Proceedings of Machine Learning Research vol 167:1–35, 2022 33rd International Conference on Algorithmic Learning Theory

Understanding Simultaneous Train and Test Robustness

Pranjal Awasthi PRANJALAWASTHI@GOOGLE.COM
Google Research

Sivaraman Balakrishnan SIVA@STAT.CMU.EDU
Carnegie Mellon University

Aravindan Vijayaraghavan ARAVINDV@NORTHWESTERN.EDU
Northwestern University

Editors: Sanjoy Dasgupta and Nika Haghtalab

Abstract

This work concerns the study of robust learning algorithms. In practical settings, it is desirable to achieve robustness to many different types of corruptions and shifts in the data distribution such as defending against adversarial examples, dealing with covariate shifts, and contamination of training data (data poisoning). While there has been extensive recent work on these topics, models and algorithms for these different notions of robustness have been largely developed in isolation.

In this paper, we propose a natural notion of robustness that allows us to simultaneously reason about train-time and test-time corruptions, that can be measured using various distance metrics (e.g., total variation distance, Wasserstein distance). We study our proposed notion in three fundamental settings in supervised and unsupervised learning (of regression, classification and mean estimation). In each case we design sample and time-efficient learning algorithms with strong simultaneous train-and-test robustness guarantees. In particular, our work shows that the two seemingly different notions of robustness at train-time and test-time are closely related, and this connection can be leveraged to develop algorithmic techniques that are applicable in both the settings.

1. Introduction

Model robustness is an important consideration in the design of modern machine learning systems from a reliability and security perspective. Robustness refers broadly to the ability of a learning algorithm to deal with various uncertainties introduced either during the training of the model, or during model deployment. For instance, when collecting data for model training one often encounters noisy labels, missing attributes, measurement errors, model misspecification and even malicious data injected by an adversary (Nettleton et al., 2010). We will use the term *train corruptions* to refer to such perturbations. Similarly, when a model is deployed into the real world it often encounters a distribution shift problem, i.e., it is tested on a data distribution that is different from the distribution of the training data (Hendrycks and Gimpel, 2016). Furthermore, recent works have demonstrated that modern systems such as deep neural networks are highly susceptible to small imperceptible perturbations made to the inputs at test time. These are called adversarial examples (Szegedy et al., 2013; Biggio et al., 2013; Goodfellow et al., 2014). We refer to the above scenarios occurring at test time as *test corruptions*.

Achieving robustness to corruptions at both train-time and test-time is important for the development of reliable machine learning systems. In this work we study this question from a theoretical perspective. There is a large body of existing work on designing principled algorithms that are robust to corruptions at train-time. Some examples include statistical and computational methods for handling train-time corruptions measured in total variation distance (Huber, 2011; Lai et al., 2016; Diakonikolas et al., 2018b), methods for missing data (Rubin, 1976), and algorithms to deal with noisy labels (Kearns et al., 1994; Kalai et al., 2008). Similarly, there is a body of work on designing algorithms that are robust to distribution shifts at test time (Ben-Tal et al., 2009; Wiesemann et al., 2014), and more recently, to adversarial examples (Awasthi et al., 2019; Diakonikolas et al., 2020a; Montasser et al., 2021).

The works mentioned above have largely focused on developing techniques for either train-time robustness or test-time robustness in isolation. This leads to the following natural question:

Question: *Is there a unified framework for developing algorithms that are simultaneously robust to both train-time and test-time corruptions?*

It is reasonable to expect that these goals are aligned – a technique that achieves robustness for one setting, say robustness to test-time corruptions, incentivizes robustness to train-time corruptions as well. As an example consider an algorithm that fits a linear classifier to the data by maximizing the margin (Cortes and Vapnik, 1995). If there indeed exists a large margin classifier, then at test-time such a classifier will be robust to small perturbations to the inputs (measured in Wasserstein W_∞ metric). However, by the same large margin property, any algorithm that finds a large margin classifier will also be able to handle small perturbations to the input at train-time (again measured in Wasserstein W_∞ metric). Studying train-time and test-time robustness under a common framework will allow us to understand how they are related to each other.

Our Contributions. We propose a new notion of robustness that captures corruptions at both train-time and test-time. In our definition we assume the existence of a ground truth distribution $P^* \in \mathcal{P}$ unknown to the learning algorithm, where $\mathcal{P}$ is a (structured) family of distributions. For instance, in classification settings, P^* would correspond to a joint distribution over (example, label) pairs. Given a hypothesis/predictor h, a loss function $L(\cdot)$ of interest, and a distance metric d over distributions, the *robust-loss* of h is the worst loss incurred by h when tested on any δ-perturbation of P^*:

$$\textit{robust-loss}(L, P^*, h, \delta) := \sup_{P \in \mathsf{Ball}_d(P^*, \delta)} L_P(h).$$

Furthermore the learning algorithm is given samples at train-time from a corrupted distribution $\widetilde{P}$ that is also a δ-perturbation of P^* (in general, the train and test perturbations could be of different magnitudes). The goal of the learning algorithm is to use samples from the corrupted training distribution $\widetilde{P}$ and output a hypothesis $\widehat{h}$ that is competitive with the *robust-loss* of the best hypothesis among a class of functions i.e.,

$$\textit{robust-loss}(L, P^*, \widehat{h}, \delta) \leq (1+\alpha) \min_{h \in H} \textit{robust-loss}(L, P^*, h, \delta), \quad \text{for some constant } \alpha > 0.$$

See Def. 2 for a more formal definition. We will call a hypothesis that satisfies the above definition to be *train-and-test robust*. We remark that the distance metrics for the test-time perturbation and the train-time perturbation can be different. Our algorithmic results will focus on two common choices for measuring corruptions, the total variation distance d_{TV}, and the Wasserstein distance W_∞.

We first consider the population setting, i.e., ignoring statistical and computational concerns. *Why should there exist hypothesis that are simultaneous train-and-test robust?* We present a simple yet general theorem (see Theorem 3) that proves that the existence of a test-time robust hypothesis yields a hypothesis that is simultaneously train-time and test-time robust. This formalizes the intuition of large margin classification discussed above, and demonstrates that simultaneous train-and-test robustness is achievable in many settings. However, the algorithm that achieves this is neither statistically nor computationally efficient in most settings. The main contribution of our work is to show how we can design efficient algorithms that are train-and-test robust in fundamental settings like linear classification, linear regression and mean estimation, as described below.

Computationally efficient algorithms for linear classification. We next design efficient train-and-test robust algorithms for the basic setting of learning linear classifiers in Section 3. We study the non-realizable setting where the ground truth distribution P^* may not admit a linear classifier of zero error. This is a challenging setting even without test-time robustness considerations, and certain distributional assumptions are needed for computational tractability. We make the standard assumption that the marginal distribution over the examples is the Gaussian distribution in n dimensions (Klivans et al., 2009; Awasthi et al., 2014; Daniely, 2015; Diakonikolas et al., 2021). We first consider train-time and test-time corruptions in the Wasserstein W_∞ distance (over some ℓ_q norm space). We establish a structural claim that any classifier that achieves low robust loss must be sparse (measured in the dual metric ℓ_{q^*}). We leverage this sparsity property along with an appropriately modified iterative hinge loss algorithm that was designed for noisy labels (Awasthi et al., 2014), to achieve simultaneous train-time and test-time robustness. Furthermore, the resulting algorithm is computationally and statistically efficient. We also consider a more challenging model where the training distribution is not only perturbed in the Wasserstein W_∞ distance, but additionally in total variation distance d_{TV} as well. In this case we design a novel outlier removal procedure that again exploits the sparsity property to simultaneously achieve both train-time and test-time robustness guarantees. Complementing our earlier general connection (Theorem 3), these results demonstrate that robust techniques designed to handle corruptions of one kind, can yield robustness benefits at both train-time and test-time.

Computationally efficient algorithms for linear regression and mean estimation. In Sections 4 and 5 we consider linear regression and mean estimation in our framework under Wasserstein W_∞ corruptions. In this case we show that minimizing a natural convex objective that carefully combines an ℓ_2 loss term and an ℓ_1 loss term leads to simultaneous train-time and test-time robustness guarantees. We also provide statistical rates of convergence that require carefully analyzing the structure of the optimal solution of the objective. These results both exploit the connection from test-time to train-and-test robustness establishes in Theorem 3. In Section 6 we end by exploring an alternate definition of simultaneous train-and-test robustness, followed by some open directions.

Comparison to related work. There is a large body of work on designing robust algorithms under various corruption models. Here we discuss the works most relevant to our results. A popular model for studying training corruptions is Huber's ε-contamination model (Huber, 2011). Here one has access to a corrupted distribution $\widetilde{P}$ that is close (in total variation distance) to an unknown structured distribution $P^* \in \mathcal{P}$ on which one wants to perform a certain task such as mean estimation or regression. The study of Huber's model has led to many insightful results, both towards characterizing minimax optimal rates of estimation (Chen et al., 2016, 2018), and the design of computationally efficient algorithms that are robust to ε-contamination (Diakonikolas et al., 2019; Lai et al., 2016; Balakrishnan et al., 2017; Diakonikolas et al., 2018a; Prasad et al., 2018).

For classification problems there are works focusing on algorithms that are robust to noisy labels. The noise models considered in these works range from understanding *random classification noise*, where the label for each data point is flipped independently and identically (Kearns, 1998; Blum et al., 1998), to the more challenging agnostic learning model (Kearns et al., 1994; Kalai et al., 2008). Some of the techniques in these works can also be extended to handle malicious noise where both the examples and the labels can be corrupted (Klivans et al., 2009; Awasthi et al., 2014). In particular, our algorithms for linear classifiers in Section 3 build on these techniques, that were designed for train corruptions, to get simultaneous train-time and test-time robustness. Other models of training corruptions include learning with missing features (Bullins et al., 2016), entry-wise perturbations (Awasthi et al., 2020), and strategic classification where individuals when viewed as data points may perturb their attributes for a desired classification or outcome (Hardt et al., 2016).

There also exist many works studying robustness to changes in the test distribution. The classical framework of robust optimization aims to find a solution that minimizes the worst case performance over an *uncertainty set* (Ben-Tal et al., 2009). In the above framework, uncertainty in the test distribution can be modeled by considering a ball (in a distance measure such as KL-divergence or Wasserstein metric) around the input data distribution. Recent works on distributionally robust optimization study this approach towards test-time robustness (Wiesemann et al., 2014; Namkoong and Duchi, 2016; Sinha et al., 2017; Kuhn et al., 2019). In particular, the work of Sinha et al. (2017) studies the problem of certifying and learning test time robust classifiers under a similar notion of test time perturbations as studied in our work, i.e., Wasserstein distances. However the work of Sinha et al. (2017) and other mentioned above do not study or provide guarantees for achieving simultaneous train and test robustness. Additionally, these works often involve relaxing the actual robust loss to a convex proxy and designing first order methods to optimize the proxy. This in general does not provide theoretical guarantees on the true robust test loss. In our work we design polynomial time algorithms to provide guarantees on the true robust test loss (while additionally being robust to train time perturbations as well). Another approach to robustness is via *domain adaptation* where the goal is to design algorithms that when trained on some domain, can adapt to related domains (Ben-David et al., 2007; David et al., 2010; Blitzer et al., 2008; Mansour et al., 2008).

Finally, motivated by the phenomenon of adversarial examples in deep learning (Szegedy et al., 2013; Biggio et al., 2013), many recent works have explored algorithmic and statistical issues in defending against small imperceptible perturbations made to the inputs at test-time. The work of Tsipras et al. (2018) explore inherent trade-offs between robustness and natural accuracy, whereas the works of Bubeck et al. (2018b,a); Nakkiran (2019) provide computational hardness results for robustly learning certain tasks that are easy to learn in the standard classification settings. There have also been works exploring the sample complexity of robust classification (Cullina et al., 2018; Yin et al., 2018; Khim and Loh, 2018; Montasser et al., 2019). Some recent works also study the problem of designing efficient learning algorithms that are robust to adversarial attacks, for natural function classes such as linear classifiers and polynomial threshold functions (Awasthi et al., 2019; Montasser et al., 2020; Diakonikolas et al., 2020a; Montasser et al., 2021; Ghazi et al., 2021)

2. Preliminaries and Definitions

We will use $\mathcal{X}$ to denote the input domain (which is typically $\mathbb{R}^n$) and $\mathcal{Y}$ to denote the label space (typically $\{+1, -1\}$ or $\mathbb{R}$). For a joint distribution P over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, we will use $P_{\mathcal{X}}$, $P_{\mathcal{Y}}$ to denote the corresponding marginals over $\mathcal{X}$ and $\mathcal{Y}$ respectively; in the unsupervised setting $\mathcal{Z} = \mathcal{X}$. H will denote a hypothesis class of functions from $\mathcal{X} \to \mathcal{Y}$. Given a class H, and a loss function $L : \mathcal{M}(\mathcal{Z}) \times H \to \mathbb{R}$, the loss of a hypothesis $h \in H$ on distribution $P \in \mathcal{M}(\mathcal{Z})$ is denoted by

$L(P, h)$ or $L_P(h)$ and defined to be

$$L_P(h) := \mathop{\mathbb{E}}_{z \sim P}[\ell(h, z)] \text{ where } \ell : H \times \mathcal{Z} \to \mathbb{R}.$$

We will primarily focus on the case when $L_P(h)$ is the 0/1 classification loss or the ℓ_2 loss.

Distances between distributions. For any space $\mathcal{Z}$, $\mathcal{M}(\mathcal{Z})$ denotes the space of valid measures defined on $\mathcal{Z}$. For any metric (or semi-metric) $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$ on distributions $\mathcal{M}$, any distribution $P \in \mathcal{M}$ and $\delta \geq 0$, we will denote by $\mathsf{Ball}_d(P, \delta) := \{Q \in \mathcal{M} | d(P, Q) \leq \delta\}$. By default, the underlying space is $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{M}$ represents $\mathcal{M}(\mathbb{R}^n)$. For $P_1, P_2 \in \mathcal{M}$, $d_{\mathsf{TV}}(P_1, P_2) := \sup_{A \subset \mathcal{X}} |P_1[A] - P_2[A]|$ denotes the total variation distance between them. The corresponding δ-ball around P in d_{TV} is denoted by $\mathsf{Ball}_{\mathsf{TV}}(P, \delta)$.

To define Wasserstein distances, we consider joint distributions over a product space $\mu \in \mathcal{M}(\mathbb{R}^n \times \mathbb{R}^n)$, and use $\mu^{(1)}, \mu^{(2)} \in \mathcal{M}(\mathbb{R}^n)$ to denote the marginal distributions imposed by μ on $\mathbb{R}^n$. The W_∞ Wasserstein distance over $\mathbb{R}^n$ [1] equipped with the ℓ_q norm is defined as follows.

