Robustly-reliable learners under poisoning attacks

Maria-Florina Balcan   NINAMF@CS.CMU.EDU
Carnegie Mellon University

Avrim Blum   AVRIM@TTIC.EDU
Toyota Technological Institute at Chicago

Steve Hanneke   STEVE.HANNEKE@GMAIL.COM
Purdue University

Dravyansh Sharma   DRAVYANS@CS.CMU.EDU
Carnegie Mellon University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Data poisoning attacks, in which an adversary corrupts a training set with the goal of inducing specific desired mistakes, have raised substantial concern: even just the possibility of such an attack can make a user no longer trust the results of a learning system. In this work, we analyze when strong robustness guarantees can be achieved even in the face of such attacks.

We define and show how to provide *robustly-reliable predictions*, in which the predicted label is guaranteed to be correct so long as the adversary has not exceeded a given corruption budget, even in the presence of *instance targeted attacks*, where the adversary aims to cause a failure on specific test examples. Our guarantees are substantially stronger than those in prior approaches, which were only able to provide certificates that the prediction of the learning algorithm does not change, as opposed to certifying that the prediction is correct, as we do here. Remarkably, we provide a complete characterization of learnability in this setting, in particular, nearly-tight matching upper and lower bounds on the region that can be certified, as well as efficient algorithms for computing this region given an ERM oracle. Moreover, for the case of linear separators over logconcave distributions, we provide efficient truly polynomial time algorithms (i.e., non-oracle algorithms) for such robustly-reliable predictions.

We also extend these results to the active setting where the algorithm adaptively asks for labels of specific informative examples, and the difficulty is that the adversary might even be adaptive to this interaction, as well as to the agnostic learning setting where there is no perfect classifier even over the uncorrupted data.

Keywords: Data poisoning, adversarial robustness, instance-targeted attacks, reliable learning

1. Introduction

Overview: There has been significant interest in machine learning in recent years in building robust learning systems that are resilient to adversarial attacks, either test time attacks (Goodfellow et al., 2015; Carlini and Wagner, 2017; Madry et al., 2018; Zhang et al., 2019) or training time attacks (Steinhardt et al., 2017; Shafahi et al., 2018) also known as *data poisoning*. A lot of the effort on provable guarantees for such systems has focused on test-time attacks (e.g., Yin et al., 2019; Attias et al., 2019; Montasser et al., 2019, 2020, 2021; Goldwasser et al., 2020), and the question of understanding what is fundamentally possible under training time attacks or data poisoning has been fairly wide open (Barreno et al., 2006; Levine and Feizi, 2021).

In data poisoning attacks, an adversary corrupts a training set used to train a learning algorithm in order to induce some desired behavior. Even just the possibility of such an attack can make a user no longer trust the results of a learning system. Particularly challenging is the prospect of providing formal guarantees for *instance targeted attacks*, where the adversary has a goal of inducing a mistake on *specific instances*; the difficulty is that the learner does not know which instance the adversary is targeting, so it must try to guard against attacks for essentially every possible instance. In this work, we devolop a general understanding of what robustness guarantees are possible in such cases while simultaneouly analyzing multiple important axes:

- *Instance targeted attacks*: The adversary can know our test example in advance, applying its full corruption budget with the goal of making the learner fail on this example, and even potentially change its corruptions from one test example to another.

- *Robustly-reliable predictions*: When our algorithm outputs a prediction y with robustness level η, this is a guarantee that y is correct so long as the target function belongs to the given class and the adversary has corrupted at most an η fraction of the training data. For any value η, we analyze for which points it is possible to provide a prediction with such a strong robustness level; we provide both sample and distribution dependent nearly matching upper and lower bounds on the size of this set, as well as efficient algorithms for constructing it given access to an ERM oracle. We note that prior work (e.g., Gao et al. (2021)) only provided certificates of stability, meaning that the prediction of the algorithm does not change, rather than correctness. Robustly-reliable predictions, which certify correctness, imply that our predictions can truly be trusted.

- *Active learning against adaptive adversaries:* We also address the challenging active learning setting where labeled data is expensive to obtain, but the learning algorithm has the power to adaptively ask for labels of select examples from a large pool of unlabeled data in order to learn accurate classifiers with fewer labeled examples. While this interaction could save labels, it also raises the concern that the adversary could generate harm by adaptively deciding which points to corrupt based on choices of the algorithm. We provide algorithms that not only learn with fewer labeled examples, but are able to operate against such adaptive adversaries that can make corruptions on the fly.

- *Agnostic learning:* Providing a reasonable extension to the agnostic case where there is no perfect classifier even over the uncorrupted data.

While prior work has considered some of these aspects separately, we provide the first results that bring these together, in particular combining reliable predictions with instance-targeting poisoning attacks. We also provide nearly-tight upper and lower bounds on guarantees achievable, as well as efficient algorithms.

Our Results: Instance targeted data poisoning attacks have been of growing concern as learned classifiers are increasingly used in society. Classic work on data poisoning attacks (e.g., Valiant (1985); Kearns and Li (1993); Bshouty et al. (2002); Awasthi et al. (2017)) only considers *non*-instance-targeted attacks, where the goal of the adversary is to increase the overall error rate, rather than to cause errors on specific test points it wishes to target. Instance targeted poisoning attacks, on the other hand, are more challenging but they are particularly relevant for modern applications like recommendation engines, fake review detectors and spam filters where training data is likely to include user-generated data, and the results of the learning algorithm could have financial consequences. For example, a company depending on a recommendation engine for advertisement of its

product has an incentive to induce a positive classification on its own product or to harm a specific competitor. To defend against such adversaries, we would like to provide predictions with instance-specific correctness guarantees, even when the adversary can use its entire corruption budget to target those instances. This has been of significant interest in recent years in machine learning.

In this work we consider algorithms that provide *robustly-reliable* predictions, which are guaranteed to be correct under well-specified assumptions even in the face of targeted attacks. In particular, given a test input x, a robustly-reliable classifier outputs both a prediction y and a robustness level η, with a guarantee that y is correct unless one of two bad events has occurred: (a) the true target function does not belong to its given hypothesis class $\mathcal{H}$ or (b) an adversary has corrupted more than an η fraction of the training data. Such a guarantee intrinsically has in it a notion of targeting, because the prediction is guaranteed to be correct even if the adversarial corruptions in (b) were designed specifically to target x. Note that it is possible to produce a trivial (and useless) robustly-reliable classifier that always outputs $\eta < 0$ (call this "abstaining" or an "unconfident prediction"). We will want to produce classifiers that, as much as possible, instead output *confident* predictions, that is, predictions with large values of η. This leads to several kinds of guarantees one might hope for, because while the adversary cannot cause the algorithm to be incorrect with a high confidence, it could potentially cause the classifier to produce low-confidence predictions in a targeted way.

In this work, we demonstrate an optimal robustly-reliable learner $\mathcal{L}$, and precisely identify the guarantees that can be acheived in this setting, with nearly matching upper and lower bounds. Specifically,

- We prove guarantees on the set of test points for which the learner will provide confident predictions, for any given adversarial corruption of the training data (Theorem 5), or even more strongly, for all bounded instance-targeted corruptions of the training set (Theorem 7). Intuitively speaking, we show that the set of points on which our learner should be confident are those that belong to the region of agreement of low error hypotheses. Our learner shows more confidence for points in the agreement regions of larger radii around the target hypothesis. We also show how $\mathcal{L}$ may be implemented efficiently given an ERM oracle (Theorem 6) by running the oracle on datasets where multiple copies of the test point are added with different possible labels. We further provide empirical estimates on the set of such test points (Theorem 9), which could help determine if the learner is ready to be fielded in a highly adversarial environment.

- We provide fundamental lower bounds on which test points a robustly-reliable learner could possibly be confident on (Theorems 10 and 11). We do this through characterizing the set of points an adversary could successfully attack for any learner, which roughly speaking is the region of disagreement of low error classifiers. In essence, ambiguity about the original dataset prior to corruption leads to ambiguity about which low error classifier is the target concept, so we cannot predict confidently on any point in the region of disagreement among these low error concepts. Our upper bound in Theorem 7 and lower bound in Theorem 10 exactly match, and imply that our learner $\mathcal{L}$ is optimal in its confident set.

- For learning linear separators under logconcave marginal distributions, we show a polynomial time robustly-reliable learning algorithm under instance-targeted attacks produced by a malicious adversary (Valiant, 1985) (see Theorem 14). The key idea is to first use the learner from the seminal work of Awasthi et al. (2017) (that was designed for non-targeted attacks) to first find a low error classifier and then to use that and the geoemetry of the underlying data distribution in order to efficiently find a good approximation of the region of agreement of the low error

3

classifiers. We also show that the robust-reliability region for this algorithm is near-optimal for this problem, even ignoring computational efficiency issues (Theorem 28).

- We also study *active learning*, extending the classic disagreement-based active learning techniques, to achieve comparable robust reliability guarantees as above, but with reduced number of labeled examples (Theorems 30, 32), even when an adversary can choose which points to corrupt in an adaptive manner. This is particularly challenging because the adversary can use its entire corruption budget on only those points whose labels are specifically requested by the learner.

- Finally, we generalize our results to the agnostic case where there is no perfect classifier even over the uncorrupted data (Theorems 37, 38, 39, and 41). In this case, the adversary could cause a robustly-reliable learner to make a mistake, but only if *every* low-error hypothesis in the class would have also made a mistake on that point as well.

1.1. Context and Related Work

Non-instance targeted poisoning attacks. The classic malicious noise model introduced in (Valiant, 1985) and subsequently analyzed in (Kearns and Li, 1993; Bshouty et al., 2002; Klivans et al., 2009; Awasthi et al., 2014, 2017) provides one approach to modeling poisoning attacks. However the malicious noise model only captures untargeted attacks — formally the adversary's goal is to maximize the learner's overall error rate.

Instance targeted poisoning attacks. Instance-targeted poisoning attacks were first considered by Barreno et al. (2006). Shafahi et al. (2018) and Suciu et al. (2018) showed empirically that such targeted attacks can be powerful even if the adversary adds only *correctly-labeled data* to the training set (called "clean-label attacks"). Targeted poisoning attacks have generated significant interest in recent years due to the damage they can cause to the trustworthiness of a learning system (Mozaffari-Kermani et al., 2014; Chen et al., 2017; Geiping et al., 2020).

Past theoretical work on defenses against instance-targeted poisoning attacks has generally focused on producing certificates of stability, indicating when an adversary with a limited budget could not have changed the prediction that was made. For example, Levine and Feizi (2021) propose partitioning training data into k portions, training separate classifiers on each portion, and then using the strength of the majority-vote over those classifiers as such a certificate (since any given poisoned point can corrupt at most one portion). Gao et al. (2021) formalize a wide variety of different kinds of adversarial poisoning attacks, and analyze the problem of providing certificates of stability against them in both distribution-independent and distribution-specific settings. In contrast to those results that certify when a budget-limited adversary could not *change* the learner's prediction, our focus is on certifying that the prediction made is *correct*. For example, a learner that always outputs the "all-negative" classifier regardless of the training data would be a certifying learner in the sense of Gao et al. (2021) (for an arbitrarily large attack budget) and its correctness region would be the probability mass of true negative examples. In contrast, in our model, outputting a prediction (y, η) means that y is guaranteed to be a correct prediction so long as the adversary corrupted at most an η fraction of the training data and the target belongs to the given class; so, a learner that always outputs (y, η) for $y = -1$ and $\eta \geq 0$ would not be robustly-reliable in our model (unless the given class only had the all-negative function). We are the first to consider such strong correctness guarantees in the presence of adversarial data poisoning.

Interestingly, for learning linear separators, our results improve over those of Gao et al. (2021) even for producing certificates of stability, in that our algorithms in run polynomial time and apply

to a much broader class of data distributions (any isotropic log-concave distribution and not only uniform over the unit ball).

Blum et al. (2021) provide a theoretical analysis for the special case of clean-label poisoning attacks (Shafahi et al., 2018; Suciu et al., 2018). They analyze the probability mass of attackable instances for various algorithms and hypothesis classes. However, they do not consider any form of certification or reliability guarantees.

Reliable useful learners. Our model can be viewed as a broad generalization of the reliable-useful learning model of Rivest and Sloan (1988), and the perfect selective classification model of El-Yaniv and Wiener (2010, 2012), which only consider learning from noiseless data, to the setting of noisy data and adversarial poisoning attacks. El-Yaniv and Wiener (2010) in particular give tight positive and negative results for the noise-free case, and exploration of the "risk-coverage" tradeoff in which the algorithm is allowed to have a nonzero error-rate in the region of confidence. Because we are concerned with instance-wise guarantees against targeted attacks, in our setting we require zero error when the algorithm claims to be confident: a prediction (y, η) must be correct whenever an adversary has poisoned at most an η fraction of the training points.

2. Formal Setup

Setup. Let $\mathcal{D}$ denote a data distribution over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y} = \{0, 1\}$ is the label space. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ be the concept space. We will primarily assume the realizable case, that is, for some $h^* \in \mathcal{H}$, for any (x, y) in the support $\mathrm{supp}(\mathcal{D})$ of the data distribution $\mathcal{D}$, we have $y = h^*(x)$ (we extend to the non-realizable case in Section 6). We will use $\mathcal{D}_{\mathcal{X}}$ to denote the marginal distribution over $\mathcal{X}$ of unlabeled examples. The learner $\mathcal{L}$ has access to a corruption S' of a sample $S \sim \mathcal{D}^m$, and is expected to output a hypothesis $h_{\mathcal{L}(S')} \in \mathcal{Y}^{\mathcal{X}}$ (*proper* with respect to $\mathcal{H}$ if $h_{\mathcal{L}(S')} \in \mathcal{H}$). We will use the 0-1 loss, i.e. $\ell(h, (x, y)) = \mathbb{1}[h(x) \neq y]$. For a fixed (possibly corrupted) sample S', let $\mathrm{err}_{S'}(h)$ denote the average empirical loss for hypothesis h, i.e. $\mathrm{err}_{S'}(h) = \frac{1}{|S'|} \sum_{(x,y) \in S'} \ell(h, (x, y))$. Similarly define $\mathrm{err}_{\mathcal{D}}(h) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(h, (x, h^*(x)))]$. For a sample S, let $\mathcal{H}_{\eta}(S) = \{h \in \mathcal{H} \mid \mathrm{err}_S(h) \leq \eta\}$ be the set of hypotheses in $\mathcal{H}$ with at most η error on S. Similarly for any distribution $\mathcal{D}$, let $\mathcal{H}_{\eta}(\mathcal{D}) = \{h \in \mathcal{H} \mid \mathrm{err}_{\mathcal{D}}(h) \leq \eta\}$. We will consider a class of attacks where the adversary can make arbitrary corruptions to up to an η fraction of the training sample S. If the adversary can also choose which examples to attack, it corresponds to the *nasty* attack model of Bshouty et al. (2002). We formalize the adversary below.

Adversary. Let $d(S, S') = 1 - \frac{|S \cap S'|}{m} \in [0, 1]$ denote the normalized Hamming distance between two samples S, S' with $m = |S| = |S'|$. Let $A(S)$ denote the sample corrupted by adversary A. For $\eta \in [0, 1]$, let $\mathcal{A}_{\eta}$ be the set of adversaries with corruption budget η and $\mathcal{A}_{\eta}(S) = \{S' \mid d(S, S') \leq \eta\}$ denotes the possible corrupted training samples under an attack from an adversary in $\mathcal{A}_{\eta}$. Intuitively, if the given sample is S', we would like to give guarantees for learning when $S' \in \mathcal{A}_{\eta}(S)$ for some (realizable) uncorrupted sample S. Also we will use the convention that $\mathcal{A}_{\eta}(S) = \{\}$ for $\eta < 0$ to allow the learner to sometimes predict without any guarantees (cf. Definition 1). Note that the adversary can change both x and y in an example (x, y) it chooses to corrupt, and can arbitrarily select which η fraction to corrupt as in Bshouty et al. (2002).

We now define the notion of a *robustly-reliable* learner in the face of instance-targeted attacks. This learner, for any given test example x, outputs both a prediction y and a robust reliability level η_x, such that y is guaranteed to be correct so long as $h^* \in \mathcal{H}$ and the adversary's corruption budget is $\leq \eta_x$. This learner then only gets credit for predictions guaranteed to at least a desired value η.

Definition 1 *A learner $\mathcal{L}$ is* **robustly-reliable** *for sample S' w.r.t. concept space $\mathcal{H}$ if, given S', the learner outputs a function $\mathcal{L}_{S'} : \mathcal{X} \to \mathcal{Y} \times \mathbb{R}$ such that for all $x \in \mathcal{X}$ if $\mathcal{L}_{S'}(x) = (y, \eta)$ and if $S' \in \mathcal{A}_\eta(S)$ for some sample S labeled by concept $h^* \in \mathcal{H}$, then $y = h^*(x)$. Note that if $\eta < 0$, then $\mathcal{A}_\eta(S) = \{\}$ and the above condition imposes no requirement on the learner's prediction. If $\mathcal{L}_{S'}(x) = (y, \eta)$ then let $h_{\mathcal{L}(S')}(x) = y$.*

*Given sample S labeled by h^*, the η-**robustly-reliable region** $RR^{\mathcal{L}}(S, h^*, \eta)$ for learner $\mathcal{L}$ is the set of points $x \in \mathcal{X}$ for which given any $S' \in \mathcal{A}_\eta(S)$ we have that $\mathcal{L}_{S'}(x) = (y, \eta')$ with $\eta' \geq \eta$. More generally, for a class of adversaries $\mathcal{A}$ with budget η, $RR^{\mathcal{L}}_{\mathcal{A}}(S, h^*, \eta)$ is the set of points $x \in \mathcal{X}$ for which given any $S' \in \mathcal{A}(S)$ we have that $\mathcal{L}_{S'}(x) = (y, \eta')$ with $\eta' \geq \eta$. We also define the* **empirical η-robustly-reliable region** $\widehat{RR}^{\mathcal{L}}(S', \eta) = \{x \in \mathcal{X} : \mathcal{L}_{S'}(x) = (y, \eta') \text{ for some } \eta' \geq \eta\}$. *So, $RR^{\mathcal{L}}_{\mathcal{A}}(S, h^*, \eta) = \cap_{S' \in \mathcal{A}(S)} \widehat{RR}^{\mathcal{L}}(S', \eta)$.*

Remark 2 *The requirement of security even to targeted attacks appears in two places in Definition 1. First, if a robustly-reliable learner outputs (y, η) on input x, then y must be correct even if an η fraction of the training data had been corrupted specifically to target x. Second, for a point x to be in the η-robustly-reliable region, it must be the case that for any $S' \in \mathcal{A}_\eta(S)$ (even if S' is a targeted attack on x) we have $\mathcal{L}_{S'}(x) = (y, \eta')$ for some $\eta' \geq \eta$. So, points in the η-robustly-reliable region are points an adversary cannot successfully target with a budget of η or less.*

Definition 1 describes a robustly-reliable learner for a particular (corrupted) sample. We now extend this to the notion of a learner being robustly reliable with high probability for an adversarially-corrupted sample drawn from a given distribution.

