Proofs as Explanations: Short Certificates for Reliable Predictions

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

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Chirag Pabbaraju CPABBARA@STANFORD.EDU
Stanford University

Donya Saless DONYA@TTIC.EDU
Toyota Technological Institute at Chicago [*]

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider a model for explainable AI in which an explanation for a prediction $h(x) = y$ consists of a subset S' of the training data (if it exists) such that *all* classifiers $h' \in \mathcal{H}$ that make at most b mistakes on S' predict $h'(x) = y$. Such a set S' serves as a *proof* that x indeed has label y under the assumption that (1) the true target function $h^\star$ belongs to $\mathcal{H}$, and (2) the set S contains at most b noisy or corrupted points. For example, if $b = 0$ and $\mathcal{H}$ is the family of linear classifiers in $\mathbb{R}^d$, and if x lies inside the convex hull of the positive data points in S (and therefore every consistent linear classifier labels x as positive), then Carathéodory's theorem states that x in fact lies inside the convex hull of $d + 1$ of those points. So, a set S' of size $d + 1$ could be released as an explanation for a positive prediction, and would serve as a short proof of correctness of the prediction under the assumption of perfect realizability.

In this work, we consider this problem more generally, for general hypothesis classes $\mathcal{H}$ and general values $b \geq 0$. We define the notion of the *robust hollow star number* of $\mathcal{H}$ (which generalizes the standard hollow star number), and show that it precisely characterizes the worst-case size of the smallest certificate achievable, and analyze its size for natural classes. We also consider worst-case distributional bounds on certificate size, as well as *distribution-dependent* bounds that we show tightly control the sample size needed to get a certificate for any given test example. In particular, we define a notion of the *certificate coefficient* ε_x of an example x with respect to a data distribution $\mathcal{D}$ and target function $h^\star$, and prove matching upper and lower bounds on sample size as a function of ε_x, b, and the VC dimension d of $\mathcal{H}$.

Keywords: Explainable AI, sample complexity, hollow star number, reliable learning

1. Introduction

There has been substantial recent interest in *explainable* AI, (Ali et al., 2023; Dwivedi et al., 2023; Doshi-Velez and Kim, 2017; Ribeiro et al., 2016; Miller, 2019). For example, in a medical setting, if a classifier $h \in \mathcal{H}$ trained on some large dataset S predicts that patient x should get treatment y, the patient's doctor may want an explanation of why. Much of the work in explainable machine learning has focused on decision-tree models, or identifying the most salient features for the prediction made (Craven and Shavlik, 1995; Breiman and Shang, 1996; Zhou and Hooker, 2016). In this work, we consider an alternative approach: outputting a subset S' of the training set S such that *all* classifiers

[*] Authors are listed alphabetically.

© 2025 A. Blum, S. Hanneke, C. Pabbaraju & D. Saless.

$h' \in \mathcal{H}$ that agree with S' (or that make at most b mistakes on S') predict $h'(x) = y$, if such an S' exists. Such a set S' would serve as a *proof* that x indeed has label y under the assumption that (1) the true target function $h^\star$ belongs to $\mathcal{H}$, and (2) the set S contains at most b noisy or corrupted points. For example, if $b = 0$ and $\mathcal{H}$ is the family of linear classifiers in $\mathbb{R}^d$, and if x lies inside the convex hull of the positive data points in S (and hence every consistent linear classifier labels x as positive), then Carathéodory's theorem states that x in fact lies inside the convex hull of $d + 1$ of those points. So, a set S' of size $d + 1$ could be released as an explanation for a positive prediction, and a proof of its correctness under the assumption of realizability.

We aim to consider such explanations for general families $\mathcal{H}$ and general values of b. Our work is inspired by Balcan et al. (2022) who propose the notion of *robustly reliable* classifiers that, given an example x, output both a label y and a value b with the guarantee that any $h' \in \mathcal{H}$ with $h'(x) \neq y$ makes strictly more than b mistakes on S (where $b < 0$ if x is not in the agreement region of the version space). Our work can be viewed as investigating the shortest *proof* that can be provided for such a guarantee.

1.1. Main Contributions

Our main contributions are the following:

1. We formalize the notion of a robust certificate: a subset S' of the training data that serves as a proof that a given example x must have label y if the target function belongs to a given class $\mathcal{H}$ and the training set has at most b noisy or corrupted points. To analyze this, we define the notion of the *robust hollow star number* of $\mathcal{H}$, which generalizes the standard hollow star number (Bousquet et al., 2020), and show that it precisely characterizes the worst-case size of the smallest certificate achievable for a class $\mathcal{H}$, and analyze its size for natural classes.

2. We examine worst-case distributional bounds on certificate size, showing that in this case, one can achieve tight bounds on certificates achievable from a finite sample in terms of the (standard) hollow star number of Bousquet et al. (2020).

3. We also consider *distribution-dependent* bounds on the sample size needed to get a certificate for any given test example in terms of how "close" the example is to the "boundary" of the target function with respect to the distribution $\mathcal{D}$ and class $\mathcal{H}$. In particular, we define a notion of the *certificate coefficient* ε_x of an example x with respect to a data distribution $\mathcal{D}$ and target function $h^\star$, and prove matching upper and lower bounds on sample size as a function of ε_x, b, and the VC-dimension d of the class $\mathcal{H}$.

4. We examine how reweighted versions of the certificate coefficient can provide better bounds on the certificate size achievable from a polynomial-sized data sample.

1.2. Context and Related Work

Explainable ML research largely focuses on decision trees or key predictive features. In these approaches, a certificate for an instance x corresponds to the root-to-leaf path of x in the tree. A widely studied method for explaining a black-box model f involves first learning a decision tree T that closely approximates f. Once this surrogate decision tree T is obtained, a certificate for any instance x can be generated by retrieving its associated path in T (Craven and Shavlik, 1995; Breiman and Shang, 1996; Zhou and Hooker, 2016). Blanc et al. (2021) proposed algorithms for implicitly

learning the surrogate decision trees that approximate the target function, with provable performance guarantees under the uniform distribution. Ribeiro et al. (2018) were the first to introduce certificates (which they term anchors) providing high precision by identifying a minimal set of rules that *anchor* a prediction, ensuring that the output remains stable under small perturbations. They provide an efficient heuristic, based on greedy search, for finding such high-precision certificates. Blanc et al. (2022) investigates the minimum number of queries required to certify a monotone function's prediction at a given point. They define as certificate a subset of input coordinates such that fixing these guarantees the function's value remains unchanged; Gupta and Manoj further investigate this. There is also an increasing interest in hybrid models that are partially interpretable as studied by Ferry et al. (2023); Frost et al. (2024).

Another related line of work considers learning in the face of malicious noise (Kearns and Li, 1988; Balcan and Haghtalab, 2021), such as the breakthrough work of Awasthi et al. (2017) on guarantees for learning linear separators under malicious noise, as well as work on instance-targeted data-poisoning attacks (Barreno et al., 2006). Suciu et al. (2018) and Shafahi et al. (2018) demonstrated empirically that data-poisoning attacks can be highly effective, even when the adversary only adds correctly-labeled data to the training set. These targeted attacks have attracted attention in recent years due to their potential to compromise the trustworthiness of systems (Geiping et al., 2021; Mozaffari-Kermani et al., 2015; Chen et al., 2017). A key concern here is: when can predictions be trusted under such attacks? Most theoretical work on this question has focused on certifying stability of predictions under small changes to the training set (Gao et al., 2021; Levine and Feizi, 2021). However, recently, Balcan et al. (2022, 2023); Blum and Saless (2025) consider certifying the actual correctness of predictions, under assumptions of realizability and bounded adversarial power. Our work can be viewed as studying the shortest *proof* that can be provided for such a guarantee.

2. Preliminaries

2.1. Notation

The input space is denoted $\mathcal{X}$, and the label space $\{+1, -1\}$. A hypothesis class $\mathcal{H}$ is a subset of $\{-1, +1\}^{\mathcal{X}}$. A sequence $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is realizable by $\mathcal{H}$ if $\exists h \in \mathcal{H}$ such that $\sum_{i=1}^{n} \mathbb{1}[h(x_i) \neq y_i] = 0$. We use $[n]$ to denote the first n natural numbers, $[n] = \{1, 2, \ldots, n\}$.

2.2. Formal Setting

The primary subject in this work is the notion of a *certificate* for predictions on test points. The certificates we consider are in terms of subsets of the training data. Concretely, suppose that we are given a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ labeled by some unknown target function $h^\star \in \mathcal{H}$ (i.e., $y_i = h^\star(x_i), \forall i$). An ideal scenario would be that there are no corruptions whatsoever in the labels, so that S is completely realizable by $h^\star$. However, in practice, label corruptions are inevitable due to a variety of reasons like noisy measurements, human errors, etc. To account for this, we allow for a corruption budget $b \geq 0$. This leads to the following definition of a *robustly realizable dataset*.

Definition 1 (Robust Realizability) *For a budget $b \geq 0$, a sequence $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is b-robustly realizable by $\mathcal{H}$ if $\exists h \in \mathcal{H}$ such that $\sum_{i=1}^{n} \mathbb{1}[h(x_i) \neq y_i] \leq b$.*

Given a b-robustly realizable dataset S, we wish to certify that the prediction on a given test point x ought to be y, and we wish to frame this certificate in terms of a subset of S that is ideally small. For this, we require the notion of an *agreement* region.

