CERTIFYING SOME DISTRIBUTIONAL ROBUSTNESS WITH PRINCIPLED ADVERSARIAL TRAINING

Aman Sinha[*,1], **Hongseok Namkoong**[*,2], **John Duchi**[1,3]
Departments of [1]Electrical Engineering, [2]Management Science & Engineering, [3]Statistics
Stanford University
Stanford, CA 94305
{amans,hnamk,jduchi}@stanford.edu

ABSTRACT

Neural networks are vulnerable to adversarial examples and researchers have proposed many heuristic attack and defense mechanisms. We address this problem through the principled lens of distributionally robust optimization, which guarantees performance under adversarial input perturbations. By considering a Lagrangian penalty formulation of perturbing the underlying data distribution in a Wasserstein ball, we provide a training procedure that augments model parameter updates with worst-case perturbations of training data. For smooth losses, our procedure provably achieves moderate levels of robustness with little computational or statistical cost relative to empirical risk minimization. Furthermore, our statistical guarantees allow us to efficiently certify robustness for the population loss. For imperceptible perturbations, our method matches or outperforms heuristic approaches.

1 INTRODUCTION

Consider the classical supervised learning problem, in which we minimize an expected loss $\mathbb{E}_{P_0}[\ell(\theta; Z)]$ over a parameter $\theta \in \Theta$, where $Z \sim P_0$, P_0 is a distribution on a space $\mathcal{Z}$, and ℓ is a loss function. In many systems, robustness to changes in the data-generating distribution P_0 is desirable, whether they be from covariate shifts, changes in the underlying domain (Ben-David et al., 2010), or adversarial attacks (Goodfellow et al., 2015; Kurakin et al., 2016). As deep networks become prevalent in modern performance-critical systems (perception for self-driving cars, automated detection of tumors), model failure is increasingly costly; in these situations, it is irresponsible to deploy models whose robustness and failure modes we do not understand or cannot certify.

Recent work shows that neural networks are vulnerable to adversarial examples; seemingly imperceptible perturbations to data can lead to misbehavior of the model, such as misclassification of the output (Goodfellow et al., 2015; Nguyen et al., 2015; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016). Consequently, researchers have proposed adversarial attack and defense mechanisms (Papernot et al., 2016a;b;c; Rozsa et al., 2016; Carlini & Wagner, 2017; He et al., 2017; Madry et al., 2017; Tramèr et al., 2017). These works provide an initial foundation for adversarial training, but it is challenging to rigorously identify the classes of attacks against which they can defend (or if they exist). Alternative approaches that provide formal verification of deep networks (Huang et al., 2017; Katz et al., 2017a;b) are NP-hard in general; they require prohibitive computational expense even on small networks. Recently, researchers have proposed convex relaxations of the NP-hard verification problem with some success (Kolter & Wong, 2017; Raghunathan et al., 2018), though they may be difficult to scale to large networks. In this context, our work is situated between these agendas: we develop efficient procedures with rigorous guarantees for *small to moderate* amounts of robustness.

We take the perspective of distributionally robust optimization and provide an adversarial training procedure with provable guarantees on its computational and statistical performance. We postulate a class $\mathcal{P}$ of distributions around the data-generating distribution P_0 and consider the problem

$$\underset{\theta \in \Theta}{\text{minimize}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]. \tag{1}$$

[*]Equal contribution

The choice of $\mathcal{P}$ influences robustness guarantees and computability; we develop robustness sets $\mathcal{P}$ with computationally efficient relaxations that apply even when the loss ℓ is non-convex. We provide an adversarial training procedure that, for smooth ℓ, enjoys convergence guarantees similar to non-robust approaches while *certifying* performance even for the worst-case population loss $\sup_{P\in\mathcal{P}} \mathbb{E}_P[\ell(\theta;Z)]$. On a simple implementation in Tensorflow, our method takes 5–10$\times$ as long as stochastic gradient methods for empirical risk minimization (ERM), matching runtimes for other adversarial training procedures (Goodfellow et al., 2015; Kurakin et al., 2016; Madry et al., 2017). We show that our procedure—which learns to protect against adversarial perturbations in the training dataset—generalizes, allowing us to train a model that prevents attacks to the test dataset.

We briefly overview our approach. Let $c : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$, where $c(z, z_0)$ is the "cost" for an adversary to perturb z_0 to z (we typically use $c(z, z_0) = \|z - z_0\|_p^2$ with $p \geq 1$). We consider the robustness region $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$, a ρ-neighborhood of the distribution P_0 under the Wasserstein metric $W_c(\cdot,\cdot)$ (see Section 2 for a formal definition). For deep networks and other complex models, this formulation of problem (1) is intractable with arbitrary ρ. Instead, we consider its Lagrangian relaxation for a fixed penalty parameter $\gamma \geq 0$, resulting in the reformulation

$$\underset{\theta\in\Theta}{\text{minimize}} \left\{ F(\theta) := \sup_P \left\{\mathbb{E}_P[\ell(\theta;Z)] - \gamma W_c(P,P_0)\right\} = \mathbb{E}_{P_0}[\phi_\gamma(\theta;Z)] \right\} \tag{2a}$$

$$\text{where } \ \phi_\gamma(\theta;z_0) := \sup_{z\in\mathcal{Z}} \left\{\ell(\theta;z) - \gamma c(z,z_0)\right\}. \tag{2b}$$

(See Proposition 1 for a rigorous statement of these equalities.) Here, we have replaced the usual loss $\ell(\theta;Z)$ by the robust surrogate $\phi_\gamma(\theta;Z)$; this surrogate (2b) allows adversarial perturbations of the data z, modulated by the penalty γ. We typically solve the penalty problem (2) with P_0 replaced by the empirical distribution $\widehat{P}_n$, as P_0 is unknown (we refer to this as the penalty problem below).

The key feature of the penalty problem (2) is that moderate levels of robustness—in particular, defense against imperceptible adversarial perturbations—are achievable at essentially no computational or statistical cost for *smooth losses* ℓ. Specifically, for large enough penalty γ (by duality, small enough robustness ρ), the function $z \mapsto \ell(\theta;z) - \gamma c(z,z_0)$ in the robust surrogate (2b) is strongly concave and hence *easy* to optimize if $\ell(\theta,z)$ is smooth in z. Consequently, stochastic gradient methods applied to problem (2) have similar convergence guarantees as for non-robust methods (ERM). In Section 3, we provide a *certificate of robustness* for any ρ; we give an efficiently computable data-dependent upper bound on the worst-case loss $\sup_{P:W_c(P,P_0)\leq\rho} \mathbb{E}_P[\ell(\theta;Z)]$. That is, the worst-case performance of the output of our principled adversarial training procedure is guaranteed to be no worse than this certificate. Our bound is tight when $\rho = \widehat{\rho}_n$, the achieved robustness for the empirical objective. These results suggest advantages of networks with smooth activations rather than ReLU's. We experimentally verify our results in Section 4 and show that we match or achieve state-of-the-art performance on a variety of adversarial attacks.

Robust optimization and adversarial training The standard robust-optimization approach minimizes losses of the form $\sup_{u\in\mathcal{U}} \ell(\theta;z+u)$ for some uncertainty set $\mathcal{U}$ (Ratliff et al., 2006; Ben-Tal et al., 2009; Xu et al., 2009). Unfortunately, this approach is intractable except for specially structured losses, such as the composition of a linear and simple convex function (Ben-Tal et al., 2009; Xu et al., 2009; 2012). Nevertheless, this robust approach underlies recent advances in adversarial training (Szegedy et al., 2013; Goodfellow et al., 2015; Papernot et al., 2016b; Carlini & Wagner, 2017; Madry et al., 2017), which heuristically perturb data during a stochastic optimization procedure.

One such heuristic uses a locally linearized loss function (proposed with $p = \infty$ as the "fast gradient sign method" (Goodfellow et al., 2015)):

$$\Delta_{x_i}(\theta) := \underset{\|\eta\|_p\leq\epsilon}{\text{argmax}} \{\nabla_x \ell(\theta;(x_i,y_i))^T\eta\} \ \text{ and perturb } \ x_i \to x_i + \Delta_{x_i}(\theta). \tag{3}$$

One form of adversarial training trains on the losses $\ell(\theta;(x_i + \Delta_{x_i}(\theta), y_i))$ (Goodfellow et al., 2015; Kurakin et al., 2016), while others perform iterated variants (Papernot et al., 2016b; Carlini & Wagner, 2017; Madry et al., 2017; Tramèr et al., 2017). Madry et al. (2017) observe that these procedures attempt to optimize the objective $\mathbb{E}_{P_0}[\sup_{\|u\|_p\leq\epsilon} \ell(\theta;Z+u)]$, a constrained version of the penalty problem (2). This notion of robustness is typically intractable: the inner supremum is generally non-concave in u, so it is unclear whether model-fitting with these techniques converges, and there are possibly worst-case perturbations these techniques do not find. Indeed, it is NP-hard to find worst-case perturbations when deep networks use ReLU activations, suggesting difficulties for fast and iterated heuristics (see Lemma 2 in Appendix B). Smoothness, which can be obtained in standard deep architectures with exponential linear units (ELU's) (Clevert et al., 2015), allows us to find Lagrangian worst-case perturbations with low computational cost.

Distributionally robust optimization To situate the current work, we review some of the substantial body of work on robustness and learning. The choice of $\mathcal{P}$ in the robust objective (1) affects both the richness of the uncertainty set we wish to consider as well as the tractability of the resulting optimization problem. Previous approaches to distributional robustness have considered finite-dimensional parametrizations for $\mathcal{P}$, such as constraint sets for moments, support, or directional deviations (Chen et al., 2007; Delage & Ye, 2010; Goh & Sim, 2010), as well as non-parametric distances for probability measures such as f-divergences (Ben-Tal et al., 2013; Bertsimas et al., 2013; Lam & Zhou, 2015; Miyato et al., 2015; Duchi et al., 2016; Namkoong & Duchi, 2016), and Wasserstein distances (Esfahani & Kuhn, 2015; Shafieezadeh-Abadeh et al., 2015; Blanchet et al., 2016). In constrast to f-divergences (e.g. χ^2- or Kullback-Leibler divergences) which are effective when the support of the distribution P_0 is fixed, a Wasserstein ball around P_0 includes distributions Q with different support and allows (in a sense) robustness to unseen data.

Many authors have studied tractable classes of uncertainty sets $\mathcal{P}$ and losses ℓ. For example, Ben-Tal et al. (2013) and Namkoong & Duchi (2017) use convex optimization approaches for f-divergence balls. For worst-case regions $\mathcal{P}$ formed by Wasserstein balls, Esfahani & Kuhn (2015), Shafieezadeh-Abadeh et al. (2015), and Blanchet et al. (2016) show how to convert the saddle-point problem (1) to a regularized ERM problem, but this is possible only for a limited class of convex losses ℓ and costs c. In this work, we treat a much larger class of losses and costs and provide direct solution methods for a Lagrangian relaxation of the saddle-point problem (1). One natural application area is in domain adaptation (Lee & Raginsky, 2017); concurrently with this work, Lee & Raginsky provide guarantees similar to ours for the empirical minimizer of the robust saddle-point problem (1) and give specialized bounds for domain adaptation problems. In contrast, our approach is to use the distributionally robust approach to both defend against imperceptible adversarial perturbations and develop efficient optimization procedures.

2 Proposed approach

Our approach is based on the following simple insight: assume that the function $z \mapsto \ell(\theta; z)$ is smooth, meaning there is some L for which $\nabla_z \ell(\theta; \cdot)$ is L-Lipschitz. Then for any $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$ 1-strongly convex in its first argument, a Taylor expansion yields

$$\ell(\theta; z') - \gamma c(z', z_0) \le \ell(\theta; z) - \gamma c(z, z_0) + \langle \nabla_z (\ell(\theta; z) - \gamma c(z, z_0)), z' - z \rangle + \frac{L - \gamma}{2} \left\| z - z' \right\|_2^2 . \tag{4}$$

For $\gamma \ge L$ this is the first-order condition for $(\gamma - L)$-strong concavity of $z \mapsto (\ell(\theta; z) - \gamma c(z, z_0))$. Thus, whenever the loss is smooth enough in z and the penalty γ is large enough (corresponding to less robustness), computing the surrogate (2b) is a strongly-concave optimization problem.

We leverage the insight (4) to show that as long as we do not require *too much* robustness, this strong concavity approach (4) provides a computationally efficient and principled approach for robust optimization problems (1). Our starting point is a duality result for the minimax problem (1) and its Lagrangian relaxation for Wasserstein-based uncertainty sets, which makes the connections between distributional robustness and the "lazy" surrogate (2b) clear. We then show (Section 2.1) how stochastic gradient descent methods can efficiently find minimizers (in the convex case) or approximate stationary points (when ℓ is non-convex) for our relaxed robust problems.

Wasserstein robustness and duality Wasserstein distances define a notion of closeness between distributions. Let $\mathcal{Z} \subset \mathbb{R}^m$, and let $(\mathcal{Z}, \mathcal{A}, P_0)$ be a probability space. Let the transportation cost $c : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ be nonnegative, lower semi-continuous, and satisfy $c(z, z) = 0$. For example, for a differentiable convex $h : \mathcal{Z} \to \mathbb{R}$, the Bregman divergence $c(z, z_0) = h(z) - h(z_0) - \langle \nabla h(z_0), z - z_0 \rangle$ satisfies these conditions. For probability measures P and Q supported on $\mathcal{Z}$, let $\Pi(P, Q)$ denote their couplings, meaning measures M on $\mathcal{Z}^2$ with $M(A, \mathcal{Z}) = P(A)$ and

$M(\mathcal{Z}, A) = Q(A)$. The Wasserstein distance between P and Q is

$$W_c(P, Q) := \inf_{M \in \Pi(P,Q)} \mathbb{E}_M[c(Z, Z')].$$

For $\rho \geq 0$ and distribution P_0, we let $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$, considering the Wasserstein form of the robust problem (1) and its Lagrangian relaxation (2) with $\gamma \geq 0$. The following duality result (Blanchet & Murthy, 2016) gives the equality (2) for the relaxation and an analogous result for the problem (1). We give an alternative proof in Appendix C.1 for convex, continuous cost functions.

Proposition 1. *Let $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ and $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be continuous. Let $\phi_\gamma(\theta; z_0) = \sup_{z \in \mathcal{Z}} \{\ell(\theta; z) - \gamma c(z, z_0)\}$ be the robust surrogate* (2b). *For any distribution Q and any $\rho > 0$,*

$$\sup_{P : W_c(P,Q) \leq \rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\gamma \geq 0} \left\{\gamma\rho + \mathbb{E}_Q[\phi_\gamma(\theta; Z)]\right\}, \tag{5}$$

and for any $\gamma \geq 0$, we have

$$\sup_P \{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, Q)\} = \mathbb{E}_Q[\phi_\gamma(\theta; Z)]. \tag{6}$$

Leveraging the insight (4), we give up the requirement that we wish a prescribed amount ρ of robustness (solving the worst-case problem (1) for $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$) and focus instead on the Lagrangian penalty problem (2) and its empirical counterpart

$$\operatorname*{minimize}_{\theta \in \Theta} \left\{ F_n(\theta) := \sup_P \left\{ \mathbb{E}[\ell(\theta; Z)] - \gamma W_c(P, \widehat{P}_n) \right\} = \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] \right\}. \tag{7}$$

2.1 Optimizing the robust loss by stochastic gradient descent

We now develop stochastic gradient-type methods for the relaxed robust problem (7), making clear the computational benefits of relaxing the strict robustness requirements of formulation (5). We begin with assumptions we require, which roughly quantify the amount of robustness we can provide.

Assumption A. *The function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is continuous. For each $z_0 \in \mathcal{Z}$, $c(\cdot, z_0)$ is 1-strongly convex with respect to the norm $\|\cdot\|$.*

To guarantee that the robust surrogate (2b) is tractably computable, we also require a few smoothness assumptions. Let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$; we abuse notation by using the same norm $\|\cdot\|$ on Θ and $\mathcal{Z}$, though the specific norm is clear from context.

Assumption B. *The loss $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ satisfies the Lipschitzian smoothness conditions*

$$\|\nabla_\theta \ell(\theta; z) - \nabla_\theta \ell(\theta'; z)\|_* \leq L_{\theta\theta} \|\theta - \theta'\|, \ \ \|\nabla_z \ell(\theta; z) - \nabla_z \ell(\theta; z')\|_* \leq L_{\mathsf{zz}} \|z - z'\|,$$
$$\|\nabla_\theta \ell(\theta; z) - \nabla_\theta \ell(\theta; z')\|_* \leq L_{\theta\mathsf{z}} \|z - z'\|, \ \ \|\nabla_z \ell(\theta; z) - \nabla_z \ell(\theta'; z)\|_* \leq L_{\mathsf{z}\theta} \|\theta - \theta'\|.$$

These properties guarantee both (i) the well-behavedness of the robust surrogate ϕ_γ and (ii) its efficient computability. Making point (i) precise, Lemma 1 shows that if γ is large enough and Assumption B holds, the surrogate ϕ_γ is still smooth. Throughout, we assume $\Theta \subseteq \mathbb{R}^d$.