Definition 1 (Wasserstein distance with ℓ_q norm) *For any $q \geq 1$, we will denote by $d_{W_\infty(\ell_q)}$ the Wasserstein distance where the underlying space is an ℓ_q metric space. For two measures ν_1, ν_2 defined over the normed metric space $(\mathbb{R}^n, \ell_q)$, the Wasserstein distances are given by*

$$d_{W_\infty(\ell_q)}(\nu_1, \nu_2) = \inf_{\substack{\mu \in \mathcal{M}(\mathbb{R}^n \times \mathbb{R}^n) \\ \mu^{(1)} = \nu_1, \mu^{(2)} = \nu_2}} \sup_{\substack{x, \widetilde{x} \in \mathbb{R}^n \times \mathbb{R}^n \\ \mu(x, \widetilde{x}) > 0}} \|x - \widetilde{x}\|_q, \tag{1}$$

$\mathsf{Ball}_{W(\ell_q)}(P_0, \delta) := \{P \in \mathcal{M}(\mathbb{R}^n) \mid d_{W_\infty(\ell_q)}(P, P_0) \leq \delta\}$ is the Wasserstein δ-ball around P_0.

2.1. Definition of Robustness

Given an uncorrupted distribution P^*, hypothesis $h \in H$, loss L and a distance metric d_{test}, the *robust test loss* (or robust loss, in short) is defined as:

$$\textit{robust-loss}(L, P^*, h, \delta) = \sup_{P \in \mathcal{M}(Z) \text{ s.t. } d_{\mathsf{test}}(P, P^*) \leq \delta} L_P(h). \tag{2}$$

We now define the main notion of robustness we study in this paper. Our definition incorporates robustness both at training time and testing time, measured using (potentially different) distance metrics $d_{\mathsf{train}}, d_{\mathsf{test}}$ on space of distributions $\mathcal{M}(\mathcal{Z})$.

Definition 2 (($\delta_1, \delta_2, \alpha$)-TRAIN-AND-TEST ROBUSTNESS) *Suppose $\delta_1, \delta_2, \alpha > 0$, and $\mathcal{M} = \mathcal{M}(\mathcal{Z})$ be the space of distributions over the space $\mathcal{Z}$, and let H be a hypothesis class equipped with the loss function L. For a distribution $P^* \in \mathcal{M}$, an algorithm Alg is $(\delta_1, \delta_2, 1 + \alpha)$-TRAIN-AND-TEST ROBUST w.r.t P^*, H i.f.f. for any $\varepsilon > 0$, given $m(\varepsilon, \delta_1, \delta_2)$ i.i.d. samples S drawn from a distribution $\widetilde{P}$ such that $d_{\mathsf{train}}(\widetilde{P}, P^*) \leq \delta_1$, Alg outputs a function $\widehat{h} = \mathsf{Alg}(S)$ such that robust-loss$(L, P^*, \widehat{h}, \delta_2) \leq (1 + \alpha) \min_{h \in H}$ robust-loss$(L, P^*, h, \delta_2) + \varepsilon$, i.e.,*

$$\mathop{\mathbb{P}}_{S \sim \widetilde{P}} \Big[\sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(\widehat{h}) \leq (1 + \alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h) + \varepsilon \Big] \geq \frac{2}{3}, \quad \textit{where } \widehat{h} = \mathsf{Alg}(S). \tag{3}$$

[1] Technically we need to consider a compact subset of $\mathbb{R}^n$. However, all our algorithmic results consider the Gaussian marginal and hence we can always restrict to a ball of large enough radius at the expense of an arbitrarily small additive error in our final bounds. For ease of exposition we ignore this technicality.

Given a family $\mathcal{P}$ of distributions, we say that Alg *is train-and-test robust w.r.t.* $(\mathcal{P}, H, d)$ *if for any* $P^* \in \mathcal{P}$*,* Alg *is* TRAIN-AND-TEST *robust w.r.t.* P^*, H *(note that neither* $\widetilde{P}$ *nor the test distribution* P *needs to be restricted to the family* $\mathcal{P}$*).*

For binary classification, when the distribution shift is measured with a distance that involves a metric structure (e.g., Wasserstein $W_\infty(\ell_q)$), it is more natural to measure these distances over the input space $\mathcal{X} = \mathbb{R}^n$, as opposed to over the space $\mathcal{Z} = \mathcal{R}^n \times \{+1, 1\}$. In this case, we will overload notation (when clear from context) and use $d_{W_\infty(\ell_q)}(P, Q)$ to denote the distance $d_{W_\infty(\ell_q)}(P_\mathcal{X}, Q_\mathcal{X})$.

We remark that $(0, \delta, 1 + \alpha)$ refers to test-time robustness without any train-time robustness considerations, while $(\delta, 0, 1 + \alpha)$ refers to train-time robustness. We will focus on two of the most common choices of distance metric — (i) the total variation distance d_{TV}, and (ii) the Wasserstein distance $d_{W_\infty(\ell_q)}$. The former is commonly used to capture contamination in training data (Huber, 2011). The latter is often used to capture adversarial perturbations. In Section A, we consider simple examples involving linear classification, regression and mean estimation. In these examples, we see that the setting when the test-time perturbations are measured in d_{TV} distance is less interesting.

2.2. Population setting: test-time to simultaneous train-and-test robustness

We start by showing that robustness as defined in Def. 2 is achievable in the population setting, when both train and test robustness are measured in the same metric. We prove that the existence of a hypothesis with good test-time robustness implies the existence of a hypothesis that has both train-time and test-time robustness in the same metric. Thus the goals of test-time robustness and train-time robustness are aligned. In fact, this simultaneous train-and-test robustness is achieved by an algorithm that given training samples from any distribution $\widetilde{P}$, loss function L and $\delta > 0$ outputs

$$\widehat{h} = \operatorname*{argmin}_{h \in H} \textit{robust-loss}(L, \widetilde{P}, h, \delta) = \operatorname*{argmin}_{h \in H} \sup_{P \in \mathsf{Ball}_d(\widetilde{P}, \delta)} L_P(h),$$

where d is the metric over distributions. We call the above algorithm $\mathsf{Robust\text{-}ERM}(\widetilde{P}, \delta)$. We first show a simple claim that suggests a connection between test-time robustness and TRAIN-AND-TEST robustness.

Claim 3 *Consider any hypothesis* H *over* $\mathcal{Z}$*, loss function* $L : \mathcal{M}(\mathcal{Z}) \times H \to \mathbb{R}$*, and a metric* $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$*. Suppose* $\alpha > 0$*,* $P^* \in \mathcal{M}(\mathcal{Z})$ *and let* $\eta(\delta) := \min_{h \in H} \textit{robust-loss}(L, P^*, h, \delta)$*. Then for any* $\delta_1, \delta_2 > 0$*, given training samples from any (corrupted) distribution* $\widetilde{P} \in \mathsf{Ball}_d(P^*, \delta_1)$*, the algorithm* $\mathsf{Robust\text{-}ERM}(\widetilde{P}, \delta_1 + \delta_2)$ *outputs* $\widehat{h} \in H$ *which is test-time robust w.r.t* P^* *i.e.,*

$$\textit{robust-loss}(L, P^*, \widehat{h}, \delta_2) = \sup_{Q \in \mathsf{Ball}_d(P^*, \delta_2)} L_Q(\widehat{h}) \leq \eta(2\delta_1 + \delta_2) \tag{4}$$

Hence for any $\delta_1, \delta_2 > 0$*,* $\mathsf{Robust\text{-}ERM}$ *achieves* $(\delta_1, \delta_2, \frac{\eta(2\delta_1 + \delta_2)}{\eta(\delta_2)})$-TRAIN-AND-TEST *robustness.*

When $\delta_1 = \delta_2 = \delta$, the above proposition provides a $(\delta, \delta, \alpha(\delta))$-TRAIN-AND-TEST robustness guarantee where $\alpha(\delta) = \eta(3\delta)/\eta(\delta)$. In general, the converse of the above proposition is not true, i.e. we do not expect that the existence of a good train robust hypothesis has implications for test robustness (i.e., test robustness is harder to achieve). The above theorem shows that a specific algorithm ($\mathsf{Robust\text{-}ERM}$) achieves train-and-test robustness to an extent that depends on the best achievable test-time robustness. However, this does not give such a translation for every algorithm.

Proof Consider the hypothesis $\widehat{h}$ output by $\textsf{Robust-ERM}(\widetilde{P}, \delta_1 + \delta_2)$. Showing that $\widehat{h}$ achieves a guarantee of the form (3) involves proving a upper bound on the robust loss of $\widehat{h}$, as well as a lower bound on the robust loss of the optimal hypothesis w.r.t distribution P^*. This is carried out using the following careful sequence of inequalities. The main observation is that the robust loss achieved by $\widehat{h} = \text{argmin}_{h \in H} \sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1+\delta_2)} L_Q(h)$ is given by

$$\sup_{Q \in \mathsf{Ball}_d(P^*, \delta_2)} L_Q(\widehat{h}) \leq \sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1+\delta_2)} L_Q(\widehat{h}) \qquad \text{(by triangle inequality)}$$

$$\leq \inf_{h \in H} \sup_{Q' \in \mathsf{Ball}_d(P^*, 2\delta_1+\delta_2)} L_{Q'}(h) = \eta(2\delta_1 + \delta_2). \qquad (5)$$

The inequality in (5) holds since $\mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2) \subseteq \mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)$; hence, the supremum over $\mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2)$ attained by (its minimizer) $\widehat{h}$ is smaller than supremum over $\mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)$ attained by any $h \in H$. Finally the last equality holds by definition of $\eta(2\delta_1 + \delta_2)$. ■

The above theorem does not involve any statistical or computational considerations. In fact, the Robust-ERM algorithm that achieves the above guarantee may not be statistical efficient or computationally efficient. The following sections show how one can design efficient algorithms that achieve train-and-test robustness according to Def. 2 in many settings. We remark that the above theorem also extends in a natural way to algorithms (other than Robust-ERM) that do not necessarily achieve the optimal robust test loss, but achieve it approximately (see Claim 17 in Appendix B).

3. Computationally efficient train-and-test-robust classification

In this section we design a polynomial time algorithm that achieves the simultaneous train-and-test robustness requirement in Eq (3) with $\alpha = O(1)$, when H is the class of linear classifiers. In particular, we will assume that the structured family $\mathcal{P}$ comprises all distributions P^* over $\mathbb{R}^n \times \{-1, +1\}$ such that the marginal of P^* is the Gaussian distribution $N(0, I)$. Let $H = \{h : x \mapsto \text{sgn}(w \cdot x) : w \in \mathbb{R}^n, \|w\|_2 = 1\}$. Let w^* be optimizer of the robust loss:

$$w^* = \underset{w \in H}{\text{argmin}}\, \textit{robust-loss}(L, P^*, w, \delta) = \underset{w \in H}{\text{argmin}} \sup_{P \in \mathsf{Ball}_d(P^*, \delta_2)} L_P(h).$$

The loss L is the $0/1$ loss, i.e., $L_P(w) = \mathbb{P}_{(x,y)\sim P}[y \neq \text{sgn}(w \cdot x)]$. We will assume that both the train and test perturbations are measured in the $d_{W_\infty(\ell_q)}$ metric. It will be instructive to think of $q = \infty$ for the results. Note that since the labels y take values in $\{-1, +1\}$ we only consider perturbations made to the marginal distribution over $\mathbb{R}^n$. Recall that our goal is to output a classifier that is competitive with w^* when given access to samples drawn from $\widetilde{P} \in \mathsf{Ball}_{W(\ell_q)}(P^*, \delta_1)$. Our main theorem is stated below.

Theorem 4 *There is an algorithm* Alg *(see Figure (1)) and absolute constants* $\alpha, c > 0$, *such that for any* $\delta_1, \delta_2, \varepsilon > 0$ *such that* $\delta_1 \leq c \cdot \delta_2$, Alg *takes as input* $m(\varepsilon, \delta_1, \delta_2) = \textit{poly}(n, \frac{1}{\varepsilon})$ *samples drawn i.i.d. from* $\widetilde{P} \in \mathsf{Ball}_{W(\ell_q)}(P^*, \delta_1)$, *runs in time polynomial in* $m(\varepsilon, \delta_1, \delta_2)$, *and outputs with probability at least* $2/3$, *a hypothesis* w *such that*

$$\textit{robust-loss}(L, P^*, w, \delta_2) \leq (1 + \alpha) \cdot \textit{robust-loss}(L, P^*, w^*, \delta_2) + \varepsilon.$$

We remark that one can use Claim 17 to obtain different tradeoffs in the constants for δ_1, δ_2 and α. In the rest of the section we outline our proposed algorithm and and ideas behind the proof of the above theorem. The formal proofs are in Appendix C. First we show that for the robust loss of a classifier to be small, it must be sparse (in an appropriate sense). This structural claim follows from the following two lemmas that prove the upper and the lower bounds on the robust loss of any classifier w. In what follows $\Phi(\cdot)$ is the CDF of a standard Gaussian, and for $a > 0$, $1 - 2\Phi(a) = \mathbb{P}_{x\sim N(0,1)}[|x| < a]$.

Lemma 5 (Upper bound on loss) *Suppose w is a classifier achieving error $\eta(w)$ on P^*. Then*

$$\textit{robust-loss}(L, P^*, w, \delta) \leq \eta + 1 - 2\Phi(\delta\|w\|_{q^*}) = \eta + O(\delta\|w\|_{q^*}). \tag{6}$$

When $\eta(w) = 0$, $\widetilde{L}(w) = 1 - 2\Phi(\delta\|w\|_{q^*})$.

In the above lemma q^* is the parameter for the dual norm of the ℓ_q norm. Before proceeding to the proof of the lemma, we state a simple claim below (see Appendix C for the proof) that will be used in establishing both the upper and the lower bounds.

Claim 6 *For a unit vector w with $\|w\|_2 = 1$, then the robust loss on P^**

$$\textit{robust-loss}(L, P^*, w, \delta) := \sup_{\substack{P: \\ d_{W_\infty(\ell_q)}(P,P^*)\leq\delta}} \mathop{\mathbb{E}}_{(\widetilde{x},y)\sim P}[\ell(w,(\widetilde{x},y))] = \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y\langle w,x\rangle < \delta\|w\|_{q^*}\Big]. \tag{7}$$

Proof [Proof of Lemma 5] First, let us handle the case $\eta(w) = 0$. Let us call this classifier w^*. This follows by direct substitution in Claim 6 and observing that for $(x,y) \sim P^*$, $y\langle w,x\rangle = |\langle w,x\rangle|$. In the general case, again using Claim 6

$$\begin{aligned}
\mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y\langle w,x\rangle < \delta\|w\|_{q^*}\Big] &= \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y = \text{sign}(\langle w,x\rangle)\ ;\ y\langle w,x\rangle < \delta\|w\|_{q^*}\Big] \\
&\quad + \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y \neq \text{sign}(\langle w,x\rangle)\ ;\ y\langle w,x\rangle < \delta\|w\|_{q^*}\Big] \\
&\leq \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[\text{sign}(\langle w,x\rangle)\langle w,x\rangle < \delta\|w\|_{q^*}\Big] + \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y \neq \text{sign}(\langle w,x\rangle)\Big] \\
&= 1 - 2\Phi(\delta\|w\|_{q^*}) + \eta.
\end{aligned}$$

■

The following claim shows a lower bound on the robust loss.

