MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Robust Machine Learning via Privacy/Rate-Distortion Theory

Wang, Ye; Aeron, Shuchin; Rakin, Adnan S; Koike-Akino, Toshiaki; Moulin, Pierre

TR2021-082 July 12, 2021

Abstract

Robust machine learning formulations have emerged to address the prevalent vulnerability of deep neural networks to adversarial examples. Our work draws the connection between optimal robust learning and the privacy-utility tradeoff problem, which is a generalization of the rate-distortion problem. The saddle point of the game between a robust classifier and an adversarial perturbation can be found via the solution of a maximum conditional entropy problem. This information-theoretic perspective sheds light on the fundamental tradeoff between robustness and clean data performance, which ultimately arises from the geometric structure of the underlying data distribution and perturbation constraints.

IEEE International Symposium on Information Theory (ISIT)

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Robust Machine Learning via Privacy/Rate-Distortion Theory

Ye Wang*, Shuchin Aeron†, Adnan Siraj Rakin‡, Toshiaki Koike-Akino*, Pierre Moulin§
*Mitsubishi Electric Research Laboratories, †Tufts University,
‡Arizona State University, §University of Illinois at Urbana-Champaign
*{yewang, koike}@merl.com, †shuchin@ece.tufts.edu, ‡asrakin@asu.edu, §pmoulin@illinois.edu

***Abstract*—Robust machine learning formulations have emerged to address the prevalent vulnerability of deep neural networks to adversarial examples. Our work draws the connection between optimal robust learning and the privacy-utility tradeoff problem, which is a generalization of the rate-distortion problem. The saddle point of the game between a robust classifier and an adversarial perturbation can be found via the solution of a maximum conditional entropy problem. This information-theoretic perspective sheds light on the fundamental tradeoff between robustness and clean data performance, which ultimately arises from the geometric structure of the underlying data distribution and perturbation constraints.**

***Index Terms*—robust learning, adversarial examples, privacy**

I. Introduction

The widespread susceptibility of neural networks to adversarial examples [1], [2] has been demonstrated through a wide variety of practical attacks [3]–[9]. This has motivated much research towards mitigating these vulnerabilities, although many earlier defenses have been shown to be ineffective [10]–[12]. We focus our attention on robust learning formulations that aim for guaranteed resiliency against the worst-case input perturbations or in a distributional sense. Our work draws the information-theoretic connections between optimal robust learning and the privacy-utility tradeoff problem. We utilize this perspective to shed light on the fundamental tradeoff between robustness and clean data performance, and to inspire novel algorithms for optimizing robust models.

The influential approach of [13] proposes the robust optimization formulation given by

$$\min_{\theta} \mathbb{E}_{P_{X,Y}} \Big[\max_{\delta \in \mathcal{S}} \ell(f_\theta(X+\delta), Y) \Big],$$

where δ represents the worst-case over some set $\mathcal{S}$ of small perturbations applied to the original input X of the model f_θ, with the aim of maximizing the loss ℓ with respect to the true label Y. This formulation has inspired a plethora of defenses: some that tackle the problem directly (albeit with limitations to scalability) [14]–[18] and others that employ approximate bounding [19]–[23] or noise injection [24]–[26] to certify robustness guarantees.

We generalize this formulation to allow stronger adversaries that may employ mixed strategies, where the perturbation can be viewed as a channel $P_{Z|X,Y}$, while focusing our study on the fundamental optimum of the ideal robust classification game. With the minimization over all decision rules $q(Y|Z)$ for the cross-entropy loss objective, we show the following minimax result that reduces the problem to a maximum conditional entropy problem,

$$\min_{q(Y|Z)} \max_{P_{Z|X,Y} \in \mathcal{D}} \mathbb{E}[-\log q(Y|Z)]$$
$$= \max_{P_{Z|X,Y} \in \mathcal{D}} \min_{q(Y|Z)} \mathbb{E}[-\log q(Y|Z)] = \max_{P_{Z|X,Y} \in \mathcal{D}} H(Y|Z).$$

This minimax result is established in Theorems 1 and 2 in terms of the more general notion of distributional robustness, which considers the worst-case data distribution over some convex set $\mathcal{D}$. This subsumes expected distortion constraints as a special case when $\mathcal{D}$ is a Wasserstein-ball with a suitably chosen ground metric. Due to space constraints, the proof of our main result is presented in an extended version of this paper [27], along with another result that gives, for the maximum conditional entropy problem over a Wasserstein-ball constraint, a fixed point characterization, which exposes the interplay between the geometry of the ground cost in the Wasserstein-ball constraint, the worst-case adversarial distribution, and the given reference data distribution.

The minimax equality establishes the connection to the privacy-utility tradeoff problem [28]–[33], where the aim is to design a distortion-constrained data perturbation mechanism $P_{Z|X,Y}$ that maximizes the uncertainty about sensitive information Y as measured by $H(Y|Z)$. The equivalence between the maximin problem and maximum conditional entropy is used by [29] to argue that conditional entropy measures privacy against an inference attacker represented by q. Figure 1 illustrates these connections.

