Certified Robust Neural Networks: Generalization and Corruption Resistance

Amine Bennouna [1]   Ryan Lucas [1]   Bart Van Parys [1]

Abstract

Recent work have demonstrated that robustness (to "corruption") can be at odds with generalization. Adversarial training, for instance, aims to reduce the problematic susceptibility of modern neural networks to small data perturbations. Surprisingly, overfitting is a major concern in adversarial training despite being mostly absent in standard training. We provide here theoretical evidence for this peculiar "robust overfitting" phenomenon. Subsequently, we advance a novel distributionally robust loss function bridging robustness and generalization. We demonstrate both theoretically as well as empirically the loss to enjoy a certified level of robustness against two common types of corruption—data evasion and poisoning attacks—while ensuring guaranteed generalization. We show through careful numerical experiments that our resulting holistic robust (HR) training procedure yields state of the art performance. Finally, we indicate that HR training can be interpreted as a direct extension of adversarial training and comes with a negligible additional computational burden. A ready-to-use python library implementing our algorithm is available at `https://github.com/RyanLucas3/HR_Neural_Networks`.

1. Introduction

Recent work has shown that many modern neural networks are vulnerable to data corruption putting into question the reliability and performance of such models at deployment time. Robustness has in recent years, next to out-of-sample generalization, become a central research topic in the machine learning community. The key question is how to develop machine learning models which are reliable and perform well despite being trained on corrupted training data and evaluated on perturbed test data.

Data corruption can be due to multiple causes ranging from benign imprecise data collection to malicious adversaries. Two types of corruption, each associated with a distinct cause, stand out in prior work which we briefly discuss.

Evasion Corruption. This type of corruption results in small, often imperceptible, perturbations of all the test or train instances and may fool a naively trained model to suffer poor performance at deployment time. Goodfellow et al. (2014) study evasion attacks at test time and show that neural network models can suffer a severe decrease in accuracy when test instances are subjected to an imperceptible amount of noise. This pioneering work spurred on a rich literature on attacks (Tramer et al., 2020) as well as defense strategies such as the popular adversarial training (Madry et al., 2018; Bai et al., 2021). Evasion corruption need not necessarily be caused by an adversary or even take place at test time. Indeed, noise is present in training data as well often as a result of measurement imperfections. Examples include street noise and background chatter in speech recognition (Li et al., 2014; Mitra et al., 2017) or impulse noise and blur in image classification (Hendrycks & Dietterich, 2019; Hendrycks et al., 2021).

Poisoning Corruption. In this type of corruption, a certain fraction of the training data points is drastically changed. This alteration of the training dataset may then mislead the learning in selecting a "bad" model. Clearly, if a large majority of data points is affected, the data set is wholly corrupted and there is no point in hoping for a good performance. Hence, poisoning corruption is assumed to affect only a small fraction of the data points. One example is label flipping in classification problems, where a small portion of the training data points are wrongly labeled or an adversary can select a small number of labels to flip before training time (Biggio et al., 2012). Poisoning corruption has also been shown to considerably degrade performance at deployment time. For instance, Nelson et al. (2008) shows that altering only 1% of email spam labels can completely fool commercial spam filters. Several related attacks and defense mechanisms have been identified in follow-up work (Goldblum et al., 2022).

[1]Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Amine Bennouna <amineben@mit.edu>, Ryan Lucas <ryanlu@mit.edu>, Bart Van Parys <vanparys@mit.edu>.

Generalization. Beyond protecting against both discussed types of attacks, an equally important challenge is out-of-sample performance. A model generalizes if it provides protection against **statistical error** which we take here to mean the difference between the finite number of data points provided during training time as opposed to the full data distribution. Indeed, all models must avoid *overfitting* to a particular training set and generalize to the test data. Several classical approaches have been developed to address statistical error such as ℓ_1, ℓ_2 regularization, early stopping, and data augmentation. Such models seek not only a good expected performance under the randomness of the samples but also low variance. To generalization well, models need to avoid being overly sensitive to a particular training data and rather seek "satisfactory" performance on all "likely" data sets of the task at hand.

Generalization and robustness to each type of corruption have been extensively studied in isolation. However, the interaction between corruption protection and generalization has been largely ignored: existing "robust" approaches typically dismiss statistical error. Recent work has shown that robustness (to corruption) can be at odds with generalization, highlighting the need to consider corruption and statistical error simultaneously. Schmidt et al. (2018) and Rice et al. (2020) exhibit theoretical and empirical evidence that adversarial training (AT)—which provides robustness against evasive corruption—suffers from notably worse generalization than natural training. Rice et al. (2020) show in particular that the surprising generalization properties of neural networks, such as improved out-of-sample performance with larger models, vanishes under AT. Schmidt et al. (2018) indicate that "understanding the interactions between robustness, classifier model, and data distribution from the perspective of generalization is an important direction for future work". Hence, while AT aims to improve generalization performance by protecting against evasive corruption, it seems inadvertently to limit generalization by exacerbating the adverse effect of statistical error. This interesting phenomenon, dubbed *robust overfitting*, and its remedies is the topic of several recent papers (Wu et al., 2020; Li & Spratling, 2022; Yu et al., 2022). Similar observations were also noted in the case of poisoning, where poisoning corruption exacerbates overfitting (Zhang & Sabuncu, 2018; Hendrycks et al., 2019).

Besides generalization issues, existing robust approaches typically protect against one single type of corruption exclusively, either evasion or poisoning. However, in practical settings, we typically are unaware of what type of corruption, if any, plagues the data. Unsurprisingly, a model designed for a specific type of corruption risks considerably poor performance when deployed in a setting with a different type of corruption. Hence, to develop practical robust models, it is important to protect against both types of corruption, and

their potential combination.

1.1. Contributions

In this paper, we study generalization when learning with data affected by evasion and/or poisoning corruption on top of statistical error simultaneously.

We first provide theoretical evidence of the significance of the interaction between corruption and statistical error. We show that the generalization gap of AT can be decomposed into an empirical risk minimization (ERM) generalization gap and a positive term which we interpret as an *adversary shift* (Section 3). This result provides theoretical evidence to the robust overfitting phenomenon observed empirically by Rice et al. (2020) and indicates that AT indeed suffers from worse generalization than ERM. This theoretical observation illustrates that protecting against corruption only while ignoring statistical error can inadvertently yield worse generalization performance.

Second, we tackle the problem of robustness against both types of corruption and statistical error simultaneously. We do so by designing and optimizing a provable "tight" upper bound on the *test loss* when learning with data subject to corruption. This is in contrast to most prior work which is either empirically motivated or, like AT, merely optimizes empirical proxies to such upper bounds—dismissing statistical error. To design this upper bound, we construct a set around the corrupted training distribution which contains the testing distribution with high probability. A valid upper bound on the test loss can then be taken as the worst-case expectation of the loss over all distributions in this ambiguity set. We show that training a neural network model based on such associated distributionally robust optimization (DRO) objective function can be interpreted as a natural generalization to standard AT which protects against the two discussed corruption sources simultaneously and particularly enjoys strong generalization. Our holistic robust (HR) objective function has three parameters $\alpha \in [0, 1]$, $r \geq 0$ and $\mathcal{N}$, and is guaranteed to be an upper bound on the test performance with high probability $1 - e^{-rn+O(1)}$ when less then a fraction α of all n samples are tampered by poisoning, and the evasion corruption is bounded within the set $\mathcal{N}$. The latter guarantee certifies robustness for *any* desired level of evasion attacks ($\mathcal{N}$) and poisoning attacks (α): our out-of-sample deployment performance is at least as good as the estimated in-sample performance with high probability.

Finally, we propose a practical HR loss training algorithm for neural networks with negligible additional computational burden compared to classical AT. We conduct careful numerical experiments which illustrate the efficacy of HR training on both MNIST and CIFAR-10 datasets in all possible corruption settings: clean, affected by poisoning and/or evasive corruption. By using validation on the protection pa-

rameters $\alpha, r, \mathcal{N}$, HR training adapts to and protects against whatever corruption plagues the data including combination of evasion and poisoning, while ensuring strong generalization. In all settings, HR training achieves state of the art (SOTA) performance and outperforms classical evasion defense mechanisms (AT of Madry et al. (2018), TRADES of Zhang et al. (2019)), poisoning defense mechanisms (DPA of Levine & Feizi (2020); Wang et al. (2022)), and the combination of poisoning and evasion defenses. Furthermore, we replicate the experiments in Rice et al. (2020) and provide numerical evidence that HR training largely circumvents the robust overfitting phenomenon experienced by AT. Finally, we show empirically that the *training loss* of HR training is an upper bound on its *adversarial test loss* with high probability, providing a generalization guarantee. Notably, this observation does not hold for any of the prior robust approaches.

We release an open-source Python library (`https://github.com/RyanLucas3/HR_Neural_Networks`) that can be directly installed through `pip`. With our library, HR training can be used to provide robustness by changing essentially only one line of the training code. We also provide an extensive Colab tutorial on using HR in the github repository.

1.2. Related work

Robust Overfitting. Understanding and eliminating the robust overfitting phenomenon has been the topic of previous work. Rice et al. (2020) suggests that conventional remedies for overfitting cannot improve upon early stopping. To compensate for poor generalization of AT (Schmidt et al., 2018), data augmentation techniques such as semi-supervised learning (Alayrac et al., 2019; Carmon et al., 2019) and data interpolation (Lee et al., 2020; Chen et al., 2021) have been considered. Alternative approaches directly modify the AT loss function by reweighing data points (Wang et al., 2019; Zhang et al., 2021), mitigating memorization effects (Dong et al., 2021) or favoring large loss data (Yu et al., 2022). Kulynych et al. (2022) recently considers differential privacy as a way to ensure strong generalization. Previous works mostly attempt to eliminate robust overfitting through empirical insight. In our work, however, we present theoretical evidence with regards to the working of the robust overfitting mechanism (Theorem 3.2) and propose a disciplined approach at its elimination. Our proposed HR training is indeed a theoretically disciplined approach which comes with a theoretical certificate against overfitting (Theorem 4.1). On a practical level, HR can be understood as a smart reweighing of the standard adversarial examples considered in Madry et al. (2018) and in doing so equips standard AT against with protection against both poisoning and evasion attacks.

Distributionally Robust Optimization. Wasserstein distributionally robust optimization (WDRO) has recently received attention in the context of AT (Staib & Jegelka, 2017; Sinha et al., 2017; Wong et al., 2019; Bui et al., 2022). WDRO safeguards against attackers with a bound on the average perturbation applied to all data points as opposed to standard AT which protects against attacks in which the perturbation on each individual data point is bounded. WDRO however still suffers from fundamentally the same robust overfitting phenomenon as standard AT as it does not take into account statistical error. However, the Kullback-Leibler (KL) divergence ambiguity sets are known in the DRO literature to be superior at providing safeguards against statistical error (Lam, 2019; Van Parys et al., 2021) than their Wasserstein counterparts in the absence of data corruption. The Lévy-Prokhorov (LP) metric has been shown in the statistical literature (Hampel, 1971) to precisely captures certain types of data corruption. Recently, Bennouna & Van Parys (2022) proposed a novel ambiguity set combining both the KL and LP metrics in an attempt to safeguard against corruption and statistical error simultaneously. Contrary to the WDRO approach of Sinha et al. (2017), the holistic robust DRO formulation based on this novel ambiguity set is proven here to guarantee an upper bound on the adversarial test loss for *any* desired corruption level and *any* loss function. Moreover, the HR formulation is not harder to train than classical AT. This paper considers the ambiguity set of Bennouna & Van Parys (2022) in the context of evasion (at *test time*) and poisoning, and proves new statistical properties. In particular, as we show later, the HR ambiguity set provides *uniform* finite sample generalization guarantees, independent of the dimentionality of the model class—a crucial novel property in the context of deep learning.

2. Preliminaries: Learning under corruption

We formalize here the setting of learning under evasion and poisoning corruption. A typical underlying assumption in prior work is that the empirical distribution of the data is close to the out-of-sample distribution, neglecting statistical error and risking poor generalization. Here, we carefully consider both types of corruption and their interaction with statistical error.

Consider a learning problem with covariates $x \in \mathcal{X}$ and outputs $y \in \mathcal{Y}$ (e.g., labels in classification problems), following a joint distribution $\mathcal{D}$ with compact support $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ which represents the distribution of the clean data. We do not observe samples directly from this clean distribution $\mathcal{D}$ but rather up to a fraction α of these n independent samples are wholly corrupted (poisoning corruption). We hence observe the n samples $\{z_i := (x_i, y_i)\}_{i=1}^n \in \mathcal{Z}^n$ post corruption and denote with $\mathcal{D}_n$ their empirical distribution.
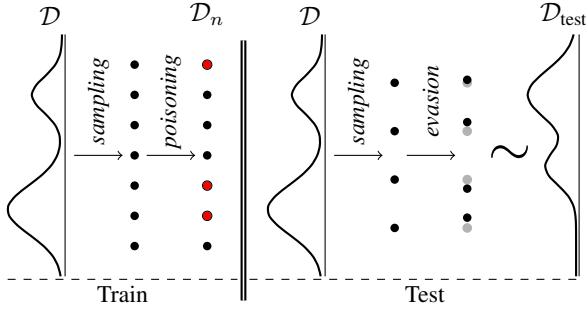
Figure 1. Illustration of the distributions $\mathcal{D}, \mathcal{D}_n$ and $\mathcal{D}_{\text{test}}$. Here, the red points indicate the modified samples by poisoning.

During test time, the test data is again sampled from $\mathcal{D}$ but is subjected to noise out of a set $\mathcal{N}$ (evasion corruption). Typical work in AT considers for example the noise bounded in an l_p ball (Madry et al., 2018). We denote the distribution of the perturbed test instances with $\mathcal{D}_{\text{test}}$. We observe $\mathcal{D}_n$ but our goal is to learn a model that minimizes the test loss

$$\mathcal{L}_{\text{test}}(\theta) := \mathbb{E}_{\mathcal{D}_{\text{test}}}[\ell(\theta, Z)],$$

where $\theta \in \Theta$ are the model parameters, ℓ is a bounded loss function (e.g., cross-entropy on neural networks); see Figure 1. Here, as in the remainder of the paper, $\mathbb{E}_{\mathcal{D}'}$ denotes expectation over a random variable Z with distribution $\mathcal{D}'$.

3. Adversarial training and generalization

We now highlight the importance of robustness against statistical error when learning under corruption. Consider the classical setting of adversarial evasion attacks in the absence of poisoning (i.e., $\alpha = 0$) and let here $\ell^{\mathcal{N}}(\theta, z) := \max_{\delta \in \mathcal{N}, \, z + \delta \in \mathcal{Z}} \ell(\theta, z + \delta)$ when the adversary can perturb the data z within the set $\mathcal{N} \ni 0$. The AT models of Madry et al. (2018) minimize the empirical adversarial loss $\mathcal{L}_{\text{AT}}(\theta) := \mathbb{E}_{\mathcal{D}_n}[\ell^{\mathcal{N}}(\theta, z)]$ and hope that its minimizer $\theta_{\mathcal{D}_n}^{\text{AT}}$ over the parameter set Θ is safeguarded against evasion attacks. We have indeed that the testing loss $\mathcal{L}_{\text{test}}(\theta)$ is upper bounded by $\mathbb{E}_{\mathcal{D}}[\ell^{\mathcal{N}}(\theta, z)]$ and hence minimizing its empirical counterpart $\mathcal{L}_{\text{AT}}(\theta)$ makes intuitive sense. However, due to statistical error there might be a considerable gap between the adversarial test loss $\mathbb{E}_{\mathcal{D}}[\ell^{\mathcal{N}}(\theta, z)]$ and the empirical adversarial loss $\mathcal{L}_{\text{AT}}(\theta)$ putting into question the intuitive appeal of this approach. AT typically does overfit to the worst-case perturbations of the *data points*—rather than the out-of-sample distribution—and thus generalizes poorly. This is akin to classical overfitting in ERM caused by the gap between the minimized loss $\mathcal{L}_{\text{ERM}}(\theta) := \mathbb{E}_{\mathcal{D}_n}[\ell(\theta, Z)]$ and the out-of-sample cost $\mathbb{E}_{\mathcal{D}}[\ell(\theta, Z)]$. In the case of AT, it has been empirically observed that overfitting is in fact exacerbated when compared to ERM (Rice et al., 2020). We now provide theoretical insight into why AT suffers from overfitting more than ERM.