Lemma 1. *Let $f : \Theta \times \mathcal{Z} \to \mathbb{R}$ be differentiable and λ-strongly concave in z with respect to the norm $\|\cdot\|$, and define $\bar{f}(\theta) = \sup_{z \in \mathcal{Z}} f(\theta, z)$. Let $\mathsf{g}_\theta(\theta, z) = \nabla_\theta f(\theta, z)$ and $\mathsf{g}_\mathsf{z}(\theta, z) = \nabla_z f(\theta, z)$, and assume g_θ and g_z satisfy Assumption B with $\ell(\theta; z)$ replaced with $f(\theta, z)$. Then $\bar{f}$ is differentiable, and letting $z^\star(\theta) = \operatorname{argmax}_{z \in \mathcal{Z}} f(\theta, z)$, we have $\nabla \bar{f}(\theta) = \mathsf{g}_\theta(\theta, z^\star(\theta))$. Moreover,*

$$\|z^\star(\theta_1) - z^\star(\theta_2)\| \leq \frac{L_{\mathsf{z}\theta}}{\lambda} \|\theta_1 - \theta_2\| \quad \textit{and} \quad \left\|\nabla \bar{f}(\theta) - \nabla \bar{f}(\theta')\right\|_\star \leq \left(L_{\theta\theta} + \frac{L_{\theta\mathsf{z}} L_{\mathsf{z}\theta}}{\lambda}\right) \|\theta - \theta'\|.$$

See Section C.2 for the proof. Fix $z_0 \in \mathcal{Z}$ and focus on the ℓ_2-norm case where $c(z, z_0)$ satisfies Assumption A with $\|\cdot\|_2$. Noting that $f(\theta, z) := \ell(\theta, z) - \gamma c(z, z_0)$ is $(\gamma - L_{\mathsf{zz}})$-strongly concave from the insight (4) (with $L := L_{\mathsf{zz}}$), let us apply Lemma 1. Under Assumptions A, B, $\phi_\gamma(\cdot; z_0)$ then has $L = L_{\theta\theta} + \frac{L_{\theta\mathsf{z}} L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+}$-Lipschitz gradients, and

$$\nabla_\theta \phi_\gamma(\theta; z_0) = \nabla_\theta \ell(\theta; z^\star(z_0, \theta)) \ \text{ where } \ z^\star(z_0, \theta) = \operatorname*{argmax}_{z \in \mathcal{Z}} \{\ell(\theta; z) - \gamma c(z, z_0)\}.$$

Algorithm 1 Distributionally robust optimization with adversarial training

INPUT: Sampling distribution P_0, constraint sets Θ and $\mathcal{Z}$, stepsize sequence $\{\alpha_t > 0\}_{t=0}^{T-1}$
for $t = 0, \ldots, T-1$ **do**
  Sample $z^t \sim P_0$ and find an ϵ-approximate maximizer $\widehat{z}^t$ of $\ell(\theta^t; z) - \gamma c(z, z^t)$
  $\theta^{t+1} \leftarrow \mathsf{Proj}_\Theta(\theta^t - \alpha_t \nabla_\theta \ell(\theta^t; \widehat{z}^t))$

This motivates Algorithm 1, a stochastic-gradient approach for the penalty problem (7). The benefits of Lagrangian relaxation become clear here: for $\ell(\theta; z)$ smooth in z and γ large enough, gradient ascent on $\ell(\theta^t; z) - \gamma c(z, z^t)$ in z converges linearly and we can compute (approximate) $\widehat{z}^t$ efficiently (we initialize our inner gradient ascent iterations with the sampled natural example z^t).

Convergence properties of Algorithm 1 depend on the loss ℓ. When ℓ is convex in θ and γ is large enough that $z \mapsto (\ell(\theta; z) - \gamma c(z, z_0))$ is concave for all $(\theta, z_0) \in \Theta \times \mathcal{Z}$, we have a stochastic monotone variational inequality, which is efficiently solvable (Juditsky et al., 2011; Chen et al., 2014) with convergence rate $1/\sqrt{T}$. When the loss ℓ is nonconvex in θ, the following theorem guarantees convergence to a stationary point of problem (7) at the same rate when $\gamma \ge L_{\mathsf{zz}}$. Recall that $F(\theta) = \mathbb{E}_{P_0}[\phi_\gamma(\theta; Z)]$ is the robust surrogate objective for the Lagrangian relaxation (2).

Theorem 2 (Convergence of Nonconvex SGD)**.** *Let Assumptions A and B hold with the ℓ_2-norm and let $\Theta = \mathbb{R}^d$. Let $\Delta_F \ge F(\theta^0) - \inf_\theta F(\theta)$. Assume $\mathbb{E}[\|\nabla F(\theta) - \nabla_\theta \phi_\gamma(\theta, Z)\|_2^2] \le \sigma^2$ and take constant stepsizes $\alpha = \sqrt{\frac{\Delta_F}{L_\phi T \sigma^2}}$ where $L_\phi := L_{\theta\theta} + \frac{L_{\theta \mathsf{z}} L_{\mathsf{z}\theta}}{\gamma - L_{\mathsf{zz}}}$. For $T \ge \frac{L_\phi \Delta_F}{\sigma^2}$, Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - \frac{4L_{\theta\mathsf{z}}^2}{\gamma - L_{\mathsf{zz}}}\epsilon \le 4\sigma\sqrt{\frac{L_\phi \Delta_F}{T}}.$$

See Section C.3 for the proof. We make a few remarks. First, the condition $\mathbb{E}[\|\nabla F(\theta) - \nabla_\theta \phi_\gamma(\theta, Z)\|_2^2] \le \sigma^2$ holds (to within a constant factor) whenever $\|\nabla_\theta \ell(\theta, z)\|_2 \le \sigma$ for all θ, z. Theorem 2 shows that the stochastic gradient method achieves the rates of convergence on the penalty problem (7) achievable in standard smooth non-convex optimization (Ghadimi & Lan, 2013). The accuracy parameter ϵ has a *fixed* effect on optimization accuracy, independent of T: approximate maximization has limited effects.

Key to the convergence guarantee of Theorem 2 is that the loss ℓ is smooth in z: the inner supremum (2b) is NP-hard to compute for non-smooth deep networks (see Lemma 2 in Section B for a proof of this for ReLU's). The smoothness of ℓ is essential so that a penalized version $\ell(\theta, z) - \gamma c(z, z_0)$ is concave in z (which can be approximately verified by computing Hessians $\nabla_{zz}^2 \ell(\theta, z)$ for each training datapoint), allowing computation and our coming certificates of optimality. Replacing ReLU's with sigmoids or ELU's (Clevert et al., 2015) allows us to apply Theorem 2, making distributionally robust optimization tractable for deep learning.

In supervised-learning scenarios, we are often interested in adversarial perturbations only to feature vectors (and not labels). Letting $Z = (X, Y)$ where X denotes the feature vector (covariates) and Y the label, this is equivalent to defining the Wasserstein cost function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$ by

$$c(z, z') := c_x(x, x') + \infty \cdot \mathbf{1}\{y \ne y'\} \tag{8}$$

where $c_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the transportation cost for the feature vector X. All of our results suitably generalize to this setting with minor modifications to the robust surrogate (2b) and the above assumptions (see Section D). Similarly, our distributionally robust framework (2) is general enough to consider adversarial perturbations to only an arbitrary subset of coordinates in Z. For example, it may be appropriate in certain applications to hedge against adversarial perturbations to a small fixed region of an image (Brown et al., 2017). By suitably modifying the cost function $c(z, z')$ to take value ∞ outside this small region, our general formulation covers such variants.

3 CERTIFICATE OF ROBUSTNESS AND GENERALIZATION

From results in the previous section, Algorithm 1 provably learns to protect against adversarial perturbations of the form (7) on the training dataset. Now we show that such procedures generalize, allowing us to prevent attacks on the test set. Our subsequent results hold uniformly over the space of parameters $\theta \in \Theta$, including θ_{WRM}, the output of the stochastic gradient descent procedure in Section 2.1. Our first main result, presented in Section 3.1, gives a data-dependent upper bound on the population worst-case objective $\sup_{P: W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta;Z)]$ for any arbitrary level of robustness ρ; this bound is optimal for $\rho = \widehat{\rho}_n$, the level of robustness achieved for the empirical distribution by solving (7). Our bound is efficiently computable and hence *certifies* a level of robustness for the worst-case population objective. Second, we show in Section 3.2 that adversarial perturbations on the training set (in a sense) generalize: solving the empirical penalty problem (7) guarantees a similar level of robustness as directly solving its population counterpart (2).

3.1 Certificate of robustness

Our main result in this section is a data-dependent upper bound for the worst-case population objective: $\sup_{P: W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta;Z)] \le \gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta;Z)] + O(1/\sqrt{n})$ for all $\theta \in \Theta$, with high probability. To make this rigorous, fix $\gamma > 0$, and consider the worst-case perturbation, typically called the *transportation map* or Monge map (Villani, 2009),

$$T_\gamma(\theta; z_0) := \operatorname*{argmax}_{z\in\mathcal{Z}} \{\ell(\theta;z) - \gamma c(z, z_0)\}. \tag{9}$$

Under our assumptions, T_γ is easily computable when $\gamma \ge L_{\mathsf{zz}}$. Letting δ_z denote the point mass at z, Proposition 1 shows the empirical maximizers of the Lagrangian formulation (6) are attained by

$$\begin{aligned} P_n^*(\theta) &:= \operatorname*{argmax}_P \left\{ \mathbb{E}_P[\ell(\theta;Z)] - \gamma W_c(P, \widehat{P}_n) \right\} = \frac{1}{n}\sum_{i=1}^n \delta_{T_\gamma(\theta,Z_i)} \quad \text{and} \\ \widehat{\rho}_n(\theta) &:= W_c(P_n^*(\theta), \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta;Z), Z)]. \end{aligned} \tag{10}$$

Our results imply, in particular, that the empirical worst-case loss $\mathbb{E}_{P_n^*}[\ell(\theta;Z)]$ gives a *certificate of robustness* to (population) Wasserstein perturbations up to level $\widehat{\rho}_n$. $\mathbb{E}_{P_n^*(\theta)}[\ell(\theta;Z)]$ is efficiently computable via (10), providing a data-dependent guarantee for the worst-case population loss.

Our bound relies on the usual covering numbers for the model class $\{\ell(\theta;\cdot) : \theta \in \Theta\}$ as the notion of complexity (e.g. van der Vaart & Wellner, 1996), so, despite the infinite-dimensional problem (7), we retain the same uniform convergence guarantees typical of empirical risk minimization. Recall that for a set V, a collection $v_1, \ldots, v_N$ is an *ϵ-cover* of V in norm $\|\cdot\|$ if for each $v \in \mathcal{V}$, there exists v_i such that $\|v - v_i\| \le \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is

$$N(V, \epsilon, \|\cdot\|) := \inf\left\{N \in \mathbb{N} \mid \text{there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\|\right\}.$$

For $\mathcal{F} := \{\ell(\theta,\cdot) : \theta \in \Theta\}$ equipped with the $L^\infty(\mathcal{Z})$ norm $\|f\|_{L^\infty(\mathcal{Z})} := \sup_{z\in\mathcal{Z}} |f(z)|$, we state our results in terms of $\|\cdot\|_{L^\infty(\mathcal{Z})}$-covering numbers of $\mathcal{F}$. To ease notation, we let

$$\epsilon_n(t) := \gamma b_1 \sqrt{\frac{M_\ell}{n}} \int_0^1 \sqrt{\log N(\mathcal{F}, M_\ell\epsilon, \|\cdot\|_{L^\infty(\mathcal{Z})})}\, d\epsilon + b_2 M_\ell \sqrt{\frac{t}{n}}$$

where b_1, b_2 are numerical constants.

We are now ready to state the main result of this section. We first show from the duality result (6) that we can provide an upper bound for the worst-case population performance for any level of robustness ρ. For $\rho = \widehat{\rho}_n(\theta)$ and $\theta = \theta_{\mathrm{WRM}}$, this certificate is (in a sense) tight as we see below.

Theorem 3. *Assume $|\ell(\theta;z)| \le M_\ell$ for all $\theta \in \Theta$ and $z \in \mathcal{Z}$. Then, for a fixed $t > 0$ and numerical constants $b_1, b_2 > 0$, with probability at least $1 - e^{-t}$, simultaneously for all $\theta \in \Theta$, $\rho \ge 0$, $\gamma \ge 0$,*

$$\sup_{P: W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta;Z)] \le \gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta;Z)] + \epsilon_n(t). \tag{11}$$

In particular, if $\rho = \widehat{\rho}_n(\theta)$ then with probability at least $1 - e^{-t}$, for all $\theta \in \Theta$

$$\begin{aligned} \sup_{P: W_c(P,P_0)\le\widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta;Z)] &\le \gamma\widehat{\rho}_n(\theta) + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta;Z)] + \epsilon_n(t) \\ &= \sup_{P: W_c(P,\widehat{P}_n)\le\widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta;Z)] + \epsilon_n(t). \end{aligned} \tag{12}$$

See Section C.4 for its proof. We now give a concrete variant of Theorem 3 for Lipschitz functions. When Θ is finite-dimensional ($\Theta \subset \mathbb{R}^d$), Theorem 3 provides a robustness guarantee scaling linearly with d despite the infinite-dimensional Wasserstein penalty. Assuming there exist $\theta_0 \in \Theta$, $M_{\theta_0} < \infty$ such that $|\ell(\theta_0; z)| \le M_{\theta_0}$ for all $z \in \mathcal{Z}$, we have the following corollary (see proof in Section C.5).

Corollary 1. *Let $\ell(\cdot; z)$ be L-Lipschitz with respect to some norm $\|\cdot\|$ for all $z \in \mathcal{Z}$. Assume that $\Theta \subset \mathbb{R}^d$ satisfies $\mathrm{diam}(\Theta) = \sup_{\theta,\theta' \in \Theta} \|\theta - \theta'\| < \infty$. Then, the bounds* (11) *and* (12) *hold with*

$$\epsilon_n(t) = b_1 \sqrt{\frac{d(L\,\mathrm{diam}(\Theta) + M_{\theta_0})}{n}} + b_2(L\,\mathrm{diam}(\Theta) + M_{\theta_0})\sqrt{\frac{t}{n}} \tag{13}$$

for some numerical constants $b_1, b_2 > 0$.

A key consequence of the bound (11) is that $\gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]$ *certifies* robustness for the worst-case population objective for any ρ and θ. For a given θ, this certificate is tightest at the achieved level of robustness $\widehat{\rho}_n(\theta)$, as noted in the refined bound (12) which follows from the duality result

$$\underbrace{\mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]}_{\text{surrogate loss}} + \underbrace{\gamma\widehat{\rho}_n(\theta)}_{\text{robustness}} = \sup_{P: W_c(P, \widehat{P}_n) \le \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] = \mathbb{E}_{P_n^*(\theta)}[\ell(\theta; Z)]. \tag{14}$$

(See Section C.4 for a proof of these equalities.) We expect θ_{WRM}, the output of Algorithm 1, to be close to the minimizer of the surrogate loss $\mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]$ and therefore have the best guarantees. Most importantly, the certificate (14) is easy to compute via expression (10): as noted in Section 2.1, the mappings $T(\theta, Z_i)$ are efficiently computable for large enough γ, and $\widehat{\rho}_n = \mathbb{E}_{\widehat{P}_n}[c(T(\theta, Z), Z)]$.

The bounds (11)–(13) may be too large—because of their dependence on covering numbers and dimension—for practical use in security-critical applications. With that said, the strong duality result, Proposition 1, still applies to any distribution. In particular, given a collection of test examples Z_i^{test}, we may interrogate possible losses under perturbations for the *test* examples by noting that, if $\widehat{P}_{\mathrm{test}}$ denotes the empirical distribution on the test set (say, with putative assigned labels), then

$$\frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n} \sup_{z: c(z, Z_i^{\mathrm{test}}) \le \rho} \{\ell(\theta; z)\} \le \sup_{P: W_c(P, \widehat{P}_{\mathrm{test}}) \le \rho} \mathbb{E}_P[\ell(\theta; Z)] \le \gamma\rho + \mathbb{E}_{\widehat{P}_{\mathrm{test}}}[\phi_\gamma(\theta; Z)] \tag{15}$$

for all $\gamma, \rho \ge 0$. Whenever γ is large enough (so that this is tight for small ρ), we may efficiently compute the Monge-map (9) and the test loss (15) to guarantee bounds on the sensitivity of a parameter θ to a particular sample and predicted labeling based on the sample.

3.2 Generalization of Adversarial Examples

We can also show that the level of robustness on the training set generalizes. Our starting point is Lemma 1, which shows that $T_\gamma(\cdot; z)$ is smooth under Assumptions A and B:

$$\|T_\gamma(\theta_1; z) - T_\gamma(\theta_2; z)\| \le \frac{L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+} \|\theta_1 - \theta_2\| \tag{16}$$

for all θ_1, θ_2, where we recall that L_{zz} is the Lipschitz constant of $\nabla_z \ell(\theta; z)$. Leveraging this smoothness, we show that $\widehat{\rho}_n(\theta) = \mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta; Z), Z)]$, the level of robustness achieved for the empirical problem, concentrates uniformly around its population counterpart.

Theorem 4. *Let $\mathcal{Z} \subset \{z \in \mathbb{R}^m : \|z\| \le M_{\mathsf{z}}\}$ so that $\|Z\| \le M_{\mathsf{z}}$ almost surely and assume either that (i) $c(\cdot, \cdot)$ is L_{c}-Lipschitz over $\mathcal{Z}$ with respect to the norm $\|\cdot\|$ in each argument, or (ii) that $\ell(\theta, z) \in [0, M_\ell]$ and $z \mapsto \ell(\theta, z)$ is γL_{c}-Lipschitz for all $\theta \in \Theta$.*

If Assumptions A and B hold, then with probability at least $1 - e^{-t}$,

$$\sup_{\theta \in \Theta} |\mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T_\gamma(\theta; Z), Z)]| \le 4B\sqrt{\frac{1}{n}\left(t + \log N\left(\Theta, \frac{[\gamma - L_{\mathsf{zz}}]_+ t}{4L_{\mathsf{c}} L_{\mathsf{z}\theta}}, \|\cdot\|\right)\right)}. \tag{17}$$

where $B = L_{\mathsf{c}} M_{\mathsf{z}}$ under assumption (i) and $B = M_\ell/\gamma$ under assumption (ii).

See Section C.6 for the proof. For $\Theta \subset \mathbb{R}^d$, we have $\log N(\Theta, \epsilon, \|\cdot\|) \le d \log(1 + \frac{\mathrm{diam}(\Theta)}{\epsilon})$ so that the bound (30) gives the usual $\sqrt{d/n}$ generalization rate for the distance between adversarial perturbations and natural examples. Another consequence of Theorem 4 is that $\widehat{\rho}_n(\theta_{\mathrm{WRM}})$ in the certificate (12) is positive as long as the loss ℓ is not completely invariant to data. To see this, note from the optimality conditions for $T_\gamma(\theta; Z)$ that $\mathbb{E}_{P_0}[c(T_\gamma(\theta; Z), Z)] = 0$ iff $\nabla_z \ell(\theta; z) = 0$ almost surely, and hence for large enough n, we have $\widehat{\rho}_n(\theta) > 0$ by the bound (30).