A similar minimax result is given in [34], however with technical limitations preventing it from addressing adversarial input perturbation, and much of their development focuses on the case where the marginal distribution for X remains fixed. The similarities between the robust learning and privacy problems are noted by [35], however, they only state the minimax inequality relating the two.

We examine the fundamental tradeoff between model robustness and clean data performance from our information-theoretic perspective. This tradeoff ultimately arises from the geometric structure of the underlying data distribution and the adversarial perturbation constraints. We illustrate these tradeoffs with the numerical analysis of a toy example. The fundamental tradeoff between clean data and adversarial loss is also theoretically addressed by [36]. This theory was further

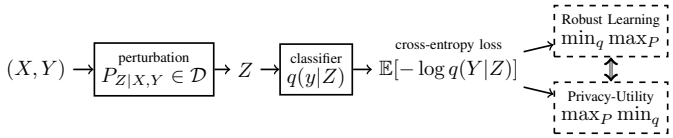

Fig. 1. Robust Learning and Privacy-Utility Tradeoff problems both involve a game between a classifier and a constrained input perturbation. The goal of robust learning is a classifier robust to the perturbation, and posed as a minimax problem. The alternative maximin optimization captures the privacy-utility tradeoff problem, where the goal is a perturbation mechanism that hides sensitive information from an adversarial classifier aiming to recover it. Our minimax result shows that these two problems are equivalent.

expanded upon by [37] and leveraged to develop an improved adversarial training defense.

Notation: We use $\mathcal{P}(\mathcal{Z}|\mathcal{X},\mathcal{Y})$ to denote the set of conditional probability distributions over $\mathcal{Z}$ given variables over the sets $\mathcal{X}$ and $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ is similarly defined.

II. Robust Machine Learning

The influential robust learning formulation of [13] addresses the worst-case attack, as given by

$$\min_{\theta} \mathbb{E}\left[\max_{\substack{Z\in\mathcal{X}:\\ d(X,Z)\le\epsilon}} \ell(f_\theta(Z),Y)\right], \tag{1}$$

where $d:\mathcal{X}\times\mathcal{X}\to[0,\infty]$ is some suitably chosen distortion metric (e.g., often ℓ_0, ℓ_p, or ℓ_∞ distance), and $\epsilon\ge 0$ represents the allowable perturbation. The robust learning formulation in (1) can be viewed as a two-player, zero-sum game, where the adversary (corresponding to the inner maximization) plays second using a pure strategy by picking a fixed Z subject to the distortion constraint. We will instead consider an adversary that utilizes a mixed strategy, where $Z\in\mathcal{X}=:\mathcal{Z}$ can be a randomized function of (X,Y) as specified by a conditional distribution $P_{Z|X,Y}$. This revised formulation is given by

$$\min_{\theta} \max_{P_{Z|X,Y}\in\mathcal{D}^*_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z),Y)], \tag{2}$$

where the expectation is over $(X,Y,Z)\sim P_{X,Y}P_{Z|X,Y}$, and the distortion limit is given by

$$\mathcal{D}^*_{d,\epsilon} := \{P_{Z|X,Y}\in\mathcal{P}(\mathcal{Z}|\mathcal{X},\mathcal{Y}) : \Pr[d(X,Z)\le\epsilon]=1\}. \tag{3}$$

Note that under this maximum distortion constraint, allowing mixed strategies does not actually strengthen the adversary, i.e., the games in (1) and (2) have the same value. However, if we replace the distortion limit constraint of (3) with an average distortion constraint, given by

$$\mathcal{D}_{d,\epsilon} := \{P_{Z|X,Y}\in\mathcal{P}(\mathcal{Z}|\mathcal{X},\mathcal{Y}) : \mathbb{E}[d(X,Z)]\le\epsilon\}, \tag{4}$$

then the adversary is potentially strengthened, i.e.,

$$\max_{P_{Z|X,Y}\in\mathcal{D}_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z),Y)] \ge \max_{P_{Z|X,Y}\in\mathcal{D}^*_{d,\epsilon}} \mathbb{E}[\ell(f_\theta(Z),Y)].$$

A. Distributional Robustness

Since the objective $\mathbb{E}[\ell(f_\theta(Z),Y)]$ only depends on the joint distribution of the variables $(Z,Y)\in\mathcal{X}\times\mathcal{Y}$, the robust learning formulation is straightforward to generalize by instead considering the maximization over an arbitrary set of joint distributions $\mathcal{D}\subset\mathcal{P}(\mathcal{X},\mathcal{Y})$. With a change of variable (replacing Z with X to simplify presentation), this becomes

$$\min_{\theta} \max_{p\in\mathcal{D}} \mathbb{E}_{(X,Y)\sim p}[\ell(f_\theta(X),Y)], \tag{5}$$

which includes the scenarios considered in (1) through (4) as special cases. However, unlike these earlier formulations, (5) allows for the label Y to also be potentially changed.

