Adversarial Robustness with Semi-Infinite Constrained Learning

Abstract

Despite strong performance in numerous applications, the vulnerability of deep
learning to input perturbations has raised serious questions about its use in critical
domains. While adversarial training has been successful at mitigating this issue in
practice, state-of-the-art methods are increasingly application-dependent, heuristic
in nature, and suffer from fundamental trade-offs between nominal performance
and robustness. Moreover, the problem of finding worst-case perturbations is non-
convex and severely underparameterized, both of which engender a non-favorable
optimization landscape. In this way, there is a gap between the theory and practice
of robust learning, particularly with respect to *when* and *why* adversarial training
works. In this paper, we take a constrained learning approach to address these
questions and provide a theoretical foundation for robust learning. In particular,
we leverage semi-infinite optimization and non-convex duality theory to show
that adversarial training is equivalent to a statistical problem over perturbation
distributions, which we fully characterize under mild conditions. Notably, we show
that a myriad of previous robust training techniques can be recovered for particular,
sub-optimal choices of these distributions. Using these insights, we then propose a
hybrid Langevin Markov Chain Monte Carlo approach, of which several common
algorithms (e.g., PGD) are special cases. Finally, we also use our framework to
limit the nominal performance degradation with generalization guarantees, yielding
comparable and, in some cases, improved results over state-of-the-art benchmarks.

1 Introduction

Learning is at the core of many modern information systems, with wide-ranging applications in
clinical research [1–4], smart grids [5–7], and robotics [8–10]. However, it has become clear that
learning-based solutions suffer from a critical lack of robustness [11–17], leading to models that
are vulnerable to malicious tampering and unsafe behavior [18–22]. While robustness has been
studied in statistics for decades [23–25], this issue has been exacerbated by the opacity, scale, and
non-convexity of modern learning models, such as convolutional neural network (CNNs), making
their behavior difficult to analyze. Thus, designing models which are robust to malicious attacks,
such as imperceptible perturbations, has become a critical challenge in machine learning (ML).

The pernicious nature of these vulnerabilities has led to rapidly-growing interest in improving the
so-called *adversarial robustness* of modern ML models. To this end, numerous approaches have
been proposed based on tools from fields such as distributionally robust optimization [26–28] and
statistical smoothing [29–31]. Yet a great deal of empirical evidence has shown that *adversarial
training* is the most effective way to obtain robust classifiers, wherein models are trained on perturbed
samples rather than directly on clean data [32–37]. And while this approach is now ubiquitous in
practice, adversarial training faces two fundamental challenges.

Firstly, it is well-known that obtaining *worst-case*, adversarial perturbations of data is challenging in the context of deep neural networks (DNNs) [38, 39]. While gradient-based methods have been shown to be empirically effective at finding perturbations that lead to misclassification, there are no guarantees that these perturbations are truly worst-case due to the non-convexity of most commonly-used ML function classes [40]. Moreover, whereas optimizing the parameters of a CNNs is typically an overparameterized problem, finding worst-case perturbations is severely underparametrized and therefore does not enjoy the benign optimization landscape of standard training [41–45]. For this reason, state-of-the-art adversarial attacks increasingly rely on heuristics such as random initializations, multiple restarts, pruning, and other *ad hoc* training procedures [46–55].

The second challenge faced by adversarial training is that it engenders a fundamental trade-off between robustness and nominal performance [56–58]. In practice, penalty-based methods that incorporate clean data into the training objective are often used to overcome this issue [59–62]. However, while empirically successful, these methods cannot typically guarantee nominal or adversarial performance outside of the training samples. Indeed, classical learning theory [63, 64] provides generalization bounds only for the aggregated objective and not each individual penalty term. Additionally, the choice of the penalty parameter is not straightforward and depends on the underlying learning task, making it difficult to transfer across applications and highly dependent on domain expert knowledge.

Contributions. To summarize, there is a significant gap between the theory and practice of robust learning, particularly with respect to *when* and *why* adversarial training works. In this paper, we study the algorithmic foundations of robust learning toward understanding the fundamental limits of adversarial training. To do so, we leverage semi-infinite constrained learning theory, providing a theoretical foundation for gradient-based attacks and mitigating the issue of nominal performance degradation. In particular, our contributions are as follows:

- We show that adversarial training is equivalent to a stochastic optimization problem over a specific, *non-atomic* distribution, which we characterize using recent non-convex duality results [65, 66]. Further, we show that a myriad of previous adversarial attacks reduce to particular, sub-optimal choices of this distribution.
- We propose an algorithm to solve this problem based on stochastic optimization and Markov chain Monte Carlo (MCMC). Gradient-based adversarial methods such as [33, 32] can be interpreted as limiting cases of this procedure.
- We show that our algorithm performs similarly, and in some cases outperforms, state-of-the-art baselines on standard benchmarks, including MNIST and CIFAR.
- We provide generalization guarantees for the empirical version of this algorithm based on dual learning theory [66], showing how to effectively limit the nominal performance degradation of robust classifiers.

2 Problem formulation

Standard training. Consider the statistical problem of learning a classifier that minimizes an expected loss according to an unknown data distribution. Explicitly, let $\mathcal{D}$ denote an *unknown* joint probability distribution over instance-label pairs $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathcal{X}$ is supported on a compact subset of $\mathbb{R}^d$, and $y \in \mathcal{Y} = \{1, \ldots, K\}$ denotes the class of $\mathbf{x}$. For example, in a handwritten digit image-classification tasks, the label y would correspond to the number depicted in the image $\mathbf{x}$. For the sake of exposition, assume that $\mathcal{D}$ admits a density $\mathfrak{p}(\mathbf{x}, y)$. Throughout this paper, we consider the measurable space $(\Omega, \mathcal{B})$, with $\Omega = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{B}$ denoting its Borel σ-algebra.

Consider the hypothesis class $\mathcal{H}$ containing functions $f_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathcal{P}^K$ parametrized by $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, where again Θ is assumed to be compact, and where $\mathcal{P}^K$ is the $(K-1)$-simplex. We assume that $f_{\boldsymbol{\theta}}$ is differentiable with respect to $\boldsymbol{\theta}$. Note that the classes of halfspaces, logistic classifiers, and CNNs with softmax outputs can all be described by this formalism. To make a prediction $\hat{y} \in \mathcal{Y}$, we assume that the simplex $\mathcal{P}^K$ is mapped to the set of classes $\mathcal{Y}$ via

$$\hat{y} \in \mathrm{argmax}_{k \in \mathcal{Y}} \, [f_{\boldsymbol{\theta}}(\mathbf{x})]_k \tag{1}$$

with ties broken arbitrarily. In this way, we can think of the k-th output of the classifier as representing the probability of $y = k$. This learning problem can then be formulated as the mathematical program

$$\underset{\boldsymbol{\theta} \in \Theta}{\mathrm{minimize}} \; \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \Big[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}), y\big) \Big], \tag{P-NOM}$$

where ℓ is a $[0, B]$-valued loss function such that $\ell(\cdot, y)$ is M-Lipschitz continuous function for all $y \in \mathcal{Y}$. Observe that $(\mathbf{x}, y) \mapsto \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$ must be integrable for (P-NOM) to be well-defined; we further assume that this map is an element of the functional L^p space for some fixed $p \in (1, \infty)$.

Adversarial training. For common choices of the hypothesis class $\mathcal{H}$, including DNNs, the classifier obtained from (P-NOM) is known to be sensitive to input perturbations [67]. In other words, it is often straightforward to find a relatively small pertubations $\boldsymbol{\delta}$ such that the classifier correctly predicts the label y of $\mathbf{x}$, but misclassifies the perturbed sample $\mathbf{x} + \boldsymbol{\delta}$. This has lead to increased interest in the robust analog of (P-NOM), namely,

$$P_{\mathrm{R}}^{\star} \triangleq \min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right], \tag{P-RO}$$

where $\Delta \subset \mathbb{R}^d$ denotes the set of valid perturbations[1]. Typically, Δ is chosen to be a ball with respect to a given metric on Euclidean space, e.g. $\Delta = \Delta(\epsilon) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\| \leq \epsilon\}$. Here, however, we make no particular assumption on the specific form of Δ. In particular, our contributions apply to arbitrary perturbation models, such as those used in [68–72].

Analyzing conditions under which (P-RO) can be (probably approximately) solved from data, e.g., using empirical risk minimization (ERM), remains an active area of research. While bounds on the Rademacher complexity [73, 74] and VC dimension [73–77] of the robust loss

$$\bar{\ell}(f_{\boldsymbol{\theta}}(\mathbf{x}), y) = \max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \tag{2}$$

have been derived for an array of losses and hypothesis classes, there are still open questions on the effectiveness and sample complexity of adversarial learning [74].

While these works justify replacing the objective of (P-RO) by its empirical counterpart, they do not address the issue of computing the adversarial loss (2). Indeed, due to the non-concavity of the map $\boldsymbol{\delta} \mapsto \ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y)$ (except in trivial cases, e.g., when $\mathcal{H}$ and ℓ are linear), evaluating the maximum in (2) is not straightforward. For this reason, a variety of heuristics are generally used to approximate (2), such as linearizing the objective [78], drawing $\boldsymbol{\delta}$ randomly from a hand-crafted distribution [79], or choosing perturbations so as to modify perceptual properties of the input $\mathbf{x}$ [80–82]. However, the most common and empirically effective strategy has been to leverage the differentiability of typical ML models (e.g., CNNs) with respect to their inputs and approximate the value of (2) using projected gradient ascent. In [33, 83], for instance, $\boldsymbol{\delta}$ is computed from $(\boldsymbol{\theta}, \mathbf{x}, y)$ by repeatedly applying

$$\boldsymbol{\delta}^+ = \textstyle\prod_{\Delta} \Big[\boldsymbol{\delta} + \eta \operatorname{sign} \big[\nabla_{\boldsymbol{\delta}} \ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) \big] \Big], \tag{3}$$

where Π_{Δ} denotes the projection onto Δ and $\eta > 0$ is a fixed step size.

