Reliable learning in challenging environments

Maria-Florina Balcan
Carnegie Mellon University
ninamf@cs.cmu.edu

Steve Hanneke
Purdue University
steve.hanneke@gmail.com

Rattana Pukdee
Carnegie Mellon University
rpukdee@cs.cmu.edu

Dravyansh Sharma
Carnegie Mellon University
dravyans@cs.cmu.edu

Abstract

The problem of designing learners that provide guarantees that their predictions are provably correct is of increasing importance in machine learning. However, learning theoretic guarantees have only been considered in very specific settings. In this work, we consider the design and analysis of reliable learners in challenging test-time environments as encountered in modern machine learning problems: namely 'adversarial' test-time attacks (in several variations) and 'natural' distribution shifts. In this work, we provide a reliable learner with provably optimal guarantees in such settings. We discuss computationally feasible implementations of the learner and further show that our algorithm achieves strong positive performance guarantees on several natural examples: for example, linear separators under log-concave distributions or smooth boundary classifiers under smooth probability distributions.

1 Introduction

The question of providing reliability guarantees on the output of learned classifiers has been studied previously in the classical learning setting where the training and test data are independent and identically distributed (i.i.d.) draws from the same distribution [RS88, EYW10, EYW12]. Conceptually, a *reliable* learner outputs a prediction and may output a correctness guarantee. We know that the learner is correct on all points with the guarantee as long as the learning-theoretic assumptions hold, e.g., realizability. While a trivial model that abstains from providing any guarantee is also a reliable learner, we are interested in a reliable learner that provides the guarantee on as many points as possible (*useful* in the sense of Rivest and Sloan [RS88]). [EYW10] provides a characterization of optimal reliable learners in this classical learning setting.

However, the assumption that the training and test data are drawn from the same distribution is often violated in practice. The mismatch may take the form of a 'natural distribution shift' when the test distribution is different from the training distribution or 'adversarial attacks' when there is an adversary that can perturb a test data point with the goal of changing the model prediction. This is frequently accompanied by a significant performance drop, as well as the inability to guarantee the usefulness of the algorithm. As a result, there is a significant interest in the study of test-time attacks [GSS15, CW17, MMS+18] and distribution shift [LWS18, RRSS19, MTR+21] among the applied machine learning community. Furthermore, recently there has been growing interest in the theoretical machine learning community for designing approaches with provable guarantees under test time attacks [AKM19, MHS19, MHS22] as well as renewed interest in distribution shift [BDBCP06, MMR08, HK19]. All the prior theoretical work in the literature has mainly focused on the effect of attacks or distribution shift on average error rate (e.g. [BDBCP06, AKM19]). However,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

this neglects a major relevant concern for users of machine learning algorithms, namely the ability to provide correctness guarantees for individual predictions: i.e., reliability.

In this work, we advance this line of work by developing a general understanding of how to learn reliably in the presence of corruptions or changes to the test set, specifically under adversarial test-time attacks as well as distribution shift between the training (source) and test (target) data.

Our results. We consider algorithms that provide robustly-reliable predictions which are guaranteed to be correct under standard assumptions from statistical learning theory, for both test-time attacks and distribution shift. Our first main set of results tackles the challenging case of adversarial test-time perturbations. For this setting, we introduce a novel compelling reliability criterion on a learner that particularly captures the challenge of reliability under the test-time attacks. Given a test point z, a *robustly-reliable* classifier either abstains from prediction, or outputs both a prediction y and a reliability guarantee η with the guarantee that y is correct unless one of two bad events has occurred: 1) the true target function does not belong to the given hypothesis set $\mathcal{H}$ or, 2) a test-point z is perturbed from its original point by adversarial strength of at least η (measured in the relevant metric). In the case of distribution shift, we provide novel analysis and a complexity measure that extend the classical notion of reliable learning to the setting when the test distribution is allowed to be an arbitrary new distribution.

1.1 Summary of contributions

1. We propose robustly-reliable learners for test-time attacks which guarantee reliable learning in the presence of test-time attacks, and characterize the region of instance space where they are simultaneously robust and reliable. Specifically, under the realizable setting, for adversarial perturbations within metric balls around the test points, we use the radius of the metric ball as a natural notion of adversarial strength. We output a reliability radius η with a guarantee that our prediction on a point is correct as long as it was perturbed with a distance less than η (under a given metric). We further show that our proposed robustly-reliable learner achieves pointwise optimal values for this reliability radius: that is, no robustly-reliable learner can output a reliability radius larger than our learner for any point in the instance space (Theorem B.1, B.2).
2. The pointwise optimal algorithm is easy to derive from our definition. We discuss a computationally efficient implementation of the optimal learners. (Section 4).
3. We discuss variants of these algorithms and guarantees appropriate for three different variants of adversarial losses studied in the literature: depending on whether the perturbed point must have the same label as the original point, or in lieu of this, whether the algorithm should predict the true label of the perturbed point, or the same label as the original point (Definition 1).
4. We further introduce a safely-reliable region, which captures the challenge caused by the adversary's ability to perturb a test point to cause a reduction in our reliability radius (Definition 6). As examples, we show that the safely-reliable region can be large for linear separators under log-concave distributions and for classifiers with smooth decision boundaries under nearly-uniform distributions and as a consequence, the robustly-reliable region is large as well (Theorem 3.3).
5. We extend this characterization to abstention-based reliable predictions for arbitrary adversarial perturbation sets, where we no longer restrict ourselves to metric balls. We again get a tight characterization of the robustly-reliable region (Theorem C.1).
6. We also consider reliability in the distribution shift setting where the test data points come from a different distribution. We introduce a novel refinement to the notion of disagreement coefficient [Han07], to measure the **transferability of reliability guarantees** across distributions. We provide bounds on the probability mass of the reliable region under transfer for several interesting examples including, when learning linear separators, transfer from β_1 log-concave to β_2 log-concave and to s-concave distributions (Theorems G.1, G.2). We additionally bound the probability of the reliable region for learning classifiers with general smooth classification boundaries, for transfer between smooth distributions (Theorem G.3).
7. We further extend our reliability results to the setting of robustness tranfer, where the test data is simultaneously under adversarial perturbations as well as distribution shift (Section J).
8. Finally, we demonstrate that it is possible extend our results into the agnostic setting. (Section 7)

Conceptual advances over prior work. Prior works on certified robustness [SKL17, CRK19, WLF22] have examined pointwise consistency guarantees. The certified robustness guarantee is only that a prediction does not change with an adversarial perturbation, but it does not guarantee that the

prediction is correct (neither for the original point nor the perturbation); in particular, a constant function is always certified robust but it may not be useful. In contrast, our notion of robustly-reliable learner guarantees that, for any test point x and perturbation z, if z has a distance less than η to x (η = reliability radius), then the prediction will be "correct" (robust loss zero) in a sense informed by which robust loss we are addressing; we discuss this idea for several different losses, leading to different interpretations of this guarantee. In particular, for the stability loss, the prediction being "correct" means that it predicts the true label of the original point x; this implies certified robustness, but is even stronger, since it also guarantees the correct label. Prior work [BBHS22] introduces the notion of a robustly-reliable learner for poisoning attacks which is different from our definition that is tailored to test-time attacks with a guarantee in terms of a reliability radius. In distribution shifts setting, we are the first to assess the transferability of reliability guarantees which differ from a widely-studied metric of average error rate. For additional related work, we refer to Appendix A.

2 Preliminaries and problem formulation

Let $\mathcal{X}$ denote the instance space and $\mathcal{Y} = \{0, 1\}$ be the label space. Let $\mathcal{H}$ be a hypothesis class. The learner $\mathcal{L}$ is given access to a labeled sample $S = \{(x_i, y_i)\}_{i=1}^m$ drawn from a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and learns a concept $h^{\mathcal{L}} : \mathcal{X} \to \mathcal{Y}$. In the realizable setting, we assume we have a hypothesis (concept) class $\mathcal{H}$ and target concept $h^* \in \mathcal{H}$ such that the *true label* of any $x \in \mathcal{X}$ is given by $h^*(x)$. In particular, $S = \{(x_i, h^*(x_i))\}_{i=1}^m$ in this setting. Given the 0-1 loss function $\ell : \mathcal{H} \times \mathcal{X} \to \{0, 1\}$, define $\mathrm{err}_S(h, \ell) = \frac{1}{m} \sum_{(x,y)\in S} \ell(h, x)$. We use $\mathcal{D}_{\mathcal{X}}$ to denote the marginal distribution over $\mathcal{X}$. We use $\mathbb{I}[\cdot]$ to denote the indicator function that takes values in $\{0, 1\}$. We also define $B_{\mathcal{D}}^{\mathcal{H}}(h^*, r) = \{h \in \mathcal{H} \mid \Pr_{\mathcal{D}}[h(x) \neq h^*(x)] \leq r\}$ as the set of hypotheses in $\mathcal{H}$ that disagree with h^* with probability at most r. During test-time, the learner makes a prediction on a test-point $z \in \mathcal{X}$. We consider the following settings

1. **Adversarial test-time attack.** We consider adversarial attacks with perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ that can perturb a test point x to an arbitrary point z from the perturbation set $\mathcal{U}(x) \subseteq \mathcal{X}$ [MHS19]. We assume that the adversary has access to the learned concept $h^{\mathcal{L}}$ as well as the test point x, and can perturb this data point to any $z \in \mathcal{U}(x)$ and then provide this perturbed data point to the learner at test-time. We want to provide pointwise robustness and reliability guarantees in this setting. We will assume that $x \in \mathcal{U}(x)$ for all $x \in \mathcal{X}$. For any point z, we have $\mathcal{U}^{-1}(z) := \{x \in \mathcal{X} | z \in \mathcal{U}(x)\}$, the set of points that can be perturbed to z. We use *perturbation* to refer to a point $z \in \mathcal{U}(x)$ and the perturbation sets $\mathcal{U}(x)$ interchangeably.
2. **Distribution shift.** We consider when a test point z is drawn from a different distribution from the training samples. In this case, we want to provide a pointwise reliability guarantee. We will discuss more on this in Section 5.

2.1 Robust loss functions

In the applied and theoretical literature, various definitions of adversarial success have been explored, each dependent on the interpretation of robustness; depending on whether the perturbed point must have the same label as the original point, or in lieu of this, whether the algorithm should predict the true label of the perturbed point, or the same label as the original point. To capture these, we formally consider the following loss functions.

Definition 1 (Robust loss functions)**.** *For a hypothesis h, a test point x, and a perturbation function $\mathcal{U}$, we consider the following adversarially successful events.*

1. ***Constrained Adversary loss** [SZS+14, BBSZ23]. There exists a perturbation z of x that does not change the true label of an original point x but $h(z)$ is incorrect.*

$$\ell_{CA}^{h^*}(h, x) = \sup_{\substack{z\in\mathcal{U}(x) \\ h^*(z)=h^*(x)}} \mathbb{I}[h(z) \neq h^*(z)].$$

   For a fixed perturbation $z \in \mathcal{U}(x)$, define $\ell_{CA}^{h^*}(h, x, z) = \mathbb{I}[h(z) \neq h^*(z) \wedge h^*(z) = h^*(x)]$.
2. ***True Label loss** [ZL19, GKKW21]. There exists a perturbation z of x such that $h(z)$ is incorrect.*

$$\ell_{TL}^{h^*}(h, x) = \sup_{z\in\mathcal{U}(x)} \mathbb{I}[h(z) \neq h^*(z)].$$

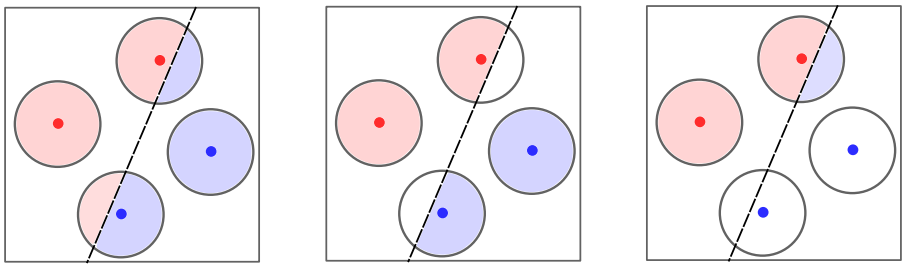

Figure 1: Different perturbation sets considered in $\ell_{\text{TL}}, \ell_{\text{ST}}$ (left), ℓ_{CA} (mid) and ℓ_{IA} (right). The dashed line represents the decision boundary of h^* and the background color of red and blue represents the label 0 and 1 respectively. The ball around each point describes the possible perturbation set $\mathcal{U}(x)$ and the shaded area inside each ball is the allowed perturbation. In $\ell_{\text{TL}}, \ell_{\text{ST}}$, we consider all perturbation in $\mathcal{U}(x)$ while in ℓ_{CA}, we consider perturbations that do not change the true label of the perturbed point. Lastly, in ℓ_{IA}, an adversary only perturb points where the original true label is 0.

In this case, if the true label of the z changes, then the learner need to match its prediction with the new label. For a fixed perturbation $z \in \mathcal{U}(x)$, define $\ell_{TL}^{h^*}(h, x, z) = \mathbb{I}[h(z) \neq h^*(z)]$.

3. ***Stability loss** [AKM19, MHS19, MHS22]. There exists a perturbation z of x such that $h(z)$ is different from $h^*(x)$.*

$$\ell_{ST}^{h^*}(h, x) = \sup_{z \in \mathcal{U}(x)} \mathbb{I}[h(z) \neq h^*(x)].$$

In this case, we focus on the consistency aspect where we want the prediction of any perturbation z to be the same as the prediction of x and this has to be correct w.r.t. x i.e. equals to $h^*(x)$. For a fixed perturbation $z \in \mathcal{U}(x)$, define $\ell_{ST}^{h^*}(h, x, z) = \mathbb{I}[h(z) \neq h^*(x)]$.

4. ***Incentive-aware Adversary loss** [ZC21]. We take inspiration from economics application where we assume that the label 1 is a more favorable outcome e.g. loan approval for which an adversary has no incentive to make any perturbation when the original label is 1. Define the perturbation set*

$$\mathcal{U}_{IA}(x, h^*) = \begin{cases} \mathcal{U}(x) & ; h^*(x) = 0 \\ \{x\} & ; h^*(x) = 1 \end{cases}$$

and define an incentive-aware adversary loss as

$$\ell_{IA}^{h^*}(h, x) = \sup_{z \in \mathcal{U}_{IA}(x, h^*)} \mathbb{I}[h(z) \neq h^*(x)].$$

For a fixed perturbation $z \in \mathcal{U}(x)$, define $\ell_{IA}^{h^*}(h, x, z) = \mathbb{I}[h(z) \neq h^*(x) \wedge z \in \mathcal{U}_{IA}(x, h^*)]$.

We say that h is robust to a perturbation function $\mathcal{U}$ at x w.r.t. a robust loss ℓ if $\ell^{h^*}(h, x) = 0$.

Remark. $\ell_{\text{ST}}, \ell_{\text{IA}}$ are robust losses that we can always evaluate in practice on the training data since we are comparing $h(z)$ with $h^*(x)$ which is known to us on the training data. For $\ell_{\text{CA}}, \ell_{\text{TL}}$, we are comparing $h(z)$ with $h^*(z)$ for which z may lie outside of the support of the natural data distribution and we may not have access to $h^*(z)$. We illustrate the relationship between these losses by making a few useful observations.

- In the robustly-realizable case [MHS20] when the perturbation function $\mathcal{U}$ does not change the *true* label of any x in the training or test data, then all the losses $\ell_{\text{CA}}, \ell_{\text{TL}}, \ell_{\text{ST}}$ are equivalent. This corresponds to a common assumption in the adversarial robustness literature, that the perturbations are "human-imperceptible", which is usually quantified as the set of perturbations within a small metric ball around the data point.
- We provide an illustration of the perturbation set considered in various robust losses in Figure 1. By considering these perturbation set, we have the following implication $\ell_{\text{TL}} \to \ell_{\text{CA}}$, $\ell_{\text{ST}} \to \ell_{\text{CA}}$ and $\ell_{\text{ST}} \to \ell_{\text{IA}}$ where $\ell_1 \to \ell_2$ means robustness w.r.t. ℓ_1 implies robustness w.r.t. ℓ_2.

3 Robustly-reliable learners w.r.t. metric ball attacks

Although our robust losses are defined for any general perturbation set, we first consider the case where the perturbation sets are balls in some metric space. Such attacks are widely studied in the literature, in particular, for balls with bounded L_p-norm. Moreover, the radius of the metric ball serves as a natural notion of adversarial strength that allows us to quantify the level of robustness. We will later (Theorem C.1) present results for general perturbation sets as well.

Let $\mathcal{M} = (\mathcal{X}, d)$ be a metric space equipped with distance metric d. We use the notation $\mathbf{B}_{\mathcal{M}}(x, r) = \{x' \in \mathcal{X} \mid d(x, x') \leq r\}$ (resp. $\mathbf{B}^o_{\mathcal{M}}(x, r) = \{x' \in \mathcal{X} \mid d(x, x') < r\}$) to denote a closed (resp. open) ball of radius r centered at x. We will sometimes omit the underlying metric $\mathcal{M}$ from the subscript to reduce notational clutter. We formally define a metric ball attack as follows.

Definition 2. ***Metric-ball attacks*** *are defined as the class of perturbation functions $\mathcal{U}_{\mathcal{M}} = \{u_\eta : \mathcal{X} \to 2^{\mathcal{X}} \mid u_\eta(x) = \mathbf{B}_{\mathcal{M}}(x, \eta)\}$, induced by the metric $\mathcal{M} = (\mathcal{X}, d)$ defined over the instance space.*

At test-time, given a test-point $z \in \mathcal{X}$, we would like to make a prediction at z with a reliability guarantee. We consider this type of learner, a *robustly-reliable* learner defined formally as follows.

Definition 3 (Robustly-reliable learner w.r.t. $\mathcal{M}$-ball attacks)**.** *A learner $\mathcal{L}$ is robustly-reliable w.r.t. $\mathcal{M}$-ball attacks for hypothesis space $\mathcal{H}$ and robust loss function ℓ if,* ***for any target*** *concept $h^* \in \mathcal{H}$, given S labeled by h^*, the learner outputs functions $h^{\mathcal{L}}_S : \mathcal{X} \to \mathcal{Y}$ and $r^{\mathcal{L}}_S : \mathcal{X} \to [0, \infty) \cup \{-1\}$ such that for all $x, z \in \mathcal{X}$ if $r^{\mathcal{L}}_S(z) = \eta > 0$ and $z \in \mathbf{B}^o_{\mathcal{M}}(x, \eta)$ then $\ell^{h^*}(h^{\mathcal{L}}_S, x, z) = 0$. Further, if $r^{\mathcal{L}}_S(z) = 0$, then $h^*(z) = h^{\mathcal{L}}_S(z)$.*

Note that $\mathcal{L}$ outputs a prediction and a real value r (the "reliability radius") for any test input. $r = -1$ corresponds to abstention (even in the absence of perturbation) i.e. when the learner is incapable of giving a reliability guarantee for that prediction), and $r = \eta > 0$ is a guarantee from the learner that if the adversary's attack is in $\mathbf{B}^o_{\mathcal{M}}(x, \eta)$ then we are correct i.e. if an adversary changes the original test point x to z, the attack will not succeed if the adversarial budget is less than η. Lastly, when $r = 0$, the learner provides a guarantee that the learner's prediction at z is correct.