Another particular case for $\mathcal{D}$ is the Wasserstein-ball around a distribution $\mu\in\mathcal{P}(\mathcal{X},\mathcal{Y})$, as given by

$$\mathcal{D}^{\mathbb{W}}_{\epsilon}(\mu) := \{\nu\in\mathcal{P}(\mathcal{X},\mathcal{Y}) : \mathbb{W}_d(\mu,\nu)\le\epsilon\}, \tag{6}$$

where $\mathbb{W}_d$ is the 1-Wasserstein distance [38]–[40] for some ground metric (or in general a cost) d on the space $\mathcal{X}\times\mathcal{Y}$. Recall that the 1-Wasserstein distance is given by

$$\mathbb{W}_d(\mu,\nu) := \inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_\gamma\big[d\big((X,Y),(X',Y')\big)\big],$$

where the set of couplings $\Gamma(\mu,\nu)$ is defined as all joint distributions with the marginals $(X,Y)\sim\mu$ and $(X',Y')\sim\nu$. Note that maximizing over $p\in\mathcal{D}^{\mathbb{W}}_{\epsilon}(P_{X,Y})$ is equivalent to maximizing over channels $P_{X',Y'|X,Y}$ subject to the distortion expected constraint $\mathbb{E}\big[d\big((X,Y),(X',Y')\big)\big]\le\epsilon$, where $(X,Y,X',Y')\sim P_{X,Y}P_{X',Y'|X,Y}$. Unlike the formulation considered in (2), this channel may also change the label Y. However, if modifying Y is prohibited by a cost of the form

$$d\big((x,y),(x',y')\big) = \begin{cases} d(x,x'), & \text{if } y=y',\\ \infty, & \text{otherwise,}\end{cases} \tag{7}$$

then the 1-Wasserstein distributionally robust formulation specializes to the earlier formulation in (2) with the average distortion constraint given by (4). Robust-ML with Wasserstein-ball constraints is also referred to as Distributional Robust Optimization (DRO) [41]–[43] and shown to be equivalent to imposing Lipschitz constraints on the classifier [43], [44]. There is however no characterization, that is considered in these papers, of the optimal value of the min-max problem in this setting.

B. Optimal Robust Learning

The specifics of the loss function ℓ and model f_θ are crucial to our analysis. Hence, we will focus specifically on learning classification models, where $X\in\mathcal{X}$ represents the data features, $Y\in\mathcal{Y}:=\{1,\dots,m\}$ represent class labels, and the model $f_\theta:\mathcal{X}\to[0,1]^m$ can be viewed as producing $q_\theta\in\mathcal{P}(\mathcal{Y}|\mathcal{X})$ that aims to approximate the underlying posterior $P_{Y|X}$. When cross-entropy is the loss function, i.e.,

$\ell(f_\theta(X), Y) = -\log q_\theta(Y|X)$, the expected loss, with respect to some distribution $(X, Y) \sim p = P_X P_{Y|X}$, is given by

$$\begin{aligned}
&\mathbb{E}_p[-\log q_\theta(Y|X)] \\
&\quad = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} P_{Y|X}(y|x) \log \frac{P_{Y|X}(y|x)}{q_\theta(y|x) P_{Y|X}(y|x)} \, \mathrm{d}P_X(x) \\
&\quad = \mathrm{KL}(P_{Y|X}(y|X) \| q_\theta(y|X) | P_X) + H(Y|X). \qquad (8)
\end{aligned}$$

Thus, the principle of learning via minimizing the expected cross-entropy loss optimizes the approximate posterior $q_\theta(y|x)$ toward the underlying posterior $P_{Y|X}$, and the loss is lower bounded by the conditional entropy $H(Y|X)$, which is arguably nonzero for nontrivial classification problems.

The robust learning problem, given by

$$\min_\theta \max_{p \in \mathcal{D}} \mathbb{E}_{(X,Y)\sim p}[-\log q_\theta(Y|X)], \qquad (9)$$

still critically depends on the specific parametric family (e.g., neural network architecture) chosen for the model $\{f_\theta\}_{\theta \in \Theta}$, which determines the corresponding parametric family of approximate posteriors, i.e., $\{q_\theta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})\}_{\theta \in \Theta}$. Motivated by the ultimate meta-objective of learning the best possible robust models, we consider the idealized optimal robust learning formulation where the minimization is performed over all conditional distributions $q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})$, as given by

$$\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \max_{p \in \mathcal{D}} \mathbb{E}_{(X,Y)\sim p}[-\log q(Y|X)], \qquad (10)$$

which clearly lower-bounds (9), which is specific to the particular parametric family.