Common pitfalls for adversarial training. Their empirical success notwithstanding, gradient-based approaches to adversarial training are not without issues. One fundamental pitfall is the fact that gradient-based algorithms are not guaranteed to provide optimal (or even near-optimal) perturbations, since $\ell(f_{\boldsymbol{\theta}}(\cdot), y)$ is typically not a convex function. In fact, notice that maximizing over $\boldsymbol{\delta}$ in (P-RO) is a severely underparametrized problem as opposed to the minimization over $\boldsymbol{\theta}$ and therefore does not enjoy the same benign optimization landscape [41–45]. In fact, heuristics such as random initializations [79] and step size adjustment [84] are often needed to improve the solutions obtained by (3). These issues are particularly critical given that models are also *evaluated* using some variant of (3). Thus, the use of heuristics generally undermines confidence with respect to the test-time robustness when training-time attacks do not match test-time attacks.

Another issue faced by the robust formulation (P-RO) is that it often degrades the nominal performance, i.e. the performance of the model on clean data. Penalty-based approaches are often used to overcome this issues by combining the objective of (P-NOM) with a penalty that promotes output invariance in a neighborhood of each sample [59]. In particular, this technique has been used to obtain state-of-the-art performance in several benchmarks [85]. These results, however, are not guaranteed to generalize outside of the training sample. Indeed, classical learning theory guarantees

---

[1]Note that $f_{\boldsymbol{\theta}}$ must now be defined on $\mathcal{X} \oplus \Delta$, where $\oplus$ denotes the Minkowski (set) sum. In a slight abuse of notation, we will refer to this set as $\mathcal{X}$ from now on.

generalization in terms of the aggregated objective and not in terms of the robustness requirements it may describe [63, 64, 66].

In the remainder of this paper, we address these two generalization issues by leveraging semi-infinite constrained learning theory. To do so, we explicitly formulate the problem of finding the most robust classifier among those that have good nominal performance. We then show that (P-RO) is equivalent to a stochastic optimization problem (Section 3.1) that can be related to many robust training methods proposed in the literature. We then leverage recent dual empirical learning results [66] to provide generalization guarantees that allow constrained robust learning problems to be solved using empirical (*unconstrained*) risk minimization (Section 3.2). Finally, we derive an algorithm based on a Langevin MCMC sampler of which (3) is a particular case (Section 4).

3 Dual robust learning

In this section, we develop the theoretical foundation to tackle the two challenges of (P-RO), namely, finding worst-case perturbations, i.e. evaluating (2), and mitigating the degradation of the nominal performance. To do so, we will rely on constrained learning and duality theory. Explicitly, we set out to tackle the following constrained learning problem

$$
\begin{aligned}
P^\star \triangleq \underset{\boldsymbol{\theta} \in \Theta}{\text{minimize}} \quad & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right] \\
\text{subject to} \quad & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}), y\big) \right] \leq \rho
\end{aligned}
\tag{PI}
$$

where $\rho \geq 0$ is the fixed desired nominal performance level. At a high level, in Problem (PI) we seek the most robust classifier among those that have good nominal performance. In this way, Problem PI is directly designed to address the well-known trade-off between robustness and accuracy.

At face value, the statistical constraint in Problem PI is challenging to enforce in practice, especially given the well-known difficult in solving the unconstrained analog of (P-RO) in practice. To this end, our approach in this work is to use duality to obtain solutions for (PI) that generalize with respect to both adversarial and nominal performance (see Section 3.2). Before describing this approach, we emphasize that while we consider the nominal loss as a constraint, the theory and algorithms that follow also apply to the constrained learning problem in which the objective and constraint of (PI) are reversed.

3.1 Computing worst-case perturbations

Before tackling the constrained problem (PI), let us begin by solving its unconstrained version, namely, (P-RO). To do so, we start by writing (P-RO) using an epigraph formulation of the maximum function to obtain the semi-infinite program

$$
\begin{aligned}
P_{\text{R}}^\star = \underset{\boldsymbol{\theta} \in \Theta,\, t \in L^p}{\text{minimize}} \quad & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[t(\mathbf{x}, y)\big] \\
\text{subject to} \quad & \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \leq t(\mathbf{x}, y), \quad \text{for all } (\mathbf{x}, \boldsymbol{\delta}, y) \in \mathcal{X} \times \Delta \times \mathcal{Y}.
\end{aligned}
\tag{PII}
$$

Note that (PII) is indeed equivalent to (P-RO) given that

$$
\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \leq t(\mathbf{x}, y) \iff \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \leq t(\mathbf{x}, y) \ \forall \boldsymbol{\delta} \in \Delta
\tag{4}
$$

While at first it may seem that we have made Problem PI more challenging to solve by transforming an unconstrained problem into an infinitely-constrained problem, notice that (PII) is no longer a composite min-max problem. Additionally, it is *linear* in t, indicating that Problem PII should be amenable to a Lagrangian duality based approach. Indeed, the following proposition shows that (PII) can be used to obtain a statistical counterpart of (P-RO).

Proposition 3.1. *If* $(\mathbf{x}, y) \mapsto \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}), y\big) \in L^p$ *for* $p \in (1, \infty)$, *then* (P-RO) *can be written as*

$$
P_{\text{R}}^\star = \min_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{\theta}),
\tag{PIII}
$$

for the primal function

$$
p(\boldsymbol{\theta}) \triangleq \max_{\lambda \in \mathcal{P}^q} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathbb{E}_{\boldsymbol{\delta} \sim \lambda(\boldsymbol{\delta}|\mathbf{x},y)} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) \right] \right],
\tag{5}
$$

where $\mathcal{P}^q$, *with* $\frac{1}{p} + \frac{1}{q} = 1$, *is the subspace of* L^q *containing almost everywhere non-negative functions such that* $\mathfrak{p}(\mathbf{x}, y) = 0 \Rightarrow \lambda(\boldsymbol{\delta} \mid \mathbf{x}, y) = 0$ *and* $\int \lambda(\boldsymbol{\delta} \mid \mathbf{x}, y)d\boldsymbol{\delta} = 1$ *for almost every* $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$.

171 *Proof.* See supplementary material (Appendix B). □

172 Proposition 3.1 establishes an equivalence between the traditional robust learning problem (P-RO),
173 where the maximum is taken over perturbations $\boldsymbol{\delta} \in \Delta$ of the input, and its stochastic version (PIII),
174 where the maximum is taken over a conditional distribution of perturbations $\boldsymbol{\delta} \sim \lambda(\boldsymbol{\delta}|\mathbf{x}, y)$. This
175 dichotomy parallels the one that arises in PAC vs. agnostic PAC learning. Indeed, while the former
176 seeks a deterministic map $(\boldsymbol{\theta}, \mathbf{x}, y) \mapsto \boldsymbol{\delta}$, the latter considers instead a distribution of perturbations
177 over $\boldsymbol{\delta}|\mathbf{x}, y$ parametrized by $\boldsymbol{\theta}$. In fact, since (PIII) is obtained from (P-RO) through semi-infinite
178 duality (see proof of Proposition 3.1), the density of this distribution is exactly characterized by
179 the dual variables λ. Several adversarial training methods can be interpreted by taking particular,
180 sub-optimal choices of this distribution (see Appendix A).

181 It is also worth noting that while (PIII) was obtained using Lagrangian duality, it can also be seen
182 as a *linear lifting* of the maximization in (2). From this perspective, while recovering (2) would
183 require λ to be atomic, Proposition 3.1 shows that this is in fact not necessary as long as $\ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y)$
184 is an element of L^p. Furthermore, observe that Proposition 3.1 does not account for $p \in \{1, \infty\}$ for
185 conciseness only, since their dual spaces are not isomorphic to L^q for any q. Still, neither of the dual
186 spaces L^{1*} or $L^{\infty*}$ contain Dirac distributions, meaning that λ would remain non-atomic.

187 **Exact solutions for the outer maximization in Problem** (PIII). While (PIII) provides a new,
188 infinitely-constrained alternative formulation for (P-RO), the fact remains that the objectives of
189 both (PIII) and (P-RO) involve the solution of a non-trivial maximization. However, whereas the
190 maximization problem in (P-RO) a (possibly non-convex) finite dimensional optimization problem,
191 the maximization in (PIII) is a linear, variational problem. We can therefore leverage variational
192 duality theory to obtain a full characterization of the optimal distribution $\lambda^\star$ when $p = 2$.

193 **Proposition 3.2** (Optimal distribution for (PIII)). *Let $p = 2$ (and $q = 2$) in Proposition 3.1. For*
194 *each $(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}$, there exists $\gamma(\mathbf{x}, y) > 0$ and $\mu(\mathbf{x}, y) \in \mathbb{R}$ such that*

$$\lambda^\star(\boldsymbol{\delta} \mid \mathbf{x}, y) = \left[\frac{\ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) - \mu(\mathbf{x}, y)}{\gamma(\mathbf{x}, y)} \right]_+, \tag{6}$$

195 *with $[z]_+ = \max(0, z)$, is a solution of the maximization in (5). The value of $\mu(\mathbf{x}, y)$ is such that*

$$\int \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) - \mu(\mathbf{x}, y) \right]_+ d\boldsymbol{\delta} = \gamma(\mathbf{x}, y), \text{for all } (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}.$$

196 *Proof.* See supplementary material (Appendix C). □

197 Hence, in the particular case of $(\mathbf{x}, y) \mapsto \ell(f_{\boldsymbol{\theta}}(\mathbf{x}), y) \in L^2$, we can obtain a closed-form expression
198 for the distribution that maximizes the objective of (PIII), which turns out to be essentially proportional
199 to a truncated version of the loss of the classifier. Note that the assumption that the loss belongs to
200 $\in L^2$ is mild given that the compactness of $\mathcal{X}$, $\mathcal{Y}$, and Δ imply that $L^{p_1} \subset L^{p_2}$ for $p_1 > p_2$. It is,
201 however, fundamental to obtain the closed-form solution in Proposition 3.2 since it allows (5) to be
202 formulated as a strongly convex constrained problem whose primal solution (6) can be recovered
203 from its dual variables (namely, γ and μ).