4 Experiments

Our technique for distributionally robust optimization with adversarial training extends beyond supervised learning. To that end, we present empirical evaluations on supervised and reinforcement learning tasks where we compare performance with empirical risk minimization (ERM) and, where appropriate, models trained with the fast-gradient method (3) (FGM) (Goodfellow et al., 2015), its iterated variant (IFGM) (Kurakin et al., 2016), and the projected-gradient method (PGM) (Madry et al., 2017). PGM augments stochastic gradient steps for the parameter θ with projected gradient ascent over $x \mapsto \ell(\theta; x, y)$, iterating (for data point x_i, y_i)

$$\Delta x_i^{t+1}(\theta) := \operatorname*{argmax}_{\|\eta\|_p \le \epsilon} \{\nabla_x \ell(\theta; x_i^t, y_i)^T \eta\} \;\text{ and }\; x_i^{t+1} := \Pi_{\mathcal{B}_{\epsilon,p}(x_i^t)} \left\{ x_i^t + \alpha_t \Delta x_i^t(\theta) \right\} \tag{18}$$

for $t = 1, \ldots, T_{\mathrm{adv}}$, where Π denotes projection onto $\mathcal{B}_{\epsilon,p}(x_i) := \{x : \|x - x_i\|_p \le \epsilon\}$.

The adversarial training literature (e.g. Goodfellow et al. (2015)) usually considers $\|\cdot\|_\infty$-norm attacks, which allow imperceptible perturbations to all input features. In most scenarios, however, it is reasonable to defend against weaker adversaries that instead perturb influential features more. We consider this setting and train against $\|\cdot\|_2$-norm attacks. Namely, we use the squared Euclidean cost for the feature vectors $c_x(x, x') := \|x - x'\|_2^2$ and define the overall cost as the covariate-shift adversary (8) for WRM (Algorithm 1), and we use $p = 2$ for FGM, IFGM, PGM training in all experiments; we still test against adversarial perturbations with respect to the norms $p = 2, \infty$. We use $T_{\mathrm{adv}} = 15$ iterations for all iterative methods (IFGM, PGM, and WRM) in training and attacks.

In Section 4.1, we visualize differences between our approach and ad-hoc methods to illustrate the benefits of certified robustness. In Section 4.2 we consider a supervised learning problem for MNIST where we adversarially perturb the test data. Finally, we consider a reinforcement learning problem in Section 4.3, where the Markov decision process used for training differs from that for testing.

WRM enjoys the theoretical guarantees of Sections 2 and 3 for large γ, but for small γ (large adversarial budgets), WRM becomes a heuristic like other methods. In Appendix A.4, we compare WRM with other methods on attacks with large adversarial budgets. In Appendix A.5, we further compare WRM—which is trained to defend against $\|\cdot\|_2$-adversaries—with other heuristics trained to defend against $\|\cdot\|_\infty$-adversaries. WRM matches or outperforms other heuristics against imperceptible attacks, while it underperforms for attacks with large adversarial budgets.

4.1 Visualizing the benefits of certified robustness

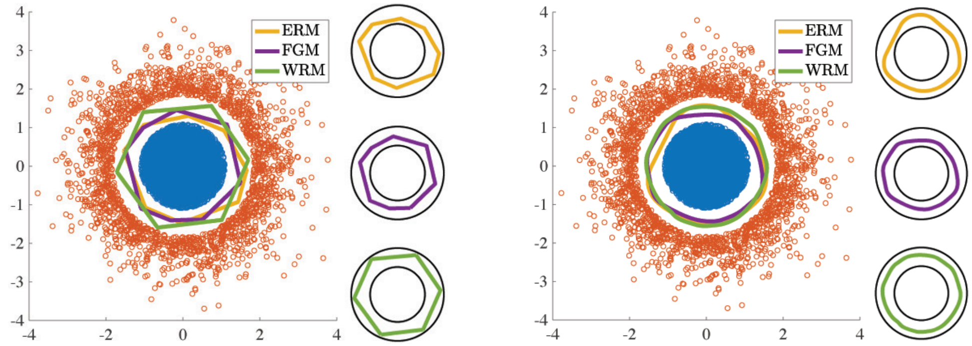
For our first experiment, we generate synthetic data $Z = (X, Y) \sim P_0$ by $X_i \overset{\mathrm{iid}}{\sim} \mathsf{N}(0_2, I_2)$ with labels $Y_i = \mathrm{sign}(\|x\|_2 - \sqrt{2})$, where $X \in \mathbb{R}^2$ and I_2 is the identity matrix in $\mathbb{R}^2$. Furthermore, to create a wide margin separating the classes, we remove data with $\|X\|_2 \in (\sqrt{2}/1.3, 1.3\sqrt{2})$. We train a small neural network with 2 hidden layers of size 4 and 2 and either all ReLU or all ELU activations between layers, comparing our approach (WRM) with ERM and the 2-norm FGM. For our approach we use $\gamma = 2$, and to make fair comparisons with FGM we use

$$\epsilon^2 = \widehat{\rho}_n(\theta_{\mathrm{WRM}}) = W_c(P_n^*(\theta_{\mathrm{WRM}}), \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T(\theta_{\mathrm{WRM}}, Z), Z)], \tag{19}$$

for the fast-gradient perturbation magnitude ϵ, where θ_{WRM} is the output of Algorithm 1.[1]

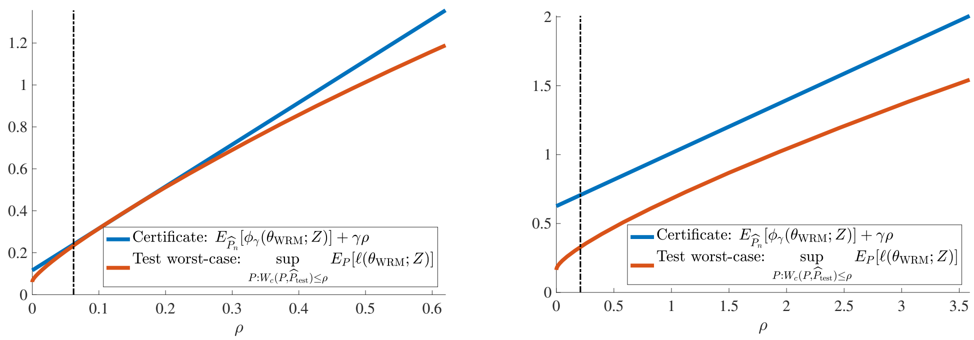
Figure 1 illustrates the classification boundaries for the three training procedures over the ReLU-activated (Figure 1(a)) and ELU-activated (Figure 1(b)) models. Since 70% of the data are of the

[1]For ELU activations with scale parameter 1, $\gamma = 2$ makes problem (2b) strongly concave over the training data. ReLU's have no guarantees, but we use 15 gradient steps with stepsize $1/\sqrt{t}$ for both activations.



(a) ReLU model (b) ELU model

Figure 1. Experimental results on synthetic data. Training data are shown in blue and red. Classification boundaries are shown in yellow, purple, and green for ERM, FGM, and WRM respectively. The boundaries are shown with the training data as well as separately with the true class boundaries.



(a) Synthetic data (b) MNIST

Figure 2. Empirical comparison between certificate of robustness (11) (blue) and test worst-case performance (red) for experiments with (a) synthetic data and (b) MNIST. We omit the certificate's error term $\epsilon_n(t)$. The vertical bar indicates the achieved level of robustness on the training set $\widehat{\rho}_n(\theta_{\rm WRM})$.

blue class ($\|X\|_2 \le \sqrt{2}/1.3$), distributional robustness favors pushing the classification boundary outwards; intuitively, adversarial examples are most likely to come from pushing blue points outwards across the boundary. ERM and FGM suffer from sensitivities to various regions of the data, as evidenced by the lack of symmetry in their classification boundaries. For both activations, WRM pushes the classification boundaries further outwards than ERM or FGM. However, WRM with ReLU's still suffers from sensitivities (e.g. radial asymmetry in the classification surface) due to the lack of robustness guarantees. WRM with ELU's provides a certified level of robustness, yielding an axisymmetric classification boundary that hedges against adversarial perturbations in all directions.

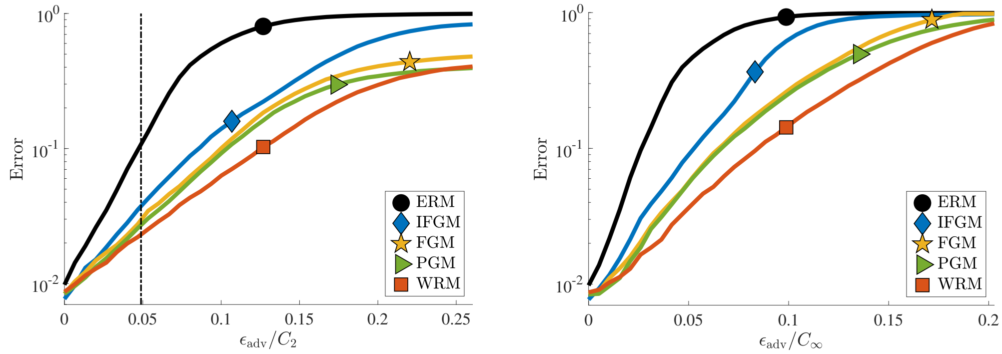
Recall that our *certificates of robustness* on the worst-case performance given in Theorem 3 applies for any level of robustness ρ. In Figure 2(a), we plot our certificate (11) against the out-of-sample (test) worst-case performance $\sup_{P: W_c(P,P_0)\le\rho} \mathbb{E}_P[\ell(\theta; Z)]$ for WRM with ELU's. Since the worst-case loss is hard to evaluate directly, we solve its Lagrangian relaxation (6) for different values of $\gamma_{\rm adv}$. For each $\gamma_{\rm adv}$, we consider the distance to adversarial examples in the test dataset

$$\widehat{\rho}_{\rm test}(\theta) := \mathbb{E}_{\widehat{P}_{\rm test}}[c(T_{\gamma_{\rm adv}}(\theta, Z), Z)], \tag{20}$$

where $\widehat{P}_{\rm test}$ is the test distribution, $c(z, z') := \|x - x'\|_2^2 + \infty \cdot \mathbf{1}\{y \ne y'\}$ as before, and $T_{\gamma_{\rm adv}}(\theta, Z) = \mathop{\rm argmax}_z\{\ell(\theta; z) - \gamma_{\rm adv} c(z, Z)\}$ is the adversarial perturbation of Z (Monge map) for the model θ. The worst-case losses on the test dataset are then given by

$$\mathbb{E}_{\widehat{P}_{\rm test}}[\phi_{\gamma_{\rm adv}}(\theta_{\rm WRM}; Z)] + \gamma_{\rm adv}\widehat{\rho}_{\rm test}(\theta_{\rm WRM}) = \sup_{P: W_c(P, P_{\rm test}) \le \widehat{\rho}_{\rm test}(\theta_{\rm WRM})} \mathbb{E}_P[\ell(\theta_{\rm WRM}; Z)].$$

As anticipated, our certificate is almost tight near the achieved level of robustness $\widehat{\rho}_n(\theta_{\rm WRM})$ for WRM (10) and provides a performance guarantee even for other values of ρ.



(a) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$ attack

(b) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$ attack

Figure 3. PGM attacks on the MNIST dataset. (a) and (b) show test misclassification error vs. the adversarial perturbation level $\epsilon_{\rm adv}$ for the PGM attack with respect to Euclidean and ∞ norms respectively. The vertical bar in (a) indicates the perturbation level used for training the PGM, FGM, and IFGM models as well as the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\rm WRM})}$. For MNIST, $C_2 = 9.21$ and $C_\infty = 1.00$.

4.2 Learning a more robust classifier

We now consider a standard benchmark—training a neural network classifier on the MNIST dataset. The network consists of 8×8, 6×6, and 5×5 convolutional filter layers with ELU activations followed by a fully connected layer and softmax output. We train WRM with $\gamma = 0.04\mathbb{E}_{\widehat{P}_n}[\|X\|_2]$, and for the other methods we choose ϵ as the level of robustness achieved by WRM (19).[2] In the figures, we scale the budgets $1/\gamma_{\rm adv}$ and $\epsilon_{\rm adv}$ for the adversary with $C_p := \mathbb{E}_{\widehat{P}_n}[\|X\|_p]$.[3]

First, in Figure 2(b) we again illustrate the validity of our certificate of robustness (11) for the worst-case test performance for arbitrary level of robustness ρ. We see that our certificate provides a performance guarantee for out-of-sample worst-case performance.

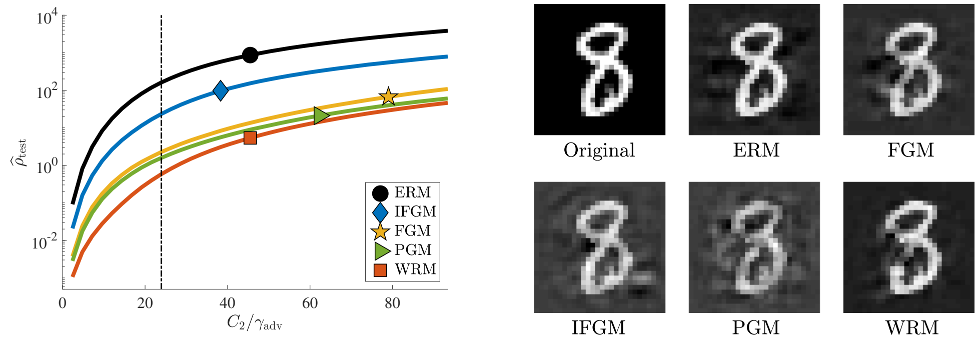
We now compare adversarial training techniques. All methods achieve at least 99% test-set accuracy, implying there is little test-time penalty for the robustness levels (ϵ and γ) used for training. It is thus important to distinguish the methods' abilities to combat attacks. We test performance of the five methods (ERM, FGM, IFGM, PGM, WRM) under PGM attacks (18) with respect to 2- and ∞-norms. In Figure 3(a) and (b), all adversarial methods outperform ERM, and WRM offers more robustness even with respect to these PGM attacks. Training with the Euclidean cost still provides robustness to ∞-norm fast gradient attacks. We provide further evidence in Appendix A.1.

Next we study stability of the loss surface with respect to perturbations to inputs. We note that small values of $\widehat{\rho}_{\rm test}(\theta)$, the distance to adversarial examples (20), correspond to small magnitudes of $\nabla_z \ell(\theta; z)$ in a neighborhood of the nominal input, which ensures stability of the model. Figure 4(a) shows that $\widehat{\rho}_{\rm test}$ differs by orders of magnitude between the training methods (models $\theta = \theta_{\rm ERM}, \theta_{\rm FGM}, \theta_{\rm IFGM}, \theta_{\rm PGM}, \theta_{\rm WRM}$); the trend is nearly uniform over all $\gamma_{\rm adv}$, with $\theta_{\rm WRM}$ being the most stable. Thus, we see that our adversarial-training method defends against gradient-exploiting attacks by reducing the magnitudes of gradients near the nominal input.

In Figure 4(b) we provide a qualitative picture by adversarially perturbing a single test datapoint until the model misclassifies it. Specifically, we again consider WRM attacks and we decrease $\gamma_{\rm adv}$ until each model misclassifies the input. The original label is 8, whereas on the adversarial examples IFGM predicts 2, PGM predicts 0, and the other models predict 3. WRM's "misclassifications" appear consistently reasonable to the human eye (see Appendix A.2 for examples of other digits); WRM defends against gradient-based exploits by learning a representation that makes gradients point towards inputs of other classes. Together, Figures 4(a) and (b) depict our method's defense mechanisms to gradient-based attacks: creating a more stable loss surface by reducing the magnitude of gradients and improving their interpretability.

[2]For this γ, $\phi_\gamma(\theta_{\rm WRM}; z)$ is strongly concave for 98% of the training data.

[3]For the standard MNIST dataset, $C_2 := \mathbb{E}_{\widehat{P}_n}\|X\|_2 = 9.21$ and $C_\infty := \mathbb{E}_{\widehat{P}_n}\|X\|_\infty = 1.00$.



(a) $\widehat{\rho}_{test}$ vs. $1/\gamma_{\text{adv}}$

(b) Perturbations on a test datapoint

Figure 4. Stability of the loss surface. In (a), we show the average distance of the perturbed distribution $\widehat{\rho}_{\text{test}}$ for a given γ_{adv}, an indicator of local stability to inputs for the decision surface. The vertical bar in (a) indicates the γ we use for training WRM. In (b) we visualize the smallest WRM perturbation (largest γ_{adv}) necessary to make a model misclassify a datapoint. More examples are in Appendix A.2.

4.3 Robust Markov decision processes

For our final experiments, we consider distributional robustness in the context of Q-learning, a model-free reinforcement learning technique. We consider Markov decision processes (MDP's) $(\mathcal{S}, \mathcal{A}, P_{sa}, r)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, state-action transition probabilities P_{sa}, and rewards $r : \mathcal{S} \to \mathbb{R}$. The goal of a reinforcement-learning agent is to maximize (discounted) cumulative rewards $\sum_t \lambda^t \mathbb{E}[r(s^t)]$ (with discount factor λ); this is analogous to minimizing $\mathbb{E}_P[\ell(\theta; Z)]$ in supervised learning. Robust MDP's consider an ambiguity set $\mathcal{P}_{sa}$ for state-action transitions. The goal is maximizing the worst-case realization $\inf_{P \in \mathcal{P}_{sa}} \sum_t \lambda^t \mathbb{E}_P[r(s^t)]$, analogous to problem (1).