Definition 4 (Robustly-reliable region w.r.t. $\mathcal{M}$-ball attacks)**.** *For a robustly-reliable learner $\mathcal{L}$ w.r.t. $\mathcal{M}$-ball attacks for sample S, hypothesis space $\mathcal{H}$ and robust loss function ℓ defined above, the robustly-reliable region of $\mathcal{L}$ at a reliability level η is defined as $RR^{\mathcal{L}}(S, \eta) = \{x \in \mathcal{X} \mid r^{\mathcal{L}}_S(x) \geq \eta\}$ for sample S and $\eta \geq 0$.*

The robustly-reliable region contains all points with a reliability guarantee of at least η. We use $\mathrm{RR}^{\mathcal{L}}_W$ to denote robustly-reliable regions with respect to losses ℓ_W for $W \in \{\mathrm{CA}, \mathrm{TL}, \mathrm{ST}, \mathrm{IA}\}$. A natural goal is to find a robustly-reliable learner $\mathcal{L}$ that has the largest robustly-reliable region possible. First, we note that predictions that are known by the learner to be correct are still known to be correct even when the test points are attacked. Therefore, a test point z lies in the robustly-reliable region w.r.t. $\ell_{\mathrm{CA}}, \ell_{\mathrm{TL}}$, as long as we can be sure that $h^{\mathcal{L}}_S(z)$ is correct. This is equivalent to z being classified perfectly, i.e. according to the true label. Therefore, the robustly-reliable region w.r.t. $\ell_{\mathrm{CA}}, \ell_{\mathrm{TL}}$ is given by the agreement region of the version space, which is the largest region where we can be sure of what the correct label is in the absence of any adversarial attack [EYW10]. We recall the definition of version space [Mit82] and agreement region [CAL94, BBL06].

Definition 5. *For a set $H \subseteq \mathcal{H}$ of hypothesis, and any set of samples S, let $\mathrm{DIS}(H) = \{x \in \mathcal{X} : \exists h_1, h_2 \in \mathcal{H}$ s.t. $h_1(x) \neq h_2(x)\}$ be the* ***disagreement region*** *and $\mathrm{Agree}(H) = \mathcal{X} \setminus \mathrm{DIS}(H)$ be the* ***agreement region****. Let $\mathcal{H}_0(S) = \{h \in \mathcal{H} | \, \mathrm{err}_S(h) = 0\}$ be a* ***version space****: the set of all hypotheses that correctly classify S. More generally, $\mathcal{H}_\nu(S) = \{h \in \mathcal{H} | \, \mathrm{err}_S(h) \leq \nu\}$ for $\nu \geq 0$.*

We can also characterize the robustly-reliable region with respect to other robust losses in terms of the agreement region in the following Theorem.

Theorem 3.1. *Let $\mathcal{H}$ be any hypothesis class. With respect to $\mathcal{M}$-ball attacks and ℓ_W, for $\eta \geq 0$,*

(a) there exists a robustly-reliable learner $\mathcal{L}$ such that $RR^{\mathcal{L}}_W(S, \eta) \supseteq A_W$, and
(b) for any robustly-reliable learner $\mathcal{L}$, $RR^{\mathcal{L}}_W(S, \eta) \subseteq A_W$.

Specifically, for the robust loss ℓ_W, the optimal robustly-reliable region A_W are

Robust loss ℓ_W	**Optimal robustly-reliable region A_W**
$\ell_{\text{CA}}, \ell_{\text{TL}}$	$\{z \mid z \in \text{Agree}(\mathcal{H}_0(S))\}$
ℓ_{ST}	$\{z \mid \mathbf{B}^o(z,\eta) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(z) = h(x), \forall x \in \mathbf{B}^o(z,\eta), \forall h \in \mathcal{H}_0(S)\}$
ℓ_{IA}	$(A_{\text{ST}} \cap \{z \mid h^*(z) = 1\}) \cup \{z \mid z \in \text{Agree}(\mathcal{H}_0(S)) \wedge h^*(z) = 0\}$

Proof. (Sketch) We provide the construction of the optimal robustly-reliable learner $\mathcal{L}_{\text{opt}}$ such that $\text{RR}_W^{\mathcal{L}_{\text{opt}}}(S,\eta) \supseteq A_W$ and later show that for any robustly-reliable learner $\mathcal{L}$, we must also have $\text{RR}_W^{\mathcal{L}}(S,\eta) \subseteq A_W$. We start with $\ell_{\text{CA}}, \ell_{\text{TL}}$, consider a learner $\mathcal{L}_{\text{opt}}$ that predicts using an ERM classifier and outputs $\eta = \infty$ for all points in the agreement region of $\mathcal{H}_0(S)$. Any prediction in $\text{Agree}(\mathcal{H}_0(S))$ is reliable because it also agrees with h^* ($h^* \in \mathcal{H}_0(S)$ by realizability). On the other hand, for $z \in \text{DIS}(\mathcal{H}_0(S))$, there exist $h_1, h_2 \in \mathcal{H}_0(S)$ that disagree on z. For any learner $\mathcal{L}$, it is not possible to guarantee that $h^{\mathcal{L}}(z)$ is correct as we may have $h^* = h_1$ or $h^* = h_2$.

Now, for ℓ_{ST}, ca The first condition guarantees that $h(x) = h^*(x), \forall x \in \mathbf{B}^o(z,\eta)$. Combined with the second condition we have $h(z) = h(x) = h^*(x), \forall x \in \mathbf{B}^o(z,\eta)$. Thus, $\mathcal{L}_{\text{opt}}$ is a robustly-reliable learner. On the other hand, for a robustly-reliable learner $\mathcal{L}$, consider $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S,\eta)$ for $\eta > 0$. We must have $h^{\mathcal{L}}(z) = h^*(x), \forall x \in \mathbf{B}^o(z,\eta)$. Using a similar argument to the case for $\ell_{\text{CA}}, \ell_{\text{TL}}$, we have $z \in \text{Agree}(\mathcal{H}_0(S))$. If there exists $x \in \mathbf{B}^o(z,\eta)$ that $x \notin \text{Agree}(\mathcal{H}_0(S))$, there exists $h_1, h_2 \in \mathcal{H}_0(S)$ that $h^{\mathcal{L}}(z) \neq h_1(x)$ or $h^{\mathcal{L}}(z) \neq h_2(x)$. It is not possible to guarantee that $h^{\mathcal{L}}(z) = h^*(x)$ as we may have $h^* = h_1$ or $h^* = h_2$. Therefore, we must have $\mathbf{B}^o(z,\eta) \subseteq \text{Agree}(\mathcal{H}_0(S))$. Finally, we cannot have $x \in \mathbf{B}^o(z,\eta)$ that $h(z) = h^*(z) \neq h^*(x)$ since this contradict with $h(z) = h^*(x)$. Therefore, we must have $h^*(x) = h^*(z)$. Since we have $x \in \text{Agree}(\mathcal{H}_0(S))$, this implies that $h(x) = h(z)$ for $h \in \mathcal{H}_0(S)$. For ℓ_{IA}, the construction is similar to ℓ_{ST}. For full proof, we refer to Appendix B. □

For ℓ_{ST}, the learner is able to certify a subset of the agreement region, which satisfies two additional conditions: $h_S^{\mathcal{L}}$ must be correct on all possible points x that could be perturbed to an observed test point z, and the true label of z should match the true label of x. We denote the second condition of $h(z) = h(x), \forall x \in \mathbf{B}^o(z,\eta), \forall h \in \mathcal{H}_0(S)$ as the **label consistency condition**. For ℓ_{IA}, since robustness w.r.t. ℓ_{ST} implies robustness w.r.t. ℓ_{IA}, we have $A_{\text{ST}} \subset A_{\text{IA}}$. In addition, with an incentive-aware adversary, whenever $h(z) = 0$, we must have $h^*(x) = 0$ since the adversary does not perturb a data point with the original label 1. Therefore, we can additionally provide a guarantee on z that lies in the agreement region and $h(z) = 0$. We remark that we can identify the term $h^*(z)$ for any point $z \in \text{Agree}(\mathcal{H}_0(S))$ due to the realizability assumption.

We have provided a robustly-reliable learner with the largest possible robustly-reliable region for losses $\ell_{\text{CA}}, \ell_{\text{TL}}, \ell_{\text{ST}}, \ell_{\text{IA}}$. However, we note that the probability mass of the robustly-reliable region may not be a meaningful way to quantify the overall reliability of a learner because a perturbation z may lie outside of the support of the natural data distribution and have zero probability mass. It seems more useful to measure the mass of points x where any perturbation z of x still lies within the robustly-reliable region. We formally define this region as the safely-reliable region.

Definition 6 (Safely-reliable region w.r.t. $\mathcal{M}$-ball attacks)**.** *Let $\mathcal{L}$ be a robustly-reliable learner w.r.t. $\mathcal{M}$-ball attacks for sample S, hypothesis class $\mathcal{H}$ and robust loss function ℓ. The safely-reliable region of learner $\mathcal{L}$ at reliability levels η_1, η_2 is defined as*

1. $SR_{CA}^{\mathcal{L}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1) \cap \{z \mid h^*(z) = h^*(x)\} \subseteq RR_{CA}^{\mathcal{L}}(S,\eta_2)\}$,
2. $SR_{W}^{\mathcal{L}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1) \subseteq RR_{W}^{\mathcal{L}}(S,\eta_2)\}$ *for* $W \in \{TL, ST\}$,
3. $SR_{IA}^{\mathcal{L}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid h^*(x) = 0 \wedge \mathbf{B}_{\mathcal{M}}(x,\eta_1) \subseteq RR_{IA}^{\mathcal{L}}(S,\eta_2)\} \cup \{x \in \mathcal{X} \mid h^*(x) = 1 \wedge x \in RR_{IA}^{\mathcal{L}}(S,\eta_2))\}$.

The safely-reliable region contains any point that retains a reliability radius of at least η_2 even after being attacked by an adversary with strength η_1. In the safely-reliable region, we consider a set of potential natural (before attack) points x, while in the robustly-reliable region, we consider a set of potential test points z. In the following subsections, we show that in interesting cases commonly studied in the literature, the probability mass of the safely-reliable region is actually quite large.

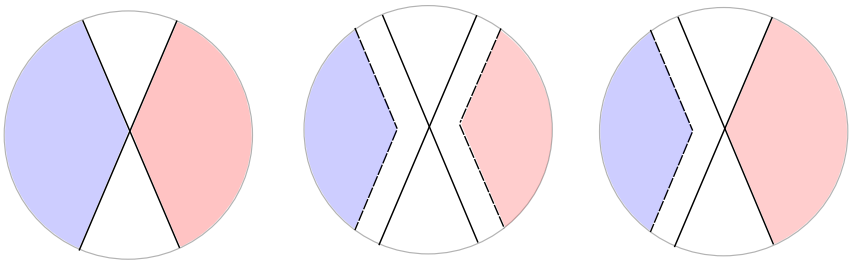

Figure 2: The robustly-reliable region for $\ell_{\text{CA}}, \ell_{\text{TL}}$ (left), ℓ_{ST} (mid) and ℓ_{IA}(right) for linear separators with an L_2-ball perturbation. The background color of blue and red represents the agreement region of class 1 and 0 respectively. In this case, we can remove the label consistency condition and reduce the robustly-reliable region into the 'η-buffered' agreement region.

3.1 Safely-reliable region for linear separators under log-concave distributions is large

We provide the probability mass of safely-reliable regions with respect to different losses for linear separators when the data distribution follows an isotropic (mean zero and identity covariance matrix) log-concave (logarithm of density function is concave) distribution under a bounded L_2-norm ball attack. For full proof, we refer to Appendix D. We will rely on the following key lemma which states that the agreement region of a linear separator cannot contain points that are arbitrarily close to the decision boundary of h^* for any sample S.

Lemma 3.2. *Let $\mathcal{D}$ be a distribution over $\mathbb{R}^d$ and $\mathcal{H} = \{h : x \to \text{sign}(\langle w_h, x\rangle) \mid w_h \in \mathbb{R}^d, \|w_h\|_2 = 1\}$ be a class of linear separators. For $h^* \in \mathcal{H}$, for a set of samples $S \sim \mathcal{D}^m$ such that there is no data point in S that lies on the decision boundary, for any $0 < c < d$, there exists $\delta(S, c, d) > 0$ such that for any x with $c \le \|x\| \le d$ and $|\langle w_{h^*}, x\rangle| < \delta$, we have $x \notin \text{Agree}(\mathcal{H}_0(S))$.*

A direct implication of the lemma is that any L_2-ball $\mathbf{B}(x, \eta)$ that lies in the agreement region must not contain the decision boundary of h^* and must contain points with the same label. This allows us to remove the *label consistency condition* and instead focus on whether the ball $\mathbf{B}(x, \eta)$ lies in the agreement region. Intuitively, the reliable region is now given by the 'η-buffered' agreement region where we only select points that have a distance at least η from the boundary of the agreement region (Figure 2). We provide bounds for the probability mass of the safely-reliable region below and we refer to the full proof in Appendix D.

Theorem 3.3. *Let $\mathcal{D}$ be isotropic log-concave over $\mathbb{R}^d$ and $\mathcal{H} = \{h : x \to \text{sign}(\langle w_h, x\rangle) \mid w_h \in \mathbb{R}^d, \|w_h\|_2 = 1\}$ be the class of linear separators. Let $\mathbf{B}(\cdot, \eta)$ be a L_2 ball perturbation with radius η. For $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\text{VCdim}(\mathcal{H}) + \ln\frac{1}{\delta}))$, for an optimal robustly-reliable learner $\mathcal{L}$,*

(a) $\Pr(SR^{\mathcal{L}}_{TL}(S, \eta_1, \eta_2)) \ge 1 - 2\eta_1 - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability at least* $1 - \delta$,
(b) $SR^{\mathcal{L}}_{CA}(S, \eta_1, \eta_2) = SR^{\mathcal{L}}_{TL}(S, \eta_1, \eta_2)$ *almost surely,*
(c) $\Pr(SR^{\mathcal{L}}_{ST}(S, \eta_1, \eta_2)) \ge 1 - 2(\eta_1 + \eta_2) - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability at least* $1 - \delta$,
(d) $\Pr(SR^{\mathcal{L}}_{IA}(S, \eta_1, \eta_2)) \ge 1 - (\eta_1 + \eta_2) - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability at least* $1 - \delta$.

The $\tilde{\mathcal{O}}$-notation suppresses dependence on logarithmic factors and distribution-specific constants.

We remark that we can't always remove the label consistency condition for a general perturbation set. For example, consider $\mathcal{U}(x) = \mathbf{B}(x - a, \eta) \cup \{x\} \cup \mathbf{B}(x + a, \eta)$, is made of two L_2 balls with center $x - a, x + a$, with appropriate value of a, η, we may have each ball lie in the different side of the agreement region so that the whole perturbation set lie in the agreement region but contain points with different labels (Figure 3). We also provide bounds on the probability mass of the safely-reliable region for more general concept spaces beyond linear separators, specifically, classifiers with smooth boundaries in Appendix E.

4 On computational efficiency

Given the definition, it is possible to implement computationally efficient robustly-reliable learners. For example, for linear separator concept classes under bounded L_2-norm attack. The optimal

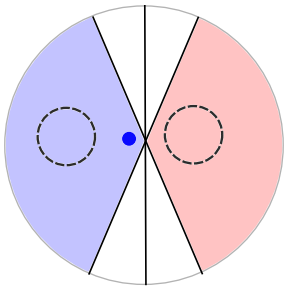

Figure 3: The perturbation set is represented by two dashed balls. This lies inside the agreement region but contains points with different labels.

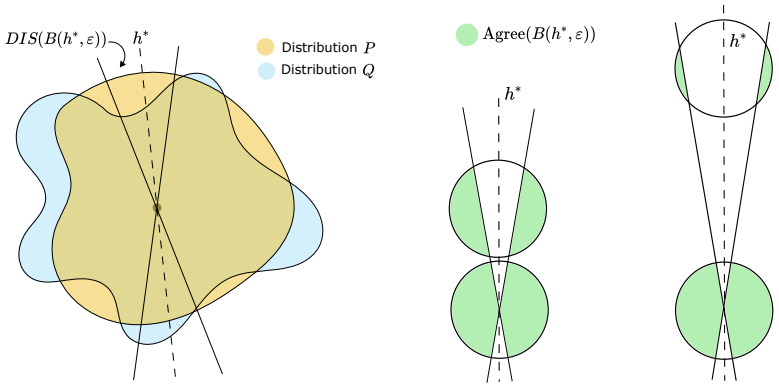

Figure 4: The disagreement region and the agreement region under a distribution shift where $\mathcal{P}$ and $\mathcal{Q}$ are isotropic (left) and where there is a mean shift (right).

robustly-reliable learner $\mathcal{L}_{\text{opt}}$, described above may be implemented via a linearly constrained quadratic program that computes the (squared) distance of the test point z to the closest point z' in the disagreement region. This gives us the reliability radius, since for linear separators one must cross the decision boundary to perturb a point to a differently labeled point

$$\begin{aligned} & \min_{w,w',z'} ||z-z'||^2 & \\ \text{s.t.} \quad & \langle w, x_i \rangle y_i \geq 0, & \text{for each } (x_i, y_i) \in S, \\ & \langle w', x_i \rangle y_i \geq 0, & \text{for each } (x_i, y_i) \in S, \\ & \langle w, z' \rangle \langle w', z' \rangle \leq 0. & \end{aligned}$$

Given training sample S, for any given test point z, the learner $\mathcal{L}$ can efficiently compute the solution s^* to the above program and output $\sqrt{s^*}$ as the reliability radius. We show that the variant of this objective also provides a reliability radius for a wide range of hypothesis classes under L_2 ball attacks (see Lemma F.1). In addition, we can relax this objective into a regularized objective that gives a lower bound on the reliability radius of $||z - z^*||^2$, when z^* is the solution of

$$h_1, h_2, z^* = \operatorname*{argmin}_{h,h',z'} ||z-z'||^2 + \lambda(\hat{R}(h, S \cup \{(z', 0)\}) + \hat{R}(h', S \cup \{(z', 1)\}))$$

when $\hat{R}(h, S)$ is the empirical risk of h on S. We provide a more detailed discussion in Appendix F.

5 Robustly-reliable learning under distribution shift

We now consider the reliability aspect for distribution shift, a different kind of test-time robustness challenge when the test data comes from a different distribution than the training data. Formally, let $\mathcal{P}$ be the training distribution and let $\mathcal{Q}$ be the test distribution. We assume the *realizable distribution shift* setting, i.e. there is a target concept $h^* \in \mathcal{H}$ such that the true label of any $x \in \mathcal{X}$ is given by $h^*(x)$ at training time and test time, or $\text{err}_{\mathcal{P}}(h^*) = \text{err}_{\mathcal{Q}}(h^*) = 0$. As observed earlier, points that

are known by the learner to be correct (reliable) are still known to be correct even when it is drawn from a different distribution. This reliability guarantee holds even when the distributions $\mathcal{P}$ and $\mathcal{Q}$ do not share a common support, a setting for which many prior theoretical works result in vacuous bounds. For example, suppose $\mathcal{X} = \mathbb{R}^n$, $\mathcal{P}$ and $\mathcal{Q}$ are supported on disjoint n-balls, and $\mathcal{H}$ is the class of linear separators. Then the total variation distance, the $\mathcal{H}$-divergence [KBDG04, BDBC+10] as well as the discrepancy distance [MMR08] between $\mathcal{P}$ and $\mathcal{Q}$ are all 1. While recent work of [HK19] does apply in this setting, they do not focus on the reliability guarantee. In this work, we are interested in quantifying the transferability of reliability guarantee transfer between distributions $\mathcal{P}$ and $\mathcal{Q}$. We recall the notion of reliable prediction [EYW10].