Lemma 7 (Lower bound on robust loss) *Suppose w is a classifier achieving error $\eta(w)$ on P^*. Then*

$$\textit{robust-loss}(L, P^*, w, \delta) \geq \frac{\eta(w)}{2} + \frac{1}{2}\Big(1 - 2\Phi(\delta\|w\|_{q^*})\Big) \geq \frac{\eta(w)}{2} + \Omega(1)\cdot\min\{\delta\|w\|_{q^*}, 1\}. \tag{8}$$

Proof Using Claim 6

$$\begin{aligned}\textit{robust-loss}(L, P^*, w, \delta) &= \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y\langle w, x\rangle < \delta\|w\|_{q^*}\Big] \\ &= \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y\langle w, x\rangle < 0\Big] + \mathop{\mathbb{P}}_{(x,y)\sim P_0}\Big[0 \le y\langle w, x\rangle < \delta\|w\|_{q^*}\Big] \\ &= \eta(w) + \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[0 \le y\langle w, x\rangle < \delta\|w\|_{q^*}\Big]. \qquad (9)\end{aligned}$$

We will lower bound the second term by coupling to $\mathcal{U} = \{u : 0 \le \langle w, u\rangle \le \delta\|w\|_{q^*}\}$. Let $h(x) = y \in \{\pm 1\}$ refer to the label of x. In particular, consider the coupling given by $\phi(x) = h(x)x$. Observe that $y\langle w, x\rangle = \langle w, \phi(x)\rangle$. Ideally this mapping is measure preserving and surjective (note that it is also at most 2-to-1, though this is not useful here). The following simple observation characterizes the "bad" points for the map.

Claim 8 *The set of $u \in \mathbb{R}^n$ with no preimages is given by*

$$\{u \in \mathbb{R}^n : \phi^{-1}(u) = \emptyset\} = \{u \in \mathbb{R}^n : h(u) = -1 \textit{ and } h(-u) = 1\}.$$

Moreover the measure of such $u \in \mathcal{U}$ with no preimage is upper bounded:

$$\mathop{\mathbb{P}}_{u\sim N(0,I)}\Big[u \in \mathcal{U};\ \phi^{-1}(u) = \emptyset\Big] \le \frac{\eta(w)}{2}.$$

We now finish the proof of Lemma 7 assuming the claim. By the second part of Claim 8,

$$\begin{aligned}\mathop{\mathbb{P}}_{x\sim N(0,I)}\Big[y\langle w, x\rangle < \delta\|w\|_1\Big] &= \mathop{\mathbb{P}}_{u\sim N(0,I)}[u \in \mathcal{U}] - \mathop{\mathbb{P}}_{u\sim N(0,I)}[u \in \mathcal{U} \text{ and } \phi^{-1}(u) = \emptyset] \\ &\ge \frac{1}{2}\Big(1 - 2\Phi(\delta\|w\|_{q^*})\Big) - \frac{\eta(w)}{2}.\end{aligned}$$

Substituting this in (9) finishes the proof of the lemma. ■

We now finish the proof of the inner claim in the above lemma.
Proof [Proof of Claim 8] Consider $x_1 = u$ and $x_2 = -u$. Since $\phi(x_1) \ne u$ we have $h(x_1) = h(u) = -1$. Similarly since $\phi(x_2) \ne u$, we have $h(-u) = 1$. This establishes the first part of the claim.

For the second part note that when $\phi^{-1}(u) = \emptyset$ and $\langle u, w\rangle \ge 0$, the first part implies that both $h(u)\langle w, u\rangle < 0$ and $h(-u)\langle w, -u\rangle < 0$. Hence both u and $-u$ are mislabeled by w, but $-u$ does not belong to $\mathcal{U}$. Hence

$$\mathop{\mathbb{P}}_{u}[u \in \mathcal{U},\ \phi^{-1}(u) = \emptyset] \le \frac{1}{2}\mathop{\mathbb{P}}_{u}[h(u)\langle w, u\rangle < 0] \le \frac{\eta(w)}{2}.$$

■

From Lemma 5 and Lemma 7 we know that w^* achieves a robust loss of $\Theta\big(\eta(w^*) + \kappa\delta_2\big)$, where $\kappa = \|w^*\|_{q^*}$. Furthermore, we can assume without loss of generality that $\kappa = O(\frac{1}{\delta_2})$ since otherwise any hypothesis w will be a constant approximation to the robust loss of w^*. Next notice that for any w such that $\|w - w^*\|_2 \le \varepsilon$ and $\|w\|_{q^*} \le \kappa$, *robust-loss*$(L, P^*, w, \delta_2) \le O(\eta(w^*) + \varepsilon + \kappa\delta_2)$.

Hence given labeled examples generated from $\widetilde{P}$, in order to get a constant factor guarantee for the objective in Eq. (3), our goal is to output a vector w such that $\|w - w^*\|_2 \leq O(\eta(w^*) + \kappa\delta_2)$ and $\|w\|_{q^*} \leq \kappa$.

Even without train-test corruptions, the above problem of learning w^* is non-trivial and corresponds to that of agnostically learning halfspaces. This problem has been widely studied and there exist efficient algorithms that can approximate w^* up to an $O(\eta(w^*))$ error (Klivans et al., 2009; Awasthi et al., 2014; Daniely, 2015; Diakonikolas et al., 2020b, 2021). In general these algorithms will not work when the train distribution is corrupted under the $d_{W_\infty(\ell_q)}$ metric. Even ignoring train-corruptions these algorithms will not achieve test-time robustness as there is no guarantee that the output hypothesis w will be sparse. However we show that a suitable adaptation of the agnostic learning algorithm of Awasthi et al. (2014) can indeed achieve the desired train-test guarantee. The algorithm of Awasthi et al. (2014) works by iteratively solving a series of hinge loss minimization problems in order to improve the current guess w_k for w^*. The algorithm focuses the minimization problem on parts of the distribution that get more and more concentrated around the current guess w_k. In particular, these regions are defined by $|w_k \cdot x| \leq b_k$ for a carefully chosen value of b_k. In the presence of corruptions in the $d_{W_\infty(\ell_q)}$ metric these regions could become very different from the "ideal" regions thereby rendering the iterative procedure meaningless. To overcome this we use the claim from Lemma 5 and Lemma 7 that the true classifier w^* is sparse. We show that if the optimization is performed over such sparse classifiers in each iteration, then one can define adjusted regions that are not too far from the ideal ones. Based on the above intuition our main algorithm is sketched in Figure 1. For simplicity we assume that $\kappa = \|w^*\|_{q^*}$ is known. A standard doubling trick can be used to remove this assumption. See Appendix C.

```
HINGELOSS($\widetilde{P}, \kappa, \delta_1, s, \varepsilon_0, w_0$)
Input: An Oracle to sample labeled examples from distribution $\widetilde{P}$, bounds on $\kappa$ and train corruption magnitude $\delta_1$, iteration bound $s$, precision value $\varepsilon_0$, initialization $w_0$.
  1. Draw $m_1$ labeled examples from $\widetilde{P}$ and add them to a set $W$.
  2. For $k = 1, \ldots s$,
       (a) Find $v_k$ such that $\|v_k - w_{k-1}\|_2 \leq r_k$, $\|v_k\|_2 \leq 1$ and $\|v_k\|_{q^*} \leq \kappa$ that achieves
             $\ell_{\tau_k}(v_k, W) \leq \min_{\substack{w:\|w-w_{k-1}\|_2 \leq r_k, \\ \|w\|_2 \leq 1, \|w\|_{q^*} \leq \kappa}} \ell_{\tau_k}(w, W) + \varepsilon_0.$
       (b) Set $w_k = \frac{v_k}{\|v_k\|_2}$.
       (c) Clear the working set $W$.
       (d) Until $m_{k+1}$ points are added to the set $W$, sample $(x, y)$ from $\widetilde{P}$ and add to $W$ if $|w_k \cdot x| \leq b_k$.
  3. Output: Return $w_s$.
```

Figure 1: Iterative Hinge Loss Minimization.

Parameter Settings. In the algorithm above we set the parameters as $b_k = c_1 2^{-k} + \kappa\delta_1$, $r_k = c_2 2^{-k} + \kappa\delta_1$ and $\tau_k = c_3 b_{k-1}$ for universal constants $c_1, c_2 > 0, c_3 \in (0, 1)$ are absolute constants.

Furthermore the hinge loss function $\ell_\tau(w, W)$ is defined as

$$\ell_\tau(w, W) = \frac{1}{|W|} \sum_{(x,y)\in W} \max\big(0, 1 - \frac{y(w \cdot x)}{\tau}\big). \tag{10}$$

Then we have the following guarantee that implies Theorem 4.

Theorem 9 *There exist universal constants $c, c_1, c_2, c_3, c_4 > 0$ such that the algorithm from Figure 1 when run with $s \geq \lceil \log(1/\varepsilon) \rceil$, $\kappa \leq O(\|w^*\|_{q^*})$, $\varepsilon_0 = c_3$, and w_0 such that $\theta(w_0, w^*) < \frac{\pi}{2}$, uses poly$(n, \frac{1}{\varepsilon})$ examples from $\widetilde{P}$ and outputs, with constant probability, w_s such that $\|w_s - w^*\|_2 \leq \varepsilon$, provided that $\eta(w^*) + \kappa\delta_2 \leq \varepsilon/c_4$ and $\delta_1 \leq c \cdot \delta_2$.*

Training corruptions in d_{TV}. We also study a more challenging setting where in addition to corruptions in the Wasserstein metric, the training data is also corrupted in the d_{TV} metric. We show that even under this stronger corruption model, we can achieve the robustness guarantee in Eq. (3). In order to do this we again follow the iterative hinge loss minimization approach outlined in Figure 1. However, we cannot simply minimize the hinge loss over the entire working set W (step 2(a) of the algorithm). This is due to the presence of "malicious" examples resulting from the fact that $\widetilde{P}$ consists of corruptions in d_{TV} as well. Instead, we first perform an outlier removal procedure that reweighs the working set W such that the variance of the data in all sparse directions is controlled. Such an outlier removal followed by a weighted hinge loss minimization lets us defend simultaneously against corruptions in $d_{W_\infty(\ell_q)}$ and d_{TV}. See Appendix C for the formal guarantee and proofs.

4. Computationally-efficient train-and-test-robust linear regression

We next turn our attention to the problem of linear regression and design a polynomial-time algorithm that achieves the simultaneous train-and-test requirement of (3) with $\alpha = O(1)$. As in the previous section, we will assume that the structured family $\mathcal{P}$ comprises of distributions P^* over $\mathbb{R}^n \times \mathbb{R}$, that follow the standard model of noisy linear regression, i.e, $(x, y) \sim P^*$ is generated by sampling $x \sim N(0, I)$, and $y = w_0^\top \cdot x + \xi$, where ξ is independent mean zero noise with bounded variance σ^2. The marginal of P^* over $\mathbb{R}^n$ is the standard normal distribution $N(0, I)$. Let $H = \{h : x \mapsto w \cdot x\}$. We assume that the loss L is the squared loss, i.e. $L_P(w) = \mathbb{E}_{(x,y)\sim P}(y - w \cdot x)^2$. Let w^* be the optimizer of the robust test loss i.e.,

$$w^* = \operatorname*{argmin}_{w\in H} \textit{robust-loss}(L, P^*, w, \delta_2) = \operatorname*{argmin}_{w\in H} \sup_{P\in \mathsf{Ball}_d(P^*,\delta_2)} L_P(h).$$

We will assume that both the train and the test perturbations are measured in the $d_{W_\infty(\ell_q)}$ metric. In contrast to the classification setting, in regression the perturbation can also affect the labels. We first observe that at the population-level one can derive a closed form expression for the robust test loss of any $w \in H$ by characterizing the worst-case perturbation, i.e. we have the following Lemma.

Lemma 10 *The robust loss of any $w \in H$ is given by,*

$$\begin{aligned}\textit{robust-loss}(L, P^*, w, \delta_2) = \mathbb{E}_{(x,y)\sim P^*}(y - w \cdot x)^2 + 2\delta_2 \|(w, 1)\|_{q^*} \mathbb{E}_{(x,y)\sim P^*}|y - w \cdot x| \\ + \delta_2^2 \|(w, 1)\|_{q^*}^2.\end{aligned}$$

Following, Claim 3, the above characterization suggests a natural strategy when, given samples from a distribution which is δ_1 far from P^*, of minimizing the robust test loss over a larger ball of radius $\delta_1 + \delta_2$. Formally, given samples from $\widetilde{P}$ we define the following estimate,

$$\widehat{w} = \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} (y_i - w \cdot x_i)^2 + 2(\delta_1 + \delta_2) \|(w, 1)\|_{q^*} \frac{1}{m} \sum_{i=1}^{m} |y_i - w \cdot x_i| \\ + (\delta_1 + \delta_2)^2 \|(w, 1)\|_{q^*}^2. \tag{11}$$

From a computational standpoint, we observe that the objective function is convex. In more detail, it can be written as the (sum of) the square of the positive convex function: $|y_i - w \cdot x_i| + (\delta_1 + \delta_2)(1 + \|w\|_{q^*})$. Thus, the optimization problem defining $\widehat{w}$ can be solved using sub-gradient descent.

Our main theorem stated below analyzes the statistical properties of the estimate $\widehat{w}$.

Theorem 11 *There are absolute constants $\alpha, c > 0$, such that for any $\delta_1, \delta_2, \varepsilon > 0$, such that $\delta_1 \leq c \cdot \delta_2$, with $m(\varepsilon, \delta_1, \delta_2) = poly(n, \frac{1}{\varepsilon}, \|w_0\|_2, \sigma)$ samples drawn i.i.d. from $\widetilde{P} \in \mathsf{Ball}_{W(\ell_q)}(P^*, \delta_1)$, the estimate $\widehat{w}$ in Eq. (11) with probability at least $2/3$, satisfies the condition that,*

$$\textit{robust-loss}(L, P^*, \widehat{w}, \delta_2) \leq (1 + \alpha) \cdot \textit{robust-loss}(L, P^*, w^*, \delta_2) + \varepsilon.$$

We outline the main ideas behind the proof of the theorem. The formal proofs are Appendix D. Our proof has two components. As a consequence of Claim 3, it is straightforward to verify that, under the assumptions of the Theorem above, the minimizer of the robust test loss at radius $\delta_1 + \delta_2$ yields an estimate that is $(\delta_1, \delta_2, O(1))$ train-and-test robust.

The technical challenge then is to analyze the statistical properties of $\widehat{w}$. We follow the standard technique, of showing uniform convergence of the objective in Eq. (11) to its population counterpart. However, we need to address several technical challenges. The first is that, even though the distribution P^* is Gaussian, the sampling distribution $\widetilde{P}$ can have very different concentration properties. We address this by showing that the relevant functionals $|y - w \cdot x|$ and its square, are still sub-Gaussian and concentrate sharply around their expectations. The second is the fact that naively we would require uniform convergence over an unbounded set (since the optimization is unconstrained). However, a careful argument shows that with high-probability the minimizer $\widehat{w}$ cannot be too far from w^* and so establishing uniform convergence over this smaller set of vectors w suffices.

5. Computationally-efficient train-and-test-robust mean estimation

Finally, we consider the problem of mean estimation. In this setting, we will make relatively weak assumptions on the structured family $\mathcal{P}$, assuming simply that the distribution P^* has (finite) mean μ^* and variance Σ^*. We again assume that the loss L is the squared loss, i.e. $L_P(w) = \mathbb{E}_{x \sim P} \|x - w\|_2^2$. and that perturbations are measured in the $d_{W_\infty(\ell_\infty)}$ metric.