Definition 3 *A learner $\mathcal{L}$ is a $(1 - \gamma)$-**probably robustly-reliable learner** for concept space $\mathcal{H}$ under marginal $\mathcal{D}_X$ at sample size m if for any target function $h^* \in \mathcal{H}$, with probability at least $1 - \gamma$ over the draw of $S \sim \mathcal{D}^m$ ($\mathcal{D}$ is the distribution over examples labeled by h^* with marginal $\mathcal{D}_X$), for all $S' \in \mathcal{A}_\eta(S)$, and for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \geq \eta$ then $y = h^*(x)$. If $\mathcal{L}$ is a $(1 - \gamma)$-probably robustly-reliable learner with $\gamma = 0$ for all marginal distributions $\mathcal{D}_X$, then we say $\mathcal{L}$ is* **strongly robustly-reliable** *for $\mathcal{H}$. Note that a strongly robustly-reliable learner is a robustly-reliable learner in the sense of Definition 1 for every sample S' of size m.*

The η-robustly-reliable correctness for learner $\mathcal{L}$ for sample S labeled by h^* is given by the probability mass of the robustly-reliable region, $RobC^{\mathcal{L}}(\mathcal{D}, \eta, S) = \Pr_{x \sim \mathcal{D}_X}[x \in RR^{\mathcal{L}}(S, h^*, \eta)]$.

In our analysis we will also need several key quantities – agreement/disagreement regions and the disagreement coefficient — that have been previously used in the disagreement-based active learning literature (Balcan et al., 2006; Hanneke, 2007; Balcan et al., 2009).

Definition 4 *For a set $H \subseteq \mathcal{H}$ of hypotheses, define the* **disagreement region** $DIS(H) = \{x \in \mathcal{X} : \exists h_1, h_2 \in H \text{ s.t. } h_1(x) \neq h_2(x)\}$, *and the* **agreement region** $\text{Agree}(H) = \mathcal{X} \setminus DIS(H)$. *For hypotheses h, h' we denote their disagreement over distribution $\mathcal{D}$ as $d_{\mathcal{D}}(h, h') = \Pr_{x \sim \mathcal{D}_X}[h(x) \neq h'(x)]$ and similarly their disagreement over sample S as $d_S(h, h') = \frac{1}{|S|} \sum_{(x,y) \in S} \mathbb{1}[h(x) \neq h'(x)]$. For a hypothesis h and a value $\epsilon \geq 0$, the ball $\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h, \epsilon)$ around h of radius ϵ is defined as the set $\{h' \in \mathcal{H} : d_{\mathcal{D}}(h, h') \leq \epsilon\}$, and similarly for sample S, $\mathcal{B}^{\mathcal{H}}_S(h, \epsilon) = \{h' \in \mathcal{H} : d_S(h, h') \leq \epsilon\}$. Lastly, the* **disagreement coefficient**, θ_ϵ, *of $\mathcal{H}$ with respect to h^* over $\mathcal{D}$ is given by*

$$\theta_\epsilon = \sup_{r > \epsilon} \frac{\Pr_{\mathcal{D}_X}[DIS(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r))]}{r}.$$

3. Robustly Reliable Learners for Instance-Targeted Adversaries

We now provide a general strongly robustly-reliable learner using the notion of agreement regions (Theorem 5) and show how it can be implemented efficiently given access to an ERM oracle for $\mathcal{H}$ (Theorem 6). We then prove that its robustly reliable region is pointwise optimal (contains the robustly reliable regions of all other such learners), for all values of the adversarial budget η (Theorems 7 and 10). Recall that a strongly robustly-reliable learner $\mathcal{L}$ is given a possibly-corrupted sample S' and outputs a function $\mathcal{L}_{S'}$ such that if $\mathcal{L}_{S'} = (y, \eta)$ and $S' = \mathcal{A}_\eta(S)$ for some (unknown) uncorrupted sample S labeled by some (unknown) target concept $h^* \in \mathcal{H}$, then $y = h^*(x)$.

Theorem 5 *Let $\mathcal{H}_\eta(S') = \{h \in \mathcal{H} \mid \mathrm{err}_{S'}(h) \leq \eta\}$. For any hypothesis class $\mathcal{H}$, there exists a strongly robustly-reliable learner $\mathcal{L}$ that given S' outputs a function $\mathcal{L}_{S'}$ such that*

$$\widehat{RR}^{\mathcal{L}}(S', \eta) \supseteq \mathsf{Agree}(\mathcal{H}_\eta(S')).$$

Proof Given sample S', the learner $\mathcal{L}$ outputs the function $\mathcal{L}_{S'}(x) = (y, \eta)$ where η is the largest value such that $x \in \mathsf{Agree}(\mathcal{H}_\eta(S'))$, and y is the common prediction in that agreement region; if $x \notin \mathsf{Agree}(\mathcal{H}_\eta(S'))$ for all $\eta \geq 0$, then $\mathcal{L}_{S'}(x) = (\bot, -1)$. This is a strongly robustly-reliable learner because if $\mathcal{L}_{S'}(x) = (y, \eta)$ and $S' \in \mathcal{A}_\eta(S)$, then $h^* \in \mathcal{H}_\eta(S')$, so $y = h^*(x)$. Also, by design, all points in $\mathsf{Agree}(\mathcal{H}_\eta(S'))$ are given a robust reliability level at least η. ∎

We now show how the strongly robustly-reliable learner $\mathcal{L}$ from Theorem 5 can be implemented efficiently given access to an ERM oracle for class $\mathcal{H}$.

Theorem 6 *Learner $\mathcal{L}$ from Theorem 5 can be implemented efficiently given an ERM oracle for $\mathcal{H}$.*

Proof Given sample S' and test point x, for each possible label $y \in \mathcal{Y}$ the learner first computes the minimum value ϵ_y of the empirical error on S' achievable using $h \in \mathcal{H}$ subject to $h(x) = y$; that is, $\epsilon_y = \min\{\mathrm{err}_{S'}(h) : h \in \mathcal{H}, h(x) = y\}$. This can be computed efficiently using an ERM oracle by simply running the oracle on a training set consisting of S' and $m + 1$ copies of the labeled example (x, y) where $m = |S'|$; this will force ERM to produce a hypothesis h such that $h(x) = y$. Now if $\epsilon_0 = \epsilon_1$, then this means that $x \notin \mathsf{Agree}(\mathcal{H}_\eta(S'))$ for any value η for which $\mathcal{H}_\eta(S')$ is nonempty, so the algorithm can output $(\bot, -1)$. Otherwise, the algorithm can output the label $y = \mathrm{argmin}_{y'}\{\epsilon_{y'}\}$ and $\eta = \max\{\epsilon_0, \epsilon_1\} - 1/m$. Specifically, by definition of ϵ_y, this is the largest value of η such that $x \in \mathsf{Agree}(\mathcal{H}_\eta(S'))$. ∎

We now analyze the η-robustly-reliable region for the algorithm above (Theorems 7 and 9) and prove that it is optimal over all strongly robustly-reliable learners (Theorem 10). Specifically, Theorem 7 provides a guarantee on the size of the η-robustly-reliable region in terms of properties of h^* and S, and then Theorem 9 gives an empirically-computable bound. Thus, these indicate when such an algorithm can be fielded with confidence even in the presence of adversaries that control an η fraction of the training data and can modify them at will.

Theorem 7 *For any hypothesis class $\mathcal{H}$, the strongly robustly-reliable learner $\mathcal{L}$ from Theorem 5 satisfies the property that for all S and for all $\eta \geq 0$,*

$$RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)\right).$$

Moreover, if $S \sim \mathcal{D}^m$ then with probability $1 - \delta$, $\mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))$ for $m = O(\frac{1}{\epsilon^2}(d + \ln\frac{1}{\delta}))$. So, whp, $RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))$. Here $d = \mathrm{VCdim}(\mathcal{H})$.

Proof By Theorem 5, the empirical η-robustly-reliable region satisfies $\widehat{\mathsf{RR}}^{\mathcal{L}}(S', \eta) \supseteq \mathsf{Agree}(\mathcal{H}_\eta(S'))$. The set $\mathsf{RR}^{\mathcal{L}}(S, h^*, \eta) = \cap_{S' \in \mathcal{A}_\eta(S)} \widehat{\mathsf{RR}}^{\mathcal{L}}(S', \eta)$ therefore contains

$$\bigcap_{S' \in \mathcal{A}_\eta(S)} \mathsf{Agree}(\mathcal{H}_\eta(S')) = \mathsf{Agree}\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)\right),$$

where the above equality holds because if $h \in \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$ then there exists $S' \in \mathcal{A}_\eta(S)$ such that $h \in \mathcal{H}_\eta(S')$, and if $h \in \mathcal{H}_\eta(S')$ for some $S' \in \mathcal{A}_\eta(S)$ then $h \in \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$. Finally, by uniform convergence (Anthony and Bartlett (2009) Theorem 4.10), if $S \sim \mathcal{D}^m$ for $m = O(\frac{1}{\epsilon^2}(d + \ln\frac{1}{\delta}))$ then with probability at least $1 - \delta$ we have $d_{\mathcal{D}}(h, h^*) \leq d_S(h, h^*) + \epsilon$ for all $h \in \mathcal{H}$. Thus, $\mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))$, and $\mathsf{RR}^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))$. ∎

Remark 8 (Linear separators for uniform distribution over the unit ball) *The above agreement region can be fairly large for the well-studied setting of learning linear separators under the uniform distribution $\mathcal{D}$ over the unit ball in $\mathbb{R}^d$, or more generally when the disagreement coefficient (Hanneke (2007)) is bounded. Since the disagreement coefficient is known to be at most $\pi\sqrt{d}$ in this setting (Hanneke (2007)), we have that $\Pr[DIS(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))] \leq \pi\sqrt{d}(2\eta + \epsilon)$ or $\Pr[\mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))] \geq 1 - \pi\sqrt{d}(2\eta + \epsilon)$. For example, if 1% of our data is poisoned and our dataset is large enough so that $\epsilon = 0.1\%$, even for $d = 9$ we are confident on over 80% of the points. We illustrate the relevant agreement region for linear separators in Figure 1.*

We can also obtain (slightly weaker) empirical estimates for the size of the robustly-reliable region for the learner $\mathcal{L}$ from Theorem 5. Such estimates could help determine when the algorithm can be safely fielded in an adversarial environment. For example, in federated learning, if an adversary controls processors that together hold an η fraction of the training data, and can also cause the learning algorithm to re-run itself at any time, then this would allow one to decide if a low abstention rate can be guaranteed or if additional steps need to be taken (like collecting more training data, or adding more security to processors).
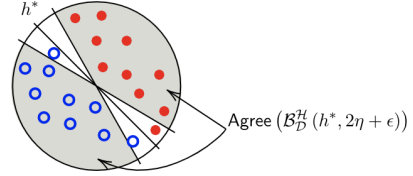

Figure 1: The agreement region of linear separators that can be given robustly-reliable predictions.

Theorem 9 *For any hypothesis class $\mathcal{H}$, the strongly robustly-reliable learner $\mathcal{L}$ from Theorem 5 satisfies the property that for all S, for all $\eta \geq 0$ and for all $S' \in \mathcal{A}_\eta(S)$,*

$$RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{H}_{3\eta}(S')).$$

Furthermore, $\mathcal{L}$ outputs a hypothesis $\hat{h}$ such that $RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}\left(\mathcal{B}_{S'}^{\mathcal{H}}(\hat{h}, 4\eta)\right)$.

Proof We use the fact that $S' \in \mathcal{A}_\eta(S)$ to conclude $\mathsf{Agree}(\mathcal{H}_{3\eta}(S')) \subseteq \mathsf{Agree}\left(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta)\right)$, which together with Theorem 7 implies the first claim. Indeed $h \in \mathcal{B}_S^\mathcal{H}(h^*, 2\eta)$ implies that $\mathrm{err}_S(h) \le 2\eta$ and therefore $\mathrm{err}_{S'}(h) \le 3\eta$. Also, if $\mathrm{err}_{S'}(h) \le 3\eta$, then $h \in \mathcal{B}_{S'}^\mathcal{H}(\hat{h}, 4\eta)$ since $\mathrm{err}_{S'}(\hat{h}) \le \eta$ again using the triangle inequality, which implies the second claim. ∎

It turns out that the bound on the robustly-reliable region from Theorem 7 is essentially optimal. We can show the following lower bound on the ability of any robustly-reliable learner for any hypothesis class $\mathcal{H}$ to be confident on any point in $\mathsf{Agree}(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta))$. Our upper and lower bounds extend the results of Gao et al. (2021) for learning halfspaces over the uniform distribution to general hypothesis classes and any distribution with bounded region of disagreement.

Theorem 10 *Let $\mathcal{L}$ be a strongly robustly-reliable learner for hypothesis class $\mathcal{H}$. Then for any $h^* \in \mathcal{H}$ and any sample S, any point in the η-robustly-reliable region must lie in the agreement region of $\mathcal{B}_S^\mathcal{H}(h^*, 2\eta)$. That is,*

$$RR^\mathcal{L}(S, h^*, \eta) \subseteq \mathsf{Agree}\left(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta)\right).$$

Moreover, if $S \sim \mathcal{D}^m$ then with probability $1 - \delta$, $\mathsf{Agree}(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta)) \subseteq \mathsf{Agree}(\mathcal{B}_\mathcal{D}^\mathcal{H}(h^*, 2\eta - \epsilon))$ for $m = O(\frac{1}{\epsilon^2}(d + \ln \frac{1}{\delta}))$, where $d = \mathrm{VCdim}(\mathcal{H})$. So, whp, $RR^\mathcal{L}(S, h^*, \eta) \subseteq \mathsf{Agree}(\mathcal{B}_\mathcal{D}^\mathcal{H}(h^*, 2\eta - \epsilon))$.

Proof Let $x \notin \mathsf{Agree}(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta))$. We will show that x cannot be in the η-robustly-reliable region. First, since $x \notin \mathsf{Agree}(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta))$, there must exist some $h' \in \mathcal{B}_S^\mathcal{H}(h^*, 2\eta)$ such that $h'(x) \ne h^*(x)$. Next, let S_X be the points in S with labels removed, and let $S_\mathcal{A}$ denote a labeling of S_X such that exactly half the points in $\Delta_S = \{x \in S_X \mid h'(x) \ne h^*(x)\}$ are labeled according to h^* (the remaining half using h', for convenience assume $|\Delta_S|$ is even). Notice that sample $S = \{(x_i, h^*(x_i)) \mid x_i \in S_X\}$ which labels points in S_X using h^* satisfies $S_\mathcal{A} \in \mathcal{A}_\eta(S)$ since $h' \in \mathcal{B}_S(h^*, 2\eta)$. Also for $S' = \{(x_i, h'(x_i)) \mid x_i \in S_X\}$ labeled by h', we have $S_\mathcal{A} \in \mathcal{A}_\eta(S')$. Now, assume for contradiction that $x \in RR^\mathcal{L}(S, h^*, \eta)$. This means that $\mathcal{L}_{S_\mathcal{A}}(x) = (y, \eta')$ for some $\eta' \ge \eta$. However, if $y \ne h^*(x)$, the learner is incorrectly confident for (true) dataset S since $S_\mathcal{A} \in \mathcal{A}_\eta(S)$. Similarly, if $y = h^*(x)$, the learner is incorrectly confident for sample S' since $h'(x) \ne h^*(x)$. Thus, $\mathcal{L}$ is not a strongly robustly-reliable learner and we have a contradiction.

Finally, by uniform convergence, if $S \sim \mathcal{D}^m$ for $m = O(\frac{1}{\epsilon^2}(d + \ln \frac{1}{\delta}))$ then with probability at least $1 - \delta$ we have $d_\mathcal{D}(h, h^*) \ge d_S(h, h^*) - \epsilon$ for all $h \in \mathcal{H}$. This implies that $\mathsf{Agree}(\mathcal{B}_S^\mathcal{H}(h^*, 2\eta)) \subseteq \mathsf{Agree}(\mathcal{B}_\mathcal{D}^\mathcal{H}(h^*, 2\eta - \epsilon))$ and so $RR^\mathcal{L}(S, h^*, \eta) \subseteq \mathsf{Agree}(\mathcal{B}_\mathcal{D}^\mathcal{H}(h^*, 2\eta - \epsilon))$ as desired. ∎

We can also extend the above result to $(1 - \gamma)$-probably robustly-reliable learners, as follows. See Appendix A for a proof.

Theorem 11 *Let $\mathcal{L}$ be a $(1 - \gamma)$-probably robustly-reliable learner for hypothesis class $\mathcal{H}$ under marginal $\mathcal{D}_X$. For any $h^* \in \mathcal{H}$, given a large enough sample size $m = |S| \ge \frac{c}{\epsilon^2} \ln \frac{1}{\delta}$, we have*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[RobC^\mathcal{L}(\mathcal{D}, \eta, S)] \le \Pr[\mathsf{Agree}(\mathcal{B}_\mathcal{D}^\mathcal{H}(h^*, 2\eta - \epsilon))] + 2\gamma + \delta,$$

where c is an absolute constant and $\mathcal{D}$ is the distribution with marginal $\mathcal{D}_X$ consistent with h^*.

We can strengthen Theorem 11 for learners that satisfy a somewhat stronger condition that with probability at least $1 - \gamma$ over the draw of the *unlabeled* sample S_X, for all targets $h^* \in \mathcal{H}$, the sample S produced by labeling S_X by h^* should be a good sample. See Theorem 16 (Appendix A).

4. Robustly-Reliable Learners for Instance-Targeted Malicious Noise

So far, our noise model has allowed the adversary to corrupt an *arbitrary* η fraction of the training examples. We now turn to the classic malicious noise model (Valiant, 1985; Kearns and Li, 1993) in which the points an adversary may corrupt are selected *at random*: each point independently with probability $1 - \eta$ is drawn from $\mathcal{D}$ and with probability η is chosen adversarially. Roughly, the malicious adversary model corresponds to a data poisoner that can *add* poisoned points to the training set (because the "clean" points are a true random sample from $\mathcal{D}$) whereas the $\mathcal{A}_\eta$ noise model corresponds to an adversary that can both add *and remove* points from the training set.