Definition 2 (Robust Agreement Region) *A point (x, y) is in the b-robust agreement region of a labeled sequence $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ if*

$$\forall h \in \mathcal{H} : \sum_{i=1}^{n} \mathbb{1}[h(x_i) \neq y_i] \leq b \implies h(x) = y. \tag{1}$$

When $b = 0$, we refer to the 0-robust agreement region simply as the agreement region.

We are now ready to define our notion of robust certificates.

Definition 3 (Robust Certificate) *A sequence $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a b-robust certificate for (x, y) if*

1. *S is b-robustly realizable by $\mathcal{H}$.*
2. *(x, y) is in the b-robust agreement region of S.*

S is furthermore a minimal b-robust certificate for (x, y) if S is a b-robust certificate for (x, y), and no proper subsequence $S' \subset S$ is a b-robust certificate for (x, y).

Our setting assumes that we are given a b-robustly realizable training dataset S, together with a test point x which satisfies that (x, y) belongs to the b-robust agreement region of S, for some $y \in \{+1, -1\}$. Namely, S is itself a b-robust certificate for (x, y).[1] Our primary objective is to analyze and obtain the smallest $S' \subseteq S$ such that S' continues to be a b-robust certificate for (x, y).

3. Worst Case Bounds on Certificate Size

The motivating question for this section is: what is the smallest certificate for x that we can extract from S, and can a certificate of such size always be extracted, even for *worst-case* instances of S and x? We begin with two simple examples for certification in the case of no corruptions (i.e., $b = 0$).

Example 1 (Halfspaces) *Consider the class of d-dimensional halfspaces passing through the origin, i.e., $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{x \mapsto \text{sign}(w^T x) : w \in \mathbb{R}^d\}$. Suppose we are given a training dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ realizable by $\mathcal{H}$. Let $S_+ = \{x_i : i \in [n], y_i = +1\}$, $S_- = \{-x_i : i \in [n], y_i = -1\}$, and let $S_+ \cup S_- = \{z_1, \ldots, z_n\}$ for convenience. Note that we negate a negatively labeled x_i in S to form S_-. A test point $(x, +1)$ belongs to the agreement region of S if and only if $x \in \text{Cone}(S_+ \cup S_-)$, where $\text{Cone}(\cdot)$ denotes the conic hull. To see this, suppose first that $x \in \text{Cone}(S_+ \cup S_-)$. This means that $x = \sum_{i=1}^{n} \alpha_i z_i$, where $\alpha_i \geq 0, \forall i$. Let $w \in \mathbb{R}^d$ represent any halfspace that labels all the examples in S correctly. Then, observe that for every $z_i \in S_+ \cup S_-$, $w^T z_i \geq 0$. But this also means that $w^T x = \sum_{i=1}^{n} \alpha_i \cdot w^T z_i \geq 0$. Thus, $(x, +1)$ belongs to the agreement region of S. In the other direction, suppose that $x \notin \text{Cone}(S_+ \cup S_-)$.*

1. The problem of obtaining small-size subsets of the training data that can serve as good-faith certificates for test-time predictions only makes sense if the test point is in the agreement region of the training data to start with.

Then, because $\mathrm{Cone}(S_+ \cup S_-)$ *is a closed convex set, the separating hyperplane theorem (e.g., see Theorem 1 in Nachbar (2018)) implies the existence of* $w \in \mathbb{R}^d$*, such that* $w^T x < 0$*, and* $w^T y \geq 0$ *for* $y \in \mathrm{Cone}(S_+ \cup S_-)$*. In particular, this means that* w *labels all examples in* S *correctly. However,* $w^T x < 0$*, so* $(x, +1)$ *is not in the agreement region of* S*.*

So, consider a test point $(x, +1)$ *that belongs to the agreement region of* S*. By the preceding argument,* $x \in \mathrm{Cone}(S_+ \cup S_-)$*. But then, by Carathéodory's Theorem,* x *can be written as a conic combination of at most* d *points*[2] *from* $S_+ \cup S_-$*, which implies that* $(x, +1)$ *is in the agreement region of these points. Thus, we can use this subset of* $S_+ \cup S_-$ *(together with labels), which has size at most* d*, as a certificate for* $(x, +1)$*.*

Note that in the above example, the size of the training data could be much larger than the ambient dimension, i.e., $n \gg d$. Even so, as long as the test point belongs to the agreement region of the data, it is possible to obtain a certificate of size at most d. One might observe that the VC dimension of d-dimensional halfspaces passing through the origin is d. Given how predominantly the VC dimension features in the characterization of learning-theoretic properties of binary hypothesis classes, one might wonder if the VC dimension also characterizes certificate size. That is, could it be possible to always extract an $O(VC(\mathcal{H}))$-size certificate from the training data, whenever the given test point belongs to its agreement region? Our next example shows that this is not the case.

Example 2 (Singletons) *Consider the class of singletons on a domain of size* n*, i.e.,* $\mathcal{X} = [n]$*,* $\mathcal{H} = \{x \mapsto (-1)^{1-\mathbb{1}[x=i]} : i \in [n]\}$*. The VC dimension of this class is 1. Suppose that we are given a training dataset* $S = \{(1, -1), (2, -1), \ldots, (n-1, -1)\}$*. Observe that the test point* $(n, +1)$ *is in the agreement region of* S*, because the only hypothesis in* $\mathcal{H}$ *that labels* S *in the given way labels* n *positively. Note however, that the test point is not in the agreement region of any proper subset of* S*. We must therefore necessarily provide all the* $n-1$ *points in* S *to certify a positive label on* n*.*

The instance in the above example has a curious property: as long as there remains a single point in the domain whose label we haven't observed, it is not possible to completely determine the label of the test point. This, however, is the defining property of another combinatorial quantity in learning theory, known as the *hollow star number* (Bousquet et al., 2020). For example, the hollow star number is known to lower bound the sample complexity of proper PAC learners (Bousquet et al., 2020, Theorem 10). Indeed, as we show ahead, the hollow star number ends up also being the relevant quantity that characterizes worst-case minimum certificate sizes. First, we define a slightly more generalized version of the hollow star number, which allows us to handle $b \geq 0$ corruptions.

Definition 4 (Robust Realizability with Weights) *A weighted sequence* $S = \{(x_i, y_i, w_i)\}_{i=1}^n$ *is* b*-robustly realizable by* $\mathcal{H}$ *if* $\exists h \in \mathcal{H}$ *such that* $\sum_{i=1}^n w_i \cdot \mathbb{1}[h(x_i) \neq y_i] \leq b$*.*

Definition 5 (Robust Hollow Star) *A sequence* $T = \{(x_1, y_1), \ldots, (x_{s_b}, y_{s_b})\}$ *is a* b*-robust hollow star for* $\mathcal{H}$*, if there exist integer weights* $w_1, \ldots, w_{s_b}$ *and* $i^* \in [s_b]$ *such that:*

1. $w_i = 1$ *for all* $i \in [s_b] \setminus \{i^*\}$*.*
2. $w_{i^*} = b + 1$*.*
3. *The weighted sequence* $T = \{(x_i, y_i, w_i)\}_{i=1}^{s_b}$ *is not* b*-robustly realizable by* $\mathcal{H}$*.*

2. There are hardness results on computing the "Carathéodory number" of a point set (Bereg and Haghpanah, 2020).

4. *For any $i \in [s_b]$, $T \setminus \{(x_i, y_i, w_i)\}$ is b-robustly realizable by $\mathcal{H}$.*

The size s_b of the largest b-robust hollow star is the b-robust hollow star number of $\mathcal{H}$.

Remark 1 *We observe that setting $b = 0$ in the above definition recovers the standard definition (e.g., Definition 3 in Bousquet et al. (2020)) of the hollow star number. We also note that repeats are allowed in the sequence T (i.e., there might be $i \neq j$ where $x_i = x_j$).*

The following claim (proof in Section A.1) lower bounds the b-robust hollow star number in terms of the 0-robust hollow star number (referred to hereon simply as the hollow star number).

Claim 1 *For $b \geq 0$, let s_b be the b-robust hollow star number of $\mathcal{H}$. Then, $s_b \geq (b+1)(s_0 - 1) + 1$.*

The b-robust hollow star number *exactly* characterizes the smallest size of a reliable certificate.

Theorem 6 (Robust Hollow Star Characterizes Minimum Certificate Size) *Let $\mathcal{H}$ be a hypothesis class that has b-robust hollow star number s_b, S be a training dataset that is b-robustly realizable by $\mathcal{H}$, and x be a test point such that for some $y \in \{-1, 1\}$, (x, y) is in the b-robust agreement region of S. Then, there exists a b-robust certificate $S' \subseteq S$ for (x, y) that has size at most $s_b - 1$. Furthermore, there exists a training dataset S of size $s_b - 1$ that is b-robustly realizable by $\mathcal{H}$, test point x and test label $y \in \{-1, 1\}$ which satisfy that (x, y) belongs to the b-robust agreement region of S, such that no proper subsequence of S is a b-robust certificate for (x, y).*

Proof We establish the upper and lower bound in order:

Upper Bound. Consider $S' \subseteq S$ to be the smallest subset of S that is a b-robust certificate for (x, y)—let $S' = \{(x_1, y_1), \ldots, (x_k, y_k)\}$. Note that S' is a minimal b-robust certificate for (x, y). We argue that $\{(x_1, y_1), \ldots, (x_k, y_k), (x, 1-y)\}$ is a b-robust hollow star, which implies $k+1 \leq s_b$.

