Learning in the Limit with Adversarial Disturbances

Constantine Caramanis* **and Shie Mannor†‡**

Abstract

We study distribution-dependent, data-dependent, learning in the limit with adversarial disturbance. We consider an optimization-based approach to learning binary classifiers from data under worst-case assumptions on the disturbance. The learning process is modeled as a decision-maker who seeks to minimize generalization error, given access only to possibly maliciously corrupted data. Two models for the nature of the disturbance are considered: disturbance in the labels of a certain fraction of the data, and disturbance that also affects the position of the data points. We provide distribution-dependent bounds on the amount of error as a function of the noise level for the two models, and describe the optimal strategy of the decision-maker, as well as the worst-case disturbance.

1 Introduction

Most of the work on learning in the presence of malicious noise has been within the PAC framework, focusing on *a priori*, distribution independent bounds on generalization error and sample complexity. This work has not fully addressed the question of what a decision-maker must do when faced with a particular realization of the data, and perhaps some knowledge of the underlying distribution and the corrupting disturbance. The main contribution of this paper is the development of a robust optimization-based, algorithmic data-dependent, distribution-dependent approach to minimizing error of learning subject to adversarial disturbance.

In the adversarial PAC setup, a decision-maker has access to IID samples from some source, only that a fraction of these points are altered by an adversary. There are several models for the noise which we discuss below. The decision-maker is given $\epsilon > 0$ and $\delta > 0$ and attempts to learn an ϵ-optimal classifier with probability of at least $1 - \delta$. The emphasis in [KL93], as well as in several follow-up works (e.g., [BEK02, ACB98, CBDF$^+$99, Ser03]) is on the sample complexity of learning in such setups and on particularly bad data sources.

The algorithmic issue of the decision-maker's optimal strategy when faced with a certain disturbance level, i.e., a certain amount of possible data corruption, and a realization of the data has not been adequately explored; see [Lai88] for an initial discussion. While there are quite a few possible disturbance models that differ on the precise setup (what the adversary know, what the adversary can do and in which order), we focus on the strongest disturbance model where the adversary has access to the actual distribution and can modify it adversarialy within a constraint on the disturbance level. This "learning in the information limit" model is used to abstract other issues such as finite sample or limited adversary (see [CBDF$^+$99] for a discussion on some relevant models). In this paper we consider two different noise models, with the intention of addressing the algorithmic aspects and the effect of the disturbance level. We note that we use the term disturbance rather than noise because in our model data are corrupted in a possibly adversarial way and the probabilistic aspect is essentially not relevant.

We deviate from the traditional learning setup in three major assumptions. First, we focus on the question of how the decision-maker should minimize error, rather than following PAC-style results of computing *a priori* bounds on that error. Moreover, our analysis is distribution specific and we do not focus on particularly bad data sources. Second, the noise level is not assumed small and the decision-maker has to incur error in all but trivial problems (this has been studied in the malicious noise setup; see [CBDF$^+$99]). Third, we do not ask how many samples are needed to obtain low generalization error, instead we assume that the distribution of the samples is provided to the decision-maker (equivalently, one may think of this as considering the large sample or "information theoretic" limit). However, this distribution is corrupted by potentially persistent noise; we may consider it as first tampered with by an adversary. After observing the modified distribution, the decision-maker has to commit to a single classifier from some predefined set $\mathcal{H}$. The performance of the classifier chosen by the decision-maker is measured on the original, true distribution (this is similar to the agnostic setup of [KSS92]). The question is what should the decision-maker do? And how much error will he incur in the worst case?

In order to answer these questions we adopt a robust

*Department of Electrical and Computer Engineering, The University of Texas at Austin, cmcaram@ece.utexas.edu

†Department of Electrical and Computer Engineering, McGill University, shie.mannor@mcgill.ca

‡This work was partially supported by NSF Grants CNS-0721532, EFRI-0735905, and the Canada Research Chairs Program

optimization-theoretic perspective, where we regard our decision-maker as trying to make optimal decisions while facing an adversary. Our aim is to provide an analysis by identifying optimal strategies, and quantify the error as a function of the adversary's strategy, i.e., the nature of the corrupting disturbance. We refer to the disturbance as selected by an adversary merely as a conceptual device, and not in strict analogy to game theory. In particular, the decision-maker does not assume that the corrupting noise is chosen with any specific aim; rather, the decision-maker selects a strategy to protect himself in the worst-case scenario.

The true probability distribution is defined over the input space and on the labels. We focus on the case of proper learning, where this amounts to a distribution and the true classifier. Then the adversary modifies the distribution of the input points and the labels. The decision-maker observes the modified distribution and chooses a classifier in $\mathcal{H}$ to minimize the *worst-case* error. We note the relationship with [KSS92] who use a slightly different model. In their model, the decision maker chooses a classifier in $\mathcal{H}$ knowing that the true classifier is in some "touchstone" class $T \subseteq \mathcal{H}$. They say that an algorithm facilitates learning (with respect to a loss function) if it learns a function from $\mathcal{H}$ that is close to a function from T in the usual PAC sense (i.e., with high probability and small error after observing a number of samples polynomial in one over the error, and one over the confidence). As opposed to [KSS92] and most subsequent works, we do not focus on small noise and we ignore the sample complexity aspect altogether. Instead, we focus on the policy chosen by the decision maker and on the informational limits. In that respect, our work is most related to [CBDF$^+$99] who considered the case of substantial noise. Their proposed strategy that deals with noise, however, is based on randomizing two strategies or using majority vote (phase 2 of the randomized Algorithm SIH in [CBDF$^+$99]). We propose a more principled approach to handling adversarial noise, leading to improved results.

If the noise level and characteristics are unlimited, the decision-maker cannot hope to do better than randomly guessing. We therefore limit the noise, and allow the adversary to change only a given fraction of the distribution, which we refer to as "the power of the adversary". An alternative view, which is common in robust optimization [BTN99], is to consider the power of the adversary as a *design parameter*. According to this view, the decision-maker tries to be resilient to a specified amount of uncertainty in the parameters of the problem.

The paper is structured as follows. In Section 2 we describe the setup. We define two types of adversaries: one that can only flip a fraction of the points, and one that can also move the points to another location. In Section 3 we consider the optimal solution pairs for the two different set-ups. We characterize the strategy of both the decision-maker and the adversary as a function of the level of noise (the power of the adversary) and the specific distribution that generates the data. Taking such a distribution-dependent perspective allows us to characterize the decision-maker's optimal strategy as the solution to a linear program if the adversary can only flip labels, or a robust optimization problem in the case of the more powerful adversary that can also modify the measure. We further bound the error that may be incurred and show that in the worst case, both adversaries can cause an error twice their power. In Section 4 we show how performance degrades with the increase of this power. A technical proof along with a somewhat surprising worked out example are deferred to the online appendix [CM08].

2 Setup and Definitions

In this section we give the basic definitions of the noisy learning setup. Also, we formulate the optimization problem which characterizes the optimal policy of the decision-maker, and the worst-case noise. The decision-maker, after observing the noisy data, and knowing the power of the adversary, outputs a decision in the classifier space. The disagreement with the true classifier, is the generalization error. The decision-maker's goal is to minimize this, in the worst case. We allow our decision-maker to output a so-called mixed strategy.[1]

Throughout this paper we focus on proper learning. We let $\mathcal{H}$ denote a predefined set of classifiers from which the true classifier is drawn, and from which the decision-maker must choose. Moreover, we assume that $\mathcal{H}$ is finite for the sake of simplicity and to avoid some (involved but straightforward) technicalities. Indeed, there are three natural extensions to our work that we postpone, primarily due to space limitations. First, while we focus on the proper learning setup, the non-proper setup (as in [KSS92]) seems to naturally follow our framework. Second, the case of an infinite set of classifiers $\mathcal{H}$ could be resolved by eliminating classifiers that are "close" according to the observed measure. This is particularly useful for the flip-only setup where the adversary cannot make two classifiers substantially different. Finally, while we do not consider sample complexity, such results should not be too difficult to derive by imitating the arguments in [CBDF$^+$99].

2.1 The Learning Model

In this paper, we deviate from the PAC learning setup, and consider an *a priori* fixed underlying distribution μ, that generates the location (not the labels) of the training data. Thus the error calculations we make are a function of the power of the adversary and also of the fixed probability measure μ. We use the symbol μ throughout this paper, exclusively in reference to the true probability distribution which generates the location (not the label) of the points, and hence, is used to determine the generalization error. Given a particular classifier $\hat{h}$, a true classifier h_{true}, and the underlying probability measure μ, the generalization error is given by the error function

$$\mathcal{E}_\mu(h_{\text{true}}; \hat{h}) \triangleq \mu\{x : h_{\text{true}}(x) \neq \hat{h}(x)\}.$$

We can extend this definition to a probability measure over $\mathcal{H}$, or, in the game-theory terminology, a mixed strategy over $\mathcal{H}$, given by a weighting vector $\alpha = (\alpha_1, \alpha_2, \dots)$ where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$. In that case, denoting the space of mixed strategies by $\Delta_{\mathcal{H}}$, and a particular mixed strategy by

[1] That is, rather than commit to a single classifier, our decision-maker can commit to a randomized strategy, involving possibly multiple classifiers.