Traditionally, the malicious noise model has been examined in a non-instance-targeted setting. That is, the concern has been on the overall error-rate of classifiers learned in this model. Here, we consider instance-targeted malicious noise, and provide efficient robustly-reliable learners building on the seminal work of Awasthi et al. (2014, 2017). To discuss the set of possible adversarial corruptions in the malicious noise model with respect to particular random draws, for training set S and indicator vector $v \in \{0, 1\}^{|S|}$, let $\mathcal{A}^{mal}(S, v)$ denote the set of all possible S' achievable by replacing the points in S indicated by v with arbitrary points in $\mathcal{X} \times \mathcal{Y}$. We will be especially interested in $\mathcal{A}^{mal}(S, v)$ for $v \sim \text{Bernoulli}(\eta)^m$. Formally,

Definition 12 *Given sample S and indicator vector $v \in \{0, 1\}^{|S|}$, let $\mathcal{A}^{mal}(S, v)$ denote the collection of corrupted samples S' achievable by replacing the points in S indicated by v with arbitrary points in $\mathcal{X} \times \mathcal{Y}$. We use $\mathcal{A}^{mal}_\eta$ to denote $\mathcal{A}^{mal}(S, v)$ for $v \sim \text{Bernoulli}(\eta)^{|S|}$.*

We now define the notion of probably robustly-reliable learners and their robustly-reliable region for the malicious adversary model.

Definition 13 *A learner $\mathcal{L}$ is a $(1 - \gamma)$-**probably robustly-reliable learner against** $\mathcal{A}^{mal}_\eta$ for class $\mathcal{H}$ under marginal $\mathcal{D}_X$ at sample size m if for any target function $h^* \in \mathcal{H}$, with probability at least $1 - \gamma$ over the draw of $S \sim \mathcal{D}^m$ and $v \sim \text{Bernoulli}(\eta)^m$, for all $S' \in \mathcal{A}^{mal}(S, v)$ and all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \geq \eta$ then $y = h^*(x)$. The η-**robustly-reliable region** $RR^{\mathcal{L}}_{mal}(S, h^*, \eta, v) = \cap_{S' \in \mathcal{A}^{mal}(S,v)} \widehat{RR}^{\mathcal{L}}(S', \eta)$, that is, the set of points x given robust-reliability level $\eta' \geq \eta$ for all $S' \in \mathcal{A}^{mal}(S, v)$. The η-robustly-reliable correctness is defined as the probability mass of the η-robustly-reliable region, $RobC^{\mathcal{L}}_{mal}(\mathcal{D}, \eta, S, v) = \Pr_{\mathcal{D}_\mathcal{X}}[RR^{\mathcal{L}}_{mal}(S, h^*, \eta, v)]$.*

In Appendix B we show how this model relates to the $\mathcal{A}_\eta$ adversary from Section 3, and to other forms of the malicious noise model of Kearns and Li (1993).

4.1. Efficient algorithm for linear separators with malicious noise

For learning linear separators, the above approaches and prior work Gao et al. (2021) provide inefficient upper bounds, since we do not generically have an efficient ERM oracle for that class. In the following we show how we can build on an algorithm of Awasthi et al. (2017) for noise-tolerant learning over log-concave distributions to give strong per-point reliability guarantees that scale well with adversarial budget η. The algorithm of Awasthi et al. (2017) operates by solving an adaptive series of convex optimization problems, focusing on data within a narrow band around its current classifier and using an inner optimization to perform a soft outlier-removal within that band.[1] We

---

1. A simplified version of the algorithm without the outlier removal step was analyzed by Awasthi et al. (2015, 2016) for the case of learning under Massart noise.

will need to modify the algorithm to fit our setting, and then build on it to produce per-point reliability guarantees.

Theorem 14 *Let $\mathcal{D}_X$ be isotropic log-concave over $\mathbb{R}^d$ and $\mathcal{H}$ be the class of linear separators. There is a polynomial-time $(1 - \delta)$-probably robustly-reliable learner $\mathcal{L}$ against $\mathcal{A}_\eta^{mal}$ for class $\mathcal{H}$ under $\mathcal{D}_X$ based on Algorithm 2 of Awasthi et al. (2017), which uses a sample of size $m = poly(d, \frac{1}{\eta}, \log(1/\delta))$. Furthermore, for any $h^* \in \mathcal{H}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$ and $v \sim Bernoulli(\eta)^m$, we have $\Pr_{\mathcal{D}_X}[RR_{mal}^{\mathcal{L}}(S, h^*, \eta, v)] \geq 1 - \tilde{O}(\sqrt{d}\eta)$. The $\tilde{O}$-notation suppresses dependence on logarithmic factors and distribution-specific constants.*

Proof We will run a deterministic version of Algorithm 2 of Awasthi et al. (2017) (see Appendix C) and let $h = h(S')$ be the halfspace output by this procedure. By Theorem 27 (an analog of Theorem 4.1 of Awasthi et al. (2017) for the $\mathcal{A}_\eta^{mal}$ noise model), there is a constant C_0 such that $\text{err}_{\mathcal{D}}(h) \leq C_0\eta$ with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}^m$ and $v \sim Bernoulli(\eta)^m$.[2] In the following, we will assume the occurrence of this $1 - \delta$ probability event where the learner outputs a low-error hypothesis for any $S' \in \mathcal{A}^{mal}(S, v)$. We now show how we extend this result to *robustly-reliable* learning.

We will first show we can in principle output a robust-reliability value η on all points in the set $\text{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta))$. (Algorithmically, we will do something slightly different because we do not have an ERM oracle and so cannot efficiently test membership in the agreement region). Indeed by Theorem 27, h^* is guaranteed to be in $\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta)$. By the definition of the disagreement region, if $h(x) \neq h^*(x)$ for some x, then $x \in \text{DIS}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta))$. Therefore every point in $\text{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta))$ can be confidently classified with robust reliability level η.

Since we cannot efficiently determine if a point lies in $\text{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta))$, algorithmically we instead do the following. First, for any halfspace h', let $w_{h'}$ denote the unit-length vector such that $h'(x) = sign(\langle w_{h'}, x \rangle)$. Next, following the argument in the proof of Theorem 14 of Balcan and Long (2013), we show that for some constant C_1, for all values α, we have:

$$\text{Agree}\left(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta)\right) \supseteq \left\{x : ||x|| < \alpha\sqrt{d}\right\} \cap \left\{x : |\langle w_h, x \rangle| \geq C_1\alpha\eta\sqrt{d}\right\}. \tag{1}$$

Algorithmically, we will provide robustness level η for all points satisfying the right-hand-side above (which we can do efficiently), and the containment above implies this is legal. To complete the proof of the theorem, we must (a) give the proof of containment (1) and (b) prove that the intersection of the right-hand-sides of (1) over all $h = h(S')$ for $S' \in \mathcal{A}^{mal}(S, v)$ has probability mass $1 - \tilde{O}(\sqrt{d}\eta)$.

We begin with (a) giving the proof of the containment in formula (1). First, by Lemma 20 (due to Balcan and Long (2013)), $\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta)$ consists of hypotheses having angle at most $C_1\eta$ to h for some constant C_1. So, if $||x|| < \alpha\sqrt{d}$, then for any $h' \in \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0\eta)$ we have:

$$|\langle w_{h'}, x \rangle - \langle w_h, x \rangle| \leq ||w_{h'} - w_h|| \cdot ||x|| < C_1\alpha\eta\sqrt{d}.$$

---

2. While Awasthi et al. (2017) describe the adversary as making its malicious choices in a sequential order, the results all hold if the adversary makes its choices after the sample S has been fully drawn; that is, if the adversary selects an arbitrary $S' \in \mathcal{A}^{mal}(S, v)$. What requires additional care is that the results of Awasthi et al. (2017) imply that for any $S' \in \mathcal{A}^{mal}(S, v)$ the algorithm succeeds with probability $1 - \delta$, whereas we want that with probability $1 - \delta$ the algorithm succeeds for any $S' \in \mathcal{A}^{mal}(S, v)$; in particular, the adversary can make its choice after observing any internal randomness in the algorithm. We address this by making the algorithm deterministic, increasing its label complexity. For more discussion, see Appendix C.

Thus, if x also satisfies $|\langle w_h, x \rangle| \geq C_1 \alpha \eta \sqrt{d}$, we have $\langle w_h, x \rangle \langle w_{h'}, x \rangle > 0$. This implies that $x \in \mathsf{Agree}\left(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h, C_0 \eta)\right)$, completing the proof of containment (1).

Now we prove (b) that the intersection of the right-hand-sides of (1) over all $h = h(S')$ for $S' \in \mathcal{A}^{mal}(S, v)$ has probability mass $1 - \tilde{O}(\sqrt{d}\eta)$. To do this, we first show that for all h such that $\mathrm{err}_{\mathcal{D}}(h) \leq C_0 \eta$ we have:

$$B_{\alpha\sqrt{d}} \cap \{x : |\langle w_h, x \rangle| \geq C_1 \alpha \eta \sqrt{d}\} \supseteq B_{\alpha\sqrt{d}} \cap \{x : |\langle w_{h^*}, x \rangle| \geq 2C_1 \alpha \eta \sqrt{d}\}, \qquad (2)$$

where $B_{\alpha\sqrt{d}} = \{x : ||x|| < \alpha\sqrt{d}\}$ is the open ball of radius $\alpha\sqrt{d}$ centered at the origin in the Euclidean d-space. This follows from the same argument used to prove (a) above. Specifically, by Lemma 20, $\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, C_0 \eta)$ consists of hypotheses having angle at most $C_1 \eta$ to w_{h^*} for the same constant C_1 as above. So, if $||x|| < \alpha\sqrt{d}$, then for any $h \in \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, C_0 \eta)$ we have

$$|\langle w_h, x \rangle - \langle w_{h^*}, x \rangle| \;\leq\; ||w_h - w_{h^*}|| \cdot ||x|| \;<\; C_1 \alpha \eta \sqrt{d}.$$

This means that if $|\langle w_{h^*}, x \rangle| \geq 2C_1 \alpha \eta \sqrt{d}$ then $|\langle w_h, x \rangle| \geq C_1 \alpha \eta \sqrt{d}$.

To complete the proof of (b), we now just need to show that the probability mass of the right-hand-side of (2) is at least $1 - \tilde{O}(\sqrt{d}\eta)$. This follows using the argument in the proof of Theorem 14 in Balcan and Long (2013). First we use the fact that for an isotropic log-concave distribution over $\mathbb{R}^d$, we have $\mathrm{Pr}_{\mathcal{D}_{\mathcal{X}}}(||x|| \geq \alpha\sqrt{d}) \leq e^{-\alpha+1}$ (Lemma 19). This means we can choose $\alpha = \ln(1/(\sqrt{d}\eta))$ and ensure that at most an $O(\sqrt{d}\eta)$ probability mass of points x fail to satisfy the first term in the right-hand-side of (2). For the second term, using the fact that marginals of isotropic log-concave distributions are also isotropic log-concave (in particular, the 1-dimensional marginal given by taking inner-product with w_{h^*}) and that the density of a 1-dimensional isotropic log-concave distribution is at most 1 (Lemma 19), at most an $O(\alpha\eta\sqrt{d})$ probability mass of points x fail to satisfy the second term in the right-hand-side of (2). Putting these together yields the theorem. ∎

Remark. Note that the above result improves over a related distribution-specific result of Gao et al. (2021) in three ways. First, we are able to certify that the prediction of our algorithm is correct rather than only certifying that the prediction would not change due to the adversary intervention. Second, we are able to handle much more general distribution — any isotropic logconcave distribution, as opposed to just the uniform distribution. We essentially achieve the same $\tilde{O}(\sqrt{d}\eta)$ upper bound on the uncertified region but for a much larger class of distributions. Finally, unlike Gao et al. (2021), our algorithm is *polynomial time*.

Lower bounds. In Appendix D we give a near-matching lower bound on robust reliability to $\mathcal{A}_{\eta}^{mal}$.

5. Active Robustly Reliable Learners

Instance-targeted attacks may also occur in situations where the labeled data is expensive to obtain and one would typically use active learners to learn accurate classifiers with fewer labeled examples. For example, consider translation to a rare language for which an expert can be expensive to obtain. We might have access to proprietary labeled data, but may be charged per label and would like to minimize the cost of obtaining the labels. It is possible that an attacker with access to the data corrupts it in an instance-targeted way.

We consider a pool-based active learning setting (Settles (2009)), where the learner draws (labeled and unlabeled) examples from a large pool of examples which may have been corrupted by

an instance-targeted adversary. That is, a dataset S is drawn according to some true distribution $\mathcal{D}$ and the instance-targeted adversary may corrupt an η fraction of the dataset arbitrarily (change both data points and labels). The learner has access to unlabeled examples from the corrupted dataset S', and is allowed to make *label* queries for examples in the dataset. The goal of the learner is to produce a robustly-reliable predictor for the marginal distribution $\mathcal{D}_X$, and using as few labels as possible. Our (passive) robustly-reliable learner in Theorem 7 needs $\tilde{O}(\frac{d}{\epsilon^2})$ examples; in this section we explore if we can actively robustly-reliably learn with fewer labeled examples.

Our results are stated formally and proven in Appendix E. Here we briefly summarize the results. In Theorem 30, we propose an active learning method that, for any given η, produces an *optimal* η-robustly-reliable region (i.e., matching the upper and lower bounds of Theorems 7 and 10), while making a number of label queries that, for small η, is significantly smaller than the number m of examples in S' (the corrupted sample), when the disagreement coefficient θ is small: specifically, it makes a number of queries $\tilde{O}(\theta\eta m + \theta d)$. We note that in this setting, because robust-reliability requires reliability for *any* corruption S' of S, we can effectively think of the adversary as being *adaptive*, meaning it can even adaptively choose its corruptions based on which points the learner will choose to query. For instance, it could choose to corrupt the labels of every one of the first ηm queries made by the learning algorithm, so that the learner might need ηm queries just to *find* a single uncorrupted sample. The algorithm we propose for this challenging setting is actually quite simple: we process the corrupted data S' in sequence, and for each new point x_i we query its label y_i iff x_i is in the region of disagreement of the set of all $h \in \mathcal{H}$ that make at most ηm mistakes among all previously-queried points. In the end, after processing all m examples in S', we use the region of agreement of this same set of classifiers as the empirical η-robustly-reliable region, predicting using their agreed-upon label for any test point x in this region. In addition to this result, in Theorem 32 we also show further reductions in label complexity are possible when the adversary is oblivious to the internal randomness of the learner.

6. Robustly reliable agnostic learners

So far we have assumed that the uncorrupted samples S are realizable under our concept class $\mathcal{H}$. We now show that our results can be extended to the non-realizable setting, i.e. $\min_{h \in \mathcal{H}} \text{err}_S(h) > 0$, with weaker but still interesting guarantees. Specifically, our algorithm might now produce an incorrect prediction (y, η) with η greater than the adversary's power, but only if *every* hypothesis in $\mathcal{H}$ with low error on the uncorrupted S would also be incorrect on that example.

We can define a ν-*tolerably robustly-reliable* learner in the non-realizable setting as the learner whose reliable predictions agree with every low error hypothesis (error at most ν) on the uncorrupted sample. If there are multiple $h \in \mathcal{H}$ satisfying $\text{err}_S(h) \leq \nu$ and they disagree on x then it must be the case that the algorithm outputs $\eta < 0$. The setting $\nu = 0$ yields the usual robustly-reliable learner. The (ν, η)-*robustly-reliable region* $\text{RR}^{\mathcal{L}}(S, \nu, \eta)$ and *empirical (ν, η)-robustly-reliable region* $\widehat{\text{RR}}^{\mathcal{L}}(S', \nu, \eta)$ are defined as in the realizable setting. Similarly to Definition 3, we can also extend this to a robustly-reliable learner with high probability for an adversarially-corrupted sample drawn from a given distribution. A $(1 - \gamma)$-*probably ν-tolerably robustly-reliable learner* for concept space $\mathcal{H}$ would be tolerably robustly-reliable with probability at least $1 - \gamma$ over the draw of samples. Formal definitions appear in Appendix F.

We provide a ν-tolerably strongly robustly-reliable learner using the notion of agreement regions (Theorem 37). Our results here generalize corresponding results from Section 3. We first present our

learner and a guarantee on the learner's empirical (ν, η)-robustly-reliable region given any (possibly corrupted) dataset. Our algorithm assumes ν is given. The learner's empirical robustly-reliable region contains the agreement region of hypotheses with error at most $\eta + \nu$ on the corrupted sample, which implies agreement with all hypotheses with error at most ν on the uncorrupted dataset. Using arguments similar to those in the proof of Theorem 7, we can show that the (ν, η)-robustly-reliable region of our learner contains the agreement region of hypotheses with error at most $2\eta + \nu$ on the uncorrupted sample, i.e. $\mathrm{Agree}\,(H_{2\eta+\nu}(S))$ (Theorem 38). We also show a matching lower bound by showing that any point not in the agreement region may be successfully attacked for any learner (Theorem 39).

We observe that any 2ν-tolerably strongly robustly-reliable learner for the concept space can be used to give a $(1 - \gamma)$-probably ν-tolerably robustly-reliable learner over a distribution $\mathcal{D}$ where the best hypothesis h^* in $\mathcal{H}$ has error $\mathrm{err}_{\mathcal{D}}(h^*) = \nu$ (Theorem 40). We can use this reduction, together with our ν-tolerably strongly robustly-reliable learner, to give a probably tolerably robustly-reliable learner for any distribution $\mathcal{D}$ along with guarantees about its distribution-averaged robust reliability region (Theorem 41). We further show a lower bound on the robust reliability region of *any* $(1 - \gamma)$-probably ν-tolerably robustly-reliable learner which implies that our learner is nearly optimal (Theorem 42). The key idea in establishing this lower bound is the creation of two distributions with the same marginal but nearly-consistent with two hypotheses which are close in error. See Appendix F for formal statements of the results and proofs.

7. Discussion

In this work, we provide the first correctness guarantees (specifically robust reliability) for instance-targeted training time attacks, along with guarantees about the size of the robustly-reliable region. One implication of our results is a combined robustness against test-time attacks as well, where an adversary makes imperceptible or irrelevant perturbations to the *test* examples in order to induce errors. Specifically, even without being given any knowledge of the kinds of perturbations such an adversary could make, our algorithms in the worst case will simply abstain (output $\eta < 0$) or produce an η less than the perturbation budget of the adversary if the adversary is performing both kinds of attacks together. In the agnostic case, the adversary could cause us to make a mistake, but only if *every* low-error hypothesis in the class would have also made a mistake. An interesting open question is to provide usefulness guarantees for test time attacks, for reliable predictors: that is, nontrivial bounds on the probability that an adversary's perturbation of the test point can cause the predictor to abstain.

Acknowledgments

References

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 449–458, 2014.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Conference on Learning Theory*, pages 167–190. PMLR, 2015.

Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory*, pages 152–192. PMLR, 2016.