First, we present a distributional perspective on AT which will be useful in what follows. Consider the ambiguity set

$$\mathcal{U}_{\mathcal{N}}(\mathcal{D}_n) := \{\mathcal{D}'_{\text{test}} : \text{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq 0\} \quad (1)$$

associated with the optimal transport metric

$$\text{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) := \quad (2)$$
$$\inf\left\{\int \mathbf{1}(z' - z \notin \mathcal{N}) \, \mathrm{d}\gamma(z, z') \; : \; \gamma \in \Gamma(\mathcal{D}_n, \mathcal{D}'_{\text{test}})\right\}$$

where $\Gamma(\mathcal{D}_n, \mathcal{D}'_{\text{test}})$ denotes the set of all couplings on $\mathcal{Z} \times \mathcal{Z}$ between $\mathcal{D}_n$ and $\mathcal{D}'_{\text{test}}$ (Villani, 2009). It is straightforward to observe that

$$\mathcal{L}_{\text{AT}}(\theta) = \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}_n)\}. \quad (3)$$

Clearly, hence, naively considering DRO formulations does not by itself eliminate robust overfitting. We will show in Section 4 that a careful generalization of the previous DRO formulation does protect against robust overfitting.

ERM overfitting gap. Let us first provide intuition on classical overfitting in ERM. As $\mathcal{D}_n$ is here the empirical distribution of a random dataset it is itself a random variable and we denote with $\mathcal{S}_n$ its distribution. Denote the associated ERM solution to some arbitrary distribution $\mathcal{D}'$ as $\theta_{\mathcal{D}'} \in \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}'}[\ell(\theta, Z)]$ which we assume to be unique for simplicity. ERM overfitting is bound to occur as the expected testing loss is larger than the expected training loss, i.e.,

$$\mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}}[\ell(\theta_{\mathcal{D}_n}, Z)]] = \mathbb{E}_{\mathcal{D}_n^1, \mathcal{D}_n^2 \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n^1}[\ell(\theta_{\mathcal{D}_n^2}, Z)]] \quad (4)$$
$$\geq \mathbb{E}_{\mathcal{D}_n^1, \mathcal{D}_n^2 \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n^1}[\ell(\theta_{\mathcal{D}_n^1}, Z)]] \quad (5)$$
$$= \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n}[\ell(\theta_{\mathcal{D}_n}, Z)]]. \quad (6)$$

Here, Equality (4) is due to Fubini's theorem. Inequality (5) follows from the fact that $\theta_{\mathcal{D}_n^1}$ is a minimizer of $\mathbb{E}_{\mathcal{D}_n^1}[\ell(\theta, Z)]$. Equality (6) follows from the fact that $\mathcal{D}_n^2$ and $\mathcal{D}_n^1$ are independent and share the same distribution $\mathcal{S}_n$. The right hand side of (6) is precisely the expected training loss. We define the ERM *overfitting gap*, between training and testing loss, as

$$\mathcal{G}_{\text{ERM}}(\mathcal{S}_n, \ell)$$
$$:= \mathbb{E}_{\mathcal{D}_n^1, \mathcal{D}_n^2 \sim \mathcal{S}_n}\left[\mathbb{E}_{\mathcal{D}_n^2}[\ell(\theta_{\mathcal{D}_n^1}, Z)] - \mathbb{E}_{\mathcal{D}_n^1}[\ell(\theta_{\mathcal{D}_n^1}, Z)]\right]$$
$$= \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}}[\ell(\theta_{\mathcal{D}_n}, Z)] - \mathbb{E}_{\mathcal{D}_n}[\ell(\theta_{\mathcal{D}_n}, Z)]].$$

This gap definition is general in that it characterizes overfitting for any data generation process $\mathcal{S}_n$. It also highlights what fundamentally causes overfitting in ERM: an expected difference in empirical distribution between training ($\mathcal{D}_n^1$) and test ($\mathcal{D}_n^2$) data.

AT overfitting gap. Let us now examine the robust overfitting phenomenon in AT. For an arbitrary distribution $\mathcal{D}'$ and an arbitrary model θ we define its associated worst-case adversary as the distribution $\mathcal{D}'(\theta) \in \arg\max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}')\}$. We denote in particular with $\mathcal{D}'^{\text{AT}} := \mathcal{D}'^{\text{AT}}(\theta^{\text{AT}}_{\mathcal{D}'})$ the worst-case adversary of the AT solution $\theta^{\text{AT}}_{\mathcal{D}'} \in \arg\min_{\theta \in \Theta} \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}')\}$ which minimizes $\mathcal{L}_{\text{AT}}$ (see (3)). We will assume in this section that the AT best model $\theta^{\text{AT}}_{\mathcal{D}'}$ and its associated worst-case adversary $\mathcal{D}'^{\text{AT}}$ form a saddle-point.

Assumption 3.1 (Saddle-Point). For any $\mathcal{D}'$ we have $\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta^{\text{AT}}_{\mathcal{D}'}, Z)] \leq \mathbb{E}_{\mathcal{D}'^{\text{AT}}}[\ell(\theta, Z)]$ for all $\theta \in \Theta$ and $\mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}')$.

The previous saddle point condition implies in particular that the AT best model also minimizes a nominal ERM cost $\mathbb{E}_{\mathcal{D}^{\text{AT}}_n}[\ell(\theta, Z)]$ with respect to a worst-case adversary $\mathcal{D}^{\text{AT}}_n \sim \mathcal{S}^{\text{AT}}_n$, i.e., $\theta^{\text{AT}}_{\mathcal{D}_n} = \theta_{\mathcal{D}^{\text{AT}}_n} \in \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}^{\text{AT}}_n}[\ell(\theta, Z)]$. Assumption 3.1 holds for any convex loss functions such as logistic regression (Sion, 1958; Bose et al., 2020). However, it is not necessarily verified for neural networks due to their inherent nonconvexity. Recent work by Gidel et al. (2021) suggests however that it may also hold for neural networks approximately. We define the *AT overfitting gap* as the expected difference between the adversarial testing loss and the AT loss, i.e.,

$$\mathcal{G}_{\text{AT}}(\mathcal{S}_n, \ell) := \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}}[\ell^{\mathcal{N}}(\theta^{\text{AT}}_{\mathcal{D}_n}, Z)] - \mathbb{E}_{\mathcal{D}_n}[\ell^{\mathcal{N}}(\theta^{\text{AT}}_{\mathcal{D}_n}, Z)]]$$
$$= \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}(\theta^{\text{AT}}_{\mathcal{D}_n})}[\ell(\theta^{\text{AT}}_{\mathcal{D}_n}, Z)] - \mathbb{E}_{\mathcal{D}^{\text{AT}}_n}[\ell(\theta^{\text{AT}}_{\mathcal{D}_n}, Z)]].$$

Our next result shows that under a saddle-point assumption, the AT overfitting gap can be decomposed into an ERM overfitting gap and a *nonnegative* term $\mathcal{G}_{\text{SHIFT}}(\mathcal{S}_n, \ell) := \mathbb{E}_{\mathcal{D}^1_n, \mathcal{D}^2_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}^2_n(\theta^{\text{AT}}_{\mathcal{D}^1_n})}[\ell(\theta^{\text{AT}}_{\mathcal{D}^1_n}, Z)] - \mathbb{E}_{\mathcal{D}^2_n(\theta^{\text{AT}}_{\mathcal{D}^2_n})}[\ell(\theta^{\text{AT}}_{\mathcal{D}^1_n}, Z)]]$.

Theorem 3.2. *Let Assumption 3.1 hold. Then,*

$$\mathcal{G}_{\text{AT}}(\mathcal{S}_n, \ell) = \mathcal{G}_{\text{ERM}}(\mathcal{S}^{AT}_n, \ell) + \underbrace{\mathcal{G}_{SHIFT}(\mathcal{S}_n, \ell)}_{\geq 0}.$$

The proof of Theorem 3.2 can be found in Appendix A and indicates in particular that AT overfitting gap is always at least as large as the overfitting gap suffered by an ERM procedure on data generated by the worst-case adversary $\mathcal{D}^{\text{AT}}_n$. Hence, while ERM overfitting is due to difference between train and test data sets, AT suffers worse overfitting as its gap is due to a difference between (adversarial) train and test datasets and an additional adversary shift term. This result gives theoretical evidence for the robust overfitting phenomenon observed empirically in a flurry of recent papers. The additional adversary shift term can be interpreted as the sensitivity of the AT solution to a change of adversary. Hence, Theorem 3.2 indicates that robust overfitting is due to the trained model adapting too much to the specific artificially added perturbations during AT.

Let us detail this adversary shift gap $\mathcal{G}_{\text{SHIFT}}$ interpretation. Recall that $\theta^{\text{AT}}_{\mathcal{D}_n}$ is an AT model trained with distribution $\mathcal{D}_n$. Intuitively, $\mathcal{D}^2_n$ can be seen as a testing distribution, and $\mathcal{D}^2_n(\theta^{\text{AT}}_{\mathcal{D}^1_n})$ (resp. $\mathcal{D}^2_n(\theta^{\text{AT}}_{\mathcal{D}^2_n})$) is a testing distribution with adversarial examples where the adversary is chosen against model $\theta^{\text{AT}}_{\mathcal{D}^1_n}$ (resp. $\theta^{\text{AT}}_{\mathcal{D}^2_n}$). Hence, each of the two terms of the gap measures the robust testing error under a *distinct* adversary. The difference measures therefore the gap of testing error when the adversary changes. This implies that the adversary shift quantifies the *sensitivity of the AT model to a change of adversary* resulting from a change of training distribution: if we train with data $\mathcal{D}^1_n$, AT trains with perturbation from adversary against $\theta^{\text{AT}}_{\mathcal{D}^1_n}$, while if we train with data $\mathcal{D}^2_n$, AT trains with perturbation from a distinct (shifted) adversary, namely against $\theta^{\text{AT}}_{\mathcal{D}^2_n}$. This gap is large when the trained model adapts too much (overfits) to the specific added perturbation to the training samples with AT.

4. A holistic robust loss

From the previous section, it is clear that practical robustness requires us to take into account statistical error in addition to poisoning and evasion corruption. The key idea, following the motivation behind AT, is to construct an upper bound on the test loss when learning with corruption. As opposed to classical AT, such upper bound should account for poisoning as well, and also, crucially, should bound the *out-of-sample* adversarial test loss $\mathcal{L}_{\text{test}}$ rather than its empirical proxy.

Let us first build intuition on our upper bound. As Section 3 demonstrate, an empirical proxy of the test loss typically underestimates the test loss of the trained model, as the expected overfitting gap is always positive $\mathcal{G}_{\text{ERM}}, \mathcal{G}_{\text{AT}} \geq 0$. This is caused by the fluctuations of the empirical distribution of samples around the out-of-sample distributions generating the samples. Hence, instead of considering the ambiguity set (1) of the AT loss (3) capturing only the corruption at hand, we augment this ambiguity set to capture the fluctuations of the empirical distribution due to sampling. These statistical fluctuations are captured by the Kullback-Leibler divergence (Van Erven & Harremos, 2014), denoted here as KL, as we will demonstrate shortly. On the other hand, the optimal transport metric LP (defined in (2)) will capture both evasion and poisoning. This augmentation will increase the estimated loss "just enough" to be a tight upper bound on the test loss.

Define here an ambiguity set $\mathcal{U}_{\mathcal{N}, \alpha, r}(\mathcal{D}_n) := \{\mathcal{D}'_{\text{test}} : \exists \mathcal{D}' \text{ s.t. } \text{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}') \leq \alpha, \text{KL}(\mathcal{D}', \mathcal{D}'_{\text{test}}) \leq r\}$, and the associated holistic robust (HR) loss as

$$\mathcal{L}^{\mathcal{N}, \alpha, r}_{\text{HR}}(\theta) := \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}, \alpha, r}(\mathcal{D}_n)\}, \quad (7)$$

where the desired protection against evasion corruption is controlled by $\mathcal{N}$, protection against poisoning corruption is

controlled by α, and statistical error is accounted for by r, respectively. Intuitively, the LP ball $\{\mathcal{D}' : LP_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}') \leq \alpha\}$ contains all the possible "training" empirical distributions $\mathcal{D}'$ directly sampled from the adversarial testing distribution $\mathcal{D}_{test}$—that is, by removing the poisoned data points and then perturbing each data point by evasion. As $\mathcal{D}'$ is an empirical distribution of samples from $\mathcal{D}_{test}$, then large deviation theory (Dembo & Zeitouni, 2009) ensures that the set $\{\mathcal{D}'_{test} : KL(\mathcal{D}', \mathcal{D}'_{test}) \leq r\}$ contains (intuitively) $\mathcal{D}_{test}$ with high probability $1 - e^{-rn + O(1)}$. This intuitive explanation is merely a simplified illustration. In the proof of subsequent Theorem 4.1 (which we defer to Appendix C) we prove indeed that the considered ambiguity set around the random observed empirical distribution contains the adversarial testing distribution (with probability $1 - e^{-rn + O(1)}$), given that less than a fraction α of all training data points are corrupted by poisoning and the evasive attack on the test data is limited to a compact set in the interior of the set $\mathcal{N}$.

Theorem 4.1 (Robustness Certificate). *Suppose that less than a fraction $\alpha \in (0, 1]$ of the independent identically distributed (IID) training data is poisoned and the evasion attack on the test set is limited to a compact set in $\mathrm{int}(\mathcal{N})$. Then, $Pr(\mathcal{L}_{HR}^{\mathcal{N}, \alpha, r}(\theta) \geq \mathcal{L}_{test}(\theta) \ \forall \theta \in \Theta) \geq 1 - e^{-rn + O(1)}$.*

Denote here the holistic robust solution as $\theta_{\mathcal{D}_n}^{\mathsf{HR}} \in \arg\min_{\theta \in \Theta} \max_{\mathcal{D}'_{\mathsf{test}} \in \mathcal{U}_{\mathcal{N}, \alpha, r}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'_{\mathsf{test}}}[\ell(\theta, Z)]$. The previous theorem certifies that asymptotically the testing error is not larger than the HR loss with high probability, i.e.,

$$\Pr\left(\mathcal{L}_{\mathsf{HR}}^{\mathcal{N}, \alpha, r}(\theta_{\mathcal{D}_n}^{\mathsf{HR}}) \geq \mathcal{L}_{\mathsf{test}}(\theta_{\mathcal{D}_n}^{\mathsf{HR}}) \right) \geq 1 - e^{-rn + O(1)}.$$

In Appendix C we prove a finite sample extension of this generalization certificate (Theorem C.4) and indicate that the results is statistically tight (Theorem C.5). Our proofs of these results are along the lines of techniques from Bennouna & Van Parys (2022). We point out that, unlike certificates founds in several prior works (for instance in Sinha et al. (2017)), our bound does not depend on the dimension of the parameter space Θ but rather only on the "size" of the event set $\mathcal{Z}$ (eg images and labels). This is critical as in the context of neural networks typically the number of weights is taken much larger than the number of data points n. The HR loss is a natural generalization of the AT loss. When no robustness against statistical error ($r = 0$) and poisoning corruption ($\alpha = 0$) is desired, we have indeed $\mathcal{L}_{\mathsf{HR}}^{\mathcal{N}, 0, 0}(\theta) = \mathcal{L}_{\mathsf{AT}}(\theta)$; See Appendix B for further details.