At the population-level one can derive a closed form expression for the robust loss.

Lemma 12 *The robust test loss is given by,*

$$\textit{robust-loss}(L, P^*, w, \delta_2) = \mathbb{E}_{x \sim P^*} \|x - w\|_2^2 + 2\delta_2 \mathbb{E}_{x \sim P^*} \|x - w\|_1 + \delta_2^2 n.$$

As in the case of linear regression we define the following estimate,

$$\widehat{w} = \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \|x_i - w\|_2^2 + \frac{2(\delta_1 + \delta_2)}{m} \sum_{i=1}^{m} \|x_i - w\|_1. \tag{12}$$

When δ_1 and/or δ_2 are non-zero, the loss is intuitive and leads to an estimate which interpolates between the (coordinatewise) mean and median of the data. This objective function is convex and straightforward to minimize, and furthermore the objective is separable across its coordinates.

Our main theorem stated below analyzes the statistical properties of the estimate $\widehat{w}$.

Theorem 13 *There are absolute constants $\alpha, c > 0$, such that for any $\delta_1, \delta_2, \varepsilon > 0$, such that $\delta_1 \leq c{\cdot}\delta_2$, with $m(\varepsilon, \delta_1, \delta_2) = poly(n, \frac{1}{\varepsilon}, \|\mu^*\|_2, \|\Sigma^*\|)$ samples drawn i.i.d. from $\widetilde{P} \in \mathsf{Ball}_{W(\ell_\infty)}(P^*, \delta_1)$, the estimate $\widehat{w}$ in (11) with probability at least $2/3$, satisfies the condition that,*

$$\textit{robust-loss}(L, P^*, \widehat{w}, \delta_2) \leq (1 + \alpha) \cdot \textit{robust-loss}(L, P^*, w^*, \delta_2) + \varepsilon.$$

Once again, we proceed in two steps. We first show that the population robust loss minimizer corresponding to Eq. (12) has favorable properties. In contrast to our previous result on linear regression, in this case we directly analyze the minimizer of the program in Eq. (12) via the KKT conditions. Exploiting the fact that the optimization problem decomposes coordinate wise, and that along each coordinate the minimizer has a relatively simple structure (trading-off a squared and absolute loss) we give a sharp analysis of the statistical properties of $\widehat{w}$ in Appendix E.

6. Alternative Definition and Discussion

An alternate definition of train-and-test robustness is one that we call the "Weak Oracle" variant. Here the optimal algorithm (oracle) is also penalized for corruptions in the training set. In contrast to Def. 2 where the optimal algorithm has knowledge of the uncorrupted distribution $P^* \in \mathcal{P}$ (hence called "Strong Oracle"), in the Weak Oracle definition, the optimal algorithm is given access only to $\widetilde{P}$. Hence our algorithm is compared against a weaker benchmark in the new definition.

While the above is less natural than Def. 2, it has the advantage that one can achieve non-trivial guarantees in settings where robust learning is infeasible against the stronger benchmark (Strong Oracle). E.g., consider the *realizable* setting where there exists an h^* such that $L_{P^*}(h^*) = 0$. If the radius of test perturbation δ_2 also equals zero, then the optimal robust loss equals the optimal loss at P^* and hence is zero. However, given only access to a corrupted distribution ($\delta_1 \neq 0$), it would be information theoretically impossible for any algorithm to achieve zero robust test loss on P^*.

In order to formalize this, for a given $\widetilde{P}$ that is a corruption of an unknown $P^* \in \mathcal{P}$, define $\mathcal{D}_{\widetilde{P}}$ to be the set of possible test distributions. This set can be obtained by first identifying all distributions P_1 within $\mathcal{P}$ that $\widetilde{P}$ could be a corruption of, and then identifying all distributions P_2 that could be test corruptions of P_1. Formally we have that

$$\mathcal{D}_{\widetilde{P}} = \{P_2 : \exists P_1, d_{\mathsf{train}}(\widetilde{P}, P_1) \leq \delta_1, \text{ and, } d_{\mathsf{test}}(P_1, P_2) \leq \delta_2\}. \tag{13}$$

Then we have the following alternate definition of train-and-test robustness.

Definition 14 (WEAK ORACLE TRAIN-AND-TEST ROBUSTNESS) *Suppose $\delta_1, \delta_2, \alpha > 0$, and $\mathcal{M} = \mathcal{M}(\mathcal{Z})$ be the space of distributions over the space $\mathcal{Z}$, and let H be a hypothesis class* *equipped with the loss function L. An algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$-*TRAIN-AND-TEST ROBUST *w.r.t $\mathcal{P}, H$ i.f.f. for any $\varepsilon > 0$, given $m(\varepsilon, \delta_1, \delta_2)$ i.i.d. samples S drawn from $\widetilde{P}$, Alg outputs a function $\widehat{h} = \mathsf{Alg}(S)$,*

$$\mathop{\mathbb{P}}_{S\sim\widetilde{P}}\left[\sup_{P\in\mathcal{D}_{\widetilde{P}}} L_P(\widehat{h}) \le (1+\alpha)\min_{h\in H}\sup_{P\in\mathcal{D}_{\widetilde{P}}} L_P(h) + \varepsilon\right] \ge \frac{2}{3}, \quad \textit{where } \widehat{h} = \mathsf{Alg}(S). \tag{14}$$

Note there is an unbounded algorithm (with access to infinite samples and time) that always satisfies the above definition with $\alpha = 1$. Hence this new definition can potentially avoid the situation where the optimal algorithm (or oracle) achieves zero loss. On the other hand, the definition of $\mathcal{D}_{\widetilde{P}}$ and the corresponding loss minimization objective makes it hard to study the design of efficient algorithms in this setting. Moreover, a guarantee as in (14) is relative, and does not necessarily tell us if we can achieve good robust test error. Nevertheless, we prove that if an algorithm is robust according to one notion of train-and-test robustness, then it also achieves train-and-test robustness in the other definition (with some loss in parameters).

(Informal) Theorem 15 *If an algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 2, then it is also $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 14. Conversely, robustness according to Definition 14 implies robustness according to Def. 2 with an extra slack factor that allows the algorithm to compete with a benchmark defined over a slightly larger perturbation radius.*

See Appendix F for a formal statement and details. To conclude, we studied the design of learning algorithms that are simultaneously robust to corruptions both at train-time and test-time. Our new notion of train-and-test robustness leads to several algorithmic and statistical implications. It allowed us to show close connections between test-time and train-time robustness, and to demonstrate that techniques designed for robustness to one form of corruption can be adapted to obtain algorithms that are simultaneously robust against both kinds of corruptions. It would be interesting to study our proposed notion of robustness in other learning settings to get reliable and efficient reliable algorithms.

Acknowledgments

The second and third authors were supported by a Google Research Scholar program. In addition the last author was supported by NSF grants CCF-1652491, CCF-1637585 and CCF-1934931.

References

Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 72–80. ACM, 2004.

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

Pranjal Awasthi, Abhratanu Dutta, and Aravindan Vijayaraghavan. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems*, pages 13737–13747, 2019.

Pranjal Awasthi, Xue Chen, and Aravindan Vijayaraghavan. Estimating principal components under adversarial perturbations. In *Conference on Learning Theory*, pages 323–362. PMLR, 2020.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. 2008.

Avrim Blum, Alan Frieze, Ravi Kannan, and Santosh Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1):35–52, 1998.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*, 2018a.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018b.

Brian Bullins, Elad Hazan, and Tomer Koren. The limits of learning with missing data. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3503–3511, 2016.

Mengjie Chen, Chao Gao, Zhao Ren, et al. A general decision theory for huber's ε-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016.

Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

Amit Daniely. A ptas for agnostically learning halfspaces. In *Conference on Learning Theory*, pages 484–502, 2015.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018a.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Learning geometric concepts with nasty noise. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1061–1073. ACM, 2018b.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 48(2):742–864, 2019.

Ilias Diakonikolas, Daniel M Kane, and Pasin Manurangsi. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *arXiv preprint arXiv:2007.15220*, 2020a.

Ilias Diakonikolas, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Non-convex sgd learns halfspaces with adversarial label noise. *arXiv preprint arXiv:2006.06742*, 2020b.

Ilias Diakonikolas, Daniel M Kane, Vasilis Kontonis, Christos Tzamos, and Nikos Zarifis. Agnostic proper learning of halfspaces under gaussian marginals. *arXiv preprint arXiv:2102.05629*, 2021.

Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Thao Nguyen. Robust and private learning of halfspaces. In *International Conference on Artificial Intelligence and Statistics*, pages 1603–1611. PMLR, 2021.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Peter J Huber. *Robust statistics*. Springer, 2011.

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

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Adam R Klivans, Philip M Long, and Rocco A Servedio. Learning halfspaces with malicious noise. *Journal of Machine Learning Research*, 10(Dec):2715–2740, 2009.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21:1041–1048, 2008.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pages 7010–7021. PMLR, 2020.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. *arXiv preprint arXiv:2102.02145*, 2021.

Preetum Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, volume 29, pages 2208–2216, 2016.

David F Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial intelligence review*, 33(4): 275–306, 2010.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

Appendix A. Simple examples

In this section we present via a simple example the behavior of our proposed notion of train-and-test robustness when perturbations are measured in total variation distance. In particular, we will show that the standard *empirical risk minimization* (ERM) based algorithm achieves competitive guarantees. More formally, we have the following theorem.

Theorem 16 *Let H be a class of functions from $\mathcal{X}$ to $\{-1,+1\}$ and $\mathcal{P}$ be a family of distributions over $\mathcal{X} \times \{-1,+1\}$. Furthermore, let the perturbations and d_{test} be measured in d_{TV}. Given access to a corrupted distribution $\widetilde{P}$, let Alg be the algorithm that performs empirical risk minimization (on the population loss) over $\widetilde{P}$, i.e., Alg outputs $\widehat{h}$ such that*

$$\widehat{h} = \operatorname*{argmin}_{h \in H} L_{\widetilde{P}}(h).$$

Here $L_{\widetilde{P}}(h)$ is the $0/1$ loss defined as $L_{\widetilde{P}}(h) = \mathbb{P}_{(x,y)\sim\widetilde{P}}[h(x) \neq y]$. Then for any $\delta_1, \delta_2 \geq 0$, Alg is $(\delta_1, \delta_2, (1+\alpha))$ TRAIN-AND-TEST ROBUST w.r.t. $\mathcal{P}$, for $\alpha = 1 + \frac{2\delta_1}{\delta_2}$.

Proof First recall that for any two distributions P, Q and any $h \in H$ we have that

$$|L_P(h) - L_Q(h)| \leq d_{\mathsf{TV}}(P, Q).$$

Next, fix a particular corrupted distribution $\widetilde{P}$ and let $P^* \in \mathcal{P}$ be such that $d_{\mathsf{TV}}(\widetilde{P}, P^*) \leq \delta_1$. Then we have

$$\begin{aligned}
\sup_{P \in \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_2)} L_P(\widehat{h}) &\leq L_{\widetilde{P}}(\widehat{h}) + \delta_1 + \delta_2 \\
&\leq \operatorname*{argmin}_{h \in H} L_{\widetilde{P}}(h) + \delta_1 + \delta_2 \\
&\leq \operatorname*{argmin}_{h \in H} L_{P^*}(h) + 2\delta_1 + \delta_2 \\
&\leq \operatorname*{argmin}_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_2)} L_{P^*}(h) + 2\delta_1 + \delta_2. \qquad (15)
\end{aligned}$$

Furthermore, notice that for any $h \in H$

$$\sup_{P \in \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_2)} L_{P^*}(h) \geq \delta_2. \qquad (16)$$

The above can be achieved by taking a distribution that is $(1-\delta_2)P^* + \delta_2 Q$, where Q is a distribution with any arbitrary marginal over $\mathcal{X}$ and for each $x \in \mathcal{X}$, $Q_{y|x}$ is supported entirely on $-h(x)$. In other words, we take an arbitrary δ_2 fraction of the mass in P^* and change its class labels to ensure that h makes a mistake in the entire region.

By combining Eq. (15) and Eq. (16) we get that

$$\sup_{P \in \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_2)} L_P(\widehat{h}) \leq (1+\alpha) \operatorname*{argmin}_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_2)} L_P(h), \qquad (17)$$

for $\alpha = (1 + \frac{2\delta_1}{\delta_2})$. ■

Finally, notice that if the loss $L_P(h)$ is unbounded, as is typically the case for problems such as mean estimation and regression, then the *robust-loss* under test perturbations in d_{TV} is ∞ for any hypothesis h, and hence getting non-trivial guarantees according to Definition 2 is not possible.

Appendix B. Translation from Test-time to Train-and-Test Robustness: Proof from Section 2.2

Claim 17 *Consider any hypothesis H over $\mathcal{Z}$, loss function $L : \mathcal{M}(\mathcal{Z}) \times H \to \mathbb{R}$, and a metric $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$. Suppose $\alpha > 0$ and $P^* \in \mathcal{M}(\mathcal{Z})$, suppose $\forall \delta > 0$, $\eta(\delta) := \min_{h \in H}$ robust-loss(L, P^*, h, δ). Let $\delta_1, \delta_2 > 0$. Suppose for any distribution $P \in \mathcal{M}$, the algorithm $\mathsf{Alg}(P, \delta = \delta_1 + \delta_2)$ with robustness parameter $\delta = \delta_1 + \delta_2$ on samples from P is $(0, \delta_1 + \delta_2, \alpha)$-*TRAIN-AND-TEST *robust (i.e., it is test-time robust up to an approximation factor of α). Then given training samples from any (corrupted) distribution $\widetilde{P} \in \mathsf{Ball}_d(P^*, \delta_1)$, the algorithm $\mathsf{Alg}(\widetilde{P}, \delta_1 + \delta_2)$ outputs $\widehat{h} \in H$ which is also test-time robust w.r.t P^* i.e.,*

$$\textit{robust-loss}(L, P^*, \widehat{h}, \delta_2) = \sup_{Q \in \mathsf{Ball}_d(P^*, \delta_2)} L_Q(\widehat{h}) \leq \alpha(\delta_1 + \delta_2)\eta(2\delta_1, \delta_2) \tag{18}$$

*Hence for any $\delta_1, \delta_2 > 0$, it achieves $(\delta_1, \delta_2, \alpha(\delta_1 + \delta_2) \cdot \frac{\eta(2\delta_1 + \delta_2)}{\eta(\delta_2)})$-*TRAIN-AND-TEST *robustness.*

Proof [Proof of Claim 17]

We are given training samples from $\widetilde{P}$ with $d(P^*, \widetilde{P}) \leq \delta_1$. Consider the hypothesis $\widehat{h}$ output by $\mathsf{Alg}(\widetilde{P}, \delta_1 + \delta_2)$. Let $\alpha := \alpha(\delta_1 + \delta_2)$ for convenience. It satisfies

$$\sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2)} L_Q(\widehat{h}) \leq \alpha \cdot \operatorname*{argmin}_{h \in H} \sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2)} L_Q(h). \tag{19}$$

The main observation is that the robust test loss achieved by $\widehat{h}$ is

$$\begin{aligned}
\sup_{Q \in \mathsf{Ball}_d(P^*, \delta_2)} L_Q(\widehat{h}) &\leq \sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2)} L_Q(\widehat{h}) && \text{(by triangle inequality)} \\
&\leq \alpha \times \inf_{h \in H} \sup_{Q \in \mathsf{Ball}_d(\widetilde{P}, \delta_1 + \delta_2)} L_Q(\widehat{h}) && \text{(by the } \alpha \text{ approximate guarantee)} \\
&\leq \alpha \cdot \inf_{h \in H} \sup_{Q' \in \mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)} L_{Q'}(h) = \alpha \cdot \eta(2\delta_1, \delta_2). && (20)
\end{aligned}$$

The inequality in (20) is the crucial one, and holds since $\mathsf{Ball}_d(\widehat{P}, \delta_1 + \delta_2) \subseteq \mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)$; hence, the supremum over $\mathsf{Ball}_d(\widehat{P}, \delta_1 + \delta_2)$ attained by its minimizer is smaller than supremum over $\mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)$ attained by any $h \in H$. Finally the last equality holds by definition of $\eta(2\delta_1 + \delta_2)$. The furthermore part holds just because

■

Appendix C. Linear Classification: Proof from Section 3

We first begin with the proof of Claim 6.