$\alpha \in \Delta_{\mathcal{H}}$, we have

$$\mathcal{E}_\mu(h_{\text{true}}; \alpha) \triangleq \sum_i \alpha_i \mathcal{E}_\mu(h_{\text{true}}; h_i).$$

We note that the mixing is often referred to as "probabilistic concepts" or "probabilistic hypotheses" in machine learning. In the context of learning with adversarial noise see [CBDF$^+$99].

2.2 The Noise Model and The Decision-Maker

We next define the possible actions of the adversary and of the decision-maker. As discussed above, in this paper we do not consider sample complexity, and effectively consider the situation where the training sample is infinite in size (the information theoretic limit). We model this situation by assuming that rather than training samples, the decision-maker receives a distribution for each of the two labels. Since the adversary modifies this object in various ways (noise is added to the observations) we make some formal definitions which facilitate discussion of this in the sequel.

Let $\mathcal{X}$ denote the space in which the training data exist. In the typical, finite training data model, the decision-maker has access to a collection of labelled points, $\{(x_i, l_i)\}$, where $x_i \in \mathcal{X}$, and $l_i \in \{+, -\}$. In our case then, the decision-maker receives a probability measure over this space $\sigma \in \mathcal{M}(\mathcal{X} \times \{+, -\})$ ($\mathcal{M}$ denotes the space of probability measures). We can represent such a measure σ by a triple (λ, μ_+, μ_-), where μ_+, μ_- are probability measures on $\mathcal{X}$, and represent the distribution of the positive and negative-labelled points respectively, and $\lambda \in [0, 1]$ is the weight (or probability) of the positively labelled region, and $(1 - \lambda)$ that of the negatively labelled region. The interpretation is that a point-label pair is generated by first choosing a label '$+$' or '$-$' with probability λ or $1-\lambda$, respectively, and then a point is generated according to the corresponding distribution, μ_+ or μ_-. Thus, the underlying distribution μ generating the location of the points (not the labels) is given by $(\lambda\mu_+ + (1-\lambda)\mu_-)$. Thus, if h_{true} is the true classifier, then in the absence of any noise, we would observe $\sigma = (\lambda, \mu_+, \mu_-)$, where μ_+ is the scaled restriction of μ to the region $h_{\text{true}}(+) \triangleq \{x : h_{\text{true}}(x) = +\}$, and similarly for μ_-:

$$\lambda = \mu(h_{\text{true}}(+)); \quad \mu_+ = \frac{\mu \cdot \chi_{\{h_{\text{true}}(+)\}}}{\lambda}; \quad \mu_- = \frac{\mu \cdot \chi_{\{h_{\text{true}}(-)\}}}{1-\lambda},$$

where if $\lambda = 0$ there is no μ_+, and if $\lambda = 1$ there is no μ_-. Indeed, the triple (λ, μ_+, μ_-) is completely defined by μ and the true classifier h_{true}. Since μ is fixed, we write $(\lambda, \mu_+, \mu_-)_{h_{\text{true}}}$ to denote the triple determined by μ and h_{true}.

Using this terminology, the adversary's action is a map

$$\begin{array}{ccc} T : \mathcal{M}(\mathcal{X} \times \{+,-\}) & \longrightarrow & \mathcal{M}(\mathcal{X} \times \{+,-\}) \\ (\lambda, \mu_+, \mu_-) & \longmapsto & (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-). \end{array}$$

We use the hat symbol, '$\hat{}$' throughout, to denote the observation of the decision-maker. Therefore, while the true probability measure generating the point location is given, as above, by $\mu = \lambda\mu_+ + (1-\lambda)\mu_-$, the decision-maker observes an underlying probability measure of the form $\hat{\mu} = \hat{\lambda}\hat{\mu}_+ + (1-\hat{\lambda})\hat{\mu}_-$.

The restrictions on this map determine the nature and level of noise. We consider two models for the noise, i.e., two adversaries. First, we have a 'flip-only' adversary, corresponding to the noise model where the adversary can flip some fixed fraction of the labels. We also consider a stronger 'move-and-flip' adversary who can not only flip a constant fraction of the points, but may also change their location. For the flip-only adversary the underlying measure μ is the same as the observed measure $\hat{\mu}$. Therefore the decision-maker minimizes the worst-case error where the worst case is over all possible $h \in \mathcal{H}$. This need not be true for the move-and-flip adversary. In this case, the decision-maker has only partial information of the measure μ against which generalization error is computed, and hence the decision-maker must protect himself against the worst-case error, considering all possible classifiers $h \in \mathcal{H}$, as well as all possible underlying measures $\tilde{\mu}$ consistent with the observations $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$.

We do not intend measurability questions to be an issue in this paper. Therefore we assume throughout that all measures (and images under the adversary's action) are measurable with respect to some natural σ-field $\mathcal{G}$.

In each of the two cases above, the level of noise is determined by how different the output probability measure $T(\lambda, \mu_+, \mu_-) = (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$ can be from the true probability measure (λ, μ_+, μ_-). A natural measure for this is the notion of total variation. The distance, in total variation, between measures ν_1, ν_2 is defined as

$$||\nu_1 - \nu_2||_{TV} = \frac{1}{2} \sup_{\substack{k, A_1, \ldots, A_k \in \mathcal{G} \\ \text{s.t. } A_i \cap A_j = \emptyset \text{ for } i \neq j}} \sum_{i=1}^{k} |\nu_1(A_i) - \nu_2(A_i)|.$$

This definition also holds for unnormalized measures. We extend this definition to triples (λ, μ_+, μ_-) by

$$||(\lambda, \mu_+, \mu_-) - (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)||_{TV} \triangleq ||\lambda\mu_+ - \hat{\lambda}\hat{\mu}_+||_{TV} + ||(1-\lambda)\mu_- - (1-\hat{\lambda})\hat{\mu}_-||_{TV}.$$

Therefore, we have:

Definition 1 *An adversary using policy T (either flip-only, or move-and-flip) has power η if given any triple (λ, μ_+, μ_-), his policy T satisfies $||T(\lambda, \mu_+, \mu_-) - (\lambda, \mu_+, \mu_-)||_{TV} \leq \eta$. We abbreviate this, and simply write $||T|| \leq \eta$.*

We can now define the two notions of adversary introduced above.

Definition 2 *A flip-only adversary of power η can choose any policy T such that $||T|| \leq \eta$, and $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-) = T(\lambda, \mu_+, \mu_-)$ satisfies*

$$\mu = \lambda\mu_+ + (1-\lambda)\mu_- = \hat{\lambda}\hat{\mu}_+ + (1-\hat{\lambda})\hat{\mu}_- = \hat{\mu}.$$

Definition 3 *A move-and-flip adversary of power η can choose any policy T such that $||T|| \leq \eta$.*

The decision-maker must base his decision on the 'noisy observations' he receives, in other words, on the triple $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-) = T(\lambda, \mu_+, \mu_-)$ which he sees. His goal is to minimize the worst-case generalization error, where the worst case is taken over consistent $h \in \mathcal{H}$, and also over consistent measures $\tilde{\mu}$. We allow our decision-maker to play a so-called mixed strategy, and rather than output a single classifier $h \in \mathcal{H}$, to output a randomized strategy, α, interpreted to mean that classifier h_i is chosen with probability α_i. We denote the set of these mixed strategies by $\Delta_{\mathcal{H}}$, and a particular mixed strategy by $\alpha \in \Delta_{\mathcal{H}}$. Then, the decision-maker's strategy is a map:

$$\begin{aligned} D_{\eta,\mathcal{H}} : \mathcal{M}(\mathcal{X} \times \{+,-\}) &\longrightarrow \Delta_{\mathcal{H}} \\ (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-) &\longmapsto \alpha. \end{aligned}$$

The idea is that if the decision-maker can eliminate some elements of $\mathcal{H}$, but cannot identify a unique optimal choice, then the resulting strategy $D_{\eta,\mathcal{H}}$ will output some measure supported over the ambiguous elements of $\mathcal{H}$. We explicitly assume that the decision-maker's policy is a function of η, the power of the adversary. In a worst-case formulation, a decision-maker without knowledge of η is necessarily powerless. We also assume that the decision-maker knows whether the adversary has flip-only, or move-and-flip power. We do not assume that the decision-maker has any knowledge of the underlying distribution μ that generates the location of the points. For the flip-only adversary, the decision-maker receives exact knowledge 'for free' since by ignoring the $\{+,-\}$-labels, he obtains the true underlying distribution μ. Therefore in this case there is only a single consistent underlying measure, namely, the correct measure μ, and the decision-maker need only protect against the worst-case $h \in \mathcal{H}$. In the case of the move-and-flip adversary, however, the decision-maker receives only partial knowledge of the probability measure that generates the location of the points.

Given a strategy D of the decision maker and a rule T for the adversary, we define the error for a given measure μ and a true classifier h_{true} as:

$$\text{Error}(\mu, h_{\text{true}}, \eta, D, T) \stackrel{\triangle}{=} \left[\mathcal{E}_\mu(h_{\text{true}}; D(T((\lambda, \mu_+, \mu_-)_{h_{\text{true}}})))\right]. \tag{2.1}$$

2.3 An Optimization-Based Characterization

In this section we characterize the optimal policy of the decision-maker, and also the worst-case policy of the adversary, i.e., the worst-case noise, given the policy of the decision-maker. The noise-selecting adversary has access to the true triple (λ, μ_+, μ_-), and seeks to maximize the true error incurred. The decision-maker sees only the corrupted version $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$, and minimizes the worst-case error, where the worst case is taken over all possible, or consistent triples $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ that the particular adversary with power η (flip-only, or move-and-flip) could, under any policy, map to the observed triple $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$.[2]

For the flip-only adversary, any consistent triple $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ the decision-maker considers must satisfy $\tilde{\lambda}\tilde{\mu}_+ + (1-\tilde{\lambda})\tilde{\mu}_- = \mu$. Therefore the worst case over all consistent triples becomes a worst case over all consistent classifiers.