5. HR training algorithm

We now introduce an algorithm—HR training—to optimize efficiently the HR loss $\mathcal{L}_{\mathsf{HR}}^{\mathcal{N}, \alpha, r}$. At first glance, minimizing the HR loss seems challenging as it requires the solution of a saddle point problem over the model parameters $\theta \in \Theta$ and potentially continuous distribution $\mathcal{D}'_{\mathsf{test}} \in \mathcal{U}_{\mathcal{N}, \alpha, r}(\mathcal{D}_n)$. We

exploit, however, a finite reformulation of the HR loss by Bennouna & Van Parys (2022). For any given training data (sub)set $\{z_i = (x_i, y_i)\}_{i \in I}$ with $I \subseteq [n]$, we can compute its associated HR loss *exactly* as

$$\mathcal{L}_{\mathsf{HR}}^{\mathcal{N}, \alpha, r}(\theta) = \begin{cases} \max \sum_{i=1}^n d'_i \ell^{\mathcal{N}}(\theta, z_i) + d'_0 \ell^{\mathcal{Z}}(\theta) \\ \text{s.t. } d' \in \Re_+^{n+1}, \ \hat{d}' \in \Re_+^{n+1}, \ s \in \Re_+^n, \\ \quad \sum_{i=0}^n d'_i = \sum_{i=0}^n \hat{d}'_i = 1, \\ \quad \hat{d}'_i + s_i = \frac{1}{n} \ \forall i \in [1, \dots, n], \\ \quad \sum_{i=0}^n \hat{d}'_i \log(\frac{\hat{d}_i}{d'_i}) \leq r, \ \sum_{i=1}^n s_i \leq \alpha \end{cases}$$

$$(8)$$

where $\ell^{\mathcal{Z}}(\theta) := \max_{z \in \mathcal{Z}} \ell(\theta, z)$ and $\ell^{\mathcal{N}}(\theta, z) := \max_{\delta \in \mathcal{N}, z + \delta \in \mathcal{Z}} \ell(\theta, z + \delta)$. In particular, the supremum of the HR loss (7) is attained in distributions $\mathcal{D}'_{test}$ and $\mathcal{D}'$ of finite support (d' and $\hat{d}'$ respectively in (8)), specifically the adversarial examples and a point maximizing the loss. The HR loss is therefore in essence simply an adversarial reweighing of loss terms comprising the AT loss and an additional loss term $\ell^{\mathcal{Z}}(\theta)$. Determining the HR loss exactly requires evaluating the adversarial loss $\ell^{\mathcal{N}}(\theta, z)$ and subsequently solving the exponential cone problem (8) with $O(n)$ variables and constraints, which can be done efficiently by off-the-shelf optimization solvers.

Practically, we train the HR loss by mimicking standard adversarial training. At each minibatch iteration, we indeed first compute an approximate adversarial example $z'_i \approx z_i + \arg\max_{\delta \in \mathcal{N}, z_i + \delta \in \mathcal{Z}} \ell(\theta, z_i + \delta)$ for each data point in the minibatch $I \subset [n]$ using the projected gradient descent (PGD) algorithm of Madry et al. (2018). With the help of the approximations $\ell^{\mathcal{N}}(\theta, z_i) \approx \ell(\theta, z'_i)$ as well as $\ell^{\mathcal{Z}}(\theta) \approx \max_{i \in I} \ell(\theta, z'_i)$ we determine an approximate maximizer $d'^{\star}$ in problem (8). The model parameters are then updated by stepping in the direction of the negative gradient of the HR loss. We remark that this gradient can be computed efficiently using Danskin's theorem as a weighted sum of the gradients on the adversarial examples, i.e., $\nabla_\theta \mathcal{L}_{\mathsf{HR}}^{\mathcal{N}, \alpha, r}(\theta; I) = \sum_{i \in I} d'^{\star}_i \nabla_\theta \ell(\theta, z'_i) + d'^{\star}_0 \nabla_\theta \ell(\theta, z'_{\arg\max_{i \in I} \ell(\theta, z'_i)})$. The algorithm is summarized in Algorithm 1 where we remark that HR training is different from standard adversarial training only in that it requires computing the maximizer $d'^{\star}$ in (8). Practically, HR training with PGD takes less than 6% more time than standard adversarial training using PGD in typical benchmarks.

6. Experiments

In this section, we investigate the efficacy of HR training through extensive experiments on the MNIST and

Algorithm 1 Holistic Robust Training

Specification: Learning rate λ, number of epochs T, mini-batches B, minibatch replays M, HR parameters $(\mathcal{N}, \alpha, r)$.

Input: Data $\{z_i = (x_i, y_i)\}_{i \in [n]}$. Initialized θ.

//iterate M times per batch, for a total of T epochs

for $t \in [T/M]$, *minibatch* $I \in B$, $\_ \in [M]$ **do**

 //adversarial attack, e.g., using PGD

 Compute $z_i' \approx z_i + \arg\max_{\delta \in \mathcal{N}} \ell(\theta, z_i + \delta) \, \forall i \in I$,

 //reweigh data points

 Compute optimal weights $d'^\star$ in problem (8) ,

 //update model with gradient descent

 $\theta \leftarrow \theta - \lambda \nabla_\theta \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta; I)$

end

Output: Model parameters θ after T epochs.

CIFAR-10 datasets[1]. We consider the crossentropy loss $\ell(\theta, (x, y)) = -\sum_{k=1}^{K} y^{(k)} \log f_\theta^{(k)}(x)$, $K = 10$, where $f_\theta(x) \in \Re^K$ represents the network with weights θ and softmax output probabilities for covariate x and one-hot encoded class label $y \in \Re^K$. We present here only the CIFAR-10 results and leave the MNIST experiments, which bear the same insights, to Appendix D. We conduct three sets of experiments. The first set evaluates each type of robustness separately and shows that each parameter provides robustness against a distinct source of overfitting: $\mathcal{N}$ against data evasion, α against data poisoning and finally r against statistical error. The second set of experiments shows that HR training does not suffer robust overfitting and also benchmarks HR training against SOTA evasion robustness methods. The last set of experiments evaluates robustness against statistical error and both corruptions simultaneously and shows that HR training significantly outperforms benchmarks. In particular, each of these sets of experiments shows the theoretical robustness certificate provided in Theorem 4.1 holds empirically.

In these experiments, we train HR with PGD (see Algorithm 1) and adopt the classical setting of covariates evasion attacks by choosing $\mathcal{N} = B(0, \epsilon) \times \{0\}$ where depending on the context $B(0, \epsilon)$ is here either a Euclidean or infinity norm ball with radius ϵ.

6.1. Evaluating Robustness

In this set of experiments, we use a ResNet18 architecture for CIFAR-10. We detail the used architectures and network characteristics in Appendix D.1.

Robustness to Statistical Error ($\epsilon = 0, \alpha = 0, r \geq 0$). To study the robustness of HR to statistical error, we train HR with various parameters r and growing data size—associated with decreasing levels of statistical error. We
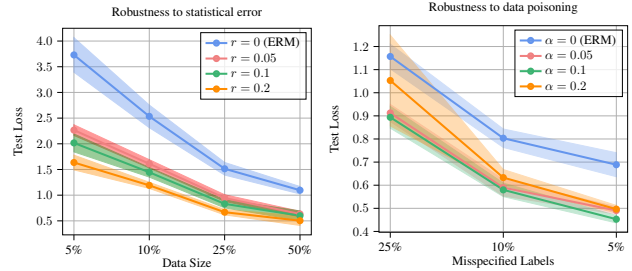

Figure 2. Robustness to statistical error / poisoning on CIFAR-10.

limit training data to $5\%, 10\%, 25\%$ and 50% of all available training data. We run 10 trials of HR training for each parameter value $r \in \{0, 0.05, 0.1, 0.25\}$ on such random subsampled training data. Figure 2 depicts the average and standard deviation of the testing loss across all trials. The parameter r (reflecting the use of KL distributional robustness) clearly improves performance, and its effect is considerably stronger with lower data sizes which are associated with more significant statistical error. Moreover, variance of the loss with respect to the random subsamples in each trail decreases with larger r. These observations are in line with the theory that KL distributional robustness provides efficient robustness against statistical error (Gotoh et al., 2018). In Appendix D.3.1 we present more extensive plots.

Robustness to Poisoning ($\epsilon = 0, \alpha \geq 0, r = 0$). To study HR robustness to poisoning, we evaluate HR training under label-flipping attacks. We select $0\%, 5\%, 10\%, 25\%$ of training images randomly and flip their labels uniformly at random to a different label. We then run 10 trials of HR training for each parameter value $\alpha \in \{0, 0.05, 0.1, 0.2\}$ on the resulting randomly corrupted data. Figure 2 present the average and standard deviation of the testing loss across all trials. HR with parameter α clearly improves performance when learning under poisoning corruption and its improvement is more significant with larger corruption levels. When α is chosen too large, the learned model is naturally overly conservative which decreases performance. We present further detailed plots of the results in Appendix D.3.1, showing that the learned model indeed provides a robustness certificate as the HR loss is an upper bound on the testing loss whenever α is chosen roughly larger or equal to the corruption level.

Robustness to Evasion ($\epsilon \geq 0, \alpha = 0, r = 0$) When $\alpha = 0$ and $r = 0$, the HR loss becomes the classical AT loss, and HR training with PGD (Algorithm 1) becomes the PGD algorithm of Madry et al. (2018) which has been well documented to provide robustness against evasion attacks.

---

[1]The code reproducing all experiments is available at: https://github.com/RyanLucas3/HR_Neural_Networks

6.2. Combatting robust overfitting

We replicate here the experiment of Rice et al. (2020) exhibiting robust overfitting for AT training. As in Rice et al. (2020) we use preactivation ResNet18 on CIFAR-10 with PGD attacks of 10 steps, and with SGD learning rate decay at epochs 100 and 150. Further experimental setup is discussed in Appendix D.1. We train AT with PGD of Madry et al. (2018), TRADES of Zhang et al. (2019), and HR with $r \geq 0$, $\epsilon > 0$ and $\alpha = 0$. For all algorithms, we use the same PGD attack on training with $\epsilon = 8/255$ and 10 attack steps. Figure 3 shows the test loss and error up to 200 training epochs when using pre-activation ResNet18 under ℓ_∞ attacks. Table 1 reports the test error at the 200 epoch mark, replicating the results reported in Table 2 in Rice et al. (2020). AT clearly exhibits robust overfitting as the test loss drastically increases with more training epochs. The test error also increases but perhaps not as drastically. This indicates that although the number of errors made eventually settles, worryingly, the errors are made with increasing confidence as quantified by the crossentropy loss. Similar to Rice et al. (2020), we find that TRADES (Zhang et al., 2019) without early stopping also experiences robust overfitting although to a lesser extent than regular PGD training. HR training with modest parameter values r effectively mitigates this phenomenon and results in a monotonically decreasing test loss as a function of training epochs eliminating the need for early stopping. It is worth noting that PGD best checkpoint performs better than HR final checkpoint. However, it might be practically hard to recover the performance of such best checkpoint with early stopping. In fact, PGD's training curve around the best checkpoint is very steep, and hence, early stopping might result in unstable performance on practical datasets, with potentially lower datasize.
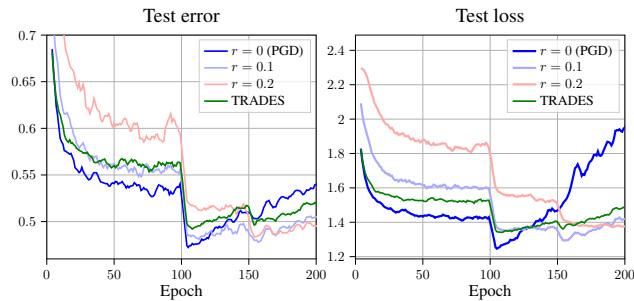

Figure 3. Standard AT training versus HR training test loss for evasion with $\epsilon = 8/255$ on CIFAR-10. Robustness to statistical error ($r > 0$) reduces robust overfitting experienced by AT.

6.3. Holistic Robustness

We finally compare the robustness of HR to benchmarks on datasets in all possible corruption settings: natural data

ROBUST TEST ERROR (%)		FINAL	BEST	DIFF
	E. Stopng.	46.9	46.7	0.2
Robust Methods	plain-PGD	53.9	46.8	7.2
	PGD+ℓ_1-Reg.	53.0	**46.4**	4.4
	PGD+ℓ_2-Reg.	55.2	48.6	8.8
	TRADES	52.0	48.8	3.2
	HR (Ours)	**49.6**	47.8	**1.8**

Table 1. Replicating Table 2 from (Rice et al., 2020). For each model, the optimal hyperparameters are selected as in Rice et al. (2020).
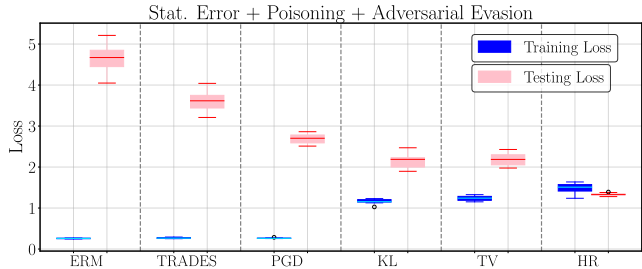

Figure 4. Training and adversarial testing loss of HR and various benchmarks in the presence of statistical error, data poisoning and evasion attacks. Training loss refers to the loss minimized during training while testing loss refers to the loss on the adversarially perturbed test set.

(no corruption), evasion corruption, evasion corruption with subsampling, poisoning corruption with subsampling, simultaneous evasion and poisoning corruption with subsampling. As in Section 6.1, subsampling allows to measure the effect of statistical error. For each setting, we test the performance of each algorithm on 10 trials of randomly sampled and corrupted data. In each trial, when statistical error is considered, we train algorithms with a random 25% subsample of the full dataset. When poisoning attacks are considered, we further randomly flip 10% of the training labels. The test loss performance is then evaluated with the help of a separate test set. When evasion is considered, the test set is adversarially perturbed by PGD-10 attacks with an ℓ_2 adversary of magnitude $\epsilon = 0.1$.