Claim 18 (Claim 6 restated.) *For a unit vector w with $\|w\|_2 = 1$, then the robust loss on P^**

$$\textit{robust-loss}(L, P^*, w, \delta) := \sup_{\substack{P: \\ d_{W_\infty(\ell_q)}(P, P^*) \leq \delta}} \mathbb{E}_{(\widetilde{x}, y) \sim P}[\ell(w, (\widetilde{x}, y))] = \mathbb{P}_{(x,y) \sim P^*}\left[y\langle w, x\rangle < \delta\|w\|_{q^*}\right]. \tag{21}$$

Proof For each point x and a fixed w, the worst perturbation is deterministic. In particular,

$$\begin{aligned}\textit{robust-loss}(L, P^*, w, \delta) &= \sup_{\substack{P: \\ d_{W_\infty(\ell_q)}(P,P^*)\le\delta}} \mathop{\mathbb{E}}_{(\widetilde{x},y)\sim P}[\ell(w,(\widetilde{x},y))] = \mathop{\mathbb{E}}_{(x,y)\sim P^*} \sup_{\|z\|_q\le\delta} \mathbb{I}\Big[y\langle w, x+z\rangle < 0\Big] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P^*} \sup_{\|z\|_q\le\delta} \mathbb{I}\Big[y\langle w, x\rangle + y\langle w, z\rangle < 0\Big] \\ &= \mathop{\mathbb{E}}_{(x,y)\sim P^*} \mathbb{I}\Big[y\langle w, x\rangle - \delta\|w\|_{q^*}\Big] = \mathop{\mathbb{P}}_{(x,y)\sim P^*}\Big[y\langle w, x\rangle < \delta\|w\|_{q^*}\Big].\end{aligned}$$

■

We next prove the main theorem, i.e., Theorem 9. We start with two simple lemmas relating the distribution $\widetilde{P}$ to P^* that will be used throughout the analysis. Below we use $\widetilde{P}_X$ to denote the marginal distribution over x.

Lemma 19 *Let $\widetilde{P}$ be such that $d_{W_\infty(\ell_q)}(\widetilde{P}_X, N(0,I)) \le \delta_1$. Then for any $u, v \in \mathbb{R}^n$ such that $\|u\|_2 = \|v\|_2 = 1$ and $\|u\|_{q^*}, \|v\|_{q^*} \le \kappa$ it holds that*

$$\mathop{\mathbb{P}}_{x\sim\widetilde{P}_X}[u\cdot x \in [a,b]] \le |b-a| + 2\kappa\delta \tag{22}$$

$$\mathop{\mathbb{P}}_{x\sim\widetilde{P}_X}[sgn(u\cdot x) \ne sgn(v\cdot x)] \ge c\cdot\Theta(u,v) - 4\kappa\delta \tag{23}$$

where $c > 0$ is a universal constant. Furthermore, for any $c_5 > 0$ there exists $c_6 > 0$ such that

$$\mathop{\mathbb{P}}_{x\sim\widetilde{P}_X}[sgn(u\cdot x) \ne sgn(v\cdot x), |u\cdot x| > c_6\Theta(u,v) + \kappa\delta] \le c_5\Theta(u,v) + 4\kappa\delta. \tag{24}$$

Proof The proof is follows easily from the following statements that were shown in Awasthi et al. (2014). For any u, v such that $\|u\|_2, \|v\|_2 = 1$ it holds that

$$\mathop{\mathbb{P}}_{x\sim P_X}[u\cdot x \in [a,b]] \le |b-a| \tag{25}$$

$$\mathop{\mathbb{P}}_{x\sim P_X}[\text{sgn}(u\cdot x) \ne \text{sgn}(v\cdot x)] \ge c\cdot\Theta(u,v). \tag{26}$$

Next we have that for any u, v such that $\|u\|_{q^*}, \|v\|_{q^*} \le \kappa$,

$$\begin{aligned}\mathop{\mathbb{P}}_{x\sim\widetilde{P}_X}[u\cdot x \in [a,b]] &\le \mathop{\mathbb{P}}_{x\sim P_X}[u\cdot x \in [a-\kappa\delta, b+\kappa\delta]] \\ &\le |b-a| + 2\kappa\delta.\end{aligned}$$

Next we have that

$$\begin{aligned}\mathop{\mathbb{P}}_{x\sim\widetilde{P}_X}[\text{sgn}(u\cdot x) \ne \text{sgn}(v\cdot x)] &\ge \mathop{\mathbb{P}}_{x\sim P_X}[\text{sgn}(u\cdot x) \ne \text{sgn}(v\cdot x)] - \mathop{\mathbb{P}}_{x\sim P_X}[|u\cdot x| \le \kappa\delta \text{ or } |v\cdot x| \le \kappa\delta] \\ &\ge c\cdot\Theta(u,v) - 4\kappa\delta.\end{aligned}$$

Finally, the authors in Awasthi et al. (2014) also proved that for any $c_5 > 0$ there exists $c_6 > 0$ such that

$$\mathop{\mathbb{P}}_{x\sim P_X}[\text{sgn}(u\cdot x) \ne \text{sgn}(v\cdot x), |u\cdot x| > c_6\Theta(u,v)] \le c_5\Theta(u,v). \tag{27}$$

From the above it follows that

$$\begin{aligned}\underset{x\sim\widetilde{P}_X}{\mathbb{P}}[\mathrm{sgn}(u\cdot x)\neq\mathrm{sgn}(v\cdot x), |u\cdot x|>c_6\Theta(u,v)+\kappa\delta] &\leq \underset{x\sim P_X}{\mathbb{P}}[\mathrm{sgn}(u\cdot x)\neq\mathrm{sgn}(v\cdot x), |u\cdot x|>c_6\Theta(u,v)]\\ &+\underset{x\sim P_X}{\mathbb{P}}[|u\cdot x|\leq\kappa\delta \text{ or } |v\cdot x|\leq\kappa\delta]\\ &\leq c_5\Theta(u,v)+4\kappa\delta.\end{aligned}$$

■

Proof [Proof of Theorem 9] Define $err(w)=\mathbb{P}_{x\sim\widetilde{P}_X}[\mathrm{sgn}(w^*\cdot x)\neq\mathrm{sgn}(w\cdot x)]$. We will prove by induction that at round k, with probability at least $1-\frac{k}{3(k+1)}$, we have that $err(w_k)\leq 2^{-k}+c\kappa\delta$ for a universal constant $c>0$. Clearly the induction hypothesis holds at $k=0$. Let's assume that the hypothesis holds at $k-1$. Then conditioned on the event that $err(w_{k-1})\leq 2^{-k}+c$, we have that for any w_k such that $\|w_k-w_{k-1}\|\leq r_{k-1}$ and $S_{k-1}=\{x:|w_{k-1}\cdot x|\leq b_{k-1}\}$ we have from (27) that

$$\underset{x\sim\widetilde{P}_X}{\mathbb{P}}[\mathrm{sgn}(w^*\cdot x)\neq\mathrm{sgn}(w_k\cdot x)]\leq\frac{2^{-k}}{4}+8\kappa\delta.$$

Hence we get that

$$err(w_k)\leq\mathbb{P}(S_{k-1})err(w_k|S_{k-1})+\frac{2^{-k}}{4}+8\kappa\delta. \tag{28}$$

The guarantee of Lemma 20 now shows that with probability at least $\frac{1}{3(k+k^2)}$ we have that $err(w_k|S_{k-1})\leq 1/8$. Substituting into (28) we get that with probability at least $1-\frac{k}{3(k+1)}$, $err(w_k)\leq 2^{-k}+c\kappa\delta$ for a universal constant $c\geq 15$. ■

Lemma 20 *With probability at least* $1-\frac{1}{3(k+k^2)}$, *in round* k *of the Algorithm in Figure 1, we have that* $err(w_k|S_{k-1})\leq\frac{1}{8}$.

Proof The proof of the lemma is analogous to the proof of Theorem 3.6 in Awasthi et al. (2014). For round k define P_k be the conditional distribution of examples in the region S_{k-1} and labeled according to $\mathrm{sgn}(w^*\cdot x)$, and define $\widetilde{P}_k$ to be the distribution with the true noisy labels. We first argue that the true classifier w^* has small hinge loss under P_k. To see this notice that

$$\underset{(x,y)\sim P_k}{\mathbb{E}}\ell_{\tau_k}(w^*)=\frac{\mathbb{P}[|w^*\cdot x|\leq\tau_k]}{\mathbb{P}[|w_{k-1}\cdot x|\leq b_{k-1}]} \tag{29}$$

$$\leq\frac{1}{64} \tag{30}$$

for an appropriate choice of the constants c_1 and c_3.

Next we bound how the hinge loss differs when looking at the true noise distribution $\widetilde{P}_k$ as compared to P_k. In particular we will show that for any $w\in B(w_{k-1},r_k)$ and $\|w\|_{q^*}\leq\kappa$ we have that

$$\Big|\underset{(x,y)\sim P_k}{\mathbb{E}}\ell_{\tau_k}(w)-\underset{(x,y)\sim\widetilde{P}_k}{\mathbb{E}}\ell_{\tau_k}(w)\Big|\leq O\Big(\sqrt{\frac{\eta}{\varepsilon}}\Big)\frac{z_k}{\tau_k},$$

where $z_k = \sqrt{r_k^2 + b_{k-1}^2}$. To establish this we follow the analysis in Lemma 3.8 of Awasthi et al. (2014) and get

$$\begin{aligned}\Big|\mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(w) - \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} \ell_{\tau_k}(w)\Big| &= \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} 1_{(x,y)\in N}\big|\ell_{\tau_k}(w,x,y) - \ell_{\tau_k}(w,x,-y)\big| && (31)\\ &\le 2\mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} 1_{(x,y)\in N}\frac{|w\cdot x|}{\tau_k} && (32)\\ &\le \frac{2}{\tau_k}\sqrt{\mathop{\mathbb{P}}_{(x,y)\sim \widetilde{P}_k}(N)}\sqrt{\mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k}|w\cdot x|^2}. && (33)\end{aligned}$$

Here N is the set of noisy examples, i.e., the examples where the label y does not match $\text{sgn}(w^* \cdot x)$. ■

The lemma follows from noticing that

$$\begin{aligned}\mathop{\mathbb{P}}_{(x,y)\sim \widetilde{P}_k}(N) &= \frac{\mathbb{P}_{(x,y)\sim \widetilde{P}}(N)}{\mathbb{P}(S_{k-1})}\\ &= O(\frac{\eta(w^*)}{\varepsilon}),\end{aligned}$$

and that

$$\begin{aligned}\mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k}|w\cdot x|^2 &\le 2\mathop{\mathbb{E}}_{(x,y)\sim P_k}|w\cdot x|^2 + 2\kappa^2\delta^2\\ &= O(z_k^2)\end{aligned}$$

The last inequality follows from Lemma 3.4 of Awasthi et al. (2014) who showed that $\mathbb{E}_{(x,y)\sim P_k}|w\cdot x|^2 = O(r_k^2 + b_{k-1}^2)$. Finally, we need the statement of Lemma 21 below that follows from standard concentration bounds and relates the hinge losses on the distributions P_k and $\widetilde{P}_k$ to their finite sample counterparts.

Finally combining everything we have

$$\begin{aligned}err(w_k|S_{k-1}) &= err(v_k|S_{k-1})\\ &\le \mathop{\mathbb{E}}_{(x,y)\sim P_k}\ell_{\tau_k}(v_k)\\ &\le \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k}\ell_{\tau_k}(v_k) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k}.\end{aligned}$$

Using Lemma 21 we get

$$\begin{aligned}err(w_k|S_{k-1}) &\le \ell(v_k, W) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \quad \text{(from Lemma 21)}\\ &\le \ell(w^*, W) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k}\\ &\le \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k}\ell_{\tau_k}(w^*) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \quad \text{(Lemma 21)}\\ &\le \mathop{\mathbb{E}}_{(x,y)\sim P_k}\ell_{\tau_k}(w^*) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \le \frac{1}{8} \quad \text{(Lemma 21).}\end{aligned}$$

Lemma 21 (Lemma 3.9 of Awasthi et al. (2014)) *Given W at round k as defined in the algorithm in Figure 1, defined*

$$\textit{cleaned}(W) = \{(x, \textit{sgn}(w^* \cdot x)) : (x, y) \in W\}.$$

Then with probability at least $1 - \frac{1}{3(k+k^2)}$ if holds that

$$\Big| \mathop{\mathbb{E}}_{(x,y)\sim P_k} [\ell_{\tau_k}(w)] - \ell_{\tau_k}(w, \textit{cleaned}(W)) \Big| \leq \frac{1}{64} \tag{34}$$

$$\Big| \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} [\ell_{\tau_k}(w)] - \ell_{\tau_k}(w, W) \Big| \leq \frac{1}{64} \tag{35}$$

```
HINGELOSS($\widetilde{P}, \kappa, \delta_1, s, \varepsilon_0, w_0$)
Input: An Oracle to sample labeled examples from distribution $\widetilde{P}$, bounds on $\kappa$ and train corruption
magnitude $\delta_1$, iteration bound $s$, precision value $\varepsilon_0$, initialization $w_0$.
  1. Draw $m_1$ labeled examples from $\widetilde{P}$ and add them to a set $W$.
  2. For $k = 1, \dots s$,
      (a) Find $v_k$ such that $\|v_k - w_{k-1}\|_2 \leq r_k$, $\|v_k\|_2 \leq 1$ and $\|v_k\|_{q^*} \leq \kappa$ that achieves
              $\ell_{\tau_k}(v_k, W) \leq \min_{\substack{w:\|w-w_{k-1}\|_2 \leq r_k, \\ \|w\|_2 \leq 1, \|w\|_{q^*} \leq \kappa}} \ell_{\tau_k}(v_k, W) + \varepsilon_0.$
      (b) Set $w_k = \frac{v_k}{\|v_k\|_2}$.
      (c) Clear the working set $W$.
      (d) Until $m_{k+1}$ points are added to the set $W$, sample $(x, y)$ from $\widetilde{P}$ and add to $W$ if
          $|w_k \cdot x| \leq b_k$.
      (e) $W \leftarrow$ OUTLIER-REMOVAL$(W, b_k^2 + r_{k+1}^2, \kappa/16, O(\frac{r_k^2}{\tau_k^2 + b_{k-1}^2}))$.
  3. Output: Return $w_s$.
```

Figure 2: Iterative Hinge Loss Minimization for robustness against train corruptions in $d_{W_\infty(\ell_q)}$ and d_{TV}.