When facing the move-and-flip adversary, it may no longer be true that $\tilde{\lambda}\tilde{\mu}_+ + (1-\tilde{\lambda})\tilde{\mu}_- = \mu$. Therefore the decision-maker must consider the worst case over all consistent classifiers, and also over all consistent underlying measures ν such that $\nu = \tilde{\lambda}\tilde{\mu}_+ + (1-\tilde{\lambda})\tilde{\mu}_-$ for some possible $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ with total variation at most η from $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$. We refer to this set of consistent underlying measures as

$$\Phi \stackrel{\triangle}{=} \Phi(\eta, (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)).$$

We define the following two setups for a fixed measure μ on $\mathcal{X}$, $h_{\text{true}} \in \mathcal{H}$, and a value η for the power of the adversary.

(S1) The flip-only setup:

$$D_1 \stackrel{\triangle}{=} \operatorname*{argmin}_{D_{\eta,\mathcal{H}}} \Big[\max_{\substack{T:||T||\le\eta \\ T \text{ flip-only}}} \big[\max_{h\in\mathcal{H}} \text{Error}(\mu, h, \eta, D, T) \big] \Big] \tag{2.2}$$

$$T_1 \stackrel{\triangle}{=} \operatorname*{argmax}_{\substack{T:||T||\le\eta \\ T \text{ flip-only}}} \left[\text{Error}(\mu, h_{\text{true}}, \eta, D_1, T)\right].$$

The decision-maker knows η and $\mathcal{H}$, and can infer μ since the adversary is flip-only. Thus he chooses D_1 to minimize the worst-case error, where the worst case is over classifiers $h \in \mathcal{H}$. The adversary has prior knowledge of μ, h_{true} and $\mathcal{H}$, and of course η, and chooses his strategy to maximize the *true* error, i.e., the error with respect to h_{true} and μ.

(S2) The move-and-flip setup:

$$D_2 \stackrel{\triangle}{=} \operatorname*{argmin}_{D_{\eta,\mathcal{H}}} \Big[\max_{T:||T||\le\eta} \big[\max_{\substack{\nu\in\Phi \\ h\in\mathcal{H}}} \text{Error}(\nu, h, \eta, D, T) \big] \Big] \tag{2.3}$$

$$T_2 \stackrel{\triangle}{=} \operatorname*{argmax}_{T:||T||\le\eta} \left[\text{Error}(\mu, h_{\text{true}}, \eta, D_2, T)\right].$$

Here the adversary is no longer constrained to pick T so that $\hat{\mu} = \mu$. In this case the decision-maker must choose a policy D_2 to minimize the worst-case generalization error, with respect to $h \in \mathcal{H}$ and also measures $\nu \in \Phi$. The adversary again tries to maximize the true error w.r.t. h_{true} and μ.

We use Error^i $(i = 1, 2)$ to denote the error in $S1$ and $S2$ when μ, h_{true}, and η are clear from the context, i.e., $\text{Error}^i = \text{Error}(\eta, h_{\text{true}}, \eta, D_i, T_i)$. We show below that the max and min in both (2.2) and (2.3) are attained, and can be computed by solving appropriate optimization problems. We interpret the argmin/argmax as selecting an arbitrary optimal solution if there are more than one.

The fact that the max and min in both (2.2) and (2.3) are attained by some rule requires a proof. We show below that this is indeed the case for both setups since the respective rules can be computed by solving appropriate optimization problems.

[2] We remark again that unlike the game-theoretic setup, the decision-maker does not assume a rational adversary. We consider this case elsewhere.

$S1$ and $S2$ are not equivalent.

We first show by example that the "flip only" setup and the "move and flip" setup are not equivalent. This is the case even for two classifiers. Indeed, consider the case $\mathcal{X} = [-5,5] \subseteq \mathbb{R}$, with threshold classifiers $\mathcal{H} = \{h_1, h_2\}$ with $h_1(+) = [0,5]$ and $h_2(+) = [1,5]$. Then the disagreement region is $[0,1)$. Suppose h_1 is the true classifier, and that the true underlying measure μ is uniform on $[-5,5]$, so that $\mu([0,1)) = 10\%$. For $\eta < 5\%$, $\mathrm{Error}^1 = \mathrm{Error}^2 = 0$. For $\eta \geq 5\%$, however, both the flip-only and move-and-flip adversaries can cause error. Suppose $\eta = 10\%$. In $S1$, the decision-maker knows the true μ, and hence knows that $\mu([0,1)) = \eta = 10\%$. Thus regardless of the action of the adversary, the decision-maker's optimal strategy is $(\alpha_1, \alpha_2) = (1/2, 1/2)$, and the error is therefore $\mathrm{Error}^1 = 10/2 = 5\%$. In $S2$, however, the optimal strategy of the adversary is unique: flip the labels of all the points in $[0,1)$. The decision-maker sees $\hat{\mu}([0,1)) = 10\%$, but because the adversary has move-power, the decision-maker does not know μ exactly. His goal is to minimize the error in the worst case, where now the worst case is over classifiers, and also over possible underlying measures. From his observations, the decision-maker can only conclude that if $h_{\mathrm{true}} = h_1$ then $0\% \leq \mu([0,1)) \leq 10\%$, and if $h_{\mathrm{true}} = h_2$, then $0\% \leq \mu([0,1)) \leq 20\%$. The worst-case error corresponding to a strategy (α_1, α_2) is therefore $\max\{10\alpha_1; 20\alpha_2\}$. Minimizing this objective function subject to $\alpha_1 + \alpha_2 = 1$, and $\alpha_1, \alpha_2 \geq 0$, we find $(\alpha_1, \alpha_2) = (1/3, 2/3)$, and the true error (as opposed to the worst-case error) is $\mathrm{Error}^2 = (1/3) \cdot 0 + (2/3) \cdot 10 = 20/3$, which is greater than Error^1.

3 Optimal Strategy and Worst-Case Noise

In this section we consider S1 and S2, and determine optimal strategies for the decision-maker, and the optimal strategy for the adversary, i.e., the worst-case noise.

3.1 The Decision-Maker in $S1$

First we consider the decision-maker's optimal strategy for $S1$, i.e., in the face of the flip-only adversary. The decision-maker outputs a mixed strategy $\alpha \in \Delta_{\mathcal{H}}$. The support of the weight vector α is the subset $\mathcal{F}$ of 'feasible' classifiers in $\mathcal{H}$, which incur at most error η. This set is often referred to as the "version space".

Definition 4 *Given the output $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-) = T(\lambda, \mu_+, \mu_-)$ of a flip-only adversary with power η, the set of* feasible, *and hence* ambiguous *classifiers, $\mathcal{F} \triangleq \mathcal{F}_\eta(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-) \subseteq \mathcal{H}$, is given by*

$$\mathcal{F} \triangleq \{h \in \mathcal{H} : \hat{\lambda}\hat{\mu}_+(h(-)) + (1-\hat{\lambda})\hat{\mu}_-(h(+)) \leq \eta\}. \tag{3.4}$$

Here we define $h(+)$ to be the positively labelled region, and $h(-)$ the negatively labelled region, so that $\hat{\lambda}\hat{\mu}_+(h(-))$ is the measure of the positive labels observed in the region $h(-)$. The measure of the region where the true classifier disagrees with the observed measure can be at most η. That is,

$$\hat{\lambda}\hat{\mu}_+(h_{\mathrm{true}}(-)) + (1-\hat{\lambda})\hat{\mu}_-(h_{\mathrm{true}}(+)) \leq \eta.$$

This follows by our assumption that the adversary has power η, and because $\lambda\mu_+(h_{\mathrm{true}}(-)) + (1-\lambda)\mu_-(h_{\mathrm{true}}(+)) = 0$. Therefore, $\mathcal{F}$ is the set of classifiers in $\mathcal{H}$ that could possibly be equal to h_{true} and thus Definition 4 above indeed gives the set of feasible, and therefore ambiguous, classifiers. In particular, under the assumption of proper learning, $h_{\mathrm{true}} \in \mathcal{F}$.

Next, the decision-maker must compute the value of α_h for every $h \in \mathcal{F}$, the feasible subset of classifiers. For any mixed strategy (this is sometimes referred to as a "probabilistic hypothesis") $\alpha \in \Delta_{\mathcal{H}}$ that the decision-maker might choose, the error incurred is

$$\mathcal{E}_\mu(h_{\mathrm{true}}; \alpha) \quad = \quad \sum_{h \neq h_{\mathrm{true}}} \alpha_h \mu(\blacktriangle(h, h_{\mathrm{true}})), \tag{3.5}$$

where for any two classifiers h', h'', we define $\blacktriangle(h', h'') \triangleq \{x : h'(x) \neq h''(x)\}$ to be the region where they differ.