Definition 7 (Reliability)**.** *A learner $\mathcal{L}$ is reliable w.r.t. concept space $\mathcal{H}$ if, for any target concept $h^* \in \mathcal{H}$, given any sample S labeled by h^*, the learner outputs functions $h_S^{\mathcal{L}} : \mathcal{X} \to \mathcal{Y}$ and $a_S^{\mathcal{L}} : \mathcal{X} \to \{0, 1\}$ such that for all $x \in \mathcal{X}$ if $a_S^{\mathcal{L}}(x) = 1$ then $h_S^{\mathcal{L}}(x) = h^*(x)$. Else, if $a_S^{\mathcal{L}}(x) = 0$, the learner abstains from prediction. The reliable region of $\mathcal{L}$ is $R^{\mathcal{L}}(S) = \{x \in \mathcal{X} \mid a_S^{\mathcal{L}}(x) = 1\}$.*

We define the following metric to measure the reliability of a learner under distribution shift.

Definition 8 ($\mathcal{P}\to\mathcal{Q}$ reliable correctness)**.** *The $\mathcal{P}\to\mathcal{Q}$-reliable correctness of $\mathcal{L}$ (at sample rate m, for distribution shift from $\mathcal{P}$ to $\mathcal{Q}$) is defined as the expected probability mass of its reliable region under Q when trained on a random training $S \sim \mathcal{P}^m$, i.e. $\Pr_{x\sim Q, S\sim P^m}[x \in R^{\mathcal{L}}(S)]$.*

The disagreement coefficient was originally introduced to study the label complexity in agnostic active learning [Han07] and is also known to characterize reliable learning in the absence of any distribution shift [EYW10]. We propose the following refinement to the notion of disagreement coefficient, which we will use to give bounds on a learner's $\mathcal{P}\to\mathcal{Q}$ reliable correctness.

Definition 9 ($\mathcal{P}\to\mathcal{Q}$ disagreement coefficient)**.** *For a hypothesis class $\mathcal{H}$, the $\mathcal{P}\to\mathcal{Q}$ disagreement coefficient of $h^* \in \mathcal{H}$ with respect to $\mathcal{H}$ is given by*

$$\Theta_{\mathcal{P}\to\mathcal{Q}}(\varepsilon) = \sup_{r\geq\varepsilon} \frac{\Pr_{\mathcal{Q}}[\mathrm{DIS}(B_{\mathcal{P}}(h^*, r))]}{r},$$

where $B_{\mathcal{P}}(h^*, r) = \{h \in \mathcal{H} \mid \Pr_{\mathcal{P}}[h(x) \neq h^*(x)] \leq r\}$.

This quantifies the rate of disagreement over $\mathcal{Q}$ among classifiers which are within disagreement-balls w.r.t. h^* under $\mathcal{P}$, relative to the version space radius. The proposed metric is asymmetric between $\mathcal{P}$ and $\mathcal{Q}$, and also depends on the target concept h^*. More simple examples are in Appendix H. We show that the $\mathcal{P}\to\mathcal{Q}$-reliable correctness of our learner may be bounded in terms of the $\mathcal{P}\to\mathcal{Q}$ disagreement coefficient using a uniform convergence based argument. The proof details are in Appendix I.

Theorem 5.1. *Let $\mathcal{Q}$ be a realizable distribution shift of $\mathcal{P}$ with respect to $\mathcal{H}$, and $h^* \in \mathcal{H}$ be the target concept. Given sufficiently large sample size $m \geq \frac{c}{\varepsilon^2}(d + \ln\frac{1}{\delta})$, the $\mathcal{P}\to\mathcal{Q}$-reliable correctness of $\mathcal{L}$, the optimal robustly-reliable learner, is at least*

$$\Pr_{x\sim Q, S\sim P^m}[x \in R^{\mathcal{L}}(S)] \geq 1 - \Theta_{\mathcal{P}\to\mathcal{Q}} \cdot \varepsilon - \delta.$$

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

In Appendix J, we show that this $\mathcal{P} \to \mathcal{Q}$ disagreement coefficient can be small for several examples which implies that it is possible to transfer the reliability guarantee from one distribution to the other. In particular, when learning linear separators, we provide bounds for transferring from β_1 log-concave to β_2 log-concave and to s-concave distributions (Theorems G.1, G.2). In addition, when learning classifiers with general smooth classification boundaries, we provide bounds for transferring between smooth distributions (Theorem G.3).

6 Safely-reliable correctness under distribution shift

There is a growing practical [SSZ+20, SIE+20] as well as recent theoretical interest [DGH+23] in the setting of 'robustness transfer', where one simultaneously expects adversarial test-time attacks as well as distribution shift. We will study the reliability aspect for this more challenging setting. We note that the definition of a robustly-reliable learner does not depend on the data distribution (see

Definition 3) as the guarantee is pointwise. Our optimality result in Section 3 applies even when a test point is drawn from a different distribution $\mathcal{Q}$. In this case, the safely-reliable region instead would have a different probability mass.

Definition 10 ($\mathcal{P}\to\mathcal{Q}$ safely-reliable correctness)**.** *The $\mathcal{P}\to\mathcal{Q}$ safely-reliable correctness of $\mathcal{L}$ (at sample rate m, for distribution shift from $\mathcal{P}$ to $\mathcal{Q}$, w.r.t. robust loss ℓ) is defined as the probability mass of its safely-reliable region under Q, on a sample $S\sim\mathcal{P}^m$, i.e.*

$$PQR_\ell^{\mathcal{L}}(S,\eta_1,\eta_2) := \Pr_{x\sim Q, S\sim P^m}[x\in SR_\ell^{\mathcal{L}}(S,\eta_1,\eta_2)].$$

We consider an example when the training distribution $\mathcal{P}$ is isotropic log-concave and the test distribution $\mathcal{Q}_\mu$ is log-concave with its mean shifted by μ but the covariance matrix is still an identity matrix (see Figure 4, right). We provide the bound on the $\mathcal{P}\to\mathcal{Q}$ safely-reliable correctness of this example in Appendix J (see Theorem J.2).

7 Reliability in the agnostic setting

In the above, we have assumed that the training samples S are realizable under our concept class $\mathcal{H}$, i.e. there is a target concept h^* consistent with our (uncorrupted) data. In the agnostic setting, we can have $\min_{h\in\mathcal{H}}\mathrm{err}_S(h)>0$, meaning no single concept is always correct. We define a *ν-tolerably robustly-reliable* learner under test-time attacks in the agnostic setting as the learner whose reliable predictions agree with every low-error hypothesis (error at most ν) on the training sample ([BBHS22] have proposed the corresponding definition for data poisoning attacks).

Definition 11 (ν-tolerably robustly-reliable learner w.r.t. $\mathcal{M}$-ball attacks)**.** *A learner $\mathcal{L}$ is robustly-reliable w.r.t. $\mathcal{M}$-ball attacks for sample S, hypothesis space $\mathcal{H}$ and robust loss function ℓ if, for **every** concept $h^*\in\mathcal{H}$ with $\mathrm{err}_S(h^*)\le\nu$, the learner outputs functions $h_S^{\mathcal{L}}:\mathcal{X}\to\mathcal{Y}$ and $r_S^{\mathcal{L}}:\mathcal{X}\to[0,\infty)\cup\{-1\}$ such that for all $x,z\in\mathcal{X}$ if $r_S^{\mathcal{L}}(z)=\eta>0$ and $z\in\mathbf{B}_{\mathcal{M}}^o(x,\eta)$ then $\ell^{h^*}(h_S^{\mathcal{L}},x,z)=0$. Further, if $r_S^{\mathcal{L}}(z)=0$, then $h^*(z)=h_S^{\mathcal{L}}(z)$. Given sample S such that some concept $h^*\in\mathcal{H}$ satisfies $\mathrm{err}_S(h^*)\le\nu$, the robustly-reliable region of $\mathcal{L}$ is defined as $RR^{\mathcal{L}}(S,\nu,\eta)=\{x\in\mathcal{X}\mid r_S^{\mathcal{L}}(x)\ge\eta\}$ for $\nu,\eta\ge0$.*

We generalize our results from Section 3 to the agnostic setting (proof details in Appendix K). Here $\mathcal{H}_\nu(S)=\{h\in\mathcal{H}\mid\mathrm{err}_S(h)\le\nu\}$.

Theorem 7.1. *Let $\mathcal{H}$ be any hypothesis class. With respect to $\mathcal{M}$-ball attacks and ℓ_{CA}, for $\eta\ge0$,*

(a) There exists a robustly-reliable learner $\mathcal{L}$ such that $RR_{CA}^{\mathcal{L}}(S,\nu,\eta)\supseteq\mathrm{Agree}(\mathcal{H}_\nu(S))$,
(b) For any robustly-reliable learner $\mathcal{L}$, $RR_{CA}^{\mathcal{L}}(S,\nu,\eta)\subseteq\mathrm{Agree}(\mathcal{H}_\nu(S))$.

8 Discussion

In this work, we generalize the classical line of works on reliable learning to address challenging test-time environments. We propose a novel robustly-reliability criterion that is applicable to several variations of test-time attacks. Our analysis leads to an easy-to-derive algorithm that can be implemented efficiently in many cases. Additionally, we introduce a $\mathcal{P}\to\mathcal{Q}$ disagreement coefficient to capture the transferability of the reliability guarantee between distributions. The proposed robustly-reliability criterion and the $\mathcal{P}\to\mathcal{Q}$ disagreement coefficient together provide a comprehensive framework for handling test-time attacks and evaluating the reliability of learning models. This contributes to the advancement of reliable learning methodologies in the face of challenging real-world scenarios, facilitating the development of more resilient and trustworthy machine learning systems. Notably, key questions remain open, including, how to efficiently implement the algorithm for a class of neural networks, and how to learn reliably with respect to any general robust loss function?

9 Acknowledgements

This work was supported in part by NSF grants CCF-1910321 and SES-1919453, the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003, a Bloomberg Data Science PhD fellowship, and a Simons Investigator Award.

References

[AB99] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 1999.

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[ABHU15] 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.

[ABL14] 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.

[AK91] David Applegate and Ravi Kannan. Sampling and integration of near log-concave functions. In *Proceedings of the twenty-third annual ACM Symposium on Theory of Computing*, pages 156–163, 1991.

[AKM19] Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR, 2019.

[BBHS22] Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dravyansh Sharma. Robustly-reliable learners under poisoning attacks. In *Conference on Learning Theory*. PMLR, 2022.

[BBL06] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 65–72, 2006.

[BBS07] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, pages 81–88, 2007.

[BBSZ23] Maria-Florina Balcan, Avrim Blum, Dravyansh Sharma, and Hongyang Zhang. On the power of abstention and data-driven decision making for adversarial robustness. *Journal of Machine Learning Research*, 2023.

[BCKW15] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[BDBC+10] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[BDBCP06] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19, 2006.

[BH21] Maria-Florina Balcan and Nika Haghtalab. Noise in classification. *Beyond the Worst-Case Analysis of Algorithms*, page 361, 2021.

[BL13] 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.

[BPRZ23] Maria-Florina Balcan, Rattana Pukdee, Pradeep Ravikumar, and Hongyang Zhang. Nash equilibria and pitfalls of adversarial training in adversarial robustness games. In *The 26nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

[BZ17] Maria-Florina Balcan and Hongyang Zhang. Sample and computationally efficient learning algorithms under s-concave distributions. *Advances in Neural Information Processing Systems*, 30, 2017.

[CAL94] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[CDG+18] Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *International Conference on Machine Learning*, pages 1059–1067. PMLR, 2018.

[CDM16] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. *Advances in Neural Information Processing Systems*, 29, 2016.

[CRK19] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[CRS+19] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[CW17] 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.

[CWG+19] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019.

[Dan16] Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the forty-eighth annual ACM Symposium on Theory of Computing*, pages 105–117, 2016.

[DGH+23] Yuyang Deng, Nidham Gazagnadou, Junyuan Hong, Mehrdad Mahdavi, and Lingjuan Lyu. On the hardness of robustness transfer: A perspective from Rademacher complexity over symmetric difference hypothesis space. *arXiv preprint arXiv:2302.12351*, 2023.

[DGT19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent pac learning of halfspaces with massart noise. *Advances in Neural Information Processing Systems*, 32, 2019.

[DN21] John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

[EYW10] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

[EYW12] Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *Journal of Machine Learning Research*, 13(2), 2012.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[GKKM20] 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.

[GKKW21] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. *The Journal of Machine Learning Research*, 22(1):12521–12549, 2021.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[GSS15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

[Han07] 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.

[HGB+06] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19, 2006.

[HK19] Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[HKLM20] Max Hopkins, Daniel Kane, Shachar Lovett, and Gaurav Mahajan. Noise-tolerant, reliable active classification with comparison queries. In *Conference on Learning Theory*, pages 1957–2006. PMLR, 2020.

[JSH20] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

[KBDG04] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *VLDB*, volume 4, pages 180–191. Toronto, Canada, 2004.

[KKMS08] Adam Tauman Kalai, Adam R Klivans, Yishay Mansour, and Rocco A Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KV94] Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

[LAG+19] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy*, pages 656–672. IEEE, 2019.

[LCWC19] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in Neural Information Processing Systems*, 32, 2019.

[LPB17] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[LV07] László Lovász and Santosh Vempala. The geometry of logconcave functions and sampling algorithms. *Random Structures & Algorithms*, 30(3):307–358, 2007.

[LWS18] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pages 3122–3130. PMLR, 2018.

[MHS19] 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.

[MHS20] Omar Montasser, Steve Hanneke, and Nati Srebro. Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33:14626–14637, 2020.

[MHS21] Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pages 3452–3482. PMLR, 2021.

[MHS22] Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. *Advances in Neural Information Processing Systems*, 2022.

[MIG+19] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[Mit82] Tom M Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[MMR08] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems*, 21, 2008.

[MMS+18] 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.

[MTR+21] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[ND16] Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in Neural Information Processing Systems*, 29, 2016.

[PY10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[PZ21] Nikita Puchkin and Nikita Zhivotovskiy. Exponential savings in agnostic active learning through abstention. In *Conference on Learning Theory*, pages 3806–3832. PMLR, 2021.

[QCSSL08] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2008.

[RRR21] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

[RRSS19] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

[RS88] Ronald L Rivest and Robert H Sloan. Learning complicated concepts reliably and usefully. In *Association for the Advancement of Artificial Intelligence*, pages 635–640, 1988.

[RWK20] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

[RXY+20] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7909–7919, 2020.

[Shi00] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[SIE+20] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[SKHL20] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

[SKL17] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in Neural Information Processing Systems*, 30, 2017.

[SNG+19] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

[SS16] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[SST+18] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 31, 2018.

[SSZ+20] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations*, 2020.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[TKP+18] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

[TS10] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[TSE+19] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[Vap98] Vladimir N Vapnik. *Statistical learning theory*. Wiley-Interscience, 1998.

[vdVW96] Aad W van der Vaart and Jon A Wellner. *Weak convergence*. Springer, 1996.

[Wan11] Liwei Wang. Smoothness, disagreement coefficient, and the label complexity of agnostic active learning. *Journal of Machine Learning Research*, 12(7), 2011.

[WLF22] Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, pages 22769–22783. PMLR, 2022.

[WR06] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[YCJ16] Songbai Yan, Kamalika Chaudhuri, and Tara Javidi. Active learning from imperfect labelers. *Advances in Neural Information Processing Systems*, 29, 2016.

[ZC21] Hanrui Zhang and Vincent Conitzer. Incentive-aware pac learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5797–5804, 2021.

[ZL19] Yuchen Zhang and Percy Liang. Defending against whitebox adversarial attacks via randomized discretization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 684–693. PMLR, 2019.

[ZLLJ19] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019.

[ZLWJ20] Yuchen Zhang, Mingsheng Long, Jianmin Wang, and Michael I Jordan. On localized discrepancy for domain adaptation. *arXiv preprint arXiv:2008.06242*, 2020.

[ZN22] Yinglun Zhu and Robert D Nowak. Efficient active learning with abstention. In *Advances in Neural Information Processing Systems*, 2022.

[ZYJ+19] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

A Additional related work

Reliability. A learning model that outputs a confidence level is valuable in practical applications, as it allows us to determine when to trust the model and when to defer the task to a human. However, it is well-known that models like neural networks can exhibit high confidence, yet still produce incorrect results [GPSW17]. To tackle this issue, there has been a line of works on learning algorithms with uncertainty estimate [WR06, BCKW15, GG16, LPB17, MIG+19]. Unlike prior work, our results take into account the relevant notion of robust loss. In particular, we extend the reliability guarantees in perfect selective classification [EYW12] and reliable-useful learning model [RS88] to different robust losses under a test-time attack. Prior work on reliability under data poisoning attacks [BBHS22] obtained similar results on training-time attacks, by providing guarantees that the learner is always correct at any point that it makes a prediction provided the training data corruption does not exceed a point-specific threshold. Our work is also related to learning algorithms with an abstention option [YCJ16, CDM16, CDG+18, PZ21, ZN22].

Robustness. Robustness against adversarial attacks is essential for the safe deployment of machine learning models in the real world. Our focus in this work is on perturbation attacks, where we aim to provide learners that remain robust even when the test data points are perturbed. It is known that many modern approaches such as deep neural networks fail in this case even when the perturbation is human-imperceptible [SZS+14, GSS15]. There has been a lot of empirical effort [MMS+18, ZYJ+19, SNG+19, RWK20, TKP+18, CRS+19] as well as theoretical effort [TSE+19, SST+18, JSH20, RXY+20, MHS19, MHS21, GKKM20, BPRZ23] to develop learners with improved robustness, and more broadly to understand various aspects of adversarial robustness. In particular, there is a line of work on certified robustness [CRK19, LAG+19, LCWC19] which provides a pointwise guarantee that the prediction does not change, so long as the attack strength is within a learner-specified 'radius' for the point. While the certified robustness research focuses on this consistency aspect, our work addresses the reliability aspect where we hope to guarantee that the prediction is also correct.

Distribution shift. A distribution shift refers to the phenomenon where the training distribution differs from the test distribution which often leads to a degradation in the learner's performance. This has been studied under several different settings [QCSSL08], ranging from covariate shift [Shi00, HGB+06, BBS07], and domain adaptation [MMR08, BDBC+10, ZLLJ19, ZLWJ20] to transfer learning [PY10, TS10, HK19]. Many algorithms have been proposed to deal with the shift which involves encouraging invariance between different domains [SS16, ABGLP19, RRR21] or taking into account the worst-case subpopulation [ND16, DN21, SKHL20, CWG+19]. While prior work typically focuses on the average performance on the target domain or subpopulation, we provide point-wise reliability guarantees.

Learning with noise. There is extensive classic literature on learning methods which are tolerant or robust to noise [KV94, Vap98]—including efficient learning under bounded or Massart noise [ABHU15, DGT19], agnostic active learning [KKMS08, Dan16], learning under malicious noise [ABL14, BH21], to list a few. Recent work has considered reliable learning under some of these classic noise models [HKLM20, BBHS22].

B Additional proof details for robustly-reliable learners w.r.t. metric ball attacks

Theorem B.1. *Let $\mathcal{H}$ be any hypothesis class. With respect to $\mathcal{M}$-ball attacks and ℓ_{CA}, for $\eta \geq 0$,*

(a) there exists a robustly-reliable learner $\mathcal{L}$ such that $RR^{\mathcal{L}}_{CA}(S, \eta) \supseteq \mathrm{Agree}(\mathcal{H}_0(S))$, and
(b) for any robustly-reliable learner $\mathcal{L}$, $RR^{\mathcal{L}}_{CA}(S, \eta) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))$.

The results hold for $RR^{\mathcal{L}}_{TL}$ as well.