Handling train corruptions in TV distance. In order to deal with corruptions to the training set in d_{TV} metric in addition to the $d_{W_\infty(\ell_q)}$ metric we extend the hinge loss based algorithm from Figure 1 by incorporating an outlier removal subroutine. The updated algorithm is shown in Figure 2. Recall from the proof of Lemma 20 that the analysis of the Algorithm from Figure 1 relies on the fact that at each iteration k, the variance of the data in any sparse direction w, i.e., $|w|_{q^*} \leq \kappa$, according to the distribution $\widetilde{P}_k$ is bounded by z_k^2. The goal of the outlier removal subroutine is to ensure that by appropriately reweighting the set W the same variance bound can be maintained. The outlier removal subrotine itself is a convex program in the weights w_i. In order to be able to efficiently solve the program we need a separation oracle for the constraint $\|M\|_{q\to 2} \leq \kappa$. Recall that the $\|\|_{q\to 2}$ norm of a matrix is defined as

$$\|M\|_{q\to 2} = \max_{v:\|v\|_q=1} \|Mv\|_2.$$

While computing the $\|\|_{q\to 2}$ norm exactly is NP-hard, there exist polynomial time algorithms that approximate the optimal value of the norm (for $q \geq 2$) up to a constant factor Alon and Naor (2004). This is enough for us to implement an approximate separation oracle. As a result we have the following guarantee for the algorithm in Figure 2 that implies Theorem 23.

Parameter Settings. In the algorithm above we set the parameters as $b_k = c_1 2^{-k} + \kappa\delta_1$, $r_k = c_2 2^{-k} + \kappa\delta_1$ and $\tau_k = c_3 b_{k-1}$ for universal constants $c_1, c_2 > 0, c_3 \in (0,1)$ to be defined later.

Theorem 22 *There exist universal constants $c, c_1, c_2, c_3, c_4 > 0$ such that the algorithm from Figure 2 when run with $s \geq \lceil \log(1/\varepsilon) \rceil$, $\kappa \leq O(\|w^*\|_{q^*})$, $\varepsilon_0 = c_3$, and w_0 such that $\theta(w_0, w^*) < \frac{\pi}{2}$, uses poly$(n, \frac{1}{\varepsilon})$ examples from $\widetilde{P}$ and outputs, with constant probability, w_s such that $\|w_s - w^*\|_2 \leq \varepsilon$, provided that $\eta(w^*) + \kappa\delta_2 \leq \varepsilon/c_4$ and $\delta_1 \leq c \cdot \delta_2$ and $\delta_3 \leq c \cdot \eta(w^*)$.*

OUTLIER-REMOVAL$(W, \sigma^2, \kappa, \xi)$
Input: A set W of examples, variance bound σ^2, sparsity bound κ.

1. Let $(x_1, y_1), \ldots (x_m, y_m)$ be the samples in W.
2. Find weights $w_1, w_2, \ldots, w_n$ such that, $w_i \in [0,1]$ and $\sum_i w_i \geq m(1-\xi)$, and for all $M \in \mathbb{R}^{n\times n}$ such that $M \succeq 0$ and $\|M\|_{q\to 2} \leq \kappa$ it holds that
$$\langle \sum_i w_i x_i x_i^\top, M \rangle \leq O(\sigma^2).$$
3. **Output:** Return the weighted set of examples.

Figure 3: Outlier removal procedure.

Theorem 23 *There is an algorithm* Alg *(see Figure 1) and absolute constants $\alpha, c > 0$, such that for any $\delta_1, \delta_2, \delta_3, \varepsilon > 0$ such that $\delta_1 \leq c \cdot \delta_2$ and $\delta_3 \leq c \cdot \eta(w^*)$,* Alg *takes as input $m(\varepsilon, \delta_1, \delta_2, \delta_3) =$ poly$(n, \frac{1}{\varepsilon})$ samples drawn i.i.d. from $\widetilde{P} \in \mathsf{Ball}_{W(\ell_q)}(P^*, \delta_1) \cap \mathsf{Ball}_{d_{\mathsf{TV}}}(P^*, \delta_3)$, runs in time polynomial in $m(\varepsilon, \delta_1, \delta_2, \delta_3)$, and outputs with probability at least $2/3$, a hypothesis w such that*

$$\textit{robust-loss}(L, P^*, w) \leq (1+\alpha) \cdot \textit{robust-loss}(L, P^*, w^*) + \varepsilon.$$

Notice that the condition on $\delta_3 \leq c \cdot \eta(w^*)$ is unavoidable even without test-time considerations Awasthi et al. (2014).

Proof We follow exactly the same proof outline as that of Theorem 9. As before the key is to argue that with probability at least $1 - \frac{1}{3(k+k^2)}$, in round k of the Algorithm in Figure 2 we have that

$err(w_k|S_{k-1}) \leq \frac{1}{8}$. For round k define P_k be the conditional distribution of examples in the region S_{k-1} and labeled according to sgn$(w^* \cdot x)$, and define $\widetilde{P}_k$ to be the distribution with the true noisy labels. As before we first argue that the true classifier w^* has small hinge loss under P_k. To see this notice that

$$\mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(w^*) = \frac{\mathbb{P}[|w^* \cdot x| \leq \tau_k]}{\mathbb{P}[|w_{k-1} \cdot x| \leq b_{k-1}]} \tag{36}$$

$$\leq \frac{1}{64} \tag{37}$$

for an appropriate choice of the constants c_1 and c_3.

Next we bound how the hinge loss differs when looking at the true noise distribution $\widetilde{P}_k$ as compared to P_k. In particular we will show that for any $w \in B(w_{k-1}, r_k)$ and $\|w\|_{q^*} \leq \kappa$ we have that

$$\big| \mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(w) - \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} \ell_{\tau_k}(w) \big| \leq O\big(\sqrt{\frac{\eta(w^*) + \delta_3}{\varepsilon}}\big)\frac{z_k}{\tau_k},$$

where $z_k = \sqrt{r_k^2 + b_{k-1}^2}$. To establish this we follow the analysis in Lemma 3.8 of Awasthi et al. (2014) and get

$$\big| \mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(w) - \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} \ell_{\tau_k}(w) \big| = \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} 1_{(x,y)\in N} \big| \ell_{\tau_k}(w, x, y) - \ell_{\tau_k}(w, x, -y) \big| \tag{38}$$

$$\leq 2 \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} 1_{(x,y)\in N} \frac{|w \cdot x|}{\tau_k} \tag{39}$$

$$\leq \frac{2}{\tau_k} \sqrt{\mathop{\mathbb{P}}_{(x,y)\sim \widetilde{P}_k}(N)} \sqrt{\mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} |w \cdot x|^2}. \tag{40}$$

Here N is the set of noisy examples, i.e., the examples where the label y does not match sgn$(w^* \cdot x)$. The lemma follows from noticing that

$$\mathop{\mathbb{P}}_{(x,y)\sim \widetilde{P}_k}(N) = \frac{\mathbb{P}_{(x,y)\sim \widetilde{P}}(N)}{\mathbb{P}(S_{k-1})}$$

$$= O\big(\frac{\eta(w^*) + \delta_3}{\varepsilon}\big),$$

and that

$$\mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} |w \cdot x|^2 \leq 2 \mathop{\mathbb{E}}_{(x,y)\sim P_k} |w \cdot x|^2 + 2\kappa^2\delta^2$$

$$= O(z_k^2)$$

The last inequality follows from Lemma 3.4 of Awasthi et al. (2014) who showed that $\mathbb{E}_{(x,y)\sim P_k} |w \cdot x|^2 = O(r_k^2 + b_{k-1}^2)$ and the fact that the outlier removal procedure in Figure 3 ensures that the variance in sparse directions is preserved. Combining everything we have

$$
\begin{aligned}
err(w_k|S_{k-1}) &= err(v_k|S_{k-1}) \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(v_k) \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} \ell_{\tau_k}(v_k) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} \\
&\leq \ell(v_k, W) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \quad \text{(from Lemma 21)} \\
&\leq \ell(w^*, W) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim \widetilde{P}_k} \ell_{\tau_k}(w^*) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \quad \text{(Lemma 21)} \\
&\leq \mathop{\mathbb{E}}_{(x,y)\sim P_k} \ell_{\tau_k}(w^*) + O(\sqrt{\frac{\eta(w^*)}{\varepsilon}})\frac{z_k}{\tau_k} + \frac{1}{64} \quad \text{(Lemma 21)} \\
&\leq \frac{1}{8}.
\end{aligned}
$$

■

Appendix D. Linear Regression: Proof of Theorem 11

Setup: Recall that we observe pairs $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ where $(x, y) \sim \widetilde{P}$ and $d_{W_\infty(\ell_q)}(P^*, \widetilde{P}) \leq \delta_1$. We further assume that P^* has $X_0 \sim N(0, I)$ and $y_0 = \langle X_0,\, w_0\rangle + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$.
Robust estimator: We compute the estimate,

$$
\begin{aligned}
\widehat{w} = \operatorname*{argmin}_{w\in\mathbb{R}^n} & \frac{1}{m}\sum_{i=1}^{m}(y_i - w\cdot x_i)^2 + 2(\delta_1+\delta_2)\|(1,w)\|_{q^*}\frac{1}{m}\sum_{i=1}^{m}|y_i - w\cdot x_i| \\
&+ (\delta_1+\delta_2)^2\|(1,w)\|_{q^*}^2.
\end{aligned}
$$

We denote by $\widetilde{w}$ the corresponding population minimizer, i.e.

$$
\begin{aligned}
\widetilde{w} = \operatorname*{argmin}_{w\in\mathbb{R}^n} & \mathbb{E}_{(x,y)\sim\widetilde{P}}(y - w\cdot x)^2 + 2(\delta_1+\delta_2)\|(1,w)\|_{q^*}\mathbb{E}_{(x,y)\sim\widetilde{P}}|y - w\cdot x| \\
&+ (\delta_1+\delta_2)^2\|(1,w)\|_{q^*}^2.
\end{aligned}
$$

For convenience we define a robust loss,

$$
\mathcal{L}(w) = \mathbb{E}_{(x,y)\sim\widetilde{P}}(y - \langle x,\, w\rangle)^2 + (\delta_1+\delta_2)^2\|(1,w)\|_{q^*}^2 + 2(\delta_1+\delta_2)\|(1,w)\|_{q^*}\mathbb{E}_{(x,y)\sim\widetilde{P}}|y - \langle x,\, w\rangle|,
$$

and by $\mathcal{L}_n(w)$ we denote its empirical counterpart. We also note, the following elementary fact, that the robust loss at $\widetilde{P}$ is always at least as large as the loss suffered at P^*, i.e.

$$
\mathcal{L}(w) \geq \mathbb{E}_{P^*}(y_0 - \langle x_0,\, w\rangle)^2.
$$

We first restate the following result characterizing the robust test loss in linear regression.

Lemma 24 *The robust loss of any $w \in H$ is given by,*

$$\text{robust-loss}(L, P^*, w, \delta_2) = \mathbb{E}_{(x,y)\sim P^*}(y - w \cdot x)^2 + 2\delta_2 \|(w,1)\|_{q^*} \mathbb{E}_{(x,y)\sim P^*}|y - w \cdot x| \quad (41)$$
$$+ \delta_2^2 \|(w,1)\|_{q^*}^2. \quad (42)$$

As a technical preliminary, we collect bounds on several useful intermediate quantities in the following lemma.

Lemma 25 *Suppose that $m \gtrsim n$, then with probability at least $1 - \eta$ all of the following results hold,*

1. *Let $\widehat{\Sigma}_0 = X_0^T X_0/m$, then*
$$\|\widehat{\Sigma}_0 - I\|_{op} \lesssim \sqrt{\frac{n}{m}} + \sqrt{\frac{\log(1/\eta)}{m}}.$$
2. *Let $\varepsilon \in \mathbb{R}^m$ denote the vector $(\varepsilon_1, \dots, \varepsilon_m)^T$. Then we have that,*
$$\frac{1}{m}\|X_0^T \varepsilon\|_2 \lesssim \sqrt{\frac{n \log(1/\eta)}{m}}.$$
3. $\left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i^2 - \sigma^2\right| \lesssim \sqrt{\log(1/\eta)/m}.$
4. $$\|\widehat{w} - w_0\|_2 \lesssim \max\left\{\sqrt{\frac{n \log(1/\eta)}{m}}, \sigma + \delta(1 + \|w_0\|_{q^*})\right\},$$
$$\|w^* - w_0\|_2 \lesssim \sigma + \delta(1 + \|w_0\|_{q^*}).$$
5. $$\|\widehat{w}\|_{q^*} \lesssim 1 + \|w_0\|_{q^*} + \frac{1}{\delta}\left[\sqrt{\frac{\log(1/\eta)}{m}} + \sigma\right],$$
$$\|w^*\|_{q^*} \lesssim 1 + \|w_0\|_{q^*} + \frac{\sigma}{\delta}.$$
6. *Define, $B := \sigma + \|w - w_0\|^2 + (\delta_1 + \delta_2)\|(1,w)\|_{q^*}$.*
$$\sup_{\|w - w_0\|_2 \le R_{q^*}, \|w\|_{q^*} \le R_2} [\mathcal{L}_m(w) - \mathcal{L}(w)] \le \textit{poly-log}(n, m, \|w_0\|_2)\left[\sqrt{\frac{B^2 \log(1/\eta)}{m}} + \frac{B^2 \log(1/\eta)}{m}\right].$$

As a consequence of this lemma and some elementary calculations we obtain the following theorem.

Theorem 26 *Define $B := \sigma + \|w - w_0\|^2 + (\delta_1 + \delta_2)\|(1,w)\|_{q^*}$, then with probability at least $1 - \eta$,*

$$\mathcal{L}(\widehat{w}) - \mathcal{L}(w^*) \lesssim \textit{poly-log}(n, m, \|w_0\|_2)\left[\sqrt{\frac{B^2 \log(1/\eta)}{m}} + \frac{B^2 \log(1/\eta)}{m}\right].$$

To complete the proof of the claimed result we need to relate the robust loss of w^* to the robust loss of $\widetilde{w}$. By definition, we know that, $\mathcal{L}(\widetilde{w}) \le \mathcal{L}(w^*)$.

