A. Proof of Theorem 1 and its corollaries

Theorem 1. *Let $N, m, k \in \mathbb{N}$ be integers, such that $k \in (N/2, N]$. Let $\alpha = \frac{N-k}{N} < \frac{1}{2}$ be the proportion of corrupted sources. Assume that $\mathcal{H}$ has the uniform convergence property with rate function s. Then there exists a learner $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$ with the following two properties.*

(a) Let G be a fixed subset of $[N]$ of size $|G| = k$. For $S^{'} = \{S_1^{'}, \ldots, S_N^{'}\} \overset{i.i.d.}{\sim} \mathcal{D}$, with probability at least $1 - \delta$ over the sampling of S':

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S^{'}))) - \min_{h \in \mathcal{H}} \mathcal{R}(h) \leq 2s\big(km, \frac{\delta}{2}, S_G\big) + 6\alpha \max_{i \in [N]} s\big(m, \frac{\delta}{2N}, S_i\big), \tag{22}$$

uniformly against all fixed-set adversaries with preserved set G, where $S = \{S_1, \ldots, S_N\} = \mathcal{A}(S^{'})$ is the dataset modified the adversary and $S_G = \cup_{i \in G} S_i$ is the set of all uncorrupted data.

(b) For $S^{'} = \{S_1^{'}, \ldots, S_N^{'}\} \overset{i.i.d.}{\sim} \mathcal{D}$, with probability at least $1 - \delta$ over the sampling of S':

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S^{'}))) - \min_{h \in \mathcal{H}} \mathcal{R}(h) \leq 2s\big(km, \frac{\delta}{2\binom{N}{k}}, S_G\big) + 6\alpha \max_{i \in [N]} s\big(m, \frac{\delta}{2N}, S_i\big), \tag{23}$$

uniformly against all flexible-set adversaries with preserved size k, where $S = \{S_1, \ldots, S_N\} = \mathcal{A}(S^{'})$ is the dataset returned by the adversary, G is the set of sources not modified by the adversary and $S_G = \cup_{i \in G} S_i$ is the set of all uncorrupted data.

Proof. Denote by $S_i' = \{(x_{i,1}', y_{i,1}'), \ldots, (x_{i,m}', y_{i,m}')\}$ for $i = 1, \ldots, N$ the initial datasets and by $S_i = \{(x_{i,1}, y_{i,1}), \ldots, (x_{i,m}, y_{i,m})\}$ for $i = 1, \ldots, N$ the datasets after the modifications of the adversary. As explained in the main body of the paper, we denote by:

$$\widehat{\mathcal{R}}_i(h) = \frac{1}{m} \sum_{j=1}^{m} \ell(h(x_{i,j}), y_{i,j}) \tag{24}$$

the empirical risk of any hypothesis $h \in \mathcal{H}$ on the dataset S_i and by:

$$d_{\mathcal{H}}(S_i, S_j) = \sup_{h \in \mathcal{H}} |\widehat{\mathcal{R}}_i(h) - \widehat{\mathcal{R}}_j(h)| \tag{25}$$

the empirical discrepancy between the datasets S_i and S_j.

We show that a learner that first runs a certain filtering algorithm (Algorithm 1) based on the discrepancy metric and then performs empirical risk minimization on the remaining data to compute a hypothesis satisfies the properties stated in the theorem. The full algorithm for the learner is therefore given in Algorithm 2.

(a) The key idea of the proof is that the clean sources are close to each other with high probability, so they get selected when running Algorithm 1. On the other hand, if a bad source has been selected, it must be close to *at least one* of the good sources, so it can not have too bad an effect on the empirical risk.

Algorithm 1

input Datasets $S_1, \ldots, S_N$
  Initialize $\mathrm{T} = \{\}$ // trusted sources
  for $i = 1, \ldots, N$ **do**
    if $d_{\mathcal{H}}\big(S_i, S_j\big) \leq s\left(m, \frac{\delta}{2N}, S_i\right) + s\left(m, \frac{\delta}{2N}, S_j\right)$,
      for at least $\lfloor \frac{N}{2} \rfloor$ values of $j \neq i$, **then**
      $\mathrm{T} = \mathrm{T} \cup \{i\}$
    end if
  end for
output $\bigcup_{i \in \mathrm{T}} S_i$ // indices of the trusted sources

Algorithm 2

input Datasets $S_1, \ldots, S_N$
Run Algorithm 1 to compute $S_{\mathrm{T}} = \bigcup_{i \in \mathrm{T}} S_i$
Compute $h^{\mathcal{A}} = \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{|S_{\mathrm{T}}|} \sum_{(x,y) \in S_{\mathrm{T}}} \ell(h(x), y)$ //empirical risk minimizer over all trusted sources
output $h^{\mathcal{A}}$

For all $i \in G$, let $\mathcal{E}_i$ be the event that:

$$\sup_{h \in \mathcal{H}} \left| \mathcal{R}(h) - \widehat{\mathcal{R}}_i(h) \right| \leq s\left(m, \frac{\delta}{2N}, S_i\right). \tag{26}$$

Further, let $\mathcal{E}_G$ be the event that:

$$\sup_{h \in \mathcal{H}} \left| \mathcal{R}(h) - \widehat{\mathcal{R}}_G(h) \right| \leq s\left(km, \frac{\delta}{2}, S_G\right), \tag{27}$$

where

$$\widehat{\mathcal{R}}_G(h) = \frac{1}{km} \sum_{i \in G} \sum_{j=1}^{m} \ell(h(x_{i,j}), y_{i,j}).$$