Proof. (Proof of Theorem B.1) The robustly-reliable learner $\mathcal{L}$ is given as follows. Set $h^{\mathcal{L}}_S = \mathrm{argmin}_{h\in\mathcal{H}} \mathrm{err}_S(h)$ i.e. an ERM over S, and $r^{\mathcal{L}}_S(z) = \infty$ if $z \in \mathrm{Agree}(\mathcal{H}_0(S))$, else $r^{\mathcal{L}}_S(z) = -1$. By realizability, $\mathrm{err}_S(h^{\mathcal{L}}_S) \leq \mathrm{err}_S(h^*) = 0$, or $h^{\mathcal{L}}_S \in \mathcal{H}_0(S)$. We first show that $\mathcal{L}$ is robustly-reliable. For $z \in \mathcal{X}$, if $r^{\mathcal{L}}_S(z) = \eta > 0$, then $z \in \mathrm{Agree}(\mathcal{H}_0(S))$. We have $h^*(z) = h^{\mathcal{L}}_S(z)$ since the classifiers

$h^*, h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$ and z lies in the agreement region of classifiers in $\mathcal{H}_0(S)$ in this case. Thus, we have $\ell_{\text{CA}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathbf{B}_{\mathcal{M}}^o(x, \eta)$. The $\eta = 0$ case corresponds to reliability in the absence of test-time attack, so [EYW10] applies. Therefore, $\text{RR}_{\text{CA}}^{\mathcal{L}}(S, \eta) \supseteq \text{Agree}(\mathcal{H}_0(S))$ for all $\eta \geq 0$ follows from the setting $r_S^{\mathcal{L}}(z) = \infty$ if $z \in \text{Agree}(\mathcal{H}_0(S))$.

Conversely, let $z \in \text{DIS}(\mathcal{H}_0(S))$. There exist $h_1, h_2 \in \mathcal{H}_0(S)$ such that $h_1(z) \neq h_2(z)$. If possible, let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{CA}}^{\mathcal{L}}(S, \eta)$ for some $\eta > 0$. By definition of the robust-reliability region, we must have $r_S^{\mathcal{L}}(z) > 0$. By definition of a ball, we have $z \in \mathbf{B}_{\mathcal{M}}^o(z, \eta)$ for any $\eta > 0$, and therefore $\ell_{\text{CA}}^{h^*}(h_S^{\mathcal{L}}, z, z) = 0$. But then we must have $h_S^{\mathcal{L}}(z) = h^*(z)$ by definition of ℓ_{CA}. But we can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. But $h_1(z) \neq h_2(z)$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. □

Theorem B.2. *Let $\mathcal{H}$ be any hypothesis class. With respect to $\mathcal{M}$-ball attacks and ℓ_{ST}, for $\eta \geq 0$,*

(a) there exists a robustly reliable learner $\mathcal{L}$ such that $RR_{ST}^{\mathcal{L}}(S, \eta) \supseteq A_{ST}$, and
(b) for any robustly-reliable learner $\mathcal{L}$, $RR_{ST}^{\mathcal{L}}(S, \eta) \subseteq A_{ST}$,

where $A_{ST} = \{z \mid \mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge \forall h \in \mathcal{H}_0(S), h(x) = h(z), \forall x \in \mathbf{B}^o(z, \eta)\}$.

Proof. (Proof of Theorem B.2) Given sample S, consider the learner $\mathcal{L}$ which outputs $h_S^{\mathcal{L}} = \text{argmin}_{h \in \mathcal{H}} \text{err}_S(h)$, and $r_S^{\mathcal{L}}(z)$ is given by the largest $\eta > 0$ for which $\mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))$ and $h(x) = h(z), \ \forall \ x \in \mathbf{B}^o(z, \eta), h \in \mathcal{H}_0(S)$, else $\eta = 0$ if $z \in \text{Agree}(\mathcal{H}_0(S))$, and -1 otherwise. Note that the supremum exists here since a union of open sets is also open. By realizability, $\text{err}_S(h_S^{\mathcal{L}}) \leq \text{err}_S(h^*) = 0$, or $h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$. We first show that $\mathcal{L}$ is robustly-reliable w.r.t. $\mathcal{M}$ for loss ℓ_{ST}. For $z \in \mathcal{X}$, if $r_S^{\mathcal{L}}(z) = \eta \geq 0$, then $\mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))$, in particular $z \in \text{Agree}(\mathcal{H}_0(S))$. Moreover, by definition, for any $x \in \mathbf{B}^o(z, \eta)$, we have $h_S^{\mathcal{L}}(z) = h_S^{\mathcal{L}}(x)$ by construction. Putting together, and using the property that distance functions of a metric are symmetric, we have $h_S^{\mathcal{L}}(z) = h^*(x)$ for any x such that $z \in \mathbf{B}^o(x, \eta)$. Thus, we have $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathbf{B}^o(x, \eta)$. Thus $\mathcal{L}$ satisfies Definition 3.

Conversely, we will show that for any robustly-reliable learner $\mathcal{L}$ w.r.t. ℓ_{ST}, for any $\eta > 0$,

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)),$$

which follows from similar arguments from Theorem B.1 which also apply to the ℓ_{ST} loss. Let $z \in \text{DIS}(H_0(S))$. There exists $h_1, h_2 \in H_0(S)$ such that $h_1(z) \neq h_2(z)$. If possible, let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta)$ for some $\eta > 0$. By definition of the robust-reliability region, we must have $r_S^{\mathcal{L}}(z) > 0$. By definition of a closed ball, we have $z \in \mathbf{B}_{\mathcal{M}}^o(z, \eta)$ for any $\eta > 0$, and therefore $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, z, z) = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h^*(z)] = 0$ which implies that $h_S^{\mathcal{L}}(z) = h^*(z)$. But we can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. But $h_1(z) \neq h_2(z)$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. Next, we will show that, for any $\eta > 0$,

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta) \subseteq \{z \mid \mathbf{B}_{\mathcal{M}}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))\}.$$

We will prove this by contradiction. Suppose $z \in \text{Agree}(H_0(S))$, but there exists $x' \in \mathbf{B}_{\mathcal{M}}^o(z, \eta)$ such that $x' \notin \text{Agree}(H_0(S))$. Let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta)$. By definition, we have $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x that $z \in \mathbf{B}_{\mathcal{M}}^o(x, \eta)$. This implies that $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x', z) = 0$ that is $h_S^{\mathcal{L}}(z) = h^*(x')$. Because $x' \notin \text{Agree}(H_0(S))$, there exists $h_1, h_2 \in H_0(S)$ such that $h_1(x') \neq h_2(x')$. We can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. But $h_1(x') \neq h_2(x')$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. Finally, we will show that, for any $\eta > 0$,

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta) \subseteq \{z \mid \mathbf{B}_{\mathcal{M}}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = h(z), \ \forall \ x \in \mathbf{B}_{\mathcal{M}}^o(z, \eta), h \in \mathcal{H}_0(S)\}.$$

Let z be a data point such that $\mathbf{B}_{\mathcal{M}}^o(z, \eta) \subseteq \text{Agree}(H_0(S))$ but there exists $x' \in \mathbf{B}_{\mathcal{M}}^o(z, \eta)$ such that $h(x') \neq h(z)$ for $h \in H_0(S)$. Let there be a robustly-reliable learner such that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S, \eta)$. This implies that $\ell_{\text{ST}}(h_S^{\mathcal{L}}, x, z) = 0$ for any x that $z \in \mathbf{B}_{\mathcal{M}}^o(z, \eta)$. However, $\ell_{\text{ST}}(h_S^{\mathcal{L}}, x', z) = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h^*(x')] = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h_S^{\mathcal{L}}(x')] \neq 0$, contradicting that $\mathcal{L}$ is robustly-reliable. □

Theorem B.3. *Let $\mathcal{H}$ be any hypothesis class. With respect to $\mathcal{M}$-ball attacks and ℓ_{IA}, for $\eta \geq 0$,*

(a) there exists a robustly reliable learner $\mathcal{L}$ such that $RR^{\mathcal{L}}_{IA}(S, \eta) \supseteq A_{IA}$, and
(b) for any robustly-reliable learner $\mathcal{L}$, $RR^{\mathcal{L}}_{IA}(S, \eta) \subseteq A_{IA}$,

where $A_{IA} = (A_{ST} \cap \{z \mid h^*(z) = 1\}) \cup \{z \mid z \in \text{Agree}(\mathcal{H}_0(S)) \wedge h^*(z) = 0\}$.

Proof. (Proof of Theorem B.3) The construction of the robustly-reliable learner for ℓ_{IA} is similar to the robustly-reliable learner for ℓ_{ST}. The key difference is that the reliability radius now depends on the predicted label. Given sample S, consider the learner $\mathcal{L}$ which outputs $h^{\mathcal{L}}_S = \text{argmin}_{h \in \mathcal{H}} \text{err}_S(h)$.

1. If $h^{\mathcal{L}}_S(z) = 1$, $r^{\mathcal{L}}_S(z)$ is given by the largest $\eta > 0$ for which $\mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))$ and $h(x) = h(z),\ \forall\ x \in \mathbf{B}^o(z, \eta), h \in \mathcal{H}_0(S)$ and $\eta = 0$ when $z \in \text{Agree}(\mathcal{H}_0(S))$, and -1 otherwise. Note that the supremum exists here since a union of open sets is also open.
2. If $h^{\mathcal{L}}_S(z) = 0$, $r^{\mathcal{L}}_S(z) = \infty$ when $z \in \text{Agree}(\mathcal{H}_0(S))$, and -1 otherwise.

We first show that $\mathcal{L}$ is robustly-reliable w.r.t. $\mathcal{M}$ for loss ℓ_{IA}. By realizability, $\text{err}_S(h^{\mathcal{L}}_S) \leq \text{err}_S(h^*) = 0$, or $h^{\mathcal{L}}_S \in \mathcal{H}_0(S)$. For $z \in \mathcal{X}$, if $h^{\mathcal{L}}_S(z) = 1$ and $r^{\mathcal{L}}_S(z) = \eta \geq 0$, then $\mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))$, in particular $z \in \text{Agree}(\mathcal{H}_0(S))$. Moreover, by definition, for any $x \in \mathbf{B}^o(z, \eta)$, we have $h^{\mathcal{L}}_S(z) = h^{\mathcal{L}}_S(x)$ by construction. Putting together, and using the property that distance functions of a metric are symmetric, we have $h^{\mathcal{L}}_S(z) = h^*(x)$ for any x such that $z \in \mathbf{B}^o(x, \eta)$. Thus, we have $\ell^{h^*}_{\text{ST}}(h^{\mathcal{L}}_S, x, z) = 0$ for any x such that $z \in \mathbf{B}^o(x, \eta)$. This also implies that $\ell^{h^*}_{\text{IA}}(h^{\mathcal{L}}_S, x, z) = 0$ since ℓ_{ST} implies ℓ_{IA}.

On the other hand, if $h^{\mathcal{L}}_S(z) = 0$ and $r^{\mathcal{L}}_S(z) = \infty$, we have $z \in \text{Agree}(\mathcal{H}_0(S))$. This implies that $h^{\mathcal{L}}_S(z) = h^*(z) = 0$. For any x that $z \in \mathcal{U}_{\text{IA}}(x, h^*)$, by the incentive-aware property of the adversary, if $h^*(x) = 1$, we must have $\mathcal{U}_{\text{IA}}(x, h^*) = \{x\}$ which implies that $z = x$ and $h^*(z) = h^*(x) = 1$. In our case, $h^*(z) = 0$ also implies that we must also have $h^*(x) = 0$. Therefore, $\ell^{h^*}_{\text{IA}}(h^{\mathcal{L}}_S, x, z) = \mathbb{I}[h^{\mathcal{L}}_S(z) \neq h^*(x) \wedge z \in \mathcal{U}_{\text{IA}}(x, h^*)] \leq \mathbb{I}[h^{\mathcal{L}}_S(z) \neq h^*(x)] = 0$. Therefore, we can conclude that $\mathcal{L}$ satisfies Definition 3.

Conversely, we will show that for any robustly-reliable learner $\mathcal{L}$ w.r.t. ℓ_{IA}, for any $\eta > 0$,

$$\text{RR}^{\mathcal{L}}_{\text{IA}}(S, \eta) \cap \{z \mid h^{\mathcal{L}}_S(z) = 0\} \subseteq \text{Agree}(\mathcal{H}_0(S)) \cap \{z \mid h^*(z) = 0\}.$$

Since $z \in \mathcal{U}_{\text{IA}}(z)$, for z to lie in the robustly-reliable region, we need $\ell^{h^*}_{\text{IA}}(h^{\mathcal{L}}_S, z, z) = \mathbb{I}[h^{\mathcal{L}}_S(z) \neq h^*(z)] = 0$ that is z must be reliable. By similar arguments from Theorem B.1, we have the result.

Next, we will show that, for any $\eta > 0$,

$$\text{RR}^{\mathcal{L}}_{\text{IA}}(S, \eta) \cap \{z \mid h^{\mathcal{L}}_S(z) = 1\} \subseteq \{z \mid \mathbf{B}^o_{\mathcal{M}}(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))\}.$$

We will prove this by contradiction. Suppose $z \in \text{Agree}(H_0(S))$, but there exists $x' \in \mathbf{B}^o_{\mathcal{M}}(z, \eta)$ such that $x' \notin \text{Agree}(H_0(S))$. Let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}^{\mathcal{L}}_{\text{IA}}(S, \eta)$ and $h^{\mathcal{L}}_S(z) = 1$. Because $x' \notin \text{Agree}(H_0(S))$, there exists $h_1 \in H_0(S)$ such that $h_1(x') = 0$. We may have $h^* = h_1$ since h_1 is consistent with S. However, we have $\mathcal{U}_{\text{IA}}(x', h_1) = \mathbf{B}^o_{\mathcal{M}}(x', \eta)$ and

$$\ell^{h^*}_{\text{IA}}(h^{\mathcal{L}}_S, x', z) = \mathbb{I}[h(z) \neq h^*(x') \wedge z \in \mathcal{U}_{\text{IA}}(x')] = 1$$

which contradicts with z lies in the robustly-reliable region. Furthermore, with a similar argument that we can't have $h^*(x') = 0$, we can show that the agreed label of any x must be 1,

$$\begin{aligned}
&\text{RR}^{\mathcal{L}}_{\text{IA}}(S, \eta) \cap \{z \mid h^{\mathcal{L}}_S(z) = 1\} \\
&\subseteq \{z \mid \mathbf{B}^o_{\mathcal{M}}(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = 1,\ \forall\ x \in \mathbf{B}^o_{\mathcal{M}}(z, \eta), h \in \mathcal{H}_0(S)\}. \\
&= A_{\text{ST}} \cap \{z \mid h^*(z) = 1\}
\end{aligned}$$

This concludes that for any robustly-reliable learner $\mathcal{L}$ with respect to ℓ_{IA}, we have

$$\text{RR}^{\mathcal{L}}_{\text{IA}}(S, \eta) \subseteq (A_{\text{ST}} \cap \{z \mid h^*(z) = 1\}) \cup \{z \mid z \in \text{Agree}(\mathcal{H}_0(S)) \wedge h^*(z) = 0\}.$$

□

C General robustly-reliable learner

Definition 12 (General robustly-reliable learner)**.** *A learner $\mathcal{L}$ is robustly-reliable for sample S w.r.t. a perturbation function $\mathcal{U}$, concept space $\mathcal{H}$ and robust loss function ℓ if, for any target concept $h^* \in \mathcal{H}$, given S labeled by h^*, the learner outputs functions $h_S^{\mathcal{L}} : \mathcal{X} \to \mathcal{Y}$ and $a_S^{\mathcal{L}} : \mathcal{X} \to \{0,1\}$ such that for all $z \in \mathcal{X}$ if $a_S^{\mathcal{L}}(z) = 1$ and $z \in \mathcal{U}(x)$ then $\ell^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$. On the other hand, if $a_S^{\mathcal{L}}(z) = 0$, our learner abstains from prediction. The robustly-reliable region of a learner $\mathcal{L}$ is defined as $RR^{\mathcal{L}}(S) = \{x \in \mathcal{X} \mid a_S^{\mathcal{L}}(x) = 1\}$, the region that the learner $\mathcal{L}$ does not abstain.*

We again obtain the pointwise optimal characterization of the robustly-reliable region in terms of the agreement region. For $\ell_{\text{CA}}, \ell_{\text{TL}}$ the robustly-reliable region would be the same as the region where we can be sure of what the correct label is: i.e. the agreement region of the version space while for ℓ_{ST}, it is the region of points z for which $\mathcal{U}^{-1}(z)$ lies inside the agreement region of the version space, and all classifiers in the version space agree on $\mathcal{U}^{-1}(z)$.

Theorem C.1. *Let $\mathcal{H}$ be any hypothesis class, and $\mathcal{U}$ be the perturbation function.*

(a) There exists a robustly-reliable learner $\mathcal{L}$ w.r.t. $\mathcal{U}$ and ℓ_{CA} such that $RR_{CA}^{\mathcal{L}}(S) \supseteq \text{Agree}(\mathcal{H}_0(S))$. Moreover, for any robustly-reliable learner $\mathcal{L}$, $RR_{CA}^{\mathcal{L}}(S) \subseteq \text{Agree}(\mathcal{H}_0(S))$.

(b) The same results hold for $RR_{TL}^{\mathcal{L}}$ as well.

(c) There exists a robustly-reliable learner $\mathcal{L}$ w.r.t. $\mathcal{U}$ and ℓ_{ST}, such that $RR_{ST}^{\mathcal{L}}(S) \supseteq A_{ST}$, and for any $\mathcal{L}$ robustly-reliable w.r.t. ℓ_{ST}, $RR_{ST}^{\mathcal{L}}(S) \subseteq A_{ST}$, where $A_{ST} = \{z \mid \mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = h(z),\ \forall\, x \in \mathcal{U}^{-1}(z), h \in \mathcal{H}_0(S)\}$.

(d) There exists a robustly-reliable learner $\mathcal{L}$ w.r.t. $\mathcal{U}$ and ℓ_{IA}, such that $RR_{IA}^{\mathcal{L}}(S) \supseteq A_{IA}$, and for any $\mathcal{L}$ robustly-reliable w.r.t. ℓ_{IA}, $RR_{IA}^{\mathcal{L}}(S) \subseteq A_{IA}$, where $A_{IA} = (A_{ST} \cap \{z \mid h^*(z) = 1\}) \cup \{z \mid z \in \text{Agree}(\mathcal{H}_0(S)) \wedge h^*(z) = 0\}$.

Proof. We first establish part (a). Given sample S, consider the learner $\mathcal{L}$ which outputs $h_S^{\mathcal{L}} = \text{argmin}_{h \in \mathcal{H}} \text{err}_S(h)$ i.e. an ERM over S, and $a_S^{\mathcal{L}}(z) = \mathbb{I}[z \in \text{Agree}(\mathcal{H}_0(S))]$. By realizability, $\text{err}_S(h_S^{\mathcal{L}}) \leq \text{err}_S(h^*) = 0$, or $h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$. We first show that $\mathcal{L}$ is robustly-reliable. For $z \in \mathcal{X}$, if $a_S^{\mathcal{L}}(z) = 1$, then $z \in \text{Agree}(\mathcal{H}_0(S))$. We have $h^*(z) = h_S^{\mathcal{L}}(z)$ since the classifiers $h^*, h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$ and z lies in the agreement region of classifiers in $\mathcal{H}_0(S)$. Thus, we have $\ell_{\text{CA}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathcal{U}(x)$. $\text{RR}_{\text{CA}}^{\mathcal{L}}(S) \supseteq \text{Agree}(\mathcal{H}_0(S))$ follows from the choice of $a_S^{\mathcal{L}}(z) = \mathbb{I}[z \in \text{Agree}(\mathcal{H}_0(S))]$.