In a standard MDP, Q-learning learns a quality function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ via the iterations

$$Q(s^t, a^t) \leftarrow Q(s^t, a^t) + \alpha_t \left(r(s^t) + \lambda \max_a Q(s^{t+1}, a) - Q(s^t, a^t) \right) \tag{21}$$

such that $\operatorname{argmax}_a Q(s, a)$ is (eventually) the optimal action in state s to maximize cumulative reward. In scenarios where the underlying environment has a continuous state-space and we represent Q with a differentiable function (e.g. Mnih et al. (2015)), we can modify the update (21) with an adversarial state perturbation to incorporate distributional robustness. Namely, we draw the nominal state-transition update $\widehat{s}^{t+1} \sim p_{sa}(s^t, a^t)$, and proceed with the update (21) using the perturbation

$$s^{t+1} \leftarrow \operatorname*{argmin}_s \left\{ r(s) + \lambda \max_a Q(s, a) + \gamma c(s, \hat{s}^{t+1}) \right\}. \tag{22}$$

For large γ, we can again solve problem (22) efficiently using gradient descent. This procedure provides robustness to uncertainties in state-action transitions. For tabular Q-learning, where we represent Q only over a discretized covering of the underlying state-space, we can either neglect the second term in the update (22) and, after performing the update, round s^{t+1} as usual, or we can perform minimization directly over the discretized covering. In the former case, since the update (22) simply modifies the state-action transitions (independent of Q), standard results on convergence for tabular Q-learning (e.g. Szepesvári & Littman (1999)) apply under these adversarial dynamics.

We test our adversarial training procedure in the cart-pole environment, where the goal is to balance a pole on a cart by moving the cart left or right. The environment caps episode lengths to 400 steps and ends the episode prematurely if the pole falls too far from the vertical or the cart translates too far from its origin. We use reward $r(\beta) := e^{-|\beta|}$ for the angle β of the pole from the vertical. We use a tabular representation for Q with 30 discretized states for β and 15 for its time-derivative $\dot{\beta}$ (we perform the update (22) without the Q-dependent term). The action space is binary: push the cart left or right with a fixed force. Due to the nonstationary, policy-dependent radius for the Wasserstein ball, an analogous ϵ for the fast-gradient method (or other variants) is not well-defined. Thus, we only compare with an agent trained on the nominal MDP. We test both models with perturbations to the physical parameters: we shrink/magnify the pole's mass by 2, the pole's length by 2, and the strength of gravity g by 5. The system's dynamics are such that the heavy, short, and strong-gravity cases are more unstable than the original environment, whereas their counterparts are less unstable.

Environment	Regular	Robust
Original	399.7 ± 0.1	400.0 ± 0.0
Easier environments		
Light	400.0 ± 0.0	400.0 ± 0.0
Long	400.0 ± 0.0	400.0 ± 0.0
Soft g	400.0 ± 0.0	400.0 ± 0.0
Harder environments		
Heavy	150.1 ± 4.7	334.0 ± 3.7
Short	245.2 ± 4.8	400.0 ± 0.0
Strong g	189.8 ± 2.3	398.5 ± 0.3

Table 1. Episode length over 1000 trials (mean ± standard error)

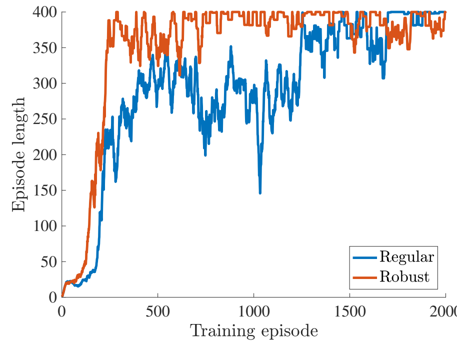


Figure 5. Episode lengths during training. The environment caps episodes to 400 steps.

Table 1 shows the performance of the trained models over the original MDP and all of the perturbed MDPs. Both models perform similarly over easier environments, but the robust model greatly outperforms in harder environments. Interestingly, as shown in Figure 5, the robust model also learns more efficiently than the nominal model in the original MDP. We hypothesize that a potential side-effect of robustness is that adversarial perturbations encourage better exploration of the environment.

5 Discussion and future work

Explicit distributional robustness of the form (5) is intractable except in limited cases. We provide a principled method for efficiently guaranteeing distributional robustness with a simple form of adversarial data perturbation. Using only assumptions about the smoothness of the loss function ℓ, we prove that our method enjoys strong statistical guarantees and fast optimization rates for a large class of problems. The NP-hardness of certifying robustness for ReLU networks, coupled with our empirical success and theoretical certificates for smooth networks in deep learning, suggest that using smooth networks may be preferable if we wish to guarantee robustness. Empirical evaluations indicate that our methods are in fact robust to perturbations in the data, and they match or outperform less-principled adversarial training techniques. The major benefit of our approach is its simplicity and wide applicability across many models and machine-learning scenarios.

There remain many avenues for future investigation. Our optimization result (Theorem 2) applies only for small values of robustness ρ and to a limited class of Wasserstein costs. Furthermore, our statistical guarantees (Theorems 3 and 4) use $\|\cdot\|_\infty$-covering numbers as a measure of model complexity, which can become prohibitively large for deep networks. In a learning-theoretic context, where the goal is to provide insight into convergence behavior as well as comfort that a procedure will "work" given enough data, such guarantees are satisfactory, but this may not be enough in security-essential contexts. This problem currently persists for most learning-theoretic guarantees in deep learning, and the recent works of Bartlett et al. (2017), Dziugaite & Roy (2017), and Neyshabur et al. (2017) attempt to mitigate this shortcoming. Replacing our current covering number arguments with more intricate notions such as margin-based bounds (Bartlett et al., 2017) would extend the scope and usefulness of our theoretical guarantees. Of course, the certificate (15) still holds regardless.

More broadly, this work focuses on small-perturbation attacks, and our theoretical guarantees show that it is possible to efficiently build models that provably guard against such attacks. Our method becomes another heuristic for protection against attacks with large adversarial budgets. Indeed, in the large-perturbation regime, efficiently training certifiably secure systems remains an important open question. We believe that conventional $\|\cdot\|_\infty$-defense heuristics developed for image classification do not offer much comfort in the large-perturbation/perceptible-attack setting: $\|\cdot\|_\infty$-attacks with a large budget can render images indiscernible to human eyes, while, for example, $\|\cdot\|_1$-attacks allow a concerted perturbation to critical regions of the image. Certainly $\|\cdot\|_\infty$-attack and defense models have been fruitful in building a foundation for security research in deep learning, but moving beyond them may be necessary for more advances in the large-perturbation regime.

ACKNOWLEDGMENTS

We thank Jacob Steinhardt for valuable feedback. AS, HN, and JD were partially supported by the SAIL-Toyota Center for AI Research. AS was also partially supported by a Stanford Graduate Fellowship and a Fannie & John Hertz Foundation Fellowship. HN was partially supported by a Samsung Fellowship. JD was partially supported by the National Science Foundation award NSF-CAREER-1553086.

REFERENCES

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6241–6250, 2017.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. *arXiv:1401.0212 [math.OC]*, 2013. URL `http://arxiv.org/abs/1401.0212`.

P. Billingsley. *Convergence of Probability Measures*. Wiley, Second edition, 1999.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *arXiv:1604.01446 [math.PR]*, 2016.

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to machine learning. *arXiv:1610.05627 [math.ST]*, 2016.

J. F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: a Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

T. Brown, D. Mane, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. In *Machine Learning and Computer Security Workshop, Neural Information Processing Systems*, 2017.

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

X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

Y. Chen, G. Lan, and Y. Ouyang. Accelerated schemes for a class of variational inequalities. *arXiv:1403.4164 [math.OC]*, 2014.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

J. C. Duchi, P. W. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv:1610.03425 [stat.ML]*, 2016. URL `https://arxiv.org/abs/1610.03425`.

G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv:1703.11008 [cs.LG]*, 2017.

P. M. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *arXiv:1505.05116 [math.OC]*, 2015.

S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

W. He, J. Wei, X. Chen, N. Carlini, and D. Song. Adversarial example defenses: Ensembles of weak defenses are not strong. *arXiv:1706.04701 [cs.LG]*, 2017.

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

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with the stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

G. Katz, C. Barrett, D. Dill, K. Julian, and M. Kochenderfer. Reluplex: An efficient SMT solver for verifying deep neural networks. *arXiv:1702.01135 [cs.AI]*, 2017a.

G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer. Towards proving the adversarial robustness of deep neural networks. *arXiv:1709.02802 [cs.LG]*, 2017b.

J. Z. Kolter and E. Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv:1711.00851 [cs.LG]*, 2017. URL `https://arxiv.org/abs/1711.00851`.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv:1611.01236 [cs.CV]*, 2016.

H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *Proceedings of the 2015 Winter Simulation Conference*. IEEE, 2015.

J. Lee and M. Raginsky. Minimax statistical learning and domain adaptation with Wasserstein distances. *arXiv:1705.07815 [cs.LG]*, 2017.

D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083 [stat.ML]*, 2017.

T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. *arXiv:1507.00677 [stat.ML]*, 2015.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, 2016.

H. Namkoong and J. C. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems 30*, 2017.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pp. 5949–5958, 2017.

A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 427–436, 2015.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv:1602.02697 [cs.CR]*, 2016a.

N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pp. 372–387. IEEE, 2016b.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pp. 582–597. IEEE, 2016c.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

A. Raghunathan, J. Steinhardt, and P. Liang. Certified defenses against adversarial examples. *arXiv:1801.09344 [cs.LG]*, 2018. URL `https://arxiv.org/abs/1801.09344`.

N. Ratliff, J. A. Bagnell, and M. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

A. Rozsa, M. Gunther, and T. E. Boult. Towards robust deep neural networks with bang. *arXiv:1612.00138 [cs.CV]*, 2016.

S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv:1312.6199 [cs.CV]*, 2013.

C. Szepesvári and M. L. Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 11(8):2017–2060, 1999.

F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv:1705.07204 [stat.ML]*, 2017.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

C. Villani. *Optimal Transport: Old and New*. Springer, 2009.

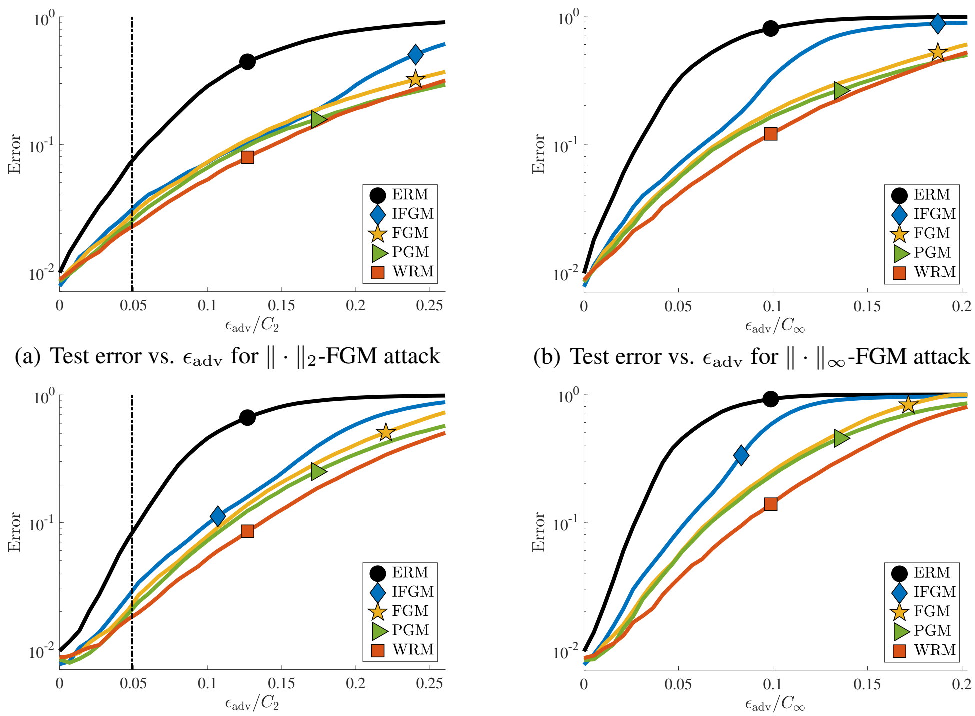
H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *The Journal of Machine Learning Research*, 10:1485–1510, 2009.

H. Xu, C. Caramanis, and S. Mannor. A distributional interpretation of robust optimization. *Mathematics of Operations Research*, 37(1):95–110, 2012.

A Additional Experiments

A.1 MNIST attacks

We repeat Figure 3 using FGM (tow row of Figure 6) and IFGM (bottom row of Figure 6) attacks. The same trends are evident as in Figure 3.



(a) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-FGM attack

(b) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-FGM attack

(c) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-IFGM attack

(d) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-IFGM attack

Figure 6. Further attacks on the MNIST dataset. We illustrate test misclassification error vs. the adversarial perturbation level $\epsilon_{\rm adv}$. Top row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (a) and (c) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\rm WRM})}$.

A.2 MNIST stability of loss surface

In Figure 7, we repeat the illustration in Figure 4(b) for more digits. WRM's "misclassifications" are consistently reasonable to the human eye, as gradient-based perturbations actually transform the original image to other labels. Other models do not exhibit this behavior with the same consistency (if at all). Reasonable misclassifications correspond to having learned a data representation that makes gradients interpretable.

A.3 MNIST Experiments with varied γ

In Figure 8, we choose a fixed WRM adversary (fixed $\gamma_{\rm adv}$) and perturb WRM models trained with various penalty parameters γ. As the bound (11) with $\eta = \gamma$ suggests, even when the adversary has more budget than that used for training ($1/\gamma < 1/\gamma_{\rm adv}$), degradation in performance is still *smooth*. Further, as we decrease the penalty γ, the amount of achieved robustness—measured here by test error on adversarial perturbations with $\gamma_{\rm adv}$—has diminishing gains; this is again consistent with our theory which says that the inner problem (2b) is not efficiently computable for small γ.

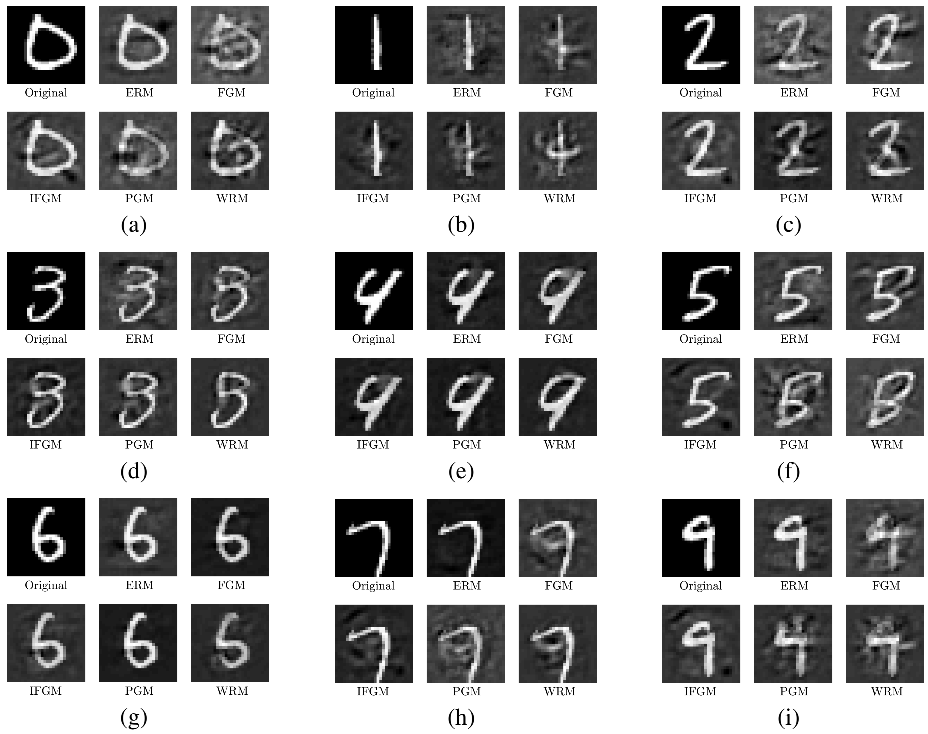
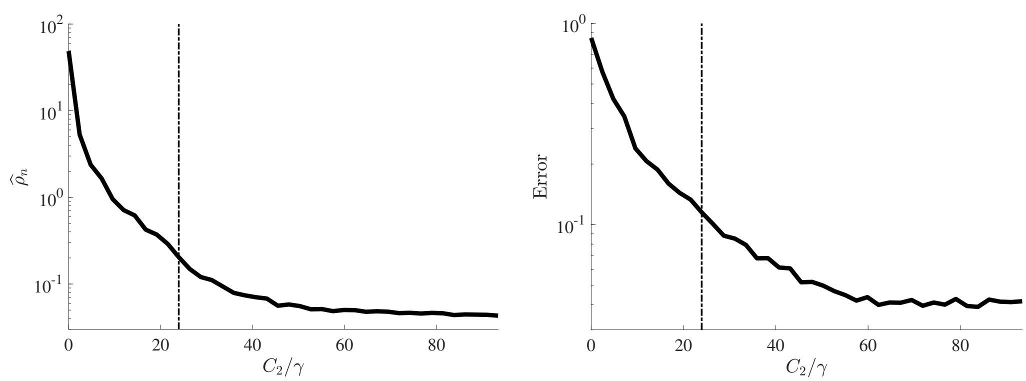


Figure 7. Visualizing stability over inputs. We illustrate the smallest WRM perturbation (largest $\gamma_{\rm adv}$) necessary to make a model misclassify a datapoint.



(a) $\widehat{\rho}_n$ vs. $1/\gamma$

(b) Test error vs. $1/\gamma$

Figure 8. (a) Stability and (b) test error for a fixed adversary. We train WRM models with various levels of γ and perturb them with a fixed WRM adversary ($\gamma_{\rm adv}$ indicated by the vertical bar).