204 For completeness, we remark that in practice, the value of γ for which (6) is a solution of (5) is
205 not known *a priori* and can be arbitrarily close to zero, in which case the support of $\lambda^\star$ vanishes.
206 Since sampling from these discontinuous distributions can be quite intricate in high dimensional
207 settings, we fix $\gamma(\mathbf{x}, y) = \int \ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) d\boldsymbol{\delta}$ in the sequel. Since ℓ is non-negative, this implies
208 that $\mu(\mathbf{x}, y) = 0$ and that the distribution induced by $\lambda^\star$ is exactly proportional to the loss. In
209 Section 4, we put forward an MCMC algorithm that samples from this distribution. While this is
210 naturally an approximation that may over-smooth the optimal distribution, it is one that is quite
211 effective in practice (see Section 6). What is more, it allows us to derive generalization guarantees for
212 dual solutions of the constrained learning problem (PI).

213 ## 3.2 Solving the constrained learning problem

214 While we have shown that (PI) exactly formulates the problem of finding the most robust model with
215 high nominal performance, the statistical and possibly non-convex nature of the constraint makes

it difficult to impose directly (e.g. by projecting onto the feasibility set). Moreover, recall that we have access to the data distribution $\mathcal{D}$ only through samples $(\mathbf{x}, y) \sim \mathcal{D}$, which means that we in practice we cannot evaluate the expectations in (PI). To overcome these obstacles, we use duality to approximate (PI) by the empirical, unconstrained saddle point problem

$$\hat{D}^\star \triangleq \max_{\nu \geq 0} \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \nu) \tag{$\widehat{\text{DI}}$}$$

where $\hat{L}(\boldsymbol{\theta}, \nu)$ defines the empirical Lagrangian:

$$\hat{L}(\boldsymbol{\theta}, \nu) = \frac{1}{N} \sum_{n=1}^{n} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n + \boldsymbol{\delta}), y_n\big) + \nu \left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n\big) - \rho \right] \right]. \tag{7}$$

The following proposition shows that solutions of $(\widehat{\text{DI}})$ are (probably approximately) near-optimal *and* feasible for (PI). Naturally, this is only possible if the objective and constraint of (PI) are learnable individually. As we discussed in Section 2, this is known to hold in a variety of scenarios (when the Rademacher complexity or VC dimension are bounded), but remains an area of active research [73–77]. This learnability condition is encapsulated in the following assumptions:

Assumption 3.3. *The parametrization $f_{\boldsymbol{\theta}}$ is rich enough so that for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$ and $\beta \in [0, 1]$, there exists $\boldsymbol{\theta} \in \Theta$ such that*

$$\sup_{\mathbf{x} \in \mathcal{X}} |\beta f_{\boldsymbol{\theta}_1}(\boldsymbol{x}) + (1 - \beta) f_{\boldsymbol{\theta}_2}(\boldsymbol{x}) - f_{\boldsymbol{\theta}}(\boldsymbol{x})| \leq \alpha. \tag{8}$$

Assumption 3.4. *There exists $\boldsymbol{\theta}' \in \Theta$ such that $\mathbb{E}_{\mathcal{D}} \left[\ell\big(f_{\boldsymbol{\theta}'}(\mathbf{x}), y\big) \right] < \rho - M\alpha$.*

Assumption 3.5. *There exists $\zeta_R(N), \zeta_N(N) \geq 0$ monotonically decreasing with N such that*

$$\left| \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right] - \frac{1}{N} \sum_{n=1}^{N} \max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n + \boldsymbol{\delta}), y_n\big) \right| \leq \zeta_R(N) \ w.p. \ 1 - \delta \tag{9a}$$

$$\left| \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}), y\big) \right] - \frac{1}{N} \sum_{n=1}^{N} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n\big) \right| \leq \zeta_N(N) \ w.p. \ 1 - \delta \tag{9b}$$

for all $\boldsymbol{\theta} \in \Theta$.

Under these assumptions, we can explicitly bound the empirical duality gap between gap (with high probability) and characterize the feasibility of the empircal dual optimal solution for (PI).

Proposition 3.6 (The empirical dual of (PI)). *Let $\ell(\cdot, y)$ be a convex function for all $y \in \mathcal{Y}$. Under Assumptions 3.3–3.5, it holds with probability $1 - 6\delta$ that*

1. $\left| P^\star - \hat{D}^\star \right| \leq M\alpha + (1 + \overline{\nu}) \max(\zeta_R(N), \zeta_N(N))$; *and*

2. there exists $\boldsymbol{\theta}^\dagger \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \hat{\nu}^\star)$ such that $\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell\big(f_{\boldsymbol{\theta}^\dagger}(\mathbf{x}), y\big) \right] \leq \rho + \zeta_N(N)$.

Here, $\hat{\nu}^\star$ denotes a solution of (DI), *$\nu^\star$ denotes an optimal dual variable of* (PI) *solved over $\overline{\mathcal{H}} = \operatorname{conv}(\mathcal{H})$ instead of $\mathcal{H}$, and $\overline{\nu} = \max(\hat{\nu}^\star, \nu^\star)$. Additionally, for any interpolating classifier $\boldsymbol{\theta}'$, i.e. such that $\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell\big(f_{\boldsymbol{\theta}'}(\mathbf{x}), y\big) \right] = 0$, it holds that*

$$\nu^\star \leq \rho^{-1} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}'}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right]. \tag{10}$$

Proof. See supplementary material (Appendix D). $\qquad\square$

Proposition 3.6 states that seeking a robust classifier with a given nominal performance is (probably approximately) equivalent to seeking a classifier that minimizes a combination of the nominal and adversarial empirical loss. Yet, while (7) resembles a penalty-based formulation such as the one in the well-known TRADES algorithm [59], notice that ν is an optimization variable in $(\widehat{\text{DI}})$ rather than a fixed hyperparameter. Though seemingly innocuous, this is the difference between guaranteeing generalization only on the aggregated loss (7) and guaranteeing generalization for the objective value

and feasibility as in Proposition 3.6. Additionally, notice that [59] seeks an invariant classifier, i.e. a classifier for which $f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta})$ and $f_{\boldsymbol{\theta}}(\mathbf{x})$ are similar, rather than one with small adversarial loss. This problem can also be accounted for in (PI) by replacing the loss in the objective by a measure of average causal effect (ACE).

Combining Propositions 3.1–3.6, we obtain an optimization problem that is considerably more amenable than (PI). Indeed, it is (i) *empirical* and does not involve unknown statistical quantities such as $\mathcal{D}$; (ii) unconstrained; and (iii) its objective does not involve a hard maximization problem in view of the closed-form characterization of $\lambda^{\star}$ in Proposition 3.2. In the next section, we leverage these properties to propose a practical algorithms based on MCMC and stochastic optimization.

4 Dual robust learning algorithm

As shown in the previous section, Propositions 3.1–3.6 allows us to transform (PI) into the following **D**ual **A**dversarial **LE**arning problem

$$\hat{D}^{\star} \triangleq \max_{\nu \geq 0} \min_{\boldsymbol{\theta} \in \Theta} \frac{1}{N} \sum_{n=1}^{n} \Big[\mathbb{E}_{\boldsymbol{\delta}_n} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}_n + \boldsymbol{\delta}_n), y_n) \right] + \nu \left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n\big) - \rho \right] \Big] \qquad \text{(P-DALE)}$$

where $\boldsymbol{\delta}_n \sim \gamma^{-1} \big[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n + \boldsymbol{\delta}_n), y_n\big) - \mu \big]_+$ for some $\gamma > 0$ and μ as in Proposition 3.2. Unlike (PI), (P-DALE) is an *unconstrained* optimization problem whose cost function no longer requires a maximization over the perturbations $\boldsymbol{\delta}$. In fact, for models that are linear in $\boldsymbol{\theta}$ but nonlinear in the input (e.g., kernel models or logistic regression), this implies that we have transformed the non-convex, composite optimization problem (PI) into the convex (P-DALE). Naturally, for many modern ML models, such as CNNs, (P-DALE) remains a non-convex program in $\boldsymbol{\theta}$. However, there is overwhelming theoretical and empirical evidence that gradient descent algorithm yield good local minimizers for such overparametrized problems [41–45].

That being said, solving (P-DALE) remains a highly nontrivial challenge. In particular, although written in empirical form over the samples $\{(x_n, y_n)\}$, (P-DALE) remains a stochastic optimization problem. Still, since we know the distribution of $\boldsymbol{\delta}_n$, we can leverage stochastic approximation methods to minimize its objective using samples from this perturbation distribution. However, obtaining such samples can be challenging, especially when the dimension of $\boldsymbol{\delta}_n$ is large (e.g. for image-classification tasks) and the distribution is discontinuous (Proposition 3.2). This issue can be addressed using Hamiltonian Monte Carlo (HMC) methods, which leverage the geometry of the distribution to ovecome the curse of dimensionality. In particular, we propose to use a projected Langevin sampler (LMC) [86] and fix $\gamma = \int \ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) d\mathbf{x} dy$ so that $\mu = 0$ and the distribution of $\boldsymbol{\delta}_n$ is exactly proportional to the loss. We leave for future work the exploration of more advanced HMC methods capable of faster mixing (e.g., proximal Langevin [86] or hit-and-run [87]) and sampling from more complex, discontinuous distributions (e.g. [88]). The resulting algorithm is summarized in Algorithm 1.

Notice that Algorithm 1 accounts for scenarios in which the losses associated with the adversarial performance ($\bar{\ell}$), the perturbation (g), and the nominal performance (ℓ) are different. It can therefore learn from perturbations that are adversarial for a different loss than the one used for training the model $\boldsymbol{\theta}$. This generality allows it to tackle different applications by, e.g., replacing the adversarial error objective in (PI) by a measure of model invariance (e.g. ACE in [59]). It will also be used to show how existing adversarial training procedures can be seen as approximations of Algorithm 1 (see Appendix A).