On the other hand, Let $z \in \text{DIS}(H_0(S))$. There exist $h_1, h_2 \in H_0(S)$ such that $h_1(z) \neq h_2(z)$. If possible, let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{CA}}^{\mathcal{L}}(S)$. That is, $a_S^{\mathcal{L}}(z) = 1$. We have $z \in \mathcal{U}(z)$, and therefore $\ell_{\text{CA}}^{h^*}(h_S^{\mathcal{L}}, z, z) = 0$. But then we must have $h_S^{\mathcal{L}}(z) = h^*(z)$ by definition of ℓ_{CA}. We can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. However, $h_1(z) \neq h_2(z)$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable.

This completes the proof of (a). Essentially the same argument may be used to establish (b), by substituting ℓ_{CA} with ℓ_{TL}. We will now turn our attention to part (c).

Given sample S, consider the learner $\mathcal{L}$ which outputs $h_S^{\mathcal{L}} = \text{argmin}_{h \in \mathcal{H}} \text{err}_S(h)$, that is an ERM over S, and $a_S^{\mathcal{L}}(z) = \mathbb{I}[\mathcal{U}^{-1}(z) \in \text{Agree}(\mathcal{H}_0(S)) \wedge h_S^{\mathcal{L}}(x_1) = h_S^{\mathcal{L}}(x_2)\ \forall\ x_1, x_2 \in \mathcal{U}^{-1}(z)]$. By realizability, $\text{err}_S(h_S^{\mathcal{L}}) \leq \text{err}_S(h^*) = 0$, or $h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$. We first show that $\mathcal{L}$ is robustly-reliable w.r.t. ℓ_{ST}. For $z \in \mathcal{X}$, if $a_S^{\mathcal{L}}(z) = 1$, then $\mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S))$, in particular $z \in \text{Agree}(\mathcal{H}_0(S))$. Moreover, by definition, for any x that $z \in U(x)$, we have $h_S^{\mathcal{L}}(z) = h_S^{\mathcal{L}}(x)$ by construction of $a_S^{\mathcal{L}}(z)$. Putting together, we have $h_S^{\mathcal{L}}(z) = h^*(x)$ for any x such that $z \in \mathcal{U}(x)$. Thus, we have $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathcal{U}(x)$.

On the other hand, for any robustly-reliable learner $\mathcal{L}$, we will show that

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S) \subseteq \text{Agree}(\mathcal{H}_0(S)).$$

Let $z \in \text{DIS}(H_0(S))$. There exists $h_1, h_2 \in H_0(S)$ such that $h_1(z) \neq h_2(z)$. If possible, let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S)$. That is, $a_S^{\mathcal{L}}(z) = 1$. We have $z \in \mathcal{U}^{-1}(z)$,

and therefore $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, z, z) = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h^*(z)] = 0$ which implies that $h_S^{\mathcal{L}}(z) = h^*(z)$. But we can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. However, $h_1(z) \neq h_2(z)$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. Next, we will show that

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S) \subseteq \{z \mid \mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S))\}.$$

We will prove this by contradiction, $z \in \text{Agree}(H_0(S))$ but there exists $x' \in \mathcal{U}^{-1}(z)$ such that $x' \notin \text{Agree}(H_0(S))$. Let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S)$. By definition, we have $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x that $z \in \mathcal{U}(x)$. This implies that $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x', z) = 0$ that is $h_S^{\mathcal{L}}(z) = h^*(x')$. Because $x' \notin \text{Agree}(H_0(S))$, there exists $h_1, h_2 \in H_0(S)$ such that $h_1(x') \neq h_2(x')$. We can set $h^* = h_1$ or $h^* = h_2$ since both are consistent with S. But $h_1(x') \neq h_2(x')$, and therefore $h_S^{\mathcal{L}}(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. Next, we will show that

$$\text{RR}_{\text{ST}}^{\mathcal{L}}(S) \subseteq \{z \mid \mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = h(z),\ \forall\, x \in \mathcal{U}^{-1}(z), h \in \mathcal{H}_0(S)\}.$$

Let z be a data point that $\mathcal{U}^{-1}(z) \subseteq \text{Agree}(H_0(S))$ but there exists $x' \in \mathcal{U}^{-1}(z)$ that $h(x') \neq h(z)$ for $h \in H_0(S)$. Let there be a robustly-reliable learner that $z \in \text{RR}_{\text{ST}}^{\mathcal{L}}(S)$. This implies that $\ell_{\text{ST}}(h_S^{\mathcal{L}}, x, z) = 0$ for any x that $z \in \mathcal{U}(x)$. However, $\ell_{\text{ST}}(h_S^{\mathcal{L}}, x', z) = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h^*(x')] = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h_S^{\mathcal{L}}(x')] \neq 0$, contradicting that $\mathcal{L}$ is robustly-reliable.

Finally, the proof of part d) is similar to the proof of part c). Given sample S, consider the learner $\mathcal{L}$ which outputs $h_S^{\mathcal{L}} = \text{argmin}_{h \in \mathcal{H}} \text{err}_S(h)$, that is an ERM over S, and

1. if $h_S^{\mathcal{L}}(z) = 1$, let $a_S^{\mathcal{L}}(z) = \mathbb{I}[\mathcal{U}^{-1}(z) \in \text{Agree}(\mathcal{H}_0(S)) \wedge h_S^{\mathcal{L}}(x_1) = h_S^{\mathcal{L}}(x_2)\ \forall\ x_1, x_2 \in \mathcal{U}^{-1}(z)]$;
2. if $h_S^{\mathcal{L}}(z) = 0$, let $a_S^{\mathcal{L}}(z) = \mathbb{I}[z \in \text{Agree}(\mathcal{H}_0(S))]$.

By realizability, $\text{err}_S(h_S^{\mathcal{L}}) \leq \text{err}_S(h^*) = 0$, or $h_S^{\mathcal{L}} \in \mathcal{H}_0(S)$. We first show that $\mathcal{L}$ is robustly-reliable w.r.t. ℓ_{IA}. For $z \in \mathcal{X}$, if $h_S^{\mathcal{L}}(z) = 1$ and $a_S^{\mathcal{L}}(z) = 1$, then $\mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S))$, in particular $z \in \text{Agree}(\mathcal{H}_0(S))$. Moreover, by definition, for any x that $z \in U(x)$, we have $h_S^{\mathcal{L}}(z) = h_S^{\mathcal{L}}(x)$ by construction of $a_S^{\mathcal{L}}(z)$. Putting together, we have $h_S^{\mathcal{L}}(z) = h^*(x)$ for any x such that $z \in \mathcal{U}(x)$. Thus, we have $\ell_{\text{ST}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathcal{U}(x)$. This also implies that $\ell_{\text{IA}}^{h^*}(h_S^{\mathcal{L}}, x, z) = 0$ for any x such that $z \in \mathcal{U}(x)$ since ℓ_{ST} implies ℓ_{IA}.

For $z \in \mathcal{X}$, if $h_S^{\mathcal{L}}(z) = 0$ and $a_S^{\mathcal{L}}(z) = 1$, we have $z \in \text{Agree}(\mathcal{H}_0(S))$ and $h_S^{\mathcal{L}}(z) = h^*(z) = 0$. By the incentive-aware property of the adversary, any x such that $z \in \mathcal{U}(x)$, we can't have $h^*(x) = 1$ since the adversary has no incentive to make any perturbation in this case. Therefore, we have $h^*(x) = 0$ and $\mathcal{U}_{\text{IA}}(h^*, x) = \mathcal{U}(x)$. We have $\ell_{\text{IA}}^{h^*}(h_S^{\mathcal{L}}, x, z) = \mathbb{I}[h(z) \neq h^*(x) \wedge z \in \mathcal{U}_{\text{IA}}(x, h^*)] = 0$. We can conclude that our learner $\mathcal{L}$ is robustly-reliable w.r.t. ℓ_{IA}.

Conversely, we will show that for any robustly-reliable learner $\mathcal{L}$ w.r.t. ℓ_{IA}, for any $\eta > 0$,

$$\text{RR}_{\text{IA}}^{\mathcal{L}}(S) \cap \{z \mid h_S^{\mathcal{L}}(z) = 0\} \subseteq \text{Agree}(\mathcal{H}_0(S)) \cap \{z \mid h^*(z) = 0\}.$$

Since $z \in \mathcal{U}_{\text{IA}}(z)$, for z to lie in the robustly-reliable region, we need $\ell_{\text{IA}}^{h^*}(h_S^{\mathcal{L}}, z, z) = \mathbb{I}[h_S^{\mathcal{L}}(z) \neq h^*(z)] = 0$ that is z must be reliable. By similar arguments from above, we have the result. Next, we will show that,

$$\text{RR}_{\text{IA}}^{\mathcal{L}}(S) \cap \{z \mid h_S^{\mathcal{L}}(z) = 1\} \subseteq \{z \mid \mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S))\}.$$

We will prove this by contradiction. Suppose $z \in \text{Agree}(H_0(S))$, but there exists $x' \in \mathcal{U}^{-1}(z)$ such that $x' \notin \text{Agree}(H_0(S))$. Let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \text{RR}_{\text{IA}}^{\mathcal{L}}(S)$ and $h_S^{\mathcal{L}}(z) = 1$. Because $x' \notin \text{Agree}(H_0(S))$, there exists $h_1 \in H_0(S)$ such that $h_1(x') = 0$. We may have $h^* = h_1$ since h_1 is consistent with S. However, we have $\mathcal{U}_{\text{IA}}(x', h_1) = \mathcal{U}(x')$ and

$$\ell_{\text{IA}}^{h^*}(h_S^{\mathcal{L}}, x', z) = \mathbb{I}[h(z) \neq h^*(x') \wedge z \in \mathcal{U}_{\text{IA}}(x')] = 1$$

which contradicts with z lies in the robustly-reliable region. Furthermore, with a similar argument that we can't have $h^*(x') = 0$, we can show that the agreed label of any x must be 1,

$$\begin{aligned}
&\text{RR}_{\text{IA}}^{\mathcal{L}}(S, \eta) \cap \{z \mid h_S^{\mathcal{L}}(z) = 1\} \\
&\subseteq \{z \mid \mathcal{U}^{-1}(z) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = 1,\ \forall\, x \in \mathcal{U}^{-1}(z), h \in \mathcal{H}_0(S)\}. \\
&= A_{\text{ST}} \cap \{z \mid h^*(z) = 1\}
\end{aligned}$$

This concludes that for any robustly-reliable learner $\mathcal{L}$ with respect to ℓ_{IA}, we have

$$\text{RR}^{\mathcal{L}}_{\text{IA}}(S) \subseteq (A_{\text{ST}} \cap \{z \mid h^*(z) = 1\}) \cup \{z \mid z \in \text{Agree}(\mathcal{H}_0(S)) \wedge h^*(z) = 0\}.$$

□

We can also define a safely-reliable region for general perturbations as follows.

Definition 13. *(General safely-reliable region) Let $\mathcal{L}$ be a robustly-reliable learner w.r.t. a perturbation function $\mathcal{U}$ for sample S, concept space $\mathcal{H}$ and robust loss function ℓ. The safely-reliable region of a learner $\mathcal{L}$ is defined as $\text{SR}^{\mathcal{L}}(S) = \{x \in \mathcal{X} \mid \mathcal{U}(x) \subseteq \text{RR}^{\mathcal{L}}(S)\}$.*

D Additional proof details for safely-reliable region

Lemma D.1. *Let $\mathcal{D}$ be isotropic log-concave over $\mathbb{R}^d$ and $\mathcal{H} = \{h : x \to \text{sign}(\langle w_h, x\rangle) \mid w_h \in \mathbb{R}^d, \|w_h\|_2 = 1\}$ be the class of linear separators. Let $\mathbf{B}(\cdot, \eta)$ be a L_2 ball perturbation with radius η. For $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\text{VCdim}(\mathcal{H}) + \ln\frac{1}{\delta}))$, with probability at least $1-\delta$, we have*

$$\Pr(\{x \mid \mathbf{B}(x, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)\}) \geq 1 - 2\eta - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon).$$

Proof. (Proof of Lemma D.1) From uniform convergence (Theorem 4.1 [AB99]), for $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\text{VCdim}(\mathcal{H}) + \ln\frac{1}{\delta}))$, with probability at least $1-\delta$, we have $\text{Agree}(B^{\mathcal{H}}_{\mathcal{D}}(h^*, \varepsilon)) \subseteq \text{Agree}(\mathcal{H}_0(S))$. From [BL13] (Theorem 14), for linear separators on a log-concave distribution $A = \{x : \|x\|_2 < \alpha\sqrt{d}\} \cap \{x : |\langle w_{h^*}, x\rangle| \geq C_1\alpha\varepsilon\sqrt{d}\} \subseteq \text{Agree}(B^{\mathcal{H}}_{\mathcal{D}}(h^*, \varepsilon))$ for some constant C_1. We claim that for any $x \in A_\eta := \{x : \|x\|_2 < \alpha\sqrt{d} - \eta\} \cap \{x : |\langle w_{h^*}, x\rangle| \geq C_1\alpha\varepsilon\sqrt{d} + \eta\}$, we have $\mathbf{B}(x, \eta) \subseteq A$. Let $x \in A_\eta$, consider $z \in \mathbf{B}(x, \eta)$. We have $||z||_2 \leq ||z - x||_2 + \|x\|_2 \leq \eta + \alpha\sqrt{d} - \eta = \alpha\sqrt{d}$ and $|\langle w_{h^*}, z\rangle| \geq |\langle w_{h^*}, x\rangle| - |\langle w_{h^*}, z - x\rangle| \geq C_1\alpha\varepsilon\sqrt{d} + \eta - ||z - x|| \geq C_1\alpha\varepsilon\sqrt{d}$. Therefore, $z \in A$ for any $z \in \mathbf{B}(x, \eta)$ which implies that for any $x \in A_\eta$, $\mathbf{B}(x, \eta) \subseteq A$ which also implies that $A_\eta \subseteq \{x \mid \mathbf{B}(x, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S))\}$. We can bound the probability mass of A_η with the following fact on isotropic log-concave distribution $\mathcal{D}$ over $\mathbb{R}^d$ [LV07]: 1) $\Pr_{x\sim\mathcal{D}}(\|x\| \geq \alpha\sqrt{d}) \leq e^{-\alpha+1}$, 2) When $d = 1$ $\Pr_{x\sim\mathcal{D}}(x \in [a, b]) \leq |b - a|$ and 3) The projection $\langle w_{h^*}, x\rangle$ follows a 1-dimensional isotropic log-concave distribution. We have $\Pr_{x\sim\mathcal{D}}(A_\eta) \geq 1 - \Pr_{x\sim\mathcal{D}}(\{x : ||x|| \geq \alpha\sqrt{d} - \eta\}) - \Pr_{x\sim\mathcal{D}}(\{x : |\langle w_{h^*}, x\rangle| \leq C_1\alpha\varepsilon\sqrt{d} + \eta\}) \geq 1 - e^{-\left(\alpha - \frac{\eta}{\sqrt{d}}\right)+1} - 2C_1\alpha\varepsilon\sqrt{d} - 2\eta = 1 - 2\eta - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$. The final line holds when we set $\alpha = \ln(\frac{1}{\sqrt{d}\varepsilon})$. □

Proof. (Proof of Lemma 3.2) First, we will show that for any sample $S \sim \mathcal{D}^m$ with no points lying on the decision boundary of h^*, there exists a constant $\delta_1(S)$ such that for any h with a small enough angle to h^*, $\theta(w_h, w_{h^*}) \leq \delta_1$, h must have the same prediction as h^* on S that is $h \in \mathcal{H}_0(S)$. Since there is no point lying on the decision boundary, we have $\min_{x\in S} \frac{|\langle w_{h^*}, x\rangle|}{\|x\|} > 0$.

For δ_1, such that $0 < \delta_1 < \min_{x\in S} \frac{|\langle w_{h^*}, x\rangle|}{\|x\|}$, if $\theta(w_h, w_{h^*}) \leq \delta_1$, for any $x \in S$,

$$\begin{aligned} |\langle w_{h^*}, x\rangle - \langle w_h, x\rangle| &\leq \|w_{h^*} - w_h\|\|x\| \\ &\leq \theta(w_{h^*}, w_h)\|x\| \\ &< |\langle w_{h^*}, x\rangle|. \end{aligned}$$

The second to last inequality holds due to the fact that the arc length cannot be smaller than corresponding chord length, and the last inequality follows from the assumption $\theta(w_h, w_{h^*}) \leq \delta_1$. This implies that $\langle w_{h^*}, x\rangle\langle w_h, x\rangle > 0$ and $h \in \mathcal{H}_0(S)$. Now, consider any x such that $c \leq \|x\| \leq d$. We will show that there exists a constant $\delta = \delta(S, c, d)$ such that if the margin of $|\langle w_{h^*}, x\rangle|$ is smaller than δ then $x \notin \text{Agree}(\mathcal{H}_0(S))$. Since $|\langle w_{h^*}, x\rangle| = \|x\|\cos(\theta(w_{h^*}, x)) \geq c\cos(\theta(w_{h^*}, x))$, $\delta > |\langle w_{h^*}, x\rangle|$ implies that $\cos(\theta(w_{h^*}, x)) < \frac{\delta}{c}$ that is the angle between w_{h^*} and x is almost $\frac{\pi}{2}$. Intuitively, we claim that if δ is small enough then there exists $h \in \mathcal{H}_0(S)$ such that $h(x) \neq h^*(x)$. Without loss of generality, let $\langle w_{h^*}, x\rangle > 0$ ($\theta(w_{h^*}, x) < \frac{\pi}{2}$). We will show that if $\theta(w_{h^*}, x)$ is close enough to $\frac{\pi}{2}$, we can rotate w_{h^*} to w_h with a small enough angle so that $\theta(w_h, w_{h^*}) \leq \delta_1$ but

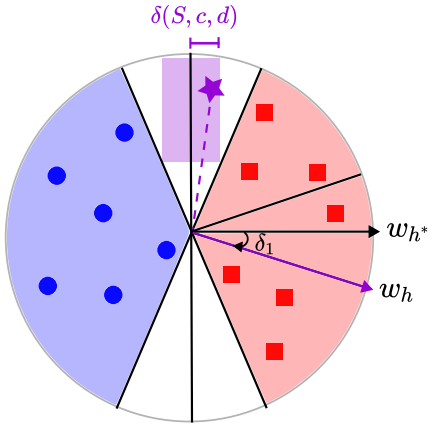

Figure 5: For any set of samples S (blue and red points) with no point lying on the decision boundary of a linear separator h^*, for any $0 < c < d$, there exists an area around the decision boundary of h^* (formally defined as $\{x \in \mathbb{R}^d \mid c \leq \|x\| \leq d, |\langle w_{h^*}, x\rangle| \leq \delta(S, c, d)\}$, illustrated by a purple rectangle) such that for any point (purple star) in this area, there exists a hypothesis h that agree with h^* on S but disagree with h^* at that point

$\langle w_h, x\rangle < 0$ ($\theta(w_h, x) > \frac{\pi}{2}$) as illustrated in Figure 5. Formally, we consider $w_h = \frac{w_{h^*} - \lambda x}{\|w_{h^*} - \lambda x\|}$ for some $\lambda > 0$ (to be specified). We will show that there exists λ such that 1) $\langle w_h, x\rangle < 0$, and 2) $\theta(w_h, w_{h^*}) \leq \delta_1$. The first condition corresponds to $\lambda > \frac{\langle x, w_{h^*}\rangle}{\|x\|^2}$. The second condition leads to the following inequality

$$\frac{\langle w_{h^*} - \lambda x, w_{h^*}\rangle}{\|w_{h^*} - \lambda x\|} \geq \cos(\delta_1)$$
$$\frac{1 - \lambda\langle x, w_{h^*}\rangle}{\sqrt{1 - 2\lambda\langle x, w_{h^*}\rangle + \|x\|^2\lambda^2}} \geq \cos(\delta_1)$$