We benchmark HR with the following algorithms: (i) algorithms with no built-in robustness: ERM, (ii) algorithms designed to protect against evasion attacks: AT with PGD (Madry et al., 2018) and TRADES (Zhang et al., 2019) (iii) algorithms designed to protect against statistical error: KL DRO (Namkoong & Duchi, 2017); a particular case of HR with $\alpha = 0, \epsilon = 0$ (see Appendix B), (iv) algorithms designed to protect against poisoning corruption: DPA (Wang et al., 2022) and Total Variations (TV) DRO; particular case of HR with $r = 0, \epsilon = 0$ (see again Appendix B), (v) Combining evasion and poisoning defenses algorithms: combination of PGD and DPA; see Appendix D.2 for de-

		Natural (No Attack)	Evasion	Stat.Error + Poisoning	Stat.Error + Evasion	Stat.Error + Poisoning + Evasion
Combined Defenses	HR	**9.0%**	**18.5%**	**25.3%** (**27.1%**)	**33.0%** (**34.4%**)	**36.6%** (**38.9%**)
	PGD + DPA	18.6%	27.8%	37.0% (38.2%)	47.1% (48.0%)	45.2% (46.1%)
AT Defenses	PGD	10.6%	20.2%	30.1% (32.4%)	33.8% (35.3%)	40.6% (41.7%)
	TRADES	13.9%	32.8%	32.6% (34.3%)	43.3% (46.9%)	49.7% (52.0%)
Poisoning Defenses	DPA	15.6%	39.2%	35.4% (37.9%)	50.3% (51.4%)	49.6% (51.4%)
	TV	9.8%	40.8%	**25.3%** (**27.1%**)	43.9% (50.5%)	50.6% (54.4%)
No Defense	ERM	9.4%	35.4%	28.8% (30.5%)	43.9% (50.5%)	52.2% (55.5%)
	KL	**9.0%**	38.1%	28.4% (31.5%)	43.9% (50.5%)	49.9% (54.7%)

Table 2. Mean and (worst-case) error rate over 10 trials with model parameters chosen based on a validation set.

tails. The hyperparameters of each algorithm are selected based on a 70/30 train/validation split of the training data as detailed in Appendix D.2.

Results. Table 2 presents the natural and adversarial classification errors of each algorithm under various corruption settings. HR training significantly surpasses all other benchmarks in *every setting*. Traditional robust algorithms, such as PGD, TRADES, and DPA, unsurprisingly falter when faced with a corruption variant different from what they were originally designed to handle, performing even worse than ERM. In stark contrast, HR training demonstrates its adaptability by automatically adjusting to the specific corruption in the data, consistently maintaining high performance even when faced with combined attack strategies. Furthermore, the superior generalization capability of HR enables it to outperform each algorithm even in the unique attack settings for which they were specifically designed. Finally, HR also significantly outperforms the combination of defense techniques (PGD+DPA). Indeed, combining these defense techniques does not necessarily yield cumulative benefits due to the potential negative interactions between them. In contrast, HR is designed to provide effective simultaneous defenses while ensuring strong generalization.

Figure 4 shows the training loss and the adversarial test loss of each algorithm in the settings of combined evasion and poisoning. Here, DPA and DPA+PGD are not represented a they have no clear notion of loss. HR training significantly outperforms all other benchmarks in two distinct ways. First, HR training enjoys the smallest adversarial test loss. Furthermore, the variance of its adversarial test loss is the smallest among the benchmarked procedures. Second, HR training effectively provides a robustness certificate as alluded to in Theorem 4.1. The training loss is indeed with high probability an upper bound on the testing loss—an observation which fails to hold for the other procedures.

Training time. We investigate the train time of HR to evaluate the computational cost of its gain in robustness and generalization. Table 3 reports the training time (in

minutes) per epoch of ERM, PGD, and HR on various datasets and architectures. While HR takes four times longer than PGD for small datasets and architectures (ConvNet on MNSIT), its additional computational cost compared to PGD is negligible for large datasets and architectures: $<$**+6%** for CIFAR-10 with ResNet and TinyImageNet with ResNet/EfficientNet. This is due to the fact that computationally, HR training only differ with PGD in solving an additional conic optimization problem (8) as described in Algorithm 1. This optimization problem depends only on the batch size (linearly) and is *independent* on the data dimensionality and the network's size/architecture. As a result, while the PGD step and the network gradient step scale with the network's size and data dimensionality, HR's additional step does not scale with these factors, keeping its additional computational burden constant compared to PGD.

Architecture	Dataset	ERM	PGD	HR
ConvNet	MNIST	0.04	0.28	1.30
ResNet-18	CIFAR-10	0.24	2.60	2.75
ResNet-18	Tiny ImageNet	3.39	6.06	6.36
EfficientNet	Tiny ImageNet	7.65	39.59	40.87

Table 3. Runtime of ERM, PGD, and HR in minutes per epoch (one full pass of the training dataset).

7. Discussion & Conclusion

We make neural network training resilient against adversarial attacks based on a disciplined robust optimization approach. We remark that our approach can be generalized to other corruption models by substituting the LP metric with a metric suitable for the considered corruption model (e.g., Wasserstein for perturbations bounded on average). The tractability properties of the resulting method may however depend on the considered metric. We believe hence that the considered approach constitutes an interesting new inroad to bridging generalization and adversarial robustness of modern deep neural networks.

References

Alayrac, J.-B., Uesato, J., Huang, P.-S., Fawzi, A., Stanforth, R., and Kohli, P. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Bennouna, A. and Van Parys, B. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022.

Bennouna, M. and Van Parys, B. P. Learning and decision-making with data: Optimal formulations and phase transitions. *arXiv preprint arXiv:2109.06911*, 2021.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Bose, J., Gidel, G., Berard, H., Cianflone, A., Vincent, P., Lacoste-Julien, S., and Hamilton, W. Adversarial example games. *Advances in neural information processing systems*, 33:8921–8934, 2020.

Bui, T. A., Le, T., Tran, Q., Zhao, H., and Phung, D. A unified wasserstein distributional robustness framework for adversarial training. *arXiv preprint arXiv:2202.13437*, 2022.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

Chen, C., Zhang, J., Xu, X., Hu, T., Niu, G., Chen, G., and Sugiyama, M. Guided interpolation for adversarial training. *arXiv preprint arXiv:2102.07327*, 2021.

Cover, T. M. and Thomas, J. A. Information theory and statistics. *Elements of information theory*, 1(1):279–335, 1991.

Dembo, A. and Zeitouni, O. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

Dong, Y., Xu, K., Yang, X., Pang, T., Deng, Z., Su, H., and Zhu, J. Exploring memorization in adversarial training. *arXiv preprint arXiv:2106.01606*, 2021.

Gidel, G., Balduzzi, D., Czarnecki, W., Garnelo, M., and Bachrach, Y. A limited-capacity minimax theorem for non-convex games or: How i learned to stop worrying about mixed-nash and love neural nets. In *International Conference on Artificial Intelligence and Statistics*, pp. 2548–2556. PMLR, 2021.

Goldblum, M., Tsipras, D., Xie, C., Chen, X., Schwarzschild, A., Song, D., Madry, A., Li, B., and Goldstein, T. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gotoh, J.-y., Kim, M. J., and Lim, A. E. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters*, 46(4):448–452, 2018.

Hampel, F. R. A general qualitative definition of robustness. *The annals of mathematical statistics*, 42(6):1887–1896, 1971.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Huber, P. J. *Robust Statistics*. Wiley series in probability and mathematical statistics. John Wiley & Sons, 1981.

Kulynych, B., Yang, Y.-Y., Yu, Y., Błasiok, J., and Nakkiran, P. What you see is what you get: Principled deep learning via distributional generalization. In *Advances in Neural Information Processing Systems*, 2022.

Lam, H. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Lee, S., Lee, H., and Yoon, S. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 272–281, 2020.

Levine, A. and Feizi, S. Deep partition aggregation: Provable defense against general poisoning attacks. *arXiv preprint arXiv:2006.14768*, 2020.

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777, 2014.

Li, L. and Spratling, M. Understanding and combating robust overfitting via input loss landscape analysis and regularization. *Pattern Recognition*, pp. 109229, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Mitra, V., Franco, H., Stern, R. M., Hout, J. v., Ferrer, L., Graciarena, M., Wang, W., Vergyri, D., Alwan, A., and Hansen, J. H. Robust features in deep-learning-based speech recognition. In *New era for robust speech recognition*, pp. 187–217. Springer, 2017.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. *Advances in neural information processing systems*, 30, 2017.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I., Saini, U., Sutton, C., Tygar, J. D., and Xia, K. Exploiting machine learning to subvert your spam filter. *LEET*, 8(1):9, 2008.

Prokhorov, Y. V. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Sinha, A., Namkoong, H., Volpi, R., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Sion, M. On general minimax theorems. *Pacific Journal of mathematics*, 8(1):171–176, 1958.

Staib, M. and Jegelka, S. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, pp. 4, 2017.

Strassen, V. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Van Parys, B. P., Esfahani, P. M., and Kuhn, D. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009.

Wang, W., Levine, A. J., and Feizi, S. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, pp. 22769–22783. PMLR, 2022.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Wong, E., Schmidt, F., and Kolter, Z. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pp. 6808–6817. PMLR, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. S. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

A. Theory of overfitting

In this section we present theoretical insights into AT overfitting and its connection to ERM overfitting. Recall the notations of Section 3. Recall that ERM overfitting gap, between training and testing loss, for a data generation process $\mathcal{S}_n$ can be written as

$$\mathcal{G}_{\mathrm{ERM}}(\mathcal{S}_n,\ell) = \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}[\ell(\theta_{\mathcal{D}_n},Z)]}_{\text{test loss}} - \underbrace{\mathbb{E}_{\mathcal{D}_n}[\ell(\theta_{\mathcal{D}_n},Z)]}_{\text{train loss}}\Big] = \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\big[\mathbb{E}_{\mathcal{D}_n^2}[\ell(\theta_{\mathcal{D}_n^1},Z)] - \mathbb{E}_{\mathcal{D}_n^2}[\ell(\theta_{\mathcal{D}_n^2},Z)]\big] \geq 0. \quad (9)$$

Here $\mathcal{D}_n^2$ can be interpreted as the empirical distribution of random test data and $\mathcal{D}_n^1$ as the empirical distribution of training data which are equal in distribution but independent. The model $\theta_{\mathcal{D}_n^1}$ is trained in the context of $\mathcal{D}_n^1$ and tested with the help of $\mathcal{D}_n^2$ and subsequently compared to its training performance on $\mathcal{D}_n^2$. Hence, the ERM gap is fundamentally due to the distribution difference (statistical error) between the empirical distribution $\mathcal{D}_n^1$ of the training data and the empirical distribution $\mathcal{D}_n^2$ of the test data.

Proof of Theorem 3.2. First, observe that we have

$$\begin{aligned}
\mathcal{G}_{\mathrm{AT}}(\mathcal{S}_n,\ell) :=& \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}(\theta_{\mathcal{D}_n}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n}^{\mathrm{AT}}, Z)]\Big] - \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n(\theta_{\mathcal{D}_n}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n}^{\mathrm{AT}}, Z)]\Big] \\
=& \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)]\Big] - \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}}, Z)]\Big] \\
=& \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)] - \mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)]\Big] \\
&+ \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)] - \mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}}, Z)]\Big]
\end{aligned}$$

Here the first equality uses Fubini's theorem with $\mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathcal{D}_n] = \mathcal{D}$ which applies as $\mathbb{E}_{\mathcal{D}}[\ell^{\mathcal{N}}(\theta, z)]$ is assumed finite for any $\theta \in \Theta$ and we have $\mathbb{E}_{\mathcal{D}(\theta)}[\ell(\theta, z)] = \mathbb{E}_{\mathcal{D}}[\ell^{\mathcal{N}}(\theta, z)] = \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n}[\ell^{\mathcal{N}}(\theta, z)]] = \mathbb{E}_{\mathcal{D}_n \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n(\theta)}[\ell(\theta, z)]]$ for every $\theta \in \Theta$ as well. The second equality follows by adding and subtracting $\mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)]]$.

Using the notations $\mathcal{D}_n^1(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}) = \mathcal{D}_n^{1,\mathrm{AT}} \sim \mathcal{S}_n^{\mathrm{AT}}$ and $\mathcal{D}_n^2(\theta_{\mathcal{D}_n^2}^{\mathrm{AT}}) = \mathcal{D}_n^{2,\mathrm{AT}} \sim \mathcal{S}_n^{\mathrm{AT}}$ we get

$$\begin{aligned}
\mathcal{G}_{\mathrm{AT}}(\mathcal{S}_n,\ell) =& \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)] - \mathbb{E}_{\mathcal{D}_n^{2,\mathrm{AT}}}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)]\Big] \\
&+ \mathbb{E}_{\mathcal{D}_n^{1,\mathrm{AT}} \sim \mathcal{S}_n^{\mathrm{AT}}}\mathbb{E}_{\mathcal{D}_n^{2,\mathrm{AT}} \sim \mathcal{S}_n^{\mathrm{AT}}}\Big[\mathbb{E}_{\mathcal{D}_n^{2,\mathrm{AT}}}[\ell(\theta_{\mathcal{D}_n^{1,\mathrm{AT}}}, Z)] - \mathbb{E}_{\mathcal{D}_n^{2,\mathrm{AT}}}[\ell(\theta_{\mathcal{D}_n^{2,\mathrm{AT}}}, Z)]\Big] \\
=& \mathbb{E}_{\mathcal{D}_n^1 \sim \mathcal{S}_n}\mathbb{E}_{\mathcal{D}_n^2 \sim \mathcal{S}_n}\Big[\mathbb{E}_{\mathcal{D}_n^2(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}})}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)] - \mathbb{E}_{\mathcal{D}_n^{2,\mathrm{AT}}}[\ell(\theta_{\mathcal{D}_n^1}^{\mathrm{AT}}, Z)]\Big] + \mathcal{G}_{\mathrm{ERM}}(\mathcal{S}_n^{\mathrm{AT}},\ell)
\end{aligned}$$

where we exploit that $\theta_{\mathcal{D}'}^{\mathrm{AT}} = \theta_{\mathcal{D}'^{\mathrm{AT}}}$ for all $\mathcal{D}'$ due to saddle point condition in Assumption 3.1. $\qquad\square$

B. Further insights on the holistic robust loss

In this section, we detail further insights on the HR loss. Recall that the HR loss is defined for all $\theta \in \Theta$ as $\mathcal{L}_{\mathrm{HR}}^{\mathcal{N},\alpha,r}(\theta) := \max\{\mathbb{E}_{\mathcal{D}_{\mathrm{test}}'}[\ell(\theta, Z)] : \mathcal{D}_{\mathrm{test}}' \in \mathcal{U}_{\mathcal{N},\alpha,r}(\mathcal{D}_n)\}$ with ambiguity set

$$\mathcal{U}_{\mathcal{N},\alpha,r}(\mathcal{D}_n) := \{\mathcal{D}_{\mathrm{test}}' : \exists \mathcal{D}' \text{ s.t. } \mathrm{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}') \leq \alpha, \ \mathrm{KL}(\mathcal{D}', \mathcal{D}_{\mathrm{test}}') \leq r\}$$

with KL denoting here the Kullback-Leibler divergence $\mathrm{KL}(\mathcal{D}', \mathcal{D}_{\mathrm{test}}') = \int \log(d\mathcal{D}'/d\mathcal{D}_{\mathrm{test}}')(z)d\mathcal{D}'(z)$ and LP the optimal transport metric $\mathrm{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}') := \inf\{\int \mathbf{1}(z' - z \notin \mathcal{N})\,d\gamma(z,z') \ : \ \gamma \in \Gamma(\mathcal{D}_n, \mathcal{D}')\}$ where $\Gamma(\mathcal{D}_n, \mathcal{D}')$ denotes set of all couplings on $\mathcal{Z} \times \mathcal{Z}$ between $\mathcal{D}_n$ and $\mathcal{D}'$ (Villani, 2009). The HR loss on an arbitrary collection of data points

$\{z_i = (x_i, y_i)\}_{i=1}^n$ admits (Bennouna & Van Parys (2022), Theorem 3.5) the tractable convex reformulation

$$\mathcal{L}_{\text{HR}}^{\mathcal{N},\alpha,r}(\theta) = \begin{cases} \max \sum_{i=1}^n d'_i \ell^{\mathcal{N}}(\theta, z_i) + d'_0 \ell^{\mathcal{Z}}(\theta) \\ \text{s.t. } d' \in \Re_+^{n+1}, \ \hat{d}' \in \Re_+^{n+1}, \ s^n \in \Re_+^n, \\ \quad \sum_{i=0}^n d'_i = \sum_{i=0}^n \hat{d}'_i = 1, \ \hat{d}'_i + s_i = \frac{1}{n} \ \ \forall i \in [1, \ldots, n], \\ \quad \sum_{i=0}^n \hat{d}'_i \log(\hat{d}_i / d'_i) \leq r, \ \ \sum_{i=1}^n s_i \leq \alpha \end{cases} \tag{10}$$

where we recall $\ell^{\mathcal{N}}(\theta, z) := \sup_{\delta \in \mathcal{N}, z+\delta \in \mathcal{Z}} \ell(\theta, z + \delta)$ for all $z \in \mathcal{Z}$ and $\ell^{\mathcal{Z}}(\theta) := \sup_{z \in \mathcal{Z}} \ell(\theta, z)$ for all $\theta \in \Theta$. Recall that the standard AT loss is $\mathcal{L}_{\text{AT}}(\theta) = \sum_{i=1}^n d'_i \ell^{\mathcal{N}}(\theta, z_i)$ The HR loss is hence in essence a smart adversarial weighing of the classical AT loss terms but also includes an additional loss term $\ell^{\mathcal{Z}}(\theta)$.