Denote by $\mathcal{E}_i^c$ and $\mathcal{E}_G^c$ the complements of these events. Then we know that $\mathbb{P}\left(\mathcal{E}_G^c\right) \leq \frac{\delta}{2}$, and $\mathbb{P}\left(\mathcal{E}_i^c\right) \leq \frac{\delta}{2N}$ for all $i \in G$. Therefore, if $\mathcal{E} = \mathcal{E}_G \wedge (\wedge_{i \in G} \mathcal{E}_i)$, we have:

$$\mathbb{P}\left(\mathcal{E}^c\right) = \mathbb{P}\left(\mathcal{E}_G^c \vee (\vee_{i \in G} \mathcal{E}_i^c)\right) \leq \mathbb{P}\left(\mathcal{E}_G^c\right) + \sum_{i \in G} \mathbb{P}\left(\mathcal{E}_i^c\right) \leq \frac{\delta}{2} + k\frac{\delta}{2N} \leq \delta. \tag{28}$$

Hence, the probability of the event $\mathcal{E}$ that all of (26) and (27) hold, is at least $1 - \delta$. We now show that under the event $\mathcal{E}$, Algorithm 2 returns a hypothesis that satisfies the condition in (a).

Indeed, fix an arbitrary fixed-set adversary $\mathcal{A}$ with preserved set G. Whenever $\mathcal{E}$ holds, for all $i, j \in G$ we have:

$$\begin{aligned} d_{\mathcal{H}}\left(S_i, S_j\right) = \sup_{h \in \mathcal{H}} (|\widehat{\mathcal{R}}_i(h) - \widehat{\mathcal{R}}_j(h)|) &\leq \sup_{h \in \mathcal{H}} \left(|\widehat{\mathcal{R}}_i(h) - \mathcal{R}(h)|\right) + \sup_{h \in \mathcal{H}} \left(|\mathcal{R}(h) - \widehat{\mathcal{R}}_j(h)|\right) \\ &\leq s\left(m, \frac{\delta}{2N}, S_i\right) + s\left(m, \frac{\delta}{2N}, S_j\right). \end{aligned} \tag{29}$$

Now since $k \geq \lfloor \frac{N}{2} \rfloor + 1$, we get that $G \subseteq \mathrm{T}$. Moreover, for any $i \in \mathrm{T} \setminus G$, there exists *at least one* $j \in G$, such that $d_{\mathcal{H}}(S_i, S_j) \leq s\left(m, \frac{\delta}{2N}, S_i\right) + s\left(m, \frac{\delta}{2N}, S_j\right)$. For any $i \in \mathrm{T} \setminus G$, denote by $f(i)$ the smallest such j. Therefore, for any $i \in (\mathrm{T} \setminus G)$:

$$|\widehat{\mathcal{R}}_i(h) - \mathcal{R}(h)| \leq |\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_{f(i)}(h)| + |\widehat{\mathcal{R}}_{f(i)}(h) - \mathcal{R}(h)| \leq d_{\mathcal{H}}\left(S_i, S_{f(i)}\right) + s\left(m, \frac{\delta}{2N}, S_{f(i)}\right) \tag{30}$$

$$\leq s\left(m, \frac{\delta}{2N}, S_i\right) + 2s\left(m, \frac{\delta}{2N}, S_{f(i)}\right) \tag{31}$$

Denote by

$$\widehat{\mathcal{R}}_{\mathrm{T}}(h) = \frac{1}{|\mathrm{T}|} \sum_{i \in \mathrm{T}} \widehat{\mathcal{R}}_i(h) = \frac{1}{|S_{\mathrm{T}}|} \sum_{(x,y) \in S_{\mathrm{T}}} \ell(h(x), y) \tag{32}$$

the loss over all the trusted data. Then for any $h \in \mathcal{H}$ we have:

$$\left|\widehat{\mathcal{R}}_{\mathrm{T}}(h) - \mathcal{R}(h)\right| \leq \frac{1}{|\mathrm{T}|m} \left(\left| \sum_{i \in G} \sum_{l=1}^{m} \left(\ell(h(x_{i,l}), y_{i,l}) - \mathcal{R}(h)\right) \right| + \sum_{i \in (\mathrm{T} \setminus G)} \left| \sum_{l=1}^{m} \left(\ell(h(x_{i,l}), y_{i,l}) - \mathcal{R}(h)\right) \right| \right) \tag{33}$$

$$= \frac{k}{|\mathrm{T}|} \left|\widehat{\mathcal{R}}_G(h) - \mathcal{R}(h)\right| + \frac{1}{|\mathrm{T}|} \sum_{i \in (\mathrm{T} \setminus G)} \left|\widehat{\mathcal{R}}_i(h) - \mathcal{R}(h)\right| \tag{34}$$

$$\leq \frac{k}{|\mathrm{T}|} s\left(km, \frac{\delta}{2}, S_G\right) + \frac{1}{|\mathrm{T}|} \sum_{i \in (\mathrm{T} \setminus G)} \left|\widehat{\mathcal{R}}_i(h) - \mathcal{R}(h)\right| \tag{35}$$

$$\leq \frac{k}{|\mathrm{T}|} s\left(km, \frac{\delta}{2}, S_G\right) + \frac{1}{|\mathrm{T}|} \sum_{i \in (\mathrm{T} \setminus G)} \left(s\left(m, \frac{\delta}{2N}, S_i\right) + 2s\left(m, \frac{\delta}{2N}, S_{f(i)}\right)\right) \tag{36}$$

$$\leq \frac{k}{|\mathrm{T}|} s\left(km, \frac{\delta}{2}, S_G\right) + 3\frac{|\mathrm{T}| - k}{|\mathrm{T}|} \max_{i \in [N]} s\left(m, \frac{\delta}{2N}, S_i\right) \tag{37}$$

$$\leq s\left(km, \frac{\delta}{2}, S_G\right) + 3\frac{N - k}{N} \max_{i \in [N]} s\left(m, \frac{\delta}{2N}, S_i\right) \tag{38}$$