The decision-maker, however, does not know h_{true}, and hence his optimal strategy is the one that minimizes the worst-case error, $\max_{h_{\mathrm{true}} \in \mathcal{H}} \mathcal{E}_\mu(h_{\mathrm{true}}; \alpha)$. In the case of the flip-only adversary, the decision-maker sees the probability measure $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$, and since he knows that $\mu = \hat{\mu}$, he can correctly compute the value $\mu(\blacktriangle(h', h''))$ for any two classifiers h', h''. In other words, the decision-maker knows the true weight of any region where two classifiers disagree, and therefore we can state the following result which is a restatement of the above.

Proposition 5 *The optimal policy of the decision-maker in $S1$ is given by computing the minimizer of:*

$$\min_{\alpha} \max_{h_{\mathrm{true}} \in \mathcal{F}} \sum_{h \neq h_{\mathrm{true}}} \alpha_h \mu(\blacktriangle(h, h_{\mathrm{true}})). \tag{3.6}$$

Enumerating the set $\mathcal{F}$ as $\{h_1, \ldots, h_k\}$, the optimal α is computed by solving the following linear optimization problem:

$$\begin{array}{lll} \min: & u & \\ \text{s.t.}: & u \geq \sum_{i \neq j} \alpha_i \mu(\blacktriangle(h_i, h_j)) & j = 1, \ldots, k \\ & \sum_i \alpha_i = 1 & \\ & \alpha_i \geq 0 & i = 1, \ldots, k. \end{array}$$

PROOF. The proof follows directly from the definition of the error associated to any mixed strategy α, given in (3.5). □

We note that in [CBDF+99] the question of how to choose the best probabilistic hypothesis was considered. The solution there was to randomize between two (maximally apart) classifiers or to choose a majority vote. We now explain why this is suboptimal. Consider three linear classifiers in general position in the plane $\mathcal{H} = \{h_1, h_2, h_3\}$ and let's suppose that there are 7 regions in the plane according to the agreement of the classifiers (assume that $h_1(+) \cap h_2(+) \cap h_3(+) \neq \emptyset$). Suppose that the decision maker observes that $\hat{\mu}_+$ has support only on $h_1(+) \cap h_2(+) \cap h_3(+)$ (assume that $\hat{\lambda} = 1 - 3\eta$ and that $\eta < 1/4$) and that $\hat{\mu}_-$ has equal support of η on $h_1(-) \cap h_2(-) \cap h_3(+)$, $h_1(-) \cap h_2(+) \cap h_3(-)$ and $h_1(+) \cap h_2(-) \cap h_3(-)$. The example is constructed so that choosing any one classifier, in the worst case can lead to an error of 2η. It is easy to see that a majority vote would lead to

a worst case error of 2η. Mixing between any two classifiers would lead to a worst case error of 2η as well. Mixing between the 3 classifiers, which is suggested by Proposition 5 leads to a worst case error of $4\eta/3$ since we will get the classifier right with probability 1/3 and incur the 2η loss with probability 2/3.

3.2 The Decision-Maker in $S2$

Next we consider the setup $S2$, with the more powerful move-and-flip adversary. Again, the goal of the decision-maker is to pick a mixed strategy $\alpha \in \Delta_{\mathcal{H}}$, that minimizes the error given in (3.5). The set $\mathcal{F}$ of ambiguous classifiers is as defined in (3.4). In this case, however, in addition to not knowing $h_{\rm true}$, the decision-maker also does not know the underlying measure μ, and hence the values $\mu(\blacktriangle(h', h''))$, exactly. As introduced in Section 2.3, we use $\Phi \stackrel{\triangle}{=} \Phi(\eta, (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-))$ to denote the set of measures consistent with $(\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$. Thus the decision-maker seeks to minimize the worst-case error, now over $\mathcal{H}$ and Φ.

Any points that have the wrong label w.r.t. h could have been both moved and flipped. Therefore, to compute the worst case possible values of $\mu(\blacktriangle(h', h''))$, for each classifier h the decision-maker considers, he must consider the observed measure of the points that have the *correct* label, and the *wrong* label, with respect to h. Thus we define:

$$\begin{aligned} \overset{\text{wrong}}{\hat{\mu}_h}(\blacktriangle(h', h'')) &\stackrel{\triangle}{=} \hat{\lambda}\hat{\mu}_+(\blacktriangle(h', h'') \cap h(-)) + \\ &\qquad (1-\hat{\lambda})\hat{\mu}_-(\blacktriangle(h', h'') \cap h(+)) \\ \overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h', h'')) &\stackrel{\triangle}{=} \hat{\mu}(\blacktriangle(h', h'')) - \overset{\text{wrong}}{\hat{\mu}_h}(\blacktriangle(h', h'')). \end{aligned} \tag{3.7}$$

In Proposition 6 below, the decision-maker uses these quantities to compute his optimal strategy that protects against the worst-case consistent classifier $h \in \mathcal{F}$, and underlying measure $\nu \in \Phi$. The worst-case classifier h and measure ν may depend on the action α the decision-maker chooses. Thus, the decision-maker must solve a $\min\max$ linear program. In doing so, he implicitly computes the worst-case measure ν as well, by computing a saddle point.

Proposition 6 (a) *The decision-maker's optimal policy, is to compute the set $\mathcal{F}$, and then compute the optimal weight-vector α that is the minimizer of*

$$\begin{aligned} \min_{\alpha} \max_{\substack{\nu\in\Phi \\ h_{\rm true}\in\mathcal{H}}} \mathcal{E}_\nu(h_{\rm true};\alpha) &= \min_{\alpha} \max_{\substack{\nu\in\Phi \\ h_{\rm true}\in\mathcal{H}}} \\ &\qquad \sum_{h\neq h_{\rm true}} \alpha_h \nu(\blacktriangle(h, h_{\rm true})), \end{aligned} \tag{3.8}$$

where the max is over $\mathcal{H}$ and Φ. The $\min$ *and the* $\max$ *are both attained.*

(b) *Moreover, the optimal strategy of the decision-maker is obtained as the solution to a robust linear optimization problem, which we reformulate as a single linear optimization.*

Recall that in $S2$, in addition to the labels, the underlying measure μ is also corrupted. Therefore the decision-maker must compute the strategy α with respect to the worst-case feasible classifier, and the worst-case consistent values for $\mu(\blacktriangle(h', h''))$, i.e., the worst-case values for $\nu(\blacktriangle(h', h''))$ for $\nu \in \Phi$.

The worst case over ν depends on the worst case over $h \in \mathcal{H}$. That is, if h_1 is the true classifier, then the worst-case values for $\nu(\blacktriangle(h', h''))$ may be different from the worst-case value if h_2 is the true classifier.

The worst-case values are computed using $\overset{\text{wrong}}{\hat{\mu}_h}(\blacktriangle(h', h''))$ and $\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h', h''))$. The idea is as follows: if some h is the true classifier, then any measure in the region $\blacktriangle(h', h'')$ that is incorrectly labelled with respect to h may have also been moved from some other region. Therefore in the case that $h = h_{\rm true}$, the weight of any particular region $\blacktriangle(h', h)$ could be as large as the weight of the correctly labeled points under $\hat{\mu}$, $\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h', h''))$, plus the weight (again under $\hat{\mu}$) of the mislabelled points with respect to h in all other regions, plus the additional weight that could be moved to $\blacktriangle(h', h)$ using any 'unused' power of the adversary. The weight of the mislabelled points is

$$\hat{\lambda}\hat{\mu}_+(h(-)) + (1-\hat{\lambda})\hat{\mu}_-(h(+)).$$

The unused power is

$$\eta - \hat{\lambda}\hat{\mu}_+(h(-)) + (1-\hat{\lambda})\hat{\mu}_-(h(+)).$$

Therefore the weight (under $\hat{\mu}$) of the mislabelled points with respect to any h, plus the unused power, must be exactly η.

If $h = h_{\rm true}$, consider some region $\blacktriangle(h', h)$. The reasoning above tells us that the worst-case measure of this region is $\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h', h)) + \eta$. The following lemma makes this intuition precise, and shows that this is indeed the case.

Lemma 7 *Assume that $\blacktriangle(h, h') \neq \emptyset$ for any $h \neq h'$. Then, if $h = h_{\rm true}$, we have*

$$\mu(\blacktriangle(h, h')) \leq \overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h, h')) + \eta.$$

This bound is tight in the sense that there is a measure $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ with total variation at most η from the observations, that attains the upper bound.