In what remains of this section we discuss several interesting special cases of our HR loss functions.

Robustness to Evasion $(\mathcal{N} \neq \{0\}, \alpha = 0, r = 0)$**.** The set $\mathcal{N}$ characterizes the desired robustness to evasion attacks. Indeed, notice that in the special case of $\alpha = 0$ and $r = 0$, we have $\mathcal{U}_{\mathcal{N},0,0}(\mathcal{D}_n) = \{\mathcal{D}'_{\text{test}} \ : \ \text{LP}_{\mathcal{N}}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq 0\} = \{\mathcal{D}'_{\text{test}} \ : \ \exists \gamma \in \Gamma(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \text{ s.t. } \int \mathbf{1}(z' - z \notin \mathcal{N}) \, d\gamma(z, z') = 0\}$. Hence,

$$\mathcal{L}_{\text{HR}}^{\mathcal{N},0,0}(\theta) = \int \left[\max_{\delta \in \mathcal{N}, z+\delta \in \mathcal{Z}} \ell(\theta, z + \delta) \right] d\mathcal{D}_n(z) = \mathbb{E}_{\mathcal{D}_n}[\ell^{\mathcal{N}}(\theta, Z)] = \mathcal{L}^{\text{AT}}(\theta),$$

which is precisely the classical adversarial training loss proposed by (Madry et al., 2018). Recall that here $Z = (X, Y)$ represents both the covariate and labels. In practice, a common choice for the set $\mathcal{N}$ is $\{\delta \ : \ \|\delta\|_p \leq \epsilon\} \times \{0\}$ which allows for covariate noise bounded in an ℓ_p ball.

Robustness to Poisoning $(\mathcal{N} = \{0\}, \alpha > 0, r = 0)$**.** The parameter α sets the desired robustness to poisoning attacks when less than αn samples are tampered with by the adversary. Notice that with $\mathcal{N} = \{0\}$ and $r = 0$, we have $\mathcal{U}_{\{0\},\alpha,0}(\mathcal{D}_n) = \{\mathcal{D}'_{\text{test}} \ : \ \text{LP}_{\{0\}}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq \alpha\} = \{\mathcal{D}'_{\text{test}} \ : \ \text{TV}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq \alpha\}$ where TV is the total variation distance associated with the optimal transport cost $c(z, z') = \mathbf{1}(z \neq z')$. Hence,

$$\mathcal{L}_{\text{HR}}^{\{0\},\alpha,0}(\theta) = \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \text{TV}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq \alpha\}.$$

Robustness to Statistical Error $(\mathcal{N} = \{0\}, \alpha = 0, r > 0)$**.** The parameter r sets the desired robustness to statistical error, guaranteeing that the computed loss is an upper bound on the out-of-sample loss with probability $1 - e^{-rn+O(1)}$. When $\mathcal{N} = \{0\}$ and $\alpha = 0$, the HR loss becomes

$$\mathcal{L}_{\text{HR}}^{\{0\},0,r}(\theta) = \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \ \text{KL}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq r\},$$

which is classical robust formulation extensively considered by Lam (2019); Namkoong & Duchi (2017); Van Parys et al. (2021); Bennouna & Van Parys (2021).

Robustness to Evasion and Statistical Error $(\mathcal{N} \neq \{0\}, \alpha = 0, r > 0)$**.** The HR loss with $\alpha = 0$ protects against evasion attacks while ensuring strong generalization. The HR loss hence combats the robust overfitting phenomenon of AT with KL distributional robustness. In this case, the HR loss reduces to

$$\mathcal{L}_{\text{HR}}^{\mathcal{N},0,r}(\theta) = \max\left\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}\left[\ell^{\mathcal{N}}(\theta, Z)\right] \ : \ \text{KL}(\mathcal{D}_n, \mathcal{D}'_{\text{test}}) \leq r\right\}.$$

Robustness to Poisoning and Statistical Error $(\mathcal{N} = \{0\}, \alpha > 0, r > 0)$**.** The HR loss with $\mathcal{N} = \{0\}$ protects against poisoning attacks while ensuring strong generalization. In this case, the HR loss becomes

$$\mathcal{L}_{\text{HR}}^{\{0\},\alpha,r}(\theta) = \max\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)] : \exists \mathcal{D}' \text{ s.t. } \text{TV}(\mathcal{D}_n, \mathcal{D}') \leq \alpha, \ \text{KL}(\mathcal{D}', \mathcal{D}'_{\text{test}}) \leq r\}. \tag{11}$$

C. Proof of the robustness certificate of the holistic robust loss

In this section, we prove formally Theorem 4.1 and its finite sample version Theorem C.4. The goal is to show that the HR training loss $\mathcal{L}_{\text{HR}}^{\mathcal{N},\alpha,r}(\theta_{\mathcal{D}_n}^{\text{HR}})$ is an upper bound on the test loss $\mathcal{L}_{\text{test}}(\theta_{\mathcal{D}_n}^{\text{HR}})$ with high probability. A similar result was shown by (Bennouna & Van Parys, 2022) in the context of evasion in the train set. Here, evasion happens at test time.

Proof of Theorem 4.1. Let us denote with $\alpha' = \alpha - \delta_1$ and $\delta_1 > 0$ the fraction of the training data corrupted by poisoning and $0 \in \mathcal{N}' \subseteq \text{int}(\mathcal{N})$ the compact evasion attack set. Define $\delta_2 := \min_{n \in \mathcal{Z}, n' \in \mathcal{Z}}\{\|n - n'\| \; : \; n \notin \text{int}(\mathcal{N}), \; n' \in \mathcal{N}'\}$. We remark that as the feasible region is compact and the objective function continuous the minimum is achieved at some $n^\star \neq n'^\star$ and hence $\delta_2 = \|\delta - \delta'\| > 0$. Set now $\delta' = \min(\delta_1, \delta_2) > 0$ and note that $\alpha' + \delta' \leq \alpha$ and $\mathcal{N}' + B(0, \delta') \subseteq \mathcal{N}$. Theorem C.4 hence guarantees that

$$\Pr\left(\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta) \geq \mathcal{L}_{\text{test}}(\theta), \; \forall \theta \in \Theta\right) \geq 1 - (4/\delta')^{m(\mathcal{Z}, \delta')} \exp(-rn)$$

for all $n \geq 1$ where $m(\mathcal{Z}, \delta') = \min\{k \geq 0 \; : \; \exists z_1, \ldots, z_k \in \mathcal{Z} \text{ s.t. } \cup_{i=1}^k B(z_i, \delta') \supseteq \mathcal{Z}\}$ denotes the internal covering number of the support set $\mathcal{Z}$ with balls of radius δ'. Hence, in particular we have for all $n \geq 1$ the bound

$$\Pr\left(\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_{\mathcal{D}_n}^{\text{HR}}) \geq \mathcal{L}_{\text{test}}(\theta_{\mathcal{D}_n}^{\text{HR}})\right) \geq 1 - (4/\delta')^{m(\mathcal{Z}, \delta')} \exp(-rn)$$

from which the claim follows immediately. □

To prove the main finite sample result in Theorem C.4 the following composition lemma is useful.

Lemma C.1. *Let $\mathcal{D}$ and $\mathcal{D}'$ be two distributions supported on $\mathcal{Z}$ and suppose $0 \in \mathcal{N}_1$ and $0 \in \mathcal{N}_2$. We have*

$$\{\mathcal{D}'' \; : \; \exists \mathcal{D}' \text{ s.t. } LP_{\mathcal{N}_1}(\mathcal{D}, \mathcal{D}') \leq \alpha_1, \; LP_{\mathcal{N}_2}(\mathcal{D}', \mathcal{D}'') \leq \alpha_2\} = \{\mathcal{D}'' \; : \; LP_{\mathcal{N}_1 + \mathcal{N}_2}(\mathcal{D}, \mathcal{D}'') \leq \alpha_1 + \alpha_2\}$$

where here $\mathcal{N}_1 + \mathcal{N}_2$ denotes the Minkowski sum of $\mathcal{N}_1$ and $\mathcal{N}_2$.

Proof. Define $\mathcal{E} := \{\mathcal{D}'' \; : \; \exists \mathcal{D}' \text{ s.t. } LP_{\mathcal{N}_1}(\mathcal{D}, \mathcal{D}') \leq \alpha_1, \; LP_{\mathcal{N}_2}(\mathcal{D}', \mathcal{D}'') \leq \alpha_2\}$ and $\mathcal{E}' := \{\mathcal{D}'' \; : \; LP_{\mathcal{N}_1 + \mathcal{N}_2}(\mathcal{D}, \mathcal{D}'') \leq \alpha_1 + \alpha_2\}$. We first prove that $\mathcal{E} \subseteq \mathcal{E}'$. Observe that we have indeed

$$\begin{aligned}
\mathcal{E} =& \{\mathcal{D}'' \; : \; \exists Z_1 \sim \mathcal{D}, \; Z_2 \sim \mathcal{D}', \; Z_3 \sim \mathcal{D}'' \text{ s.t. } \Pr(Z_2 - Z_1 \in \mathcal{N}_1) \geq 1 - \alpha_1, \; \Pr(Z_3 - Z_2 \in \mathcal{N}_2) \geq 1 - \alpha_2\} \\
\subseteq& \{\mathcal{D}'' \; : \; \exists Z_1 \sim \mathcal{D}, \; Z_2 \sim \mathcal{D}', \; Z_3 \sim \mathcal{D}'' \text{ s.t. } \Pr(Z_2 - Z_1 \in \mathcal{N}_1, \; Z_3 - Z_2 \in \mathcal{N}_2) \geq 1 - \alpha_1 - \alpha_2\}. \\
\subseteq& \{\mathcal{D}'' \; : \; \exists Z_1 \sim \mathcal{D}, \; Z_3 \sim \mathcal{D}'' \text{ s.t. } \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) \geq 1 - \alpha_1 - \alpha_2\} = \mathcal{E}'
\end{aligned}$$

where the first inclusion follows from Fréchet inequality $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$.

We now prove that also $\mathcal{E}' \subseteq \mathcal{E}$. Consider an arbitrary distribution $\mathcal{D}'' \in \mathcal{E}'$. Hence, we can find two random variables $Z_1 \sim \mathcal{D}$ and $Z_3 \sim \mathcal{D}''$ so that $\Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) \geq 1 - \alpha_1 - \alpha_2$. Suppose that $z_3 - z_1 \in \mathcal{N}_1 + \mathcal{N}_2$ then by definition of the Minkowski sum we can always find z_2 such that $z_2 - z_1 \in \mathcal{N}_1$ and $z_3 - z_2 \in \mathcal{N}_2$. Hence, we can construct a random variable Z_2' so that $Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2 \implies Z_2' - Z_1 \in \mathcal{N}_1, \; Z_3 - Z_2' \in \mathcal{N}_2$. Let s be an independent Bernoulli random variable with success parameter $\alpha_1 / (\alpha_1 + \alpha_2)$. Construct now $Z_2'' = Z_1(1 - s) + Z_3 s$ and let finally $Z_2 = Z_2' \mathbf{1}(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + Z_2'' \mathbf{1}(Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2)$. Standard manipulations guarantee that

$$\begin{aligned}
\Pr(Z_2 - Z_1 \in \mathcal{N}_1) =& \Pr(Z_2' \mathbf{1}(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + Z_2'' \mathbf{1}(Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2) - Z_1 \in \mathcal{N}_1) \\
=& \Pr(Z_2' - Z_1 \in \mathcal{N}_1, \; Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(Z_2'' - Z_1 \in \mathcal{N}_1, \; Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2) \\
=& \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(Z_2'' - Z_1 \in \mathcal{N}_1, \; Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2) \\
=& \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 0, \; Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2) \\
=& \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 0)\Pr(Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2) \\
=& \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 0)(1 - \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2)) \\
=& \Pr(s = 0) + \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2)\Pr(s = 1) \\
\geq& (1 - \alpha_1 / (\alpha_1 + \alpha_2)) + (1 - \alpha_2 - \alpha_1)\alpha_1 / (\alpha_1 + \alpha_2) = 1 - \alpha_1
\end{aligned}$$

and similarly

$$\begin{aligned}
\Pr(Z_3 - Z_2 \in \mathcal{N}_2) =& \Pr(Z_3 - Z_2' \mathbf{1}(Z_1 - Z_3 \in \mathcal{N}_1 + \mathcal{N}_2) - Z_2'' \mathbf{1}(Z_1 - Z_3 \notin \mathcal{N}_1 + \mathcal{N}_2) \in \mathcal{N}_2) \\
=& \Pr(Z_3 - Z_2' \in \mathcal{N}_2, \; Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(Z_3 - Z_2'' \in \mathcal{N}_2, \; Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2)
\end{aligned}$$

$$=\Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(Z_3 - Z_2'' \in \mathcal{N}_2,\ Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2)$$
$$=\Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 1,\ Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2)$$
$$=\Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 1)\Pr(Z_3 - Z_1 \notin \mathcal{N}_1 + \mathcal{N}_2)$$
$$=\Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2) + \Pr(s = 1)(1 - \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2))$$
$$=\Pr(s = 1) + \Pr(Z_3 - Z_1 \in \mathcal{N}_1 + \mathcal{N}_2)\Pr(s = 0)$$
$$\geq (1 - \alpha_2/(\alpha_2 + \alpha_1)) + (1 - \alpha_2 - \alpha_1)\alpha_2/(\alpha_2 + \alpha_1) = 1 - \alpha_2.$$

Hence we have that $\mathcal{D}'' \in \mathcal{E}$ as well. As $\mathcal{D}''$ was arbitrary we also have that $\mathcal{E}' \subseteq \mathcal{E}$. $\qquad\square$

The goal now is to show that a slightly inflated KL ball captures the deviation of an empirical distribution from the out-of-sample distribution that generated it. Consider the random empirical distribution $\mathcal{D}_n''$ of n (uncorrupted) independent samples from a data generating distribution $\mathcal{D}$. It is well known that if a data generating distribution is continuous then the KL ball $\{\mathcal{D}'' : \mathrm{KL}(\mathcal{D}_n'', \mathcal{D}'') \leq r\}$ fails to be a confidence set of the data generating distribution $\mathcal{D}$. Indeed, the KL divergence between any distribution supported on a finite set of points and a continuous distribution is unbounded from above. Consequently, if the generating distribution $\mathcal{D}$ is continuous we have $\Pr(\mathcal{D} \in \{\mathcal{D}'' : \mathrm{KL}(\mathcal{D}_n'', \mathcal{D}'') \leq r\}) = 0$ for any $r \in \Re$.

