Regularized Robustly Reliable Learners and Instance Targeted Attacks

Avrim Blum [1] Donya Saless [1]

[1]Toyota Technological Institute at Chicago

A Robustly Reliable Learner

Given $\mathcal{H}$ the hypothesis class, and the training set, $\mathcal{S}$, they output a prediction y, and η with the following guarantee: As long as the adversary's budget is less than η, and h^* belongs to $\mathcal{H}$, y is the correct prediction, [1].

Regulization: Previous Approach versus Ours

We address two key limitations:

1. **Their definition becomes vacuous** for highly flexible hypothesis classes. We introduce *Regularized* Robustly-Reliable learners, enabling meaningful guarantees that depend on the complexity level.
2. **Their generic algorithm requires *retraining*** for each test point. We design algorithms that can produce their guarantee in time sublinear in training time, by using techniques from dynamic algorithm design, in some cool cases.

Regularized Robustly Reliable Learner

Definition 1. A learner $\mathcal{L}$ is regularized-robustly-reliable with respect to complexity measure $\mathcal{C}$ if, given training set S', the learner outputs a function $\mathcal{L}_{S'} : \mathcal{X} \times \mathbb{Z}^{\geq 0} \to \mathcal{Y} \times \mathbb{R} \times \mathbb{R}$ with the following properties: Given a test point x_{test}, and mistake budget b, $\mathcal{L}_{S'}(x_{\text{test}}, b)$ outputs a label y along with complexity levels $c_{\text{low}}, c_{\text{high}}$ such that

(a) There exists a classifier h of complexity c_{low} with at most b mistakes on S' such that $h(x_{test}) = y$,

(b) There is no classifier h' of complexity less than c_{high} with at most b mistakes on S' such that $h'(x_{test}) \neq y$.

Generic Algorithm

1. Given S', find the classifier $h_{S'}$ of minimum complexity that makes at most b mistakes on S'.

2. Given test point x_{test}, output (y, c_{low}, c_{high}) where $y = h_{S'}(x)$, $c_{low} = \mathcal{C}(h_{S'})$, and $c_{high} = \min\{\mathcal{C}(h) : h \text{ makes at most } b \text{ mistakes on } S' \text{ and } h(x) \neq h_{S'}(x)\}$.

We propose ways to make this generic algorithm efficient. See the paper :-)

A Simple Example

Definition 2. The number of alterations of a function $f : \mathbb{R} \to \{-1, +1\}$ is the number of times the function's output changes between +1 and -1 as the input variable increases from negative to positive infinity.

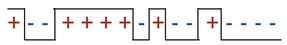

Figure 1. Number of Alterations

The test point shows up. If the mistake budget is 0, what can we guarantee?

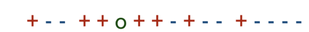

Figure 2. Test Point Arrives

Mistake Budget	Label	$(c_{\text{low}}, c_{\text{high}})$
$b = 0$	+	$[7, 9)$
$b = 1$	+	$[5, 7)$
$b = 2$	+	$[3, 5)$
$b = 3$	+	$[2, 4)$
$b = 4$	+	$[1, 3)$
$b = 5$	+	$[1, 2)$
$b = 6$	I don't know!	$\{1\}$
$b = 7, 8$	−	$[0, 1)$
$b = 9, 10, 11, 12, 13, 14, 15, 16$	I don't know!	$\{1\}$

Table 1. Guarantee for the Test Point and the Complexity Measure Number of Alterations.

Analyzing the Robustly Reliable Region

Definition 3. Given dataset S', poisoning budget b, and complexity bound c, the *optimal empirical regularized robustly reliable region* $\widehat{\text{OPTR}}^4(S', b, c)$ is the agreement region of the set of functions of complexity at most c that make at most b mistakes on S'. If there are no such functions, then $\widehat{\text{OPTR}}^4(S', b, c)$ is undefined.

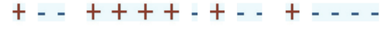

Figure 3. The blue regions depict $\widehat{\text{OPTR}}^4(S', 0, 8)$ for the number-of-alternations complexity measure, mistake budget $b = 0$, and complexity level $c = 8$.

See the paper for sample complexity bounds on the number of training examples needed in order for $\widehat{\text{OPTR}}^4$ to w.h.p. have a large probability mass! :-)

Some Examples of Other Measures

- Number of Alterations, Degree of Polynomials, **Data Independent**.
- **Local Margin, Test Data Dependent** Given a metric space $(\mathcal{M}, d_{\mathcal{M}})$, for a classifier with a decision function $h : \mathcal{X} \to \mathcal{Y}$, the local margin of the classifier with respect to a point $x^* \in \mathcal{X}$ is the distance between x^* and the nearest point $x' \in \mathcal{X}$ such that $h(x') \neq h(x^*)$.

$$r(h, x^*) = \inf_{\{x' \in \mathcal{X} : h(x') \neq h(x^*)\}} d(x^*, x')$$

We define the local margin complexity measure $\mathcal{C}(h, x^*)$ as $1/r(h, x^*)$.

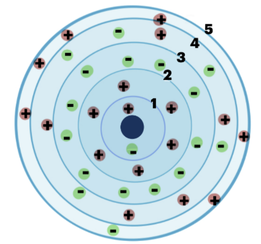

Figure 4. Local margin with x_{test} at the center

- **Global Margin, Train and Test Data Dependent** Given a metric space $(\mathcal{M}, d_{\mathcal{M}})$, a set $\tilde{S} = \{(x, y) | x \in \mathcal{X}, y \in \mathcal{Y}\}$, and a classifier $h : \mathcal{X} \to \mathcal{Y}$ that realizes $\tilde{S}$, we define the global margin of h with respect to $\tilde{S}$ as

$$r(h, \tilde{S}) = \min_{x_i \in \tilde{S}} \inf_{\{x' \in \mathcal{X} : h(x') \neq h(x_i)\}} d(x_i, x').$$

Figure 5. Global margin

See the paper for more! :-)

References

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

[2] Avrim Blum and Donya Saless. Regularized robustly reliable learners and instance targeted attacks. *arXiv preprint arXiv:2410.10572*, 2024.