Before proceeding, we provide a few notes on the convergence properties of Algorithm 1. To begin, observe that it Algorithm 1 is primal-dual algorithm [89] in which the sampling procedure in steps 3–7 is used to obtain an estimate of the stochastic gradient of the primal problem. When $\boldsymbol{\theta} \mapsto \ell\big(f_{\boldsymbol{\theta}}(\cdot), \cdot\big)$ is convex (e.g., for linear, kernel, or logistic models), it is well-known that SGD converges almost surely as long as this gradient estimate is unbiased [90]. As is typical with LMC, we omitted the Metropolis-Hastings acceptance step in Algorithm 1 that would guarantee unbiased estimates [91]. Still, when g is log-concave (e.g., the softmax output of a CNN), this procedure approaches the true distribution in total variation norm, which implies that its bias can be made arbitrarily small [86]. This is enough to guarantee almost sure convergence to a neighborhood of the optimum [92, 93].

Algorithm 1 Semi-Infinite Dual Adversarial Learning (DALE)

Initialize $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$ and $\nu \leftarrow 0$

1: **repeat**
2:    **for** Batch $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ **do**
3:       $\boldsymbol{\delta}_i \leftarrow \mathbf{0}$, for $i = 1, \ldots, m$
4:       **for** L steps **do**
5:          $U_i \leftarrow T^{-1} \log \left[g\big(f_{\boldsymbol{\theta}}(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big) \right]$, for $i = 1, \ldots, m$
6:          $\boldsymbol{\delta}_i \leftarrow \prod_{\Delta} \left[\boldsymbol{\delta}_i + \eta \nabla_{\boldsymbol{\delta}_i} U_i + \sqrt{2\eta} \boldsymbol{\xi}_i \right]$, where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, I)$ and $i = 1, \ldots, m$
7:       **end for**
8:       $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{\eta_p}{m} \sum_{i=1}^m \nabla_{\boldsymbol{\theta}} \left[\tilde{\ell}\big(f_{\boldsymbol{\theta}}(\mathbf{x}_i + \boldsymbol{\delta}_i), y_i\big) + \nu \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i\big) \right]$
9:    **end for**
10:    $\nu \leftarrow \left[\nu + \eta_d \left(\frac{1}{N} \sum_{n=1}^n \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n\big) - \rho \right) \right]_+$
11: **until** convergence

The convergence properties of primal-dual methods are less well understood when $\boldsymbol{\theta} \mapsto \ell\big(f_{\boldsymbol{\theta}}(\cdot), \cdot\big)$ is non-convex. Still, a good estimate of the primal minimizer is enough to obtain an approximate gradient for dual ascent [65, 66]. There is overwhelming empirical and theoretical evidence that this is the case for overparametrized models, such as CNNs, trained using gradient descent [41–45]. We can then run the primal (step 8) and dual (step 10) updates at different timescales so as to obtain a good estimate of the primal minimizer before performing dual ascent.

5 Related work

Adversarial robustness. As described in Section 1, it is well-know that state-of-the-art classifiers are susceptible to adversarial attacks [11–17, 32]. Toward addressing this challenging, a rapidly-growing body of work has provided *attack algorithms* to generate data perturbations that fool classifiers and *defense algorithms* which are designed to train robust classifiers to be robust against these perturbations. However, despite the myriad of work in this field and significant improvements on a number of well-known benchmarks [26–31, 33–37], there are still many open questions on when adversarial learning is even possible and in what sense [73–77]. Unlike the majority of these works, we exploit duality to derive a principled primal-dual style algorithm from first principles for the adversarial robustness setting.

Constrained optimization. Also related are works that seek to enforce constraints on learning problems [94]. While several heuristic algorithms exist for this setting, many focus on restricted classes of constraints [95–99] and those that can handle more general constraints come at the cost of added computation complexity [100, 101]. Moreover, each of these works seeks to enforce constraints on a particular parameterization for the learning problem (such as directly on the weights of a neural network) rather than on the underlying statistical problem, as we do in this paper. In this way, our work is more related to the primal-dual style algorithms which often arise in convex optimization [89, 102].

6 Experiments

In this section, we include an empirical evaluation of the DALE algorithm presented in Algorithm 1. We defer further details concerning implementation and hyperparameter selection to the Appendix.

6.1 MNIST

We first consider the MNIST dataset [103]. All defense models use a four-layer convolutional neural network (CNN) trained using with the Adadelta optimizer [104]. Throughout, we use a learning rate of 1.0, a batch size of 128, and we use a maximum perturbation radius of $\epsilon = 0.3$. In Table 1, we

show the accuracies obtained by running the DALE algorithm for $L = 1$ and $L = 100$ steps; we also offer comparison to a range of baselines. Observe that DALE achieves both strong nominal and adversarial performance, beating these state-of-the-art baselines in both categories.

Table 1: **MNIST.** Classification accuracies for MNIST with $\epsilon = 0.3$.

Training Algorithm	Clean Accuracy	Adv. Accuracy	
		PGD	FGSM
PGD	98.1	93.1	95.5
FGSM	98.3	0.65	98.1
FAB	99.2	91.6	95.3
TRADES	98.9	94.0	96.5
A-PGD	99.1	92.5	96.1
MART	98.9	93.5	96.1
DALE-1	99.1	83.7	97.7
DALE-100	98.8	97.1	99.1

Figure 1: **Distribution of adversarial perturbations.** We visualize the distribution of adversarial perturbations generated by PGD, FGSM, and DALE by embedding the perturbations in a 2-dimensional space created using principle components analysis (PCA).

Visualization the distribution of adversarial perturbations. In Section 3, we introduced a new perspective on the problem of generating adversarial examples in which we characterized adversarial perturbations with respect to a worst-case distribution $\lambda^\star$. To this end, Algorithm 1 is designed to sample from this distribution. In Figure 1, we attempt to visualize the perturbations generated by our algorithm by using principal components analysis (PCA) to embed the adversarial perturbations in a two-dimensional space. In particular, we performed PCA on the MNIST dataset to extract the principle components, and then we projected the perturbations δ generated by PGD, FGSM, and DALE in the last iteration of training onto the first two principle components. Notice that the perturbations generated by PGD and FGSM are much more concentrated nearer to the boundary of the region, whereas the perturbations generated by LMC are closer to the origin. Further, PGD and FGSM both vary much more in the second principal component (y-axis) than the first principal component (x-axis) relative to DALE. This indicates that PGD and FGSM may overfit to the first few principal components of the data; in practice, Figure 1 shows that DALE is able to overcome this issue more successfully.

6.2 CIFAR-10

We next consider the CIFAR-10 dataset [105]. All defense models were trained using SGD with momentum using the ResNet-18 architecture [106]. Following [60], we use an initial learning rate of 0.01, which was decayed to 0.001 at epoch 75, and then to 0.0001 at epoch 90. All models were trained for 100 epochs using a perturbation radius of $\epsilon = 8/255$. The clean and robust accuracies for DALE as well as several state-of-the-art baselines are shown in Table 2. Note that the DALE is able to mitigate the trade-off between robustness and accuracy more successfully that each of the baselines. This highlights the utility of performing dual-ascent in line 11 of Algorithm 1, as both TRADES and MART use similar objectives to DALE with a *fixed* penalty weight.

Table 2: **CIFAR-10.** Classification accuracies for CIFAR with $\epsilon = 8/255$.

Training Algorithm	Clean Accuracy	Adv. Accuracy
		PGD
PGD	85.7	52.4
TRADES	85.6	54.0
MART	83.8	55.2
DALE-10	86.4	54.9

7 Conclusion

In this paper, we studied robust learning from a constrained learning perspective. We rigorously proved an equivalence between the standard adversarial training paradigm and a stochastic optimization problem over a specific, *non-atomic* distribution. This insight provides a new perspective on robust learning, and engenders a natural Langevin Markov Chain Monte Carlo approach for adversarial robustness. We validate experimentally that this algorithm performs similarly, and in some cases outperforms the state-of-the-art on standard benchmarks. In future work, we improve the sampling procedure in Algorithm 1 by incorporating more sophisticated samplers.

References

[1] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark De-Pristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[2] Li Yao, Jordan Prosky, Ben Covington, and Kevin Lyman. A strong baseline for domain adaptation and generalization in medical imaging. *arXiv preprint arXiv:1904.01638*, 2019.

[3] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *arXiv preprint arXiv:2009.12829*, 2020.

[4] Vishnu M Bashyam, Jimit Doshi, Guray Erus, Dhivya Srinivasan, Ahmed Abdulkadir, Mohamad Habes, Yong Fan, Colin L Masters, Paul Maruff, Chuanjun Zhuo, et al. Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. *arXiv preprint arXiv:2010.05355*, 2020.

[5] Dongxia Zhang, Xiaoqing Han, and Chunyu Deng. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE Journal of Power and Energy Systems*, 4(3):362–370, 2018.

[6] Hadis Karimipour, Ali Dehghantanha, Reza M Parizi, Kim-Kwang Raymond Choo, and Henry Leung. A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids. *IEEE Access*, 7:80778–80788, 2019.

[7] Tariq Samad and Anuradha M Annaswamy. Controls for smart grids: Architectures and applications. *Proceedings of the IEEE*, 105(11):2244–2261, 2017.

[8] Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Never stop learning: The effectiveness of fine-tuning in robotic reinforcement learning. *arXiv e-prints*, pages arXiv–2004, 2020.

[9] Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[10] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, Pieter Abbeel, Wolfram Burgard, Michael Milford, et al. The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5):405–420, 2018.

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

[12] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[14] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. *arXiv preprint arXiv:2007.08558*, 2020.

[15] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.

[17] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[18] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

[19] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828, 2015.

[20] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

[21] Anoopkumar Sonar, Vincent Pacelli, and Anirudha Majumdar. Invariant policy optimization: Towards stronger generalization in reinforcement learning. *arXiv preprint arXiv:2006.01096*, 2020.