Consider the weighted sequence $T = \{(x_1, y_1, 1), \ldots, (x_k, y_k, 1), (x, 1 - y, b + 1)\}$. Now consider any $h \in \mathcal{H}$. If $h(x) = y$, then the weighted error of h on x is $b + 1$, and h does not b-robustly realize T. Otherwise, we have that $h(x) = 1 - y$. In this case, it must be the case that $\sum_{i=1}^{k} \mathbb{1}[h(x_i) \neq y_i] > b$, otherwise S' would not be a b-robust certificate for (x, y). Summarily, the weighted sequence T is not b-robustly realizable by any $h \in \mathcal{H}$.

Now, consider $T' = T \setminus \{(x, 1 - y, b + 1)\}$. Since S is b-robustly realizable by $\mathcal{H}$, so is S', which implies that the weighted sequence T' is also b-robustly realizable by $\mathcal{H}$. Now, consider $T' = T \setminus \{(x_i, y_i, 1)\}$ for any $i \in [k]$. Then, since S' is a minimal b-robust certificate for (x, y), there must exist some $h \in \mathcal{H}$ such that $\sum_{j=1, j\neq i}^{k} \mathbb{1}[h(x_j) \neq y_j] \leq b$ and $h(x) = 1 - y$. Otherwise, $S' \setminus \{(x_i, y_i)\}$ would be a smaller b-robust certificate for (x, y). Consequently, such an h b-robustly realizes the weighted sequence T', and we are done.

Lower Bound. Let $\{(x_1, y_1), \ldots, (x_{s_b}, y_{s_b})\}$ be the largest b-robust hollow star for $\mathcal{H}$, with corresponding weights $w_1, \ldots, w_{s_b}$. Let i^* be such that $w_{i^*} = b + 1$. We will set the training dataset $S = \{(x_i, y_i) : i \neq i^*\}$, test point to be $x_{i^\star}$ and test label to be $1 - y_{i^\star}$, and argue that S is a minimal b-robust certificate for $(x_{i^\star}, 1 - y_{i^\star})$.

Let T be the weighted sequence $\{(x_1, y_1, w_1), \ldots, (x_{s_b}, y_{s_b}, w_{s_b})\}$. By definition of the b-robust hollow star, $T' = T \setminus \{(x_{i^*}, y_{i^*}, w_{i^*})\}$ is b-robustly realizable by $\mathcal{H}$. But note that all the weights in T' are equal to 1. This means that there exists $h \in \mathcal{H}$ such that $\sum_{i \in [s_b], i \neq i^*} \mathbb{1}[h(x_i) \neq y_i] \leq b$. Namely, S is b-robustly realizable by $\mathcal{H}$.

Next, consider any $h \in \mathcal{H}$ that satisfies $\sum_{i \in [s_b], i \neq i^*} \mathbb{1}[h(x_i) \neq y_i] \leq b$. Then, it must be the case that $h(x_{i^*}) = 1 - y_{i^*}$; otherwise, the weighted sequence

$$T = \{(x_1, y_1, w_1), \ldots, (x_{i^*}, y_{i^*}, w_{i^*}), \ldots, (x_{s_b}, y_{s_b}, w_{s_b})\}$$

would be b-robustly realizable by $\mathcal{H}$, which would contradict that $\{(x_1, y_1), \ldots, (x_{s_b}, y_{s_b})\}$ is a b-robust hollow star. Thus, S is a b-robust certificate for $(x_{i^*}, 1 - y_{i^*})$.

Finally, we argue that S is minimal. Consider any $S' \subsetneq S$. From above, we know $\exists h \in \mathcal{H}$ such that S' is b-robustly realizable by h and $h(x_{i^*}) = 1 - y_{i^*}$. Furthermore, observe that $S' \cup \{(x_{i^*}, y_{i^*})\}$ excludes at least one of the members of the hollow star S. Let T' be the weighted version of S', and consider $T' \cup \{(x_{i^*}, y_{i^*}, w_{i^*})\}$. By the requirements of the b-robust hollow star, $T' \cup \{(x_{i^*}, y_{i^*}, w_{i^*})\}$ must be b-robustly realizable by $\mathcal{H}$. Because $w_{i^*} = b + 1$, this means that $\exists h \in \mathcal{H}$ which b-robustly realizes S', and satisfies $h(x_{i^*}) = y_*$. Thus, S' is not a b-robust certificate for $(x_{i^*}, 1 - y_{i^*})$. ■

Remark 2 *Let us revisit the examples of halfspaces and singletons considered earlier for the setting with no corruptions. For d-dimensional halfspaces passing through the origin (Example 1), our certificate based on Carathéodory's theorem had size d. Indeed, this class has hollow star number equal to $d + 1$, which means that this certificate size is optimal in general. On the other hand, the class of singletons on a domain of size n has hollow star number equal to n. This validates why we were unable to obtain a certificate of size smaller than $n - 1$ in Example 2.*

Thus, we have shown that in general, the b-robust hollow star number optimally characterizes minimum reliable certificate size. While Claim 1 relates the b-robust hollow star to the hollow star number (which is known for many natural classes, e.g., see the examples in Section 2.1 in Bousquet et al. (2020)), it merely gives a lower bound. It would be interesting to see if an upper bound can also be obtained, and more generally to chart out the b-robust hollow star for natural classes.

For example, we can show that the b-robust hollow star number for singletons on a domain of size n is *exactly* $(b + 1)(n - 1) + 1$ (Section A.2). In addition, for d-dimensional halfspaces, one can use the same reasoning as in Example 1, together with the *Tolerance* Carathéodory theorem (Thereom 4.1 in Montejano and Oliveros (2011), Theorem 2 in Tuza (1989)) to obtain certificates of size $< (d+b)^{O(b)}$ for test points in the b-robust agreement of the training data. From Theorem 6, we then know that the b-robust hollow star number of halfspaces is at most $(d + b)^{O(b)}$ (the lower bound from Claim 1 is only $\Omega(bd)$.) One way to close the gap for halfspaces would be to show a lower bound for the Tolerance Carathéodory theorem. The simplest phrasing of this for $b = 1$ amounts to the following purely convex-analytic question:

Open Question 1 *Can one construct a set S of $\Omega(d^2)$ points in $\mathbb{R}^d$, and a test point x, such that:*
(1) no matter which point x' is removed from S, x is still in the convex hull of $S \setminus \{x'\}$,
(2) S is a minimal set satisfying (1). That is, no matter which point x_1 is removed from S, there exists another point x' that can be removed, such that x is no longer in the convex hull of $S \setminus \{x_1, x'\}$?
Equivalently, S is a minimal set such that any hyperplane through x has at least two points in S on either side.

Such a result exists for the Tolerance *Helly* Theorem (Montejano and Oliveros, 2011, Theorem 3.2).

4. Worst-case Distributional Bounds on Certificate Size

The certificate size bounds from the previous section were from a worst-case perspective—the training data S was arbitrary, and the test point (x, y) was also an arbitrary point in its agreement region. A standard assumption in learning theory however is that S is sampled i.i.d. from some marginal distribution $\mathcal{D}$ over $\mathcal{X}$, and then labeled by the unknown $h^\star$. Under this assumption, we can think about certifying a test point x with the label $h^\star(x)$ as the sample S gets large (we will study quantitative bounds on the size of S in Section 5).

A first observation is that if the distribution $\mathcal{D}$ is discrete, and the test point x belongs to the support of $\mathcal{D}$, we will eventually observe more than $2b + 1$ copies of it in S. With a corruption budget of b, an adversary can potentially corrupt the labels on b of these copies; however, $b + 1$ copies with the true label $h^\star(x)$ still remain in S. We can use these copies, all of which have the true label, to certify that x should be labeled as $h^\star(x)$, since any hypothesis that labels x differently makes strictly more than b mistakes on these copies. Furthermore, this certificate size is optimal—any certificate of smaller size is b-robustly realizable by every hypothesis in the class.

We now turn our attention to the more interesting case, where either the test point x is *not* in the support of $\mathcal{D}$ (e.g., if there is train-test distribution shift), or $\mathcal{D}$ is a continuous distribution. Observe that in either case, it is possible to certify x *only if* x eventually belongs to the agreement region of the (uncorrupted) sample S with positive probability. Interestingly, for distributions satisfying this property, we obtain the following sharp bound on the reliable (minimum) certificate size.