III. The Privacy-Utility Tradeoff Problem

In the information-theoretic treatment of the privacy-utility tradeoff problem [28]–[33], the random variables $(X, Y) \sim P_{X,Y}$ respectively denote useful and sensitive data, and the goal is to release data Z produced from a randomized algorithm viewed as a channel $P_{Z|X,Y}$, while simultaneously preserving the privacy of the sensitive Y and maintaining utility by conveying X. Privacy is measured by $I(Y; Z)$, where smaller is better to preserve privacy. Utility is quantified with a distortion function, $d : \mathcal{X} \times \mathcal{Z} \to [0, \infty)$, given by the particular application. Minimizing (or limiting) the distortion $d(X, Z)$ captures the objective of maintaining the utility of the data release. Since the useful and sensitive data (X, Y) are correlated (and indeed the problem is uninteresting if they are independent), a tradeoff naturally emerges between the two objectives of preserving privacy and utility.

A. Optimal Privacy-Utility Tradeoff

The optimal privacy-utility tradeoff problem is formulated as an information-theoretic optimization problem in [28], [29], and is given by

$$\operatorname*{arg\,min}_{P_{Z|X,Y} \in \mathcal{D}_{d,\epsilon}} I(Y; Z) = \operatorname*{arg\,max}_{P_{Z|X,Y} \in \mathcal{D}_{d,\epsilon}} H(Y|Z), \qquad (11)$$

where $(X, Y, Z) \sim P_{X,Y} P_{Z|X,Y}$, the constraint $\mathcal{D}_{d,\epsilon}$, as given in (4), captures the expected distortion budget, and the equivalence follows from $I(Y; Z) = H(Y) - H(Y|Z)$ since $H(Y)$ is constant. Similarly, one could consider the alternative maximum distortion constraint $\mathcal{D}^*_{d,\epsilon}$, given in (3).

B. Adversarial Formulation of Privacy

In [29], the privacy-utility problem in (11), is derived from a broader perspective that poses privacy as maximizing the loss of an adversary that mounts a statistical inference attack attempting to recover the sensitive Y from the release Z. Their framework considers an adversary that can observe the release Z and choose a conditional distribution $q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})$ to minimize its expected loss. As observed in [29], when cross-entropy (or "self-information") is the loss, we have that

$$\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}[-\log q(Y|Z)] = H(Y|Z), \qquad (12)$$

with the optimum $q^* = p_{Y|Z}$, which follows from a derivation similar to (8). Thus, the optimal privacy-utility tradeoff given in (11) is equivalent to a maximin problem, as stated in Lemma 1.

Lemma 1 (equivalence of privacy formulations [29])**.** *For any joint distribution $P_{X,Y}$ and closed, convex constraint set $\mathcal{D} \subset \mathcal{P}(\mathcal{Z}|\mathcal{X}, \mathcal{Y})$, e.g., $\mathcal{D}^*_{d,\epsilon}$ or $\mathcal{D}_{d,\epsilon}$, as given by* (3) *or* (4)*, we have*

$$\begin{aligned}
&\max_{P_{Z|X,Y} \in \mathcal{D}} \min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}[-\log q(Y|Z)] \\
&\quad = \max_{P_{Z|X,Y} \in \mathcal{D}} H(Y|Z) = H(Y) - \min_{P_{Z|X,Y} \in \mathcal{D}} I(Y; Z),
\end{aligned}$$

where $(X, Y, Z) \sim P_{X,Y} P_{Z|X,Y}$.

The privacy-utility tradeoff problem is also highly related to rate-distortion theory, which considers the efficiency of lossy data compression. When $X = Y$, the optimization problem in (11) immediately reduces to the single-letter characterization of the optimal rate-distortion tradeoff. However, the privacy problem considers an inherently single-letter scenario, where we deal with just a single instance of the variables (X, Y, Z), which could be high-dimensional, but have no restrictions placed on their statistical structure across these dimensions.

IV. Main Results – Duality between Optimal Robust Learning and Privacy-Utility Tradeoffs

The solution to the optimal minimax robust learning problem can be found via a maximum conditional entropy problem related to the privacy-utility tradeoff problem.

Theorem 1. *For any finite sets $\mathcal{X}$ and $\mathcal{Y}$, and closed, convex set of joint distributions $\mathcal{D} \subset \mathcal{P}(\mathcal{X}, \mathcal{Y})$, we have*

$$\begin{aligned}
&\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \max_{p \in \mathcal{D}} \mathbb{E}[-\log q(Y|X)] && (13) \\
&\quad = \max_{p \in \mathcal{D}} \min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \mathbb{E}[-\log q(Y|X)] && (14) \\
&\quad = \max_{p \in \mathcal{D}} H(Y|X) =: h^* \le \log |\mathcal{Y}|, && (15)
\end{aligned}$$

where the expectations and entropy are with respect to $(X,Y) \sim p$*. Further, the solutions for* $q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ *that minimize* (13) *are given by*

$$\bigcap_{p\in\mathcal{D}} \big\{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \mathbb{E}_{(X,Y)\sim p}[-\log q(Y|X)] \le h^*\big\} \neq \varnothing. \tag{16}$$

Proof. See Appendix in the extended version [27]. □

Intuitively, the optimal minimax robust decision rule q that solves (13) must be consistent with the posterior $p(y|x)$ corresponding to the solution of the maximum conditional entropy problem in (15). However, a given posterior $p(y|x)$ is well-defined only over the support of the marginal distribution of X, whereas the robust decision rule needs to be defined over the entire space $\mathcal{X}$. Hence, generally, determining the robust decision rule over the entirety of $\mathcal{X}$ requires considering the solution set in (16), which seems cumbersome, but can be simplified in many cases via the following corollary.