Finally, let $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \mathcal{R}(h)$ and $h^{\mathcal{A}} = \mathcal{L}(\mathcal{A}(S')) = \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathcal{R}}_{\mathrm{T}}(h)$. Then:

$$\mathcal{R}(h^{\mathcal{A}}) - \mathcal{R}(h^*) = \left(\mathcal{R}(h^{\mathcal{A}}) - \widehat{\mathcal{R}}_{\mathrm{T}}(h^{\mathcal{A}})\right) + \left(\widehat{\mathcal{R}}_{\mathrm{T}}(h^{\mathcal{A}}) - \mathcal{R}(h^*)\right) \leq \left(\mathcal{R}(h^{\mathcal{A}}) - \widehat{\mathcal{R}}_{\mathrm{T}}(h^{\mathcal{A}})\right) + \left(\widehat{\mathcal{R}}_{\mathrm{T}}(h^*) - \mathcal{R}(h^*)\right) \tag{39}$$

$$\leq 2 \sup_{h \in \mathcal{H}} \left|\widehat{\mathcal{R}}_{\mathrm{T}}(h) - \mathcal{R}(h)\right|. \tag{40}$$

Since we showed this result for an arbitrary fixed-set adversary with preserved set G, the result follows.

(b) The crucial difference in the case of the flexible-set adversary is that the set G is chosen after the clean data is observed. We thus need concentration results for *all* of the subsets of $[N]$ of size k, as well as all individual sources.

For all $i \in [N]$, let $\mathcal{E}_i$ be the event that:

$$\sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}'_i(h)\right| \leq s\left(m, \frac{\delta}{2N}, S'_i\right), \tag{41}$$

where

$$\widehat{\mathcal{R}}'_i = \frac{1}{m} \sum_{j=1}^{m} \ell(h(x'_{i,j}), y'_{i,j}) \tag{42}$$

Further, for *any* $A \subseteq [N]$ of size $|A| = k$, let $\mathcal{E}_A$ be the event that:

$$\sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}'_A(h)\right| \leq s\left(km, \frac{\delta}{2\binom{N}{k}}, S'_A\right), \tag{43}$$

where $S'_A = \cup_{i \in A} S'_i$ and

$$\widehat{\mathcal{R}}'_A(h) = \frac{1}{km} \sum_{i \in A} \sum_{l=1}^{m} \ell(h(x'_{i,l}), y'_{i,l}). \tag{44}$$

Then we know that $\mathbb{P}\left(\mathcal{E}_i^c\right) \leq \frac{\delta}{2N}$ for all $i \in [N]$ and $\mathbb{P}\left(\mathcal{E}_G^c\right) \leq \frac{\delta}{2\binom{N}{k}}$ for all $A \subseteq [N]$ with $|A| = k$. Therefore, if $\mathcal{E} = (\wedge_A \mathcal{E}_A) \wedge \left(\wedge_{i \in [N]} \mathcal{E}_i\right)$, we have:

$$\mathbb{P}\left(\mathcal{E}^c\right) = \mathbb{P}\left((\vee_A \mathcal{E}_A^c) \vee \left(\vee_{i \in [N]} \mathcal{E}_i^c\right)\right) \leq \sum_A \mathbb{P}\left(\mathcal{E}_A^c\right) + \sum_{i \in [N]} \mathbb{P}\left(\mathcal{E}_i^c\right) \leq \binom{N}{k} \frac{\delta}{2\binom{N}{k}} + N\frac{\delta}{2N} = \delta. \tag{45}$$

Hence, the probability of the event $\mathcal{E}$ that all of (41) and (43) hold, is at least $1 - \delta$. In particular, under $\mathcal{E}$:

$$\sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}_G(h)\right| = \sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}'_G(h)\right| \leq s\left(km, \frac{\delta}{2\binom{N}{k}}, S'_G\right) = s\left(km, \frac{\delta}{2\binom{N}{k}}, S_G\right) \tag{46}$$

and

$$\sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}_i(h)\right| = \sup_{h \in \mathcal{H}} \left|\mathcal{R}(h) - \widehat{\mathcal{R}}'_i(h)\right| \leq s\left(m, \frac{\delta}{2N}, S'_i\right) = s\left(m, \frac{\delta}{2N}, S_i\right), \tag{47}$$

for all $i \in G$.

Now, for any flexible-set adversary with preserved size k, the same argument as in (a) shows that:

$$\mathcal{R}(h^{\mathcal{A}}) - \mathcal{R}(h^*) \leq 2s\left(km, \frac{\delta}{2\binom{N}{k}}, S_G\right) + 6\frac{N-k}{N}\max_{i\in[N]} s\left(m, \frac{\delta}{2N}, S_i\right) \tag{48}$$

holds under the event $\mathcal{E}$. □

We now show how to obtain data-dependent guarantees, via the notion of Rademacher complexity. Let

$$\mathfrak{R}_S\left(\ell \circ \mathcal{H}\right) = \mathbb{E}_\sigma\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(h(x_i), y_i)\right) \tag{49}$$

be the Rademacher complexity of $\mathcal{H}$ with respect to the loss function ℓ on a sample $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Let $S_G = \cup_{i\in G} S_i$, $\mathfrak{R}_i = \mathfrak{R}_{S_i}\left(\ell \circ \mathcal{H}\right)$ and $\mathfrak{R}_G = \mathfrak{R}_{S_G}\left(\ell \circ \mathcal{H}\right)$. Then we have:

Corollary 2. *In the setup of Theorem 1, against a fixed-set adversary, it holds that*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S^{'}))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 4\mathfrak{R}_G + 6\sqrt{\frac{\log(\frac{4}{\delta})}{2km}} + \alpha\Big(18\sqrt{\frac{\log\left(\frac{4N}{\delta}\right)}{2m}} + 12\max_{i\in[N]}\mathfrak{R}_i\Big). \tag{50}$$

Proof. We use the standard generalization bound based on Rademacher complexity. Assume that $S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \sim \mathcal{D}$, then with probability at least $1-\delta$ over the data (Mohri et al., 2018):

$$\sup_{h\in\mathcal{H}} \left|\mathbb{E}\left(\ell(h(x), y)\right) - \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)\right| \leq 2\mathfrak{R}_S\left(\ell \circ \mathcal{H}\right) + 3\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}. \tag{51}$$

Substituting into the result of Theorem 1 gives the result. □

Corollary 3. *In the setup of Theorem 1, against a flexible-set adversary, it holds that*

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S^{'}))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 4\mathfrak{R}_G + 12\alpha\max_{i\in[N]}\mathfrak{R}_i + \widetilde{\mathcal{O}}\left(\frac{\sqrt[4]{\alpha}}{\sqrt{m}}\right). \tag{52}$$

Proof. Using the concentration result from Corollary 2 and $\binom{N}{k} = \binom{N}{(1-\alpha)N} = \binom{N}{\alpha N} \leq 2^{H(\alpha)N}$, where $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ is the binary entropy function, we obtain:

$$\begin{aligned}
\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) &\leq 4\mathfrak{R}_G + 6\sqrt{\frac{\log(\frac{4\binom{N}{k}}{\delta})}{2km}} + \alpha\left(18\sqrt{\frac{\log\left(\frac{4N}{\delta}\right)}{2m}} + 12\max_{i\in[N]}\mathfrak{R}_i\right) && (53)\\
&= 4\mathfrak{R}_G + 6\sqrt{\frac{\log(\binom{N}{k})}{2km} + \frac{\log(\frac{4}{\delta})}{2km}} + \alpha\left(18\sqrt{\frac{\log\left(\frac{4N}{\delta}\right)}{2m}} + 12\max_{i\in[N]}\mathfrak{R}_i\right) && (54)\\
&\leq 4\mathfrak{R}_G + 6\sqrt{\frac{H(\alpha)N\log(2)}{2(1-\alpha)Nm} + \frac{\log(\frac{4}{\delta})}{2(1-\alpha)Nm}} + \alpha\left(18\sqrt{\frac{\log\left(\frac{4N}{\delta}\right)}{2m}} + 12\max_{i\in[N]}\mathfrak{R}_i\right) && (55)\\
&\leq 4\mathfrak{R}_G + 12\alpha\max_{i\in[N]}\mathfrak{R}_i + \widetilde{\mathcal{O}}\left(\frac{\sqrt[4]{\alpha}}{\sqrt{m}}\right) && (56)
\end{aligned}$$

where for the last inequality we used $H(\alpha) \leq 2\sqrt{\alpha(1-\alpha)}$, $1-\alpha \in (\frac{1}{2}, 1]$ and $\sqrt[4]{\alpha} > \alpha$. □

For the case of binary classifiers, we also provide a simpler bound in terms of the VC dimension of $\mathcal{H}$.

Corollary 4. *Assume that $Y = \{-1, 1\}$ and that $\mathcal{H}$ has finite VC-dimension d. Then:*

(a) In the case of the fixed-set adversary there exists a universal constant C, such that:

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 2C\sqrt{\frac{d}{km}} + 2\sqrt{\frac{2\log(\frac{4}{\delta})}{km}} + \alpha\left(6C\sqrt{\frac{d}{m}} + 6\sqrt{\frac{2\log\left(\frac{4N}{\delta}\right)}{m}}\right). \tag{57}$$

(b) In the case of the flexible-set adversary:

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq \mathcal{O}\left(\sqrt{\frac{d}{km}} + \frac{\sqrt[4]{\alpha}}{\sqrt{m}} + \alpha\sqrt{\frac{d}{m}} + \alpha\sqrt{\frac{\log(N)}{m}}\right). \tag{58}$$

Proof. (a) Whenever $\mathcal{H}$ is of finite VC-dimension d, there exists a constant C, such that the following generalization bound holds (Bousquet et al., 2004):

$$\sup_{h\in\mathcal{H}} |\mathbb{E}\left(\ell(h(x), y)\right) - \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)| \leq C\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}} \tag{59}$$

and hence $\mathcal{H}$ has the uniform convergence property with rate function $s = C\sqrt{\frac{d}{n}} + \sqrt{\frac{2\log\left(\frac{2}{\delta}\right)}{n}}$. Substituting into the result of Theorem 1 gives the result.

(b) Using the concentration result from (a) and $\binom{N}{k} = \binom{N}{(1-\alpha)N} = \binom{N}{\alpha N} \leq 2^{H(\alpha)N}$, where $H(p) = -p\log_2(p) - (1-p)\log_2(1-p)$ is the binary entropy function, we obtain:

$$\mathcal{R}(\mathcal{L}(\mathcal{A}(S'))) - \min_{h\in\mathcal{H}} \mathcal{R}(h) \leq 2C\sqrt{\frac{d}{km}} + 2\sqrt{\frac{2\log(\frac{4\binom{N}{k}}{\delta})}{km}} + \alpha\left(6C\sqrt{\frac{d}{m}} + 6\sqrt{\frac{2\log\left(\frac{4N}{\delta}\right)}{m}}\right) \tag{60}$$