Pranjal Awasthi, Maria Florina Balcan, and Philip M Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM (JACM)*, 63(6):1–27, 2017.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316. PMLR, 2013.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, pages 16–25, 2006.

Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack. In *Conference on Learning Theory (COLT)*, 2021.

Nader H Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Sanjoy Dasgupta, Daniel J Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems*, 20, 2007.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010.

Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. *The Conference on Uncertainty in Artificial Intelligence (UAI)*, 2021.

Jonas Geiping, Liam H Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *International Conference on Learning Representations*, 2020.

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *Advances in Neural Information Processing Systems*, 33:15859–15870, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*, 2015.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, pages 353–360, 2007.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(12), 2009.

Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defense against general poisoning attacks. *International Conference on Learning Representations (ICLR)*, 2021.

László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing adversarially robust learning to non-robust PAC learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pages 3452–3482. PMLR, 2021.

Mehran Mozaffari-Kermani, Susmita Sur-Kolay, Anand Raghunathan, and Niraj K Jha. Systematic poisoning attacks on and defenses for machine learning in healthcare. *IEEE journal of biomedical and health informatics*, 19(6):1893–1905, 2014.

Ronald L Rivest and Robert H Sloan. Learning complicated concepts reliably and usefully. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 635–640, 1988.

Burr Settles. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pages 6103–6113, 2018.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems*, 30, 2017.

Octavian Suciu, Radu Marginean, Yigitcan Kaya, Hal Daume III, and Tudor Dumitras. When does machine learning FAIL? generalized transferability for evasion and poisoning attacks. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1299–1316, 2018.

Leslie G Valiant. Learning disjunctions of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial intelligence*, pages 560–566, 1985.

Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094. PMLR, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Appendix A. Additional lower bounds in the realizable setting

Proof of Theorem 11 Let $\mathcal{L}$ be a $(1 - \gamma)$-probably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$ and let $h^* \in \mathcal{H}$. Let $x \notin \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta - \epsilon))$. To prove the theorem, it suffices to prove that $\Pr_{S \sim \mathcal{D}^m}[x \in \mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)] \leq 2\gamma + \delta$.

Select some $h' \in \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta - \epsilon)$ such that $h'(x) \neq h^*(x)$; such an h' exists by definition of the disagreement region. Let $S \sim \mathcal{D}^m$ and define $S' = \{(x, h'(x)) \mid (x, h^*(x)) \in S\}$. Note that $S' \sim \mathcal{D}'^m$, where $\mathcal{D}'$ is a data distribution with the same marginal as $\mathcal{D}$ but consistent with h'. We now consider three bad events of total probability at most $2\gamma + \delta$: (A) $\mathcal{L}$ is not robustly-reliable for all datasets in $\mathcal{A}_\eta(S)$, (B) $\mathcal{L}$ is not robustly-reliable for all datasets in $\mathcal{A}_\eta(S')$, and (C) $d_S(h', h^*) > 2\eta$. Indeed events (A) and (B) occur with probability at most γ each since $\mathcal{L}$ is given

17

to be a $(1 - \gamma)$-probably robustly-reliable learner. Since $h' \in \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta - \epsilon)$, for any $x \sim \mathcal{D}_X$, the probability that $h'(x) \neq h(x)$ is at most $2\eta - \epsilon$. By Hoeffding's inequality, if $m \geq \frac{1}{2\epsilon^2} \ln \frac{1}{\delta}$, we have that $d_S(h', h^*) > 2\eta$ (i.e., event (C)) occurs with probability at most δ.

We claim that if none of these bad events occur, then $x \notin \mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)$. To prove this, assume for contradiction that none of the bad events occur and $x \in \mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)$. Let $\tilde{S}_{h',h^*}$ denote a relabeling of S such that exactly half the points in $\Delta_S = \{x \mid (x, y) \in S, h'(x) \neq h^*(x)\}$ are labeled according to h^* (the remaining half using h', for convenience assume $|\Delta_S|$ is even). Note that since bad event (C) did not occur, this means that $\tilde{S}_{h',h^*} \in \mathcal{A}_\eta(S)$ and $\tilde{S}_{h',h^*} \in \mathcal{A}_\eta(S')$, Now, since $x \in \mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)$ and $\tilde{S}_{h',h^*} \in \mathcal{A}_\eta(S)$, it must be the case that $\mathcal{L}_{\tilde{S}_{h',h^*}}(x) = (y, \eta')$ for some $\eta' \geq \eta$. But, if $y \neq h^*(x)$ then this implies bad event (A), and if $y \neq h'(x)$ then this implies bad event (B). So, x cannot be in $\mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)$ as desired. ∎

We show how we can strengthen the lower bound of Theorem 11 for learners that satisfy the somewhat stronger condition of being a $(1 - \gamma)$-uniformly robustly-reliable learner (defined below). This condition requires that with probability at least $1 - \gamma$ over the draw of the *unlabeled* sample $S_X \in \mathcal{H}$, for all targets $h^* \in \mathcal{H}$, the sample S produced by labeling S_X by h^* should be a good sample. In contrast, Theorem 11 requires only that for any h^*, with probability at least $1 - \gamma$ over the draw of S, S is a good sample (different samples may be good for different target functions).

Definition 15 $\mathcal{L}$ *is a* $(1 - \gamma)$**-uniformly robustly-reliable learner** *for class $\mathcal{H}$ under marginal distribution $\mathcal{D}_X$ if with probability at least $1 - \gamma$ over the draw of an unlabeled sample $S_X \sim \mathcal{D}_X^m$, for all $h^* \in \mathcal{H}$, for all $S' \in \mathcal{A}_\eta(S)$ (where $S = \{(x, h^*(x)) : x \in S_X\}$), for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \geq \eta$ then $y = h^*(x)$.*

Theorem 16 *Let $\mathcal{L}$ be a $(1 - \gamma)$-uniformly robustly-reliable learner for hypothesis class $\mathcal{H}$ under marginal $\mathcal{D}_X$. Then, for any $h^* \in \mathcal{H}$, for $S \sim \mathcal{D}^m$ (where $\mathcal{D}$ is the distribution over examples labeled by h^* with marginal $\mathcal{D}_X$), with probability $1 - \gamma$ over the draw of S we must have*

$$RR^{\mathcal{L}}(S, h^*, \eta) \subseteq \mathsf{Agree}\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)\right).$$

Moreover, if $m \geq \frac{c}{\epsilon^2}(d + \ln \frac{1}{\delta})$, where c is an absolute constant and d is the VC-dimension of $\mathcal{H}$, then with probability $1 - \delta$, $\mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)) \subseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta - \epsilon))$. So, $\mathbb{E}_{S \sim \mathcal{D}^m}[RobC^{\mathcal{L}}(\mathcal{D}, \eta, S)] \leq \Pr[\mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta - \epsilon))] + \gamma + \delta$.

Proof The proof is similar to the proof for Theorem 10. We are given that $\mathcal{L}$ is a $(1 - \gamma)$-uniformly robustly-reliable learner. This means that with probability $1 - \gamma$, the unlabeled sample S_X has the property that the learner would be robustly-reliable for any target function in $\mathcal{H}$. Assume this event occurs, and consider any $x \notin \mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta))$. We will show that x cannot be in the η-robustly-reliable region. First, since $x \notin \mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta))$, there must exist some $h' \in \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$ such that $h'(x) \neq h^*(x)$. Next, let $S_{\mathcal{A}}$ denote a labeling of S_X such that exactly half the points in $\Delta_S = \{x \in S_X \mid h'(x) \neq h^*(x)\}$ are labeled according to h^* (the remaining half using h', for convenience assume $|\Delta_S|$ is even). Notice that S satisfies $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$ since $h' \in \mathcal{B}_S(h^*, 2\eta)$. Also for $S' = \{(x_i, h'(x_i)) \mid x_i \in S_X\}$ labeled by h', we have $S_{\mathcal{A}} \in \mathcal{A}_\eta(S')$. Now, assume for contradiction that $x \in \mathrm{RR}^{\mathcal{L}}(S, h^*, \eta)$. This means that $\mathcal{L}_{S_{\mathcal{A}}}(x) = (y, \eta')$ for some $\eta' \geq \eta$. However, if $y \neq h^*(x)$, the learner is incorrectly confident for (true) dataset S since $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$. Similarly, if $y = h^*(x)$, the learner is incorrectly confidnet for sample S' since $h'(x) \neq h^*(x)$. This contradicts our assumption about S_X.

The second part of the theorem statement follows directly from standard uniform convergence bounds. ∎

Appendix B. Relating $\mathcal{A}_\eta^{mal}$ to other models

In this section we provide some observations that relate our instance-targeted malicious noise adversarial model from Section 4 to models from prior and current work. We begin by noting that learnability against the sample-based attacks defined above implies learnability under attacks which directly corrupt the distribution $\mathcal{D}$ with malicious noise.

Lemma 17 *Suppose $\mathcal{D}$ is a distribution over $\mathcal{X} \times \mathcal{Y}$. Under the malicious noise model of Kearns and Li (1993), the learner has access to samples from $\mathcal{D}_\eta(\mathcal{A}') = (1 - \eta)\mathcal{D} + \eta\mathcal{A}'$, where $\mathcal{A}'$ is an arbitrary adversarial distribution over $\mathcal{X} \times \mathcal{Y}$. Let $S \sim \mathcal{D}^m$ and $S' \sim \mathcal{D}_\eta(\mathcal{A}')^m$. Then there exists adversary $A_1 \in \mathcal{A}_\eta^{mal}$ such that $A_1(S)$ is distributed identically as S'.*

Proof Select A_1 as the adversary which flips a coin for each example in S and with probability η replaces the example with one generated by $\mathcal{A}'$. ∎

This implies that if we have a learner that is robust to adversaries in $\mathcal{A}_\eta^{\mathrm{mal}}$, it will also work for the malicious noise model of Kearns and Li (1993). Next, we observe that the $\mathcal{A}_\eta^{mal}$ model is a special case of the $\mathcal{A}_{\eta+\epsilon}$ model in the following sense.

Lemma 18 *Let S be a sample of large enough size $m = |S| \geq \frac{1}{\epsilon^2} \log \frac{2}{\delta}$, and S' be a corruption of S by malicious adversary $\mathcal{A}_\eta^{mal}$. Then with probability at least $1 - \delta$ over the randomness of $\mathcal{A}_\eta^{mal}$, the corrupted sample satisfies $S' \in \mathcal{A}_{\eta+\epsilon}(S)$.*

Proof By Hoeffding's inequality, with probability at least $1 - \delta$, the $\mathcal{A}_\eta^{\mathrm{mal}}$ adversary may corrupt at most $\eta + \epsilon$ fraction of examples in S. ∎

Appendix C. Further discussion of Awasthi et al. (2017) and modification to our setting

Awasthi et al. (2017) define the adversary in the malicious noise model as making its choices in a sequential order as training examples are drawn, with knowledge of the points drawn so far but not of those still to come or of the random choices the algorithm will subsequently make. In our setting, we want with high probability over the sample $S \sim \mathcal{D}^m$, the draw of $v \sim \mathrm{Bernoulli}(\eta)^m$ and any randomness in the algorithm, for the algorithm to succeed for all $S' \in \mathcal{A}^{mal}(S, v)$. This corresponds to an adversary that can make its choices after both (a) the entire sample S is drawn and (b) any internal randomness in the algorithm has been fixed. Difference (a) is easy to handle, and indeed the analysis of Awasthi et al. (2017) goes through directly. To address (b) we will modify the algorithm to make it deterministic; this increases its label complexity, though it remains polynomial. Here, we review the relevant portions of the algorithm and analysis in Awasthi et al. (2017) and describe needed changes. We begin with three useful facts about log-concave distributions:

Lemma 19 (Lovász and Vempala (2007)) *Assume $\mathcal{D}$ is an isotropic log-concave distribution over $\mathbb{R}^d$. Then $\Pr_{x\sim\mathcal{D}}(\|x\| \geq \alpha\sqrt{d}) \leq e^{-\alpha+1}$. Furthermore, if $d = 1$ then $\Pr_{x\sim\mathcal{D}}(x \in [a,b]) \leq |b - a|$.*

Lemma 20 (Balcan and Long (2013) Lemma 3) *Assume $\mathcal{D}$ is an isotropic log-concave distribution over $\mathbb{R}^d$. Then there exists c such that for any two unit vectors u and v in $\mathbb{R}^d$ we have $c\theta(u, v) \leq dis_{\mathcal{D}}(u, v) = Pr_{(x,y)\sim\mathcal{D}}sgn(\langle u, x\rangle) \neq sgn(\langle v, x\rangle)$.*

Theorem 21 (Balcan and Long (2013) Theorem 14) *Let $\beta \geq 0$ be a sufficiently small constant. Assume that $\mathcal{D}$ is an isotropic β logconcave distribution in $\mathbb{R}^d$. Then the disagreement coefficient (Hanneke, 2007) is given by $\theta(\epsilon) = O(d^{\frac{1}{2}+\frac{\beta}{2\ln 2}}\log(1/\epsilon))$.*

Now we show how we adapt the algorithm and analysis from Awasthi et al. (2017). Algorithms 2 and 3 below contain the algorithm from Awasthi et al. (2017) for learning from malicious noise, adapted to our model and restated to our notation. As noted above, one key change to the algorithm is to make it deterministic: specifically, Step 2(b) of Algorithm 2 examines the labels of all points in set W rather than just a random sample in computing and optimizing weighted hinge-loss. We verify that the arguments in Awasthi et al. (2017) continue to apply to this algorithm in the A_η^{mal} model (in particular, that the noisy and non-noisy points continue to satisfy the properties used in the analysis), and highlight where particular care or modifications are needed.

Algorithm 2 EFFICIENT LEARNING OF LINEAR SEPARATORS IN THE $\mathcal{A}_\eta^{mal}$ MODEL

Input: allowed error rate ϵ, probability of failure δ, sequences of sample sizes $n_k > 0$, $k = 1, 2, 3, ...$, a sequence of cut-off values $b_k > 0$, a sequence of hypothesis space radii $r_k > 0$, a sequence of removal rates ξ_k, a sequence of variance bounds σ_k^2, precision value κ; weight vector w_0. Note that compared to Awasthi et al. (2017), we have replaced the random sampling inside W with just directly using loss with respect to weighting p.

1. Place the first n_1 examples in S' into a working set W.

2. For $k = 1, \ldots, s = \lceil \log_2(1/\epsilon) \rceil$

   a. Apply Algorithm 3 to W with parameters $u \leftarrow w_{k-1}$, $\gamma \leftarrow b_{k-1}$, $r \leftarrow r_k$, $\xi \leftarrow \xi_k$, $\sigma^2 \leftarrow \sigma_k^2$ and let q be the output function $q : W \rightarrow [0,1]$. Normalize q to form a probability distribution p over W.

   b. Find $v_k \in B(w_{k-1}, r_k)$, $\|v_k\|_2 \leq 1$, to approximately minimize hinge loss in W weighted by p: $\ell_{\tau_k}(v_k, p) \leq \min_{w\in B(w_{k-1},r_k)\cap B(0,1)} \ell_{\tau_k}(w, p) + \kappa/32$. Normalize v_k to have unit length, yielding $w_k = \frac{v_k}{\|v_k\|_2}$.

   c. Let $W = \{\}$. Scan S' until n_{k+1} data points x have been found such that $|w_k \cdot x| < b_k$, and put them into W. (If S' is exhausted before this step is completed, then output failure.)

Output: weight vector w_s of error at most ϵ with probability $1 - \delta$.

The first step in the analysis (Theorem 4.2 in Awasthi et al. (2017)) is to show that with high probability, the convex program in Algorithm 3 has a feasible solution q^*, specifically one in which

---
Algorithm 3 LOCALIZED SOFT OUTLIER REMOVAL PROCEDURE Awasthi et al. (2017)

Input: a set W of samples; the reference unit vector u; desired radius r; a parameter ξ specifying the desired bound on the fraction of clean examples removed; a variance bound σ^2

1. Find $q : W \to [0, 1]$ satisfying the following constraints:

   a. for all $x \in W$, $0 \leq q(x) \leq 1$

   b. $\frac{1}{|W|} \sum_{(x,y) \in W} q(x) \geq 1 - \xi$

   c. for all $w \in B(u, r) \cap B(\mathbf{0}, 1)$, $\frac{1}{|W|} \sum_{x \in W} q(x)(w \cdot x)^2 \leq \sigma^2$.

Output: A function $q : W \to [0, 1]$.

$q^*(x) = 0$ for each noisy point and $q^*(x) = 1$ for each clean point. This analysis makes the worst-case assumption that every adversarial point is placed inside the band $|w_k \cdot x| < b_k$. Moreover, since $q^*(x) = 0$ for each noisy point, there is no dependence on its location or label, so this q^* remains feasible in our setting. Also, the analysis uses the fact that the non-noisy points are a true random sample from $\mathcal{D}$ (which is true in the $\mathcal{A}_\eta^{mal}$ model but not the $\mathcal{A}_\eta$ model used in Section 3). For completeness, we present the theorem and summarize its analysis.

First, let η' denote the value η divided by the probability mass under $\mathcal{D}$ of the band $|w_{k-1} \cdot x| < b_{k-1}$. This is an upper bound on the probability a random example in the band is noisy even if the adversary places all of its points inside the band. By Chernoff bounds, $n_k \geq \text{poly}(1/\eta', 1/\delta)$ suffices so that with probability at least $1 - \delta$, at most a $2\eta'$ fraction of the points in W are controlled by the adversary. Assume this is indeed the case. Also, let $D_{w_{k-1}, b_{k-1}}$ denote the distribution $\mathcal{D}$ restricted to the band $|w_{k-1} \cdot x| < b_{k-1}$.

Theorem 22 (Analog of Theorem 4.2 of Awasthi et al. (2017) for the $\mathcal{A}_\eta^{mal}$ noise model) *For any $C > 0$, there is a constant c and a polynomial p such that, for all $\xi > 2\eta'$ and all $0 < \gamma < C$, if $n_{k+1} \geq p(1/\eta', d, 1/\xi, 1/\delta, 1/\gamma, 1/r)$, then, with probability $1 - \delta$, the output q of Algorithm 3 satisfies the following:*

- *$\sum_{x \in W} q(x) \geq (1 - \xi)|W|$ (a fraction $1 - \xi$ of the weight is retained)*

- *For all unit length w such that $\|w - u\|_2 \leq r$,*

$$\frac{1}{|W|} \sum_{x \in W} q(x)(w \cdot x)^2 \leq c(r^2 + \gamma^2). \tag{6}$$

Furthermore, the algorithm can be implemented in polynomial time.

Proof (Sketch) The theorem is proven by analyzing the clean data points, namely those that are not under the adversary's control, using a pseudodimension-based uniform convergence analysis and properties of log-concave distributions. This uses the fact that the *non-noisy* points are a true random sample from the underlying distribution, so this does not apply to the $\mathcal{A}_\eta$ model used in Section 3.