Corollary 1. *Under the paradigm of Theorem 1, let*

$$\mathcal{D}^* := \big\{p \in \mathcal{D} : H(Y|X) = h^*, (X,Y) \sim p\big\}.$$

For all $p^* \in \mathcal{D}^*$*, the corresponding terms of* (16) *are given by*

$$\begin{aligned} Q(p^*) &:= \big\{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \mathbb{E}_{(X,Y)\sim p^*}[-\log q(Y|X)] \le h^*\big\} \\ &= \big\{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X}) : \forall (x,y), q(y|x)p^*(x) = p^*(x,y)\big\}. \end{aligned}$$

Further, if

$$\bigcup_{p^*\in\mathcal{D}^*} \big\{x \in \mathcal{X} : p^*(x) > 0\big\} = \mathcal{X},$$

then the solution set given by (16)*, for the minimization of* (13)*, contains exactly one point and is given by*

$$\bigcap_{p^*\in\mathcal{D}^*} Q(p^*) = \bigcap_{p\in\mathcal{D}} Q(p).$$

In the simplest case, if there exists a $p^* \in \mathcal{D}^*$ that has full support over $\mathcal{X}$ (in the marginal distribution for X), then the optimal robust decision rule that solves the minimization of (13) is simply given by the posterior $p^*(y|x)$, which is defined for all $x \in \mathcal{X}$.

A. Generalization to Arbitrary Alphabets

Extending the result in the previous section to continuous $\mathcal{X}$ requires one to expand the set of allowable Markov kernels, i.e., conditional probabilities, to what is referred to as the set of generalized decision rules in statistical decision theory [45]–[48]. This is because the set of Markov kernels is not compact, while the set of generalized decision rules is. For any $f \in C_b(\mathcal{Y})$, set of bounded continuous functions, and any bounded signed measure φ on $\mathcal{X}$, given a mapping $q(Y|X)$ (interpret this as a measurable function q_x over $\mathcal{Y}$ for each fixed x), define a bilinear functional via,

$$\beta_{q(Y|X)}(f,\varphi) = \int_{\mathcal{X}}\int_{\mathcal{Y}} f(y)q(dy|dx)\,\mathrm{d}\varphi(x). \tag{17}$$

Definition 1. [45] A generalized decision function is a bilinear function $\beta : C_b(\mathcal{Y}) \times \varphi \to \mathbb{R}$ that satisfies, (a) if $f \ge 0, \varphi \ge 0 \implies \beta(f,\varphi) \ge 0$, (b) $|\beta(f,\varphi)| \le \|f\|_\infty \|\varphi\|_{TV}$, (c) $\beta(1,\varphi) = \|\varphi\|_{TV}$ if $\varphi \ge 0$.

Define the set of generalized decision rules as the set of bilinear functions defined via (17) and satisfying the properties (a), (b), (c) above.

$$\mathcal{M} = \{q(Y|X) : q(Y|X) \text{ satisfies a. b. c. in Def. 1 via (17)}\}$$

Applying these results, we obtain the following theorem for the case of general alphabets $\mathcal{X}$. Note that in contrast to Theorem 1, here the results hold with $\inf, \sup$ instead of $\min, \max$.

Theorem 2. *Under the paradigm of Theorem 1, for continuous alphabets* $\mathcal{X}$ *and discrete* $\mathcal{Y}$*,*

$$\inf_{q\in\mathcal{M}} \sup_{p\in\mathcal{D}} \mathbb{E}_p[-\log q(Y|X)] = \sup_{p\in\mathcal{D}} H(Y|X) \tag{18}$$

Proof. Using the fact that the set $\mathcal{M}$ is convex and compact for the weak topology (Theorem 42.3, [45]), that the function $\mathbb{E}_p[-\log q(Y|X)]$ is convex in q for all $q \in \mathcal{M}$, and applying the minimax theorem [49], we have

$$\inf_{q\in\mathcal{M}} \sup_{p\in\mathcal{D}} \mathbb{E}_p[-\log q(Y|X)] = \sup_{p\in\mathcal{D}} \inf_{q\in\mathcal{M}} \mathbb{E}_p[-\log q(Y|X)], \tag{19}$$

and noting that $\inf_{q\in\mathcal{M}} \mathbb{E}_p[-\log q(Y|X)] = H(Y|X)$, the result follows. Hence, even in the case of continuous alphabets, the worst case *algorithm-independent* adversarial perturbation can be found by solving $\sup_{p\in\mathcal{D}} H(Y|X)$. □