$$= 2C\sqrt{\frac{d}{km}} + 2\sqrt{\frac{2\log(\binom{N}{k})}{km} + \frac{2\log(\frac{4}{\delta})}{km}} + \alpha\left(6C\sqrt{\frac{d}{m}} + 6\sqrt{\frac{2\log\left(\frac{4N}{\delta}\right)}{m}}\right) \tag{61}$$

$$\leq 2C\sqrt{\frac{d}{km}} + 2\sqrt{\frac{2H(\alpha)N\log(2)}{(1-\alpha)Nm} + \frac{2\log(\frac{4}{\delta})}{(1-\alpha)Nm}} + \alpha\left(6C\sqrt{\frac{d}{m}} + 6\sqrt{\frac{2\log\left(\frac{4N}{\delta}\right)}{m}}\right) \tag{62}$$

$$\leq \mathcal{O}\left(\sqrt{\frac{d}{km}} + \frac{\sqrt[4]{\alpha}}{\sqrt{m}} + \alpha\sqrt{\frac{d}{m}} + \alpha\sqrt{\frac{\log(N)}{m}}\right), \tag{63}$$

where for the last inequality we used $H(\alpha) \leq 2\sqrt{\alpha(1-\alpha)}$ and $1 - \alpha \in (\frac{1}{2}, 1]$. □

B. Proof of Theorem 2

Theorem 2. *Let $\mathcal{H}$ be a non-trivial hypothesis space. Let m and N be any positive integers and let G be a fixed subset of $[N]$ of size $k \in \{1, \dots, N-1\}$. Let $\mathcal{L} : (\mathcal{X} \times \mathcal{Y})^{N \times m} \to \mathcal{H}$ be a multi-source learner that acts by merging the data from all sources and then calling a single-source learner. Let $S' \in (\mathcal{X} \times \mathcal{Y})^{N \times m}$ be drawn i.i.d. from $\mathcal{D}$. Then there exists a distribution $\mathcal{D}$ with $\min_{h \in \mathcal{H}} \mathcal{R}(h) = 0$ and a fixed-set adversary $\mathcal{A}$ with index set G, such that:*

$$\mathbb{P}_{S' \sim \mathcal{D}}\Big(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S')\big) > \frac{\alpha}{8(1-\alpha)}\Big) > \frac{1}{20}, \tag{64}$$

where $\alpha = \frac{N-k}{N}$ is the power of the adversary.

We use a similar proof technique as in the lower bound results in (Bshouty et al., 2002) and in the classic sample complexity lower bound for binary classification, e.g. Theorem 3.20 in (Mohri et al., 2018). An overview is as follows. Consider a distribution on $\mathcal{X}$ that has support only at two points - the common point x_1 and the rare point x_2. Take $\mathbb{P}(x_2) = \mathcal{O}(\frac{\alpha}{1-\alpha})$. Then the expected number of occurrences of the point x_2 in G is $\mathcal{O}\left(\frac{\alpha}{1-\alpha}(1-\alpha)Nm\right) = \mathcal{O}\left(\alpha Nm\right)$. Thus, one can show that with constant probability the number of x_2's in G is at most αNm and hence the adversary (that has access to exactly αNm points in total) can insert the same number of x_2's, but wrongly labelled, into the final dataset. Therefore, based on the union of the corrupted datasets, no algorithm can guess with probability greater than $1/2$ what the true label of x_2 was.

Proof. We prove that there exists a distribution $\mathcal{D}$ on $\mathcal{X}$ and a labelling function $f \in \mathcal{H}$, such that the resulting joint distribution on $\mathcal{X} \times \mathcal{Y}$, defined by $x \sim \mathcal{D}$ and $y = f(x)$, satisfies the desired property.

Without loss of generality, let $G = [1, 2, \dots, k]$. Since $\mathcal{H}$ is non-trivial, there exist $h_1, h_2 \in \mathcal{H}$ and $x_1, x_2 \in \mathcal{X}$, such that $h_1(x_1) = h_2(x_1)$, while $h_1(x_2) = 1$, but $h_2(x_2) = -1$. Consider the following distribution on $\mathcal{X}$:

$$\mathbb{P}_{\mathcal{D}}(x_1) = 1 - 4\epsilon \quad \text{and} \quad \mathbb{P}_{\mathcal{D}}(x_2) = 4\epsilon, \tag{65}$$

where $\epsilon = \frac{1}{8}\frac{\alpha}{1-\alpha}$. Assume that the points are labelled by a function $f \in \mathcal{H}$ (to be chosen later as either h_1 or h_2). Denote the initial uncorrupted collection of datasets by $S' = (S'_1, \dots, S'_N)$, with $S'_i = \{(x'_{i,1}, f(x'_{i,1})), \dots, (x'_{i,m}, f(x'_{i,m}))\}$ and $x'_{i,j}$ being i.i.d. samples from $\mathcal{D}$.

First we show that with constant probability the point x_2 appears at most αNm times in G. Indeed, let C be this number of appearances. Then C is a binomial random variable with probability of success 4ϵ and number of trials $(1-\alpha)Nm$. Therefore, by the Chernoff bound:

$$\mathbb{P}_{S'}(C \geq \alpha Nm) = \mathbb{P}_{S'}(C \geq (1+1)4\epsilon(1-\alpha)Nm) \leq e^{-\frac{\alpha Nm}{6}} \leq e^{-1/6} < \frac{17}{20} \tag{66}$$

and so:

$$\mathbb{P}_{S'}(C \leq \alpha Nm) > \frac{3}{20}. \tag{67}$$

Now consider the following policy for the fixed-set adversary $\mathcal{A}^s : S' \to S$. For any index $i \in [N]$ the adversary replaces $S'_i = \{(x'_{i,1}, f(x'_{i,1})), \dots, (x'_{i,m}, f(x'_{i,m}))\}$ with a dataset $S_i = \{(x_{i,1}, y_{i,1}) \dots, (x_{i,m}, y_{i,m})\}$, such that:

$$(x_{i,j}, y_{i,j}) = \begin{cases} (x'_{i,j}, f(x'_{i,j})), & \text{if } i \in G = [1, 2, \dots, k] \\ (x_2, -f(x_2)), & \text{if } i \in [k+1, \dots, N] \text{ and } (i-k-1)m + j \leq C \\ (x_1, f(x_1)), & \text{otherwise} \end{cases} \tag{68}$$

Then the adversary returns $S = (S_1, \dots, S_N)$. That is, the adversary keeps the datasets in G untouched, and fills the datasets in $[N] \backslash G$ with as many x_2's as there are in G, but wrongly labelled.