A.4 MNIST EXPERIMENTS WITH A LARGER ADVERSARIAL BUDGET

Figures 9 and 10 repeat Figures 2(b), 3, and 6 for a larger training adversarial budget ($\gamma = 0.02C_2$) as well as larger test adversarial budgets. The distinctions in performance between various methods are less apparent now. For our method, the inner supremum is no longer strongly concave for over 10% of the data, indicating that we no longer have guarantees of performance. For large adversaries (i.e. large desired robustness values) our approach becomes a heuristic just like the other approaches.

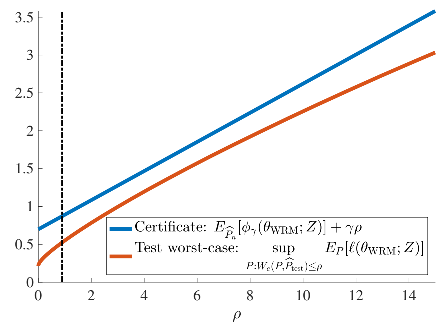
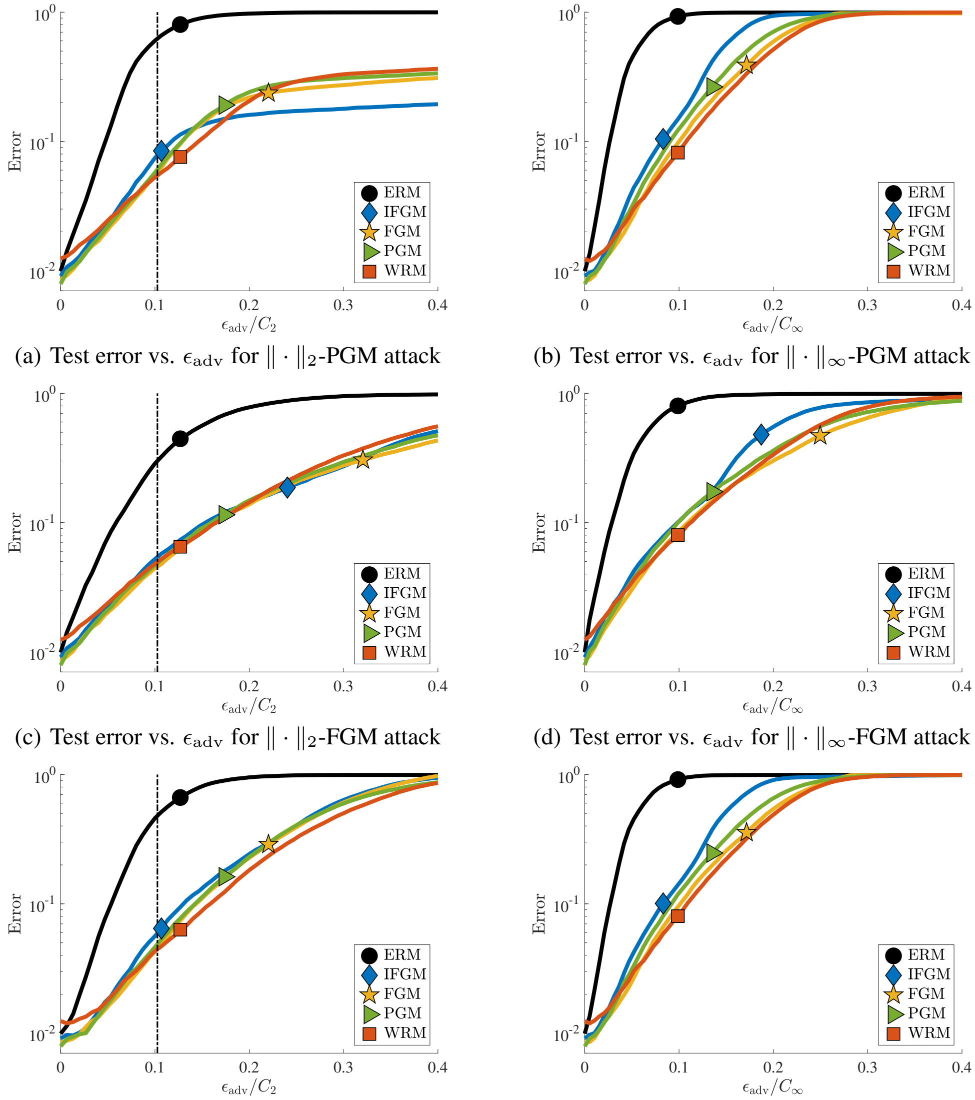


Figure 9. Empirical comparison between certificate of robustness (11) (blue) and out-of-sample (test) worst-case performance (red) for the experiments on MNIST with a larger training adversary. The statistical error term $\epsilon_n(t)$ is omitted from the certificate. The vertical bar indicates the achieved level of robustness on the training set $\widehat{\rho}_n(\theta_{\rm WRM})$.



(a) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-PGM attack

(b) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-PGM attack

(c) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-FGM attack

(d) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-FGM attack

(e) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-IFGM attack

(f) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-IFGM attack

Figure 10. Attacks on the MNIST dataset with larger (training and test) adversarial budgets. We illustrate test misclassification error vs. the adversarial perturbation level ϵ_{adv}. Top row: PGM attacks, middle row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (a), (c), and (e) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\mathrm{WRM}})}$.

A.5 MNIST ∞-NORM EXPERIMENTS

We consider training FGM, IFGM, and PGM with $p = \infty$. We first compare with WRM trained in the same manner as before—with the squared Euclidean cost. Then, we consider a heuristic Lagrangian approach for training WRM with the squared ∞-norm cost.

A.5.1 COMPARISON WITH STANDARD WRM

Our method (WRM) is trained to defend against $\|\cdot\|_2$-norm attacks by using the cost function

$$c((x,y),(x_0,y_0)) = \|x - x_0\|_2^2 + \infty \cdot \mathbf{1}\{y \neq y_0\}$$

with the convention that $0 \cdot \infty = 0$. Standard adversarial training methods often train to defend against $\|\cdot\|_\infty$-norm attacks, which we compare our method against in this subsection. Direct comparison between these approaches is not immediate, as we need to determine a suitable ϵ to train

FGM, IFGM, and PGM in the ∞-norm that corresponds to the penalty parameter γ for the $\|\cdot\|_2$-norm that we use. Similar to the expression (19), we use

$$\epsilon := \mathbb{E}_{\widehat{P}_n}[\|T(\theta_{\mathrm{WRM}}, Z) - Z\|_\infty] \tag{23}$$

as the adversarial training budget for FGM, IFGM and PGM with $\|\cdot\|_\infty$-norms. Because 2-norm adversaries tend to focus budgets on a subset of features, the resulting ∞-norm perturbations are relatively large. In Figure 11 we show the results trained with a small training adversarial budget. In this regime, (large γ, small ϵ), WRM matches the performance of other techniques.

In Figure 12 we show the results trained with a large training adversarial budget. In this regime (small γ, large ϵ), performance between WRM and other methods diverge. WRM, which provably defends against small perturbations, outperforms other heuristics against imperceptible attacks for both Euclidean and ∞ norms. Further, it outperforms other heuristics on natural images, showing that it consistently achieves a smaller price of robustness. On attacks with large adversarial budgets (large ϵ_{adv}), however, the performance of WRM is worse than that of the other methods (especially in the case of ∞-norm attacks). These findings verify that WRM is a practical alternative over existing heuristics for the moderate levels of robustness where our guarantees hold.

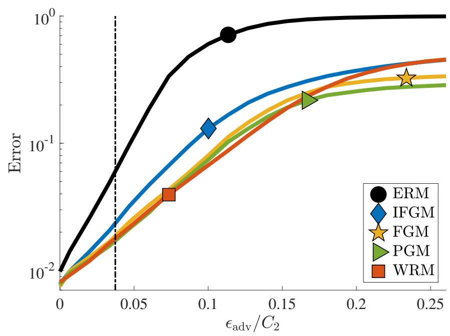
A.5.2 Comparison with $\|\cdot\|_\infty$-WRM

Our computational guarantees given in Theorem 2 does not hold anymore when we consider ∞-norm adversaries:

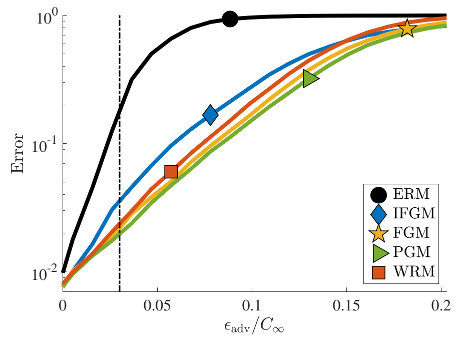
$$c((x, y), (x_0, y_0)) = \|x - x_0\|_\infty^2 + \infty \cdot \mathbf{1}\{y \neq y_0\}. \tag{24}$$

Optimizing the Lagrangian formulation (2b) with the ∞-norm is difficult since subtracting a multiple of the ∞-norm does not add (negative) curvature in all directions. In Appendix E, we propose a heuristic algorithm for solving the inner supremum problem (2b) with the above cost function (24). Our approach is based on a variant of proximal algorithms.

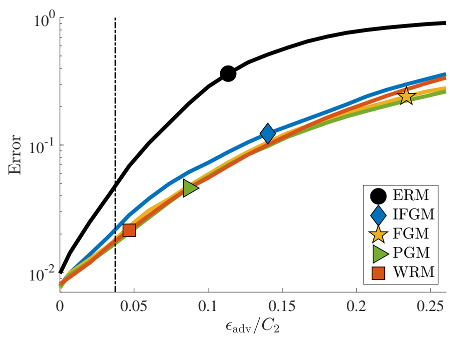
We compare our proximal heuristic introduced in Appendix E with other adversarial training procedures that were trained against ∞-norm adversaries. Results are shown in Figure 13 for a small training adversary and Figure 14 for a large training adversary. We observe that similar trends as in Section A.5.1 hold again.



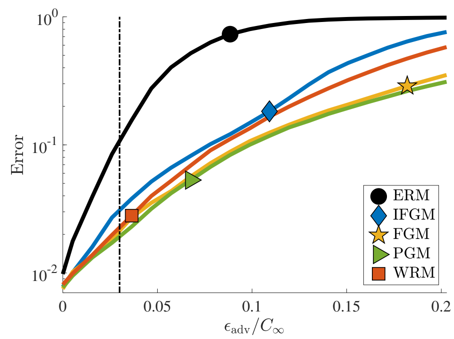
(a) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-PGM attack



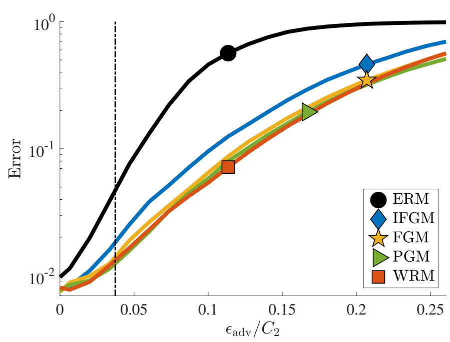
(b) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-PGM attack



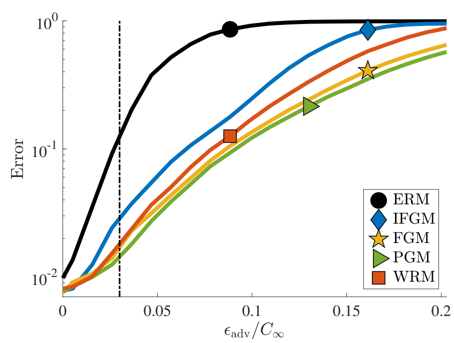
(c) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-FGM attack



(d) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-FGM attack

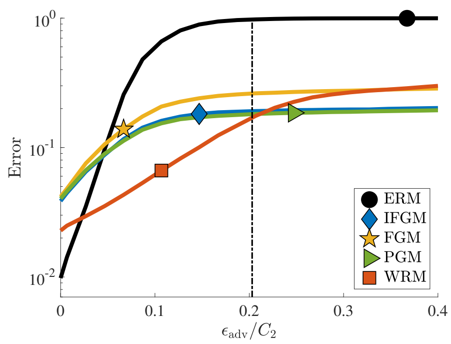


(e) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-IFGM attack

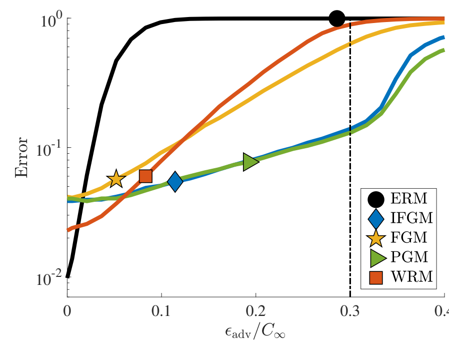


(f) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-IFGM attack

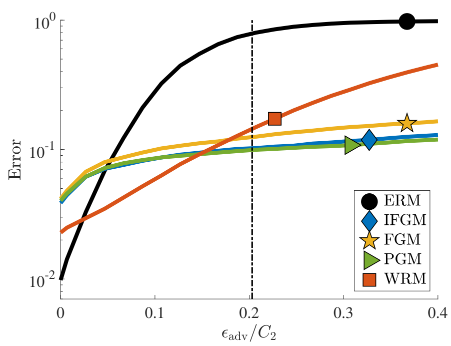
Figure 11. Attacks on the MNIST dataset. We compare standard WRM with ∞-norm PGM, FGM, IFGM. We illustrate test misclassification error vs. the adversarial perturbation level $\epsilon_{\rm adv}$. Top row: PGM attacks, middle row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (a), (c), and (e) indicates the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\rm WRM})}$. The vertical bar in (b), (d), and (f) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models via (23).



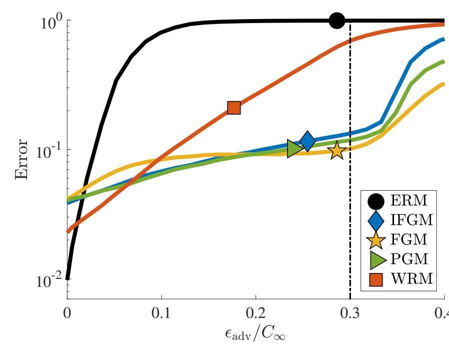
(a) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-PGM attack



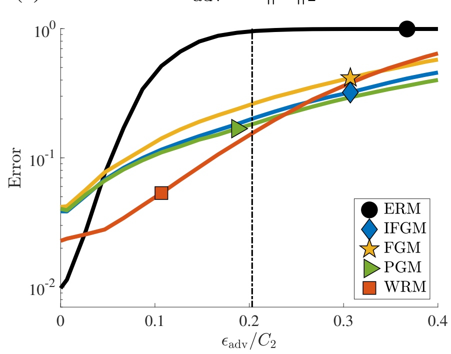
(b) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-PGM attack



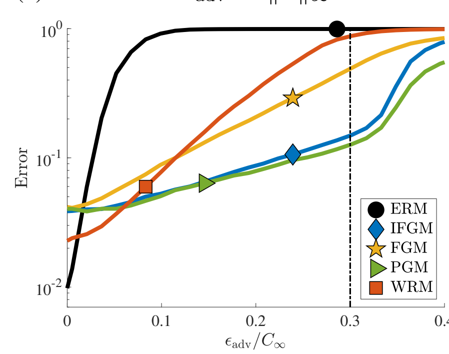
(c) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-FGM attack



(d) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-FGM attack

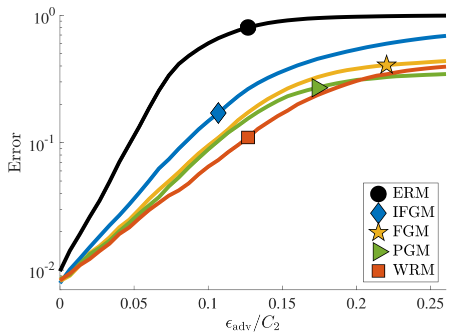


(e) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-IFGM attack

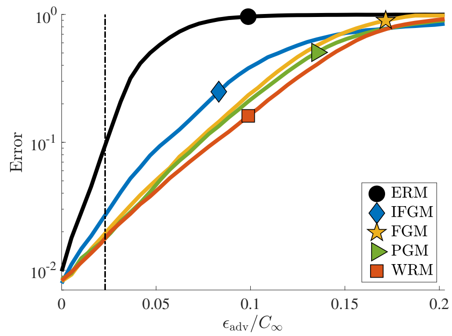


(f) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-IFGM attack

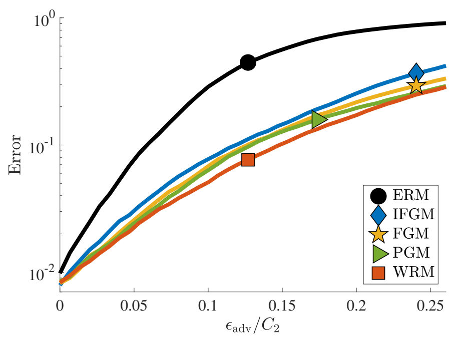
Figure 12. Attacks on the MNIST dataset with larger (training and test) adversarial budgets. We compare standard WRM with ∞-norm PGM, FGM, IFGM models. We illustrate test misclassification error vs. the adversarial perturbation level ϵ_{adv}. Top row: PGM attacks, middle row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (a), (c), and (e) indicates the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$. The vertical bar in (b), (d), and (f) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models via (23).