V. Implications of the Main Results

A. Necessity of Stochastic Perturbation

In the original robust learning formulation, as given in (1), the attacker is restricted to a pure strategy, and this is not suboptimal (i.e., this game has the same value as the mixed strategy formulation given by (2)), since the attacker has the advantage of "playing second" with the inner maximization. However, we emphasize that the original formulation given by (1), even in the basic case of optimal robust classification, is not necessarily a saddle point problem, that is,

$$\min_{q\in\mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\left[\max_{\substack{Z\in\mathcal{X}:\\ d(X,Z)\le\epsilon}} -\log q(Y|Z)\right] \tag{20}$$

$$\ge \max_{\substack{g:\mathcal{X}\times\mathcal{Y}\to\mathcal{X}\\ d(X,g(X,Y))\le\epsilon}} \min_{q\in\mathcal{P}(\mathcal{Y}|\mathcal{Z})} \mathbb{E}\Big[-\log q\big(Y|g(X,Y)\big)\Big] \tag{21}$$

will often be a strict inequality due to the determinism of the attack mapping g. In contrast, our minimax result of Theorem 1 establishes that with a stochastic attacker (or, more generally, distributional robustness constrained to a convex set), such as formulated in (2), swapping the min and max does not disadvantage the attacker for "playing first".

We illustrate the necessity of a stochastic attacker with the following example. Consider $\mathcal{X} = \mathcal{Y} = \{0,1,2,3,4\}$, where $P_{X,Y}(x,y) = 1/3$ for $(x,y) \in \{(0,0),(2,2),(4,4)\}$, and let

$\epsilon = 1$ be the distortion limit under the metric $d(x,z) = |x-z|$. For this setup, the optimal stochastic attack will clearly lie within the family parameterized by $\alpha \in [0,1]$ and given by

$$p^{\alpha}_{Z|X}(z|x) := \begin{cases} 1, & \text{if } (x,z) \in \{(0,1),(4,3)\}, \\ \alpha, & \text{if } (x,z) = (2,1), \\ 1-\alpha, & \text{if } (x,z) = (2,3), \end{cases}$$

however, the optimal deterministic attack is limited to only α equal to zero or one. The optimal stochastic attack that solves (15), and hence also (13) and (14) due to Theorem 1 and Corollary 1, is found at $\alpha = 0.5$ yielding the optimal value of $h^* = h_2(1/3)$, where $h_2(p) := -p\log(p) - (1-p)\log(1-p)$ is the binary entropy function. For deterministic attacks, the optimal value of (20) is also $h_2(1/3)$, however, the optimal value of (21) is equal to $(2/3)\log(2) < h_2(1/3)$.

B. Tradeoffs between Robustness vs Clean Data Loss

A natural question to ask is whether robustness comes at a price. It has been observed empirically that robust models will underperform on clean data in comparison to conventional, non-robust models. To understand why this is fundamentally unavoidable, we examine the loss for robust and non-robust models in combination with clean data or adversarial attack.

Let $\mu \in \mathcal{D}$ denote the unperturbed (clean data) distribution within the set of potential adversarial attacks $\mathcal{D}$. For a given decision rule $q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})$ and distribution $\nu = \nu_X \nu_{Y|X} \in \mathcal{P}(\mathcal{X},\mathcal{Y})$, recall that the cross-entropy loss is given by (8) as

$$\begin{aligned} \mathcal{L}(\nu, q) &:= \mathbb{E}_p[-\log q(Y|X)] \\ &= H_\nu(Y|X) + \mathrm{KL}(\nu_{Y|X} \| q(y|X) | \nu_X). \end{aligned}$$

The baseline loss of the ideal non-robust model for clean data is given by $\min_q \mathcal{L}(\mu, q) = H_\mu(Y|X)$. Under adversarial attack, the ideal loss of the robust model is given by Theorem 1 as

$$\min_q \max_{\nu \in \mathcal{D}} \mathcal{L}(\nu, q) = \max_{\nu \in \mathcal{D}} H_\nu(Y|X).$$

The loss of a robust model q^* that solves (13), as characterized by (16), under the clean data distribution μ is given by

$$\mathcal{L}(\mu, q^*) = H_\mu(Y|X) + \mathrm{KL}(\mu_{Y|X} \| q^*(y|X) | \mu_X).$$

The KL-divergence term must be finite, since we have

$$\begin{aligned} H_\mu(Y|X) &= \min_q \mathcal{L}(\mu, q) \le \mathcal{L}(\mu, q^*) \\ &\le \min_q \max_{\nu \in \mathcal{D}} \mathcal{L}(\nu, q) = \max_{\nu \in \mathcal{D}} H_\nu(Y|X), \end{aligned}$$

where the second inequality follows from q^* being the minimax solution.