The following results however states that a slightly inflated version of an appropriately sized KL ball around the empirical distribution of a clean training data set does serve as a confidence region of the data generating distribution.

Lemma C.2 (Dembo & Zeitouni (2009)). *Let $\mathcal{D}_n''$ be the empirical distribution of n independent samples with distribution $\mathcal{D}$ supported on a compact set $\mathcal{Z}$. Then, for all $\delta > 0$*

$$Pr(\mathcal{D} \in \{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } LP_{B(0,\delta)}(\mathcal{D}_n'', \mathcal{D}') \leq \delta,\ KL(\mathcal{D}', \mathcal{D}'') \leq r\}) \geq 1 - \left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \exp(-rn)$$

where $m(\mathcal{Z}, \delta) = \min\{k \geq 0 : \exists z_1, \ldots, z_k \in \mathcal{Z} \text{ s.t. } \cup_{i=1}^k B(z_i, \delta) \supseteq \mathcal{Z}\}$ denotes the internal covering number of the support set $\mathcal{Z}$.

Proof. For any given set $\mathcal{A}$, let $\mathcal{A}^\delta := \{\mathcal{D}' : \pi(\mathcal{D}'', \mathcal{D}') \leq \delta,\ \mathcal{D}'' \in \mathcal{A}\}$ denote the δ-inflation of the set $\mathcal{A}$ where π denotes here the Lévy-Prokhorov distance (Huber, 1981). We remark that here the LP balls $\mathcal{B}(\mathcal{D}') := \{\mathcal{D}'' : \pi(\mathcal{D}'', \mathcal{D}') \leq \delta\}$ are compact as π is continuous in the weak topology and $\mathcal{Z}$ is compact (Prokhorov, 1956). Dembo & Zeitouni (2009, Exercise 4.5.5) hence establish using a covering argument that for any set $\mathcal{A}$ and $\delta > 0$ we have for all $n \geq 1$ the upper bound

$$\Pr(\mathcal{D}_n'' \in \mathcal{A}) \leq m_{\mathrm{LP}}(\mathcal{A}, \delta) \exp\left(-n \inf_{\mathcal{D}' \in \mathcal{A}^\delta} \mathrm{KL}(\mathcal{D}', \mathcal{D})\right) \tag{12}$$

where $m_{\mathrm{LP}}(\mathcal{A}, \delta) = \min\{k \geq 0 : \exists \mathcal{D}_1, \ldots, \mathcal{D}_k \in \mathcal{A} \text{ s.t. } \cup_{i=1}^k \mathcal{B}(\mathcal{D}_i, \delta) \supseteq \mathcal{A}\}$ denotes the internal covering number of the set $\mathcal{A}$ with LP balls of radius δ.

Dembo & Zeitouni (2009, Exercise 6.2.19) also establish that with respect to the Lévy-Prokhorov distance the covering number of a subset $\mathcal{A}$ of the probability simplex on $\mathcal{Z}$ satisfies for all $\delta > 0$ the inequality $m_{\mathrm{LP}}(\mathcal{A}, \delta) \leq (4/\delta)^{m(\mathcal{Z},\delta)}$ where here $m(\mathcal{Z}, \delta)$ in turn denotes the covering number of the compact event set $\mathcal{Z}$ itself. Consider now a particular set $\mathcal{A}$ with complement $\{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } \pi(\mathcal{D}'', \mathcal{D}') \leq \delta,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r\}$. We hence have

$$\Pr(\mathcal{D} \in \{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } \mathrm{LP}_{B(0,\delta)}(\mathcal{D}_n'', \mathcal{D}') \leq \delta,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}'') \leq r\})$$
$$=\Pr(\mathcal{D} \in \{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } \pi(\mathcal{D}_n'', \mathcal{D}') \leq \delta,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}'') \leq r\})$$
$$=\Pr(\mathcal{D}_n'' \in \{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } \pi(\mathcal{D}'', \mathcal{D}') \leq \delta,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r\})$$
$$=1 - \Pr(\mathcal{D}_n'' \notin \{\mathcal{D}'' : \exists \mathcal{D}' \text{ s.t. } \pi(\mathcal{D}'', \mathcal{D}') \leq \delta,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r\})$$
$$=1 - \Pr(\mathcal{D}_n'' \in \mathcal{A})$$
$$\geq 1 - m_{\mathrm{LP}}(\mathcal{A}, \delta) \exp\left(-n \inf_{\mathcal{D}' \in \mathcal{A}^\delta} \mathrm{KL}(\mathcal{D}', \mathcal{D})\right)$$
$$\geq 1 - (4/\delta)^{m(\mathcal{Z},\delta)} \exp(-rn).$$

The first equality is due to a classical result of Strassen (1965) that implies $\{\mathcal{D}' : \mathrm{LP}_{B(0,\delta)}(\mathcal{D}_n'', \mathcal{D}') \leq \delta\} = \{\mathcal{D}' : \pi(\mathcal{D}_n'', \mathcal{D}') \leq \delta\}$. The first inequality uses the upper bound (12). The ultimate inequality follows from the implication

$\mathcal{D}' \in \mathcal{A}^\delta \implies \mathrm{KL}(\mathcal{D}', \mathcal{D}) > r$ which we now prove by contradiction. Suppose indeed that there exists $\mathcal{D}'' \in \mathcal{A}^\delta$ and $\mathrm{KL}(\mathcal{D}'', \mathcal{D}) \leq r$. From the definition of $\mathcal{A}^\delta$ there exists $\mathcal{D}' \in \mathcal{A}$ such that $\pi(\mathcal{D}'', \mathcal{D}') = \pi(\mathcal{D}', \mathcal{D}'') \leq \delta$. As we have both $\pi(\mathcal{D}', \mathcal{D}'') \leq \delta$ and $\mathrm{KL}(\mathcal{D}'', \mathcal{D}) \leq r$ it follows that $\mathcal{D}'$ is also in the complement of $\mathcal{A}$; we have reached a contradiction. $\quad\square$

The support function of a set $\mathcal{A}$ is defined as the function $h_{\mathcal{A}}(\ell) := \sup_{\mathcal{D}' \in \mathcal{A}} \int \ell(z) \, \mathrm{d}\mathcal{D}'(z)$. For all distribution $\mathcal{D}'$, denote $\mathcal{E}(\mathcal{D}') := \{\mathcal{D}'_{\text{test}} : \exists \mathcal{D} \text{ s.t. } \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r, \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}, \mathcal{D}'_{\text{test}}) \leq 0\}$ and $\mathcal{E}'(\mathcal{D}') := \{\mathcal{D}'_{\text{test}} : \exists \mathcal{D} \text{ s.t. } \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0, \, \mathrm{KL}(\mathcal{D}, \mathcal{D}'_{\text{test}}) \leq r\}$. We show that both sets have the same support function.

Lemma C.3. *For all distribution $\mathcal{D}'$, the support function of the sets $\mathcal{E}(\mathcal{D}')$ and $\mathcal{E}'(\mathcal{D}')$ coincide, i.e., $h_{\mathcal{E}(\mathcal{D}')}(\ell) = h_{\mathcal{E}'(\mathcal{D}')}(\ell)$ for all ℓ with $-\infty < \inf_{z \in \mathcal{Z}} \ell(z) \leq \sup_{z \in \mathcal{Z}} \ell(z) < \infty$ and $0 \in \mathcal{N}$.*

Proof. Remark that if $r = 0$ the result is trivial as then clearly $\mathcal{E}(\mathcal{D}') = \mathcal{E}'(\mathcal{D}') = \{\mathcal{D}'_{\text{test}} : \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}'_{\text{test}}) \leq 0\}$. Let now $r > 0$ and assume without loss of generality that $\inf_{z \in Z} \ell(z) \geq 0$. Indeed, we may remark that $h_{\mathcal{E}(\mathcal{D}')}(\ell - a) = h_{\mathcal{E}'(\mathcal{D}')}(\ell - a) = h_{\mathcal{E}(\mathcal{D}')}(\ell) - a = h_{\mathcal{E}'(\mathcal{D}')}(\ell) - a$ for all $a \in \Re$ and we can consider for instance $a = \inf_{z \in \mathcal{Z}} \ell(z)$.

Remark first that for all ℓ, we have

$$
\begin{aligned}
& h_{\mathcal{E}(\mathcal{D}')}(\ell) \\
&:= \sup\{\textstyle\int \ell(z) \, \mathrm{d}\mathcal{D}'_{\text{test}}(z) \, : \, \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r, \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}, \mathcal{D}'_{\text{test}}) \leq 0\} \\
&= \sup\{\textstyle\int \ell^{\mathcal{N}}(z) \, \mathrm{d}\mathcal{D}(z) \, : \, \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq r\} \\
&= \inf\{\alpha - \exp(\textstyle\int \log(\alpha - \ell^{\mathcal{N}}(z)) \, \mathrm{d}\mathcal{D}'(z) - r) : \sup_{z \in \mathcal{Z}} \ell^{\mathcal{N}}(z) \leq \alpha \leq \sup_{z \in \mathcal{Z}} \ell^{\mathcal{N}}(z)/(1 - e^{-r})\} \\
&= \inf\{\alpha - \exp(\textstyle\int \log(\alpha - \ell^{\mathcal{N}}(z)) \, \mathrm{d}\mathcal{D}'(z) - r) : \sup_{z \in \mathcal{Z}} \ell(z) \leq \alpha \leq \sup_{z \in \mathcal{Z}} \ell(z)/(1 - e^{-r})\}.
\end{aligned}
$$

The first equality is due to the observation that as remarked before $\max\{\int \ell(z) \, \mathrm{d}\mathcal{D}'_{\text{test}}(z) \, : \, \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}')\} = \int \ell^{\mathcal{N}}(z) \, \mathrm{d}\mathcal{D}'(u)$ for any distribution $\mathcal{D}'$. The second equality follows from a dual representation of the KL ball given in Proposition 5 by Van Parys et al. (2021) and holds for $r > 0$. The final equality follows from $\sup_{z \in \mathcal{Z}} \ell^{\mathcal{N}}(z) = \sup_{z \in \mathcal{Z}} \ell(z)$ as $0 \in \mathcal{N}$. Second, we observe that

$$
\begin{aligned}
& h_{\mathcal{E}'(\mathcal{D}')}(\ell) \\
&:= \sup\{\textstyle\int \ell(z) \, \mathrm{d}\mathcal{D}'_{\text{test}}(z) \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0, \, \mathrm{KL}(\mathcal{D}, \mathcal{D}'_{\text{test}}) \leq r\} \\
&= \sup\{\inf\{\alpha - \exp(\textstyle\int \log(\alpha - \ell(z)) \, \mathrm{d}\mathcal{D}(z) - r) : \sup_{z \in \mathcal{Z}} \ell(z) \leq \alpha \leq \sup_{z \in \mathcal{Z}} \ell(z)/(1 - e^{-r})\} \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\} \\
&= \inf\{\sup\{\alpha - \exp(\textstyle\int \log(\alpha - \ell(z)) \, \mathrm{d}\mathcal{D}(z) - r) : \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\} \, : \, \sup_{z \in \mathcal{Z}} \ell(z) \leq \alpha \leq \sup_{z \in \mathcal{Z}} \ell(z)/(1 - e^{-r})\} \\
&= \inf\{\alpha - \exp(\inf\{\textstyle\int \log(\alpha - \ell(z)) \, \mathrm{d}\mathcal{D}(z) \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\} - r) \, : \, \sup_{z \in \mathcal{Z}} \ell(z) \leq \alpha \leq \sup_{z \in \mathcal{Z}} \ell(z)/(1 - e^{-r})\} \\
&= \min\{\textstyle\int \lambda \log(\lambda/(\ell^{\mathcal{N}}(z) - \eta)) \, \mathrm{d}\mathcal{D}'(z) + (r - 1)\lambda + \eta \, : \, \eta \in \Re, \, \lambda \in \Re_+, \, \sup_{z \in \mathcal{Z}} \ell(z) \leq \eta\}.
\end{aligned}
$$

Here, the first equality from the same dual representation of the KL ball as used before. We remark that the objective function is convex and lower semicontinuous in α and concave in $\mathcal{D}'$. Hence, the second equality follows hence from a standard minimax theorem (c.f., Theorem 4.2 in Sion (1958)). The third equality uses the fact that the exponential function is nondecreasing. The final equality follows similarly from the fact the logarithm is an non-increasing function. Indeed, we can write $\inf\{\int \log(\alpha - \ell(z)) \, \mathrm{d}\mathcal{D}(z) \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\} = -\sup\{\int -\log(\alpha - \ell(z)) \, \mathrm{d}\mathcal{D}(z) \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\}$ which after using again $\max\{\int \ell(z) \, \mathrm{d}\mathcal{D}'_{\text{test}}(z) : \mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}}(\mathcal{D}')\} = \int \ell^{\mathcal{N}}(z) \, \mathrm{d}\mathcal{D}'(u)$ yields $-\int \sup_{\delta \in \mathcal{N}, z+\delta \in \mathcal{Z}} -\log(\alpha - \ell(z + \delta)) \, \mathrm{d}\mathcal{D}'(z) = -\int -\log(\alpha - \sup_{\delta \in \mathcal{N}, z+\delta \in \mathcal{Z}} \ell(z + \delta)) \, \mathrm{d}\mathcal{D}'(z) = \int \log(\alpha - \ell^{\mathcal{N}}(z)) \, \mathrm{d}\mathcal{D}(z) \, : \, \mathrm{LP}_{\mathcal{N}}(\mathcal{D}', \mathcal{D}) \leq 0\}$. $\quad\square$

We are finally ready to state and prove the main result.

Theorem C.4. *Consider a margin $\delta > 0$. Suppose at most a fraction α' with $\alpha' + \delta \leq \alpha$ of the IID training data is poisoned and the evasion attack on the test set is limited to a set $0 \in \mathcal{N}'$ with $\mathcal{N}' + B(0, \delta) \subseteq \mathcal{N}$. Denote the holistic robust solution as $\theta_{\mathcal{D}_n}^{HR} \in \arg\min_{\theta \in \Theta} \max_{\mathcal{D}'_{\text{test}} \in \mathcal{U}_{\mathcal{N}, \alpha, r}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta, Z)]$. Then, we have*

$$
Pr\Big(\mathcal{L}_{HR}^{\mathcal{N}, \alpha, r}(\theta) \geq \mathcal{L}_{test}(\theta), \, \forall \theta \in \Theta\Big) \geq 1 - (4/\delta)^{m(\mathcal{Z}, \delta)} \exp(-rn)
$$

for all $n \geq 1$, where $m(\mathcal{Z}, \delta) = \min\{k \geq 0 \, : \, \exists z_1, \dots, z_k \in \mathcal{Z} \text{ s.t. } \cup_{i=1}^{k} B(z_i, \delta) \supseteq \mathcal{Z}\}$ denotes the internal covering number of the support set $\mathcal{Z}$.