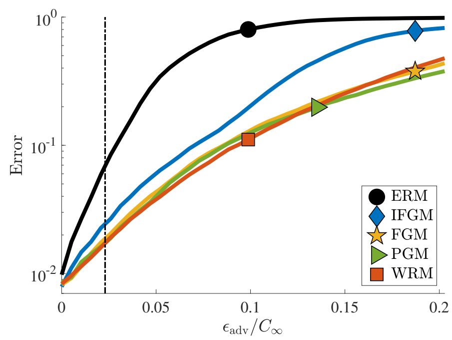
(a) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-PGM attack



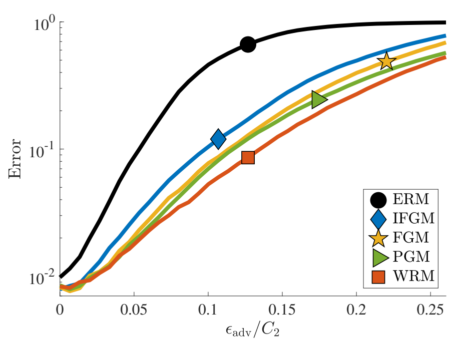
(b) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-PGM attack



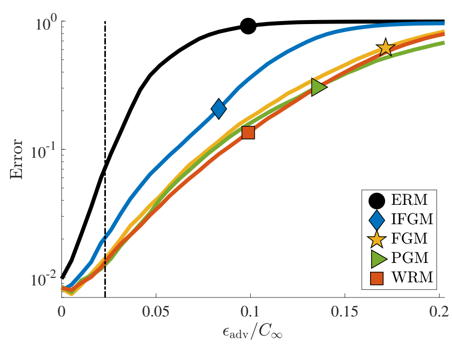
(c) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-FGM attack



(d) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-FGM attack

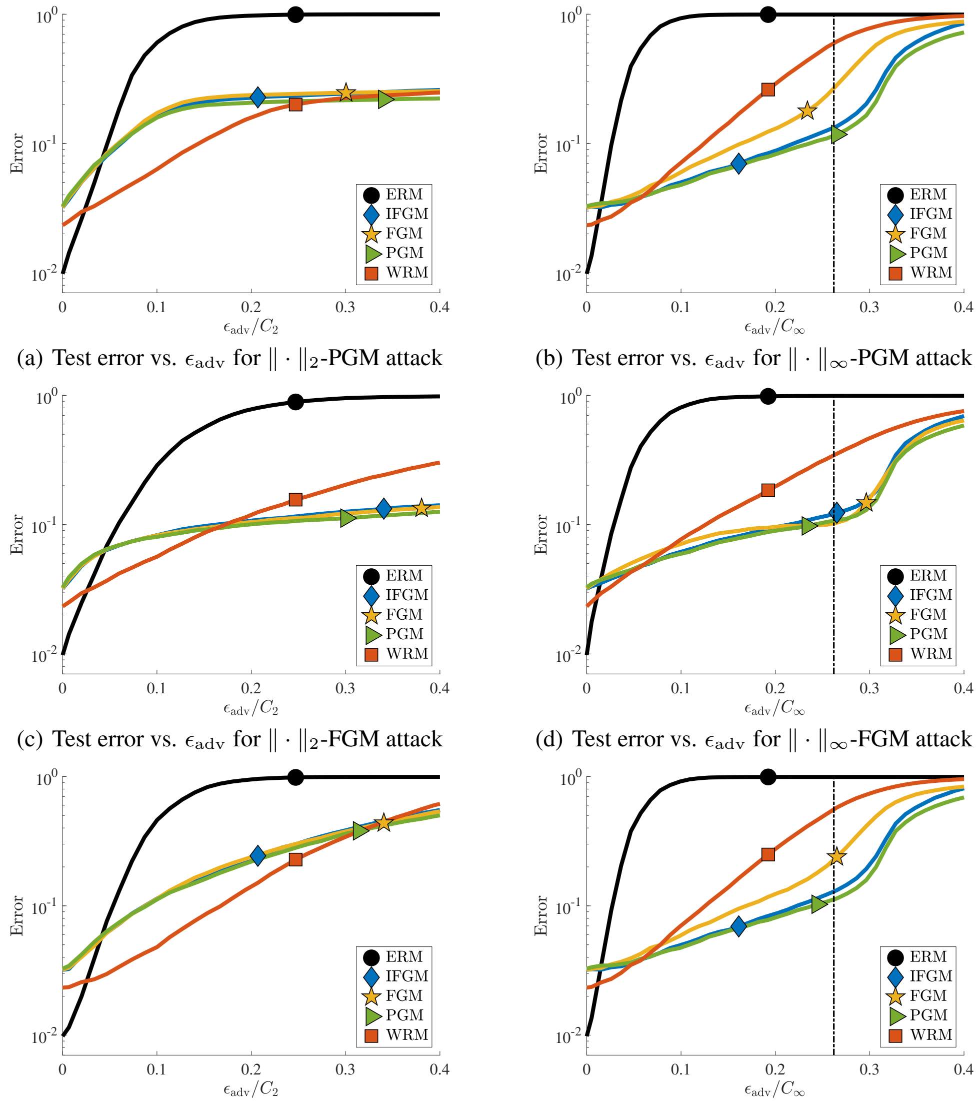


(e) Test error vs. ϵ_{adv} for $\|\cdot\|_2$-IFGM attack



(f) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$-IFGM attack

Figure 13. Attacks on the MNIST dataset. All models are trained in the ∞-norm. We illustrate test misclassification error vs. the adversarial perturbation level ϵ_{adv}. Top row: PGM attacks, middle row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (b), (d), and (f) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$.



(a) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-PGM attack (b) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-PGM attack

(c) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-FGM attack (d) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-FGM attack

(e) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_2$-IFGM attack (f) Test error vs. $\epsilon_{\rm adv}$ for $\|\cdot\|_\infty$-IFGM attack

Figure 14. Attacks on the MNIST dataset with larger (training and test) adversarial budgets. All models are trained in the ∞-norm. We illustrate test misclassification error vs. the adversarial perturbation level $\epsilon_{\rm adv}$. Top row: PGM attacks, middle row: FGM attacks, bottom row: IFGM attacks. Left column: Euclidean-norm attacks, right column: ∞-norm attacks. The vertical bar in (b), (d), and (f) indicates the perturbation level that was used for training the PGM, FGM, and IFGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\rm WRM})}$.

B FINDING WORST-CASE PERTURBATIONS WITH RELU'S IS NP-HARD

We show that computing worst-case perturbations $\sup_{u\in\mathcal{U}} \ell(\theta; z+u)$ is NP-hard for a large class of feedforward neural networks with ReLU activations. This result is essentially due to Katz et al. (2017a). In the following, we use polynomial time to mean polynomial growth with respect to m, the dimension of the inputs z.

An optimization problem is *NPO* (NP-Optimization) if (i) the dimensionality of the solution grows polynomially, (ii) the language $\{u \in \mathcal{U}\}$ can be recognized in polynomial time (i.e. a deterministic algorithm can decide in polynomial time whether $u \in \mathcal{U}$), and (iii) ℓ can be evaluated in polynomial time. We restrict analysis to feedforward neural networks with ReLU activations such that the corresponding worst-case perturbation problem is NPO.[4] Furthermore, we impose separable structure on $\mathcal{U}$, that is, $\mathcal{U} := \{v \le u \le w\}$ for some $v < w \in \mathbb{R}^m$.

Lemma 2. *Consider feedforward neural networks with ReLU's and let $\mathcal{U} := \{v \le u \le w\}$, where $v < w$ such that the optimization problem $\max_{u \in \mathcal{U}} \ell(\theta; z + u)$ is NPO. Then there exists θ such that this optimization problem is also NP-hard.*

Proof First, we introduce the decision reformulation of the problem: for some b, we ask whether there exists some u such that $\ell(\theta; z + u) \ge b$. The decision reformulation for an NPO problem is in NP, as a certificate for the decision problem can be verified in polynomial time. By appropriate scaling of θ, v, and w, Katz et al. (2017a) show that 3-SAT Turing-reduces to this decision problem: given an oracle D for the decision problem, we can solve an arbitrary instance of 3-SAT with a polynomial number of calls to D. The decision problem is thus NP-complete.

Now, consider an oracle O for the optimization problem. The decision problem Turing-reduces to the optimization problem, as the decision problem can be solved with one call to O. Thus, the optimization problem is NP-hard. □

C Proofs

C.1 Proof of Proposition 1

For completeness, we provide an alternative proof to that given in Blanchet & Murthy (2016) using convex analysis. Our proof is less general, requiring the cost function c to be continuous and convex in its first argument. The below general duality result gives Proposition 1 as an immediate special case. Recalling Rockafellar & Wets (1998, Def. 14.27 and Prop. 14.33), we say that a function $g : X \times Z \to \mathbb{R}$ is a *normal integrand* if for each α, the mapping

$$z \mapsto \{x \mid g(x, z) \le \alpha\}$$

is closed-valued and measurable. We recall that if g is continuous, then g is a normal integrand (Rockafellar & Wets, 1998, Cor. 14.34); therefore, $g(x, z) = \gamma c(x, z) - \ell(\theta; x)$ is a normal integrand. We have the following theorem.

Theorem 5. *Let f, c be such that for any $\gamma \ge 0$, the function $g(x, z) = \gamma c(x, z) - f(x)$ is a normal integrand. (For example, continuity of f and closed convexity of c is sufficient.) For any $\rho > 0$ we have*

$$\sup_{P: W_c(P,Q)} \int f(x) dP(x) = \inf_{\gamma \ge 0} \left\{ \int \sup_{x \in X} \{f(x) - \gamma c(x, z)\} \, dQ(z) + \gamma\rho \right\}.$$

Proof First, the mapping $P \mapsto W_c(P, Q)$ is convex in the space of probability measures. As taking $P = Q$ yields $W_c(Q, Q) = 0$, Slater's condition holds and we may apply standard (infinite dimensional) duality results (Luenberger, 1969, Thm. 8.7.1) to obtain

$$\begin{aligned}
\sup_{P: W_c(P,Q)} \int f(x) dP(x) &= \sup_{P: W_c(P,Q)} \inf_{\gamma \ge 0} \left\{ \int f(x) dP(x) - \gamma W_c(P, Q) + \gamma\rho \right\} \\
&= \inf_{\gamma \ge 0} \sup_{P: W_c(P,Q)} \left\{ \int f(x) dP(x) - \gamma W_c(P, Q) + \gamma\rho \right\}.
\end{aligned}$$

Now, noting that for any $M \in \Pi(P, Q)$ we have $\int f dP = \iint f(x) dM(x, z)$, we have that the rightmost quantity in the preceding display satisfies

$$\int f(x) dP(x) - \gamma \inf_{M \in \Pi(P,Q)} \int c(x, z) dM(x, z) = \sup_{M \in \Pi(P,Q)} \left\{ \int [f(x) - \gamma c(x, z)] dM(x, z) \right\}.$$

That is, we have

$$\sup_{P: W_c(P,Q)} \int f(x) dP(x) = \inf_{\gamma \ge 0} \sup_{P, M \in \Pi(P,Q)} \left\{ \int [f(x) - \gamma c(x, z)] dM(x, z) + \gamma\rho \right\}. \tag{25}$$

[4] Note that $z, u \in \mathbb{R}^m$, so trivially the dimensionality of the solution grows polynomially.

Now, we note a few basic facts. First, because we have a joint supremum over P and measures $M \in \Pi(P, Q)$ in expression (25), we have that

$$\sup_{P, M \in \Pi(P,Q)} \int [f(x) - \gamma c(x,z)] dM(x,z) \le \int \sup_x [f(x) - \gamma c(x,z)] dQ(z).$$

We would like to show equality in the above. To that end, we note that if $\mathcal{P}$ denotes the space of regular conditional probabilities (Markov kernels) from Z to X, then

$$\sup_{P, M \in \Pi(P,Q)} \int [f(x) - \gamma c(x,z)] dM(x,z) \ge \sup_{P \in \mathcal{P}} \int [f(x) - \gamma c(x,z)] dP(x \mid z) dQ(z).$$

Recall that a conditional distribution $P(\cdot \mid z)$ is regular if $P(\cdot \mid z)$ is a distribution for each z and for each measurable A, the function $z \mapsto P(A \mid z)$ is measurable. Let $\mathcal{X}$ denote the space of all measurable mappings $z \mapsto x(z)$ from Z to X. Using the powerful measurability results of Rockafellar & Wets (1998, Theorem 14.60), we have

$$\sup_{x \in \mathcal{X}} \int [f(x(z)) - \gamma c(x(z), z)] dQ(z) = \int \sup_{x \in X} [f(x) - \gamma c(x,z)] dQ(z)$$

because $f - c$ is upper semi-continuous, and the latter function is measurable. Now, let $x(z)$ be any measurable function that is ϵ-close to attaining the supremum above. Define the conditional distribution $P(\cdot \mid z)$ to be supported on $x(z)$, which is evidently measurable. Then using the preceding display, we have

$$\begin{aligned}
\int [f(x) - \gamma c(x,z)] dP(x \mid z) dQ(z) &= \int [f(x(z)) - \gamma c(x(z), z)] dQ(z) \\
&\ge \int \sup_{x \in X} [f(x) - \gamma c(x,z)] dQ(z) - \epsilon \\
&\ge \sup_{P, M \in \Pi(P,Q)} \int [f(x) - \gamma c(x,z)] dM(x,z) - \epsilon.
\end{aligned}$$

As $\epsilon > 0$ is arbitrary, this gives

$$\sup_{P, M \in \Pi(P,Q)} \int [f(x) - \gamma c(x,z)] dM(x,z) = \int \sup_{x \in X} [f(x) - \gamma c(x,z)] dQ(z)$$

as desired, which implies both equality (6) and completes the proof. □

C.2 Proof of Lemma 1

First, note that $z^\star(\theta)$ is unique and well-defined by the strong convexity of $f(\theta, \cdot)$. For Lipschitzness of $z^\star(\theta)$, we first argue that $z^\star(\theta)$ is continuous in θ. For any θ, optimality of $z^\star(\theta)$ implies that $\mathsf{g}_\mathsf{z}(\theta, z^\star(\theta))^T (z - z^\star(\theta)) \le 0$. By strong concavity, for any θ_1, θ_2 and $z_1^\star = z^\star(\theta_1)$ and $z_2^\star = z^\star(\theta_2)$, we have

$$\frac{\lambda}{2} \left\| z_1^\star - z_2^\star \right\|^2 \le f(\theta_2, z_2^\star) - f(\theta_2, z_1^\star) \text{ and } f(\theta_2, z_2^\star) \le f(\theta_2, z_1^\star) + \mathsf{g}_\mathsf{z}(\theta_2, z_1^\star)^T (z_2^\star - z_1^\star) - \frac{\lambda}{2} \left\| z_1^\star - z_2^\star \right\|^2.$$

Summing these inequalities gives

$$\lambda \left\| z_1^\star - z_2^\star \right\|^2 \le \mathsf{g}_\mathsf{z}(\theta_2, z_1^\star)^T (z_2^\star - z_1^\star) \le (\mathsf{g}_\mathsf{z}(\theta_2, z_1^\star) - \mathsf{g}_\mathsf{z}(\theta_1, z_1^\star))^T (z_2^\star - z_1^\star),$$

where the last inequality follows because $\mathsf{g}_\mathsf{z}(\theta_1, z_1^\star)^T (z_2^\star - z_1^\star) \le 0$. Using a cross-Lipschitz condition from above and Holder's inequality, we obtain

$$\lambda \left\| z_1^\star - z_2^\star \right\|^2 \le \left\| \mathsf{g}_\mathsf{z}(\theta_2, z_1^\star) - \mathsf{g}_\mathsf{z}(\theta_1, z_1^\star) \right\|_\star \left\| z_1^\star - z_2^\star \right\| \le L_{\mathsf{z}\theta} \left\| \theta_1 - \theta_2 \right\| \left\| z_1^\star - z_2^\star \right\|,$$

that is,

$$\left\| z_1^\star - z_2^\star \right\| \le \frac{L_{\mathsf{z}\theta}}{\lambda} \left\| \theta_1 - \theta_2 \right\|. \tag{26}$$

To see the second inequality, we show that $\bar{f}$ is differentiable with $\nabla \bar{f}(\theta) = \mathrm{g}_\theta(\theta, z^\star(\theta))$. By using a variant of the envelope (or Danskin's) theorem, we first show directional differentiability of $\bar{f}$. Recall that we say f is *inf-compact* if for all $\theta_0 \in \Theta$, there exists $\alpha > 0$ and a compact set $C \subset \Theta$ such that

$$\emptyset \neq \{z \in \mathcal{Z} : f(\theta, z) \leq \alpha\} \subset C$$

for all θ in some neighborhood of θ_0 (Bonnans & Shapiro, 2013). See Bonnans & Shapiro (2013, Theorem 4.13) for a proof of the following result.

Lemma 3. *Suppose that $f(\cdot, z)$ is differentiable in θ for all $z \in \mathcal{Z}$, and f, $\nabla_z f$ are continuous on $\Theta \times \mathcal{Z}$. If f is inf-compact, then $\bar{f}$ is directionally differentiable with*

$$\bar{f}'(\theta, d) = \sup_{z \in S(\theta)} \nabla_z f(\theta, z)^\top d$$

where $S(\theta) = \mathrm{argmin}_z\, f(\theta, z)$.

Now, note that from Assumption B, we have

$$|f(\theta, z) - f(\theta_0, z) - \nabla_\theta f(\theta_0, z)^\top (\theta - \theta_0)| \leq L_{\theta\theta} \|\theta - \theta_0\|$$

from which it is easy to see that f is inf-compact. Applying Lemma 3 to $\bar{f}$ and noting that $S(\theta)$ is unique by strong convexity of $f(\theta, \cdot)$, we have that $\bar{f}$ is directionally differentiable with $\nabla \bar{f}(\theta) = \mathrm{g}_\theta(\theta, z^\star(\theta))$. Since g_θ is continuous by Assumption B and $z^\star(\theta)$ is Lipschitz (26), we conclude that $\bar{f}$ is differentiable.

Finally, we have

$$\begin{aligned}
\|\mathrm{g}_\theta(\theta_1, z_1^\star) - \mathrm{g}_\theta(\theta_2, z_2^\star)\|_\star &\leq \|\mathrm{g}_\theta(\theta_1, z_1^\star) - \mathrm{g}_\theta(\theta_1, z_2^\star)\|_\star + \|\mathrm{g}_\theta(\theta_1, z_2^\star) - \mathrm{g}_\theta(\theta_2, z_2^\star)\|_\star \\
&\leq L_{\theta \mathrm{z}} \|z_1^\star - z_2^\star\| + L_{\theta\theta} \|\theta_1 - \theta_2\| \\
&\leq \left(L_{\theta\theta} + \frac{L_{\theta \mathrm{z}} L_{\mathrm{z}\theta}}{\lambda} \right) \|\theta_1 - \theta_2\|,
\end{aligned}$$

where we have used inequality (26) again. This is the desired result.