PROOF. We exhibit the following triple $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ that satisfies $||(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-) - (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)||_{TV} \leq \eta$: Assume, without loss of generality, that $\blacktriangle(h, h') \subseteq h(+)$. Let θ be any probability measure over $\mathcal{X}$, supported on $\blacktriangle(h, h')$. Then, define:

$$\begin{aligned} \tilde{\lambda} &= \hat{\lambda} + (1-\hat{\lambda})\hat{\mu}_-(h(+)), \\ \tilde{\mu}_- &= \hat{\mu}_- - \hat{\mu}_-\Big|_{h(+)}, \\ \tilde{\mu}_+ &= \frac{\left(\hat{\lambda}\left(\hat{\mu}_+ - \hat{\mu}_+\Big|_{h(-)}\right) + \kappa\theta\right)}{\hat{\lambda} + (1-\hat{\lambda})\hat{\mu}_-(h(+))}, \end{aligned}$$

where $\kappa = \Big((1-\hat{\lambda})\hat{\mu}_-(h(+)) + \hat{\lambda}\hat{\mu}_+(h(-))\Big)$. For the triple $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$, there exists a move-and-flip policy T with $||T|| \leq \eta$, such that $T(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-) = (\hat{\lambda}, \hat{\mu}_+, \hat{\mu}_-)$, hence the scalar upper bound is attainable. □

For the vector case, if $\blacktriangle(h,h_1) \cap \cdots \cap \blacktriangle(h,h_k) \neq \emptyset$, there exists a triple $(\tilde{\lambda}, \tilde{\mu}_+, \tilde{\mu}_-)$ that satisfies

$$(\mu(\blacktriangle(h,h_1)),\ldots,\mu(\blacktriangle(h,h_k))) = (\overset{\text{correct}}{\hat{\mu}}(\blacktriangle(h,h_1)),\ldots,\overset{\text{correct}}{\hat{\mu}}(\blacktriangle(h,h_k))) + \eta(1,\ldots,1).$$

This follows by replacing $\blacktriangle(h,h')$ by $\blacktriangle(h,h_1)\cap\cdots\cap\blacktriangle(h,h_k)$ in the proof above. In general, however, the tightness result does not hold simultaneously for many classifiers. That is to say, given classifiers $\{h_1,\ldots,h_k\}$ different from some h, if $\blacktriangle(h,h_1) \cap \cdots \cap \blacktriangle(h,h_k) = \emptyset$ (as is in general the case), then, while the lemma tells us that $\mu(\blacktriangle(h,h_i)) \leq \overset{\text{correct}}{\hat{\mu}}(\blacktriangle(h,h_i)) + \eta$ for each i, there will be no measure $\nu \in \Phi$ which realizes these upper bounds simultaneously. Moreover, the worst-case values then will depend on the decision-maker's particular choice of α. The α-dependent worst-case consistent values for $\mu(\blacktriangle(h',h''))$ are computed implicitly in the robust LP below.

With this intuition, and the result of the lemma, we can now prove the proposition, and explicitly give the LP that yields the optimal strategy of the decision-maker.

PROOF. (**of Proposition 6**) The proof proceeds in three main steps:

(i) First we show that the error, and hence the optimal strategy of the decision-maker, depends only on a finite dimensional equivalence class of measures $\nu \in \Phi$. The first part of the proof is to characterize this finite dimensional set.

(ii) Next, we establish the connection to robust optimization, and write a robust optimization problem that we claim yields the decision-maker's optimal strategy. Proving this claim is the second part of the proof.

(iii) Finally, we show that the robust optimization problem may in fact be rewritten as a single LP, using duality theory of linear programming.

For $\mathcal{F}$ the set of ambiguous classifiers, the decision-maker's policy is given by

$$\min_{\alpha} \max_{\substack{\nu\in\Phi \\ h_{\text{true}}\in\mathcal{F}}} \mathcal{E}_\nu(h_{\text{true}};\alpha) = \min_{\alpha} \max_{\substack{\nu\in\Phi \\ h_{\text{true}}\in\mathcal{F}}} \sum_{h\neq h_{\text{true}}} \alpha_h \nu(\blacktriangle(h,h_{\text{true}})),$$

where α is supported on $\mathcal{F}$. While the worst case is over classifiers $h \in \mathcal{F}$ and all measures $\nu \in \Phi$, the worst-case error incurred for any particular strategy α in fact can only depend on the values of $\nu(\blacktriangle(h',h''))$ for every $h',h'' \in \mathcal{F}$. Therefore we can consider equivalence classes of measures in Φ that have the same values $\nu(\blacktriangle(h',h''))$. This reduces the inner maximization to a finite dimensional one. Enumerate the set $\mathcal{F}$ as $\{h_1,\ldots,h_k\}$. Then for any fixed $h_j \in \mathcal{F}$, if $h_{\text{true}} = h_j$, then the regions whose measure is important for the error computation, are those that can be written as

$$\left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right),$$

for some $S \subseteq \{1,\ldots,k\}$. We use $\blacktriangle(h_i,h_j)^c$ to denote the complement of the set. We define a variable $\hat{\xi}_{S,j}$ to represent the amount of mass that can be added (in the worst case) to the region $\left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right)$ in the case where h_j is the true classifier. We can consider these as components of a vector in $\mathbb{R}^{2^k-1}$, indexed by nonempty subsets $S \subseteq \{1,\ldots,k\}$. Any such vector corresponds to an equivalence class of measures $\nu \in \Phi$, that are indistinguishable to the decision-maker, in the sense that they induce precisely the same error. Given such a vector, the weight of the region $\blacktriangle(h_i,h_j)$ is then $\left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j}\right]$ and thus for a given α, the error would be

$$\sum_{i\neq j} \alpha_i \left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j}\right].$$

For any fixed j, the collection of variables $\left(\hat{\xi}_{S,j}\right)_{S\subseteq\{1,\ldots,k\}}$ must satisfy four properties in order to correspond to some measure $\nu \in \Phi$. The variables must be nonnegative, and the sum over S of $\hat{\xi}_{S,j}$ must be at most η. This follows since the total amount of mass moved or flipped must be at most η, by definition of the power of the adversary. Third, if the set $\left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right)$ is empty, then the corresponding variable $\hat{\xi}_{S,j}$ must be zero. Finally, the weight of each region $\blacktriangle(h_i,h_j)$ can be at most 100%, and thus we must have

$$\left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j}\right] \leq 100\%.$$

Therefore, if $h_j = h_{\text{true}}$, the possible values of $\hat{\xi}_{\cdot,j} \in \mathbb{R}^{2^k-1}$ are given by:

$$\Xi(j) = \left\{ \left(\hat{\xi}_{\cdot,j}\right) : \begin{array}{l} \sum_S \hat{\xi}_{S,j} \leq \eta, \\ \hat{\xi}_{S,j} \geq 0, \forall S \subseteq \{1,\ldots,k\}, \ S \neq \emptyset, \\ \hat{\xi}_{S,j} = 0, \forall S : \left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \\ \qquad \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right) = \emptyset, \\ \sum_{S\ni i} \hat{\xi}_{S,j} \leq 100 - \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) \\ \qquad \forall i \neq j. \end{array} \right\}$$

For every j, the set $\Xi(j)$ is a polytope. The decision-maker must choose some α that minimizes the worst-case error, where the worst case is over possible $h_{\text{true}} \in \mathcal{F} = \{h_1,\ldots,h_k\}$, and then once that h_j is fixed, the worst case over all possible $(\hat{\xi}_{\cdot,j}) \in \Xi(j)$. Therefore the optimal strategy α of the decision-maker is the solution to the following robust opti-

mization problem:

$$
\begin{aligned}
\min: \quad & u \\
\text{s.t.}: \quad & u \geq \max_{\left\{(\hat{\xi}_{S,j})_{S\subseteq\{1,\ldots,k\}} \in \Xi(j)\right\}} \sum_{i\neq j} \alpha_i \\
& \left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j} \right], \; j = 1,\ldots,k \\
& \sum_i \alpha_i = 1, \quad \alpha_i \geq 0,
\end{aligned}
$$

$$
\Xi(j) \quad = \quad \left\{ \left(\hat{\xi}_{\cdot,j}\right) : \begin{array}{l} \sum_S \hat{\xi}_{S,j} \leq \eta, \\ \hat{\xi}_{S,j} \geq 0, \forall S \subseteq \{1,\ldots,k\}, \;\; S \neq \emptyset, \\ \hat{\xi}_{S,j} = 0, \forall S \;:\; \left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \\ \qquad \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right) = \emptyset, \\ \sum_{S\ni i} \hat{\xi}_{S,j} \leq 100 \; - \; \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) \\ \qquad \forall i \neq j. \end{array} \right\}
$$

First we prove that this robust optimization indeed yields the strategy of the decision-maker that minimizes the worst-case effort. The proof of this follows by a combination of the methods used to prove Proposition 5 and Lemma 7. Certainly, for any $h_j \in \mathcal{F}$ and $\nu \in \Phi$, there exists a vector $(\hat{\xi}_{\cdot,j}) \in \Xi(j)$ such that

$$
\nu(\blacktriangle(h_i,h_j) = \left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j} \right], \; \forall i \neq j.
$$

The technique of Lemma 7 establishes the converse, namely, for any feasible vector $(\hat{\xi}_{\cdot,j}) \in \Xi(j)$ there exists a measure $\nu \in \Phi$ that is consistent with the observed measure, and such that for any $i \in \{1,\ldots,k\}$,

$$
\nu(\blacktriangle(h_i,h_j)) = \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j}.
$$

Thus we have shown that the sets $\Xi(j)$ are indeed the sets we should be considering. Next we show that the optimization we write down is the correct one. The proof of this follows that of Proposition 5. Let α^* be the minimizer of the expression above, and let u^* be the optimal value of the optimization. If the decision-maker chooses some mixed strategy ρ that is not a minimizer of the above, then there must exist some $r \in \{1,\ldots,k\}$, corresponding to some $h_{\text{true}} \in \mathcal{F}$, and also a vector $(\hat{\xi}_{\cdot,r}) \in \Xi(r)$ feasible for the above linear optimization, for which

$$
\sum_{i\neq r} \rho_i \left[\overset{\text{correct}}{\hat{\mu}}_{h_r}(\blacktriangle(h_i,h_r)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,r} \right] > u^*.
$$