Crucially, whenever $C \leq \alpha Nm$, the union of the data in all N sets will look the same no matter if the original labelling function was h_1 or h_2. In particular, $\mathcal{L}(\mathcal{A}^s(S'))$ will be identical in both cases.

Finally, we argue that under the event $C \leq \alpha Nm$ and the chosen adversary, the learner would incur high loss and show that this implies the result in (19). Let $\mathcal{S}$ be the set of all datasets in $(\mathcal{X} \times \mathcal{Y})^{N \times m}$, such that $C \leq \alpha Nm$ holds. We just showed

that $\mathbb{P}_{S'}(S' \in \mathcal{S}) > \frac{3}{20}$ and that whenever $S' \in \mathcal{S}$, $\mathcal{L}(\mathcal{A}^s(S'))$ is independent of whether the original labelling function was h_1 or h_2.

Consider a fixed set $S' \in \mathcal{S}$ and let $S = \mathcal{A}^s(S')$ and $h_S = \mathcal{L}(S)$. Denote by $\mathcal{R}(h_S, f) = \mathbb{P}_{\mathcal{D}}(h_S(x) \neq f(x) \cap x \neq x_1)$ and note that $\mathcal{R}(h_S, f) \leq \mathbb{P}_{\mathcal{D}}(h_S(x) \neq f(x)) = \mathcal{R}(\mathcal{L}(\mathcal{A}^s(S')))$. Notice that:

$$\mathcal{R}(h_S, h_1) + \mathcal{R}(h_S, h_2) = \sum_{i=1,2} \mathbb{1}_{h_S(x_i) \neq h_1(x_i)} \mathbb{1}_{x_i \neq x_1} \mathbb{P}(x_i) + \sum_{i=1,2} \mathbb{1}_{h_S(x_i) \neq h_2(x_i)} \mathbb{1}_{x_i \neq x_1} \mathbb{P}(x_i) \tag{69}$$

$$= \mathbb{1}_{h_S(x_2) \neq h_1(x_2)} 4\epsilon + \mathbb{1}_{h_S(x_2) \neq h_2(x_2)} 4\epsilon \tag{70}$$

$$= 4\epsilon, \tag{71}$$

where we used that $h_1(x_2) = 1 = -h_2(x_2)$ and that h_S is independent of the underlying labelling function.

Since the above holds for any $S' \in \mathcal{S}$, it also holds in expectation, conditioned on $S' \in \mathcal{S}$:

$$\mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, h_1) + \mathcal{R}(h_S, h_2) \right) \geq 4\epsilon. \tag{72}$$

Therefore, $\mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, h_i) \right) \geq 2\epsilon$ for at least one of $i = 1, 2$. Take f to be h_1, if h_1 satisfies the inequality, and h_2 otherwise. Conditioning on $\{\mathcal{R}(h_S, f) \geq \epsilon\}$ and using $\mathcal{R}(h_S, f) \leq \mathbb{P}_{\mathcal{D}}(x \neq x_1) = 4\epsilon$:

$$2\epsilon \leq \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \right) = \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) | \mathcal{R}(h_S, f) \geq \epsilon \right) \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \tag{73}$$

$$+ \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) | \mathcal{R}(h_S, f) < \epsilon \right) \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) < \epsilon \right) \tag{74}$$

$$\leq 4\epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) + \epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) < \epsilon \right) \tag{75}$$

$$= \epsilon + 3\epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right). \tag{76}$$

Hence,

$$\mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \geq \frac{1}{3\epsilon} \left(2\epsilon - \epsilon \right) = \frac{1}{3} \tag{77}$$

Finally,

$$\mathbb{P}_{S'} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}^s(S'))) \geq \epsilon \right) \geq \mathbb{P}_{S'} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \geq \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \mathbb{P}_{S'} \left(S' \in \mathcal{S} \right) > \frac{1}{3} \frac{3}{20} = \frac{1}{20}. \tag{78}$$

□

C. Proof of Theorem 3

Theorem 3. *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis space, let m and N be any integers and let G be a fixed subset of $[N]$ of size $k \in \{1, \ldots, N-1\}$. Let $S' \in (\mathcal{X} \times \mathcal{Y})^{N \times m}$ be drawn i.i.d. from $\mathcal{D}$. Then the following statements hold for any multi-source learner $\mathcal{L}$:*

(a) Suppose that $\mathcal{H}$ is non-trivial. Then there exists a distribution $\mathcal{D}$ on $\mathcal{X}$ with $\min_{h \in \mathcal{H}} \mathcal{R}(h) = 0$, and a fixed-set adversary $\mathcal{A}$ with index set G, such that:

$$\mathbb{P}_{S'}\Big(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S'))\big) > \frac{\alpha}{8m}\Big) > \frac{1}{20}. \tag{79}$$

(b) Suppose that $\mathcal{H}$ has VC dimension $d \geq 2$. Then there exists a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ and a fixed-set adversary $\mathcal{A}$ with index set G, such that:

$$\mathbb{P}_{S'}\left(\mathcal{R}\big(\mathcal{L}(\mathcal{A}(S'))\big) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \sqrt{\frac{d}{1280Nm}} + \frac{\alpha}{16m}\right) > \frac{1}{64}. \tag{80}$$

In both cases, $\alpha = \frac{N-k}{N}$ is the power of the adversary.