Suppose $\eta(\delta) \coloneqq \min_{w\in\mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta)$. We first observe that from (41), we have

$$\forall w \in \mathbb{R}^n, \quad \frac{\textit{robust-loss}(L, P^*, w, 2\delta_1+\delta_2)}{\textit{robust-loss}(L, P^*, w, \delta_2)} \le \max\left\{\Big(\frac{2\delta_1+\delta_2}{\delta_2}\Big)^2, \frac{2\delta_1+\delta_2}{\delta_2}, 1\right\}$$

$$\begin{aligned}\text{Hence,}\quad \frac{\eta(2\delta_1+\delta_2)}{\eta(\delta_2)} &= \frac{\min_{w\in\mathbb{R}^n} \textit{robust-loss}(L, P^*, w, 2\delta_1+\delta_2)}{\min_{w\in\mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta_2)} \\ &\le \max\left\{\Big(\frac{2\delta_1+\delta_2}{\delta_2}\Big)^2, \frac{2\delta_1+\delta_2}{\delta_2}, 1\right\} \\ &\le (2c+1)^2,\end{aligned}$$

since $\delta_1 \le c\delta_2$. Now, we combine this with Claim 3 to conclude that

$$\begin{aligned}\textit{robust-loss}(L, P^*, \widetilde{w}, \delta_2) \le \eta(2\delta_1+\delta_2) &= \frac{\eta(2\delta_1+\delta_2)}{\eta(\delta_2)} \cdot \min_{w\in\mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta_2) \\ &\le (2c+1)^2 \cdot \min_{w\in\mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta_2). \qquad (43)\end{aligned}$$

Finally, as a consequence of this fact, and Theorem 27 we obtain the desired result.

D.1. Proof of Lemma 25

Claims (1), (2) and (3) are straightforward to show. Particularly, (1) follows from Theorem 6.1 of Wainwright (2019), while (2) and (3) follow from standard tail bounds for χ^2 random variables. We prove claims (4) and (5) for $\widehat{w}$, noting that similar reasoning applies to prove the corresponding claims for $\widetilde{w}$, replacing the empirical loss $\mathcal{L}_m$ with the population loss $\mathcal{L}$. will aim to show that if $\widehat{w}$ is far from w_0 then it cannot minimize the empirical loss. More formally, we have the following lower bounds on $\mathcal{L}_m(\widehat{w})$.

$$\begin{aligned}\mathcal{L}_m(\widehat{w}) &\ge \max\left\{\frac{1}{m}\sum_{i=1}^{n}(\langle X_{0i},\, w_0 - \widehat{w}\rangle + \varepsilon_i)^2, \delta^2(1+\|\widehat{w}\|_{q^*})^2\right\} \\ &\gtrsim \max\left\{\|\widehat{w}-w_0\|_2^2 - C_1\sqrt{\frac{n\log(1/\eta)}{m}}\|\widehat{w}-w_0\|_2, \delta^2(1+\|\widehat{w}\|_{q^*})^2\right\}\end{aligned}$$

for some $C_1 > 0$. We similarly note that,

$$\begin{aligned}\mathcal{L}_m(w_0) &\le \frac{2}{m}\sum_{i=1}^{m}(\varepsilon_i + \delta(1+\|w_0\|_{q^*}))^2 + 2\delta^2(1+\|w_0\|_{q^*}^2) \\ &\lesssim \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i^2 + \delta^2(1+\|w_0\|_{q^*}^2) \\ &\lesssim \sigma^2 + \delta^2(1+\|w_0\|_{q^*}^2) + \frac{\log(1/\eta)}{m}.\end{aligned}$$

Putting these bounds together, using the fact that $\mathcal{L}_m(\widehat{w}) \le \mathcal{L}_m(w_0)$, we obtain that,

$$\|\widehat{w}\|_{q^*} \lesssim 1 + \|w_0\|_{q^*} + \frac{1}{\delta}\left[\sqrt{\frac{\log(1/\eta)}{m}} + \sigma\right],$$

and that,

$$\|\widehat{w} - w_0\|_2 \lesssim \max\left\{\sqrt{\frac{n\log(1/\eta)}{m}}, \sigma + \delta(1 + \|w_0\|_{q^*})\right\}.$$

This yields claims (4) and (5). Now, we turn our attention to the final claim of the Lemma. We first fix a w in the ball $\|w - w_0\|_2 \leq R$ and upper bound $|\mathcal{L}_m(w) - \mathcal{L}(w)|$, the result will then follow from a discretization argument and the union bound. For a fixed w, we have that,

$$\begin{aligned}|\mathcal{L}_m(w) - \mathcal{L}(w)| \leq \left|\frac{1}{m}\sum_{i=1}^{m}(y_i - \langle X_i, w\rangle)^2 - \mathbb{E}(y - \langle X, w\rangle)^2\right| + \\ + 2\delta(1 + \|w\|_{q^*})\left|\frac{1}{m}\sum_{i=1}^{m}|y_i - \langle X_i, w\rangle| - \mathbb{E}|y - \langle X, w\rangle|\right|\end{aligned}$$

First suppose we define $Z := (y - \langle X, w\rangle)^2$ and $\widetilde{Z} := |y - \langle X, w\rangle|$, then we can write $Z := (Z_0 + Z_1)^2$ and $\widetilde{Z} := |Z_0 + Z_1|$ where $Z_0 := y_0 - \langle X_0, w\rangle \sim N(0, \|w_0 - w\|_2^2 + \sigma^2)$ and $Z_1 := Z_y + \langle Z_x, w\rangle$, and $|Z_1| \leq \delta + \delta\|w\|_1$. It is straighforward to verify that $Z_0 + Z_1$ and $\widetilde{Z}$ are sub-Gaussian random variables. Particularly, recalling the definitions of sub-Gaussian random variables and their corresponding Orlicz norm from Chapter 2 of Vershynin (2018), we have that $\|Z_0 + Z_1\|_{\psi_2} = \|\widetilde{Z}\|_{\psi_2}$ and futher that,

$$\|\widetilde{Z}\|_{\psi_2} = \|y - \langle X, w\rangle\|_{\psi_2} = \|Z_0 + Z_1\|_{\psi_2} \leq \|Z_0\|_{\psi_2} + \|Z_1\|_{\psi_2} \lesssim \sigma + \|w - w_0\|_2 + \delta + \delta\|w\|_{q^*}.$$

As a straightforward consequence, we obtain that Z is sub-exponential, i.e.

$$\|Z\|_{\psi_1} \lesssim \sigma^2 + \|w - w_0\|_2^2 + \delta^2 + \delta^2\|w\|_{q^*}^2.$$

Let us denote by $B := \sigma + \|w - w_0\|^2 + \delta(1 + \|w\|_{q^*})$. Now, as a straightforward consequence of Hoeffding's inequality for sub-Gaussians (Theorem 2.6.3 in Vershynin (2018)) and Bernstein's inequality for sub-exponentials (Theorem 2.8.1 in Vershynin (2018)) we obtain that with probability at least $1 - \eta$, for any fixed w,

$$|\mathcal{L}_m(w) - \mathcal{L}(w)| \lesssim \sqrt{\frac{B^2\log(1/\eta)}{m}} + \frac{B^2\log(1/\eta)}{m}.$$

Now, let us derive a bound on, $|\mathcal{L}(w + \Delta) - \mathcal{L}(w)|$, and note that similar arguments yield bounds on $|\mathcal{L}_n(w + \Delta) - \mathcal{L}_m(w)|$ and consequently on $|\mathcal{L}(w + \Delta) - \mathcal{L}_m(w + \Delta)| - |\mathcal{L}(w) - \mathcal{L}_m(w)|$. Let $\Sigma = \mathbb{E}xx^T$ and $\theta = \mathbb{E}xy$ then a straightforward but tedious calculation yields that,

$$|\mathcal{L}(w + \Delta) - \mathcal{L}(w)| \leq \text{poly}(n, \delta, \|\Sigma\|_{op}, \|\theta\|_2, \|w\|_{q^*})\max\{\|\Delta\|_2^2, \|\Delta\|_2\}.$$

Observing further that, $\|\Sigma\|_{op}, \|\theta\|_2$ are both bounded by polynomial functions of n and δ we have that, $|\mathcal{L}(w+\Delta) - \mathcal{L}(w)| \leq \text{poly}(n, \delta, \|w\|_{q^*})\max\{\|\Delta\|_2^2, \|\Delta\|_2\} \leq \text{poly}(n, \delta, R_2)\max\{\|\Delta\|_2^2, \|\Delta\|_2\}$. Now, this in turn implies that if we can show that over an ε-net of w vectors, with sufficiently small ε (inverse polynomial in n, R_2, δ and m) that $\mathcal{L}_m(w) - \mathcal{L}(w)$ is small, then we may conclude it is uniformly small over all w by leveraging the fact that $\mathcal{L}$ and $\mathcal{L}_m$ does not change much between net points. This net has cardinality at most $\mathcal{O}(R_1/\text{poly}(n, \delta, R_2, m))^n$ and we finally obtain the claimed result via a union bound over this ε-net.

D.2. Proof of Theorem 26

We condition on the event that all claims in Lemma 1 hold, and on this event we have that,

$$\begin{aligned}\mathcal{L}(\widehat{w}) &= \mathcal{L}(\widehat{w}) - \mathcal{L}_m(\widehat{w}) + \mathcal{L}_m(\widehat{w}) \\ &= \mathcal{L}(\widehat{w}) - \mathcal{L}_m(\widehat{w}) + \mathcal{L}_m(w^*) \\ &= \mathcal{L}(w^*) + \mathcal{L}(\widehat{w}) - \mathcal{L}_m(\widehat{w}) + \mathcal{L}_m(w^*) - \mathcal{L}(w^*) \\ &\leq \mathcal{L}(w^*) + \sup_{\|w-w_0\|_2 \leq R_1, \|w\|_{q^*} \leq R_2} [\mathcal{L}_m(w) - \mathcal{L}(w)] + \mathcal{L}_m(w^*) - \mathcal{L}(w^*),\end{aligned}$$

and the claim of the Theorem then follows from the final claim of Lemma 25.

Appendix E. Mean Estimation: Proof of Theorem 13

Setup: We briefly recall the setup. We observe samples, $X_1, \ldots, X_m \sim \widetilde{P}$ where $\widetilde{P} \in \mathsf{Ball}_{W(\ell_\infty)}(P^*, \delta_1)$. Here P^* has (finite) mean μ^* and coordinate-wise variance at most σ^2.
Robust Estimator: We compute the estimate,

$$\widehat{w} = \operatorname*{argmin}_{w \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^m \|x_i - w\|_2^2 + \frac{2(\delta_1 + \delta_2)}{m} \sum_{i=1}^m \|x_i - w\|_1.$$

Let us denote by $\widetilde{w}$ the corresponding population minimizer, i.e.

$$\widetilde{w} = \operatorname*{argmin}_{w \in \mathbb{R}^n} \mathbb{E}_{X \sim \widetilde{P}} \|X - w\|_2^2 + 2(\delta_1 + \delta_2)\mathbb{E}_{X \sim \widetilde{P}} \|X - w\|_1,$$

and let us denote by $\widetilde{\mu}$ the mean of $\widetilde{P}$. For convenience, we denote,

$$\mathcal{L}(w) = \mathbb{E}_{X \sim \widetilde{P}} \|X - w\|_2^2 + 2(\delta_1 + \delta_2)\mathbb{E}_{X \sim \widetilde{P}} \|X - w\|_1,$$

and note that,

$$\begin{aligned}\mathcal{L}(w_1) - \mathcal{L}(w_2) &= \mathbb{E}_{X \sim \widetilde{P}} \|X - w_1\|_2^2 + 2(\delta_1 + \delta_2)\mathbb{E}_{X \sim \widetilde{P}} \|X - w_1\|_1 - \mathbb{E}_{X \sim \widetilde{P}} \|X - w_2\|_2^2 \\ &\quad - 2(\delta_1 + \delta_2)\mathbb{E}_{X \sim \widetilde{P}} \|X - w_2\|_1 \\ &\leq \|w_1 - w_2\|_2^2 - 2\langle \mu^*, w_1 - w_2 \rangle + 2(\delta_1 + \delta_2)\|w_1 - w_2\|_1 \\ &\leq \|w_1 - w_2\|_2^2 + 2\|\mu^*\|\|w_1 - w_2\|_2 + 2(\delta_1 + \delta_2)\|w_1 - w_2\|_1. \qquad (44)\end{aligned}$$

We now analyze the finite-sample performance of the estimate $\widehat{w}$.

Theorem 27 *Suppose that P^* has (finite) mean μ^* and coordinate-wise variance at most σ^2 then we show the following pair of results:*

$$\mathbb{E}\|\widehat{w} - \widetilde{w}\|_2^2 \leq \mathcal{O}\left(\frac{\sigma^2 n}{m} + \frac{(\delta_1 + \delta_2)^2 n}{m}\right), \qquad (45)$$

and furthermore,

$$\mathbb{E}\left[\mathcal{L}(\widehat{w}) - \mathcal{L}(\widetilde{w})\right] \leq \mathcal{O}\left(\frac{\sigma^2 n}{m} + \|\mu^*\|_2 \cdot \sigma\sqrt{\frac{n}{m}} + (\delta_1 + \delta_2)\sigma\sqrt{\frac{n^2}{m}}\right). \qquad (46)$$

Proof We begin by proving the first claim. We analyze each coordinate separately, analyzing first the j-th coordinate. First, we denote the subgradient of the ℓ_1 norm by,

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ [-1, 1] & \text{if } x = 0. \end{cases}$$

We define further, $\overline{X}_m$ to be sample average, and $\mathbb{P}_m$ to be the empirical measure of the samples. From the KKT conditions for the optimization problems defining $\widehat{w}$ and $\widetilde{w}$ we obtain that,

$$\begin{aligned} 0 &\in 2(\widetilde{w}_j - \widetilde{\mu}_j) - 2(\delta_1 + \delta_2)\mathbb{E}[\text{sign}(X_j - \widetilde{w}_j)], \\ 0 &\in 2(\widehat{w}_j - \overline{X}_{mj}) - 2(\delta_1 + \delta_2)\mathbb{P}_m[\text{sign}(X_j - \widehat{w}_j)]. \end{aligned}$$

Let us first suppose that $\widetilde{w}_j > \widehat{w}_j$ and upper bound $\widetilde{w}_j - \widehat{w}_j$. An analogous bound holds in the case when $\widetilde{w}_j < \widehat{w}_j$. We obtain,

$$\begin{aligned} \widetilde{w}_j &\leq \widetilde{\mu}_j + (\delta_1 + \delta_2)\left[\mathbb{P}(X_j \geq \widetilde{w}_j) - \mathbb{P}(X_j < \widetilde{w}_j)\right] \\ &\leq \widetilde{\mu}_j + (\delta_1 + \delta_2)\left[\mathbb{P}(X_j > \widehat{w}_j) - \mathbb{P}(X_j \leq \widehat{w}_j)\right] \quad \text{and} \\ \widehat{w}_j &\geq \overline{X}_{mj} + (\delta_1 + \delta_2)\left[\mathbb{P}_m(X_j > \widehat{w}_j) - \mathbb{P}_m(X_j \leq \widehat{w}_j)\right]. \end{aligned}$$