Assume that $\lambda < \frac{1}{\langle x, w_{h^*}\rangle}$, the inequality is equivalent to

$$(1 - \lambda\langle x, w_{h^*}\rangle)^2 \geq \cos^2(\delta_1)(1 - 2\lambda\langle x, w_{h^*}\rangle + \|x\|^2\lambda^2)$$
$$(\cos^2(\delta_1)\|x\|^2 - \langle x, w_{h^*}\rangle^2)\lambda^2 + 2\lambda\langle x, w_{h^*}\rangle \sin^2(\delta_1) - \sin^2(\delta_1) \leq 0.$$

Solving this inequality leads to

$$\lambda \leq \lambda_{\max} = \frac{-2\sin^2(\delta_1)\langle x, w_{h^*}\rangle + 2\sin(\delta_1)\cos(\delta_1)\sqrt{(\|x\|^2 - \langle x, w_{h^*}\rangle^2)}}{2(\cos^2(\delta_1)\|x\|^2 - \langle x, w_{h^*}\rangle^2)}.$$

Therefore, there exists λ that satisfies both of conditions 1), 2) if $\lambda_{\max} > \frac{\langle x, w_{h^*}\rangle}{\|x\|^2}$. Finally, we claim that if $|\langle x, w_{h^*}\rangle| \leq \delta(S, c, d) \leq \frac{c^2 \tan(\delta_1)}{\sqrt{(d + d\tan(\delta_1))^2 + (c^2\tan(\delta_1))^2}}$ then $\lambda_{\max} > \frac{\langle x, w_{h^*}\rangle}{\|x\|^2}$. For x with $|\langle x, w_{h^*}\rangle| \leq \delta$, we have $\frac{\langle x, w_{h^*}\rangle}{\|x\|^2} \leq \frac{\delta}{\|x\|^2} \leq \frac{\delta}{c^2}$ and also

$$\begin{aligned}
\lambda_{\max} &= \frac{-2\sin^2(\delta_1)\langle x, w_{h^*}\rangle + 2\sin(\delta_1)\cos(\delta_1)\sqrt{(\|x\|^2 - \langle x, w_{h^*}\rangle^2)}}{2(\cos^2(\delta_1)\|x\|^2 - \langle x, w_{h^*}\rangle^2)} \\
&> \frac{-\sin^2(\delta_1)\langle x, w_{h^*}\rangle + \sin(\delta_1)\cos(\delta_1)\sqrt{(\|x\|^2 - \langle x, w_{h^*}\rangle^2)}}{\cos^2(\delta_1)\|x\|^2} \\
&= \frac{-\sin^2(\delta_1)\frac{\langle x, w_{h^*}\rangle}{\|x\|^2} + \sin(\delta_1)\cos(\delta_1)\sqrt{(1 - (\frac{\langle x, w_{h^*}\rangle}{\|x\|})^2)}}{\cos^2(\delta_1)} \\
&\geq \frac{-\sin^2(\delta_1)\frac{\delta}{c^2} + \sin(\delta_1)\cos(\delta_1)\sqrt{(1 - (\frac{\delta}{c^2})^2)}}{\cos^2(\delta_1)d}.
\end{aligned}$$

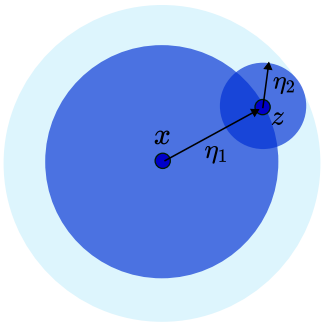

Figure 6: The safely-reliable region contains any point that retains a reliability radius of at least η_2 even after being attacked by an adversary with strength η_1.

The last inequality follows from $\frac{\langle x, w_{h^*}\rangle}{\|x\|^2} \leq \frac{\delta}{c^2}$. It is sufficient to show that

$$\begin{aligned}
&\frac{-\sin^2(\delta_1)\frac{\delta}{c^2} + \sin(\delta_1)\cos(\delta_1)\sqrt{(1-(\frac{\delta}{c^2})^2)}}{\cos^2(\delta_1)d} \geq \frac{\delta}{c^2} \\
\Leftrightarrow \quad & -\tan^2(\delta_1)\frac{\delta}{c^2} + \tan(\delta_1)\sqrt{(1-(\frac{\delta}{c^2})^2)} \geq d\frac{\delta}{c^2} \\
\Leftrightarrow \quad & \tan(\delta_1)\sqrt{(1-(\frac{\delta}{c^2})^2)} \geq (d+\tan^2(\delta_1))\frac{\delta}{c^2} \\
\Leftrightarrow \quad & \delta(S,c,d) \leq \frac{c^2\tan(\delta_1)}{\sqrt{(d+\tan^2(\delta_1))^2 + c^2\tan^2(\delta_1)}}.
\end{aligned}$$

□

Proof. (Proof of Theorem 3.3) We know that for $\ell_{\text{CA}}, \ell_{\text{TL}}$, the robustly-reliable region is the same as the reliable region. This is also a reason why the probability mass does not depend on η_2. Consider the optimal robustly-reliable learner $\mathcal{L}$, we have $\text{RR}^{\mathcal{L}}_{\text{CA}}(S,\eta_2) = \text{RR}^{\mathcal{L}}_{\text{TL}}(S,\eta_2) = \text{Agree}(\mathcal{H}_0(S))$ (Theorem B.1). On the other hand, $\text{RR}^{\mathcal{L}}_{\text{ST}}(S,\eta_2) = \{z \mid \mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = h(z),\ \forall\, x \in \mathbf{B}(z,\eta_2), h \in \mathcal{H}_0(S)\}$ (Theorem B.2). **a)** Since $\text{SR}^{\mathcal{L}}_{\text{TL}}(S,\eta_1,\eta_2) = \{x \mid \mathbf{B}(x,\eta_1) \subseteq \text{Agree}(\mathcal{H}_0(S))\}$. Applying Lemma D.1, we have the first result. **b)** Recall that $\text{SR}^{\mathcal{L}}_{\text{CA}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}(x,\eta_1) \cap \{z \mid h^*(z) = h^*(x)\} \subseteq \text{Agree}(\mathcal{H}_0(S))\}$. We will show that for any $x \in \text{SR}^{\mathcal{L}}_{\text{CA}}(S,\eta_1,\eta_2)$, $\mathbf{B}(x,\eta_1) = \mathbf{B}(x,\eta_1) \cap \{z \mid h^*(z) = h^*(x)\}$ by contradiction. Let $x \in \text{SR}^{\mathcal{L}}_{\text{CA}}(S,\eta_1,\eta_2)$ and $\mathbf{B}(x,\eta_1)$ contain two points with a different label, this implies that this ball must contain the decision boundary of h^* ($\mathbf{B}(x,\eta_1) \cap \{x \mid \langle w^*_h, x\rangle = 0\} \neq \emptyset$). The ball must also contain a point that has the same label as x with an arbitrarily small margin w.r.t. h^*. For any $a > 0$, there exists $z \in \mathbf{B}(x,\eta_1) \cap \{z \mid h^*(z) = h^*(x)\}$ with $|\langle z, w_{h^*}\rangle| < a$. This is impossible because by Lemma 3.2 the agreement region $\text{Agree}(\mathcal{H}_0(S))$ can not contain point with arbitrarily small margin if S does not contain any point on the decision boundary of h^*. This event has a probability 1 as the projection $\langle w_{h^*}, x\rangle$ also follows a log-concave distribution which implies that $\Pr(\langle w_{h^*}, x\rangle = 0) = 0$. Therefore, with probability 1, $\mathbf{B}(x,\eta_1)$ must contain points with the same label and we can conclude that $\text{SR}^{\mathcal{L}}_{\text{CA}}(S,\eta_1,\eta_2) = \text{SR}^{\mathcal{L}}_{\text{TL}}(S,\eta_1,\eta_2)$. **c)** Similarly, by Lemma 3.2, we can show that if $\mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))$, $\mathbf{B}(z,\eta_2)$ then every point in $\mathbf{B}(z,\eta_2)$ must have same label with probability 1. Therefore, $\text{RR}^{\mathcal{L}}_{\text{ST}}(S,\eta_2) = \{z \mid \mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\}$. We have $\text{SR}^{\mathcal{L}}_{\text{ST}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1) \subseteq \{z \mid \mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\}\} = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1+\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\}$ by a triangle inequality (see Figure 6). Applying Lemma D.1, we have the result. **d)** With the result above we have $\text{RR}^{\mathcal{L}}_{\text{IA}}(S,\eta_2) = (\{z \mid \mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z) = 1\}) \cup (\text{Agree}(\mathcal{H}_0(S)) \cap \{z \mid h^*(z) = 0\})$. Recall that $\text{SR}^{\mathcal{L}}_{\text{IA}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid h^*(x) = 0 \wedge \mathbf{B}_{\mathcal{M}}(x,\eta_1) \subseteq \text{RR}^{\mathcal{L}}_{\text{IA}}(S,\eta_2)\} \cup \{x \in \mathcal{X} \mid h^*(x) = 1 \wedge x \in \text{RR}^{\mathcal{L}}_{\text{CA}}(S,\eta_2))\}$. Therefore, we have $\text{SR}^{\mathcal{L}}_{\text{IA}}(S,\eta_1,\eta_2) = (\{z \mid \mathbf{B}(z,\eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z) = 1\}) \cup (\{z \mid \mathbf{B}(z,\eta_1) \subseteq \text{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z) = 0\})$ We can conclude the result by applying Lemma D.1 and symmetry. □

E Safely-reliable region for classifiers with smooth boundaries

We also bound the probability mass of the safely-reliable region for more general concept spaces beyond linear separators. Specifically, we consider classifiers with smooth boundaries in the sense of [vdVW96].

Definition 14 (α-norm)**.** *Let $f : C \to \mathbb{R}$ be a function on $C \subset \mathbb{R}^d$, and let $\alpha \in \mathbb{R}^+$. For $k = (k_1, \ldots, k_d) \in \mathbf{Z}^d_{\geq 0}$, let $||k||_1 = \sum_{i=1}^d k_i$ and let $D^k = \frac{\partial^k}{\partial^{k_1} x_1 \ldots \partial^{k_d} x_d}$. We define α-norm of f as*

$$||f||_\alpha := \max_{||k||_1 < \lceil \alpha \rceil} \sup_{x \in C} |D^k f(x)| + \max_{||k||_1 = \lceil \alpha \rceil - 1} \sup_{x \neq x' \in C} \frac{|D^k f(x) - D^k f(x')|}{|x - x'|^{\alpha - \lceil \alpha \rceil + 1}}.$$

We define αth order smooth functions to be those which have a bounded α-norm. More precisely, we define the class of αth order smooth functions $F^C_\alpha := \{f \mid ||f||_\alpha \leq C\}$. For example, 1st order smoothness corresponds to Lipschitz continuity. We now define concept classes with smooth classification boundaries.

Definition 15 (Concepts with Smooth Classification Boundaries, [Wan11])**.** *A set of concepts $\mathcal{H}^C_\alpha$ defined on $\mathcal{X} = [0,1]^{d+1}$ is said to have αth order smooth classification boundaries, if for every $h \in \mathcal{H}^C_\alpha$ the classification boundary is the graph of function $x_{d+1} = f(x_1, \ldots, x_d)$, where $f \in F^C_\alpha$ and $(x_1, \ldots, x_{d+1}) \in \mathcal{X}$ i.e. the predicted label is given by* $\operatorname{sign}(x_{d+1} - f(x_1, \ldots, x_d))$.

If we further assume that the probability density may be upper and lower bounded by some absolute positive constants (i.e. "nearly" uniform density), we can bound the safely-reliable region of our learner even in this setting. We start with analogues of Lemmas D.1 and 3.2 for concepts with smooth classification boundaries.

We first bound the probability mass of points x for which $\mathbf{B}(x, \eta)$ is contained in the agreement region of sample-consistent classifiers. We use the Lipschitzness of smooth functions to show that such point x must lie outside of a 'ribbon' around the boundary of target concept h^*, and adapt and extend the arguments of [Wan11] to bound the probability mass of this ribbon.

Lemma E.1. *Let the instance space be $\mathcal{X} = [0,1]^{d+1}$ and $\mathcal{D}$ be a distribution over $\mathcal{X}$ with a "nearly" uniform density where there exist positive constants $0 < a < b$ such that $a \leq p(x) \leq b$ for all $x \in [0,1]^{d+1}$ when $p(x)$ is the probability density of $\mathcal{D}$. Let $\mathcal{H}^C_\alpha$ be the hypothesis space of concepts with smooth classification boundaries with $d < \alpha < \infty$, and $\mathbf{B}(\cdot, \eta)$ be a L_2 ball perturbation with radius η. For $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\operatorname{VCdim}(\mathcal{H}) + \ln \frac{1}{\delta}))$, with probability at least $1 - \delta$, we have*

$$\Pr(\{x \mid \mathbf{B}(x, \eta) \subseteq \operatorname{Agree}(\mathcal{H}_0(S)\}) \geq 1 - 2b(C+1)\eta - \mathcal{O}\left(ba^{-\frac{\alpha}{d+\alpha}} \varepsilon^{\frac{\alpha}{d+\alpha}}\right).$$

Proof. By uniform convergence (Theorem 4.1, [AB99]), for $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\operatorname{VCdim}(\mathcal{H}) + \ln \frac{1}{\delta}))$, with probability at least $1 - \delta$, we have $\operatorname{Agree}(B^{\mathcal{H}}_{\mathcal{D}}(h^*, \varepsilon)) \subseteq \operatorname{Agree}(\mathcal{H}_0(S))$. Therefore, it suffices to lower bound $\pi := \Pr\{x \mid \mathbf{B}(x, \eta) \subseteq \operatorname{Agree}(B^{\mathcal{H}}_{\mathcal{D}}(h^*, \varepsilon))\}$. Let $h^* \in \mathcal{H}^C_\alpha$ be the target concept and denote $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0,1]^d$. Recall that the predicted label of $(\boldsymbol{x}, x_{d+1})$ from h, h^* is given by $\operatorname{sign}(x_{d+1} - f_h(\boldsymbol{x}))$ and $\operatorname{sign}(x_{d+1} - f_{h^*}(\boldsymbol{x}))$ respectively. Therefore, h, h^* would disagree on $(\boldsymbol{x}, x_{d+1})$ when x_{d+1} lies between $f_h(\boldsymbol{x})$ and $f_{h^*}(\boldsymbol{x})$. Denote $\Phi_h(\boldsymbol{x}) = |\int_{f_{h^*}(\boldsymbol{x})}^{f_h(\boldsymbol{x})} p(\boldsymbol{x}, x_{d+1}) dx_{d+1}|$ be the probability mass of points that h disagree with h^* over $(\boldsymbol{x}, x_{d+1})$ for a fixed $\boldsymbol{x} \in [0,1]^d$. With this notation, the probability mass of points $(\boldsymbol{x}, x_{d+1})$ that h and h^* disagree with is given by $\int_{[0,1]^d} |\Phi_h(\boldsymbol{x})| d\boldsymbol{x}$. Furthermore, from $a \leq p(\boldsymbol{x}, x_{d+1}) \leq b$, we know that $a|f_h(\boldsymbol{x}) - f_{h^*}(\boldsymbol{x})| \leq |\Phi_h(\boldsymbol{x})| \leq b|f_h(\boldsymbol{x}) - f_{h^*}(\boldsymbol{x})|$.

Consider $h \in B_{\mathcal{D}}(h^*, \varepsilon)$, we have $\int_{[0,1]^d} |\Phi_h(\boldsymbol{x})| d\boldsymbol{x} \leq \varepsilon$. This implies that $\int_{[0,1]^d} |f_h(\boldsymbol{x}) - f_{h^*}(\boldsymbol{x})| d\boldsymbol{x} \leq \int_{[0,1]^d} |\frac{\Phi(\boldsymbol{x})}{a}| d\boldsymbol{x} \leq \frac{\varepsilon}{a}$. Since the classification boundaries are assumed to be αth order smooth with $\alpha > d$, Lemma 11 of [Wan11] implies that $||f_h - f_{h^*}||_\infty = O\left((\frac{\varepsilon}{a})^{\frac{\alpha}{d+\alpha}}\right)$ where $||g||_\infty := \sup_{\boldsymbol{x} \in [0,1]^d} |g(\boldsymbol{x})|$. Consider

$$1 - \pi = \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}(\exists z \in \mathbf{B}(x, \eta), \exists h \in B(h^*, \varepsilon), h(z) \neq h^*(z)).$$

Recall that for $z = (\mathbf{z}, z_{d+1})$, $h(z) \neq h^*(z)$ when z_{d+1} lies between $h(\mathbf{z}), h^*(\mathbf{z})$ which implies $f_{h^*}(\mathbf{z}) - |f_{h^*}(\mathbf{z}) - f_h(\mathbf{z})| < z_{d+1} < f_{h^*}(\mathbf{z}) + |f_{h^*}(\mathbf{z}) - f_h(\mathbf{z})|$. Since $z \in \mathbf{B}(x, \eta)$ and the boundary

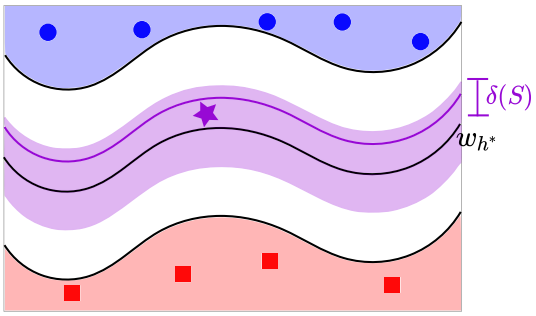

Figure 7: For any set of samples S (blue and red points) with no point lying on the decision boundary of a concept with smooth boundary h^*, there exists a band around the decision boundary of h^* (formally defined as $\{x \in \mathbb{R}^d \mid |f_{h^*}(\boldsymbol{x}) - x_{d+1}| < \delta\}$ (purple)) such that for any point (purple star) in this area, there exists a hypothesis h (by translation) that agree with h^* on S but disagree with h^* at that point.

functions f_h are C-Lipschitz (by Definition 14) we have $|f_h(\mathbf{z}) - f_h(\boldsymbol{x})| \leq C\|\mathbf{z} - \boldsymbol{x}\| \leq C\eta$ and $|z_{d+1} - x_{d+1}| \leq \eta$. This implies that $f_{h^*}(\boldsymbol{x}) - |f_{h^*}(\mathbf{z}) - f_h(\mathbf{z})| - (C+1)\eta < x_{d+1} < f_{h^*}(\boldsymbol{x}) + |f_{h^*}(\mathbf{z}) - f_h(\mathbf{z})| + (C+1)\eta$. We are interested in the set $\{x \mid \exists z \in \mathbf{B}(x,\eta), \exists h \in B(h^*,\varepsilon), h(z) \neq h^*(z)\}$, this leads to the inequality

$$f_{h^*}(\boldsymbol{x}) - D - (C+1)\eta < x_{d+1} < f_{h^*}(\boldsymbol{x}) + D + (C+1)\eta$$

when $D = \sup_{\substack{z\in\mathbf{B}(x,\eta)\\ h\in B(h^*,\varepsilon)}} |f_{h^*}(\mathbf{z}) - f_h(\mathbf{z})| \leq \sup_{h\in B(h^*,\varepsilon)} \|f_{h^*} - f_h\|_\infty = O\left((\frac{\varepsilon}{a})^{\frac{\alpha}{d+\alpha}}\right)$. Therefore,

$$\begin{aligned}
1 - \pi &\leq \int_{[0,1]^d} \int_{f_{h^*}(\boldsymbol{x})-D-(C+1)\eta}^{f_{h^*}(\boldsymbol{x})+D+(C+1)\eta} p(\boldsymbol{x}, x_{d+1}) dx_{d+1} d\boldsymbol{x} \\
&\leq 2b((C+1)\eta + D) \\
&= 2b(C+1)\eta + O\left(ba^{-\frac{\alpha}{d+\alpha}} \varepsilon^{\frac{\alpha}{d+\alpha}}\right).
\end{aligned}$$

□

The above bound immediately implies a bound on the probability mass of the safely-reliable region for ℓ_{TL}, in combination with our previous results. The following lemma allows us to easily handle the extension of our results to losses ℓ_{CA} and ℓ_{ST} as well, where the safely-reliable region involves additional constraints.