Specifically, the proof uses the following two properties of isotropic log-concave distributions.

Lemma 23 (Lemma 4.3 of Awasthi et al. (2017)) *If we draw ℓ times i.i.d. from $D_{w_{k-1}, b_{k-1}}$ to form X_C, with probability $1 - \delta$, we have that for any unit length a,*

$$\frac{1}{\ell} \sum_{x \in X_C} (a \cdot x)^2 \leq E[(a \cdot x)^2] + \sqrt{\frac{O(d \log(\ell/\delta)(d + \log(1/\delta)))}{\ell}}.$$

Lemma 24 (Lemma 3.4 of Awasthi et al. (2017)) *Assume that D is isotropic log-concave. For any c_3, there is a constant c_4 such that, for all $0 < \gamma \leq c_3$, for all a such that $\|u - a\|_2 \leq r$ and $\|a\|_2 \leq 1$*

$$\mathbf{E}_{x \sim D_{u,\gamma}}((a \cdot x)^2) \leq c_4(r^2 + \gamma^2).$$

The two lemmas above, along with the fact that the non-noisy points in the $\mathcal{A}_\eta^{mal}$ model are indeed distributed according to $\mathcal{D}$, imply that for all $w \in B(u, r)$,

$$\frac{1}{|W|} \sum_{x \in W} q^*(x)(a \cdot x)^2 \leq 2\mathbf{E}[(a \cdot x)^2] \leq c(r^2 + \gamma^2),$$

for an appropriate constant c, where $q^*(x) = 0$ for each noisy point and $q^*(x) = 1$ for each non-noisy point as defined above. Finally, an efficient separation oracle for the convex program yields the theorem. ∎

The next step is a further analysis of the working set W in each round k of Algorithm 2. The analysis partitions W into the "clean" set W_C drawn from the true distribution $D_{w_{k-1}, b_{k-1}}$ of points from $\mathcal{D}$ subject to lying inside the previous band, and the "dirty" set W_D controlled by the adversary. Lemma 4.5 in Awasthi et al. (2017) upper-bounds the size of W_D, and its analysis makes the worst-case assumption that every adversarial point lies inside the band. Since our setting is identical in terms of the probabilistic selection of which points are controlled by the adversary (differing only in allowing the adversary to make its choice after all of S is drawn), this argument also goes through directly. Specifically, Lemma 4.5, adapted to our setting, is as follows:

Lemma 25 (Analog of Lemma 4.5 of Awasthi et al. (2017) for the $\mathcal{A}_\eta^{mal}$ noise model) *There is an absolute positive constant c such that, with probability $1 - \frac{\delta}{6(k+k^2)}$,*

$$|W_D| \leq c\eta n_k M^k \leq \frac{cM\eta n_k}{\epsilon} \tag{7}$$

Proof (Sketch) Since the underlying distribution of non-noisy points is isotropic log-concave, the probability that a random example from $\mathcal{D}$ falls in the band is $\Omega(M^{-k})$. By Chernoff bounds, this implies that $O(n_k M^k)$ examples from $\mathcal{D}$ are sufficient so that with probability at least $(1 - \frac{\delta}{12(k+k^2)})$, n_k examples fall inside the band. In the $\mathcal{A}_\eta^{mal}$ model, the probability that each example drawn is noisy is η, and in the worst case, the adversary can place each noisy example inside the band. Since with probability at least $(1 - \frac{\delta}{12(k+k^2)})$ the total number of examples drawn is $O(n_k M^k)$, Chernoff bounds imply that for some constant c, with probability at least $1 - \frac{\delta}{12(k+k^2)}$ we have $|W_D| \leq c\eta n_k M^k$. The second inequality follows from the fact that $k \leq \lceil \log_M(1/\epsilon) \rceil$. ∎

The bound of Lemma 25 above is then used in Lemma 4.7 of Awasthi et al. (2017), which relates the average hinge loss of examples in W when weighted according to the solution q found

by Algorithm 3 (scaled to sum to 1 over W) to the average hinge loss of the clean examples W_C weighted uniformly. In particular, it shows that the soft outlier-removal performed by Algorithm 3 is comparable to correctly identifying and removing the adversarial data with respect to the loss of any $w \in B(w_{k-1}, r_k)$. Lemma 25 is used in equation (11) of the proof of Lemma 4.7 (see below), and is the only property of W_D used in the proof.

Specifically, let p denote the weighting q over W produced by Algorithm 3, scaled to sum to 1. Constraints a and b in the definition of q ensure that the total variation distance between p and the uniform distribution over W is at most ξ. Now, let $\ell(w, p)$ be the average hinge-loss of points in W weighted by p, and let $\ell(w, W_C)$ denote the average hinge loss of the clean examples W_C weighted uniformly. Lemma 4.7 of Awasthi et al. (2017) relates these as follows:

Lemma 26 (Analog of Lemma 4.7 of Awasthi et al. (2017) for the $\mathcal{A}_\eta^{mal}$ noise model) *There are absolute constants C_1, C_2 and C_3 such that, with probability $1 - \frac{\delta}{2(k+k^2)}$, if we define $z_k = \sqrt{r_k^2 + b_{k-1}^2}$, then for any $w \in B(w_{k-1}, r_k)$, we have*

$$\ell(w, W_C) \leq \ell(w, p) + \frac{C_1 \eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/32 \tag{8}$$

and

$$\ell(w, p) \leq 2\ell(w, W_C) + \kappa/32 + \frac{C_2 \eta}{\epsilon} + C_3\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k}{\tau_k}. \tag{9}$$

Proof (Sketch) Given the lemmas shown above, the proof now follows exactly as in Awasthi et al. (2017). In particular, the key properties of the data used are the size of the set W_D, which comes from Lemma 25 analyzed above and is used inequation (11) below, and the fact that the clean data is distributed according to $\mathcal{D}$ (which is true in the $\mathcal{A}_\eta^{mal}$ model). Specifically, from the lemmas above, with probability at least $1 - \frac{\delta}{2(k+k^2)}$, there are constants K_1, K_2 and K_3 such that

$$\frac{1}{|W|}\sum_{x \in W} q(x)(w \cdot x)^2 \leq K_1 z_k^2 \tag{10}$$

$$|W_D| \leq \frac{K_2 \eta n_k}{\epsilon} \tag{11}$$

$$\frac{1}{|W_C|}\sum_{(x,y) \in W_C} (w \cdot x)^2 \leq K_3 z_k^2. \tag{12}$$

Assume these indeed hold. Equation (10) and the fact that $\sum_{x \in W} q(x) \geq (1 - \xi_k)|W| \geq |W|/2$ imply $\sum_{x \in W} p(x)(w \cdot x)^2 \leq 2K_1 z_k^2$. Combining this with equation (11) and that fact that the total variation distance between p and the uniform distribution over W is at most ξ, and using Cauchy-Schwartz, yields the following bound on the weighted loss of the noisy examples in W:

$$\sum_{(x,y) \in W_D} p(x)\ell(w, x, y) \leq K_2\eta/\epsilon + \xi_k + \sqrt{2K_1 K_2 \eta/\epsilon + \xi_k}\left(\frac{z_k}{\tau_k}\right). \tag{13}$$

A similar argument gives the following bound on the weighted loss over all the examples in W:

$$\sum_{(x,y) \in W} p(x)\ell(w, x, y) \leq 1 + \sqrt{2K_1}\left(\frac{z_k}{\tau_k}\right). \tag{14}$$

Next, using (12) together with Cauchy-Schwartz and the fact that $\sum_{x\in W} q(x) \geq (1-\xi_k)|W|$ and $q(x) \in [0,1]$, we can upper-bound the average loss over the clean examples:

$$\ell(w, W_C) \leq \frac{1}{|W_C|}\left(\sum_{(x,y)\in W} q(x)\ell(w,x,y)\right) + 2\xi_k + \sqrt{2\xi_k K_3}\left(\frac{z_k}{\tau_k}\right).$$

Finally, ξ_k is chosen small enough so the last two terms above are at most $\kappa/32$. Applying further manipulation together with (11) and (14) yields the first inequality in the Lemma. The second inequality follows from partitioning $\ell(w,p)$ into a sum over clean points and a sum over noisy points, and then applying (13) together with (11) and the fact that $p(x) \leq 1$ for all x. ∎

These results now can be combined to give the algorithm's guarantee for the $\mathcal{A}_\eta^{mal}$ noise model.

Theorem 27 (Analog of Theorem 4.1 in Awasthi et al. (2017) for the $\mathcal{A}_\eta^{mal}$ noise model) *Let distribution $\mathcal{D}$ over R^d be isotropic log-concave. Let w^* be the (unit length) target weight vector. There are settings of the parameters of Algorithm 2, and positive constants M, C and ϵ_0, such that in the $\mathcal{A}_\eta^{mal}$ noise model, for all $\epsilon < \epsilon_0$, for any $\delta > 0$, if $\eta < C\epsilon$, $n_k = \text{poly}(d, M^k, \log(1/\delta))$, and $\theta(w_0, w^*) < \pi/2$, then after $s = O(\log(1/\epsilon))$ iterations, the algorithm finds w_s satisfying $\text{err}(w_s) \leq \epsilon$ with probability $\geq 1 - \delta$.*

Proof (Sketch) First, by a uniform convergence analysis, with high probability for all $w \in B(w_{k-1}, r_k)$ we have $|\text{err}_{D_{w_{k-1},b_{k-1}}}(w) - \ell(w, W_C)| \leq \kappa/16$. Assume this holds, as well as equations (8) and (9) from Lemma 26. Next, we analyze as in Awasthi et al. (2017) but with one key change: Algorithm 2 deterministically computes $\ell(v_k, p)$ in step 2(b) rather than using a random labeled sample from p as in Awasthi et al. (2017); thus, we do not need to relate the sampled loss and true expected loss with respect to p. This change in the algorithm was crucial because in the $\mathcal{A}_\eta^{mal}$ model, the adversary can effectively choose its corruptions *after* observing any internal randomness of the algorithm. Specifically, we now have:

$$
\begin{aligned}
\text{err}_{D_{w_{k-1},b_{k-1}}}(w_k) &\leq \ell(v_k, W_C) + \kappa/16 \\
&\leq \ell(v_k, p) + \frac{C_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/8 \\
&\leq \ell(w^*, p) + \frac{C_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/8 \quad \text{(since } w^* \in B(w_{k-1}, r_k)) \\
&\leq 2\ell(w^*, W_C) + \frac{C_2\eta}{\epsilon} + C_3\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k}{\tau_k} + \frac{C_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/4 \\
&\leq \kappa/3 + \frac{C_2\eta}{\epsilon} + C_3\sqrt{\frac{\eta}{\epsilon}} \times \frac{z_k}{\tau_k} + \frac{C_1\eta}{\epsilon}\left(1 + \frac{z_k}{\tau_k}\right) + \kappa/2.
\end{aligned}
$$

Using the fact that $z_k = O(\tau_k)$, an $\Omega(\epsilon)$ bound on η is sufficient so that $\text{err}_{D_{w_{k-1},b_{k-1}}}(w_k) \leq \kappa$ with probability $(1 - \frac{\delta}{k+k^2})$.

Finally, the theorem follows from an inductive argument showing that after k iterations, with high probability we have $\text{err}_{\mathcal{D}}(w_k) \leq M^{-k}$. In particular, since $\text{err}_{\mathcal{D}}(w_{k-1}) \leq M^{-(k-1)}$ by

induction, the angle between w_{k-1} and w^* is $O(M^{k-1})$. Also, r_k is chosen to be proportional to $M^{-(k-1)}$ so the angle between v_k and w_{k-1} is also $O(M^{k-1})$. M is chosen a sufficiently large constant so that by properties of isotropic log-concave distributions (Theorem 4 of Balcan and Long (2013)), w_k has error at most $M^{-k}/4$ with respect to w_{k-1} outside the band $|w_{k-1} \cdot x| \leq b_{k-1}$, and w_{k-1} has error at most $M^{-k}/4$ with respect to w^* outside the band. So, w_k has error at most $M^{-k}/2$ with respect to w^* outside the band. From above, we have that $\mathrm{err}_{D_{w_{k-1}, b_{k-1}}}(w_k) \leq \kappa$. Combining this with a bound on the probability mass in the band gives error at most $M^{-k}/2$ inside the band. So, $\mathrm{err}_{\mathcal{D}}(w_k) \leq M^{-k}$ as desired, proving the theorem. $\blacksquare$

Appendix D. Lower bound on robust reliability against $\mathcal{A}_\eta^{mal}$

We also have a near-matching lower bound on robust reliability in the $\mathcal{A}_\eta^{mal}$ model. This lower bound states that points given robust-reliability level η must be in $\mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, \frac{\eta}{1-\eta} - \epsilon))$ with a high probability, which differs from our upper bound above by a constant factor in the agreement radius.

Theorem 28 *Let $\mathcal{L}$ be a $(1-\gamma)$-probably robustly-reliable learner against $\mathcal{A}_\eta^{mal}$ for hypothesis class $\mathcal{H}$ under marginal $\mathcal{D}_X$. For any $h^* \in \mathcal{H}$, given a large enough sample size $m = |S| \geq \frac{c}{\epsilon^2} \ln \frac{1}{\delta}$ and any $\eta \leq \frac{1}{2}$, we have*

$$\mathbb{E}_{S \sim \mathcal{D}^m, v \sim Bernoulli(\eta)^m} RobC_{mal}^{\mathcal{L}}(D, \eta, S, v)] \leq \Pr \left[\mathsf{Agree} \left(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}} \left(h^*, \frac{\eta}{1-\eta} - \epsilon \right) \right) \right] + 2\gamma + 2\delta,$$

where c is an absolute constant and $\mathcal{D}$ is the distribution with marginal $\mathcal{D}_X$ consistent with h^*.

Proof Let $\mathcal{L}$ be a $(1-\gamma)$-probably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$ and let $h^* \in \mathcal{H}$. Let $x \notin \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, \frac{\eta}{1-\eta} - \epsilon))$. It is sufficient to prove that $\Pr_{S \sim \mathcal{D}^m, v \sim Bernoulli(\eta)^m}[x \in RR^{\mathcal{L}}(S, h^*, \eta, v)] \leq 2\gamma + 2\delta$.

Select some $h' \in \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, \frac{\eta}{1-\eta} - \epsilon)$ such that $h'(x) \neq h^*(x)$; such an h' exists by definition of the disagreement region. Let $S \sim \mathcal{D}^m$ and define $S' = \{(x, h'(x)) \mid (x, h^*(x)) \in S\}$. Note that $S' \sim \mathcal{D}'^m$, where $\mathcal{D}'$ is a data distribution with the same marginal as $\mathcal{D}$ but consistent with h'. Let $\Delta_S = \{(x, y) \in S \mid h'(x) \neq h^*(x)\}$ denote the disagreement of h^*, h' on sample S and $S^v = \{(x_i, y_i) \in S \mid v_i = 1\}$ be the points with indicator v set to 1 which the malicious adversary gets to corrupt.

We now consider four bad events of total probability at most $2\gamma + 2\delta$: (A) $\mathcal{L}$ is not robustly-reliable for all datasets in $\mathcal{A}^{mal}(S, v)$, (B) $\mathcal{L}$ is not robustly-reliable for all datasets in $\mathcal{A}^{mal}(S', v)$, (C) $d(S, S^v) < \eta - \frac{\epsilon}{3}$, and (D) $|\Delta_S \setminus S^v| > (\eta - \frac{\epsilon}{3})|S|$. Indeed events (A) and (B) occur with probability at most γ each since $\mathcal{L}$ is given to be a $(1-\gamma)$-probably robustly-reliable learner. An application of Hoeffding's inequality implies (C) occurs with probability at most δ for sufficiently large $m \geq \frac{c}{\epsilon^2} \ln \frac{1}{\delta}$ for some absolute constant c since $v \sim Bernoulli(\eta)^m$. Finally, observe that the event $(x, y) \in \Delta_S \setminus S^v$ for any (x, y) in the sample S is a Bernoulli event with probability of occurrence at most $\left(\frac{\eta}{1-\eta} - \epsilon \right)(1-\eta) = \eta - (1-\eta)\epsilon \leq \eta - \frac{\epsilon}{2}$, since the events $(x, y) \in \Delta_S$ and $(x, y) \notin S^v$ are independent and $\eta \leq \frac{1}{2}$. Another application of Hoeffding's inequality implies a δ-probability bound on (D).

We claim that if none of these bad events occur, then $x \notin \mathrm{RR}^{\mathcal{L}}_{mal}(S, h^*, \eta, v)$. Assume for contradiction that none of the bad events occur and $x \in \mathrm{RR}^{\mathcal{L}}_{mal}(S, h^*, \eta, v)$. We will now proceed to describe an adversarial corruption $\tilde{S}^v_{h', h^*}$ of S which guarantees at least one of the bad event occurs if x is in the stipulated η-robustly-reliable region. We select a uniformly random sequence of $|\Delta_S \setminus S^v|$ corruptible points from S^v and replace them $(x, 1 - y)$ for each $(x, y) \in \Delta_S \setminus S^v$. We substitute the remaining points in S^v by $(x, h^*(x))$ for some fixed $x \in \mathsf{Agree}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, \frac{\eta}{1-\eta} - \epsilon))$. We specify this substitution to ensure that all remaining points in the disagreement region occur in pairs with opposite labels, and we do not reveal the corruptible points. The first claim follows by noting that $|\Delta_S \setminus S^v| \le |S|(\eta - \frac{\epsilon}{3}) \le |S^v|$ since bad events (C) and (D) did not occur. Clearly $\tilde{S}^v_{h', h^*} \in \mathcal{A}^{mal}(S, v)$, the above observation further implies that $\tilde{S}^v_{h', h^*} \in \mathcal{A}^{mal}(S', v)$. Now, since $x \in \mathrm{RR}^{\mathcal{L}}_{mal}(S, h^*, \eta, v)$ and $\tilde{S}^v_{h', h^*} \in \mathcal{A}^{mal}(S, v)$, it must be the case that $\mathcal{L}_{\tilde{S}^v_{h', h^*}}(x) = (y, \eta')$ for some $\eta' \ge \eta$. But, if $y \ne h^*(x)$ then this implies bad event (A), and if $y \ne h'(x)$ then this implies bad event (B). So, x cannot be in $\mathrm{RR}^{\mathcal{L}}(S, h^*, \eta, v)$ as desired. ∎

Appendix E. Active Learning: Detailed Results

To formalize our results on active learning, we have the following definition.