Thus, we must have

$$
\sum_{i\neq r} \rho_i \left[\overset{\text{correct}}{\hat{\mu}}_{h_r}(\blacktriangle(h_i,h_r)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,r} \right] > \max_{\substack{j\in\{1,\ldots,k\} \\ \hat{\xi}_{S,j}\in\Xi(j)}} \sum_{i\neq j} \alpha_i^* \left[\overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j} \right].
$$

But then there must exist a measure $\nu \in \Phi$ consistent with the observed measure, for which

$$
\nu(\blacktriangle(h_i,h_j)) = \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) + \sum_{\substack{S\subseteq\{1,\ldots,k\} \\ i\in S}} \hat{\xi}_{S,j},
$$

and thus we have:

$$
\begin{aligned}
\max_{\substack{\mu\in\Phi \\ h_{\text{true}}\in\mathcal{H}}} \mathcal{E}_\mu(h_{\text{true}};\rho) \quad &\geq \quad \mathcal{E}_\nu(h_r;\rho) \\
&> \quad \max_{\substack{\nu\in\Phi \\ h_{\text{true}}\in\mathcal{H}}} \mathcal{E}_\nu(h_{\text{true}};\alpha^*).
\end{aligned}
$$

Therefore, if ν is indeed the true probability measure generating the location of the points, and if h_r is the true classifier, then the error incurred by using strategy ρ is strictly greater than the error incurred using strategy α^*. Since both ν and h_r are consistent with the observed probability measure and labels, respectively, the mixed strategy ρ does not minimize the worst-case error.

On the other hand, by similar reasoning, if ρ is not an optimal strategy, i.e., if it is does not minimize the worst-case error as given in (3.8), then it is a strictly suboptimal solution to the linear optimization. This completes the proof that the robust optimization above indeed yields the strategy of the adversary which minimizes the worst-case error, where the worst case is over $h \in \mathcal{F}$ and also $\nu \in \Phi$. This concludes the proofs of parts **(i)** and **(ii)**.

We have left to prove the second part of the proposition, and part **(iii)** in the outline, namely, that we can rewrite the robust optimization problem as a single LP. First, we remark that for each j, the set $\Xi(j)$ is a polytope. The problem then, is a robust linear optimization problem. Using standard results from duality theory [BTN99], this can be reformulated as an ordinary linear optimization problem.

We have the robust linear optimization problem:

$$\begin{array}{lll}\min: & u & \\ \text{s.t.}: & u \geq \max\limits_{\left\{\left(\hat{\xi}_{S,1}\right)_{S\subseteq\{1,\ldots,k\}} \in \Xi(1)\right\}} \sum\limits_{i\neq 1} \alpha_i & \\ & \qquad \left[\overset{\text{correct}}{\hat{\mu}}_{h_1}(\blacktriangle(h_i,h_1)) + \sum\limits_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,1}\right] & \\ & u \geq \max\limits_{\left\{\left(\hat{\xi}_{S,2}\right)_{S\subseteq\{1,\ldots,k\}} \in \Xi(2)\right\}} \sum\limits_{i\neq 2} \alpha_i & \\ & \qquad \left[\overset{\text{correct}}{\hat{\mu}}_{h_2}(\blacktriangle(h_i,h_2)) + \sum\limits_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,2}\right] & \\ & \qquad \vdots & \\ & u \geq \max\limits_{\left\{\left(\hat{\xi}_{S,k}\right)_{S\subseteq\{1,\ldots,k\}} \in \Xi(k)\right\}} \sum\limits_{i\neq k} \alpha_i & \\ & \qquad \left[\overset{\text{correct}}{\hat{\mu}}_{h_k}(\blacktriangle(h_i,h_k)) + \sum\limits_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,k}\right] & \\ & \sum\limits_i \alpha_i = 1, \quad \alpha_i \geq 0. & \end{array}$$

Note that the robustification here is constraintwise-rectangular, that is, the uncertainty set has the form

$$\Xi = \Xi(1) \times \cdots \times \Xi(k).$$

Therefore, we can consider each constraint individually. Indeed, each inequality can be rewritten as

$$u - \sum_{i\neq j} \alpha_i \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j)) \geq \max_{\left\{\left(\hat{\xi}_{S,j}\right)_{S\subseteq\{1,\ldots,k\}} \in \Xi(j)\right\}} \sum_{i\neq j} \alpha_i \left[\sum_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,j}\right],$$

and thus we can consider the linear optimization:

$$\begin{array}{lll}\max: & \sum\limits_{i\neq j} \alpha_i \left[\sum\limits_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,j}\right] & (3.9)\\ \text{s.t.}: & \left(\hat{\xi}_{S,j}\right)_{S\subseteq\{1,\ldots,k\}} \in \Xi(j). & \end{array}$$

The objective function is bilinear in both α_i and $\hat{\xi}_{S,j}$. We have

$$\sum_{i\neq j} \alpha_i \left[\sum_{\substack{S\subseteq\{1,\ldots,k\}\\ i\in S}} \hat{\xi}_{S,j}\right] = \sum_{S\subseteq\{1,\ldots,k\}} \hat{\xi}_{S,j} \left[\sum_{i\in S} \alpha_i\right],$$

and hence defining the vector c by $c_S = \sum_{i\in S} \alpha_i$ we can write the objective function as $c'\hat{\xi}_{\cdot,j}$. The polytope $\Xi(j)$ is defined by equalities and inequalities among the variables. Writing these in vector form, we have:

$$\begin{array}{rcl} -[I]\hat{\xi}_{\cdot,j} & \leq & 0, \\ [Q^{(j)}]\hat{\xi}_{\cdot,j} & = & 0, \\ (1,1,\ldots,1)'\hat{\xi}_{\cdot,j} & \leq & \eta, \\ [R^{(j)}]\hat{\xi}_{\cdot,j} & \leq & (100 - \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_1,h_j)), \ \ldots \\ & & \quad ,100 - \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_k,h_j))). \end{array} \tag{3.10}$$

Here, I is the identity matrix, $Q^{(j)}$ is a subset of the identity matrix corresponding to the sets S for which we have $\left(\bigcap_{i\in S} \blacktriangle(h_i,h_j)\right) \cap \left(\bigcap_{i\notin S} \blacktriangle(h_i,h_j)^c\right) = \emptyset$, and the generic row of $R^{(j)}$ contains a 1 in every index containing a particular i. Writing the equality as $[Q^{(j)}]\hat{\xi}_{\cdot,j} \leq 0$, and $-[Q^{(j)}]\hat{\xi}_{\cdot,j} \leq 0$, we can express the constraints defining $\Xi(j)$ more compactly as

$$\Xi(j) = \left\{\left(\hat{\xi}_{S,j}\right)_{S\subseteq\{1,\ldots,k\}} : A^{(j)}\hat{\xi}_{\cdot,j} \leq b\right\}.$$

The matrices $A^{(j)}$, and the vector b, are given by the vector inequalities in (3.10) above:

$$A^{(j)} = \left[\begin{array}{c} -I \\ \hline Q^{(j)} \\ \hline -Q^{(j)} \\ \hline R^{(j)} \\ \hline 1\ 1\ \cdots\ 1\ 1 \end{array}\right] \quad b = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 100 - \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_1,h_j)) \\ \vdots \\ 100 - \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_k,h_j)) \\ \eta \end{pmatrix}.$$

Note that while the vector $c^{(j)}$ is a linear function of α, the matrices $A^{(j)}$ and the vector b are constant. We can then rewrite the linear optimization (3.9) as

$$\begin{array}{ll}\max: & c'\hat{\xi}_{\cdot,j} \\ \text{s.t.}: & A^{(j)}\hat{\xi}_{\cdot,j} \leq b. \end{array}$$

The linear programming dual to this program is then

$$\begin{array}{ll}\min: & (b)'\, p^{(j)} \\ \text{s.t.}: & \left(p^{(j)}\right)' A^{(j)} = c \\ & p_S^{(j)} \geq 0, \qquad \forall S \subseteq \{1,\ldots,k\}. \end{array}$$

Recalling that $c_S = \sum_{i\in S} \alpha_i$, the robust linear optimization problem determining the optimal strategy of the decision-maker can now be rewritten:

$$\begin{array}{ll}\min: & u \\ \text{s.t.}: & \left(u - \sum_{i\neq j} \alpha_i \overset{\text{correct}}{\hat{\mu}}_{h_j}(\blacktriangle(h_i,h_j))\right) \geq (b)'\, p^{(j)}, \\ & \qquad j = 1,\ldots,k \\ & \left[\left(p^{(j)}\right)' A^{(j)}\right]_S = \sum_{i\in S} \alpha_i \ , \ \forall S, \ \ j = 1,\ldots,k \\ & p_S^{(j)} \geq 0, \quad \forall S, \ \ j = 1,\ldots,k \\ & \sum_i \alpha_i = 1 \\ & \alpha_i \geq 0. \end{array}$$

The variables of optimization are $\{u, \alpha_i, p_S^{(j)}\}$. The matrices $A^{(j)}$ and the vector b are constants, determined by (3.10). Therefore this is indeed a linear optimization. Thus the proof of parts **(i), (ii)** and **(iii)** is complete, as is the proof of the proposition. □

From a complexity perspective the linear program is exponential in the size of $\mathcal{F}$ since all subsets are considered. This complexity is not an added feature of our development, as even linear classification over non-separable data becomes combinatorial. Still, in spite of this exponential nature the linear program can consider several approximation schemes such as constraint sampling. Moreover, pruning can be used for the classifiers in $\mathcal{F}$; this is pursued elsewhere.