Theorem 7 (Eventual Certification with Small Certificate) *Let $\mathcal{H}$ be a hypothesis class having hollow star number s_0, $h^\star \in \mathcal{H}$ be the target hypothesis, and x be a test point. Let $\mathcal{D}$ be a distribution over samples labeled by $h^\star$, and suppose $\mathcal{D}, h^\star$ satisfy the property that: there exists m large enough such that $\Pr_{S\sim\mathcal{D}^m}[(x, h^\star(x))$ in agreement region of $S] > 0$. Then, for any $\delta > 0$, there exists $m(\delta)$ such that a b-robust certificate for $(x, h^\star(x))$ of size at most $(b + 1)(s_0 - 1)$ can be extracted from $S \sim \mathcal{D}^{m(\delta)}$ with probability $\geq 1 - \delta$. Moreover, there exists an instantiation of the setting where a b-robust certificate of smaller size is not possible.*

Remark 3 *Recall that once we draw a large enough sample S so that $(x, h^\star(x))$ is in the agreement region of S, Theorem 6 guarantees the existence of a b-robust certificate of size at most $s_b - 1$ in S, where s_b is the b-robust hollow star number of $\mathcal{H}$. However, by Claim 1, we know that $s_b - 1 \geq (b+1)(s_0 - 1)$. Thus, the certificate given by Theorem 7 potentially has a better size than that implied by Theorem 6.*

Proof Let m' be such that $\Pr_{S\sim\mathcal{D}^{m'}}[(x, h^\star(x))$ in agreement region of $S] = \delta' > 0$; such an m' exists by assumption that $(x, h^\star(x))$ belongs to the agreement region of a drawn sample S eventually. Consider drawing a sample S of size $(2b + 1)nm'/\delta'$, and let $S_1, S_2, \ldots, S_{(2b+1)n/\delta'}$ denote independent partitions of S into chunks of size m'. The expected number of chunks that satisfy that $(x, h^\star(x))$ is in the agreement region of the chunk is $(2b + 1)n$. By a Chernoff bound, the probability that at least $2b + 1$ chunks have $(x, h^\star(x))$ in their agreement region is at least $1 - e^{-\Omega(bn)}$. Now, with a corruption budget of size b, an adversary can insert a corruption in at most b chunks. Even so, the remaining $\geq b + 1$ chunks continue to have $(x, h^\star(x))$ in their agreement region. Without loss of generality, suppose $S_1, \ldots, S_{b+1}$ are all uncorrupted. Then, from Theorem 6 (with $b = 0$), we know that each of $S_1, \ldots, S_{b+1}$ will contain a certificate for $(x, h^\star(x))$ of size at most $s_0 - 1$; let these certificates be $C_1, \ldots, C_{b+1}$. The final b-robust certificate is simply a concatenation of $C_1, \ldots, C_{b+1}$.

To see that this is a b-robust certificate, observe that it is realizable by $\mathcal{H}$ since it is uncorrupted. Now consider any $h \in \mathcal{H}$ that makes at most b mistakes on the concatenation. Then, observe that h makes no mistakes on some C_i. Since C_i is a certificate for $(x, h^\star(x))$, h must label x as $h^\star(x)$. Setting $n = O\left(\log(1/\delta)/b\right)$ ensures that a sample S' of size $m(\delta) = O\left(m' \log(1/\delta)/\delta'\right)$ suffices for failure probability at most δ.

Finally, we argue that a b-robust certificate size of $(b+1)(s_0 - 1)$ is optimal in general for this setting. Let $\{(x_1, y_1), \ldots, (x_{s_0}, y_{s_0})\}$ be a hollow star for $\mathcal{H}$. Note that by the hollow star property, $\{(x_1, y_1), \ldots, (x_{s_0}, -y_{s_0})\}$ is realizable by some $h^\star \in \mathcal{H}$—let this $h^\star$ be the target hypothesis, and x_{s_0} be the test point. Let $\mathcal{D}$ be the distribution that puts all of its mass uniformly on $x_1, \ldots, x_{s_0-1}$. Note that $\mathcal{D}$ satisfies the specifications in the theorem statement.

Suppose that the adversary creates no corruptions. We claim that in this case, any b-robust certificate for $(x_{s_0}, -y_{s_0})$ must necessarily contain at least $b+1$ copies of (x_i, y_i) for every $i \in \{1, \ldots, s_0 - 1\}$. To see this, suppose that for some i, there are at most b copies of (x_i, y_i) included in the certificate. Since the adversary does not corrupt any samples, the rest of the samples included in the certificate comprise of (copies of) $(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_{s_0-1}, y_{s_0-1})$. Now, by the hollow star property, observe that there exists $h \in \mathcal{H}$, which realizes

$$\{(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_{s_0-1}, y_{s_0-1}), (x_{s_0}, y_{s_0})\},$$

and this h must label x_i as $-y_i$ (by non-realizability of the hollow star). Then, because the certificate only consists of at most b copies of (x_i, y_i), it is b-robustly realizable by h. This contradicts it being a b-robustly realizable certificate for $(x_{s_0}, -y_{s_0})$, since h labels x_{s_0} as y_{s_0}. ■

We end this section with a final observation. The conclusion of Theorem 7 implies that any distribution $\mathcal{D}$ which satisfies that $(x, h^\star(x))$ eventually belongs to the agreement region of a sample drawn from $\mathcal{D}$, actually satisfies the seemingly *stronger* condition: that $(x, h^\star(x))$ eventually belongs to the b-*robust* agreement region of a sample drawn from $\mathcal{D}$. Note that this stronger condition implies the weaker condition. Therefore, these conditions are equivalent.

5. Distribution-dependent Bounds on Sample Size

The previous section establishes a sharp bound on the b-robust certificate size for a given test point that can be obtained from a sample drawn from a distribution. This bound is *distribution-independent*, in that it holds for *all* distributions that merely satisfy that the test point eventually belongs to the agreement region of the drawn sample. However, it does not quantify the number of samples that need to be seen for the test point to belong to the b-robust agreement region.

In this section, we derive *distribution-dependent* bounds on the size of the sample S that needs to be drawn for a test point x to lie in the b-robust agreement region of S. Notably, observe that once S contains x in its b-robust agreement region, it is by definition a certificate for x. Thus, our bound for the sample size is also a bound for the b-robust certificate size for x. The sample complexity that we state is in terms of a quantity that depends on the distribution $\mathcal{D}$, target hypothesis $h^\star$ and test point x we wish to certify. We term this quantity the "Certificate Coefficient" of x.

Definition 8 (Certificate Coefficient) *Let $\mathcal{H}$ be a hypothesis class, $h^\star \in \mathcal{H}$ the ground truth hypothesis, and $\mathcal{D}$ a marginal distribution over the domain $\mathcal{X}$. For a test point $x \in \mathcal{X}$, denote the set*

of hypotheses that disagree with $h^\star$ on x by:

$$\mathcal{H}_x = \{h \in \mathcal{H} : h(x) \neq h^\star(x)\}. \tag{2}$$

The certificate coefficient $\varepsilon_x = \varepsilon_x(\mathcal{D}, h^\star)$ of x is defined as:

$$\varepsilon_x = \inf_{h \in \mathcal{H}_x} \Pr_{z \sim \mathcal{D}}[h(z) \neq h^\star(z)]. \tag{3}$$

We will refer to $\varepsilon_x(\mathcal{D}, h^\star)$ as simply ε_x when the parameters involved can be deduced from context. Intuitively, the certificate coefficient ε_x measures how quickly the point x gets embedded in the agreement region of a sample from $\mathcal{D}$. If ε_x is substantial, then x is in the "interior" of the (labeled) distribution, and we should quickly expect x to be in the agreement region. If ε_x is tiny, then x is at the boundary, and we have to see many points before x falls in the agreement region.

5.1. Upper Bound on the Sample Complexity

While our bounds above on certificate size were in terms of the b-robust hollow star number, our sample complexity bound is more familiar-looking and is in terms of the VC dimension of the class.

Theorem 9 (Sample Complexity for Robust Certification) *Let $\mathcal{H}$ be a hypothesis class having VC dimension $d < \infty$ and $h^\star \in \mathcal{H}$ be the target hypothesis. For any marginal data distribution $\mathcal{D}$, test point x satisfying $\varepsilon_x > 0$, corruption budget $b \geq 0$ and failure probability $\delta \in (0, 1)$, obtaining an i.i.d. sample S (labeled by $h^\star$) of size $m = O\left(\frac{b + d\log(1/\varepsilon_x) + \log(1/\delta)}{\varepsilon_x}\right)$ suffices to ensure that with probability $1 - \delta$, x belongs to the b-robust agreement region of S.*

Proof sketch We sketch the proof for the case when $\mathcal{H}$ is finite. Fix some $h \in \mathcal{H}_x$. Upon drawing a sample S of size m i.i.d. from $\mathcal{D}$ labeled by $h^\star$, the expected number of mistakes that h makes on S is, by definition of the certificate coefficient, at least $\varepsilon_x m$. Then, by a Chernoff bound, the probability that h makes less than $\varepsilon_x m/2$ mistakes is at most $\exp(-\varepsilon_x m/8)$. Thus, setting $m = \frac{8\log(|\mathcal{H}|/\delta)}{\varepsilon_x}$, together with a union bound, ensures that with probability at least $1-\delta$, every $h \in \mathcal{H}_x$ makes strictly more than $\varepsilon_x m/2$ mistakes on S. If in addition, $m \geq 2b/\varepsilon_x$, then this means that every $h \in \mathcal{H}_x$ makes more than b mistakes on S. But this also means that $(x, h^\star(x))$ is in the b-robust agreement region of S. Thus, setting $m = \max\left(2b/\varepsilon_x, \frac{8\log(|\mathcal{H}|/\delta)}{\varepsilon_x}\right) = O\left(\frac{b + \log|\mathcal{H}| + \log(1/\delta)}{\varepsilon_x}\right)$ is sufficient for the required guarantee. The extension to infinite classes having bounded VC dimension involves the standard double sampling + symmetrization trick, and is given in Section A.3. ■

Remark 4 (Simultaneous Certification of Multiple Points) *We observe that the proof of Theorem 9 can be adapted to ensure that S robustly certifies multiple test points simultaneously. Concretely, for a class $\mathcal{H}$ having VC dimension d, target hypothesis $h^\star \in \mathcal{H}$, and marginal distribution $\mathcal{D}$, let $\mathcal{X}_\varepsilon = \{x \in \mathcal{X} : \varepsilon_x(\mathcal{D}, h^\star) \geq \varepsilon\}$. Observe that any h that mislabels at least one point—say x_1—in $\mathcal{X}_\varepsilon$ satisfies that $h \in \mathcal{H}_{x_1} \implies \Pr_{z\sim\mathcal{D}}[h(z) \neq h^\star(z)] \geq \varepsilon$ by definition of ε_{x_1}. From here, the analysis in the proof of Theorem 9 gives us that for a sample S of size $O\left(\frac{b + d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right)$, with probability $1 - \delta$, every point in $\mathcal{X}_\varepsilon$ belongs to the b-robust agreement region of S.*[3]

3. One can also analyze the probability mass of $\mathcal{X}_\epsilon$, as it is directly related to the *disagreement coefficient* θ introduced by Hanneke (2007). Specifically, it holds that $\Pr[x \in \mathcal{X}_\varepsilon] = 1 - \Pr[x \in \mathcal{X} \setminus \mathcal{X}_\varepsilon] \geq 1 - \theta\varepsilon$.