Definition 29 *A learner $\mathcal{L}$ is said to be an* **active robustly-reliable** *learner w.r.t. concept space $\mathcal{H}$ if for any training sample S', the learner has access to unlabeled examples from S' and the ability to request labels, and must output a function $\mathcal{L}_{S'} : \mathcal{X} \to \mathcal{Y} \times \mathbb{R}$. The* **label complexity** *of $\mathcal{L}$ is the number of labels the learner requests before outputting $\mathcal{L}_{S'}$.*

An active robustly-reliable learner $\mathcal{L}$ is said to be **strongly robustly-reliable** *against $\mathcal{A}_\eta$ for class $\mathcal{H}$ if for any target function $h^* \in \mathcal{H}$, for any dataset S consistent with h^*, for all $S' \in \mathcal{A}_\eta(S)$, for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \ge \eta$ then $y = h^*(x)$. The robustly-reliable correctness and robustly-reliable region, for this value η, are defined the same as for passive learning above.*

*An active robustly-reliable learner $\mathcal{L}$ is said to be $(1 - \delta)$-***strongly robustly-reliable** *against $\mathcal{A}_\eta$ for class $\mathcal{H}$ if for any target function $h^* \in \mathcal{H}$, for any dataset S consistent with h^*, for all $S' \in \mathcal{A}_\eta(S)$, with probability at least $1 - \delta$ over the internal randomness of the learner, for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \ge \eta$ then $y = h^*(x)$. The empirical η-robustly-reliable region is defined as for passive learning above.*

We will now show two results indicating that it is possible for the learner to query the labels intelligently to obtain a robustly-reliable algorithm, first in the stronger sense of strongly robustly-reliable learning, and second in the weaker sense of $(1 - \delta)$-strongly robustly-reliable learning (i.e., with high probability over the randomness of the learner). In both cases, we are also able to obtain a bound on the number of queries made by the active learner, holding with high probability over the draw of the uncorrupted dataset S (and in the $(1 - \delta)$-strongly robustly-reliable case, also the randomness of the learner).

First, let us consider the case of strongly robustly-reliable active learning. Note that this case is quite challenging, since the adversary effectively *knows* what queries the learner will make. For instance, the adversary may choose to set it up so that the learner's first ηm queries are *all* corrupted, so that the learner cannot even *find* an uncorrupted point until it has made at least ηm queries. In particular, this intuitively means that nontrivial reductions in label complexity compared to passive learning are only possible in the case of small η.

Our result (Theorem 30 below) proposes an active learning method that, for any given η, produces an *optimal* η-robustly-reliable region (i.e., matching the upper and lower bounds of Theorems 7 and 10), while making a number of label queries that, for small η, is significantly smaller than the number m of examples in S' (the corrupted sample), when the disagreement coefficient θ is small: specifically, it makes a number of queries $\tilde{O}(\theta \eta m + \theta d)$. The algorithm we propose is actually quite simple: we process the corrupted data S' in sequence, and for each new point x_i we query its label y_i iff x_i is in the region of disagreement of the set of all $h \in \mathcal{H}$ that make at most ηm mistakes among all previously-queried points. In the end, after processing all m examples in S', we use the region of agreement of this same set of classifiers as the empirical η-robustly-reliable region, predicting using their agreed-upon label for any test point x in this region. The result is stated formally as follows.

Theorem 30 *For any hypothesis class $\mathcal{H}$ and $\eta \geq 0$, there is an active learner $\mathcal{L}$ that is strongly robustly-reliable against $\mathcal{A}_\eta$ which, for any data set S consistent with some $h^* \in \mathcal{H}$,*

$$RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)).$$

Moreover, if $\mathcal{D}$ is a realizable distribution with some $h^* \in \mathcal{H}$ as target concept, then for any $\delta \in (0, 1)$, if $m = \Omega\big(\frac{\eta + \epsilon}{\epsilon^2}\big(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\big)\big)$ (where d is the VC dimension), and $S \sim \mathcal{D}^m$, with probability at least $1 - \delta$,

$$RR^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon)),$$

and with probability at least $1 - \delta$, the number of label queries made by the algorithm is at most

$$O\left(\theta\left(\eta m + d \log \frac{m}{d} + \log \frac{1}{\delta}\right) \log\left(\min\left\{m, \frac{1}{\eta}\right\}\right)\right) = \tilde{O}(\theta \eta m + \theta d),$$

where $\theta = \theta_{4\eta}$ is the disagreement coefficient of $\mathcal{H}$ with respect to h^* over the distribution $\mathcal{D}$ (Definition 4).

Proof Given any hypothesis class $\mathcal{H}$ and $\eta \geq 0$, consider the following active learning algorithm $\mathcal{L}$. The algorithm initializes sets $V_0 = \mathcal{H}$, $Q_0 = \{\}$. Given a data set $S' = \{(x_1', y_1'), \ldots, (x_m', y_m')\}$, where the algorithm initially observes only $x_1', \ldots, x_m'$, the algorithm, processes the data in sequence: for each $t \leq m$, the algorithm queries for the label y_t' iff $x_t' \in \mathrm{DIS}(V_{t-1})$; if it queries, then $Q_t = Q_{t-1} \cup \{(x_t', y_t')\}$, and otherwise $Q_t = Q_{t-1}$; in either case, $V_t = \{h \in \mathcal{H} : \sum_{(x,y) \in Q_t} \mathbb{1}[h(x) \neq y] \leq \eta m\}$. To define the predictions of the learner in the end, for any $x \in \mathsf{Agree}(V_m)$, define $\mathcal{L}_{S'}(x) = (y, \eta)$ for the label y agreed upon for x by every $h \in V_t$; for any $x \notin \mathsf{Agree}(V_m)$, we can define $\mathcal{L}_{S'}(x) = (\perp, -1)$.

Note that if $S' \in \mathcal{A}_\eta(S)$ for a data set $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ consistent with some $h^* \in \mathcal{H}$, then (since Q_t is a subsequence of S', and h^* makes at most ηm mistakes on S'), we maintain the invariant that $h^* \in V_t$ for all t. In particular, this implies $\mathcal{L}_{S'}(x)$ is well-defined since V_m is non-empty. Moreover, this implies that for any x, any label y agreed-upon by all $h \in V_m$ is necessarily $h^*(x)$. Thus, $\mathcal{L}$ satisfies the definition of a strongly robustly-reliable learner. Additionally, for any $h \in V_m$, since we also have $h^* \in V_m$, the algorithm would have queried for the label y_t' of every $x_t' \in \mathrm{DIS}(\{h, h^*\})$. Since h remains in V_m, it must be that $h(x_t') \neq y_t'$ on at most ηm of these points. Since any $h \in \mathcal{H}$ not in $\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$ has strictly greater than $2\eta m$ points in

S on which it disagrees with h^* and is incorrect, and the number of such points in S' can be smaller by at most ηm, every $h \in V_m$ is necessarily in $\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$: that is, $V_m \subseteq \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$. It follows immediately that $\mathrm{RR}^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta))$.

To address the remaining claims, let $\mathcal{D}$ be a realizable distribution with some $h^* \in \mathcal{H}$ as target concept, and fix $\epsilon, \delta \in (0, 1)$ and $\eta \geq 0$, and suppose $m = \Omega\left(\frac{\eta+\epsilon}{\epsilon^2}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$ with an appropriate numerical constant factor. Let $S \sim \mathcal{D}^m$. By classic relative uniform convergence guarantees Vapnik and Chervonenkis (1974) (see Theorem 4.4 of Vapnik (1998)), with probability at least $1 - \delta$, every $h \in \mathcal{H}$ satisfies

$$|\mathrm{err}_S(h) - \mathrm{err}_{\mathcal{D}}(h)| \leq O\left(\sqrt{\mathrm{err}_{\mathcal{D}}(h)\frac{1}{m}\left(d \log \frac{m}{d} + \log \frac{1}{\delta}\right)} + \frac{1}{m}\left(d \log \frac{m}{d} + \log \frac{1}{\delta}\right)\right),$$

so that, for an appropriate numerical constant, if $m = \Omega\left(\frac{\eta+\epsilon}{\epsilon^2}\left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta}\right)\right)$, every $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{D}}(h) > 2\eta + \epsilon$ has $\mathrm{err}_S(h) > 2\eta$. Thus, on this event, $\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta) \subseteq \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon)$. In particular, by the above analysis of the algorithm under any fixed S, we have $V_m \subseteq \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)$. Together, we have that with probability at least $1 - \delta$, $V_m \subseteq \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon)$, which immediately implies that $\mathrm{RR}^{\mathcal{L}}(S, h^*, \eta) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 2\eta + \epsilon))$.

Finally, we turn to bounding the number of label queries. For every $t \in \{1, \ldots, m\}$, define $S_t = \{(x_1, y_1), \ldots, (x_t, y_t)\}$ and $S_t' = \{(x_1', y_1'), \ldots, (x_t', y_t')\}$. Note that the number of queries equals

$$\sum_{t=1}^{m} \mathbb{1}[x_t' \in \mathrm{DIS}(V_{t-1})] \leq \eta m + \sum_{t=1}^{m} \mathbb{1}[x_t \in \mathrm{DIS}(V_{t-1})]. \tag{15}$$

Following a similar argument to the analysis of $\mathrm{RR}^{\mathcal{L}}$ above, for any $t \leq m$, since any $h \in \mathcal{H}$ not in $\mathcal{B}_{S_t}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t}\right)$ has strictly greater than $2\eta m$ points in S_t on which it disagrees with h^* and is incorrect, and the number of such points in S_t' can be smaller by at most ηm, every $h \in V_t$ is necessarily in $\mathcal{B}_{S_t}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t}\right)$: that is, $V_t \subseteq \mathcal{B}_{S_t}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t}\right)$. Together with (15), this implies the total number of queries is at most

$$\eta m + \sum_{t=1}^{m} \mathbb{1}\left[x_t \in \mathrm{DIS}\left(\mathcal{B}_{S_{t-1}}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t-1}\right)\right)\right].$$

By Bernstein's inequality for martingale difference sequences, with probability at least $1 - \delta/2$, the right hand side above is at most

$$\eta m + \log_2 \frac{2}{\delta} + 2e \sum_{t=1}^{m} \mathrm{Pr}\left(x_t \in \mathrm{DIS}\left(\mathcal{B}_{S_{t-1}}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t-1}\right)\right)\Big| S_{t-1}\right)$$

$$= \eta m + \log_2 \frac{2}{\delta} + 2e \sum_{t=1}^{m} \mathcal{D}_X\left(\mathrm{DIS}\left(\mathcal{B}_{S_{t-1}}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t-1}\right)\right)\right). \tag{16}$$

Let $\alpha_t = \frac{c}{t}\left(d \log \frac{t}{d} + \log \frac{1}{\delta}\right)$, for a numerical constant $c > 0$. Again by relative uniform convergence guarantees, together with a union bound (over all values of t), for an appropriate numerical constant $c > 0$, with probability at least $1 - \delta/2$, for every $t \leq m$ and $h \in \mathcal{H}$,

$$\mathrm{err}_{\mathcal{D}}(h) \leq 2\mathrm{err}_{S_t}(h) + \alpha_t.$$

Thus, on this event, for every $t \leq m$, $\mathcal{B}_{S_t}^{\mathcal{H}}\left(h^*, \frac{2\eta m}{t}\right) \subseteq \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}\left(h^*, \frac{4\eta m}{t} + \alpha_t\right)$. Plugging into (16), by a union bound, with probability at least $1 - \delta$, the total number of queries is at most

$$\eta m + \log_2 \frac{2}{\delta} + 2e \sum_{t=1}^m \mathcal{D}_X\left(\text{DIS}\left(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}\left(h^*, \frac{4\eta m}{t-1} + \alpha_{t-1}\right)\right)\right)$$

$$\leq (1 + 8e)\eta m + \log_2 \frac{2}{\delta} + 2e \sum_{4\eta m < t \leq m} \mathcal{D}_X\left(\text{DIS}\left(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}\left(h^*, \frac{4\eta m}{t-1} + \alpha_{t-1}\right)\right)\right)$$

$$\leq (1 + 8e)\eta m + \log_2 \frac{2}{\delta} + 2e \sum_{4\eta m < t \leq m} \theta_{4\eta}\left(\frac{4\eta m}{t-1} + \alpha_{t-1}\right)$$

$$= O\left(\theta_{4\eta}\left(\eta m + d\log\frac{m}{d} + \log\frac{1}{\delta}\right)\log\left(\min\left\{m, \frac{1}{\eta}\right\}\right)\right),$$

where the inequality on the second-to-last line is by the definition of the disagreement coefficient, and the final expression results from summing the harmonic series: $\sum_{t=a}^b \frac{1}{t} = O\left(\log\frac{b}{a}\right)$. ∎

Next, we consider the weaker case of $(1 - \delta)$-strongly robustly-reliable active learning. Note that in this case, the adversary only corrupts the data pool, but is not directly involved with the learner's queries. That is, the corruption may depend on the dataset S and the target test instance x, but it does not depend on the randomness in the algorithm used by the learner to query labels and output a hypothesis.

In this case, our result (Theorem 32 below) provides a method that yields an empirical robustly-reliable region of comparable size to the results for passive learning above (up to constant factors), using a number of queries that is significantly smaller than the sample complexity of passive learning, when η and the disagreement coefficient are small. However, we note that unlike Theorem 30 above, in this case the number of queries is *bounded*, in that it does not depend on the size m of the data set S: that is, it can be expressed purely as a function of $\eta, \epsilon, \delta, d,$ and θ. However, it is also worth noting that the number of queries in Theorem 30 is actually of the *same* order as in Theorem 32 (up to log factors) in the case that $m = \Theta\left(\frac{\eta+\epsilon}{\epsilon^2}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$, so that for the minimal sample size sufficient for the result on $\text{RR}^{\mathcal{L}}(S, h^*, \eta)$ in Theorem 30, we essentially lose nothing compared to the guarantee achievable by the weaker $(1 - \delta)$-strongly robustly-reliable learner in Theorem 32.

Our $(1 - \delta)$-strongly robustly-reliable active learner in Theorem 32 will use an agnostic active learning algorithm (in the usual learning setting, without robust-reliability requirements). We will use the algorithm proposed by Dasgupta et al. (2007). The algorithm partitions points seen so far into sets T (labels explicitly queried) and U (labels not queried). For a new point x, ERMs h_y on T consistent with $U \cup (x, y)$ for $y \in \{0, 1\}$ are learned and the label for x is requested only if the difference in errors of h_0, h_1 is small. The algorithm has the following guarantee.

Theorem 31 (Dasgupta et al. (2007)) *Let $\mathcal{H}$ denote the hypothesis space. Let $\mathcal{D}$ be a distribution such that $h^* = \text{argmin}_{h \in \mathcal{H}} \text{err}_{\mathcal{D}}(h)$ and $\text{err}_{\mathcal{D}}(h^*) = \eta$. If $\theta_{\eta+\epsilon}$ is the disagreement coefficient of $\mathcal{H}$ w.r.t. h^* over $\mathcal{D}$, then given $S \sim \mathcal{D}^m$ with $m \geq c\left(\frac{\eta+\epsilon}{\epsilon^2}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$ unlabeled examples for sufficiently large constant c, with probability at least $1 - \delta$, the algorithm queries at most $O\left(\theta_{\eta+\epsilon}\frac{(\eta+\epsilon)^2}{\epsilon^2}\left(d\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\log\frac{1}{\epsilon}\right)$ labeled examples and returns a hypothesis $h \in \mathcal{H}$ with $\text{err}_{\mathcal{D}}(h) \leq \text{err}_{\mathcal{D}}(h^*) + \epsilon$.*

In Theorem 32 we define an $(1 - \delta)$-strongly robustly-reliable active learner which requires fewer labels than the passive learner.

Theorem 32 *Let $\mathcal{H}$ be a hypothesis class, and S be a sample consistent with $h^* \in \mathcal{H}$. If $\hat{\theta}_{\eta+\epsilon}$ is the disagreement coefficient of $\mathcal{H}$ with respect to h^* over the uniform distribution $U(S)$ over S (see Definition 4), there exists a $(1 - \delta)$-strongly robustly-reliable active learner $\mathcal{L}$ based on the agnostic active learning algorithm of Dasgupta et al. (2007), which for some constant c, given any $S' \in \mathcal{A}_\eta(S)$, for sample size $m \geq c \left(\frac{\eta+\epsilon}{\epsilon^2} \left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \right)$, with probability $1 - \delta$ the learner queries $O\left(\hat{\theta}_{\eta+\epsilon} \frac{(\eta+\epsilon)^2}{\epsilon^2} \left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \log \frac{1}{\epsilon} \right)$ labeled examples and returns $\mathcal{L}_{S'}$ which satisfies, with probability at least $1 - \delta$,*

$$\widehat{RR}^{\mathcal{L}}(S', \eta) \supseteq \mathsf{Agree}(\mathcal{B}_S^{\mathcal{H}}(h^*, 5\eta + 2\epsilon)),$$

where d is the VC dimension of $\mathcal{H}$. Moreover, if $\mathcal{D}$ is a realizable distribution consistent with $h^* \in \mathcal{H}$ and $\theta_{\eta+\epsilon}$ is the disagreement coefficient of $\mathcal{H}$ with respect to h^* over $\mathcal{D}$, if $m = \Omega\left(\frac{1}{\epsilon^2} \left(d + \log \frac{1}{\delta} \right) \right)$, for any $S \sim \mathcal{D}^m$, given any $S' \in \mathcal{A}_\eta(S)$, w.p. $1 - 2\delta$, $\mathcal{L}_{S'}$ satisfies $\widehat{RR}^{\mathcal{L}}(S', \eta) \supseteq \mathsf{Agree}(\mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 5\eta + 3\epsilon))$, and queries $O\left(\theta_{\eta+\epsilon} \frac{(\eta+\epsilon)^2}{\epsilon^2} \left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \log \frac{1}{\epsilon} \right)$ labeled examples w.p. $1 - 3\delta$.

Proof The learner samples points uniformly randomly from the (corrupted) data pool S' and applies the agnostic active learning algorithm of Dasgupta et al. (2007), i.e. the agnostic active learning algorithm receives unlabeled points drawn uniformly from the data pool S' and labels are revealed for points requested by the algorithm. Let $\hat{h}$ denote the hypothesis output by the algorithm. The learner provides test point x with the largest η such that $x \in \mathsf{Agree}(\mathcal{B}_{S'}^{\mathcal{H}}(\hat{h}, 2\eta + \epsilon))$ and outputs the common y in the agreement region, and outputs $(\perp, -1)$ if $x \notin \mathsf{Agree}(\mathcal{B}_{S'}^{\mathcal{H}}(\hat{h}, 2\eta + \epsilon))$ for any $\eta \geq 0$. Note that this computation does not need the knowledge of labels and can be performed over the unlabeled pool which the learner has access to.