We have thus derived the optimal policy of the decision-maker for both $S1$ and $S2$. We denote these as D_1^* and D_2^*, respectively.

3.3 Bounding the Decision-Maker's Error

As defined above, the decision-maker's policy is a mixed strategy – a randomized policy. In the setting of the worst-case analysis which we consider, the decision-maker stands to benefit from the randomization. For example, suppose $\mathcal{H} = \{h_1, h_2\}$, and $\mu(\blacktriangle(h_1, h_2)) = 2\eta$, where the adversary's power is η. We consider the general optimal strategy for the adversary in the next section. In this case, however, it is clear that the optimal strategy for both the flip-only and the move-and-flip adversary, is to flip half of the 'points', or measure, in $\blacktriangle(h_1, h_2)$. Then the decision-maker cannot distinguish between h_1 and h_2, and the optimal policy is $\frac{1}{2}h_1 + \frac{1}{2}h_2$ The expected worst-case error is $\frac{1}{2}\mu(\blacktriangle(h_1, h_2)) = \eta$. If not for randomization, the worst-case error would have been 2η. Thus there is a concrete benefit to randomization. The next proposition quantifies this benefit (this is similar to Proposition 4.1 from [CBDF+99], but has a slightly tighter lower bound[3]), and obtains bounds on the error an adversary with power η can obtain in any possible setup.

Proposition 8 *In both $S1$ and $S2$, for an adversary with power $\eta \leq 1/2$, there is a setup where $\mathrm{Error}^i \geq (1-\eta)2\eta$ for $i = 1, 2$. On the other hand, we always have $\mathrm{Error}^i \leq 2\eta$ for $i = 1, 2$ and if $\mathcal{F}$ is finite we have $\mathrm{Error}^i \leq (1-1/|\mathcal{F}|)2\eta$ for $i = 1, 2$.*

PROOF. We need to show that the lower bound can be approached arbitrarily closely in the case of the weaker adversary (flip-only), and the upper bound can never be exceeded by the more powerful adversary (move-and-flip). Let $\mathcal{X}$ be the unit circle in $\mathbb{R}^2$, with μ the uniform measure on the disk. If $\eta = p/q$ is rational, divide the disk into q equal, numbered wedges, and define q classifiers, so that classifier i assigns positive labels to wedges $\{i, i+1, \ldots, i+p-1\}$ $\mod q$, and negative labels to the remaining $q-p$ wedges. As in Figure 1 suppose the true classifier is h_1. The optimal action of the adversary with power η is to flip all positive labels. Now all classifiers are indistinguishable, and thus the decision-maker's optimal strategy is the uniform measure over all $\{h_i\}$. The probability of full overlap with h_{true} is $1/q$, the probability of no overlap is $(q-2p+1)/q$, and of overlap r for $0 < r < p$ is $2/q$. Computing the expectation, we have $\mathrm{Error}^1 = (2p(q-p))/q^2 = (1-\eta)2\eta$, as claimed. For η irrational, we can approximate it arbitrarily closely with a rational number. In this case we can approach the lower bound arbitrarily closely.

Next we show that even the more powerful move-and-flip adversary can never exceed the upper bound. Observe that if the power of the adversary is η, then for any two classifiers h_i and h_j, we must have $\mu(\blacktriangle(h_i, h_j)) \leq 2\eta$. Then, if the decision-maker uses the possibly sub-optimal strategy of choosing $\alpha = (1/n, 1/n, \ldots, 1/n)$ (where $n = |\mathcal{F}|$), then since by definition $\blacktriangle(h, h) = \emptyset$ for all h, from expression (3.5) above, it follows that the expected error will never exceed $(1-1/n)2\eta$.

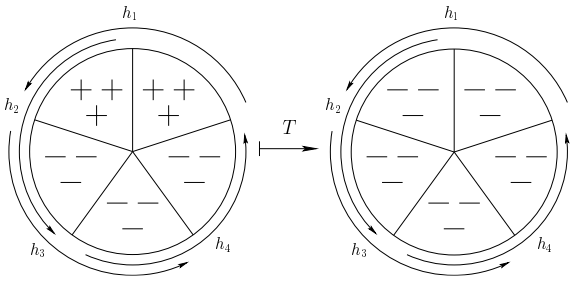

Figure 1: Here we have $\eta = 2/5$, so $p = 2$ and $q = 5$. The figure on the left shows the correct labels. The adversary flips all $+$-labels to $-$. In the figure on the right, all classifiers are indistinguishable to the decision-maker. The decision-maker, therefore, outputs a randomized strategy that is uniform over all n classifiers (here $n = 5$).

3.4 The Adversary

First we consider $S1$ and the flip-only adversary. From Proposition 5, the optimal strategy of the decision-maker is specified by the subset of ambiguous classifiers, $\mathcal{F}$. We call this $\alpha(\mathcal{F})$. Therefore the true error is also a function of $\mathcal{F}$. By an abuse of notation, we can denote this by $\mathcal{E}(\mathcal{F}) \triangleq \sum_{h \neq h_{\mathrm{true}}} \alpha_h(\mathcal{F})\mu(\blacktriangle(h, h_{\mathrm{true}}))$. Then the optimal strategy of the adversary is to create an ambiguous set $\mathcal{F}$ with as large an error as possible. Given any legal strategy T of the adversary, we denote by $\mathcal{F}_T$ the resulting set of ambiguous classifiers. Therefore we have:

Proposition 9 *In $S1$, the adversary's optimal strategy is to maximize $\mathcal{E}(\mathcal{F})$:*

$$T_1^* = \arg \max_{\substack{\mathrm{T}:||\mathrm{T}||\leq\eta \\ T \text{ flip-only}}} \mathcal{E}(\mathcal{F}_{\mathrm{T}}). \tag{3.11}$$

The $\max$ here is attained since there are only finitely many different sets $\mathcal{F}$. If there are more than one (as in general there will be) maps T corresponding to the optimal $\mathcal{F}$, we arbitrarily choose one. Therefore T_1^* is well-defined, and is the optimal strategy for the adversary in $S1$, and the proposition follows.

Next we consider $S2$, and the case of the move-and-flip adversary. From Proposition 6, the decision-maker's optimal

[3] The lower bound of Proposition 4.1 from [CBDF+99] translates to $\mathrm{Error}^i \geq \eta/(2-\eta)$ which is smaller than the bound of Proposition 8.

action is given by an LP that is a function of the ambiguity set $\mathcal{F}$, and the values $\{\overset{\text{correct}}{\hat{\mu}_{h'}}(\blacktriangle(h'',h'))\}$ for $h', h'' \in \mathcal{F}$. As above, we denote this optimal solution by $\beta \triangleq \beta(\overset{\text{correct}}{\hat{\mu}_{h'}}(\blacktriangle(h'',h')))$, and the associated true generalization error is then $\mathcal{E}_\mu(h_{\text{true}};\beta)$. For a given triple (λ,μ_+,μ_-), and power η of the adversary, not all ambiguity sets $\mathcal{F}$, and values for $\{\overset{\text{correct}}{\hat{\mu}_{h'}}(\blacktriangle(h'',h'))\}$ are attainable. We define the set of such attainable values.

Definition 10 *Let $\mathcal{A}$ be the set of values $\{\overset{\text{correct}}{\hat{\mu}_{h'}}(\blacktriangle(h'',h'))\}$, for $h', h'' \in \mathcal{F}$ for some $\mathcal{F}$, such that there exists a triple $(\hat{\lambda},\hat{\mu}_+,\hat{\mu}_-)$ that meets three conditions:*

(a) *$\mathcal{F}$ must be the ambiguity set corresponding to $(\hat{\lambda},\hat{\mu}_+,\hat{\mu}_-)$, as in (3.4).*

(b) *The triple must satisfy*

$$\begin{aligned}\overset{\text{wrong}}{\hat{\mu}_h}(\blacktriangle(h',h'')) &= \hat{\lambda}\hat{\mu}_+(\blacktriangle(h',h'')\cap h(-)) + \\ &\quad (1-\hat{\lambda})\hat{\mu}_-(\blacktriangle(h',h'')\cap h(+)) \\ \overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h'')) &= \hat{\mu}(\blacktriangle(h',h'')) - \overset{\text{wrong}}{\hat{\mu}_h}(\blacktriangle(h',h'')).\end{aligned}$$

(c) *We must have $||(\lambda,\mu_+,\mu_-)-(\hat{\lambda},\hat{\mu}_+,\hat{\mu}_-)||_{TV} \le \eta$.*

Lemma 11 **(a)** *The set $\mathcal{A}$ is a finite union of polyhedral sets, and it is compact.*

(b) *The function $\mathcal{E}_\mu(h_{\text{true}};\beta(\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h''))))$ is piecewise continuous, with finitely many discontinuities.*

We defer the proof of this lemma to [CM08]. The next proposition gives the optimal policy of the adversary for $S2$:

Proposition 12 *The adversary's optimal strategy T_2^* maps (λ,μ_+,μ_-) to a triple $(\hat{\lambda},\hat{\mu}_+,\hat{\mu}_-)$ that matches the values $\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h''))$ from the solution to the (nonlinear) program:*

$$\begin{aligned}\max: &\quad \mathcal{E}_\mu(h_{\text{true}};\beta(\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h'')))) \\ \text{s.t.}: &\quad \{\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h''))\} \in \mathcal{A}.\end{aligned} \tag{3.12}$$

PROOF. By Lemma 11, $\mathcal{A}$ is compact, and $\mathcal{E}_\mu(h_{\text{true}};\beta)$ is piecewise continuous. Therefore, the optimal value is attained for some ambiguity set $\mathcal{F}$, and corresponding element $\{\overset{\text{correct}}{\hat{\mu}_h}(\blacktriangle(h',h''))\}$ of $\mathcal{A}$. By the definition of $\mathcal{A}$, there exists at least one such map T_2^*, with $||T_2^*|| \le \eta$, that attains this value. □

Thus the optimal policies for the decision-maker and adversary are each given by respective optimization problems. In summary, we have:

Theorem 13 *The pair of strategies (D_i^*, T_i^*) $(i = 1, 2)$ for the decision-maker and the adversary, gives optimal solutions to $S1$ and $S2$, respectively.*

4 Error and the Power of the Adversary

While we treat the noise as generated by an adversary, we may also consider it to be a design parameter chosen according to how we care to trade off optimality for robustness. Indeed, upon seeing some realization $(\hat{\lambda},\hat{\mu}_+,\hat{\mu}_-)$, the decision-maker may have partial knowledge of the level η of noise. Equally, the decision-maker may specifically be interested in choosing a solution appropriate for some particular level $\tilde{\eta}$ of noise. For any fixed level $\tilde{\eta}$, from the results in Section 3, the decision-maker obtains the resulting optimal policy. When $\tilde{\eta} = 0$, the optimal strategy of the decision-maker is to deterministically choose the single classifier that minimizes the empirical error. If indeed $\eta = 0$, then this is the optimal strategy. As $\tilde{\eta}$ grows, the optimal strategy of the decision-maker becomes increasingly random, and in the limit as $\tilde{\eta} \to 100\%$, the optimal policy approaches the uniform distribution over all classifiers.

For a fixed measure μ, $\mathcal{H}$, and $h_{\text{true}} \in \mathcal{H}$ we consider the error as a function of η. Graphing this function allows the decision-maker, in the scenario described above, to consider the tradeoff of robustness and optimality, and thus may choose the desirable design parameter $\tilde{\eta}$, with respect to which the optimal mixed strategy is obtained. In addition, this graph provides other information that is of interest. The graph of the error is not continuous. Rather, it is piecewise continuous (not necessarily linear), with certain break points. The location of these break points is important, and it is a function of the structure of $\mathcal{H}$. A particular solution α of the decision-maker might be optimal for any $\tilde{\eta}$ in some interval $[\eta_1,\eta_2)$, but not optimal for $\tilde{\eta} \ge \eta_2$.

We consider the example from the end of Section 2.3 where h_1 is the true classifier. There, the move-and-flip adversary is strictly more powerful than the flip-only adversary when $\eta > 5$, and hence the setups $S1$ and $S2$ are not equivalent. The graphs in Figure 2 show $\text{Error}^i(\mu, h_{\text{true}}, \eta, T_i^*, D_i^*)$ for fixed μ and h_{true}, and varying values of η. In the left side of Figure 2 we have the superimposed graphs for this example, for $S1$ and $S2$ for $0 \le \eta \le 11$. In the right side of Figure 2 we show the full graph of the true error Error^2, for $0 \le \eta \le 100$.

The graph for $S2$ is obtained by using the results of Propositions 6 and 12. The optimal policy of the move-and-flip adversary differs for the three regions $0 \le \eta < 5$, $5 \le \eta \le 10$, $10 \le \eta \le 100$. In the first region, the adversary is powerless regardless of his action. In the second region, the optimal strategy is to flip $\eta\%$ of the labels in $\blacktriangle(h_1,h_2)$. For $10 \le \eta \le 100$, the adversary's optimal strategy is to flip all the points in $\blacktriangle(h_1,h_2)$, and also move and label '$-$' a $(\eta - 10)$ fraction of the mass into $\blacktriangle(h_1,h_2)$, so that $\hat{\mu}(\blacktriangle(h_1,h_2)) = \eta$.

The decision-maker's policy, as given by Proposition 6, protects the decision-maker against the worst possible (consistent) triple $(\tilde{\lambda},\tilde{\mu}_+,\tilde{\mu}_-)$. Solving the robust LP from the proposition reveals both the true error, and the worst-case error. Both of these quantities may be of interest. In [CM08] we show, for this example, both the true error, and the worst-case error, for all values of η. The true error exhibits numerous interesting properties. For instance, as shown in the figure, the true error is *not monotonic* in the power of the

adversary (the worst-case error over measures and classifiers is, of course, monotonic). This is a direct consequence of Proposition 6. In [CM08] we pay particular attention to this, and other properties of the graph. Also, we give the details of the computations.

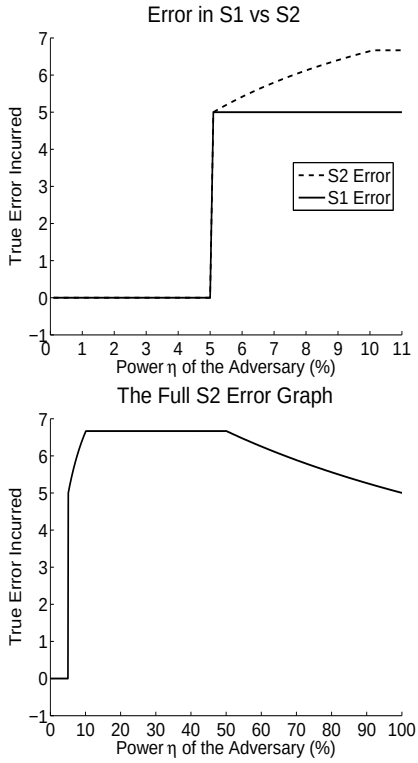

Figure 2: The graph above shows the error incurred in $G1$ on the same axes as the error incurred in $S2$, for $0 \leq \eta \leq 11$. As soon as $\eta > 5$, we see that the move-and-flip adversary is more powerful. Note that Error^2 grows sublinearly for $\eta \geq 5$. In the graph on the right we show the error graph for the more powerful adversary for $0 \leq \eta \leq 100$. The true error is not monotonic, as it decreases (non-linearly) for $\eta \geq 50\%$.

5 Discussion

This work takes a learning in the information theoretic limit view of learning with adversarial disturbance. Our main contribution is the introduction of an optimization-theoretic algorithmic framework for finding a classifier in the presence of such disturbance. We characterized the optimal policy of the decision-maker (as a function of the data) in terms of a tractable and easily solved optimization problem. This is a first step in developing the theory for a range of setups. For example, the Bayesian setup may be of interest. Here, the decision-maker has a prior over the possible classifiers, and instead of minimizing generalization error with respect to the worst-case consistent classifier and (in $S2$) underlying measure $\tilde{\mu}$, he considers minimizing expected (under the Bayesian posterior) error. Extending this algorithmic approach to the game-theoretic setup, where the decision-maker plays against a rational adversary, is also of interest, and allows the possibility of more complex information structures.

Considering the noise level as a design parameter and viewing the resulting error as a function of it yielded surprising results that show how counterintuitive the mini-max formulation of learning with adversarial noise could be. We showed for a simple example that while the worst-case error is monotone in the power of the adversary, the actual error (which depends on the particular underlying true probability measure) may not be monotone in the power of the adversary! This is because even though the adversary is more powerful, the decision maker is also better prepared.

There are three natural extensions to our work that we did not pursue here mostly due to space limits. First, while we considered the proper learning setup, the non-proper setup (as in [KSS92]) seems to naturally follow our framework. Second, the case of infinite set of classifier $\mathcal{H}$ could be resolved by eliminating classifiers that are "close" according to the observed measure. This is particularly useful for the flip-only setup where the adversary cannot make two classifiers substantially different. Finally, while we do not consider sample complexity, such results should not be too difficult to derive by imitating the arguments in [CBDF$^+$99].

References

[ACB98] P. Auer and N. Cesa-Bianchi. On-line learning with malicious noise and the closure algorithm. *Annals of AI and Mathematics*, 23(1):83–99, 1998.

[BEK02] N. H. Bshouty, N. Eiron, and E. Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[BTN99] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, August 1999.

[CBDF$^+$99] N. Cesa-Bianchi, E. Dichterman, P. Fischer, E. Shamir, and H. Ulrich Simon. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 46(5):684–719, 1999.

[CM08] C. Caramanis and S. Mannor. Beyond PAC: A robust optimization approach for learning in the presence of noise: Online appendix. Available from http://users.ece.utexas.edu/˜cmcaram/pubs/RobustLearningOnlineApp.pdf, 2008.

[KL93] M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

[KSS92] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. In *Computational Learing Theory*, pages 341–352, 1992.

[Lai88] P. D. Laird. *Learning from good and bad data*. Kluwer Academic Publishers, Norwell, MA, USA, 1988.

[Ser03] R. Servedio. Smooth boosting and learning with malicious noise. *Journal of Machine Learning Research*, 4:633–648, 2003.