From this we see that,

$$\widetilde{w}_j - \widehat{w}_j \leq \widetilde{\mu}_j - \overline{X}_{mj} + (\delta_1 + \delta_2)\left[\mathbb{P}(X_j > \widehat{w}_j) - \mathbb{P}_m(X_j > \widehat{w}_j) - \mathbb{P}(X_j \leq \widehat{w}_j) + \mathbb{P}_m(X_j \leq \widehat{w}_j)\right].$$

By the Dvoretzky-Kiefer-Wolfowitz-Massart inequality we know that with probability at least $1 - \eta$,

$$\sup_w |\mathbb{P}(X_j > w) - \mathbb{P}_m(X_j > w)| \leq \mathcal{O}\left(\sqrt{\frac{\log(1/\eta)}{m}}\right),$$

and integrating this bound we obtain a bound in expectation. Using the fact that P^* has variance at most σ^2, we obtain that $\widetilde{P}$ has variance at most,

$$\begin{aligned} \widetilde{\sigma}_j^2 = \mathbb{E}_{\widetilde{P}}\left[(X_j - \widetilde{\mu}_j)^2\right] &\leq \mathbb{E}_{P^*} \sup_{|Z| \leq \delta_1} \left[(X_j + Z - \widetilde{\mu}_j)^2\right] \\ &\leq \mathcal{O}(\sigma_j^2 + \delta_1^2 + (\widetilde{\mu}_j - \mu_j^*)^2) \\ &\leq \mathcal{O}(\sigma_j^2 + \delta_1^2). \end{aligned}$$

Putting all of these together we obtain that,

$$\mathbb{E}\left[|\widetilde{w}_j - \widehat{w}_j|\right] \leq \mathcal{O}((\sigma_j + (\delta_1 + \delta_2))/\sqrt{m}),$$

and that,

$$\mathbb{E}\left[(\widetilde{w}_j - \widehat{w}_j)^2\right] \leq \mathcal{O}((\sigma_j^2 + (\delta_1 + \delta_2)^2)/m).$$

The claim on the loss of $\widehat{w}$ then follows from the argument in (44). ■

The following result relates the robust loss of w^* to the robust loss of $\widetilde{w}$. By definition, we know that, $\mathcal{L}(\widetilde{w}) \leq \mathcal{L}(w^*)$. Recall that for any $w \in \mathbb{R}^n$,

$$\begin{aligned} \textit{robust-loss}(L, P^*, w, \delta) &= \sup_{P \in \mathsf{Ball}_{W(\ell_\infty)}(P^*, \delta)} \mathbb{E}_P \|X - w\|_2^2] \\ &= \mathbb{E}_{x \sim P}\|x - w\|_2^2 + 2\delta \mathbb{E}_{x \sim \widetilde{P}} \|x - w\|_1 + \delta^2 n. \end{aligned} \tag{47}$$

Suppose $\eta(\delta) := \min_{w \in \mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta)$. We first observe that from (47), we have

$$\begin{aligned} \forall w \in \mathbb{R}^n, \quad \frac{\textit{robust-loss}(L, P^*, w, 2\delta_1 + \delta_2)}{\textit{robust-loss}(L, P^*, w, \delta_2)} &\leq \max\left\{ (\frac{2\delta_1 + \delta_2}{\delta_2})^2, \frac{2\delta_1 + \delta_2}{\delta_2}, 1 \right\} \\ \text{Hence,} \quad \frac{\eta(2\delta_1 + \delta_2)}{\eta(\delta_2)} &\leq \max\left\{ \left(\frac{2\delta_1 + \delta_2}{\delta_2}\right)^2, \frac{2\delta_1 + \delta_2}{\delta_2}, 1 \right\} \\ &\leq (2c + 1)^2, \end{aligned}$$

since $\delta_1 \leq c\delta_2$. Now, we combine this with Claim 3 to conclude that

$$\begin{aligned} \textit{robust-loss}(L, P^*, \widetilde{w}, \delta_2) \leq \eta(2\delta_1 + \delta_2) &= \frac{\eta(2\delta_1 + \delta_2)}{\eta(\delta_2)} \cdot \min_{w \in \mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta_2) \\ &\leq (2c + 1)^2 \cdot \min_{w \in \mathbb{R}^n} \textit{robust-loss}(L, P^*, w, \delta_2). \end{aligned} \tag{48}$$

Finally, as a consequence of this fact, and Theorem 27 we obtain the desired result.

Appendix F. Alternate Definition and Discussion: Proof of Theorem 15

We first recall (and restate) the two definitions of train-and-test robustness. The first definition is the one we propose and study in this work.

Definition 28 **($(\delta_1, \delta_2, \alpha)$-TRAIN-AND-TEST ROBUSTNESS)** *Suppose $\delta_1, \delta_2, \alpha > 0$, and $\mathcal{M} = \mathcal{M}(\mathcal{Z})$ be the space of distributions over the space $\mathcal{Z}$, and let H be a hypothesis class equipped with the loss function L. For a distribution $P^* \in \mathcal{M}$, an algorithm Alg is $(\delta_1, \delta_2, 1 + \alpha)$-TRAIN-AND-TEST ROBUST w.r.t P^*, H i.f.f. for any $\varepsilon > 0$, given $m(\varepsilon, \delta_1, \delta_2)$ i.i.d. samples S drawn from a distribution $\widetilde{P}$ such that $d_{\mathsf{train}}(\widetilde{P}, P^*) \leq \delta_1$, Alg outputs a function $\widehat{h} = \mathsf{Alg}(S)$ such that $\textit{robust-loss}(L, P^*, \widehat{h}, \delta_2) \leq (1 + \alpha)\min_{h \in H} \textit{robust-loss}(L, P^*, h, \delta_2) + \varepsilon$, i.e.,*

$$\mathbb{P}_{S \sim \widetilde{P}}\left[\sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(\widehat{h}) \leq (1 + \alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h) + \varepsilon \right] \geq \frac{2}{3}, \quad \textit{where } \widehat{h} = \mathsf{Alg}(S). \tag{49}$$

Given a family $\mathcal{P}$ of distributions, we say that Alg is train-and-test robust w.r.t. $(\mathcal{P}, H, d)$ if for any $P^* \in \mathcal{P}$, Alg is TRAIN-AND-TEST *robust w.r.t. P^*, H (note that neither $\widetilde{P}$ nor the test distribution P needs to be restricted to the family $\mathcal{P}$).*

The second definition is based on a weaker benchmark, and hence we call it as a "weak oracle" variant.

Definition 29 (WEAK ORACLE TRAIN-AND-TEST ROBUSTNESS) *Suppose $\delta_1, \delta_2, \alpha > 0$, and $\mathcal{M} = \mathcal{M}(\mathcal{Z})$ be the space of distributions over the space $\mathcal{Z}$, and let H be a hypothesis class equipped with the loss function L. An algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$-TRAIN-AND-TEST ROBUST w.r.t $\mathcal{P}, H$ i.f.f. for any $\varepsilon > 0$, given $m(\varepsilon, \delta_1, \delta_2)$ i.i.d. samples S drawn from $\widetilde{P}$, Alg outputs a function $\widehat{h} = \mathsf{Alg}(S)$,*

$$\mathop{\mathbb{P}}_{S \sim \widetilde{P}} \left[\sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(\widehat{h}) \leq (1+\alpha) \min_{h \in H} \sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(h) + \varepsilon \right] \geq \frac{2}{3}, \quad \text{where } \widehat{h} = \mathsf{Alg}(S). \tag{50}$$

Here $\mathcal{D}_{\widetilde{P}}$ is defined as,

$$\mathcal{D}_{\widetilde{P}} = \{P_2 : \exists P_1, d_{\mathsf{train}}(\widetilde{P}, P_1) \leq \delta_1, \text{ and, } d_{\mathsf{test}}(P_1, P_2) \leq \delta_2\}. \tag{51}$$

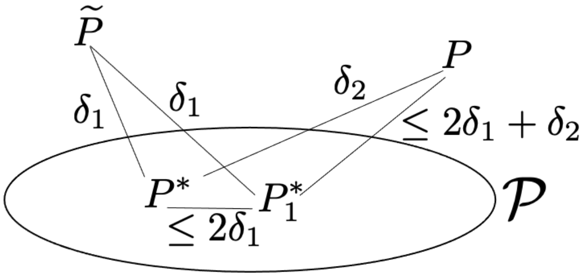

Figure 4: An illustration of the structured family $\mathcal{P}$, the train distribution $\widetilde{P}$, and the test distribution P.

The connection among the two definitions claimed in Section 6 will follow from the following two Lemmas.

Lemma 30 *If an algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust w.r.t. $\mathcal{P}$ according to Definition 28, then it is also $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 29.*

Proof Fix a particular corrupted distribution $\widetilde{P}$ as shown in Figure 4. Suppose Alg is $(\delta_1, \delta_2, \alpha)$ train-and-test robust w.r.t. $\mathcal{P}$ according to Definition 28. Note that Alg is given samples from $\widetilde{P}$ but does not know the uncorrupted distribution P^*. Then given $S \sim \widetilde{P}$, it holds with probability at least $2/3$, that $\forall P^* \in \mathcal{P}$ s.t. $d_{\mathsf{train}}(P^*, \widetilde{P}) \leq \delta_1$,

$$\sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(\widehat{h}) \leq (1+\alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h) + \varepsilon. \tag{52}$$

From the above we get that with probability at least $2/3$,

$$\begin{aligned} \sup_{P^* : d_{\mathsf{train}}(P^*, \widetilde{P}) \leq \delta_1} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(\widehat{h}) &\leq (1+\alpha) \sup_{P^* : d_{\mathsf{train}}(P^*, \widetilde{P}) \leq \delta_1} \min_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h) + \varepsilon \\ &\leq (1+\alpha) \min_{h \in H} \sup_{P^* : d_{\mathsf{train}}(P^*, \widetilde{P}) \leq \delta_1} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h). \end{aligned} \tag{53}$$

Finally, notice that from the definition of $\mathcal{D}_{\widetilde{P}}$, for any $h \in H$ we have

$$\sup_{P^*: d_{\mathsf{train}}(P^*, \widetilde{P}) \le \delta_1} \ \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(h) = \sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(h). \tag{54}$$

As a result we get that with probability at least $2/3$,

$$\sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(\widehat{h}) \le (1+\alpha) \min_{h \in H} \sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(h). \tag{55}$$

■

The above lemma shows that our proposed definition of train-and-test robustness (Definition 28) is a provably stronger notion than the alternate definition of robustness (Definition 29). However, one can also obtain implications of the stronger definition provided an algorithm that competes with the weak oracle benchmark. To see this we first relax the definition of $(\delta_1, \delta_2, 1+\alpha)$-train-and-test-robust under Definition 28 to allow for a slack factor $\gamma \ge 1$.

Definition 31 **($(\delta_1, \delta_2, \alpha)$-TRAIN-AND-TEST ROBUSTNESS with slack)** *Suppose $\delta_1, \delta_2, \alpha > 0$, and $\mathcal{M} = \mathcal{M}(\mathcal{Z})$ be the space of distributions over the space $\mathcal{Z}$, and let H be a hypothesis class equipped with the loss function L. For a distribution $P^* \in \mathcal{M}$, an algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$-TRAIN-AND-TEST ROBUST with slack γ, w.r.t P^*, H i.f.f. for any $\varepsilon > 0$, given $m(\varepsilon, \delta_1, \delta_2)$ i.i.d. samples S drawn from a distribution $\widetilde{P}$ such that $d_{\mathsf{train}}(\widetilde{P}, P^*) \le \delta_1$, Alg outputs a function $\widehat{h} = \mathsf{Alg}(S)$ such that robust-loss$(L, P^*, \widehat{h}, \delta_2) \le (1+\alpha)\min_{h \in H}$ robust-loss$(L, P^*, h, \delta_2) + \varepsilon$, i.e.,*

$$\mathop{\mathbb{P}}_{S \sim \widetilde{P}}\left[\sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \delta_2)} L_P(\widehat{h}) \le (1+\alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_{d_{\mathsf{test}}}(P^*, \gamma \cdot \delta_2)} L_P(h) + \varepsilon\right] \ge \frac{2}{3}, \quad \textit{where } \widehat{h} = \mathsf{Alg}(S). \tag{56}$$

Given a family $\mathcal{P}$ of distributions, we say that Alg is train-and-test robust w.r.t. $(\mathcal{P}, H, d)$ if for any $P^* \in \mathcal{P}$, Alg is TRAIN-AND-TEST robust w.r.t. P^*, H (note that neither $\widetilde{P}$ nor the test distribution P needs to be restricted to the family $\mathcal{P}$).

Hence, in the definition above the output $\widehat{h}$ of the algorithm is tested on distributions that are δ_2 away from P^*, whereas the benchmark is test on. distributions that are $\gamma\delta_2$ away for a given $\gamma \ge 1$. When. γ equals one, we recover the original definition (Definition 28).

Next we will assume that both the metrics d_{train} and d_{test} are the same, and will use the symbol d to denote them. Then we have the following implication.

Lemma 32 *If an algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 29, then it is also $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust with slack $\gamma = 1 + \frac{2\delta_1}{\delta_2}$, according to Definition 31.*

Proof Fix a particular corrupted distribution $\widetilde{P}$ as shown in Figure 4. Since Alg is $(\delta_1, \delta_2, \alpha)$ train-and-test robust according to Definition 29, given $S \sim \widetilde{P}$, it holds with probability at least $2/3$ that

$$\sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(\widehat{h}) \le (1+\alpha) \min_{h \in H} \sup_{P \in \mathcal{D}_{\widetilde{P}}} L_P(h) + \varepsilon. \tag{57}$$

Next fix any $P_1^* \in \mathcal{P}$ such that $d(P_1^*, \widetilde{P}) \leq \delta_1$. Then we have by triangle inequality (see Figure 4) that

$$\{P : d(P, P_1^*) \leq \delta_2\} \subseteq \mathcal{D}_{\widetilde{P}} \subseteq \{P : d(P, P_1^*) \leq 2\delta_1 + \delta_2\}. \tag{58}$$

Using the above we get that with probability at least $2/3$, $\forall P^* \in \mathcal{P}$ s.t. $d(P^*, \widetilde{P}) \leq \delta_1$,

$$\begin{aligned}\sup_{P \in \mathsf{Ball}_d(P^*, \delta_2)} L_P(\widehat{h}) &\leq (1+\alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_d(P^*, 2\delta_1 + \delta_2)} L_P(h) + \varepsilon \\ &= (1+\alpha) \min_{h \in H} \sup_{P \in \mathsf{Ball}_d\left(P^*, (1+\frac{2\delta_1}{\delta_2})\delta_2\right)} L_P(h) + \varepsilon. \end{aligned} \tag{59}$$

In other words we get that Alg is $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust with slack $\gamma = 1 + \frac{2\delta_1}{\delta_2}$. ■

The above two lemmas can be combined to obtain the following theorem.

Theorem 33 *If an algorithm Alg is $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 28, then it is also $(\delta_1, \delta_2, 1+\alpha)$ train-and-test robust according to Definition 29. Conversely, $(\delta_1, \delta_2, 1+\alpha)$ robustness according to Definition 29 implies $(\delta_1, \delta_2, 1+\alpha)$ robustness according to Def. 28 with slack $\gamma = 1 + \frac{2\delta_1}{\delta_2}$.*