By Theorem 31, we have an upper bound on the number of labels requested by the agnostic active learner given by $O\left(\theta'_{\eta+\epsilon} \frac{(\eta+\epsilon)^2}{\epsilon^2} \left(d \log \frac{1}{\epsilon} + \log \frac{1}{\delta} \right) \log \frac{1}{\epsilon} \right)$ with probability at least $1 - \delta$ over the algorithm's internal randomness, where $\theta'_{\eta+\epsilon}$ is the disagreement coefficient of $\mathcal{H}$ w.r.t. h^* over $\mathcal{D}'$, with $\mathcal{D}'$ being the distribution corresponding to uniform draws from S'. We will show that $\theta'_{\eta+\epsilon} = O(\hat{\theta}_{\eta+\epsilon})$, which implies the desired bound. Let $\hat{\mathcal{D}}$ denote $U(S)$. Observe that for each $h \in \mathcal{H}$ if $d_{\mathcal{D}'}(h, h^*) \leq r$, then $d_{\hat{\mathcal{D}}}(h, h^*) = \Pr_{\hat{\mathcal{D}}}[h(x) \neq h^*(x)] = d_S(h, h^*) \leq d_{S'}(h, h^*) + \eta = \Pr_{\mathcal{D}'}[h(x) \neq h^*(x)] + \eta \leq r + \eta$. Thus, $\mathcal{B}_{\mathcal{D}'}^{\mathcal{H}}(h^*, r) \subseteq \mathcal{B}_{\hat{\mathcal{D}}}^{\mathcal{H}}(h^*, r + \eta)$ and $\mathrm{DIS}(\mathcal{B}_{\mathcal{D}'}^{\mathcal{H}}(h^*, r)) \subseteq$

$\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, r + \eta))$. Now,

$$\theta'_{\eta+\epsilon} = \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\mathcal{D}'_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}'}(h^*, r))]}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}'}(h^*, r))] + \eta}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, r + \eta))] + \eta}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, 2r))] + r}{r}$$

$$\leq 2\hat{\theta}_{\eta+\epsilon} + 1 = O(\hat{\theta}_{\eta+\epsilon})$$

The agnostic active learning algorithm returns a hypothesis $\hat{h}$ with error at most $\text{err}_{\mathcal{D}'}(\hat{h}) \leq \text{err}_{\mathcal{D}'}(h^*) + \epsilon \leq \eta + \epsilon$ on S' with failure probability δ. Since $\text{err}_{S'}(\hat{h}) \leq \eta + \epsilon$ (if the failure event does not occur) and $\text{err}_{S'}(h^*) \leq \eta$, using the triangle inequality $h^* \in \mathcal{B}^{\mathcal{H}}_{S'}(\hat{h}, 2\eta + \epsilon)$ and we can provide robust-reliability η to x if $x \in \text{Agree}(\mathcal{B}^{\mathcal{H}}_{S'}(\hat{h}, 2\eta + \epsilon))$. Notice $\mathcal{B}^{\mathcal{H}}_{S'}(\hat{h}, 2\eta + \epsilon) \subseteq \mathcal{B}^{\mathcal{H}}_{S'}(h^*, 4\eta + 2\epsilon) \subseteq \mathcal{B}^{\mathcal{H}}_{S}(h^*, 5\eta + 2\epsilon)$. Therefore the empirical η-robustly-reliable region of the learner contains $\text{Agree}(\mathcal{B}^{\mathcal{H}}_{S}(h^*, 5\eta + 2\epsilon))$ with probability $1 - \delta$ over the learner's randomness.

To establish the distributional results, we need to relate $\theta_{\eta+\epsilon}$ with $\hat{\theta}_{\eta+\epsilon}$ and the robustly-reliable agreement region over the sample to one over the distribution, both using uniform convergence bounds. By uniform convergence (Anthony and Bartlett (2009) Theorem 4.10), with probability at least $1 - \delta$, for each $h \in \mathcal{H}$ if $d_{\hat{\mathcal{D}}}(h, h^*) \leq r$, then $d_{\mathcal{D}}(h, h^*) \leq d_S(h, h^*) + \epsilon = d_{\hat{\mathcal{D}}}(h, h^*) + \epsilon \leq r + \epsilon$. Thus, on this event, $\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, r) \subseteq \mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r + \epsilon)$ and $\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, r)) \subseteq \text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r + \epsilon))$. Moreover, note that the family $\{\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r)) : r \geq 0\}$ has VC dimension 1, and therefore by uniform convergence, with probability at least $1 - \delta$, we have $\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r))] \leq \text{Pr}_{\mathcal{D}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r))] + \epsilon$ for all $r \geq 0$. Supposing both of these events, we have

$$\hat{\theta}_{\eta+\epsilon} = \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\hat{\mathcal{D}}}(h^*, r))]}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\hat{\mathcal{D}}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r + \epsilon))]}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\mathcal{D}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, r + \epsilon))] + \epsilon}{r}$$

$$\leq \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\mathcal{D}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, 2r))] + r}{r}$$

$$= 1 + 2 \sup_{r > \eta+\epsilon} \frac{\text{Pr}_{\mathcal{D}_X}[\text{DIS}(\mathcal{B}^{\mathcal{H}}_{\mathcal{D}}(h^*, 2r))]}{2r}$$

$$\leq 2\theta_{\eta+\epsilon} + 1 = O(\theta_{\eta+\epsilon}),$$

By a union bound, the above 3 events occur simultaneously, with probability at least $1 - 3\delta$. This implies the desired label complexity bound. The bound on the trusted region follows from another

application of the uniform convergence bound which implies $\mathcal{B}_S^{\mathcal{H}}(h^*, 5\eta + 2\epsilon) \subseteq \mathcal{B}_{\mathcal{D}}^{\mathcal{H}}(h^*, 5\eta + 3\epsilon)$ with probability $1 - \delta$, from which the stated result follows by a union bound. ∎

Remark 33 *Our above result is essentially a reduction of finding a $(1-\delta)$-strongly robustly-reliable active learner to a general agnostic active learner. We have used the algorithm from Dasgupta et al. (2007) in Theorem 32, but we can substitute any other agnostic active learning algorithm to get the corresponding label complexity guarantees.*

We demonstrate an application of the above theorem to the well-studied setting of learning a linear separator, with the data distributed according to the uniform distribution over the unit ball.

Remark 34 (Linear separators for uniform distribution over the unit ball) *It was shown by Hanneke (2007) that $\theta_{\eta+\epsilon} = O(\sqrt{d})$ for this setting. Theorem 32 now implies an exponential gain in the label complexity for $\epsilon \approx \eta$ and a gain by a factor of $\tilde{O}(\frac{1}{\sqrt{d}\eta})$ for $\epsilon \ll \eta$.*

Appendix F. Extension to the agnostic case

So far we have assumed that the uncorrupted samples S are realizable under our concept class $\mathcal{H}$. We will now show that our results can be extended to the non-realizable setting, i.e. $\min_{h \in \mathcal{H}} \text{err}_S(h) > 0$, with weaker but still interesting guarantees. Specifically, our algorithm might now produce an incorrect prediction (y, η) with η greater than the adversary's power, but only if *every* hypothesis in $\mathcal{H}$ with low error on the uncorrupted S would also be incorrect on that example.

We can define a ν-*tolerably robustly-reliable* learner in the non-realizable setting as the learner whose reliable predictions agree with every low error hypothesis (error at most ν) on the uncorrupted sample.

Definition 35 *A learner $\mathcal{L}$ is ν-**tolerably robustly-reliable** for sample S' w.r.t. concept space $\mathcal{H}$ if, given S', the learner outputs a function $\mathcal{L}_{S'} : \mathcal{X} \to \mathcal{Y} \times \mathbb{R}$ such that for all $x \in \mathcal{X}$ if $\mathcal{L}_{S'}(x) = (y, \eta)$ then for all $h^* \in \mathcal{H}$ such that $\text{err}_S(h^*) \leq \nu$ for some S with $S' \in \mathcal{A}_\eta(S)$, we have $y = h^*(x)$.*

*Given sample S such that there is some $h^* \in \mathcal{H}$ satisfying $\text{err}_S(h^*) \leq \nu$, the (ν, η)-**robustly-reliable region** $RR^{\mathcal{L}}(S, \nu, \eta)$ for learner $\mathcal{L}$ is the set of points $x \in \mathcal{X}$ for which given any $S' \in \mathcal{A}_\eta(S)$ we have that $\mathcal{L}_{S'}(x) = (y, \eta')$ with $\eta' \geq \eta$. More generally, for a class of adversaries $\mathcal{A}$ with budget η, $RR_{\mathcal{A}}^{\mathcal{L}}(S, \nu, \eta)$ is the set of points $x \in \mathcal{X}$ for which given any $S' \in \mathcal{A}(S)$ we have that $\mathcal{L}_{S'}(x) = (y, \eta')$ with $\eta' \geq \eta$. We also define the **empirical** (ν, η)-**robustly-reliable region** $\widehat{RR}^{\mathcal{L}}(S', \nu, \eta) = \{x \in \mathcal{X} : \mathcal{L}_{S'}(x) = (y, \eta') \text{ for some } \eta' \geq \eta\}$. So, $RR_{\mathcal{A}}^{\mathcal{L}}(S, \nu, \eta) = \cap_{S' \in \mathcal{A}(S)} \widehat{RR}^{\mathcal{L}}(S', \nu, \eta)$.*

Definition 35 describes the notion of a robustly-reliable learner for a particular (corrupted) sample. Notice that if there are multiple $h \in \mathcal{H}$ satisfying $\text{err}_S(h) \leq \nu$ and they disagree on x then it must be the case that the algorithm outputs $\eta < 0$. Notice also that setting $\nu = 0$ in the above definition yields the usual robustly-reliable learner. Similarly to Definition 3, we now extend this to robustly-reliable with high probability for an adversarially-corrupted sample drawn from a given distribution.

Definition 36 *A learner $\mathcal{L}$ is a $(1-\gamma)$-**probably ν-tolerably robustly-reliable learner** for concept space $\mathcal{H}$ under marginal $\mathcal{D}_X$ (where $\mathcal{D}$ is the distribution over examples with marginal $\mathcal{D}_X$) if with probability at least $1 - \gamma$ over the draw of $S \sim \mathcal{D}^m$, for any concept $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) \leq \nu$, for all $S' \in \mathcal{A}_\eta(S)$, and for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \geq \eta$ then $y = h^*(x)$. If $\mathcal{L}$ is a $(1 - \gamma)$-probably ν-tolerably robustly-reliable learner with $\gamma = 0$ for all marginal distributions $\mathcal{D}_X$, then we say $\mathcal{L}$ is ν-**tolerably strongly robustly-reliable** for $\mathcal{H}$. Note that a ν-tolerably strongly robustly-reliable learner is a ν-tolerably robustly-reliable learner in the sense of Definition 35 for every sample S'. Given distribution $\mathcal{D}$, the (ν, η)-robustly-reliable correctness for learner $\mathcal{L}$ for sample S is given by the probability mass of the robustly-reliable region, $RobC^{\mathcal{L}}(\mathcal{D}, \nu, \eta, S) = \Pr_{x \sim \mathcal{D}_X}[x \in RR^{\mathcal{L}}(S, \nu, \eta)]$.*

We now provide a general ν-tolerably strongly robustly-reliable learner using the notion of agreement regions (Theorem 37). Our results here generalize corresponding results from Section 3. We first present our learner and a guarantee on the learner's empirical (ν, η)-robustly-reliable region given any (possibly corrupted) dataset. Our algorithm assumes ν is given.

Theorem 37 *Let $\mathcal{H}_{\eta+\nu}(S') = \{h \in \mathcal{H} \mid \mathrm{err}_{S'}(h) \leq \eta + \nu\}$. For any hypothesis class $\mathcal{H}$, there exists a ν-tolerably strongly robustly-reliable learner $\mathcal{L}$ (Definition 36) that given S' outputs a function $\mathcal{L}_{S'}$ such that*

$$\widehat{RR}^{\mathcal{L}}(S', \nu, \eta) \supseteq \mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S')).$$

Proof Given sample S', the learner $\mathcal{L}$ outputs the function $\mathcal{L}_{S'}(x) = (y, \eta)$ where η is the largest value such that $x \in \mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S'))$, and y is the common prediction in that agreement region; if $x \notin \mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S'))$ for all $\eta \geq 0$, then $\mathcal{L}_{S'}(x) = (\perp, -1)$. This is a strongly robustly-reliable learner because if $\mathcal{L}_{S'}(x) = (y, \eta)$ and $S' \in \mathcal{A}_\eta(S)$ and $\mathrm{err}_S(h^*) \leq \nu$, then $h^* \in \mathcal{H}_{\eta+\nu}(S')$, so $y = h^*(x)$. Also, notice that by design of the algorithm, all points in $\mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S'))$ will be given robust-reliability level at least η. The learner may be implemented using an ERM oracle using knowledge of ν and the construction in Theorem 6. ∎

We now analyze the (ν, η)-robustly-reliable region for the algorithm above which we will prove is pointwise optimal over all ν-tolerably strongly robustly-reliable learners.

Theorem 38 *For any hypothesis class $\mathcal{H}$, the ν-tolerably strongly robustly-reliable learner $\mathcal{L}$ from Theorem 37 satisfies the property that for all S and for all $\eta \geq 0$ and for any h^* with $\mathrm{err}_S(h^*) \leq \nu$,*

$$RR^{\mathcal{L}}(S, \nu, \eta) \supseteq \mathsf{Agree}\left(H_{2\eta+\nu}(S)\right) \supseteq \mathsf{Agree}\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta + 2\nu)\right),$$

where $\mathcal{H}_\alpha(T) = \{h \in \mathcal{H} \mid \mathrm{err}_T(h) \leq \alpha\}$ for any $\alpha \geq 0$, where T may be a sample or a distribution. Moreover, if $S \sim \mathcal{D}^m$ for $m = O(\frac{1}{\epsilon^2}(d + \ln \frac{1}{\delta}))$ then with probability at least $1 - \delta$, $RR^{\mathcal{L}}(S, \nu, \eta) \supseteq \mathsf{Agree}(H_{2\eta+\nu+\epsilon}(\mathcal{D}))$. Here d denotes the VC dimension of $\mathcal{H}$.

Proof By Theorem 37, the empirical η-robustly-reliable region $\widehat{RR}^{\mathcal{L}}(S', \nu, \eta) \supseteq \mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S'))$ for any dataset S'. The (ν, η)-robustly-reliable region, which is the set of points given robustness level at least η for *all* $S' \in \mathcal{A}_\eta(S)$ given that there is $h^* \in \mathcal{H}$ with $\mathrm{err}_S(h^*) \leq \nu$, is therefore at least

$$\bigcap_{S' \in \mathcal{A}_\eta(S)} \mathsf{Agree}(\mathcal{H}_{\eta+\nu}(S')) = \mathsf{Agree}\left(H_{2\eta+\nu}(S)\right)$$

where the above holds because if $h \in \mathcal{H}_{\eta+\nu}(S')$ for some $S' \in \mathcal{A}_\eta(S)$ then $h \in \mathcal{H}_{2\eta+\nu}(S)$, and conversely if $h \in \mathcal{H}_{2\eta+\nu}(S)$ then $h \in \mathcal{H}_{\eta+\nu}(S')$ for some $S' \in \mathcal{A}_\eta(S)$. Further, if $h \in \mathcal{H}_{2\eta+\nu}(S)$ then $h \in \mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta + 2\nu)$ since $\mathrm{err}_S(h^*) \leq \nu$. This implies Agree $(H_{2\eta+\nu}(S)) \supseteq$ Agree $\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta + 2\nu)\right)$. Finally, by uniform convergence, if $S \sim \mathcal{D}^m$ for $m = O(\frac{1}{\epsilon^2}(d + \ln \frac{1}{\delta}))$ then with probability at least $1 - \delta$ we have $\mathrm{err}_{\mathcal{D}}(h) \leq \mathrm{err}_S(h) + \epsilon$ for all $h \in \mathcal{H}$. This implies that Agree$(H_{2\eta+\nu}(S)) \supseteq$ Agree$(H_{2\eta+\nu+\epsilon}(\mathcal{D}))$ and so $\mathrm{RR}^{\mathcal{L}}(S, \nu, \eta) \supseteq$ Agree$(H_{2\eta+\nu+\epsilon}(\mathcal{D}))$. ∎

We will now show a matching lower bound for Theorem 38. In contrast to the realizable case, our matching bounds here are in terms of agreement regions of hypotheses with low error on the sample (more precisely Agree $(H_{2\eta+\nu}(S))$), instead of balls around a fixed hypothesis in $\mathcal{H}$. In Theorems 7 and 10, the size of the robustly-reliable region is Agree $\left(\mathcal{B}_S^{\mathcal{H}}(h^*, 2\eta)\right)$, which in the realizable case is the same as Agree $(H_{2\eta}(S))$. This also implies we do not lose too much of the robustly-reliable region in the agnostic case.

Theorem 39 *Let $\mathcal{L}$ be a ν-tolerably strongly robustly-reliable learner for hypothesis class $\mathcal{H}$. Then for any sample S, any point in the (ν, η)-robustly-reliable region must lie in the agreement region of $\mathcal{H}_{2\eta+\nu}(S) = \{h \in \mathcal{H} \mid \mathrm{err}_S(h) \leq 2\eta + \nu\}$. That is,*

$$RR^{\mathcal{L}}(S, \nu, \eta) \subseteq \mathsf{Agree}\left(\mathcal{H}_{2\eta+\nu}(S)\right).$$

Proof Let $x \notin \mathsf{Agree}(\mathcal{H}_{2\eta+\nu}(S))$. We will show that x cannot be in the (ν, η)-robustly-reliable region. First, since $x \notin \mathsf{Agree}(\mathcal{H}_{2\eta+\nu}(S))$, there must exist some $h_1, h_2 \in \mathcal{H}_{2\eta+\nu}(S)$ such that $h_1(x) \neq h_2(x)$. Since we are interested in the (ν, η)-robustly-reliable region, suppose we have hypothesis $h^* \in \mathcal{H}$ such that $\mathrm{err}_S(h^*) \leq \nu$. Now either $h_1(x) \neq h^*(x)$ or $h_2(x) \neq h^*(x)$. Assume WLOG that $h_1(x) \neq h^*(x)$. Let $S_e^* = \{(x, y) \in S \mid h^*(x) \neq y\}$ denote the points in the uncorrupted sample where h^* is incorrect. Further let $S_e = \{(x, y) \in S \mid h_1(x) \neq y\}$ be the points where h_1 is incorrect on S. Finally let $\tilde{S}$ be a fixed subset of $S_e \setminus S_e^*$ of size $\min\{2\eta m, |S_e \setminus S_e^*|\}$ and let $S' \subset \tilde{S}$ with $|S'| = |\tilde{S}|/2$.