5.2. Tightness of the Upper Bound

Our next proposition shows that the sample complexity bound in Theorem 9 is tight in general.

Proposition 10 (Tightness of Sample Complexity for Robust Certification) *For any failure probability $\delta < \frac{1}{100}$, the bound in Theorem 9 is tight up to a constant.*

Proof We will separately show that each individual term in the bound from Theorem 9 is necessary.

Tightness of $\frac{b}{\varepsilon_x}$: Consider any class $\mathcal{H}$, target $h^\star \in \mathcal{H}$, distribution $\mathcal{D}$ and test point x that satisfies $\varepsilon_x > 0$. Let $\mathcal{H}_x = \{h \in \mathcal{H} : h(x) \neq h^\star(x)\}$. By definition of ε_x (3), there exists $h \in \mathcal{H}_x$ that satisfies $\Pr_{z\in\mathcal{D}}[h(z) \neq h^\star(x)] \leq 2\varepsilon_x$. If we draw m samples i.i.d. from $\mathcal{D}$, labeled by $h^\star$, the expected number of mistakes that h makes on these samples is at most $2\varepsilon_x m$. By a Chernoff bound, the probability that h makes at least $4\varepsilon_x m$ mistakes on the samples is at most $\exp(-2\varepsilon_x m/3)$. Therefore, with probability at least $1 - \exp(-2\varepsilon_x m/3)$, h makes strictly less than $4\varepsilon_x m$ mistakes. If $m = \frac{b+1}{4\varepsilon_x}$, we get that with probability at least $1 - \exp(-(b+1)/6) \geq \frac{1}{100} > \delta$, h makes strictly less than $b+1$ mistakes, which implies that $(x, h^\star(x))$ is not in the b-robust agreement region.

Tightness of $\frac{d}{\varepsilon_x}\log\left(\frac{1}{\varepsilon_x}\right)$ and $\frac{1}{\varepsilon_x}\log\left(\frac{1}{\delta}\right)$: To show the tightness of the other two terms, we will work with the following class $\mathcal{H}$. For any given $d \geq 1$, the domain $\mathcal{X}$ of the class will be $\{x_0, \ldots, x_k\}$, where we can pick k to be any integer that satisfies $1 + \sqrt{\frac{100}{d}} < \frac{1}{2}\log\left(\frac{k}{d}\right)$.[4] For example, $k = (100d)^{20}$ is a valid choice. The hypothesis class $\mathcal{H}$ comprises of $\binom{k}{d} + 1$ hypotheses. For every $S \subset \{x_1, \ldots, x_k\}, |S| = d$, there is a hypothesis $h_S \in \mathcal{H}$ that labels exactly the set S as 1, and the rest of the domain as -1. Additionally, there is a special hypothesis $h^\star \in \mathcal{H}$ which satisfies that $h^\star(x_0) = 1, h^\star(x_i) = -1, \forall i \in [k]$. That is, $h^\star$ only labels x_0 as 1. Moreover, $h^\star$ is the only hypothesis in $\mathcal{H}$ that labels x_0 as 1; all the other $\binom{k}{d}$ hypotheses label x_0 as -1. It can be readily verified that the VC dimension of $\mathcal{H}$ is d. In what follows, we will set $b = 0$ for convenience.

We will set the target hypothesis to be $h^\star$, the data distribution $\mathcal{D}$ to be uniform on $\{x_1, \ldots, x_k\}$, and the test point x to be x_0. Crucially, $\mathcal{D}$ has zero mass on x_0. [5] Recall that $h^\star(x_0) = 1$ and $h^\star(x_i) = -1, \forall i \in [k]$. Since every other hypothesis in $\mathcal{H}$ labels x_0 as -1, and also labels some d points in $\{x_1, \ldots, x_k\}$ as 1, we have that $\varepsilon_x = \frac{d}{k}$. Now suppose that we obtain a sample of size $m = \frac{d}{2\varepsilon_x}\log\left(\frac{1}{\varepsilon_x}\right) = \frac{k}{2}\log\left(\frac{k}{d}\right)$. Then, by a coupon collector argument (see Lemma 19 in Bousquet et al. (2020)), we have that with probability at least $\frac{1}{100} > \delta$, the number of distinct elements from $\{x_1, \ldots, x_k\}$ that we see in the sample are at most $k - d$. But this means that there is some subset $S \subset \{x_1, \ldots, x_k\}$ of d elements whose labels we have not seen. Thus, h_S is consistent with the labeled data seen so far, rendering $(x, h^\star(x))$ to not be in the agreement region of the sample.

Finally, note also that upon drawing m samples, the probability that we do not see any point in $S = \{x_1, \ldots, x_d\}$ is $\left(1 - \frac{d}{k}\right)^m \geq \exp(-2dm/k)$, which is larger than δ when $m < \frac{k}{2d}\log\left(\frac{1}{\delta}\right) = \frac{1}{2\varepsilon_x}\log\left(\frac{1}{\delta}\right)$. In the absence of any point in S, $(x, h^\star(x))$ will not be in the agreement region.

Summarily, we have shown that any (generic) upper bound on the sample size m, which guarantees that a test point x belongs to the b-robust agreement region of the sample necessarily satisfies

$$m \geq \max\left\{\frac{b+1}{4\varepsilon_x}, \frac{d}{2\varepsilon_x}\log\left(\frac{1}{\varepsilon_x}\right), \frac{1}{2\varepsilon_x}\log\left(\frac{1}{\delta}\right)\right\} = \Omega\left(\frac{b + d\log(1/\varepsilon_x) + \log(1/\delta)}{\varepsilon_x}\right),$$

4. This is for the purpose of instantiating a coupon collector bound, e.g., as in Remark 20, Bousquet et al. (2020).
5. We can also allow a mass of $o(\varepsilon_x/(d\log(1/\varepsilon_x)))$ on x_0—this ensures x_0 is not seen in the sample with good chance.

which completes the proof of the proposition and establishes the tightness of Theorem 9. ■

6. Shorter Certificates by Reweighting

Theorem 9 gives a quantitative bound on the number of samples required to ensure that a test point x is in the b-robust agreement region (and hence, also certifies x); however, the bound scales inversely with the certificate coefficient ε_x. While our bound from Theorem 6 guarantees the existence of a potentially smaller certificate of size equal to the b-robust hollow star number, we might not have tight estimates of this quantity, as well as a constructive procedure to extract the shorter certificate. This motivates us to consider cleaner algorithmic primitives that might lead to short certificates.

Example 3 (Rejection Sampling) *Consider the case when $\mathcal{H}$ is the class of halfspaces in $\mathbb{R}^d$ (not necessarily passing through the origin), and $\mathcal{D}$ is uniform on the unit ball $\{x : \|x\|_2 \leq 1\}$. Suppose that the target halfspace $h^\star$ labels every point in the ball positively (i.e., it does not cut through the ball), and that the test point x is at $(1/2, 0, 0, \ldots, 0)$. We can verify that $\varepsilon_x \leq 2^{-\Omega(d)}$ (as realized by the halfspace $x_1 < 1/2$; see Theorem 2.7 in Blum et al. (2020) for the volume of the ball excluded by this halfspace), and so, the upper bound in Theorem 9 would suggest that we draw $b \cdot 2^{\Omega(d)}$ samples to ensure that x lies in the b-robust agreement region of the samples with constant probability. We might extract a shorter certificate from this sample using the worst-case bound from Theorem 6, but even this scales as the b-robust hollow star number of $\mathcal{H}$, for which the best upper bound is $(d+b)^{O(b)}$. Consider the following alternate recipe. Suppose we discard any samples from $\mathcal{D}$ that are not contained in the ball of radius $1/2$ centered at x. This induces the uniform distribution $\mathcal{D}_w$ on the smaller ball, and in expectation, we obtain one sample from $\mathcal{D}_w$ for every 2^d samples from $\mathcal{D}$. Crucially, notice that $\varepsilon_x(\mathcal{D}_w, h^\star) = 1/2$. By Theorem 9, $O(b+d)$ samples from $\mathcal{D}_w$ are sufficient to ensure that x lies in the b-robust agreement region of the samples with constant probability. Because of our rejection sampling, this really requires obtaining $(b+d) \cdot 2^{O(d)}$ samples from $\mathcal{D}$. The final result is that we have a b-robust certificate of size only $O(b+d)$ for x. This required us to draw only a polynomially larger sample than we would have if we directly applied Theorem 9.*