[22] Eugene Vinitsky, Yuqing Du, Kanaad Parvate, Kathy Jang, Pieter Abbeel, and Alexandre Bayen. Robust reinforcement learning using adversarial populations. *arXiv preprint arXiv:2008.01825*, 2020.

[23] John W Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

[24] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[25] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

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

[27] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv e-prints*, pages arXiv–1712, 2017.

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

[29] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*, 2019.

[30] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[31] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pages 5458–5467. PMLR, 2020.

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

[33] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[34] Eric Wong and J Zico Kolter. Provable Defenses Against Adversarial Examples Via the Convex Outer Adversarial Polytope. *arXiv preprint arXiv:1711.00851*, 2017.

[35] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[36] Ayan Sinha, Zhao Chen, Vijay Badrinarayanan, and Andrew Rabinovich. Gradient adversarial training of neural networks. *arXiv preprint arXiv:1806.08028*, 2018.

[37] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

[38] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[39] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

[40] Yan Li, Ethan X Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *International Conference on Learning Representations*, 2019.

[41] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[43] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, pages 233–242. PMLR, 2017.

[44] Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017.

[45] Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. In *International conference on machine learning*, pages 605–614. PMLR, 2017.

[46] Dongxian Wu, Shu-tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *arXiv preprint arXiv:2004.05884*, 2020.

[47] Minhao Cheng, Qi Lei, Pin-Yu Chen, Inderjit Dhillon, and Cho-Jui Hsieh. Cat: Customized adversarial training for improved robustness. *arXiv preprint arXiv:2002.06789*, 2020.

[48] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[49] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[50] Uri Shaham, James Garritano, Yutaro Yamada, Ethan Weinberger, Alex Cloninger, Xiuyuan Cheng, Kelly Stanton, and Yuval Kluger. Defending against adversarial images using basis functions transformations. *arXiv preprint arXiv:1803.10840*, 2018.

[51] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[52] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

[53] Yang Bai, Yan Feng, Yisen Wang, Tao Dai, Shu-Tao Xia, and Yong Jiang. Hilbert-based generative defense for adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4784–4793, 2019.

[54] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *arXiv preprint arXiv:1904.12843*, 2019.

[55] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

[56] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[57] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

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

[59] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019.

[60] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

[61] Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

[62] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

[63] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[64] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[65] Santiago Paternain, Luiz FO Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *arXiv preprint arXiv:1910.13393*, 2019.

[66] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[67] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning*, pages 9561–9571. PMLR, 2020.

[68] Alexander Robey, Hamed Hassani, and George J Pappas. Model-based robust deep learning. *arXiv preprint arXiv:2005.10247*, 2020.

[69] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *arXiv preprint arXiv:2102.11436*, 2021.

[70] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22:646–654, 2009.

[71] Eric Wong and J Zico Kolter. Learning perturbation sets for robust machine learning. *arXiv preprint arXiv:2007.08450*, 2020.

[72] Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1211–1220, 2020.

[73] Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

[74] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

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

[76] Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. *arXiv preprint arXiv:2005.07652*, 2020.

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

[78] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. *arXiv preprint arXiv:1907.02610*, 2019.

[79] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[80] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

[81] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. *arXiv preprint arXiv:1906.00001*, 2019.

[82] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2020.

[83] Uri Shaham, Yutaro Yamada, and Sahand Negahban. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015.

[84] Yucheng Shi, Yahong Han, Quanxin Zhang, and Xiaohui Kuang. Adaptive iterative attack towards explainable adversarial robustness. *Pattern Recognition*, 105:107309, 2020.

[85] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

[86] Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected langevin monte carlo. In *Advances in Neural Information Processing Systems*, pages 1243–1251. Citeseer, 2015.

[87] László Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[88] Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, 2020.

[89] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[90] J Frédéric Bonnans. *Convex and Stochastic Optimization*. Springer, 2019.

[91] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[92] Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[93] Ahmad Ajalloeian and Sebastian U Stich. Analysis of sgd with biased gradient estimators. *arXiv preprint arXiv:2008.00051*, 2020.

[94] Priya L Donti, David Rolnick, and J Zico Kolter. Dc3: A learning method for optimization with hard constraints. *arXiv preprint arXiv:2104.12225*, 2021.

[95] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1796–1804, 2015.

[96] Steven Chen, Kelsey Saulnier, Nikolay Atanasov, Daniel D Lee, Vijay Kumar, George J Pappas, and Manfred Morari. Approximating explicit model predictive control using constrained neural networks. In *2018 Annual American control conference (ACC)*, pages 1520–1527. IEEE, 2018.

[97] Thomas Frerix, Matthias Nießner, and Daniel Cremers. Homogeneous linear inequality constraints for neural network activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 748–749, 2020.

[98] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[99] Sathya N Ravi, Tuan Dinh, Vishnu Lokhande, and Vikas Singh. Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*, 2018.

[100] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and Zico Kolter. Differentiable convex optimization layers. *arXiv preprint arXiv:1910.12430*, 2019.

[101] Dimitris A Karras and Stavros J Perantonis. An efficient constrained training algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 6(6):1420–1434, 1995.

[102] Luiz FO Chamon, Santiago Paternain, Miguel Calvo-Fullana, and Alejandro Ribeiro. The empirical duality gap of constrained statistical learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8374–8378. IEEE, 2020.

[103] The MNIST database of handwritten digits Home Page. `http://yann.lecun.com/exdb/mnist/`.

[104] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[105] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[106] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[107] Lasse Holmstrom, Petri Koistinen, et al. Using additive noise in back-propagation training. *IEEE transactions on neural networks*, 3(1):24–38, 1992.

[108] Raphael Gontijo Lopes, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk. Improving robustness without sacrificing accuracy with patch gaussian augmentation. *arXiv preprint arXiv:1906.02611*, 2019.

[109] Evgenia Rusak, Lukas Schott, Roland S Zimmermann, Julian Bitterwolf, Oliver Bringmann, Matthias Bethge, and Wieland Brendel. A simple way to make neural networks robust against diverse image corruptions. In *European Conference on Computer Vision*, pages 53–69. Springer, 2020.

[110] Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

[111] Murtaza Eren Akbiyik. Data augmentation in training cnns: Injecting noise to images. 2019.

[112] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[113] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[114] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[115] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[116] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Data augmentation using random image cropping and patching for deep cnns. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9):2917–2931, 2019.

[117] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

[118] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *arXiv preprint arXiv:1805.12018*, 2018.

[119] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *arXiv preprint arXiv:2010.08001*, 2020.

[120] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[121] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[122] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

Table 3: Caption

A Connections to other problems

A.1 Relation to probabilistic and generative data augmentation

When $\lambda = \mathcal{N}(0, \sigma^2 I)$, the problem turns into Gaussian data augmentation [107]:

$$\min_{\phi \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\mathbb{E}_{\delta \sim \mathcal{N}(0, \sigma^2 I)} \left[\ell(\phi(x + \delta), y) \right] \right] \tag{11}$$

It has been claimed in past work that Gaussian augmentation improves generalization [108]; in particular, [108] advocates for adding a patch of Gaussian noise. Indeed, according to [109], "data augmentation with Gaussian... noise serves as a simple yet very strong baseline that is sufficient to surpass almost all previously proposed defenses against common corruptions on ImageNet-C for ResNet50." This paper also proposes a generative architecture that learns to add noise to data, which would correspond to the case when λ is a generative model, i.e.

$$\min_{\theta \in \mathcal{F}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\mathbb{E}_{\delta \sim G \# \mathcal{N}(0, \sigma^2 I)} \left[\ell(\phi(x + \delta), y) \right] \right] \tag{12}$$

where $G : \mathcal{Z} \to \mathcal{X}$ is a generative model. Several other studies have considered using generative models to either denoise corrupted images [110] or to corrupt images for robust training [cite me].

Additive and multiplicative noise are also considered [111]. This paper considered learning a data augmentation scheme using a recurrent neural network [112].

A.2 Relation to ad-hoc data augmentation

When λ is chosen to be a distribution that removes patches from a given image, this method reduces to `Cutout` [113, 114]. When λ is chosen to be a distribution that removes patches from a given image and replaces them with patches from different images from $\mathcal{D}$, we recover `CutMix` [115, 116]

A.3 Relation to (group-theoretic) data augmentation

Note that none of the analyses presented so far rely on the fact that the perturbation is applied linearly. That is, rather than considering perturbations of the form $\mathbf{x} \mapsto \mathbf{x} + \boldsymbol{\delta}$, we could have considered changes induced by known group structure, i.e. $\mathbf{x} \mapsto g \cdot \mathbf{x}$ where g is the element of some group G. For example, when G is the group of rotations about a fixed axis, we can think of the elements $g \in G$ as parameterizing angles and thus the application of the group element of g on x can be expressed as $g \cdot \mathbf{x} := R(g)x$, where $R(g)$ is a rotation matrix. Given such a group G, the dual problem to (**??**) for group-based perturbations would take the form

$$\operatorname*{maximize}_{\lambda_i \in \mathcal{P}(G)} \min_{\phi \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{g \sim \lambda_i} \left[\ell(\phi(g \cdot \mathbf{x}_i), y_i) \right] \tag{13}$$

For a fixed distribution $\mathbb{Q}$ in place of λ_i, this leads to the following *augmented ERM* problem:

$$\min_{\phi \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{g \sim \mathbb{Q}} \left[\ell(\phi(g \cdot \mathbf{x}_i), y_i) \right] = \frac{1}{N} \sum_{i=1}^{N} \int_G \left[\ell(\phi(g \cdot \mathbf{x}_i), y_i) \right] \mathrm{d}\mathbb{Q}(g). \tag{14}$$

This problem exactly corresponds to the group-theoretic data-augmentation scheme discussed in [117]. In this sense, we can think of (13) as seeking "worst-case" augmentation schemes WRT a given group G.

A.4 Relation to adversarial data augmentation

Closer to our work is the work of [118, 119], in which the authors propose an adversarial data augmentation scheme.