To prove part (a), we use a similar technique as in the lower bound results in (Bshouty et al., 2002) and in the classic sample complexity lower bound for binary classification, e.g. Theorem 3.20 in (Mohri et al., 2018). An overview is as follows. Consider a distribution on $\mathcal{X}$ that has support only at two points - the common point x_1 and the rare point x_2. Take $\mathbb{P}(x_2) = \mathcal{O}(\frac{\alpha}{m})$. Then one can show that with constant probability the number of datasets that contain x_2 is at most αN. We show that in this case there exists an algorithm for the strong adversary that will return the same unordered collection of datasets, regardless of the true label of x_2. Thus no learner can guess with probability greater than $1/2$ what the true label of x_2 was.

Part (b) follows from part (a) and the standard lower bound for agnostic binary classification.

Proof. a) As in Theorem 2, we prove that there exists a distribution $\mathcal{D}$ on $\mathcal{X}$ and a labeling function $f \in \mathcal{H}$, such that the resulting joint distribution on $\mathcal{X} \times \mathcal{Y}$, defined by $x \sim \mathcal{D}$ and $y = f(x)$, satisfies the desired property.

Without loss of generality, let $G = [1, 2, \ldots, k]$. Since $\mathcal{H}$ is non-trivial ($d \geq 2$), there exist $h_1, h_2 \in \mathcal{H}$ and $x_1, x_2 \in \mathcal{X}$, such that $h_1(x_1) = h_2(x_1)$, while $h_1(x_2) = 1$, but $h_2(x_2) = -1$. Consider the following distribution on $\mathcal{X}$:

$$\mathbb{P}_{\mathcal{D}}(x_1) = 1 - 4\epsilon \quad \text{and} \quad \mathbb{P}_{\mathcal{D}}(x_2) = 4\epsilon, \tag{81}$$

where $\epsilon = \frac{\alpha}{8m}$. Assume that the points are labelled by a function $f \in \mathcal{H}$ (to be chosen later as either h_1 or h_2). Denote the initial uncorrupted collection of datasets by $S' = (S'_1, \ldots, S'_N)$, with $S'_i = \{(x'_{i,1}, f(x'_{i,1})), \ldots, (x'_{i,m}, f(x'_{i,m}))\}$ and $x'_{i,j}$ being i.i.d. samples from $\mathcal{D}$.

First we show that with constant probability the point x_2 is contained in no more than αN sources. Indeed, let C_b be the number of sources that contain x_2 and let C_p be the number of points (out of the Nm in total) that are equal to x_2. Clearly $C_b \leq C_p$. Note that C_p is a binomial random variable with probability of success 4ϵ and number of trials Nm. Therefore, by the Chernoff bound:

$$\mathbb{P}_{S'}(C_p \geq \alpha N) = \mathbb{P}_{S'}(C_p \geq (1+1)4\epsilon Nm) \leq e^{-\frac{\alpha N}{6}} \leq e^{-1/6} < \frac{17}{20} \tag{82}$$

and so:

$$\mathbb{P}_{S'}(C_b \leq \alpha N) \geq \mathbb{P}_{S'}(C_p \leq \alpha N) > \frac{3}{20}. \tag{83}$$

Now consider the following policy for the adversary $\mathcal{A}^s : S' \to S$. Whenever $C_b \leq \alpha N$, let $M \subset G$ be the list of indexes $i \in G$, such that S'_i contains x_2. Let $l = |M|$ and note that $l \leq C_b \leq \alpha N$. For any index $i \in [N]$ the adversary replaces $S'_i = \{x'_{i,1}, f(x'_{i,1}), \ldots, (x'_{i,m}, f(x'_{i,m}))\}$ with a dataset $S_i = \{(x_{i,1}, y_{i,1}) \ldots, (x_{i,m}, y_{i,m})\}$, such that:

$$(x_{i,j}, y_{i,j}) = \begin{cases} (x'_{i,j}, f(x'_{i,j})), & \text{if } i \in G = [1, 2, \ldots, k] \\ (x_1, f(x_1)), & \text{if } i \in [k+1, \ldots, k+l] \text{ and } x'_{M[i-k],j} = x_1 \\ (x_2, -f(x_2)), & \text{if } i \in [k+1, \ldots, k+l] \text{ and } x'_{M[i-k],j} = x_2 \\ (x_1, f(x_1)), & \text{if } i \in [k+l+1, \ldots, N] \end{cases} \tag{84}$$

Then the adversary returns $S = (S_1, \ldots, S_N)$. That is, the adversary keeps the datasets in G untouched, copies all of the datasets in M into its own data, flipping the labels of the x_2's, and, in case there are additional sources at its disposal, it fills them with (correctly labelled) x_1's only.

Crucially, the resulting (unordered) collection is the same no matter if the original labelling function was h_1 or h_2. In particular, $\mathcal{L}(S)$ will be the same in both cases.

In the case when $C_b > \alpha N$, the adversary leaves the data unchanged, i.e. $S = S'$.

Finally, we argue that under the event $C_b \leq \alpha N$ and the chosen adversary, the learner would incur high loss and show that this implies the result in (20). Let $\mathcal{S}$ be the set of all datasets in $(\mathcal{X} \times \mathcal{Y})^{N \times m}$, such that $C_b \leq \alpha N$ holds. We just showed that $\mathbb{P}_{S'}(S' \in \mathcal{S}) > \frac{3}{20}$ and that whenever $S' \in \mathcal{S}$, $\mathcal{L}(\mathcal{A}^s(S'))$ is independent of whether the original labelling function was h_1 or h_2.