Proof. Denote again with $\mathcal{D}_n''$ the empirical distribution of the training data before poisoning corruption and likewise denote with $\mathcal{D}$ the data generating distribution of our IID samples and $\mathcal{D}_{\text{test}}$ the corrupted test, following Theorem 2.1 in Bennouna & Van Parys (2022) we have

$$\mathcal{D}_n'' \in \{\mathcal{D}'' \,:\, \text{LP}_{\{0\}}(\mathcal{D}_n, \mathcal{D}'') \leq \alpha'\}, \ \mathcal{D}_{\text{test}} \in \{\mathcal{D}_{\text{test}}' \,:\, \text{LP}_{\mathcal{N}'}(\mathcal{D}, \mathcal{D}_{\text{test}}') \leq 0\}. \tag{13}$$

Furthermore, according to Lemma C.2 we have that

$$\Pr(\mathcal{D} \in \{\mathcal{D}''' \,:\, \exists \mathcal{D}' \text{ s.t. } \text{LP}_{B(0,\delta)}(\mathcal{D}_n'', \mathcal{D}') \leq \delta, \ \text{KL}(\mathcal{D}', \mathcal{D}''') \leq r\}) \geq 1 - (4/\delta)^{m(\mathcal{Z}, \delta)} \exp(-rn). \tag{14}$$

Chaining together the previous inclusions will give us the desired result. Let us indeed denote here $\mathcal{U}_{\mathcal{N}', \alpha', r}'(\mathcal{D}_n) := \{\mathcal{D}_{\text{test}}' \,:\, \exists \mathcal{D}', \mathcal{D}'', \mathcal{D}''' \text{ s.t. } \text{LP}_{\{0\}}(\mathcal{D}_n, \mathcal{D}'') \leq \alpha', \ \text{LP}_{B(0,\delta)}(\mathcal{D}'', \mathcal{D}') \leq \delta, \ \text{KL}(\mathcal{D}', \mathcal{D}''') \leq r, \ \text{LP}_{\mathcal{N}'}(\mathcal{D}''', \mathcal{D}_{\text{test}}') \leq 0\}$ Chaining together inclusions (13) and (14) we get that $\Pr(\mathcal{D}_{\text{test}} \in \mathcal{U}_{\mathcal{N}', \alpha', r}'(\mathcal{D}_n)) \geq 1 - \exp(-rn + m(\mathcal{Z}, \delta) \log(4/\delta))$. From Lemma C.1 we have furthermore that the previous set can be characterized as $\mathcal{U}_{\mathcal{N}', \alpha', r}'(\mathcal{D}_n) = \{\mathcal{D}_{\text{test}}' \,:\, \exists \mathcal{D}', \mathcal{D}''' \text{ s.t. } \text{LP}_{B(0,\delta)}(\mathcal{D}_n, \mathcal{D}') \leq \alpha' + \delta, \ \text{KL}(\mathcal{D}', \mathcal{D}''') \leq r, \ \text{LP}_{\mathcal{N}'}(\mathcal{D}''', \mathcal{D}_{\text{test}}') \leq 0\}$

Consider now the associated robust loss $\mathcal{L}_{\text{HR}'}^{\mathcal{N}', \alpha', r}(\theta) := \max\{\mathbb{E}_{\mathcal{D}_{\text{test}}'}[\ell(\theta, Z)] : \mathcal{D}_{\text{test}}' \in \mathcal{U}_{\mathcal{N}', \alpha', r}'(\mathcal{D}_n)\}$. Recall that $\mathcal{L}_{\text{test}}(\theta) = \mathbb{E}_{\mathcal{D}_{\text{test}}}[\ell(\theta, Z)]$ and hence it follows immediately that

$$\Pr(\mathcal{L}_{\text{HR}'}^{\mathcal{N}', \alpha', r}(\theta) \geq \mathcal{L}_{\text{test}}(\theta), \ \forall \theta \in \Theta) \geq 1 - \exp(-rn + m(\mathcal{Z}, \delta) \log(4/\delta)).$$

It remains to show that $\mathcal{L}_{\text{HR}'}^{\mathcal{N}', \alpha', r}(\theta) \leq \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta)$ for all $\theta \in \Theta$. We have

$$\mathcal{L}_{\text{HR}'}^{\mathcal{N}', \alpha', r}(\theta) := \sup_{\text{LP}_{B(0,\delta)}(\mathcal{D}_n, \mathcal{D}') \leq \alpha' + \delta} \sup\{\mathbb{E}_{\mathcal{D}_{\text{test}}'}[\ell(\theta, Z)] : \exists \mathcal{D}''', \mathcal{D}_{\text{test}}' \text{ s.t. } \text{KL}(\mathcal{D}', \mathcal{D}''') \leq r, \ \text{LP}_{\mathcal{N}'}(\mathcal{D}''', \mathcal{D}_{\text{test}}') \leq 0\}$$

$$= \sup_{\text{LP}_{B(0,\delta)}(\mathcal{D}_n, \mathcal{D}') \leq \alpha' + \delta} \sup\{\mathbb{E}_{\mathcal{D}_{\text{test}}'}[\ell(\theta, Z)] : \exists \mathcal{D}''', \mathcal{D}_{\text{test}}' \text{ s.t. } \text{LP}_{\mathcal{N}'}(\mathcal{D}', \mathcal{D}''') \leq 0, \ \text{KL}(\mathcal{D}''', \mathcal{D}_{\text{test}}') \leq r\}$$

$$= \sup\{\mathbb{E}_{\mathcal{D}_{\text{test}}'}[\ell(\theta, Z)] : \exists \mathcal{D}''', \mathcal{D}_{\text{test}}' \text{ s.t. } \text{LP}_{\mathcal{N}' + B(0,\delta)}(\mathcal{D}_n, \mathcal{D}''') \leq \alpha' + \delta, \ \text{KL}(\mathcal{D}''', \mathcal{D}_{\text{test}}') \leq r\} \leq \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta).$$

The first equality follows from Lemma C.3. The second equality follows again from Lemma C.1. The final inequality follows from the assumption that $\alpha' + \delta \leq \alpha$ and $\mathcal{N}' + B(0, \delta) \subseteq \mathcal{N}$. $\square$

Denote with $\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta, \mathcal{D}_n)$ the HR cost in which we make its dependence on the distribution of the training data explicit. Theorem C.4 indicates that this cost is with high probability $1 - \exp(-nr + O(1))$ an upper bound on the adversarial test loss $\mathcal{L}_{\text{test}}(\theta)$ uniformly over $\theta \in \Theta$. The next result indicates that our HR cost is (in some sense) tight under relatively mild assumption.

Theorem C.5. *Let $0 \in int(\mathcal{N})$, $cl(int(\mathcal{N})) = \mathcal{N}$, $\alpha > 0$, $r > 0$ and $z \mapsto \ell(\theta, z)$ continuous and bounded for all $\theta \in \Theta$. Consider a function $\mathcal{L}_A^{\mathcal{N}, \alpha, r}$. Suppose that for some $\theta_0 \in \Theta$ and $\mathcal{D}_0' \in \cup_{n \geq 1}\{\sum_{i=1}^{n} \delta_{z_i}/n \,:\, z_i \in \mathcal{Z} \ \forall i \in [1, \ldots, n]\}$ we have*

$$\mathcal{L}_{HR}^{\mathcal{N}, \alpha, r}(\theta_0, \mathcal{D}_0') > \mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0, \mathcal{D}_0').$$

Then, there exists a data generating distribution $\mathcal{D}$ and an adversary which poisons less than a fraction α of all training data points and perturbs the testing data with evasion bounded in a compact set in the interior of $\mathcal{N}$ for which we have

$$\limsup_{n \to \infty} \frac{1}{n} \log Pr\left(\mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta) < \mathcal{L}_{test}(\theta), \ \exists \theta \in \Theta\right) \geq \limsup_{n \to \infty} \frac{1}{n} \log Pr\left(\mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0) < \mathcal{L}_{test}(\theta_0)\right) > -r.$$

Proof. Consider here $n_0 \geq 1$ so that we have $\mathcal{D}_0' \in \{\sum_{i=1}^{n_0} \delta_{z_i}/n_0 \,:\, z_i \in \mathcal{Z}(n_0)\}$ for some $\mathcal{Z}(n_0) := \{z_i \in \mathcal{Z}\}_{i=1}^{n_0}$. In other words, the distribution $\mathcal{D}_0'$ is the empirical distribution of n_0 data points in $\mathcal{Z}$. Let here $\epsilon := \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_0, \mathcal{D}_0') - \mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0, \mathcal{D}_0') > 0$. Without loss of generality we assume that we have ordered the points in $\mathcal{Z}(n_0)$ so that $\ell^{\mathcal{N}}(\theta_0, z_1) \leq \cdots \leq \ell^{\mathcal{N}}(\theta_0, z_{n_0})$.

We have by definition

$$\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}_0') := \sup\{\mathbb{E}_{\mathcal{D}_{\text{test}}'}[\ell(\theta_0, Z)] : \exists \mathcal{D}', \mathcal{D}_{\text{test}}' \text{ s.t. } \text{LP}_{\mathcal{N}}(\mathcal{D}_0', \mathcal{D}') \leq \alpha, \ \text{KL}(\mathcal{D}', \mathcal{D}_{\text{test}}') \leq r\}.$$

We first remark that as both the LP and KL metrics are jointly convex in the optimization variables $\mathcal{D}'$ and $\mathcal{D}'_{\text{test}}$ and the objective $\mathbb{E}_{\mathcal{D}'_{\text{test}}}[\ell(\theta_0, Z)]$ a linear function of the variable $\mathcal{D}'_{\text{test}}$ the HR loss $\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}'_0)$ is a concave function of the robustness parameters α and r. As concave functions are continuous on the interior of their domain the HR loss is a continuous function at $r > 0$ and $\alpha > 0$. Hence, we can consider $0 < r' < r$ and α' rational, $0 < \alpha' < \alpha$ so that $\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha', r'}(\theta_0; \mathcal{D}'_0) + \epsilon/4 \geq \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}'_0)$. Consider an integer $k' \geq 1$ such that $n'_0 \alpha' = n_0 k' \alpha'$ is integers and denote $n'_0 := n_0 k' \geq 1$. This choice of k' implies that $\mathcal{D}'_0$ can be seen as the empirical distribution of n'_0 data points in $\mathcal{Z}$ and our adversary will be able to precisely corrupt $n'_0 \alpha'$ data points.

Let us consider now a corrupted version of the support set $\mathcal{Z}'(n_0)$ defined here as

$$\mathcal{Z}'(n_0) = \{z'_i := z_i + \arg\max_{\delta \in \mathcal{N}, z_i + \delta \in \mathcal{Z}} \ell(\theta_0, z_i + \delta)\}_{i=1}^{n_0} \cup \{z'_\infty \in \arg\max_{z \in \mathcal{Z}} \ell(\theta_0, z)\}$$

which is well defined as $z \mapsto \ell(\theta_0, z)$ is continuous and $\mathcal{N}$ compact. Each of the points $z'_i \in \mathcal{Z}'(n_0)$ represents an evasion corrupted version of its counterpart $z_i \in \mathcal{Z}(n_0)$ while the point z'_∞ can be interpreted as a poisoned data point maximizing the loss. As the empirical distribution $\mathcal{D}'_0$ is finitely supported, from Lemma 3.7 and Corollary 2.3 in Bennouna & Van Parys (2022) it follows that for

$$\mathcal{D}'^\star = \sum_{i=\tau+1}^{n_0} \delta_{z'_i} \mathcal{D}'_0(z_i) + (1 - \alpha' - \sum_{i=\tau+1}^{n_0} \mathcal{D}'_0(z_i)) \delta_{z'_\tau} + \alpha' \delta_{z'_\infty}, \tag{15}$$

with τ the smallest integer so that $1 - \alpha' - \sum_{i=\tau+1}^{n_0} \mathcal{D}'_0(z_i) > 0$, we can find a distribution $\mathcal{D}'^\star_{\text{test}} \in \{\mathcal{D}' : \text{supp}(\mathcal{D}') \subseteq \mathcal{Z}'(n_0)\}$ with

$$\text{LP}_{\mathcal{N}}(\mathcal{D}'_0, \mathcal{D}'^\star) \leq \alpha', \quad \text{KL}(\mathcal{D}'^\star, \mathcal{D}'^\star_{\text{test}}) \leq r', \quad \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha', r'}(\theta_0; \mathcal{D}'_0) = \mathbb{E}_{\mathcal{D}'^\star_{\text{test}}}[\ell(\theta_0, Z)].$$

In other words, the distributions $\mathcal{D}'^\star$ and $\mathcal{D}'^\star_{\text{test}}$ represent maximizers in the optimization problem characterizing the HR cost $\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha', r'}(\theta_0; \mathcal{D}'_0)$. As $n'_0 \alpha'$ is integer and $\mathcal{D}'_0 \in \{\sum_{i=1}^{n_0} \delta_{z_i}/n_0 : z_i \in \mathcal{Z}(n_0)\}$ it also follows immediately that $\mathcal{D}'^\star \in \{\mathcal{D}' : \text{supp}(\mathcal{D}') \subseteq \mathcal{Z}'(n_0), \mathcal{D}'(z) n'_0 \in [0, \ldots, n'_0] \ \forall z \in \mathcal{Z}'(n_0)\}$. This last observation implies that also the distribution $\mathcal{D}'^\star$, just as $\mathcal{D}'_0$, can be seen as the empirical distribution of n'_0 points in $\mathcal{Z}$.

Introduce a distribution $\bar{\mathcal{D}}'^\star$ supported on $\mathcal{Z}(n_0) \cup \{z'_\infty\}$ by defining $\bar{\mathcal{D}}'^\star(z_i) = \mathcal{D}'^\star(z'_i)$ for all $i \in [1, \ldots, n_0]$ and $\bar{\mathcal{D}}'^\star(z'_\infty) = \mathcal{D}'^\star(z'_\infty)$. In essence, we move each point in the support of $\mathcal{D}'^\star$ from a perturbed position z'_i back to z_i with the exception of z'_∞ which we leave as is. From Equation (15) we have $\alpha' \geq \text{LP}_{\mathcal{N}}(\mathcal{D}'_0, \mathcal{D}'^\star) = \mathcal{D}'^\star(z'_\infty) = \bar{\mathcal{D}}'^\star(z'_\infty) = \text{LP}_{\{0\}}(\mathcal{D}'_0, \bar{\mathcal{D}}'^\star)$. Likewise, define a distribution $\bar{\mathcal{D}}''^\star$ supported on $\mathcal{Z}(n_0) \cup \{z'_\infty\}$ though $\bar{\mathcal{D}}''^\star(z_i) = \mathcal{D}'^\star_{\text{test}}(z'_i)$ for all $i \in [1, \ldots, n_0]$ and $\bar{\mathcal{D}}''^\star(z'_\infty) = \mathcal{D}'^\star_{\text{test}}(z'_\infty)$. In essence, we move each point in the support of $\mathcal{D}'^\star_{\text{test}}$ from a perturbed position z'_i back to z_i with the exception of z'_∞ which we leave as is. It is easy to verify that by construction $\text{KL}(\bar{\mathcal{D}}'^\star, \bar{\mathcal{D}}''^\star) = \sum_{i=1}^{n_0} \log(\bar{\mathcal{D}}'^\star(z_i)/\bar{\mathcal{D}}''^\star(z_i)) \bar{\mathcal{D}}'^\star(z_i) + \log(\bar{\mathcal{D}}'^\star(z'_\infty)/\bar{\mathcal{D}}''^\star(z'_\infty)) \bar{\mathcal{D}}'^\star(z'_\infty) = \sum_{i=1}^{n_0} \log(\mathcal{D}'^\star(z'_i)/\mathcal{D}'^\star_{\text{test}}(z'_i)) \mathcal{D}'^\star(z'_i) + \log(\mathcal{D}'^\star(z'_\infty)/\mathcal{D}'^\star_{\text{test}}(z'_\infty)) \mathcal{D}'^\star(z'_\infty) = \text{KL}(\mathcal{D}'^\star, \mathcal{D}'^\star_{\text{test}}) \leq r'$ and $\text{LP}_{\mathcal{N}}(\bar{\mathcal{D}}''^\star, \mathcal{D}'^\star_{\text{test}}) = 0$. In fact, as $z \mapsto \ell(\theta_0, z)$ is continuous and $\text{cl}(\text{int}(\mathcal{N})) = \mathcal{N}$ we can find points z''_i with $z''_i - z_i \in \text{int}(\mathcal{N})$ so that for $\mathcal{D}''^\star_{\text{test}}$ defined as $\mathcal{D}''^\star_{\text{test}}(z''_i) = \mathcal{D}'^\star_{\text{test}}(z'_i)$, $i \in [1, \ldots, n_0]$ satisfies $\text{LP}_{\mathcal{N}'}(\bar{\mathcal{D}}''^\star, \mathcal{D}'^\star_{\text{test}}) = 0$ and

$$\mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha', r'}(\theta_0; \mathcal{D}'_0) = \mathbb{E}_{\mathcal{D}'^\star_{\text{test}}}[\ell(\theta_0, Z)] \leq \mathbb{E}_{\mathcal{D}''^\star_{\text{test}}}[\ell(\theta_0, Z)] + \epsilon/4$$

with $\mathcal{N}' := \{0\} \cup \{\delta'_i := z''_i - z_i\}_{i=1}^{n_0} \subseteq \text{int}(\mathcal{N})$. In essence, the points z''_i are chosen to present a perturbation of the points z_i of an evasion adversary limited to a evasion set within a compact subset in the interior of $\mathcal{N}$.