Given a data point $(\mathbf{x}, y) \sim \mathcal{D}$, they seek to find hard-to-classify data points via the following scheme:

$$x^\star \in \operatorname*{argmax}_{x \in \mathcal{X}} \ell(f(x), y) - \gamma \left\| \tilde{f}(\mathbf{x}) - \tilde{f}(\overline{\mathbf{x}}) \right\|_2^2 \tag{15}$$

where $\tilde{f} : \mathbb{R}^d \to \mathbb{R}^q$ is a mapping to some intermediate representation in f, e.g. the final layer of a neural network.

A.5 Relation to universal adversarial perturbations

In [120], the authors seek *image-agnostic* adversarial perturbations. That is, they seek small, norm-bounded perturbations that, when applied to *any* image in a given dataset, can fool classifiers trained on that dataset. This problem can be formulated as the following min-max problem:

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \ \max_{\boldsymbol{\delta} \in \Delta} \ \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell(\phi(\mathbf{x} + \boldsymbol{\delta}), y) \right] \tag{16}$$

Proposition A.1 (Universal adversarial perturbations). *Consider the optimization problem in* (16), *and assume that $\mathcal{F}$ is convex and that ℓ is convex in its first argument. Then* (16) *is equivalent to the (empirical) dual problem for* (??) *when $\lambda_i = \lambda_j \ \forall i \neq j$ and $\mathcal{P}$ is the set of Dirac distributions over* Δ.

A.6 Connections to domain generalization

More generally, in light of Sections A.3 and A.5, we could consider the more general problem of

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \ \max_{\boldsymbol{\delta} \in \Delta} \ \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell(\phi(G(\mathbf{x}, \boldsymbol{\delta})), y) \right] \tag{17}$$

where $G : \mathcal{X} \times \Delta \to \mathcal{X}$ is a fixed transformation of data. In the notation of [121], we can write $\mathbf{x}^{\boldsymbol{\delta}} := G(\mathbf{x}, \boldsymbol{\delta})$ and think of $\mathbf{x}^{\boldsymbol{\delta}}$ as being the *observation* of $\mathbf{x}$ in a domain or environment characterized by $\boldsymbol{\delta}$. Here Δ would encode a set of such environments. Thus, we consider the problem

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \ \max_{\boldsymbol{\delta} \in \Delta} \ \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell(\phi(\mathbf{x}^{\boldsymbol{\delta}}), y) \right]. \tag{18}$$

This problem exactly corresponds to the domain generalization problem studied in [121, 69].

When $G(\mathbf{x}, \boldsymbol{\delta}) = \mathbf{x} + \boldsymbol{\delta}$, we can think of the (18) as characterizing a domain generalization problem in which we seek robustness to universal adversarial perturbations. More specifically, assuming access to a finite set of training domains $\Delta_{\text{train}} := \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_k\}$, we seek to generalize to perform well against unseen domains in $\Delta / \Delta_{\text{train}}$. If one was interested in running IRM in this setting (not sure why, but still), you could show in several settings that you would need a very large number of domains to generalize. In particular, when $\mathcal{F}$ is the class of linear classifiers, [122] tells us that you would need $k > d$ domains to generalize to unseen domains.

Another line of work considers the "average-case" domain generalization problem, i.e.

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \ \mathbb{E}_{\delta \sim \mathbb{P}_\delta} \left[\mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\ell(\phi(\mathbf{x}^{\boldsymbol{\delta}}), y^{\boldsymbol{\delta}}) \right] \right] \tag{19}$$

which corresponds to the selection of $\lambda = \mathbb{P}_{\boldsymbol{\delta}}$.

We could also connect this to the causal view of domain generalization/distribution generalization.

A.7 Relation to randomized smoothing

The primal problem for randomized smoothing can be written as

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \ \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\max_{\boldsymbol{\delta} \in \Delta} \ell \left(\mathbb{E}_{\mathbf{t} \sim \mathcal{N}(0, \sigma^2 I)} [\phi(\mathbf{x} + \mathbf{t} + \boldsymbol{\delta})], y \right) \right] \tag{20}$$

It follows from the above proposition that the dual problem can be written as

$$\underset{\lambda \in \mathcal{P}(\Delta)}{\text{maximize}} \ \min_{\phi \in \mathcal{F}} \ \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[\mathbb{E}_{\boldsymbol{\delta} \sim \lambda | \mathbf{x}} \left[\ell \left(\mathbb{E}_{\mathbf{t} \sim \mathcal{N}(0, \sigma^2 I)} [\phi(\mathbf{x} + \mathbf{t} + \boldsymbol{\delta})], y \right) \right] \right] \tag{21}$$

For the choice of the cross-entropy loss, this can be written as

$$\underset{\lambda \in \mathcal{P}(\Delta)}{\text{maximize}} \; \underset{\phi \in \mathcal{F}}{\min} \; \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathbb{E}_{\boldsymbol{\delta}\sim\lambda|\mathbf{x}} \left[-\log\left(\mathbb{E}_{\mathbf{t}\sim\mathcal{N}(0,\sigma^2 I)}[\phi(\mathbf{x}+\mathbf{t}+\boldsymbol{\delta})_y] \right) \right] \right] \tag{22}$$

This seems have more to do with the choice of loss function than it does with the optimization procedure itself. Therefore is a question of whether the normal distribution is the best distribution to smooth over, and indeed there are some papers that discuss optimal smoothing distributions for various choices of Δ.

A.8 Relation to (robust) self-training (RST)

Given a classifier ϕ^{old}, RST can be written as

$$\underset{\phi \in \mathcal{F}}{\min} \; \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\underset{\boldsymbol{\delta}\in\Delta}{\max} \ell(\phi(\mathbf{x}+\boldsymbol{\delta}),y) \Big| x \in \mathcal{X}' \right] + \alpha \, \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\underset{\boldsymbol{\delta}\in\Delta}{\max} \ell(\phi(\mathbf{x}+\boldsymbol{\delta}),\phi^{\text{old}}(x)) \Big| x \in (\mathcal{X}')^c \right] \tag{23}$$

We could even think of this as a constrained problem

$$\underset{\phi \in \mathcal{F}}{\text{minimize}} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\underset{\boldsymbol{\delta}\in\Delta}{\max} \ell(\phi(\mathbf{x}+\boldsymbol{\delta}),y) \Big| x \in \mathcal{X}' \right] \tag{24}$$

$$\text{subject to} \quad \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\underset{\boldsymbol{\delta}\in\Delta}{\max} \ell(\phi(\mathbf{x}+\boldsymbol{\delta}),\phi^{\text{old}}(x)) \Big| x \in (\mathcal{X}')^c \right] < \gamma \tag{25}$$

B Proof of Proposition 3.1

Start by writing the primal problem (PII) in Lagragian form [REF]. Explicitly,

$$P_R^\star = \underset{\boldsymbol{\theta}\in\Theta, t\in L^p}{\min} \; \underset{\overline{\lambda}\in L_+^q}{\max} \; L_{\text{PV}}(\boldsymbol{\theta}, t, \overline{\lambda}), \tag{PIV}$$

where L_+^q denotes the subspace of almost everywhere non-negative functions of L^q, for $\frac{1}{p}+\frac{1}{q}=1$, for the Lagrangian

$$L_{\text{PV}}(\boldsymbol{\theta}, t, \overline{\lambda}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[t(\mathbf{x},y) \right] + \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) \left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}),y\big) - t(\mathbf{x},y) \right] d\mathbf{x} d\boldsymbol{\delta} dy$$

$$= \int t(\mathbf{x},y) \left[\mathfrak{p}(\mathbf{x},y) - \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) d\boldsymbol{\delta} \right] d\mathbf{x} dy \tag{26}$$

$$+ \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}),y\big) d\mathbf{x} d\boldsymbol{\delta} dy,$$

where we used the density $\mathfrak{p}$ of the data distribution $\mathcal{D}$. Then, notice that (PIV) can be separated into two parts as

$$P_R^\star = \underset{\boldsymbol{\theta}\in\Theta}{\min} \; p(\boldsymbol{\theta}) \tag{PV}$$

for the primal function

$$p(\boldsymbol{\theta}) = \underset{t\in L^p}{\min} \; \underset{\overline{\lambda}\in L_+^q}{\max} \; L_{\text{PV}}(\boldsymbol{\theta}, t, \overline{\lambda}). \tag{27}$$

Notice that in the objective of (27), $\boldsymbol{\theta}$ is constant. Since (26) is a linear function of t, $p(\boldsymbol{\theta})$ is the optimal value of a linear program parametrized by $\boldsymbol{\theta}$. Hence, strong duality holds [REF] and we obtain that

$$p(\boldsymbol{\theta}) = \underset{\overline{\lambda}\in L_+^q}{\max} \; d(\overline{\lambda}) \triangleq \underset{t\in L^p}{\min} \; L_{\text{PV}}(\boldsymbol{\theta}, t, \overline{\lambda}), \tag{28}$$

for the dual function $d(\overline{\lambda})$. Since t is unconstrained and L is linear in t, the dual function either vanishes if $\mathfrak{p}(\mathbf{x},y) = \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) d\boldsymbol{\delta}$ or diverges to $-\infty$. From (26) and (28), we thus obtain that

$$p(\boldsymbol{\theta}) = \underset{\overline{\lambda}\in L_+^q}{\max} \quad \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}+\boldsymbol{\delta}),y\big) d\mathbf{x} d\boldsymbol{\delta} dy$$

$$\text{subject to} \quad \int \overline{\lambda}(\mathbf{x},\boldsymbol{\delta},y) d\boldsymbol{\delta} = \mathfrak{p}(\mathbf{x},y) \tag{29}$$

To conclude, notice that since $\overline{\lambda}$ is almost everywhere non-negative, it must be that $\overline{\lambda}(\mathbf{x}, \boldsymbol{\delta}, y) = 0$ for all $\boldsymbol{\delta} \in \Delta$ whenever $\mathfrak{p}(\mathbf{x}, y) = 0$. The measure induced by $\overline{\lambda}$ is therefore absolutely continuous with respect $\mathcal{D}$. We can therefore rewrite (29) in terms of the Radon-Nykodim derivative $\lambda(\boldsymbol{\delta} \mid \mathbf{x}, y) = \overline{\lambda}(\mathbf{x}, \boldsymbol{\delta}, y) / \mathfrak{p}(\mathbf{x}, y)$, which yields (5) as desired.