Now the proof proceeds just as in Theorem 2. Consider a fixed set $S' \in \mathcal{S}$ and let $S = \mathcal{A}^s(S')$ and $h_S = \mathcal{L}(S)$. Denote by $\mathcal{R}(h_S, f) = \mathbb{P}_{\mathcal{D}}(h_S(x) \neq f(x) \cap x \neq x_1)$ and note that $\mathcal{R}(h_S, f) \leq \mathbb{P}_{\mathcal{D}}(h_S(x) \neq f(x)) = \mathcal{R}(\mathcal{L}(\mathcal{A}^s(S')))$. Notice that:

$$\begin{aligned} \mathcal{R}(h_S, h_1) + \mathcal{R}(h_S, h_2) &= \sum_{i=1,2} \mathbb{1}_{h_S(x_i) \neq h_1(x_i)} \mathbb{1}_{x_i \neq x_1} \mathbb{P}(x_i) + \sum_{i=1,2} \mathbb{1}_{h_S(x_i) \neq h_2(x_i)} \mathbb{1}_{x_i \neq x_1} \mathbb{P}(x_i) && (85) \\ &= \mathbb{1}_{h_S(x_2) \neq h_1(x_2)} 4\epsilon + \mathbb{1}_{h_S(x_2) \neq h_2(x_2)} 4\epsilon && (86) \\ &= 4\epsilon, && (87) \end{aligned}$$

where we used that $h_1(x_2) = 1 = -h_2(x_2)$ and that h_S is independent of the underlying labelling function.

Since the above holds for any $S' \in \mathcal{S}$, it also holds in expectation, conditioned on $S' \in \mathcal{S}$:

$$\mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, h_1) + \mathcal{R}(h_S, h_2) \right) \geq 4\epsilon. \tag{88}$$

Therefore, $\mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, h_i) \right) \geq 2\epsilon$ for at least one of $i = 1, 2$. Take f to be h_1, if h_1 satisfies the inequality, and h_2 otherwise. Conditioning on $\{\mathcal{R}(h_S, f) \geq \epsilon\}$ and using $\mathcal{R}(h_S, f) \leq \mathbb{P}_{\mathcal{D}}(x \neq x_1) = 4\epsilon$:

$$\begin{aligned} 2\epsilon \leq \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \right) &= \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) | \mathcal{R}(h_S, f) \geq \epsilon \right) \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) && (89) \\ &\quad + \mathbb{E}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) | \mathcal{R}(h_S, f) < \epsilon \right) \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) < \epsilon \right) && (90) \\ &\leq 4\epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) + \epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) < \epsilon \right) && (91) \\ &= \epsilon + 3\epsilon \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right). && (92) \end{aligned}$$

Hence,

$$\mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \geq \frac{1}{3\epsilon} \left(2\epsilon - \epsilon \right) = \frac{1}{3} \tag{93}$$

Finally,

$$\begin{aligned} \mathbb{P}_{S'} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}^s(S'))) \geq \epsilon \right) &\geq \mathbb{P}_{S'} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) && (94) \\ &\geq \mathbb{P}_{S' \in \mathcal{S}} \left(\mathcal{R}(h_S, f) \geq \epsilon \right) \mathbb{P}_{S'} \left(S' \in \mathcal{S} \right) && (95) \\ &> \frac{1}{3} \frac{3}{20} = \frac{1}{20}. && (96) \end{aligned}$$

b) First we argue that there exists a distribution $\mathcal{D}_1$ on $\mathcal{X} \times \mathcal{Y}$ and a fixed-set adversary $\mathcal{A}_1^s$, such that:

$$\mathbb{P}_{S' \sim \mathcal{D}_1} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}_1^s(S'))) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \sqrt{\frac{d}{320Nm}} \right) > \frac{1}{64}. \tag{97}$$

This follows directly from the classic lower bound for binary classification in the unrealizable case. Indeed, applying Theorem 3.23 in (Mohri et al., 2018) and setting the adversary to be the identity mapping gives the result.

Now, since any hypothesis space with VC dimension $d \geq 2$ is non-trivial, we also know from a) that there exists an adversary $\mathcal{A}_2^s$ and a distribution $\mathcal{D}_2$ on $\mathcal{X} \times \mathcal{Y}$, such that:

$$\mathbb{P}_{S' \sim \mathcal{D}_2} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}_2^s(S'))) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \frac{\alpha}{8m} \right) > \frac{1}{20}. \tag{98}$$

Fix any set of values for N, m, d, k. Then at least one of the pairs $(\mathcal{A}_1^s, \mathcal{D}_1)$ and $(\mathcal{A}_2^s, \mathcal{D}_2)$ satisfies:

$$\mathbb{P}_{S'} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}^s(S'))) - \min_{h \in \mathcal{H}} \mathcal{R}(h) > \sqrt{\frac{d}{1280Nm}} + \frac{\alpha}{16m} \right) \geq \mathbb{P}_{S'} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}^s(S'))) > 2 \max\{ \sqrt{\frac{d}{1280Nm}}, \frac{\alpha}{16m} \} \right) \tag{99}$$

$$= \mathbb{P}_{S'} \left(\mathcal{R}(\mathcal{L}(\mathcal{A}^s(S'))) > \max\{ \sqrt{\frac{d}{320Nm}}, \frac{\alpha}{8m} \} \right) \tag{100}$$

$$> \frac{1}{64}. \tag{101}$$

□