C.3 Proof of Theorem 2

Our proof is based on that of Ghadimi & Lan (2013). For shorthand, let $f(\theta, z; z_0) = \ell(\theta; z) - \gamma c(z, z_0)$, noting that we perform gradient steps with

$$g^t = \nabla_\theta f(\theta^t, \widehat{z}^t; z^t)$$

for $\widehat{z}^t$ an ϵ-approximate maximizer of $f(\theta, z; z^t)$ in z, and $\theta^{t+1} = \theta^t - \alpha_t g^t$. We assume $\alpha_t \leq \frac{1}{L_\phi}$ in the rest of the proof, which is satisfied for the constant stepsize $\alpha = \sqrt{\frac{\Delta_F}{L_\phi T \sigma^2}}$ and $T \geq \frac{L_\phi \Delta_F}{\sigma^2}$. By a Taylor expansion using the L_ϕ-smoothness of the objective F, we have

$$\begin{aligned}
F(\theta^{t+1}) &\leq F(\theta^t) + \left\langle \nabla F(\theta^t), \theta^{t+1} - \theta^t \right\rangle + \frac{L_\phi}{2} \left\|\theta^{t+1} - \theta^t\right\|_2^2 \\
&= F(\theta^t) - \alpha_t \left\|\nabla F(\theta^t)\right\|_2^2 + \frac{L_\phi \alpha_t^2}{2} \left\|g^t\right\|_2^2 + \alpha_t \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - g^t \right\rangle \\
&= F(\theta^t) - \alpha_t \left(1 - \frac{1}{2} L_\phi \alpha_t\right) \left\|\nabla F(\theta^t)\right\|_2^2 \\
&\qquad + \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - g^t \right\rangle + \frac{L_\phi \alpha_t^2}{2} \left\|g^t - \nabla F(\theta^t)\right\|_2^2.
\end{aligned} \tag{27}$$

Recalling the definition (2b) of $\phi_\gamma(\theta; z_0) = \sup_{z \in \mathcal{Z}} f(\theta, z; z_0)$, we define the potentially biased errors $\delta^t = g^t - \nabla_\theta \phi_\gamma(\theta^t; z^t)$. Substituting the this into the progress guarantee (27), we have

$$\begin{aligned}
F(\theta^{t+1}) &\le F(\theta^t) - \alpha_t \left(1 - \frac{1}{2} L_\phi \alpha_t\right) \left\|\nabla F(\theta^t)\right\|_2^2 + \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - \nabla_\theta \phi_\gamma(\theta; z^t)\right\rangle \\
&\quad - \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \delta^t \right\rangle + \frac{L_\phi \alpha_t^2}{2} \left\|\nabla_\theta \phi_\gamma(\theta; z^t) + \delta^t - \nabla F(\theta^t)\right\|_2^2 \\
&= F(\theta^t) - \alpha_t \left(1 - \frac{1}{2} L_\phi \alpha_t\right) \left\|\nabla F(\theta^t)\right\|_2^2 + \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - \nabla_\theta \phi_\gamma(\theta; z^t)\right\rangle \\
&\quad - \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \delta^t \right\rangle \\
&\quad + \frac{L_\phi \alpha_t^2}{2} \left(\left\|\delta^t\right\|_2^2 + \left\|\nabla_\theta \phi_\gamma(\theta; z^t) - \nabla F(\theta^t)\right\|_2^2 + 2 \left\langle \nabla_\theta \phi_\gamma(\theta; z^t) - \nabla F(\theta^t), \delta^t \right\rangle \right).
\end{aligned}$$

Using $\pm \langle a, b \rangle \le \frac{1}{2} \|a\|_2^2 + \frac{1}{2} \|b\|_2^2$ in the preceding display, we get

$$\begin{aligned}
F(\theta^{t+1}) &\le F(\theta^t) - \frac{\alpha_t}{2} \left\|\nabla F(\theta^t)\right\|_2^2 + \alpha_t \left(1 - L_\phi \alpha_t\right) \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - \nabla_\theta \phi_\gamma(\theta; z^t)\right\rangle \\
&\quad + \frac{\alpha_t \left(1 + L_\phi \alpha_t\right)}{2} \left\|\delta^t\right\|_2^2 + L_\phi \alpha_t^2 \left\|\nabla_\theta \phi_\gamma(\theta; z^t) - \nabla F(\theta^t)\right\|_2^2 \qquad (28)
\end{aligned}$$

Letting $z_\star^t = \mathop{\mathrm{argmax}}_z f(\theta^t, z; z^t)$, note that the error δ^t satisfies

$$\begin{aligned}
\left\|\delta^t\right\|_2^2 = \left\|\nabla_\theta \phi_\gamma(\theta^t; z^t) - \nabla_\theta f(\theta, \widehat{z}^t; z^t)\right\|_2^2 &= \left\|\nabla_\theta \ell(\theta, z_\star^t) - \nabla_\theta \ell(\theta, \widehat{z}^t)\right\|_2^2 \\
&\le L_{\theta\mathbf{z}}^2 \|\widehat{z}^t - z_\star^t\|_2^2 \le \frac{2L_{\theta\mathbf{z}}^2}{\lambda} \epsilon,
\end{aligned}$$

where the final inequality uses the $\lambda = \gamma - L_{\mathbf{zz}}$ strong-concavity of $z \mapsto f(\theta, z; z_0)$. For shorthand, let $\widehat{\epsilon} = \frac{2L_{\theta\mathbf{z}}^2}{\gamma - L_{\mathbf{zz}}} \epsilon$. Taking conditional expectations in the bound (28) and using $\mathbb{E}[\nabla_\theta \phi_\gamma(\theta^t; z^t) \mid \theta^t] = \nabla F(\theta^t)$, we have

$$\begin{aligned}
\mathbb{E}[F(\theta^{t+1}) - F(\theta^t) \mid \theta^t] &\le -\frac{\alpha_t}{2} \left\|\nabla F(\theta^t)\right\|_2^2 + \frac{\alpha_t (1 + L_\phi \alpha_t)}{2} \widehat{\epsilon} + L_\phi \alpha_t^2 \left\|\nabla_\theta \phi_\gamma(\theta; z^t) - \nabla F(\theta^t)\right\|_2^2 \\
&\le -\frac{\alpha_t}{2} \left\|\nabla F(\theta^t)\right\|_2^2 + \alpha_t \widehat{\epsilon} + L_\phi \alpha_t^2 \left\|\nabla_\theta \phi_\gamma(\theta; z^t) - \nabla F(\theta^t)\right\|_2^2,
\end{aligned}$$

where we use the fact that $\alpha_t \le \frac{1}{L_\phi}$. For a fixed stepsize α, taking the total expectation yields

$$\mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - 2\widehat{\epsilon} \le \frac{2}{\alpha} \mathbb{E}[F(\theta^t) - F(\theta^{t+1})] + 2L_\phi \alpha \sigma^2$$

since we have $\mathbb{E}[\|\nabla \phi_\gamma(\theta; Z) - \nabla F(\theta)\|_2^2] \le \sigma^2$ by assumption. Summing over t, we have

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - 2\widehat{\epsilon} &\le \frac{2}{\alpha T} \left(F(\theta^0) - \mathbb{E}[F(\theta^T)]\right) + 2L_\phi \alpha \sigma^2 \\
&\le \frac{2\Delta_F}{\alpha T} + 2L_\phi \alpha \sigma^2,
\end{aligned}$$

where the latter inequality holds since $\inf_\theta F(\theta) \le F(\theta^T)$. Plugging in $\alpha = \sqrt{\frac{\Delta_F}{L_\phi T \sigma^2}}$ gives the result.

C.4 Proof of Theorem 3

We first show the bound (11). From the duality result (5), we have the deterministic result that

$$\sup_{P : W_c(P, Q) \le \rho} \mathbb{E}_Q[\ell(\theta; Z)] \le \gamma \rho + \mathbb{E}_Q[\phi_\gamma(\theta; Z)]$$

for all $\rho > 0$, distributions Q, and $\gamma \ge 0$. Next, we show that $\mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]$ concentrates around its population counterpart at the usual rate (Boucheron et al., 2005).

First, we have that
$$\phi_\gamma(\theta; z) \in [-M_\ell, M_\ell],$$
because $-M_\ell \le \ell(\theta; z) \le \phi_\gamma(\theta; z) \le \sup_z \ell(\theta; z) \le M_\ell$. Thus, the functional $\theta \mapsto F_n(\theta)$ satisfies bounded differences (Boucheron et al., 2013, Thm. 6.2), and applying standard results on Rademacher complexity (Bartlett & Mendelson, 2002) and entropy integrals (van der Vaart & Wellner, 1996, Ch. 2.2) gives the result.

To see the second result (12), we substitute $\rho = \widehat{\rho}_n$ in the bound (11). Then, with probability at least $1 - e^{-t}$, we have
$$\sup_{P : W_c(P, P_0) \le \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] \le \gamma \widehat{\rho}_n(\theta) + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + \epsilon_n(t).$$
Since we have
$$\sup_{P : W_c(P, \widehat{P}_n) \le \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] = \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + \gamma \widehat{\rho}_n(\theta).$$
from the strong duality in Proposition 1, our second result follows.

C.5 Proof of Corollary 1

The result is essentially standard (van der Vaart & Wellner, 1996), which we now give for completeness. Note that for $\mathcal{F} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$, any $(\epsilon, \|\cdot\|)$-covering $\{\theta_1, \ldots, \theta_N\}$ of Θ guarantees that $\min_i |\ell(\theta; z) - \ell(\theta_i; z)| \le L\epsilon$ for all θ, z, or
$$N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{Z})}) \le N(\Theta, \epsilon/L, \|\cdot\|) \le \left(1 + \frac{\operatorname{diam}(\Theta) L}{\epsilon}\right)^d,$$
where $\operatorname{diam}(\Theta) = \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\|$. Noting that $|\ell(\theta; Z)| \le L \operatorname{diam}(\Theta) + M_0 =: M_\ell$, we have the result.

C.6 Proof of Theorem 4

Define
$$P_n^*(\theta) := \operatorname*{argmax}_P \left\{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, \widehat{P}_n)\right\},$$
$$P^*(\theta) := \operatorname*{argmax}_P \left\{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, P_0)\right\}.$$

First, we show that $P^*(\theta)$ and $P_n^*(\theta)$ are attained for all $\theta \in \Theta$. We omit the dependency on θ for notational simplicity and only show the result for $P^*(\theta)$ as the case for $P_n^*(\theta)$ is symmetric. Let P^ϵ be an ϵ-maximizer, so that
$$\mathbb{E}_{P^\epsilon}[\ell(\theta; Z)] - \gamma W_c(P^\epsilon, P_0) \ge \sup_P \left\{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P_n, P_0)\right\} - \epsilon.$$

As $\mathcal{Z}$ is compact, the collection $\{P^{1/k}\}_{k \in \mathbb{N}}$ is a uniformly tight collection of measures. By Prohorov's theorem (Billingsley, 1999, Ch 1.1, p. 57), (restricting to a subsequence if necessary), there exists some distribution P^* on $\mathcal{Z}$ such that $P^{1/k} \xrightarrow{d} P^*$ as $k \to \infty$. Continuity properties of Wasserstein distances (Villani, 2009, Corollary 6.11) then imply that
$$\lim_{k \to \infty} W_c(P^{1/k}, P_0) = W_c(P^*, P_0). \tag{29}$$

Combining (29) and the monotone convergence theorem, we obtain
$$\begin{aligned}\mathbb{E}_{P^*}[\ell(\theta; Z)] - \gamma W_c(P^*, P_0) &= \lim_{k \to \infty} \left\{\mathbb{E}_{P^{1/k}}[\ell(\theta; Z)] - \gamma W_c(P^{1/k}, P_0)\right\} \\ &\ge \sup_P \left\{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, P_0)\right\}.\end{aligned}$$

We conclude that P^* is attained for all P_0.

Next, we show the concentration result (30). Recall the definition (9) of the transportation mapping

$$T(\theta, z) := \underset{z' \in \mathcal{Z}}{\operatorname{argmax}} \left\{ \ell(\theta; z') - \gamma c(z', z) \right\},$$

which is unique and well-defined under our strong concavity assumption that $\gamma > L_{\mathsf{zz}}$, and smooth (recall Eq. (16)) in θ. Then by Proposition 1 (or by using a variant of Kantorovich duality (Villani, 2009, Chs. 9–10)), we have

$$\mathbb{E}_{P_n^*(\theta)}[\ell(\theta; Z) = \mathbb{E}_{\widehat{P}_n}[\ell(\theta; T(\theta; Z))] \text{ and } \mathbb{E}_{P^*(\theta)}[\ell(\theta; Z) = \mathbb{E}_{P_0}[\ell(\theta; T(\theta; Z))]$$
$$W_c(P_n^*(\theta), \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), Z)] \text{ and } W_c(P^*(\theta), P_0) = \mathbb{E}_{P_0}[c(T(\theta; Z), Z)].$$

We now proceed by showing the uniform convergence of

$$\mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), Z)] \quad \text{to} \quad \mathbb{E}_{P_0}[c(T(\theta; Z), Z)]$$

under both cases (i), that c is Lipschitz, and (ii), that ℓ is Lipschitz in z, using a covering argument on Θ. Recall inequality (16) (i.e. Lemma 1), which is that

$$\|T(\theta_1; z) - T(\theta_2; z)\| \leq \frac{L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+} \|\theta_1 - \theta_2\|.$$

We have the following lemma.

Lemma 4. *Assume the conditions of Theorem 4. Then for any $\theta_1, \theta_2 \in \Theta$,*

$$|c(T(\theta_1; z), z) - c(T(\theta_2; z), z)| \leq \frac{L_{\mathsf{c}} L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+} \|\theta_1 - \theta_2\|.$$

Proof In the first case, that c is L_{c}-Lipschitz in its first argument, this is trivial: we have

$$|c(T(\theta_1; z), z) - c(T(\theta_2; z), z)| \leq L_{\mathsf{c}} \|T(\theta_1; z) - T(\theta_2; z)\| \leq \frac{L_{\mathsf{c}} L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+} \|\theta_1 - \theta_2\|$$

by the smoothness inequality (16) for T.

In the second case, that $z \mapsto \ell(\theta, z)$ is L_{c}-Lipschitz, let $z_i = T(\theta_i; z)$ for shorthand. Then we have

$$\begin{aligned} \gamma c(z_2, z) - \gamma c(z_1, z) &= \gamma c(z_2, z) - \ell(\theta_2, z_2) + \ell(\theta_2, z_2) - \gamma c(z_1, z) \\ &\leq \gamma c(z_1, z) - \ell(\theta_2, z_1) + \ell(\theta_2, z_2) - \gamma c(z_1, z) = \ell(\theta_2, z_2) - \ell(\theta_2, z_1), \end{aligned}$$

and similarly,

$$\begin{aligned} \gamma c(z_2, z) - \gamma c(z_1, z) &= \gamma c(z_2, z) - \ell(\theta_1, z_1) + \ell(\theta_1, z_1) - \gamma c(z_1, z) \\ &\geq \gamma c(z_2, z) - \ell(\theta_1, z_1) + \ell(\theta_1, z_2) - \gamma c(z_2, z) = \ell(\theta_1, z_2) - \ell(\theta_1, z_1). \end{aligned}$$

Combining these two inequalities and using that

$$|\ell(\theta, z_2) - \ell(\theta, z_1)| \leq \gamma L_{\mathsf{c}} \|z_2 - z_1\|$$

for any θ gives the result. □

Using Lemma 4 we obtain that $\theta \mapsto |\mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), \theta)] - \mathbb{E}_{P_0}[c(T(\theta; Z), Z)]|$ is $2L_{\mathsf{c}} L_{\mathsf{z}\theta} / [\gamma - L_{\mathsf{zz}}]_+$-Lipschitz. Let $\Theta_{\text{cover}} = \{\theta_1, \cdots, \theta_N\}$ be a $\frac{[\gamma - L_{\mathsf{zz}}]_+ t}{4 L_{\mathsf{c}} L_{\mathsf{z}\theta}}$-cover of Θ with respect to $\|\cdot\|$. From Lipschitzness of $|\mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T(\theta; Z), Z)]|$, we have that if for all $\theta \in \{\Theta_{\text{cover}}\}$,

$$|\mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T(\theta; Z), \theta)]| \leq \frac{t}{2},$$

then it follows that

$$\sup_{\theta \in \Theta} |\mathbb{E}_{\widehat{P}_n}[c(T(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T(\theta; Z), Z)]| \leq t.$$

Under the first assumption (i), we have $|c(T(\theta;Z),Z)| \le 2L_{\mathsf{c}}M_{\mathsf{z}}$. Applying Hoeffding's inequality, for any fixed $\theta \in \Theta$

$$\mathbb{P}\left(|\mathbb{E}_{\widehat{P}_n}[c(T(\theta;Z),Z)] - \mathbb{E}_{P_0}[c(T(\theta;Z),Z)]| \ge \frac{t}{2}\right) \le 2\exp\left(-\frac{nt^2}{32L_{\mathsf{c}}^2M_{\mathsf{z}}^2}\right).$$

Taking a union bound over $\theta_1, \cdots, \theta_N$, we conclude that

$$\mathbb{P}\left(\sup_{\theta\in\Theta}|\mathbb{E}_{\widehat{P}_n}[c(T(\theta;Z),Z)] - \mathbb{E}_{P_0}[c(T(\theta;Z),Z)]| \ge t\right) \le 2N\left(\Theta, \frac{[\gamma - L_{\mathsf{zz}}]_+ t}{4L_{\mathsf{c}}L_{\mathsf{z}\theta}}, \|\cdot\|\right)\exp\left(-\frac{nt^2}{32L_{\mathsf{c}}^2M_{\mathsf{z}}^2}\right)$$

which was our desired result (30).

Under the second assumption (ii), we have from the definition of the transport map T

$$\gamma c(T(\theta;z),z) \le \ell(\theta;z) \le M_\ell$$

and hence $|c(T(\theta;Z),Z)| \le M_\ell/\gamma$. The result for the second case follows from an identical reasoning.