C Proof of Proposition 3.2

Start by noticing that $\mathcal{P}^q$ is decomposable in $\mathcal{X} \times \mathcal{Y}$ [REF], i.e., if $h, h' \in \mathcal{P}^q$, then $\overline{h} \in \mathcal{P}^q$ for any measurable $\mathcal{Z} \in \mathcal{X} \times \mathcal{Y}$ with

$$\overline{h}(\mathbf{x}, \boldsymbol{\delta}, y) = \begin{cases} h(\mathbf{x}, \boldsymbol{\delta}, y), & (\mathbf{x}, \boldsymbol{\delta}, y) \in \mathcal{Z} \\ h'(\mathbf{x}, \boldsymbol{\delta}, y), & (\mathbf{x}, \boldsymbol{\delta}, y) \in \mathcal{Z}^c \end{cases}.$$

Indeed, L^q is decomposable for all q and all other constraints are pointwise on $\mathcal{X} \times \mathcal{Y}$. Namely, h and h' are non-negative, thus so is $\overline{h}$; they are absolutely continuous with respect to $p(\mathbf{x}, y)$, thus so is $\overline{h}$; and they are normalized along $\boldsymbol{\delta}$, thus so is $\overline{h}$. Additionally, notice that the integrand is *normal* in $\boldsymbol{x}$, i.e., everywhere continuous and measurable [REF], and $\mathcal{Y}$ is finite. Hence, we can use the variational results in [REF] to push the maximization inside the expectation and obtain

$$p(\boldsymbol{\theta}) \triangleq \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[u(\boldsymbol{\theta}, \mathbf{x}, y) \right]$$

for

$$u(\boldsymbol{\theta}, \mathbf{x}, y) \triangleq \max_{\substack{\lambda \in L_+^2, \\ \int \lambda(\boldsymbol{\delta}) d\boldsymbol{\delta} = 1}} \mathbb{E}_{\boldsymbol{\delta} \sim \lambda} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y) \right]. \tag{30}$$

Notice that the optimization in (30) is now performed separately for each data point. We therefore proceed by considering the maximization problem for a fixed, arbitrary point $(\mathbf{x}', y')$ and write $u'(\boldsymbol{\theta}) = u(\boldsymbol{\theta}, \mathbf{x}', y')$.

Notice now that since $\lambda \in L^2$, there exists a constant $c < \infty$ such that $\|\lambda^\star\|_{L_2}^2 < c$, where $\lambda^\star$ is the solution of the optimization problem in (30). Thus, we do not modify the value of (30) by including this constraint as in

$$u'(\boldsymbol{\theta}) = \max_{\lambda \in L_+^2} \mathbb{E}_{\boldsymbol{\delta} \sim \lambda} \left[\ell(f_{\boldsymbol{\theta}}(\mathbf{x}' + \boldsymbol{\delta}), y') \right]$$

$$\text{subject to} \quad \int \lambda(\boldsymbol{\delta}) d\boldsymbol{\delta} = 1, \quad \int \lambda(\boldsymbol{\delta})^2 d\boldsymbol{\delta} \le c, \tag{PVI}$$

Problem (PVI) is a convex (quadratic) program and due to our choice of c, it possesses a strictly feasible solution, namely $\lambda^\star$. Hence, (PVI) is strongly dual [REF] and we can write

$$u'(\boldsymbol{\theta}) = \min_{\gamma \ge 0, \, \mu \in \mathbb{R}} d_{\text{PVII}}(\boldsymbol{\theta}, \gamma, \mu), \tag{31}$$

for the dual function

$$d_{\text{PVII}}(\boldsymbol{\theta}, \gamma, \mu) = \max_{\lambda \in L_+^2} \int \left[\lambda(\boldsymbol{\delta}) \ell(f_{\boldsymbol{\theta}}(\mathbf{x}' + \boldsymbol{\delta}), y') - \gamma \lambda(\boldsymbol{\delta})^2 - \mu \lambda(\boldsymbol{\delta}) \right] d\boldsymbol{\delta} + \gamma c + \mu.$$

Once again employing the variational result from [REF] yields

$$\lambda^\star(\boldsymbol{\delta}) = \left[\frac{\ell(f_{\boldsymbol{\theta}}(\mathbf{x}' + \boldsymbol{\delta}), y') - \mu}{\gamma} \right]_+, \tag{32}$$

where once again $[z]_+ = \max(0, z)$ denotes the projection onto the non-negative orthant. From (PVI), μ is chosen so as to meet the normalization constraint, i.e., so that

$$\int \left[\frac{\ell(f_{\boldsymbol{\theta}}(\mathbf{x}' + \boldsymbol{\delta}), y') - \mu}{\gamma} \right]_+ d\boldsymbol{\delta} = 1.$$

To conclude, notice that due to the strong duality of (PVI), for each value of $c < \infty$ there is a value of $\gamma \ge 0$ such that (32) is a solution of (PVI) [REF]. What is more, $\gamma = 0$ implies that the constraint is not binding, in which case a smaller c can be used without modifying the solution of the problem. Finally, notice that $(\mathbf{x}', y')$ were chosen arbitrarily, so that (32) holds for all data point. Since the space is decomposable, these solutions can be "pieced together," yielding the desired result.

D Proof of Proposition 3.6

We prove the proposition in two steps. First, we characterize the parametrization gap and then the estimation gap. The result in Proposition 3.6 is obtained by combining Lemmata D.1 and D.2 using the union bound.

D.1 Parametrization gap

In this section, we are interested in the relation between the statistical problem (PI) and its dual problem. Namely,

$$D^\star \triangleq \max_{\nu \geq 0} \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \nu) \tag{DI}$$

for the Lagrangian

$$L(\boldsymbol{\theta}, \nu) = \mathbb{E}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x} + \boldsymbol{\delta}), y\big)\right] + \nu\Big[\mathbb{E}\left[\ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}), y\big)\right] - \epsilon\Big]. \tag{33}$$

We prove the following lemma.

Lemma D.1. *Under the conditions of Theorem 3.6, the value $D^\star$ of* (DI) *is related to the value $P^\star$ of* (PI) *as in*

$$P^\star - M\nu \leq D^\star \leq P^\star. \tag{34}$$

Proof. The result in Lemma D.1 is trivial when the hypothesis class $\mathcal{H}$ induced by the parametrization is convex. In this case, (PI) is a convex program and Assumption 3.4 (Slater's condition) implies that it is strongly dual [REF]. In other words, $P^\star = D^\star$. Hence, we are interested in the setting in which $\mathcal{H}$ is not convex, e.g., for CNNs, but is still a rich parametrization as per (8).

Now the upper bound in (34) is trivial from weak duality [?]. For the lower bound, consider the variational problem

$$\tilde{P}^\star \triangleq \min_{\phi \in \overline{\mathcal{H}}} \quad \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(\phi(\mathbf{x} + \boldsymbol{\delta}), y\big)\right]$$
$$\text{subject to} \quad \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}}\left[\ell\big(\phi(\mathbf{x}), y\big)\right] \leq \epsilon - M\alpha \tag{PVII}$$

for $\overline{\mathcal{H}} = \text{conv}(\mathcal{H})$. Let $\tilde{\phi}^\star \in \overline{\mathcal{H}}$ be a solution of (PVII) associated with $\delta^\star(\mathbf{x}, y)$, the perturbations that attains the maximum in its objective. Notice that (PVII) is now a convex optimization problem (recall that the pointwise maximum of convex functions is convex [REF:boyd chap 3.2]) that has a strictly feasible point $f_{\boldsymbol{\theta}'} \in \mathcal{H} \subset \overline{\mathcal{H}}$ (Assumption 3.4). Hence, it is strongly dual [REF] and

$$\tilde{P}^\star = \max_{\tilde{\nu} \geq 0} \min_{\phi \in \overline{\mathcal{H}}} \tilde{L}(\phi, \tilde{\nu}) = \tilde{L}(\tilde{\phi}^\star, \tilde{\nu}^\star), \tag{35}$$

where $\tilde{\nu}^\star$ achieves the maximum in (35) for the Lagrangian

$$\tilde{L}(\phi, \tilde{\nu}) = \mathbb{E}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(\phi(\mathbf{x} + \boldsymbol{\delta}), y\big)\right] + \nu\Big[\mathbb{E}\left[\ell\big(\phi(\mathbf{x}), y\big)\right] - \epsilon + M\alpha\Big] \tag{36}$$

For clarity, we omit the distribution $\mathcal{D}$ over which the expectations are taken.

To proceed, notice from (DI) that

$$D^\star \geq \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \nu), \quad \text{for all } \nu \geq 0,$$

and that since $\mathcal{H} \subseteq \overline{\mathcal{H}} = \text{conv}(\mathcal{H})$, we obtain

$$D^\star \geq \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \tilde{\nu}^\star) \geq \min_{\phi \in \overline{\mathcal{H}}} \tilde{L}(\phi, \tilde{\nu}^\star). \tag{37}$$

Using the strong duality of (PVII), (37) yields

$$D^\star \geq \min_{\phi \in \overline{\mathcal{H}}} \tilde{L}(\phi, \tilde{\nu}^\star) = \tilde{P}^\star = \mathbb{E}\Big[\ell\big(\phi^\star(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big)\Big]. \tag{38}$$

To obtain the lower bound in (34), suffices it then to show that $\mathbb{E}\Big[\ell\big(\phi^\star(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big)\Big] \geq P^\star - M\nu$.