Consider the data generation process with distribution $\bar{\mathcal{D}}''^\star$, where $n \geq 1$ uncorrupted data points, of empirical distribution $\mathcal{D}''_n$, are sampled from $\bar{\mathcal{D}}''^\star$ supported on $\mathcal{Z}(n_0) \cup \{z'_\infty\}$. First, for this process, if an evasion adversary implements test data evasion with $\mathcal{N}'$ then by construction of $\mathcal{D}''^\star_{\text{test}}$ we get $\mathcal{L}_{\text{test}}(\theta_0) \geq \mathbb{E}_{\mathcal{D}''^\star_{\text{test}}}[\ell(\theta_0, Z)]$ simply by adding perturbation δ'_i when encountering realization z_i, $i \in [1, \ldots, n_0]$ and 0 in case z'_∞ is observed. Remark that from $\mathcal{D}'^\star \in \{\mathcal{D}' : \text{supp}(\mathcal{D}') \subseteq \mathcal{Z}'(n_0), \mathcal{D}'(z) n'_0 \in [0, \ldots, n'_0] \ \forall z \in \mathcal{Z}'(n_0)\}$ and the construction of $\bar{\mathcal{D}}'^\star$ it follows that $\bar{\mathcal{D}}'^\star \in \{\mathcal{D}' : \text{supp}(\mathcal{D}') \subseteq \mathcal{Z}(n_0) \cup \{z'_\infty\}, \mathcal{D}'(z) n'_0 \in [0, \ldots, n'_0] \ \forall z \in \mathcal{Z}(n_0) \cup \{z'_\infty\}\}$. In other words, the distribution $\bar{\mathcal{D}}'^\star$ can be seen as the empirical distribution of n'_0 data points in $\mathcal{Z}(n_0) \cup \{z'_\infty\}$. Large deviation theory indicates that the event $\mathcal{D}''_{kn'_0} = \bar{\mathcal{D}}'^\star$ must occur with a nonzero probability. In particular, following Theorem 11.1.4 in Cover & Thomas (1991), we have the lower bound

$$\Pr(\mathcal{D}''_{kn'_0} = \bar{\mathcal{D}}'^\star) \geq \frac{\exp(-n'_0 k \text{KL}(\bar{\mathcal{D}}'^\star, \bar{\mathcal{D}}''^\star))}{(kn'_0 + 1)^{n_0 + 1}} \geq \frac{\exp(-n'_0 k r')}{(kn'_0 + 1)^{n_0 + 1}} \quad \forall k \geq 1.$$

However, if the event $\mathcal{D}''_{kn'_0} = \bar{\mathcal{D}}'^\star$ occurs then from Equation (15) it is clear that a poisoning adversary can perturb precisely $\alpha' kn'_0$ training data points to transform the training distribution $\mathcal{D}''_{kn'_0}$ into a corrupted distribution, denoted here with $\mathcal{D}_{kn'_0}$

that satisfies $\mathcal{D}_{kn_0'} = \mathcal{D}_0'$. Hence, it follows that with such a poisoning adversary which corrupts fewer than a fraction α of all data points we have $\Pr(\mathcal{D}_{kn_0'}'' = \bar{\mathcal{D}}'^\star) \leq \Pr(\mathcal{D}_{kn_0'} = \mathcal{D}_0')$. If however the event $\mathcal{D}_{kn_0'} = \mathcal{D}_0'$ occurs, we have by chaining the inequalities derived in the previous paragraph that

$$\mathcal{L}_{\text{test}}(\theta_0) \geq \mathbb{E}_{\mathcal{D}_{\text{test}}''^\star}[\ell(\theta_0, Z)] \geq \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha', r'}(\theta_0; \mathcal{D}_0') - \epsilon/4 \geq \mathcal{L}_{\text{HR}}^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}_0') - \epsilon/2 = \mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}_0') + \epsilon/2 > \mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}_0').$$

Consequently,

$$\Pr(\mathcal{L}_{\text{test}}(\theta_0) > \mathcal{L}_A^{\mathcal{N}, \alpha, r}(\theta_0; \mathcal{D}_0')) \geq \Pr(\mathcal{D}_{kn_0'}'' = \bar{\mathcal{D}}'^\star) \geq \frac{\exp(-n_0' k r')}{(kn_0' + 1)^{n_0 + 1}} \quad \forall k \geq 1$$

from which the result follows by passing to the limit $k \to \infty$ and recalling that $r' < r$. $\qquad\square$

D. Further details on experiments

D.1. Experimental Setup

MNIST As stated in the main text, we use a CNN architecture specifically designed for MNIST and provided in the main Pytorch library 🔗. This network is similar to the tensorflow variant originally used by (Madry et al., 2018), consisting of two convolutional layers with 32 and 64 filters respectively, followed by 2×2 max-pooling, and a fully connected layer of size 9216. We use the ADAM optimizer with a learning rate of 1×10^{-3} and without weight decay or dropout throughout the experiments.

CIFAR-10 For the majority of experiments in Section 6, we use the ResNet18 architecture trained with the help of the ADAM optimizer with a starting learning rate of 1×10^{-2} and without weight decay. We use ADAM, rather than stochastic gradient descent (SGD), to avoid need for manual tuning of the learning rate in our varied set of models and experiments. In Section 6.2, exceptionally, we use the same configurations as (Rice et al., 2020) for training. That is, we use here SGD with a starting learning rate of 0.1 and a fixed learning rate decay schedule, as well as implementing their chosen weight decay of 5×10^{-4}. We use the same learning rate decay schedule provided by (Rice et al., 2020). The attacks are carried out by a 10-step PGD adversary. For the robust overfitting experiments in Section 6.2, we use an ℓ_∞ adversary of size $\epsilon = 8/255$ with a step size of $2/255$ mimicking the radius and step size considered by (Rice et al., 2020). For the holistic robustness experiments in Section 6.3, we use an ℓ_2 adversary of size $\epsilon = 25.5/255 = 0.1$. This is smaller than, for instance, the attack size used by (Rice et al., 2020), but we remark that such an adversary causes catastrophic overfitting in our setting, since we additionally face data poisoning and statistical error. We use an adversarial step size of 0.2, which is the default for the ℓ_2 attack implemented in the torchattacks 🔗 library.

Except where stated otherwise, we train throughout all experiments for 300 epochs, which is longer than (Madry et al., 2018) and (Rice et al., 2020). We do so because in many cases we are working with a reduced number of batches (due to subsampling only a fraction of the available data), and hence the number of training steps scales down proportionally with the size of the subsampled dataset. We train for more epochs to ensure in particular convergence given fewer training batches per epoch.

D.2. Benchmarking

We benchmark our proposed HR training method against other procedures. We describe here briefly how each of these procedures are ran and in particular how each of their hyperparameters are tuned.

- **ERM**: there are no hyperparameters to tune.

- **PGD**: we validate over the parameter range $\epsilon \in \{0, 0.05, 0.1, 0.2\}$ associated with protection against ℓ_2 evasion attacks. As stated before, we use an ℓ_2 adversary with $\epsilon = 0.1$ to perturb the test data and hence the considered range contains the actual attack radius. When selecting the appropriate value for ϵ for HR training we consider the same range.

- **KL DRO**: KL DRO counts just one hyperparameter r which controls its robustness to statistical error. We consider here $r \in \{0, 0.05, 0.1, 0.2\}$. When selecting the appropriate value for r for HR training we consider again the same range.

- **TV DRO**: For TV DRO, there is also just one hyperparameter α controlling the robustness to data poisoning. We consider $\alpha \in \{0, 0.05, 0.1, 0.2\}$. When selecting the appropriate value for α for HR training we consider again the same range.

- **TRADES**: We use the ℓ_2-adversary version of TRADES. We grid search over the two TRADES hyperparameters β and ϵ. We use $\beta \in \{1, 6\}$, the two values used in their comparison of defense methods, and $\epsilon \in \{0, 0.05, 0.1, 0.2\}$, the same range implemented for other methods.

- **DPA**: A single parameter k controls the number of partitions over which each of the base classifiers are fit. We try $k \in \{5, 10, 50\}$. Since we are testing DPA on sub-sampled data, we avoid using larger k as each partition is already comparatively smaller than when using the full dataset. Computing the training and testing loss for DPA (e.g. Figure 4) differs slightly compared to other methods, due to the nature of the DPA partitioning procedure. Training loss is taken as the average cross-entropy for each of the base classifiers. Testing loss is computed as follows: we first aggregate to find the average value of the output layer across the base classifiers. We then compute the cross-entropy loss of these outputs on the test set.

- **DPA + PGD**: DPA splits the data into partitions, trains a separate model on each partition, and then ensembles the models to give the final prediction. We combine DPA and PGD naturally by training the models on each partition with PGD. Hence, the overall training of DPA+PGD splits the data into partitions, train a robust model with PGD (as opposed to natural training) on each partition, and then ensemble the models.

  We cross-validate over the individual parameters used for DPA and PGD individually. That is $k \in \{5, 10, 50\}$ and $\epsilon \in \{0, 0.05, 0.1, 0.2\}$, giving 12 models in total.

- **HR**: We search over the grid $(\alpha, r, \epsilon) \in \{0, 0.05, 0.1, 0.2\}^3$, giving 4 options per robustness parameter and 64 models in total.

D.3. Experimental Results

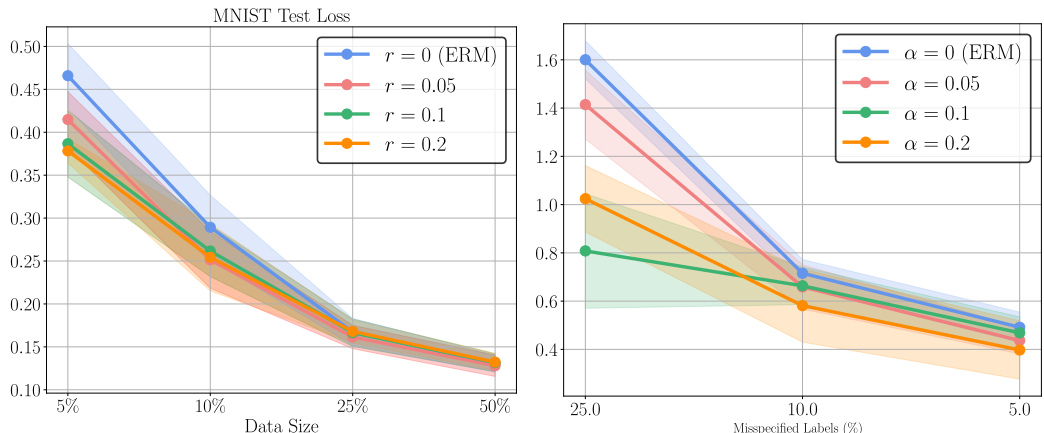#### D.3.1. Single parameter experiments

Figure 5. Robustness to statistical error (left) and data poisoning (right) for MNIST. The same results are shown for CIFAR-10 in Figure 2 of the main text.
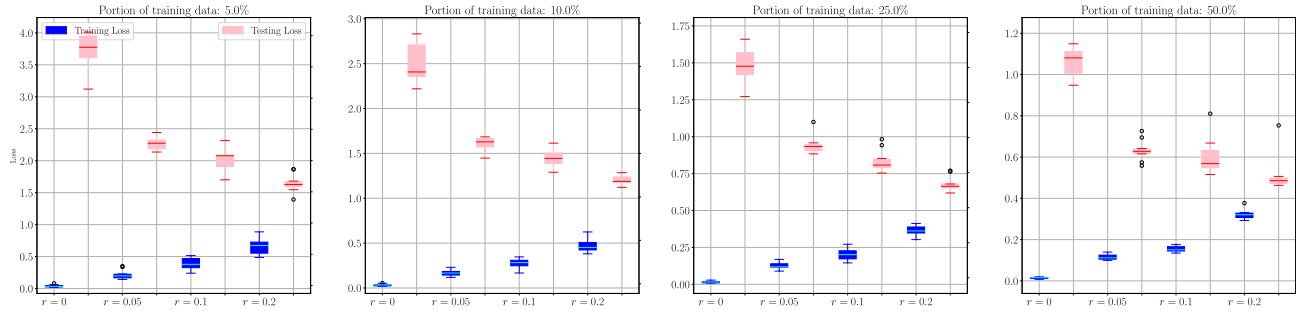
Figure 6. Experimenting HR with various values of r when learning under statistical error on CIFAR-10. Here, $\alpha = 0$ and $\mathcal{N} = \{0\}$. See section 6.1 for setup.
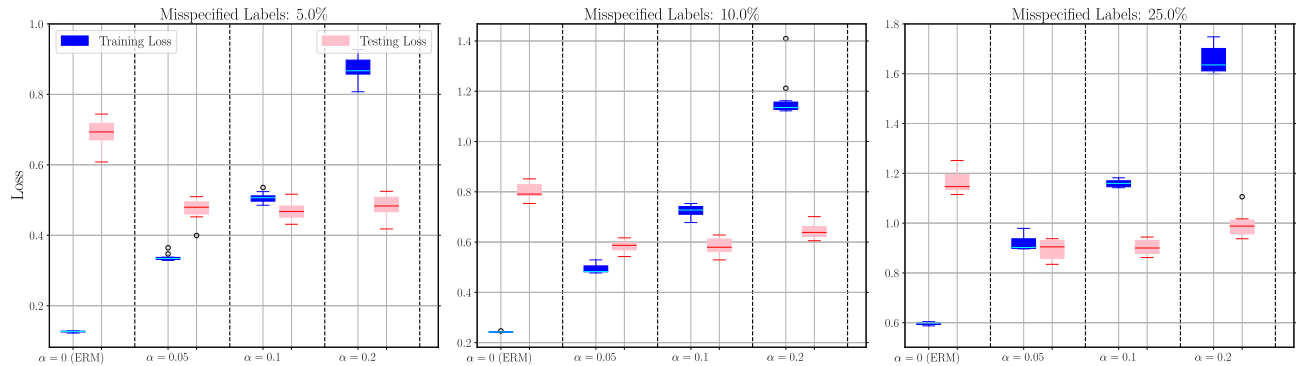

Figure 7. Experimenting HR with various values of α when learning under poisoning on CIFAR-10. Here, $r = 0$ and $\mathcal{N} = \{0\}$. See section 6.1 for setup.