In the example above, we manipulated the distribution of the data to boost up the certificate coefficient. This allowed us to obtain a better sample size bound, resulting also in a better *certificate* size! In the process, we maintained that the number of samples required is only a polynomial factor larger in the (original) problem parameters. This procedure motivates the following definition:

Definition 11 (Optimal Reweighted Coefficient) *Let $\mathcal{H}$ be a hypothesis class having VC dimension d, $h^\star \in \mathcal{H}$ the target hypothesis, $b \geq 0$ a corruption budget, and $\mathcal{D}$ a distribution over $\mathcal{X}$. Let x be a test point having $\varepsilon_x(\mathcal{D}, h^\star) > 0$. We say that $w : \mathcal{X} \to [0,1]$ is a valid reweighting of distribution $\mathcal{D}$, resulting in the distribution $\mathcal{D}_w$ where, $\mathcal{D}_w(z) \propto w(z) \cdot \mathcal{D}(z)$, if it satisfies:*

$$Z = \int_z w(z)\mathcal{D}(z)\,dz \geq \frac{1}{\mathrm{poly}(b, d, 1/\varepsilon_x)}. \tag{4}$$

Then, we define $\varepsilon_x^\star$ as the supremum of certificate coefficients over valid reweightings of $\mathcal{D}$:

$$\varepsilon_x^\star = \sup_{\substack{w:\mathcal{X}\to[0,1] \\ w \text{ valid}}} \varepsilon_x(\mathcal{D}_w, h^\star) = \sup_{\substack{w:\mathcal{X}\to[0,1] \\ w \text{ valid}}} \left\{ \inf_{h\in\mathcal{H}_x} \Pr_{z\sim\mathcal{D}_w}[h(z) \neq h^\star(z)] \right\}. \tag{5}$$

Here, $\mathcal{H}_x$ is as defined in (2).

For any valid reweighting w of $\mathcal{D}$, one can draw a sample from $\mathcal{D}_w$ as follows: until a sample gets accepted, draw $z \sim \mathcal{D}$, and accept z with probability $w(z)$. Then, the rejection sampling procedure sketched out in Example 3 gives the following theorem, whose proof is in Section A.4.

Theorem 12 (Certificates by Reweighting) *Let $\mathcal{H}$ be a hypothesis class having VC dimension $d < \infty$ and $h^\star \in \mathcal{H}$ be the target hypothesis. For any marginal data distribution $\mathcal{D}$, test point x satisfying $\varepsilon_x = \varepsilon_x(\mathcal{D}, h^\star) > 0$, corruption budget $b \geq 0$ and failure probability $\delta \in (0,1)$, with probability $1-\delta$, we can obtain a certificate for $(x, h^\star(x))$ of size $O\left(\frac{b+d\log(1/\varepsilon_x^\star)+\log(1/\delta)}{\varepsilon_x^\star}\right)$ by obtaining an i.i.d. sample S (labeled by $h^\star$) of size $\mathrm{poly}(b, d, 1/\varepsilon_x, 1/\delta)$.*

Remark 5 *Given that we were able to boost the certificate coefficient all the way up to a constant in Example 3, one might ask: is this always possible? If it were, the rejection sampling procedure from Theorem 12 would always guarantee a certificate of size just $O(b+d)$ once the sample gets sufficiently large. Unfortunately, the lower bound from Theorem 7 prevents this from happening; there are instances where $\varepsilon_x^\star \lesssim \frac{1}{s_0} + \frac{d}{bs_0}$ (ignoring log factors), where s_0 is the hollow star number.*

Example 3 and the guarantee of Theorem 12 suggest that reweighting/rejection sampling can be fruitful in obtaining short certificates. However, as stated, the computation of $\varepsilon_x^\star$ requires knowledge of the distribution $\mathcal{D}$—this is indeed a very strong assumption, leading to our next open question.

Open Question 2 *Can we construct general-purpose reweighting schemes (possibly employing some form of iterated multiplicative weights like AdaBoost) that do not require knowing $\mathcal{D}$, but implicitly converge to a weighting w for which $\varepsilon_x(\mathcal{D}_w, h^\star) \approx \varepsilon_x^\star$?*

7. Discussion

In this work, we introduced the notion of short certificates: subsets of the training data that provably determine the correct label of a test point x even under up to b corruptions. To characterize their worst-case size, we proposed the robust hollow star number, a generalization of the hollow star number of Bousquet et al. (2020). We then studied worst-case distributional bounds and introduced the certificate coefficient ε_x, which captures the distribution-dependent difficulty of certifying a given point. For this setting, we established bounds on the sample size required to certify x, as a function of ε_x, the corruption budget b, and the VC-dimension d of the class $\mathcal{H}$. We further showed that reweighted variants of ε_x, can lead to improved bounds from polynomial-sized samples. Our framework also naturally subsumes the agnostic learning setting as a special case of adversarial model, under the assumption that the true label of x is given by the hypothesis closest to the target function. Following Balcan et al. (2022), a b-robust certificate becomes a b-agnostic certificate when the corruption budget reflects the error rate of the best-in-class hypothesis: a set S certifies (x, y) if every hypothesis predicting $1-y$ on x incurs more than b errors on S. We believe Open Question 1 and Open Question 2 pose exciting directions for future research.

Acknowledgments

This work was supported in part by the National Science Foundation under grants CCF-2212968, ECCS-2216899, and ECCS-2217023, by the Simons Foundation under the Simons Collaboration on the Theory of Algorithmic Fairness, and by the Office of Naval Research MURI Grant N000142412742.

References

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

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 Nika Haghtalab. Noise in classification. *Beyond the Worst-Case Analysis of Algorithms*, page 361, 2021.

Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dravyansh Sharma. Robustly-reliable learners under poisoning attacks. In *Conference on Learning Theory*, pages 4498–4534. PMLR, 2022.

Maria-Florina F Balcan, Steve Hanneke, Rattana Pukdee, and Dravyansh Sharma. Reliable learning in challenging environments. *Advances in Neural Information Processing Systems*, 36:48035–48050, 2023.

Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ASIACCS '06, page 16–25, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595932720. doi: 10.1145/1128817.1128824. URL https://doi.org/10.1145/1128817.1128824.

Sergey Bereg and Mohammadreza Haghpanah. Computing the caratheodory number of a point. In *Canadian Conference on Computational Geometry*, 2020.

Guy Blanc, Jane Lange, and Li-Yang Tan. Provably efficient, succinct, and precise explanations. *Advances in Neural Information Processing Systems*, 34:6129–6141, 2021.

Guy Blanc, Caleb Koch, Jane Lange, and Li-Yang Tan. The query complexity of certification. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 623–636, 2022.

Avrim Blum and Donya Saless. Regularized robustly reliable learners and instance targeted attacks, 2025. URL https://arxiv.org/abs/2410.10572.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In *Conference on Learning Theory*, pages 582–609. PMLR, 2020.

L. Breiman and Nong Shang. Born again trees. 1996. URL https://api.semanticscholar.org/CorpusID:2145744.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Xiaodong Song. Targeted backdoor attacks on deep learning systems using data poisoning. *ArXiv*, abs/1712.05526, 2017. URL https://api.semanticscholar.org/CorpusID:36122023.

Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995. URL https://proceedings.neurips.cc/paper_files/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. URL https://arxiv.org/abs/1702.08608.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Comput. Surv.*, 55(9), January 2023. ISSN 0360-0300. doi: 10.1145/3561048. URL https://doi.org/10.1145/3561048.

Julien Ferry, Gabriel Laberge, and U. Aïvodji. Learning hybrid interpretable models: Theory, taxonomy, and methods. *ArXiv*, abs/2303.04437, 2023. URL https://api.semanticscholar.org/CorpusID:257405014.

Nave Frost, Zachary Lipton, Yishay Mansour, and Michal Moshkovitz. Partially interpretable models with guarantees on coverage and accuracy. In Claire Vernade and Daniel Hsu, editors, *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pages 590–613. PMLR, 25–28 Feb 2024. URL https://proceedings.mlr.press/v237/frost24a.html.

Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. In *Uncertainty in Artificial Intelligence*, pages 2135–2145. PMLR, 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*, 2021. URL https://openreview.net/forum?id=01olnfLIbD.

Meghal Gupta and Naren Sarayu Manoj. *An Optimal Algorithm for Certifying Monotone Functions*, pages 207–212. doi: 10.1137/1.9781611977585.ch19. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977585.ch19.

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. In *Proceedings of the 20th ACM Symposium on Theory of Computing*, pages 267–280, 1988.

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

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Luis Montejano and Deborah Oliveros. Tolerance in helly-type theorems. *Discrete & Computational Geometry*, 45(2):348–357, 2011.

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, 2015. doi: 10.1109/JBHI.2014.2344095.

John Nachbar. A Basic Separation Theorem for Cones. https://bpb-us-w2.wpmucdn.com/sites.wustl.edu/dist/3/2139/files/2019/09/conesepkkt.pdf, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

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. *Advances in neural information processing systems*, 31, 2018.