To do so, notice from Assumption 3.3 that there exists $\tilde{\boldsymbol{\theta}}^\dagger \in \Theta$ such that $\sup_{\mathbf{x}} \left| \tilde{\phi}^\star(\mathbf{x}) - f_{\tilde{\boldsymbol{\theta}}^\dagger}(\mathbf{x}) \right| \le \alpha$. For these parameters, it holds that

$$
\left| \mathbb{E}\left[\ell\big(\phi(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) \right] - \mathbb{E}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right] \right| \le
$$
$$
\mathbb{E}\left[\left| \ell\big(\phi(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) - \max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right| \right] \le
$$
$$
\mathbb{E}\left[\left| \ell\big(\phi(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) - \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) \right| \right],
$$

where the first inequality is due to the convexity of the absolute value (Jensen's inequality) and in the second inequality stems from the fact that $\delta^\star$ is a suboptimal solution of $\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big)$. Using the Lipschitz continuity of the loss and Assumption 3.3, we conclude that

$$
\left| \tilde{P}^\star - \mathbb{E}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right] \right| \le M\, \mathbb{E}\left[\left| \phi(\mathbf{x} + \delta^\star(\mathbf{x}, y)) - f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \delta^\star(\mathbf{x}, y)) \right| \right] \le M\alpha. \tag{39}
$$

Using a similar argument, we also obtain that

$$
\left| \mathbb{E}\left[\ell(\tilde{\phi}^\star(\mathbf{x}), y) \right] - \mathbb{E}\left[\ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x}), y\big) \right] \right| \le M\alpha. \tag{40}
$$

Hence, given that $\tilde{\phi}^\star(\boldsymbol{x})$ is feasible for (PVII), (40) implies that $\tilde{\boldsymbol{\theta}}^\star$ is feasible for (PI). By optimality, $P^\star \le \mathbb{E}\left[\max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right]$. Back in (38), we conclude that

$$
\begin{aligned}
D^\star &\ge \mathbb{E}\left[\ell\big(\phi^\star(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) \right] \\
&\ge P^\star + \mathbb{E}\left[\ell\big(\phi^\star(\mathbf{x} + \delta^\star(\mathbf{x}, y)), y\big) - \max_{\boldsymbol{\delta} \in \Delta} \ell\big(f_{\tilde{\boldsymbol{\theta}}^\star}(\mathbf{x} + \boldsymbol{\delta}), y\big) \right] \\
&\ge P^\star - M\nu,
\end{aligned}
$$

where the last inequality stems from (39). $\qquad\square$

D.2 Empirical gap

We now proceed to evaluate the empirical gap between the statistical dual problem (DI) and its empirical version $(\widehat{\text{DI}})$.

Proposition D.2. *Let $\hat{\nu}$ be a solution of $(\widehat{\text{DI}})$ with a finite $D^\star$. Under the conditions of Theorem 3.6, there exists $\hat{\boldsymbol{\theta}}^\star \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \hat{\nu})$ such that*

$$
\left| D^\star - \hat{D}^\star \right| \le (1 + \overline{\nu}) \max(\zeta_R(N), \zeta_N(N)) \quad \text{and} \tag{41}
$$
$$
\mathbb{E}_{(\boldsymbol{x}, y) \sim \mathfrak{D}_i}\left[\ell_i\big(f_{\hat{\boldsymbol{\theta}}^\star}(\boldsymbol{x}), y\big) \right] \le c_i + \zeta_i(N_i). \tag{42}
$$

hold with probability $1 - 5\delta$, where $\overline{\nu} = \max(\hat{\nu}^\star, \nu^\star)$, for $\hat{\nu}^\star$ a solution of $(\widehat{\text{DI}})$ and $\nu^\star$ a solution of (DI), and $D^\star$ and $\hat{D}^\star$ as in (DI) and $(\widehat{\text{DI}})$ respectively.

Proof. **Near-optimality.** The proof relies on the following lemma characterizing the superdifferential of the dual function $\hat{d}(\nu) = \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \nu)$ from $(\widehat{\text{DI}})$. Explicitly, we say $p \in \mathbb{R}$ is a *supergradient* of $\hat{d}$ at ν if

$$
\hat{d}(\nu') \le \hat{d}(\nu) + p(\nu' - \nu), \text{ for all } \nu' \ge 0. \tag{43}
$$

The set of all supergradients of $\hat{d}$ at ν is called the *superdifferential* of $\hat{d}$ at ν and is denoted $\partial \hat{d}(\nu)$. Additionally, let $\Theta^\dagger(\nu) = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \nu)$ for the Lagrangian defined in (7) and define the constraint slack as

$$
s(\boldsymbol{\theta}) = \left[\frac{1}{N} \sum_{n=1}^{N} \ell\big(f_{\boldsymbol{\theta}}(\mathbf{x}_n), y_n\big) - \epsilon \right]_+. \tag{44}
$$

The following is sometimes known as Danskin's theorem.

Lemma D.3. *Assume that ℓ, it holds that*

$$\partial d(\boldsymbol{\mu}) = \text{conv}\left(\bigcup_{\boldsymbol{\theta}^\dagger \in \Theta^\dagger(\boldsymbol{\mu})} \boldsymbol{s}(\boldsymbol{\theta}^\dagger)\right). \tag{45}$$

Proof. See, e.g., [? , Thm. 2.87], noting that $\hat{L}(\boldsymbol{\theta}, \cdot)$ is affine for all $\boldsymbol{\theta} \in \Theta$, $\hat{L}(\cdot, \nu)$ is continuous for all $\nu \geq 0$, and Θ is compact. Hence, we meet conditions (i)–(iii) of the theorem. Considering that $\hat{d}(\nu) = \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \nu) = -\max_{\boldsymbol{\theta} \in \Theta} -\hat{L}(\boldsymbol{\theta}, \nu)$ yields the desired result. $\square$

By contradiction, suppose that for all $\hat{\boldsymbol{\theta}}^\dagger \in \hat{\Theta}^\dagger(\hat{\nu}^\star)$ it holds that

$$\frac{1}{N} \sum_{n=1}^{N} \ell\big(f_{\hat{\boldsymbol{\theta}}^\dagger}(\mathbf{x}_n), y_n\big) > \epsilon.$$

Then, from Lemma D.3, $\mathbf{0} \notin \partial d(\hat{\nu}^\star)$, which contradicts the optimality of $\hat{\nu}^\star$. Hence, there must be $\hat{\boldsymbol{\theta}}^\dagger \in \hat{\Theta}^\dagger(\hat{\boldsymbol{\mu}}^\star)$ such that $\frac{1}{N} \sum_{n=1}^{N} \ell\big(f_{\hat{\boldsymbol{\theta}}^\dagger}(\mathbf{x}_n), y_n\big) \leq \epsilon$. For those parameters, the uniform bound in Assumption 3.5, yields that, with probability $1 - \delta$ over the data,

$$\mathbb{E}_{(\boldsymbol{x},y) \sim \mathcal{D}}\Big[\ell\big(f_{\hat{\boldsymbol{\theta}}^\dagger}(\mathbf{x}), y\big)\Big] \leq \frac{1}{N} \sum_{n=1}^{N} \ell\big(f_{\hat{\boldsymbol{\theta}}^\dagger}(\boldsymbol{x}_n), y_n\big) + \zeta_N(N) \leq \epsilon + \zeta_N(N). \tag{46}$$

Near-optimality. Let $\nu^\star$ and $\hat{\nu}^\star$ be solutions of (DI) and $\widehat{\text{(DI)}}$ respectively and consider the set of dual minimizers

$$\Theta^\dagger(\nu) = \underset{\theta \in \Theta}{\arg\min}\, L(\boldsymbol{\theta}, \nu) \quad \text{and} \quad \hat{\Theta}^\dagger(\hat{\nu}) = \underset{\theta \in \Theta}{\arg\min}\, \hat{L}(\boldsymbol{\theta}, \hat{\nu})$$

for the Lagrangians defined in (33) and (7) respectively. Using the optimality $\nu^\star$, it holds that

$$D^\star - \hat{D}^\star = \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \nu^\star) - \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \hat{\nu}^\star)$$
$$\leq \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \nu^\star) - \min_{\boldsymbol{\theta} \in \Theta} \hat{L}(\boldsymbol{\theta}, \nu^\star).$$

Since $\hat{\boldsymbol{\theta}}^\dagger \in \hat{\Theta}^\dagger(\nu^\star)$ is suboptimal for $L(\boldsymbol{\theta}, \nu^\star)$, we get

$$D^\star - \hat{D}^\star \leq L(\hat{\boldsymbol{\theta}}^\dagger, \nu^\star) - \hat{L}(\hat{\boldsymbol{\theta}}^\dagger, \nu^\star). \tag{47}$$

Using a similar argument yields

$$D^\star - \hat{D}^\star \geq L(\boldsymbol{\theta}^\dagger, \hat{\nu}^\star) - \hat{L}(\boldsymbol{\theta}^\dagger, \hat{\nu}^\star) \tag{48}$$

for $\boldsymbol{\theta}^\dagger \in \Theta^\dagger(\hat{\nu}^\star)$. Thus, we obtain that

$$\left| D^\star - \hat{D}^\star \right| \leq \max\left\{ \left| L(\hat{\boldsymbol{\theta}}^\dagger, \nu^\star) - \hat{L}(\hat{\boldsymbol{\theta}}^\dagger, \nu^\star) \right|, \left| L(\boldsymbol{\theta}^\dagger, \hat{\nu}^\star) - \hat{L}(\boldsymbol{\theta}^\dagger, \hat{\nu}^\star) \right| \right\} \tag{49}$$

Using the empirical bound from Assumption 3.5, we obtain that

$$\left| L(\boldsymbol{\theta}, \nu) - \hat{L}(\boldsymbol{\theta}, \nu) \right| \leq \zeta_R(N) + \nu \zeta_N(N), \tag{50}$$

holds uniformly over $\boldsymbol{\theta}$ with probability $1 - 2\delta$.

Combining 46, (49), and (50) using the union bound concludes the proof. $\square$