Lemma E.2. *Let the instance space be $\mathcal{X} = [0,1]^{d+1}$ and $\mathcal{D}$ be a distribution over $\mathcal{X}$. Let $\mathcal{H}_\alpha^C$ be the hypothesis space of concepts with smooth classification boundaries with $d < \alpha < \infty$. For $h^* \in \mathcal{H}_\alpha^C$, for a set of samples $S \sim \mathcal{D}^m$ such that there is no data point in S that lies on the decision boundary, there exists $\delta(S) > 0$ such that for any $x = (\boldsymbol{x}, x_{d+1})$ with $|f_{h^*}(\boldsymbol{x}) - x_{d+1}| < \delta$, we have $x \notin \text{Agree}(\mathcal{H}_0(S))$.*

Proof. Note that translation preserves smoothness, i.e. any concept h_t with $f_{h_t}(\boldsymbol{x}) = f_{h^*}(\boldsymbol{x}) + t$ would also lie in $\mathcal{H}_\alpha^C$. If $|t| < \delta(S) = \min_{x\in S} |f_{h^*}(\boldsymbol{x}) - x_{d+1}|$, $f_{h_t}(\boldsymbol{x}) - x_{d+1}$ must have the same sign as $(f_{h^*}(\boldsymbol{x}) - x_{d+1})$ for any $x \in S$, that is h_t would agree with h^* on every point in S. Furthermore, for any x with $|f_{h^*}(\boldsymbol{x}) - x_{d+1}| < \delta$, we can always translate h^* to h_t that $h_t(x) \neq h^*(x)$ (see Figure 7). Therefore, $x \notin \text{Agree}(\mathcal{H}_0(S))$. □

Equipped with the above lemmas, we establish bounds on the probability mass of the safely reliable region for concept classes with smooth classification boundaries.

Theorem E.3. *Let the instance space be $\mathcal{X} = [0,1]^{d+1}$ and $\mathcal{D}$ be a distribution over $\mathcal{X}$ with a "nearly" uniform density where there exist positive constants $0 < a < b$ such that $a \leq p(x) \leq b$ for all $x \in [0,1]^{d+1}$ when $p(x)$ is the probability density of $\mathcal{D}$. Let $\mathcal{H}_\alpha^C$ be the hypothesis space of concepts with smooth classification boundaries with $d < \alpha < \infty$, and $\mathbf{B}(\cdot,\eta)$ be a L_2 ball perturbation with radius η. For $S \sim \mathcal{D}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\text{VCdim}(\mathcal{H}) + \ln\frac{1}{\delta}))$ with no point lying on the decision boundary of h^*, for an optimal robustly-reliable learner $\mathcal{L}$, we have*

(a) $\Pr(SR^{\mathcal{L}}_{TL}(S,\eta_1,\eta_2)) \geq 1-2b(C+1)\eta_1 - \mathcal{O}\left(ba^{-\frac{\alpha}{d+\alpha}}\varepsilon^{\frac{\alpha}{d+\alpha}}\right)$ *with probability* $1-\delta$,
(b) $SR^{\mathcal{L}}_{CA}(S,\eta_1,\eta_2) = SR^{\mathcal{L}}_{TL}(S,\eta_1,\eta_2)$,
(c) $\Pr(SR^{\mathcal{L}}_{ST}(S,\eta_1,\eta_2)) \geq 1-2b(C+1)(\eta_1+\eta_2) - \mathcal{O}\left(ba^{-\frac{\alpha}{d+\alpha}}\varepsilon^{\frac{\alpha}{d+\alpha}}\right)$ *with probability* $1-\delta$,
(d) $\Pr(SR^{\mathcal{L}}_{IA}(S,\eta_1,\eta_2)) \geq 1-b(C+1)(\eta_1+\eta_2) - \mathcal{O}\left(ba^{-\frac{\alpha}{d+\alpha}}\varepsilon^{\frac{\alpha}{d+\alpha}}\right)$ *with probability* $1-\delta$.

Proof. From Lemma E.2, the agreement region does not contain points that are arbitrarily close to the boundary of h^*. Therefore, we can remove the additional conditions on labels for $\mathrm{SR}^{\mathcal{L}}_{\mathrm{CA}}(S,\eta_1,\eta_2)$ and $\mathrm{SR}^{\mathcal{L}}_{\mathrm{ST}}(S,\eta_1,\eta_2)$, that is $\mathrm{SR}^{\mathcal{L}}_{\mathrm{CA}}(S,\eta_1,\eta_2) = \mathrm{SR}^{\mathcal{L}}_{\mathrm{TL}}(S,\eta_1,\eta_2) = \{x \mid \mathbf{B}(x,\eta_1) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\}$ and $\mathrm{SR}^{\mathcal{L}}_{\mathrm{ST}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1) \subseteq \{z \mid \mathbf{B}(z,\eta_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\}\} = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1+\eta_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\}$. Similarly, we have $\mathrm{SR}^{\mathcal{L}}_{\mathrm{IA}}(S,\eta_1,\eta_2) = (\{z \mid \mathbf{B}(z,\eta_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z)=1\}) \cup (\{z \mid \mathbf{B}(z,\eta_1) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z)=0\})$. The result follows from Lemma E.1.

□

F More on computational efficiency

It is possible to extend the optimization objective to a wide range of hypothesis classes under the following assumption.

Assumption 1. *For a hypothesis class $\mathcal{H}$, we assume that for any $h^* \in \mathcal{H}$, a set of data points S labeled by h^* then for any points x, y that $h^*(x) \neq h^*(y)$, the line that connects between x, y must pass through a disagreement region of $\mathcal{H}_0(S)$,*

$$\{\lambda x + (1-\lambda)y \mid \lambda \in [0,1]\} \cap \mathrm{DIS}(\mathcal{H}_0(S)) \neq \emptyset.$$

For example, a class of linear separators and a class of classifiers with smooth boundaries satisfies this assumption.

Lemma F.1. *Let $\mathcal{H}$ be a hypothesis class and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. If $\mathcal{H}$ satisfies Assumption 1 then for a set of samples $S \sim \mathcal{D}^m$, the reliability radius of a test point z is given by*

$$\begin{aligned} &\min_{h,h',z'} ||z-z'||^2 \\ \textit{s.t.} \quad &h \in \mathcal{H}_0(S), \\ &h' \in \mathcal{H}_0(S), \\ &h(z') \neq h'(z'). \end{aligned}$$

Proof. Let r be the largest reliability radius of a test point z that is if we perturb z by a radius at most r then the perturbed point is still in the agreement region. Consider for any perturbation z' that there exists h, h' that $h'(z')$ has a different label from $h'(z)$. If $\mathcal{H}$ satisfies Assumption 1, $h'(z) \neq h'(z')$ implies that the line between z, z' must pass through the disagreement region. Therefore, $r \leq \|z-z'\|$. On the other hand, let r_0^2 be the solution of the optimization given above. This implies that for any point z' that $\|z-z'\| < r_0$, for any $h, h' \in \mathcal{H}_0(S)$, we must have that $h(z) = h(z')$ that is $z' \in \mathrm{Agree}(\mathcal{H}_0(S))$. Therefore, we have $r_0 \leq r$ and we can conclude that $r = r_0$. □

Remark. This could be solved efficiently in practice. For example, in the case of linear separators, the problem takes the form of solving a quadratic program for each test point. We observe that our proof of Theorem 3.3 above suggests an alternate margin-based approach which might be even more practical to implement, while retaining high probability reliability guarantees under the distributional assumption. If $\hat{h}$ denotes an ERM classifier on sample S, one could check the membership $z \in \hat{A}_\eta := \{x : \|x\|_2 < \alpha\sqrt{d} - \eta\} \cap \{x : |\langle w_{\hat{h}}, x\rangle| \geq C_1\alpha\varepsilon\sqrt{d} + \eta\}$ for any $\eta \geq 0$ by just computing the norm and margin w.r.t. $\hat{h}$. A simple halving search (e.g. starting with $\eta = \frac{1}{2}$) could be used to estimate the reliability radius.

Further, we can relax this constrained objective into a regularized objective that can be solved using empirical risk minimization. In the following Lemma, we show that this provides a lower bound on the reliability radius.

Lemma F.2. *(Relaxation of the optimization objective for reliability radius) Let $\mathcal{H}$ be a hypothesis class and $\mathcal{D}$ be a distribution over $\mathbb{R}^d$. If $\mathcal{H}$ satisfies Assumption 1 then for a set of samples $S \sim \mathcal{D}^m$, let h_1, h_2, z^* be the optimal solution of the objective*

$$h_1, h_2, z^* = \operatorname*{argmin}_{h,h',z'} \|z - z'\|^2 + \lambda(\hat{R}(h, S \cup \{(z', 0)\}) + \hat{R}(h', S \cup \{(z', 1)\}))$$

when $\hat{R}(h, A)$ is an empirical risk of h on the sample A then $\|z - z^*\|^2 \leq r$ when r the reliability radius of z.

Proof. Let h, h', z' be the optimal solution of the optimization objective in Lemma F.1 so that the reliability radius r is given by $\|z - z'\|$. Without loss of generality, let $h(z') = 0$ and $h'(z') = 1$. By definition, we have $\hat{R}(h, S \cup \{(z', 0)\}) = 0$ and $\hat{R}(h', S \cup \{(z', 1)\})) = 0$. Let h_1, h_2, z^* be an optimal solution of the objective in Lemma F.2 then we have

$$\begin{aligned}
&\|z - z^*\|^2 + \lambda(\hat{R}(h_1, S \cup \{(z^*, 0)\}) + \hat{R}(h_2, S \cup \{(z^*, 1)\})) \\
&\leq \|z - z'\|^2 + \lambda(\hat{R}(h, S \cup \{(z', 0)\}) + \hat{R}(h', S \cup \{(z', 1)\})) \\
&= \|z - z'\|^2
\end{aligned}$$

Since the empirical risk is non-negative, we can conclude that $\|z - z^*\|^2 \leq \|z - z'\|^2 = r^2$. □

Similar observations also apply to the computation of the safely-reliable region. It may be useful to compute the safely-reliable region when we want to estimate the robust reliability performance of an algorithm on test data. In particular, this may be helpful in determining how often and where our learner gives a bad reliability radius, which can inform its safe deployment in practice.

G Bounds on the $\mathcal{P} \to \mathcal{Q}$ disagreement coefficient

We will now consider some commonly studied concept spaces, and bound the $\mathcal{P} \to \mathcal{Q}$ disagreement coefficient for broad classes of distribution shifts.

Linear separators and nearly log-concave or s-concave distributions. We give a bound on $\Theta_{\mathcal{P} \to \mathcal{Q}}$ for $\mathcal{P}$ and $\mathcal{Q}$ isotropic nearly log-concave distributions [AK91] over $\mathbb{R}^d$, a broad class that includes isotropic log-concave distributions.

Definition 16 (β-log-concavity)**.** *A density function $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is β-log-concave if for any $\lambda \in [0, 1]$ and any $x_1, x_2 \in \mathbb{R}^d$, we have $f(\lambda x_1 + (1 - \lambda)x_2) \geq e^{-\beta} f(x_1)^{\lambda} f(x_2)^{1-\lambda}$.*

For example, 0-log-concave densities are also log-concave. Also, since the condition in the definition holds for $\lambda = 0$, we have $\beta \geq 0$. Note that a smaller value of β makes the distribution closer to logconcave. We have the following bound on $\Theta_{\mathcal{P} \to \mathcal{Q}}$ for distribution shift involving nearly log-concave densities. At a high level, we bound the angle between the normal vectors of the linear separators with small disagreement with h^* under $\mathcal{P}$, and bound the probability mass of their disagreement region under $\mathcal{Q}$, by refining and generalizing the arguments from [BL13]. In particular, we quantify how much near-logconcavity is sufficient for the angle bound to hold, which further implies that any point in the disagreement region is either far away from the mean or close to the margin.

Theorem G.1. *Let the concept space $\mathcal{H}$ be the class of linear separators in $\mathbb{R}^d$. Let $\mathcal{P}$ be isotropic β_1-log-concave and $\mathcal{Q}$ be isotropic β_2-log-concave, over $\mathbb{R}^d$. Then for $0 \leq \beta_1, \beta_2 \leq \frac{1}{56\lceil \log_2(d+1) \rceil}$, we have $\Theta_{\mathcal{P} \to \mathcal{Q}}(\varepsilon) = O(d^{1/2 + \frac{\beta_2}{2 \ln 2}} \log(d/\varepsilon))$.*

Proof. Our proof builds on and generalizes the arguments used in the proof of Theorem 14 in [BL13]. Let $h \in B_{\mathcal{P}}(h^*, r)$, i.e. $d(h, h^*) \leq r$. We can apply the whitening transform from Theorem 16 of [BL13] provided $(1/20 + c_1)\sqrt{1/C_1 - c_1^2} \leq 1/9$, where $C_1 = e^{\beta_1 \lceil \log_2(d+1) \rceil}$ and

$c_1 = e(C_1 - 1)\sqrt{2C_1}$. It may be verified that this condition holds for $0 \leq \beta_1 \leq \frac{1}{56\lceil \log_2(d+1)\rceil}$. Now, by Theorem 11 of [BL13] we can bound the angle between their normal vectors as $\theta(w_h, w_{h^*}) \leq cr$ where c is an absolute constant. Now if $x \in \mathcal{X}$ has a large margin $|w_{h^*} \cdot x| \geq cr\alpha$ and small norm $||x|| \leq \alpha$, for some $\alpha > 0$, we have

$$|w_h \cdot x - w_{h^*} \cdot x| \leq ||w_h - w_{h^*}|| \cdot ||x|| < cr\alpha.$$

Now the large margin condition $|w_{h^*} \cdot x| \geq cr\alpha$ implies $\langle w_h, x\rangle\langle w_{h^*}, x\rangle > 0$, or $h(x) = h^*(x)$. Since $h \in B_{\mathcal{P}}(h^*, r)$ was arbitrary, we have $x \notin \text{DIS}(B_{\mathcal{P}}(h^*, r)))$. Therefore, the set $\{x \mid ||x|| > \alpha\} \cup \{x \mid |w_{h^*} \cdot x| \leq cr\alpha\}$ contains the disagreement region $\text{DIS}(B_{\mathcal{P}}(h^*, r)))$.

By Theorem 11 of [BL13], since $\mathcal{Q}$ is an isotropic β_2-log-concave distribution, we have $\Pr_Q[||x|| > R\sqrt{Cd}] < Ce^{-R+1}$, for $C = e^{\beta_2\lceil \log_2(d+1)\rceil}$. Thus setting $\alpha = \sqrt{Cd}\log\frac{\sqrt{C}}{r}$ gives $\Pr_{\mathcal{Q}}[||x|| > \alpha] < e\sqrt{C}r$. Also, by Theorem 11 of [BL13], for sufficiently small non-negative $\beta_2 \leq \frac{1}{56\lceil \log_2(d+1)\rceil}$, we have $\Pr_Q[|w_{h^*} \cdot x| \leq cr\alpha] \leq c'r\sqrt{Cd}\log\frac{\sqrt{C}}{r}$ for constant c'. The proof is concluded by a union bound and applying Definition 9. □

We further consider the case where the distributions belong to the broad class of isotropic s-concave distributions. In particular, unlike β-log-concave distributions, the distributions from this class can potentially be fat-tailed.

Definition 17 (s-concavity)**.** *A density function $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ is s-concave for $s \in (-\infty, 1] \cup \{-\infty\}$ if for any $\lambda \in [0, 1]$ and any $x_1, x_2 \in \mathbb{R}^d$, we have $f(\lambda x_1 + (1-\lambda)x_2) \geq (\lambda f(x_1)^s + (1-\lambda)f(x_2)^s)^{1/s}$.*

Note that any s-concave function is also s'-concave if $s > s'$. Moreover, concave functions are 1-concave and log-concave functions are s-concave for any $s < 0$. Using results from [BZ17], we adapt the arguments in Theorem G.1 to show a bound on the disagreement coefficient when $\mathcal{P}$ is isotropic β-log-concave and $\mathcal{Q}$ is isotropic s-concave.

Theorem G.2. *Let the concept space $\mathcal{H}$ be the class of linear separators in $\mathbb{R}^d$. Let $\mathcal{P}$ be isotropic β-log-concave and $\mathcal{Q}$ be isotropic s-concave, over $\mathbb{R}^d$. Then for $s \geq -1/(2d+3)$ and sufficiently small non-negative $\beta \leq \frac{1}{56\lceil \log_2(d+1)\rceil}$, we have $\Theta_{\mathcal{P}\to\mathcal{Q}}(\varepsilon) = O\left(\sqrt{d}\frac{2(1+ds)^2}{s+s^2(d+2)}(1-\varepsilon^{s/(1+ds)})\right)$.*

Proof. Similar to the proof of Theorem G.1, we can apply the whitening transform from Theorem 16 of [BL13] provided $(1/20 + c_1)\sqrt{1/C_1 - c_1^2} \leq 1/9$, where $C_1 = e^{\beta\lceil \log_2(d+1)\rceil}$ and $c_1 = e(C_1 - 1)\sqrt{2C_1}$. It may be verified that this condition holds for $0 \leq \beta \leq \frac{1}{56\lceil \log_2(d+1)\rceil}$. We can also show that the set $\{x \mid ||x|| > \alpha\} \cup \{x \mid |w_{h^*} \cdot x| \leq cr\alpha\}$ contains the disagreement region $\text{DIS}(B_{\mathcal{P}}(h^*, r)))$.

By Theorem 11 of [BZ17], we have $\Pr_{\mathcal{Q}}[|w_{h^*} \cdot x| \leq cr\alpha] \leq \frac{2(1+ds)}{1+s(d+2)} \cdot cr\alpha$. By Theorem 5 of [BZ17], since $\mathcal{Q}$ is an isotropic s-concave distribution, we have $\Pr_Q[||x|| > t\sqrt{d}] < \left(1 - \frac{c_1 st}{1+ds}\right)^{(1+ds)/s}$, for any $t \geq 16$ and absolute constant c_1. This implies $\Pr_{\mathcal{Q}}[||x|| > c_1\sqrt{d}\frac{1+ds}{s}(1 - r^{s/(1+ds)})] < Cr$ for some constant C. Thus setting $\alpha = c_1\sqrt{d}\frac{1+ds}{s}(1 - r^{s/(1+ds)})$ gives $\Pr_{\mathcal{Q}}[||x|| > \alpha] < Cr$. Also, for this α, we have $\Pr_{\mathcal{Q}}[|w_{h^*} \cdot x| \leq cr\alpha] \leq c\frac{2(1+ds)}{1+s(d+2)} \cdot c_1\sqrt{d}\frac{1+ds}{s}(1 - r^{s/(1+ds)})r = c'\sqrt{d}\frac{2(1+ds)^2}{s+s^2(d+2)}(1 - r^{s/(1+ds)})r$ for constant c'. The proof is concluded by a union bound and applying Definition 9. □

Smooth classification boundaries. We also illustrate our notion for more general concept spaces beyond linear separators. Specifically, we consider classifiers with smooth boundaries (Definition 15). If we further assume that the probability density may be upper and lower bounded by an αth order smooth function, we can bound the disagreement coefficient for shift from $\mathcal{P}$ to $\mathcal{Q}$. Interestingly, while [Wan11] need the distribution to be sandwiched between smooth functions, our result only needs a lower bound on the smoothness of $\mathcal{P}$ and an upper bound on the smoothness of $\mathcal{Q}$.