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, Baltimore, MD, August 2018. USENIX Association. ISBN 978-1-939133-04-5. URL https://www.usenix.org/conference/usenixsecurity18/presentation/suciu.

Zsolt Tuza. Minimum number of elements representing a set system of given rank. *Journal of Combinatorial Theory, Series A*, 52(1):84–89, 1989. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(89)90064-2. URL https://www.sciencedirect.com/science/article/pii/0097316589900642.

Yichen Zhou and Giles Hooker. Interpreting models via single tree approximation. *arXiv: Methodology*, 2016. URL https://api.semanticscholar.org/CorpusID:88515329.

Appendix A. Supplementary Proofs

A.1. Proof of Claim 1

Let $\{(x_1, y_1), \ldots, (x_{s_0}, y_{s_0})\}$ be a hollow star for $\mathcal{H}$, and consider $S = \{(x_1, y_1), \ldots, (x_{s_0-1}, y_{s_0-1})\}$ which contains all but the last element of the hollow star. We will show that the sequence $T = \{S_1, \ldots, S_{b+1}, (x_{s_0}, y_{s_0})\}$, where each S_i is a copy of S, is a b-robust hollow star for $\mathcal{H}$. Since T has size $(b+1)(s_0-1)+1$, this will prove the claim.

Consider assigning a weight 1 to every element in $S_1, \ldots, S_{b+1}$, and the weight $b+1$ to the last element (x_{s_0}, y_{s_0}). Then, the weighted sequence is not b-robustly realizable by $\mathcal{H}$. To see this, consider any $h \in \mathcal{H}$. If $h(x_{s_0}) \neq y_{s_0}$, then the weighted error on x_{s_0} is already $b+1$. If $h(x_{s_0}) = y_{s_0}$, then there must exist some $k \in \{1, 2, \ldots, s_0 - 1\}$ such that $h(x_k) \neq y_k$. Otherwise, the sequence $\{(x_1, y_1), \ldots, (x_{s_0}, y_{s_0})\}$ would be realizable by h, which violates the fact that this sequence is a hollow star. Since there exist $b+1$ copies of (x_k, y_k) in T, the total weighted error of h is again at least $b+1$.

We now argue that removing any (weighted) element of T makes it b-robustly realizable by $\mathcal{H}$. If we remove (x_{s_0}, y_{s_0}), we are simply left with $b+1$ copies of S. But S is realizable by $\mathcal{H}$, since it excludes the last element of the hollow star. On the other hand, suppose we remove (x_k, y_k) from S_i in T, for some $k \in \{1, 2, \ldots, s_0 - 1\}, i \in \{1, 2, \ldots, b+1\}$. Then, observe (by the hollow star property again) that there must exist some $h \in \mathcal{H}$ which realizes $(S_i \setminus \{(x_k, y_k)\}) \cup \{(x_{s_0}, y_{s_0})\}$. This h therefore makes a mistake only on the remaining b copies of (x_k, y_k), each of which has weight 1. We conclude the proof. ■

A.2. b-robust Hollow Star Number for Singletons

Claim 2 (Singletons b-robust hollow star) *Consider the class $\mathcal{H}$ of singletons on a domain of size n, i.e., $\mathcal{X} = [n]$, $\mathcal{H} = \{x \mapsto (-1)^{1-\mathbb{1}[x=i]} : i \in [n]\}$. The b-robust hollow star number of $\mathcal{H}$ is exactly equal to $(b+1)(n-1)+1$.*

Proof Let s_b be the b-robust hollow star number of $\mathcal{H}$. The (0-robust) hollow star number of $\mathcal{H}$ is equal to n, and hence by Claim 1, we know that $s_b \geq (b+1)(n-1)+1$. We will now show that $s_b \leq (b+1)(n-1)+1$, which will prove the claim.

Suppose that $T = (x_1, y_1, b+1), (x_2, y_2, 1), \ldots, (x_{s_b}, y_{s_b}, 1)$ is a b-robust hollow star for $\mathcal{H}$, where as per Definition 5, the weight on (x_1, y_1) is $b+1$, and the weight on the rest of the points is 1. Note that no matter what y_1 is, there exists some $h_{i'} \in \mathcal{H}$ that labels x_1 as $-y_1$.

Now, consider the sequence $T \setminus \{(x_{s_b}, y_{s_b}, 1)\}$. There should exist $h_{i_1} \in \mathcal{H}$ that b-robustly realizes this sequence. In particular, h_{i_1} must label x_1 as y_1 (hence, h_{i_1} cannot be $h_{i'}$). Furthermore, it must be the case that h_{i_1} labels x_{s_b} as $-y_{s_b}$; otherwise, h_{i_1} would b-robustly realize all of T. Thus, h_{i_1} is a hypothesis in $\mathcal{H}$ which: (1) is not $h_{i'}$, (2) makes a mistake on (x_{s_b}, y_{s_b}), and (3) makes at most b mistakes on $(x_2, y_2), \ldots, (x_{s_b-1}, y_{s_b-1})$.

Repeating this argument with $T \setminus \{(x_{s_b-1}, y_{s_b-1}, 1)\}, T \setminus \{(x_{s_b-2}, y_{s_b-2}, 1)\}, \ldots, T \setminus \{(x_2, y_2, 1)\}$, we will have obtained $h_{i_1}, h_{i_2}, \ldots, h_{i_{s_b-1}}$, such that for every $j \in [s_b - 1]$, it holds that (1) h_{i_j} is not $h_{i'}$, (2) h_{i_j} makes a mistake on $(x_{s_b-j+1}, y_{s_b-j+1})$, and (3) h_{i_j} makes at most b mistakes on

$$\{(x_2, y_2), \ldots, (x_{s_b}, y_{s_b})\} \setminus \{(x_{s_b-j+1}, y_{s_b} - j + 1)\}.$$

Thus, we can think of the n hypotheses in $\mathcal{H}$ as n bins, and the $s_b - 1$ points $(x_2, y_2), \ldots, (x_{s_b}, y_{s_b})$ as $s_b - 1$ balls. From conditions (1) and (2), we get that each of the $s_b - 1$ balls is assigned one of

$n-1$ bins from the set $\mathcal{H} \setminus \{h_{i'}\}$. From condition (3), each of the $n-1$ bins is assigned no more than $b+1$ balls. Thus, it must hold that $(b+1)(n-1) \geq s_b - 1$, which completes the proof. ■

A.3. Proof of Theorem 9

Using the standard double sampling+symmetrization argument, we will generalize the bound sketched in the main body above for finite hypothesis classes to arbitrary (potentially infinite) hypothesis classes $\mathcal{H}$ having VC dimension d.

Again, let $h^\star \in \mathcal{H}$ be the target hypothesis, $\mathcal{D}$ the marginal distribution on the data, and x the test point. Let $\mathcal{H}_x$ and ε_x be defined as in (2) and (3).

Suppose we draw a sample $S \sim \mathcal{D}^m$ of size m, labeled by $h^\star$. Let A be the event: there exists $h \in \mathcal{H}_x$ which makes at most $\varepsilon_x m/4$ mistakes on S. We want to show that $\Pr_{S\sim\mathcal{D}^m}[A] \leq \delta$.

Towards this, consider (purely for the sake of analysis) drawing an additional sample $S' \sim \mathcal{D}^m$ independently of S. Let B be the event: there exists $h \in \mathcal{H}_x$ which makes at most $\varepsilon_x m/4$ mistakes on S, and *at least* $\varepsilon_x m/2$ mistakes on S'. We claim that $\Pr[A] \leq 2\Pr[B]$. To see this, observe that

$$\Pr[B] \geq \Pr[A \text{ and } B] \geq \Pr[A] \cdot \Pr[B|A].$$

Now, to bound $\Pr[B|A]$, let $h \in \mathcal{H}_x$ be the hypothesis which makes at most $\varepsilon_x m/4$ mistakes on S (this h exists since we condition on A). Since $h \in \mathcal{H}_x$, its expected mistakes on S' are at least $\varepsilon_x m$. So, since S' is independent of S, the probability that this h makes at least $\varepsilon_x m/2$ mistakes on S' is, by a Chernoff bound, at least $1/2$, provided $m \geq 6/\varepsilon_x$.

So, we continue to argue that $\Pr[B] \leq \delta/2$. First, instead of drawing S and S' independently from $\mathcal{D}^m$, we simply draw $S'' \sim \mathcal{D}^{2m}$, and think of the first m points as S and the second m points as S'. This is a distributionally identical way of obtaining S and S'. Now, we make an observation. Once we have drawn S'' (and therefore, S and S'), the event B only depends on the projection of $\mathcal{H}_x$ on S''. That is,

$$\begin{aligned}
&\Pr_{S''}[B] \\
&= \Pr_{S''}[\exists h \in \mathcal{H}_x|_{S''} \text{ s.t. } h \text{ makes at most } \varepsilon_x m/4 \text{ mistakes on } S, \text{ but at least } \varepsilon_x m/2 \text{ mistakes on } S'] \\
&\leq \sum_{h\in\mathcal{H}_x|_{S''}} \Pr_{S''}[h \text{ makes at most } \varepsilon_x m/4 \text{ mistakes on } S, \text{ but at least } \varepsilon_x m/2 \text{ mistakes on } S'].
\end{aligned}$$

Fix any $h \in \mathcal{H}_x|_{S''}$. In the form above, once S'' is drawn, the event in the parentheses either happens or it doesn't. So, we introduce some additional randomness. Suppose

$$S'' = (x_1, h^\star(x_1)), \ldots, (x_m, h^\star(x_m)), (x_{m+1}, h^\star(x_{m+1}), \ldots, (x_{2m}, h^\star(x_{2m})).$$

We now flip m independent fair coins $p_1, \ldots, p_m$. If p_i lands heads, we swap $(x_i, h^\star(x_i))$ and $(x_{m+i}, h^\star(x_{m+i}))$, otherwise we let this pair be as is.

