2 - d_1)$ and $||\mu_1' - x_0|| > d_2$. Taking $\epsilon \to 0$, we have $n/k \geq \Theta(1/\delta(U, V)) = \Theta(1/\Delta) \Rightarrow n \geq \Theta(k/\Delta)$ using Theorem 1. □

H Proof of Lemma 7

Proof. We will use Theorem 2. Given x_0, for any N and any learner f, let $\mu_2' = \mu_2 - (d_2 - d_1)/d_2 \cdot (1 + \epsilon)(\mu_2 - x_0)$ for some $\epsilon > 0$, as shown in Figure 1(b). With Theorem 2, there is a transform T' making $U = \mathbb{N}(\mu_2, I)$ and $V = \mathbb{N}(\mu_2', I)$ indistinguishable. We define a transform T by applying T' to the inputs of all samples from class 2 while others remain unchanged.

Note that $f \in \mathcal{F}$ and $T(P)$ is also a plausible distribution (in a sense that it can be expressed in the same form as Equation 1). Since $||\mu_2' - x_0|| < d_1$, we have $Pr[f_{T(D_N)}(x_0) = y_0] \leq 1/k$. In addition, since $||\mu_2 - \mu_2'|| = (1 + \epsilon)(d_2 - d_1)$ and there are in expectation N/k samples from class 2, we have $\mathbb{E}[|T(D_N) - D_N|] \leq \delta(U, V)/k = \Theta(\Delta/k)$ by taking $\epsilon \to 0$. □

Intuition for taking $\epsilon \to 0$: When ϵ is actually 0, the distributions we construct for different classes will be 'symmetric' to x_0, meaning that there will be a tie in defining the maximum likelihood prediction. For any $\epsilon > 0$, the tie will be broken. By letting $\epsilon \to 0$, we find the tightest bound of the number of poisoned samples needed from our construction.