We numerically evaluate these tradeoffs by considering a family of Wasserstein-ball constraint sets $\mathcal{D}(\epsilon)$, as given by (6), with varying radius $\epsilon \ge 0$ around a distribution μ over finite alphabets $\mathcal{X} = \mathcal{Y} = \{1, \ldots, 5\}$. The ground metric is of the form given in (7), which effectively limits the perturbation to only changing X within an expected squared-distance distortion constraint of ϵ, as equivalent to (4). The distribution μ was randomly chosen, and has entropies $H_\mu(Y) \approx 1.6$ and $H_\mu(Y|X) \approx 0.34$ (in nats).

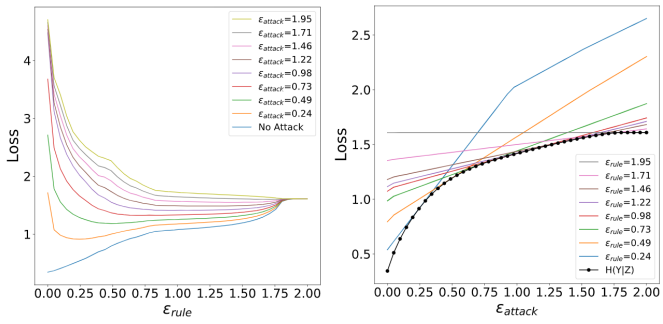

Fig. 2. *Left:* Loss as a function of decision rule, varying ϵ_{rule}, and across attacks varying ϵ_{attack}. *Right:* Loss as a function of attack distortion, varying ϵ_{attack}, and across decision rules varying ϵ_{rule}.

Leveraging Theorem 1 and Corollary 1, we numerically solve for the robust decision rules,

$$q^*_{\epsilon_{\text{rule}}} = \arg\min_{q \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \max_{\nu \in \mathcal{D}(\epsilon_{\text{rule}})} \mathcal{L}(\nu, q),$$

across a range distortion constraints $\epsilon_{\text{rule}} \in [0,2]$. In combination with each decision rule, we consider the loss under attacks at varying distortion limits $\epsilon_{\text{attack}} \in [0,2]$, as given by

$$\mathcal{L}(\epsilon_{\text{attack}}, \epsilon_{\text{rule}}) := \max_{\nu \in \mathcal{D}(\epsilon_{\text{attack}})} \mathcal{L}(\nu, q^*_{\epsilon_{\text{rule}}}).$$

Figure 2 plots the loss $\mathcal{L}(\epsilon_{\text{attack}}, \epsilon_{\text{rule}})$ across the combination of ϵ_{attack} and ϵ_{rule}. On the *left* of Figure 2, each curve is a fixed attack distortion ϵ_{attack}, over which the decision rule $q^*_{\epsilon_{\text{rule}}}$ is varied, with the optimal loss obtained when $\epsilon_{\text{rule}} = \epsilon_{\text{attack}}$. As ϵ_{rule} increases, the loss for all curves converge to $H_\mu(Y)$. In the *right* of Figure 2, the dotted black curve is the maximum conditional entropy $H_\nu(Y|X)$ over $\nu \in \mathcal{D}(\epsilon_{\text{attack}})$ at each ϵ_{attack}, which corresponds to the ideal robust loss when $\epsilon_{\text{rule}} = \epsilon_{\text{attack}}$. The other curves are each a fixed decision rule $q^*_{\epsilon_{\text{rule}}}$, over which the attack distortion ϵ_{attack} is varied, which exhibits suboptimal loss for mismatched $\epsilon_{\text{rule}} \neq \epsilon_{\text{attack}}$. The beginning of each curve, at $\epsilon_{\text{attack}} = 0$, is the clean data loss for each rule, and we can see that clean data loss is degraded as robustness for higher distortions ϵ_{attack} is improved. In the extreme of a decision rule designed to be robust for very high $\epsilon_{\text{rule}} = 1.95$, the loss is uniformly equal to $H_\mu(Y)$ across all ϵ_{attack}, since this robust decision rule $q^*_{1.95}$ only simply guesses the prior μ_Y.

REFERENCES

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

[2] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[3] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1528–1540.

[4] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *ICLR Workshop*, 2017. [Online]. Available: https://arxiv.org/abs/1607.02533

[5] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1765–1773.

[6] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[7] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 284–293.

[8] W. Van Ranst, S. Thys, and T. Goedemé, "Fooling automated surveillance cameras: adversarial patches to attack person detection," in *CVPR Workshop on The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security*, 2019.

[9] J. B. Li, F. R. Schmidt, and J. Z. Kolter, "Adversarial camera stickers: A physical camera attack on deep learning classifier," *arXiv preprint arXiv:1904.00759*, 2019.

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[11] ——, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[12] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 274–283.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1706.06083

[14] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *International Conference on Computer Aided Verification*, 2017, pp. 3–29.