We will now construct two sets S_1 and $S_{\mathcal{A}}$. S_1 will be such that $\mathrm{err}_{S_1}(h_1) \leq \nu$ and $S_{\mathcal{A}}$ will satisfy $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$ as well as $S_{\mathcal{A}} \in \mathcal{A}_\eta(S_1)$. For any $U \subseteq T$, let $\mathrm{FLIP}(T, U) = \{(x, 1 - y) \mid (x, y) \in U\} \cup T \setminus U$ denote a sample T with flipped labels for subset U. We define $S_1 = \mathrm{FLIP}(S, \tilde{S})$ and $S_{\mathcal{A}} = \mathrm{FLIP}(S, S')$. To show $\mathrm{err}_{S_1}(h_1) \leq \nu$, note that we have one of two cases. Either $|S_e \setminus S_e^*| \leq 2\eta m$, in which case $\tilde{S} = S_e \setminus S_e^*$ and h_1 will we correct on these in S_1. So h_1 is incorrect on at most $|S_e^*|$ points on S_1 and $\mathrm{err}_{S_1}(h_1) \leq \nu$. The other case is $|S_e \setminus S_e^*| > 2\eta$, which implies $\tilde{S} \geq 2\eta m$. Since $\tilde{S} \subseteq S_e$, we have that $\mathrm{err}_{S_1}(h_1) \leq \mathrm{err}_S(h_1) - 2\eta \leq \nu$. Finally since $|\tilde{S}| \leq 2\eta m$ and $|S'| = |\tilde{S}|/2$, we have $d(S, S_1) \leq 2\eta$ and $d(S, S_{\mathcal{A}}) \leq \eta$. Also, S_1 and $S_{\mathcal{A}}$ differ on points corresponding to $\tilde{S} \setminus S'$ with $|\tilde{S} \setminus S'| = |\tilde{S}|/2 \leq \eta m$. Thus, $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$ as well as $S_{\mathcal{A}} \in \mathcal{A}_\eta(S_1)$.

Now notice that sample S has $\mathrm{err}_S(h^*) \leq \nu$ and $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$. Also S has $\mathrm{err}_{S_1}(h_1) \leq \nu$ and $S_{\mathcal{A}} \in \mathcal{A}_\eta(S_1)$. Now, assume for contradiction that $x \in \mathrm{RR}^{\mathcal{L}}(S, \nu, \eta)$. This means that $\mathcal{L}_{S_{\mathcal{A}}}(x) = (y, \eta')$ for some $\eta' \geq \eta$. However, if $y \neq h^*(x)$, the learner is incorrectly confident for (true) dataset S since $S_{\mathcal{A}} \in \mathcal{A}_\eta(S)$. Similarly, if $y = h^*(x)$, the learner is incorrectly confident for sample S_1 since $h_1(x) \neq h^*(x)$. Thus, $\mathcal{L}$ is not a ν-tolerably strongly robustly-reliable learner and we have a contradiction. ∎

We can show that any 2ν-tolerably strongly robustly-reliable learner for the concept space can be used to give a $(1-\gamma)$-probably ν-tolerably robustly-reliable learner over a distribution $\mathcal{D}$ where the best hypothesis h^* in $\mathcal{H}$ has error $\mathrm{err}_\mathcal{D}(h^*) = \nu$.

Theorem 40 *Let $\mathcal{H}$ be the concept space, and $\mathcal{D}$ be a distribution such that $\min_{h \in \mathcal{H}} \mathrm{err}_\mathcal{D}(h) = \nu^*$, and let $\mathcal{L}$ be a $2\nu^*$-tolerably robustly-reliable learner for $\mathcal{H}$ for any sample S'. If $m \geq \frac{c}{\nu^{*2}}(d + \ln\frac{1}{\gamma})$, then $\mathcal{L}$ is also $(1-\gamma)$-probably ν^*-tolerably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$.*

Proof Let $\mathcal{D}'$ be a distribution with the same marginal $\mathcal{D}_X$ as $\mathcal{D}$. Let $S \sim \mathcal{D}'^m$. By uniform convergence (Anthony and Bartlett (2009) Theorem 4.10), for each $h' \in \mathcal{H}_{\nu^*}(\mathcal{D}') = \{h \in \mathcal{H} \mid \mathrm{err}_{\mathcal{D}'}(h) \leq \nu^*\}$, we have that $\mathrm{err}_S(h') \leq 2\nu^*$ with probability at least $1-\gamma$ over the draw of S. For each such S, for all $S' \in \mathcal{A}_\eta(S)$, and for all $x \in \mathcal{X}$, if $\mathcal{L}_{S'}(x) = (y, \eta')$ for $\eta' \geq \eta$ then $y = h(x)$ for any $h \in \mathcal{H}$ with $\mathrm{err}_S(h) \leq 2\nu^*$ since $\mathcal{L}$ is $2\nu^*$-tolerably robustly-reliable for any sample S'. In particular, the prediction agrees with each $h' \in \mathcal{H}_{\nu^*}(\mathcal{D}')$. Therefore $\mathcal{L}$ is a $(1-\gamma)$-probably ν^*-tolerably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$. $\blacksquare$

We can use the above reduction, together with our ν-tolerably strongly robustly-reliable learner, to give a probably tolerably robustly-reliable learner for any distribution $\mathcal{D}$ along with guarantees about its distribution-averaged robust reliability region.

Theorem 41 *Let $\mathcal{H}$ be the concept space, and $\mathcal{D}$ be a distribution such that $\min_{h \in \mathcal{H}} \mathrm{err}_\mathcal{D}(h) = \nu^*$. If $m \geq \frac{c}{\nu^{*2}}(d + \ln\frac{1}{\gamma})$, then there is a $(1-\gamma)$-probably ν^*-tolerably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$, with $(2\nu^*, \eta)$-robustly-reliable correctness that satisfies*

$$\mathbb{E}_{S \sim \mathcal{D}^m}[RobC^\mathcal{L}(\mathcal{D}, 2\nu^*, \eta, S)] \geq (1 - 2\gamma) \Pr_{x \sim \mathcal{D}_X}[\mathsf{Agree}\,(H_{2\eta+3\nu^*}(\mathcal{D}))],$$

where d is the VC-dimension of $\mathcal{H}$, and c is an absolute constant.

Proof Let $h^* \in \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{err}_\mathcal{D}(h)$. Let $S \sim \mathcal{D}^m$ for some $m \geq \frac{c}{\nu^{*2}}(d + \ln\frac{1}{\gamma})$. By uniform convergence (Anthony and Bartlett (2009) Theorem 4.10), with probabililily at least $1 - \gamma$ over the draw of S, we have that $\mathrm{err}_S(h^*) \leq \mathrm{err}_\mathcal{D}(h^*) + \nu^* \leq 2\nu^*$. By Theorem 38, there exists a $2\nu^*$-tolerably strongly robustly-reliable learner $\mathcal{L}$ such that $\mathrm{RR}^\mathcal{L}(S, 2\nu^*, \eta) \supseteq \mathsf{Agree}\,(H_{2\eta+2\nu^*}(S))$. Further, by Theorem 40, $\mathcal{L}$ is $(1-\gamma)$-probably ν^*-tolerably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$.

Finally, let's consider the η-robustly-reliable correctness of $\mathcal{L}$. By another application of uniform convergence, with probability at least $1 - \gamma$, $\mathsf{Agree}\,(H_{2\eta+2\nu^*}(S)) \supseteq \mathsf{Agree}\,(H_{2\eta+3\nu^*}(\mathcal{D}))$. As shown above, with probability $1 - \gamma$, we also have $\mathrm{RR}^\mathcal{L}(S, 2\nu^*, \eta) \supseteq \mathsf{Agree}\,(H_{2\eta+2\nu^*}(S))$. By a union bound over the failure probabilities, with probability at least $1 - 2\gamma$ over the draw of S, we have $\mathrm{RR}^\mathcal{L}(S, 2\nu^*, \eta) \supseteq \mathsf{Agree}\,(H_{2\eta+3\nu^*}(\mathcal{D}))$. Finally, by the definition of (ν, η)-robustly-reliable correctness, we have that

$$\mathbb{E}_{S \sim \mathcal{D}^m}[RobC^\mathcal{L}(\mathcal{D}, 2\nu^*, \eta, S)] = \mathbb{E}_{S \sim \mathcal{D}^m}\left[\Pr_{x \sim \mathcal{D}_X}[\mathrm{RR}^\mathcal{L}(S, 2\nu^*, \eta)]\right]$$
$$\geq (1 - 2\gamma) \Pr_{x \sim \mathcal{D}_X}[\mathsf{Agree}\,(H_{2\eta+3\nu^*}(\mathcal{D}))].$$

$\blacksquare$

Finally we have the following lower bound on the robustly-reliable correctness for any $(1-\gamma)$-probably ν-tolerably robustly-reliable learner. The key idea in establishing this lower bound is the creation of two distributions with the same marginal but nearly-consistent with two hypotheses which are close in error.

Theorem 42 *Let $\mathcal{L}$ be a $(1-\gamma)$-probably ν-tolerably robustly-reliable learner for hypothesis class $\mathcal{H}$ under marginal $\mathcal{D}_X$. Given a large enough sample size $m = |S| \geq \frac{c}{\epsilon^2}\left(d + \ln\frac{1}{\delta}\right)$, we have*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[RobC^{\mathcal{L}}(\mathcal{D},\nu+\epsilon,\eta,S)] \leq \Pr[\mathsf{Agree}(\mathcal{H}_{2\eta+\nu-\epsilon}(\mathcal{D}))] + 2\gamma + 3\delta,$$

where c is an absolute constant and $\mathcal{D}$ is the distribution with marginal $\mathcal{D}_X$ consistent with h^*.

Proof Let $\mathcal{L}$ be a $(1-\gamma)$-probably ν-tolerably robustly-reliable learner for $\mathcal{H}$ under marginal $\mathcal{D}_X$ and let $h^* \in \mathcal{H}$ such that $\mathrm{err}_{\mathcal{D}}(h^*) \leq \nu$. Let $x \notin \mathsf{Agree}(\mathcal{H}_{2\eta+\nu-\epsilon}(\mathcal{D}))$. To prove the theorem, it suffices to prove that $\Pr_{S\sim\mathcal{D}^m}[x \in \mathrm{RR}^{\mathcal{L}}(S,\nu+\epsilon,\eta)] \leq 2\gamma + 3\delta$.

By definition of the disagreement region, we can select some $h_1, h_2 \in \mathcal{H}_{2\eta+\nu-\epsilon}(\mathcal{D})$ such that $h_1(x) \neq h_2(x)$. Assume WLOG that $h_1(x) \neq h^*(x)$. Let $E^* = \{(x,y) \mid h^*(x) \neq y\}$ and $E_1 = \{(x,y) \mid h_1(x) \neq y\}$. Note that $\Pr_{\mathcal{D}}[E^*] \leq \nu$ and $\Pr_{\mathcal{D}}[E_1] \leq 2\eta + \nu - \epsilon$. Let $\mathcal{D}'$ be a distribution which modifies $\mathcal{D}$ by re-assigning a probability mass of $p = \min\{2\eta-\epsilon, \Pr_{\mathcal{D}}[E_1\setminus E^*]\}$ from points $(x,y) \in E_1 \setminus E^*$ to corresponding points with flipped labels $(x, 1-y)$. Formally, using Corollary 44, there are distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $\mathcal{D} = p\mathcal{D}_1 + (1-p)\mathcal{D}_2$ with $\mathrm{supp}(\mathcal{D}_1) \subseteq E_1 \setminus E^*$, and we define $\mathcal{D}' = p\mathcal{D}_1' + (1-p)\mathcal{D}_2$ where $\mathcal{D}_1'$ is the distribution corresponding to drawing from $\mathcal{D}_1$ but with flipped labels. Notice that $\mathcal{D}'$ has the same marginal $\mathcal{D}_X$ distribution as $\mathcal{D}$. Moreover, $\mathrm{err}_{\mathcal{D}'}(h_1) \leq \nu$. Indeed, if $\Pr_{\mathcal{D}}[E_1 \setminus E^*] \leq 2\eta - \epsilon$, we have $\mathrm{err}_{\mathcal{D}'}(h_1) \leq \Pr_{\mathcal{D}'}[E^*] = \Pr_{\mathcal{D}}[E^*] \leq \nu$. Else, observe that $\mathrm{err}_{\mathcal{D}'}(h_1) \leq \Pr_{\mathcal{D}'}[E_1] \leq \Pr_{\mathcal{D}}[E_1] - (2\eta - \epsilon) \leq \nu$.

Drawing from $\mathcal{D}$ is equivalent to first flipping a random biased coin, with probability p drawing according to $\mathcal{D}_1$ (call these points 'red'), and otherwise drawing from $\mathcal{D}_2$. Let $S \sim \mathcal{D}^m$ and S' be the dataset obtained by flipping labels of all the 'red' points in S. Notice that $S' \sim \mathcal{D}'^m$. We now consider three bad events of total probability at most $2\gamma + 3\delta$: (A) $\mathcal{L}$ is not $(\nu+\epsilon)$-tolerably robustly-reliable for all datasets in $\mathcal{A}_\eta(S)$, (B) $\mathcal{L}$ is not $(\nu+\epsilon)$-tolerably robustly-reliable for all datasets in $\mathcal{A}_\eta(S')$, and (C) $d(S,S') > 2\eta$. Indeed events (A) and (B) occur with probability at most $\gamma + \delta$ each since $\mathcal{L}$ is given to be a $(1-\gamma)$-probably ν-tolerably robustly-reliable learner for $\mathcal{H}$ under $\mathcal{D}_X$, and by uniform convergence (Anthony and Bartlett (2009) Theorem 4.10) for each h with $\mathrm{err}_{\mathcal{D}}(h) \leq \nu$ we have $\mathrm{err}_S(h) \leq \nu + \epsilon$ with probability at least $1 - \delta$. Finally, a point drawn according to $\mathcal{D}$ is red with probability $p \leq 2\eta - \epsilon$. By an application of the Hoeffding's inequality, there are at most 2η red points in S with probability at least $1 - \delta$, i.e. (C) occurs with probability at most δ.

We claim that if none of these bad events occur, then $x \notin \mathrm{RR}^{\mathcal{L}}(S,\nu+\epsilon,\eta)$. To prove this, assume for contradiction that none of the bad events occur and $x \in \mathrm{RR}^{\mathcal{L}}(S,\nu+\epsilon,\eta)$. Let $\tilde{S}_{h_1,h^*}$ denote a relabeling of S such that exactly half the red points are labeled according to h^* (the remaining half using h_1). Note that since bad event (C) did not occur, this means that $\tilde{S}_{h_1,h^*} \in \mathcal{A}_\eta(S)$ and $\tilde{S}_{h_1,h^*} \in \mathcal{A}_\eta(S')$, Now, since $x \in \mathrm{RR}^{\mathcal{L}}(S,\nu+\epsilon,\eta)$ and $\tilde{S}_{h_1,h^*} \in \mathcal{A}_\eta(S)$, it must be the case that $\mathcal{L}_{\tilde{S}_{h_1,h^*}}(x) = (y,\eta')$ for some $\eta' \geq \eta$. But, if $y \neq h^*(x)$ then this implies bad event (A), and if $y \neq h'(x)$ then this implies bad event (B). So, x cannot be in $\mathrm{RR}^{\mathcal{L}}(S,\nu+\epsilon,\eta)$ as desired. $\blacksquare$

Lemma 43 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $A \in \mathcal{F}$ and $\alpha \leq P(A)$. Then we have probability spaces $(A, 2^A \cap \mathcal{F}, P_1)$ and $(\Omega, \mathcal{F}, P_2)$ such that for any $S \in \mathcal{F}$, $P(S) = \alpha P_1(A \cap S) + (1 - \alpha) P_2(S)$.*

Proof If $P(A) = 0$, we have $\alpha = 0$ and the result holds for $P_2 = P$ and any P_1. Also, if $\alpha = 1$, we have $P(A) = 1$ (or, $P(\Omega \setminus A) = 0$) and $P_1(T) = P(T)$ for any $T \in 2^A \cap \mathcal{F}$ defines a probability space over A. Further, for any $S \in \mathcal{F}$, $P_1(A \cap S) = P(A \cap S) = P(S) - P((\Omega \setminus A) \cap S) = P(S)$ and the desired result holds for any P_2.

Assume $P(A) > 0$ and $\alpha < 1$. Define $P_1(T) = P(T)/P(A)$ for any $T \in 2^A \cap \mathcal{F}$. Notice $(A, 2^A \cap \mathcal{F}, P_1)$ is a probability space. In particular, $P_1(A) = 1$ and $0 \leq P_1(T) \leq 1$ for any $T \in 2^A \cap \mathcal{F}$. For any $S \in \mathcal{F}$, define

$$P_2(S) = \frac{1}{1 - \alpha} P(S) - \frac{\alpha}{1 - \alpha} P_1(S \cap A).$$

We have $P_2(\Omega) = \frac{1}{1-\alpha} P(\Omega) - \frac{\alpha}{1-\alpha} \frac{P(\Omega \cap A)}{P(A)} = 1$. Also for any $S \in \mathcal{F}$,

$$\begin{aligned} P_2(S) &= \frac{1}{1 - \alpha} P(S) - \frac{\alpha}{1 - \alpha} \frac{P(S \cap A)}{P(A)} \\ &\geq \frac{1}{1 - \alpha} P(S) - \frac{1}{1 - \alpha} P(S \cap A) \geq 0, \end{aligned}$$

since $P(A) \geq \alpha$ and $P(S) \geq P(S \cap A)$. Also,

$$\begin{aligned} P_2(S) &= \frac{1}{1 - \alpha} \left(P(S) - \alpha \frac{P(S \cap A)}{P(A)} \right) \\ &= \frac{1}{1 - \alpha} \left(1 - P(\Omega \setminus S) - \alpha \frac{P(A) - P(A \setminus (S \cap A))}{P(A)} \right) \\ &= 1 - \frac{1}{1 - \alpha} \left(P(\Omega \setminus S) - \alpha \frac{P(A \setminus S)}{P(A)} \right) \\ &\leq 1 - \frac{1}{1 - \alpha} \left(P(\Omega \setminus S) - P(A \setminus S) \right) \leq 1. \end{aligned}$$

σ-additivity of P_2 follows from that of P and P_1. Thus, $(\Omega, \mathcal{F}, P_2)$ is a probability space and the desired relation holds. ∎

We have the following corollary to Lemma 43.

Corollary 44 *Let $\mathcal{D}$ be a distribution over set S. Let $\alpha \leq \Pr_{\mathcal{D}}[A]$ for some $A \subseteq S$. Then there exist distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, with $supp(\mathcal{D}_1) \subseteq A$ and $supp(\mathcal{D}_2) \subseteq S$, such that $\mathcal{D} = \alpha \mathcal{D}_1 + (1 - \alpha) \mathcal{D}_2$.*