Theorem G.3. *Let the instance space be $\mathcal{X} = [0,1]^{d+1}$. Let the hypothesis space be $\mathcal{H}_\alpha^C$, with $d < \alpha < \infty$. If the marginal distributions $\mathcal{P}_\mathcal{X}, \mathcal{Q}_\mathcal{X}$ have densities $p(x)$ and $q(x)$ on $[0,1]^{d+1}$ such that there exists an αth order smooth function $g(x)$ and $a_p, b_q \in \mathbb{R}_+$ such that $a_p g(x) \leq p(x)$ and $q(x) \leq b_q g(x)$ for all $x \in [0,1]^{d+1}$, then $\Theta_{\mathcal{P}\to\mathcal{Q}}(\varepsilon) = O\left(b_q a_p^{-\frac{\alpha}{d+\alpha}} \varepsilon^{-\frac{d}{d+\alpha}}\right)$.*

Proof. We will extend the arguments from [Wan11] to the distribution shift setting. Let $\boldsymbol{x} = (x_1, \ldots, x_d) \in [0,1]^d$ and let $h \in B_P(h^*, r)$ where $h^* \in \mathcal{H}_\alpha$ is the target concept. Denote $\Phi_h^p(\boldsymbol{x}) = \int_{f_{h^*}(\boldsymbol{x})}^{f_h(\boldsymbol{x})} p(\boldsymbol{x}, x_{d+1}) dx_{d+1}$, and $\Phi_h^q(\boldsymbol{x}) = \int_{f_{h^*}(\boldsymbol{x})}^{f_h(\boldsymbol{x})} q(\boldsymbol{x}, x_{d+1}) dx_{d+1}$. It is easy to verify by taking derivatives that $\tilde{\Phi}_h(\boldsymbol{x}) = \int_{f_{h^*}(\boldsymbol{x})}^{f_h(\boldsymbol{x})} g(\boldsymbol{x}, x_{d+1}) dx_{d+1}$ is αth order smooth. Since $h \in B_P(h^*, r)$,

$$\int_{[0,1]^d} |\tilde{\Phi}_h(\boldsymbol{x})| d\boldsymbol{x} \leq \int_{[0,1]^d} \frac{1}{a_p} |\Phi_h^p(\boldsymbol{x})| d\boldsymbol{x} \leq \frac{r}{a_p}.$$

By Lemma 11 of [Wan11], this implies $||\tilde{\Phi}_h||_\infty = O\left((\frac{r}{a_p})^{\frac{\alpha}{d+\alpha}}\right)$, and therefore $||\Phi_h^q||_\infty \leq b_q ||\tilde{\Phi}_h||_\infty = O\left(b_q a_p^{-\frac{\alpha}{d+\alpha}} r^{\frac{\alpha}{d+\alpha}}\right)$. Since this holds for any $h \in B_\mathcal{P}(h^*, r)$, we have

$$\sup_{h \in B_P(h^*, r)} ||\Phi_h^q||_\infty = O\left(b_q a_p^{-\frac{\alpha}{d+\alpha}} r^{\frac{\alpha}{d+\alpha}}\right).$$

By definition of region of disagreement, we have

$$\begin{aligned}
\Pr_{x \sim \mathcal{Q}_\mathcal{X}}[x \in DIS(B_\mathcal{P}(h^*, r))] &= \Pr_{x \sim \mathcal{Q}_X}[x \in \cup_{h \in B_\mathcal{P}(h^*, r)} \{x' \mid h(x') \neq h^*(x')\}] \\
&\leq 2 \int_{[0,1]^d} \sup_{h \in B_\mathcal{P}(h^*, r)} ||\Phi_h^q||_\infty d\boldsymbol{x} \\
&= O\left(b_q a_p^{-\frac{\alpha}{d+\alpha}} r^{\frac{\alpha}{d+\alpha}}\right).
\end{aligned}$$

The result follows from definition of $\Theta_{\mathcal{P}\to\mathcal{Q}}(\epsilon)$. □

H Simple examples for the $\mathcal{P}\to\mathcal{Q}$ disagreement coefficient

Example 1. (Non-overlapping spheres the same center) Let $\mathcal{P}$, $\mathcal{Q}$ be uniform distribution over a sphere with the center at the origin with radius 1 and 2 respectively. Let $\mathcal{H}$ be a class of linear separators that pass through the origin and $h^* \in \mathcal{H}$. By symmetry, we have

$$\begin{aligned}
\Theta_{\mathcal{P}\to\mathcal{Q}}(\varepsilon) &= \sup_{r \geq \varepsilon} \frac{\Pr_\mathcal{Q}(\mathrm{DIS}(B_\mathcal{P}(h^*, r)))}{r} \\
&= \sup_{r \geq \varepsilon} \frac{\Pr_\mathcal{P}(\mathrm{DIS}(B_\mathcal{P}(h^*, r)))}{r} \\
&= \Theta_\mathcal{P}(\varepsilon).
\end{aligned}$$

The disagreement coefficient from $\mathcal{P}$ to $\mathcal{Q}$ is the same as the disagreement coefficient on $\mathcal{P}$.

Example 2. (Thresholds) Let $\mathcal{P}$, $\mathcal{Q}$ be uniform distribution over an interval $[-\frac{1}{2}, \frac{1}{2}]$ and $[-1, 1]$ respectively. Let $\mathcal{H}$ be a class of a threshold function. Let h^* have a thereshold at 0. We have $\mathrm{DIS}(B_\mathcal{P}(h^*, r)) = [-r, r]$ and

$$\begin{aligned}
\Theta_{\mathcal{P}\to\mathcal{Q}}(\varepsilon) &= \sup_{r \geq \varepsilon} \frac{\Pr_\mathcal{Q}([-r, r])}{r} \\
&= \frac{\frac{2r}{2}}{r} = 1
\end{aligned}$$

compared to the disagreement coefficient $\Theta_\mathcal{P}(\varepsilon) = 2$.

I Additional proof details for distribution shift

Proof. (of Theorem 5.1) If $S \sim \mathcal{P}^m$, with $m \geq \frac{c}{\epsilon^2}(d + \ln \frac{1}{\delta})$ for some sufficiently large constant c, we have by uniform convergence ([AB99], Theorem 4.10) that with probability at least $1-\delta$, we have $d_{\mathcal{P}}(h, h^*) \leq d_S(h, h^*) + \epsilon$ for all $h \in \mathcal{H}$. Here $d_{\mathcal{D}}(h_1, h_2) = \Pr_{x \sim \mathcal{D}_X}[h_1(x) \neq h_2(x)]$, and $d_S(h_1, h_2) = \frac{1}{|S|}\sum_{x \in S} \mathbb{I}[h_1(x) \neq h_2(x)]$. Therefore, $\text{Agree}(B^{\mathcal{H}}_{\mathcal{P}}(h^*, \epsilon)) \subseteq \text{Agree}(\mathcal{H}_0(S)) \subseteq R^{\mathcal{L}}(S)$ in this event. Denoting this event by 'E' and its complement by '$\bar{E}$', we have

$$\begin{aligned}
\Pr_{x \sim \mathcal{Q}, S \sim \mathcal{P}^m}[x \in R^{\mathcal{L}}(S)] &= \Pr_{x \sim \mathcal{Q}}[x \in R^{\mathcal{L}}(S) \mid E] \Pr[E] + \Pr_{x \sim \mathcal{Q}}[x \in R^{\mathcal{L}}(S) \mid \bar{E}] \Pr[\bar{E}] \\
&\geq \Pr_{x \sim \mathcal{Q}}[x \in R^{\mathcal{L}}(S) \mid E] \cdot (1 - \delta) \\
&\geq \Pr_{x \sim \mathcal{Q}}[x \in R^{\mathcal{L}}(S) \mid E] - \delta \\
&\geq \Pr_{x \sim \mathcal{Q}}[x \in \text{Agree}(B^{\mathcal{H}}_{\mathcal{P}}(h^*, \epsilon))] - \delta.
\end{aligned}$$

Noting $\Pr_{x \sim \mathcal{Q}}[x \in \text{Agree}(B^{\mathcal{H}}_{\mathcal{P}}(h^*, \epsilon))] = 1 - \Pr_{x \sim \mathcal{Q}}[x \in \text{DIS}(B^{\mathcal{H}}_{\mathcal{P}}(h^*, \epsilon))]$ and using Definition 9 completes the proof. □

J Safely-reliable correctness under distribution shift

There is a growing practical [SSZ+20, SIE+20] as well as recent theoretical interest [DGH+23] in the setting of 'robustness transfer', where one simultaneously expects adversarial test-time attacks as well as distribution shift. We will study the reliability aspect for this more challenging setting. We note that the definition of a robustly-reliable learner does not depend on the data distribution (see Definition 3) as the guarantee is pointwise. Our optimality result in Section 3 applies even when a test point is drawn from a different distribution $\mathcal{Q}$. In this case, the safely-reliable region instead would have a different probability mass.

Definition 18 ($\mathcal{P} \to \mathcal{Q}$ safely-reliable correctness)**.** *The $\mathcal{P} \to \mathcal{Q}$ safely-reliable correctness of $\mathcal{L}$ (at sample rate m, for distribution shift from $\mathcal{P}$ to $\mathcal{Q}$, w.r.t. robust loss ℓ) is defined as the probability mass of its safely-reliable region under Q, on a sample $S \sim \mathcal{P}^m$, i.e. $PQR^{\mathcal{L}}_{\ell}(S, \eta_1, \eta_2) := \Pr_{x \sim Q, S \sim P^m}[x \in SR^{\mathcal{L}}_{\ell}(S, \eta_1, \eta_2)]$.*

We will now combine our results on test-time attacks and distribution shift to give a general bound on the $\mathcal{P} \to \mathcal{Q}$ safely-reliable correctness for the different robust losses (Definition 1).

Theorem J.1. *Let $\mathcal{Q}$ be a realizable distribution shift of $\mathcal{P}$ with respect to $\mathcal{H}$, and $h^* \in \mathcal{H}$ be the target concept. There exist learners for robust losses $\ell_{CA}, \ell_{TL}, \ell_{ST}, \ell_{IA}$ with $\mathcal{P} \to \mathcal{Q}$ safely-reliable correctness given by*

(a) $PQR^{\mathcal{L}}_{CA}(S, \eta_1, \eta_2) = \Pr_{x \sim Q}[\mathbf{B}_{\mathcal{M}}(x, \eta_1) \cap \{z \mid h^*(z) = h^*(x)\} \subseteq \text{Agree}(\mathcal{H}_0(S))]$,

(b) $PQR^{\mathcal{L}}_{TL}(S, \eta_1, \eta_2) = \Pr_{x \sim Q}[\mathbf{B}_{\mathcal{M}}(x, \eta_1) \subseteq \text{Agree}(\mathcal{H}_0(S))]$,

(c) $PQR^{\mathcal{L}}_{ST}(S, \eta_1, \eta_2) = \Pr_{x \sim Q}[\mathbf{B}_{\mathcal{M}}(x, \eta_1) \subseteq \{z \mid \mathbf{B}^o(z, \eta) \subseteq \text{Agree}(\mathcal{H}_0(S)) \wedge h(x) = h(z),\ \forall\, x \in \mathbf{B}^o(z, \eta), h \in \mathcal{H}_0(S)\}]$,

(d) $PQR^{\mathcal{L}}_{IA}(S, \eta_1, \eta_2) = \Pr_{x \sim Q}[(\{z \mid \mathbf{B}(z, \eta_2) \subseteq \text{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z) = 1\}) \cup (\{z \mid \mathbf{B}(z, \eta_1) \subseteq \text{Agree}(\mathcal{H}_0(S))\} \cap \{z \mid h^*(z) = 0\})]$.

Proof. The proof follows by applying Theorems B.1 and B.2, and using Definitions 6 and 18. □

We consider an example when the training distribution $\mathcal{P}$ is isotropic log-concave and the test distribution $\mathcal{Q}_\mu$ is log-concave with its mean shifted by μ but the covariance matrix is still an identity matrix (see Figure 4, right).

Theorem J.2. *Let $\mathcal{P}, \mathcal{Q}$ be isotropic log-concave over $\mathbb{R}^d$. Let $\mathcal{Q}_\mu$ be a distribution after shifting the mean of $\mathcal{Q}$ by $\mu \in \mathbb{R}^d$. Let $\mathcal{H} = \{h : x \to \text{sign}(\langle w_h, x \rangle) \mid w_h \in \mathbb{R}^d, \|w_h\|_2 = 1\}$ be the class of linear separators. Let $\mathbf{B}(\cdot, \eta)$ be a L_2 ball perturbation with radius η. For $S \sim \mathcal{P}^m$, for $m = \mathcal{O}(\frac{1}{\varepsilon^2}(\text{VCdim}(\mathcal{H}) + \ln \frac{1}{\delta}))$, for an optimal robustly-reliable learner $\mathcal{L}$, we have*

(a) $\Pr_{\mathcal{Q}_\mu}(SR^{\mathcal{L}}_{TL}(S,\eta_1,\eta_2)) \geq 1 - 2(\eta_1 + \|\mu\|_2) - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability* $1-\delta$,
(b) $SR^{\mathcal{L}}_{CA}(S,\eta_1,\eta_2) = SR^{\mathcal{L}}_{TL}(S,\eta_1,\eta_2)$,
(c) $\Pr_{\mathcal{Q}_\mu}(SR^{\mathcal{L}}_{ST}(S,\eta_1,\eta_2)) \geq 1 - 2(\eta_1 + \eta_2 + \|\mu\|_2) - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability* $1-\delta$,
(d) $\Pr_{\mathcal{Q}_\mu}(SR^{\mathcal{L}}_{IA}(S,\eta_1,\eta_2)) \geq 1 - (\eta_1 + \eta_2 + 2\|\mu\|_2) - \tilde{\mathcal{O}}(\sqrt{d}\varepsilon)$ *with probability* $1-\delta$.

The $\tilde{\mathcal{O}}$-notation suppresses dependence on logarithmic factors and distribution-specific constants.

Proof. From triangle inequality, we know that $\mathbf{B}(x-\mu, r+\|\mu\|_2) \supseteq \mathbf{B}(x,r)$. We can simply extend the proofs from Theorem 3.3. Recall that $\mathrm{SR}^{\mathcal{L}}_{\mathrm{CA}}(S,\eta_1,\eta_2) = \mathrm{SR}^{\mathcal{L}}_{\mathrm{TL}}(S,\eta_1,\eta_2) = \{x \mid \mathbf{B}(x,\eta_1) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\} \supseteq \{x \mid \mathbf{B}(x-\mu,\eta_1+\|\mu\|_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\}$ and $\mathrm{SR}^{\mathcal{L}}_{\mathrm{ST}}(S,\eta_1,\eta_2) = \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x,\eta_1+\eta_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\} \supseteq \{x \in \mathcal{X} \mid \mathbf{B}_{\mathcal{M}}(x-\mu,\eta_1+\eta_2+\|\mu\|_2) \subseteq \mathrm{Agree}(\mathcal{H}_0(S))\}$. When x is drawn from a distribution $\mathcal{Q}_\mu$, we know that $x-\mu$ follows a distribution $\mathcal{Q}$ which is isotropic log-concave. We can apply Lemma D.1 to bound the probability mass of the safely-reliable region under $\mathcal{Q}_\mu$. Similarly, we can do the same for $\mathrm{SR}^{\mathcal{L}}_{\mathrm{IA}}(S,\eta_1,\eta_2)$. □

Similar bounds on reliability under robustness transfer may be given for linear separators under more general source or target distributions, including isotropic β-log-concave or s-concave distributions, as well as for concept classes with smooth classification boundaries, by applying Theorem J.1 to the examples from previous sections.

K Agnostic setting

Proof of Theorem 7.1. The robustly-reliable learner $\mathcal{L}$ is given as follows. Set $h^{\mathcal{L}}_S = \mathrm{argmin}_{h\in\mathcal{H}} \mathrm{err}_S(h)$ i.e. an ERM over S, and $r^{\mathcal{L}}_S(z) = \infty$ if $z \in \mathrm{Agree}(\mathcal{H}_\nu(S))$, else $r^{\mathcal{L}}_S(z) = -1$. To study the robustly-reliable region, we assume there is some concept $h^* \in \mathcal{H}$ which satisfies $\mathrm{err}_S(h^*) \leq \nu$. By definition of ERM, $\mathrm{err}_S(h^{\mathcal{L}}_S) \leq \mathrm{err}_S(h^*) = \nu$, or $h^{\mathcal{L}}_S \in \mathcal{H}_\nu(S)$. We first show that $\mathcal{L}$ is robustly-reliable. For $z \in \mathcal{X}$, if $r^{\mathcal{L}}_S(z) = \eta > 0$, then $z \in \mathrm{Agree}(\mathcal{H}_\nu(S))$. We have $h^*(z) = h^{\mathcal{L}}_S(z)$ since the classifiers $h^*, h^{\mathcal{L}}_S \in \mathcal{H}_\nu(S)$ and z lies in the agreement region of classifiers in $\mathcal{H}_\nu(S)$ in this case. Thus, we have $\ell^{h^*}_{\mathrm{CA}}(h^{\mathcal{L}}_S, x, z) = 0$ for any x such that $z \in \mathbf{B}^o_{\mathcal{M}}(x,\eta)$. In the $\eta = 0$ case, $h^*(z) = h^{\mathcal{L}}_S(z)$ by definiton and the same argument applies. Therefore, $\mathrm{RR}^{\mathcal{L}}_{\mathrm{CA}}(S,\nu,\eta) \supseteq \mathrm{Agree}(\mathcal{H}_\nu(S))$ for all $\eta \geq 0$ follows from the setting $r^{\mathcal{L}}_S(z) = \infty$ if $z \in \mathrm{Agree}(\mathcal{H}_\nu(S))$.

Conversely, let $z \in \mathrm{DIS}(\mathcal{H}_\nu(S))$. There exist $h_1, h_2 \in \mathcal{H}_\nu(S)$ such that $h_1(z) \neq h_2(z)$. By definition, robustly-reliable learning with $\eta = 0$ is not possible for z. If possible, let there be a robustly-reliable learner $\mathcal{L}$ such that $z \in \mathrm{RR}^{\mathcal{L}}_{\mathrm{CA}}(S,\nu,\eta)$ for some $\eta > 0$. By definition of the robust-reliability region, we must have $r^{\mathcal{L}}_S(z) > 0$. By definition of a ball, we have $z \in \mathbf{B}^o_{\mathcal{M}}(z,\eta)$ for any $\eta > 0$, and therefore $\ell^{h^*}_{\mathrm{CA}}(h^{\mathcal{L}}_S, z, z) = 0$ for every $h^* \in \mathcal{H}$ such that $\mathrm{err}_S(h^*) \leq \nu$. But then we must have $h^{\mathcal{L}}_S(z) = h^*(z)$ by definition of ℓ_{CA}. But we can set $h^* = h_1$ or $h^* = h_2$ since both are in $\mathcal{H}_\nu(S)$. But $h_1(z) \neq h_2(z)$, and therefore $h^{\mathcal{L}}_S(z) \neq h^*(z)$ for one of the above choices for h^*, contradicting that $\mathcal{L}$ is robustly-reliable. □