[15] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, "Reluplex: An efficient smt solver for verifying deep neural networks," in *International Conference on Computer Aided Verification*, 2017, pp. 97–117.

[16] R. Ehlers, "Formal verification of piece-wise linear feed-forward neural networks," in *International Symposium on Automated Technology for Verification and Analysis*, 2017, pp. 269–286.

[17] C.-H. Cheng, G. Nührenberg, and H. Ruess, "Maximum resilience of artificial neural networks," in *International Symposium on Automated Technology for Verification and Analysis*, 2017, pp. 251–268.

[18] V. Tjeng, K. Xiao, and R. Tedrake, "Evaluating robustness of neural networks with mixed integer programming," in *International Conference on Learning Representations*, 2019.

[19] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, 2018, pp. 5283–5292.

[20] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," in *Advances in Neural Information Processing Systems*, 2018, pp. 8400–8409.

[21] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *International Conference on Machine Learning*, 2018.

[22] A. Raghunathan, J. Steinhardt, and P. S. Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 877–10 887.

[23] E. Wong, F. R. Schmidt, and J. Z. Kolter, "Wasserstein adversarial examples via projected sinkhorn iterations," *arXiv preprint arXiv:1902.07906*, 2019.

[24] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," *arXiv preprint arXiv:1802.03471*, 2018.

[25] B. Li, C. Chen, W. Wang, and L. Carin, "Second-order adversarial attack and certifiable robustness," *arXiv preprint arXiv:1809.03113*, 2018.

[26] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," *arXiv preprint arXiv:1902.02918*, 2019.

[27] Y. Wang, S. Aeron, A. S. Rakin, T. Koike-Akino, and P. Moulin, "Robust machine learning via privacy/rate-distortion theory," *arXiv preprint arXiv:2007.11693*, 2020.

[28] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623–1636, 2010.

[29] F. d. P. Calmon and N. Fawaz, "Privacy against statistical inference," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2012, pp. 1401–1408.

[30] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[31] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*, 2014, pp. 501–505.

[32] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "Managing your private and public data: Bringing down inference attacks against your privacy," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1240–1255, 2015.

[33] Y. O. Basciftci, Y. Wang, and P. Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *2016 Information Theory and Applications Workshop (ITA)*, 2016, pp. 1–6.

[34] F. Farnia and D. Tse, "A minimax approach to supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 4240–4248.

[35] J. Hamm and A. Mehra, "Machine vs machine: Minimax-optimal defense against adversarial examples," *arXiv preprint arXiv:1711.04368*, 2017.

[36] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *International Conference on Learning Representations*, 2019.

[37] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning*, 2019, pp. 7472–7482.

[38] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Springer, 2015. [Online]. Available: https://www.math.u-psud.fr/~filippo/OTAM-cvgmt.pdf

[39] C. Villani, *Optimal Transport: Old and New*. Springer, Berlin, Heidelberg, 2009.

[40] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11 (5-6), pp. 355–602, 2019. [Online]. Available: https://arxiv.org/abs/1803.00567

[41] J. Blanchet and K. Murthy, "Quantifying Distributional Model Risk Via Optimal Transport," *SSRN Electronic Journal*, 2016.

[42] R. Gao and A. J. Kleywegt, "Distributionally Robust Stochastic Optimization with Wasserstein Distance," *arXiv preprint arXiv:1604.02199*, 2016.

[43] R. Gao, X. Chen, and A. J. Kleywegt, "Wasserstein distributional robustness and regularization in statistical learning," *CoRR*, vol. abs/1712.06050, 2017. [Online]. Available: http://arxiv.org/abs/1712.06050

[44] Z. Cranko, Z. Shi, X. Zhang, R. Nock, and S. Kornblith, "Generalised Lipschitz Regularisation Equals Distributional Robustness," *arXiv preprint arXiv:2002.04197*, 2020.

[45] H. Strasser, *Mathematical theory of statistics: statistical experiments and asymptotic decision theory*. Walter de Gruyter, 2011, vol. 7.

[46] L. LeCam, "An extension of Wald's theory of statistical decision functions," *Ann. Math. Statist.*, vol. 26, no. 1, pp. 69–81, 03 1955. [Online]. Available: https://doi.org/10.1214/aoms/1177728594

[47] L. Cam, *Asymptotic Methods in Statistical Decision Theory*, ser. Springer series in statistics. Springer My Copy UK, 1986. [Online]. Available: https://books.google.com/books?id=BcDxoAEACAAJ

[48] A. v. d. Vaart, "The statistical work of lucien le cam," *Ann. Statist.*, vol. 30, no. 3, pp. 631–682, 06 2002. [Online]. Available: https://doi.org/10.1214/aos/1028674836

[49] D. Pollard, *Asymptopia*. Unpublished manuscript, 2003. [Online]. Available: http://www.stat.yale.edu/~pollard/Courses/602.spring07/MmaxThm.pdf

[50] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill, 1964.