After doing this for all the m coins, we let the first m points be S and the second m points be S'. Again, this is a distributionally identical way of obtaining S and S'. So, want to bound

$$\Pr_{S''} \Pr_{p_1,\ldots,p_m}[h \text{ makes at most } \varepsilon_x m/4 \text{ mistakes on } S, \text{ but at least } \varepsilon_x m/2 \text{ mistakes on } S'].$$

Now, after drawing S'' (and conditioning on it/fixing it), the event in the parentheses depends on how the coins $p_1, \ldots, p_m$ pan out. So, for any fixed S'', we will bound the probability of the event in the parentheses, where the randomness is *only* over the coin flips.

First, we can assume that S'' satisfies: for at most $\varepsilon_x/4$ pairs $(i, m+i)$, h makes a mistake at both x_i and x_{m+i}, and for at least $\varepsilon_x m/2$ pairs $(i, m+i)$, h makes a mistake on at least one of x_i and x_{m+i}. If S'' does not satisfy the first condition, then h makes strictly more than $\varepsilon_x m/4$ mistakes on S. Similarly, if S'' does not satisfy the second condition, then h makes strictly less than $\varepsilon_x m/2$ mistakes on S'. Thus, if S'' does not satisfy either of these conditions, the required event cannot happen.

However, these two conditions together mean that there are at least $\varepsilon_x m/4$ pairs $(i, m+i)$, such that h makes a mistake on *exactly* one of x_i or x_{m+i}. And then, for the event to occur, namely for h to make at least $\varepsilon_x m/2$ mistakes on S', it must be the case that the coin flips at *all* these indices direct the mistake towards S' (namely the second half of S''). The probability of this happening is at most $2^{-\varepsilon_x m/4}$, which is at most $\delta/2\tau_{\mathcal{H}}(2m)$, provided $m \geq \frac{4\log(2\tau_{\mathcal{H}}(2m)/\delta)}{\varepsilon_x}$. Here, $\tau_{\mathcal{H}}$ denotes the *growth function* of $\mathcal{H}$; namely, $\tau_{\mathcal{H}}(2m) = \max_{T \in \mathcal{X}^{2m}} |\mathcal{H}|_T|$.

Thus, we can conclude saying that so long as $m \geq \frac{10\log(2\tau_{\mathcal{H}}(2m)/\delta)}{\varepsilon_x}$,

$$\Pr[A] \leq 2\Pr[B] \leq 2 \sum_{h \in \mathcal{H}_x|_{S''}} \delta/2\tau_{\mathcal{H}}(2m) \leq \tau_{\mathcal{H}}(2m) \cdot \delta/2\tau_{\mathcal{H}}(2m) \leq \delta,$$

as required. That is, for m satisfying this bound, every $h \in \mathcal{H}_x$ makes strictly more than $\varepsilon_x m/4$ mistakes on S. By ensuring the strong condition of $m \geq \max\left\{\frac{4b}{\varepsilon_x}, \frac{10\log(2\tau_{\mathcal{H}}(2m)/\delta)}{\varepsilon_x}\right\}$, we will have that with probability at least $1-\delta$, every $h \in \mathcal{H}_x$ makes strictly more than b mistakes on S, meaning also that $(x, h^\star(x))$ is in the b-robust agreement region of S.

Finally, since the VC dimension of $\mathcal{H}$ is d, the Sauer-Shelah-Perles lemma ensures that $\tau_{\mathcal{H}}(2m) \leq (2me/d)^d$. Plugging this in our bound gives us that $m = O\left(\frac{b+d\log(1/\varepsilon_x)+\log(1/\delta)}{\varepsilon_x}\right)$ is sufficient to ensure that with probability at least $1-\delta$, $(x, h^\star(x))$ is in the b-robust agreement region of $S \sim \mathcal{D}^m$, finishing the proof. ■

A.4. Proof of Theorem 12

We recall that $\varepsilon_x = \varepsilon_x(\mathcal{D}, h^\star)$. By definition of $\varepsilon_x^\star$ (5), there exists a valid reweighting $w : \mathcal{X} \to [0,1]$ and corresponding $\mathcal{D}_w$ which satisfies that $\varepsilon_x^w = \varepsilon_x^w(\mathcal{D}_w, h^\star) \geq \varepsilon_x^\star/2$. Consider the rejection sampling procedure for drawing a sample from $\mathcal{D}_w$: until a sample gets accepted, draw $z \sim \mathcal{D}$, and accept z with probability $w(z)$.

We will show that drawing $\mathrm{poly}(b, d, 1/\varepsilon_x, 1/\delta)$ samples from $\mathcal{D}$ is sufficient to ensure that with probability $1 - \delta/2$, the number of samples accepted is at least $6\left(\frac{b+d\log(1/\varepsilon_x^w)+\log(2/\delta)}{\varepsilon_x^w}\right)$. Conditioned on this event, note that all the accepted samples are distributed according to $\mathcal{D}_w$. Denote the first $m_w = 6\left(\frac{b+d\log(1/\varepsilon_x^w)+\log(2/\delta)}{\varepsilon_x^w}\right)$ of the accepted samples as S'. Then, Theorem 9 immediately gives us that with probability at least $1 - \delta/2$, $(x, h^\star(x))$ belongs to the b-robust agreement region of S'. A union bound over both the $\delta/2$ failure probability events ensures that S' thus obtained is the required certificate from the theorem statement, with probability $1 - \delta$. The size of S' is $m_w = 6\left(\frac{b+d\log(1/\varepsilon_x^w)+\log(2/\delta)}{\varepsilon_x^w}\right) \leq 12\left(\frac{b+d\log(2/\varepsilon_x^\star)+\log(2/\delta)}{\varepsilon_x^\star}\right)$ as required.

So, we continue to argue that if we draw $\mathrm{poly}(b, d, 1/\varepsilon_x, 1/\delta)$ samples from $\mathcal{D}$, then at least $6\left(\frac{b+d\log(1/\varepsilon_x^w)+\log(2/\delta)}{\varepsilon_x^w}\right)$ samples are accepted. Towards this, note that by definition of the sampling process, a sample drawn from $\mathcal{D}$ gets accepted with probability $p = \int_z w(z)\mathcal{D}(z)dz$. Thus, if we draw m samples i.i.d. from $\mathcal{D}$, the expected number of accepted samples is pm. By a Chernoff bound, the probability that less than $pm/2$ samples get accepted is at most $\exp(-pm/12)$, which is at most $\delta/2$ provided $m \geq \frac{12\log(2/\delta)}{p}$. Thus, for m satisfying this condition, we have that with probability at least $1-\delta/2$, at least $pm/2$ samples are accepted.

Consider setting $m = 12 \cdot \mathrm{poly}(b, d, 1/\varepsilon_x)\left(\frac{b+d\log(1/\varepsilon_x^w)+\log(2/\delta)}{\varepsilon_x^w}\right)$. First, notice that this satisfies the condition $m \geq \frac{12\log(2/\delta)}{p}$, since

$$\begin{aligned} 12 \cdot \mathrm{poly}(b, d, 1/\varepsilon_x)\left(\frac{b + d\log(1/\varepsilon_x^w) + \log(2/\delta)}{\varepsilon_x^w}\right) &\geq \frac{12}{p} \cdot \left(\frac{b + d\log(1/\varepsilon_x^w) + \log(2/\delta)}{\varepsilon_x^w}\right) \\ &\geq \frac{12\log(2/\delta)}{p}, \end{aligned}$$

where in the first inequality, we used the condition that w is a valid reweighting (4). Therefore, for this setting of m, with probability at least $1-\delta/2$, the number of accepted samples is at least

$$\begin{aligned} pm/2 &\geq \frac{6}{\mathrm{poly}(b, d, 1/\varepsilon_x)} \cdot \mathrm{poly}(b, d, 1/\varepsilon_x)\left(\frac{b + d\log(1/\varepsilon_x^w) + \log(2/\delta)}{\varepsilon_x^w}\right) \\ &\geq 6\left(\frac{b + d\log(1/\varepsilon_x^w) + \log(2/\delta)}{\varepsilon_x^w}\right), \end{aligned}$$

as required. In the first inequality above, we again used that w is a valid reweighting. To conclude the proof, we note that since $2\varepsilon_x^w \geq \varepsilon_x^\star \geq \varepsilon_x$,

$$\begin{aligned} m &= 12 \cdot \mathrm{poly}(b, d, 1/\varepsilon_x)\left(\frac{b + d\log(1/\varepsilon_x^w) + \log(2/\delta)}{\varepsilon_x^w}\right) \\ &\leq 24 \cdot \mathrm{poly}(b, d, 1/\varepsilon_x)\left(\frac{b + d\log(2/\varepsilon_x^\star) + \log(2/\delta)}{\varepsilon_x^\star}\right) = \mathrm{poly}(b, d, 1/\varepsilon_x, 1/\delta). \end{aligned}$$

■