D Supervised Learning

In supervised learning settings, it is often natural—for example, in classification—to only consider adversarial perturbations to the feature vectors (covariates). In this section, we give an adapation of the results in Sections 2 and 3 (Theorems 2 and 4) to such scenarios. Let $Z = (X, Y) \in \mathcal{X} \times \mathbb{R}$ where $X \in \mathcal{X}$ is a feature vector[5] and $Y \in \mathbb{R}$ is a label. In classification settings, we have $Y \in \{1, \dots, K\}$. We consider an adversary that can only perturb the feature vector X (Goodfellow et al., 2015), which can be easily represented in our robust formulation (2) by defining the cost function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$ as follows: for $z = (x, y)$ and $z' = (x', y')$, recall the covariate shift cost function (8)

$$c(z, z') := c_x(x, x') + \infty \cdot \mathbf{1}\{y \neq y'\},$$

where $c_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is the transportation cost for the feature vector X. As before, we assume that c_x is nonnegative, continuous, convex in its first argument and satisfies $c_x(x, x) = 0$.

Under the cost function (8), the robust surrogate loss in the penalty problem (2) and its empirical counterpart (7) becomes

$$\phi_\gamma(\theta; (x_0, y_0)) = \sup_{x\in\mathcal{X}} \{\ell(\theta; (x, y_0)) - \gamma c_x(x, x_0)\}.$$

Similarly as in Section 2.1, we require the following two assumptions that guarantee efficient computability of the robust surrogate ϕ_γ.

Assumption C. *The function $c_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is continuous. For each $x_0 \in \mathcal{X}$, $c_x(\cdot, x_0)$ is 1-strongly convex with respect to the norm $\|\cdot\|$.*

Let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$; we again abuse notation by using the same norm $\|\cdot\|$ on Θ and $\mathcal{X}$, though the specific norm is clear from context.

Assumption D. *The loss $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ satisfies the Lipschitzian smoothness conditions*

$$\|\nabla_\theta \ell(\theta; (x, y)) - \nabla_\theta \ell(\theta'; (x, y))\|_* \le L_{\theta\theta}\|\theta - \theta'\|, \quad \|\nabla_x \ell(\theta; (x, y)) - \nabla_x \ell(\theta; (x', y))\|_* \le L_{\mathsf{xx}}\|x - x'\|,$$
$$\|\nabla_\theta \ell(\theta; (x, y)) - \nabla_\theta \ell(\theta; (x', y))\|_* \le L_{\theta\mathsf{x}}\|x - x'\|, \quad \|\nabla_x \ell(\theta; (x, y)) - \nabla_x \ell(\theta'; (x, y))\|_* \le L_{\mathsf{x}\theta}\|\theta - \theta'\|.$$

Under Assumptions C and D, an analogue of Lemma 1 still holds. The proof of the following result is nearly identical to that of Lemma 1; we state the full result for completeness.

Lemma 5. *Let $f : \Theta \times \mathcal{X} \to \mathbb{R}$ be differentiable and λ-strongly concave in x with respect to the norm $\|\cdot\|$, and define $\bar{f}(\theta) = \sup_{x\in\mathcal{X}} f(\theta, x)$. Let $\mathsf{g}_\theta(\theta, x) = \nabla_\theta f(\theta, x)$ and $\mathsf{g}_{\mathsf{x}}(\theta, x) = \nabla_x f(\theta, x)$, and assume g_θ and g_{x} satisfy the Lipschitz conditions of Assumption B. Then $\bar{f}$ is differentiable, and letting $x^\star(\theta) = \mathrm{argmax}_{x\in\mathcal{X}} f(\theta, x)$, we have $\nabla \bar{f}(\theta) = \mathsf{g}_\theta(\theta, x^\star(\theta))$. Moreover,*

$$\|x^\star(\theta_1) - x^\star(\theta_2)\| \le \frac{L_{\mathsf{x}\theta}}{\lambda}\|\theta_1 - \theta_2\| \quad \textit{and} \quad \left\|\nabla\bar{f}(\theta) - \nabla\bar{f}(\theta')\right\|_\star \le \left(L_{\theta\theta} + \frac{L_{\theta\mathsf{x}}L_{\mathsf{x}\theta}}{\lambda}\right)\|\theta - \theta'\|.$$

[5] We assume that $\mathcal{X}$ is a subset of normed vector space.

From Lemma 5, our previous results (Theorems 2 and 4) follow. The following is an analogue of Theorem 2 for the cost function (8).

Theorem 6 (Convergence of Nonconvex SGD)**.** *Let Assumptions C and D hold with the ℓ_2-norm and let $\Theta = \mathbb{R}^d$. Let $\Delta_F \geq F(\theta^0) - \inf_\theta F(\theta)$. Assume $\mathbb{E}[\|\nabla F(\theta) - \nabla_\theta \phi_\gamma(\theta, Z)\|_2^2] \leq \sigma^2$ and take constant stepsizes $\alpha = \sqrt{\frac{\Delta_F}{L_\phi T \sigma^2}}$ where $L_\phi := L_{\theta\theta} + \frac{L_{\theta \mathsf{x}} L_{\mathsf{x}\theta}}{\gamma - L_{\mathsf{xx}}}$. For $T \geq \frac{L_\phi \Delta_F}{\sigma^2}$, Algorithm 1 satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - \frac{4L_{\theta\mathsf{x}}^2}{\gamma - L_{\mathsf{xx}}}\epsilon \leq 4\sigma\sqrt{\frac{L_\phi \Delta_F}{T}}.$$

Similarly, an analogous result to Theorem 4 holds. Define the transport map for the covariate shift

$$T_\gamma(\theta; (x_0, y_0)) := \operatorname*{argmax}_{x \in \mathcal{X}} \{\ell(\theta; (x, y_0)) - \gamma c_x(x, x_0)\}.$$

Theorem 7. *Let $\mathcal{Z} \subset \{z \in \mathbb{R}^m : \|z\| \leq M_{\mathsf{z}}\}$ so that $\|Z\| \leq M_{\mathsf{z}}$ almost surely and assume either that (i) $c_x(\cdot, \cdot)$ is L_{c}-Lipschitz over $\mathcal{X}$ with respect to the norm $\|\cdot\|$ in each argument, or (ii) that $\ell(\theta, z) \in [0, M_\ell]$ and $x \mapsto \ell(\theta, (x, y))$ is γL_{c}-Lipschitz for all $\theta \in \Theta$. If Assumptions C and D hold, then with probability at least $1 - e^{-t}$,*

$$\sup_{\theta \in \Theta} |\mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T_\gamma(\theta; Z), Z)]| \leq 4D\sqrt{\frac{1}{n}\left(t + \log N\left(\Theta, \frac{[\gamma - L_{\mathsf{xx}}]_+ t}{4L_{\mathsf{c}} L_{\mathsf{x}\theta}}, \|\cdot\|\right)\right)}. \tag{30}$$

where $B = L_{\mathsf{c}} M_{\mathsf{z}}$ under assumption (i) and $B = M_\ell / \gamma$ under assumption (ii).

For both results, the proofs are essentially identical as before, but with an application of Lemma 5 instead of Lemma 1.

E Proximal algorithm for $\|\cdot\|_\infty$-norm robustness

In this section, we give a efficient training algorithm that learns to defend against $\|\cdot\|_\infty$-norm perturbations. For simplicity, we assume $\mathcal{Z} = \mathbb{R}^m$ for the rest of this section. Let $\theta \in \Theta$ be some fixed model, $z^0 \in \mathcal{Z}$ a natural example[6] and define $f(z) := \ell(\theta; z)$ to ease notation. Concretely, we are interested in solving the optimization problem

$$\operatorname*{maximize}_z f(z) - \frac{\alpha}{2}\left\|z - z^0\right\|_\infty^2$$

Note that this is equivalent to computing the surrogate loss $\phi_\gamma(\theta; z^0) = \sup_{z \in \mathcal{Z}}\{\ell(\theta; z) - \gamma c(z, z^0)\}$ for $c(z, z^0) = \left\|z - z^0\right\|_\infty^2$ and $\alpha = 2\gamma$. Our following treatment can easily be modified for the supervised learning scenario $c((x, y), (x^0, y^0)) = \left\|x - x^0\right\|_\infty^2 + \infty \cdot \mathbf{1}\left\{y = y^0\right\}$ with the convention that $\infty \cdot 0 = 0$. To make our notation consistent with the optimization literature, we consider the minimization problem

$$\operatorname*{minimize}_z -f(z) + \frac{\alpha}{2}\left\|z - z^0\right\|_\infty^2. \tag{31}$$

A simple gradient descent algorithm applied to the problem (31) may be slow to converge in practice. Intuitively, this is because the subgradient of $z \mapsto \frac{1}{2}\left\|z - z^0\right\|_\infty^2$ is given by $\left\|z - z^0\right\|_\infty \cdot s$ where s is a m-dimensional vector taking values in $[-1, 1]$ whose coordinates are non-zero only when $|z_j - z_{0,j}| = \left\|z - z^0\right\|_\infty$. Hence, at any given iteration of gradient descent, the $\|\cdot\|_\infty$-norm penalty term only gets accounted for by at most a few coordinates.

To remedy this issue, we consider a proximal algorithm for solving the problem (31) (see, for example, Parikh & Boyd (2013) for an comprehensive review of proximal algorithms). For a function $g : \mathcal{Z} \to \mathbb{R}$ and a positive number $\lambda > 0$, the proximal operator for λg is defined by

$$\mathrm{prox}_{\lambda g}(v) := \operatorname*{argmin}_z \left\{g(z) + \frac{1}{2\lambda}\|z - v\|_2^2\right\}.$$

[6] We depart from our convention of denoting original datapoints as z_0 to ease forthcoming notation.

Algorithm 2 Proximal Algorithm for Maximizing $f(z) - \frac{\alpha}{2} \left\| z - z^0 \right\|_\infty^2$

INPUT: Stepsizes λ^t
for $t = 0, \ldots, T-1$ **do**
$z^{t+\frac{1}{2}} \leftarrow z^t + \lambda^t \nabla f(z^t)$
$v^t \leftarrow \text{sort}(|z^{t+\frac{1}{2}} - z^0|, \text{dec})$
Compute β^t as in (33)
$z^{t+1} \leftarrow z^{t+\frac{1}{2}} - \left[|z^{t+\frac{1}{2}} - z^0| - \beta^t \right]_+ \text{sign}\left(z^{t+\frac{1}{2}} - z^0 \right)$

Then, the proximal algorithm on the problem (31) consists of two steps at each iteration t: (i) for the smooth function $-f(z)$, take a gradient descent step at the current iterate z^t ($z^{t+\frac{1}{2}}$ below) and (ii) for the non-smooth function $\left\| z - z^0 \right\|_\infty^2$, take a proximal step for the function $\frac{\lambda^t \alpha}{2} \left\| \cdot - z^0 \right\|_\infty^2$ at $z^{t+\frac{1}{2}}$ (z^{t+1} below):

$$z^{t+\frac{1}{2}} = z^t + \lambda^t \nabla f(z^t), \qquad z^{t+1} = \text{prox}_{\frac{\lambda^t \alpha}{2} \| \cdot - z^0 \|_\infty^2} \left(z^{t+\frac{1}{2}} \right). \tag{32}$$

The following proposition shows that we can compute the proximal step z^{t+1} efficiently, simply by sorting the vector $|z^{t+\frac{1}{2}} - z^0|$. We denote by v^t, the sorted vector of $|z^{t+\frac{1}{2}} - z^0|$ in **decreasing** order. In the proposition, we use the notation $[\cdot]_+ = \max(\cdot, 0)$.

Proposition 8. *Define the scale parameter $\beta^t > 0$ by*

$$\beta^t := \frac{1}{1 + \alpha \lambda^t j^t} \sum_{i=1}^{j^t} v_i^t \ \ \textit{where} \ \ j^t := \max \left\{ j \in [m] : \sum_{i=1}^{j-1} v_i - \left(\frac{1}{\alpha \lambda^t} + (j-1) \right) v_j < 0 \right\}. \tag{33}$$

Then, z^{t+1} in the proximal update (32) *is given by*

$$z^{t+1} = z^{t+\frac{1}{2}} - \left[|z^{t+\frac{1}{2}} - z^0| - \beta^t \right]_+ \text{sign}\left(z^{t+\frac{1}{2}} - z^0 \right). \tag{34}$$

See Section E.1 for the proof of the proposition. From the proposition, we obtain the proximal procedure in Algorithm 2 that can be used to solve for the approximate maximizer of $\ell(\theta; z) - \gamma c(z, z^0)$ in Algorithm 1. Heuristically, ignoring the truncation term in the proximal update (34), we have

$$z^{t+1} \approx z^0 + \beta^t \, \text{sign}(z^t + \lambda^t \nabla f(z^t) - z^0).$$

Here, we move towards the sign of $z^t + \lambda^t \nabla f(z^t) - z^0$ modulated by the term β^t, as opposed to just the sign of $\nabla f(z^t)$ for the iterated fast sign gradient method (Goodfellow et al., 2015; Kurakin et al., 2016).

E.1 Proof of Proposition 8

In this proof, we drop the subscript on the iteration t to ease notation. We assume without loss of generality that $z^{t+\frac{1}{2}} - z^0 \neq 0$. For some convex, lower semi-continuous function $g : \mathbb{R}^m \to \mathbb{R}$, let $g^*(s) = \sup_s \{ s^\top t - g(t) \}$ be the Fenchel conjuagte of g. From the Moreau decomposition (Parikh & Boyd, 2013, Section 2.5), we have

$$\text{prox}_g(w) + \text{prox}_{g^*}(w) = w$$

for any $w \in \mathbb{R}^m$. Noting that the conjugate of $z \mapsto \frac{\alpha\lambda}{2} \| z - z_0 \|_\infty^2$ is given by $z \mapsto z^\top z_0 + \frac{1}{2\alpha\lambda} \| z \|_1^2$, we have

$$\text{prox}_{\frac{\alpha\lambda}{2} \| \cdot - z_0 \|_\infty^2}(w) = w - \text{prox}_{\langle z^0, \cdot \rangle + \frac{1}{2\alpha\lambda} \| \cdot \|_1^2}(w) = w - \text{prox}_{\frac{1}{2\alpha\lambda} \| \cdot \|_1^2}(w - z^0)$$

Let us denote the sorted vector (in decreasing order) of $|w - z^0|$ by v. Then, in light of the preceeding display, it suffices to show that

$$\text{prox}_{\frac{1}{2\alpha\lambda} \| \cdot \|_1^2}(w - z^0) = [|w - z| - \beta^\star]_+ \, \text{sign}\left(w - z^0 \right) \tag{35}$$

where $\beta^\star$ is defined as in (33). To show that equality (35) holds, note that the first order optimality conditions for

$$\mathrm{prox}_{\frac{1}{2\alpha\lambda}\|\cdot\|_1^2}(w - z^0) = \operatorname*{argmin}_z \left\{ \frac{1}{2}\|z\|_1^2 + \frac{\alpha\lambda}{2}\left\|z - (w - z^0)\right\|_2^2 \right\}$$

is given by

$$\|z\|_1 \operatorname{sign}(z_i) + \alpha\lambda(z_i - w_i + z_i^0) = 0 \quad \text{if } |z_i| \neq 0 \tag{36a}$$

$$\|z\|_1 [-1, 1] - \alpha\lambda(w_i - z_i^0) \ni 0 \quad \text{if } |z_i| = 0. \tag{36b}$$

Now, we use the following elementary lemma.

Lemma 6. *For $0 \neq v \geq 0$ with decreasing coordinates, the solution to the equation*

$$\sum_{i: v_i > \beta} (v_i - \beta) = \alpha\lambda\beta$$

exists and is given by

$$\beta^\star := \frac{1}{1 + \alpha\lambda j^\star} \sum_{i=1}^{j} v_i \ \ \textit{where} \ \ j^\star := \max\left\{ j \in [m] : \sum_{i=1}^{j-1} v_i - \left(\frac{1}{\alpha\lambda} + (j-1)\right) v_j < 0 \right\}.$$

Proof of Lemma First, note that $\beta \mapsto \sum_{i: v_i > \beta}(v_i - \beta) - \alpha\lambda\beta =: h(\beta)$ is decreasing. Noting that $\|v\|_1 > 0$ and $-\alpha\lambda\|v\|_\infty < 0$, there exists β' such that $h(\beta') = 0$ and $\beta' \in (0, \|v\|_\infty)$. Since v_i's are decreasing and nonnegative, there exists j' such that $v_{j'} > \beta' \geq v_{j'+1}$ (we abuse notation and let $v_{m+1} := 0$). Then, we have

$$\sum_{i=1}^{j'-1}(v_i - v_{j'}) - \alpha\lambda v_{j'} < 0 \leq \sum_{i=1}^{j'}(v_i - v_{j'+1}) - \alpha\lambda v_{j'+1}.$$

That is, $j' = j^\star$. Solving for β' in

$$0 = h(\beta') = \sum_{i=1}^{j^\star} v_i - (\alpha\lambda + j^\star)\,\beta',$$

we obtain $\beta' = \beta^\star$ as claimed. □

Now, define

$$z^\star = \left[|w - z^0| - \beta^\star\right]_+ \operatorname{sign}\left(w - z^0\right).$$

Then, we have from Lemma 6 that

$$\|z^\star\|_1 = \sum_{i: |w_i - z_i^0| > \beta^\star} (|w_i - z_i^0| - \beta^\star) = \sum_{i=1}^{j^\star}(v_i - \beta^\star) = \alpha\lambda\beta^\star.$$

If $z_i^\star > 0$, then $\operatorname{sign}(z_i^\star) = \operatorname{sign}(w_i - z_i^0)$ so that

$$\|z^\star\|_1 \operatorname{sign}(z_i) + \alpha\lambda(z_i^\star - w_i + z_i^0) = 0.$$

If $z_i^\star = 0$, then $|w_i - z_i^0| \leq \beta^\star$ and

$$\|z^\star\|_1 [-1, 1] - \alpha\lambda(w_i - z_i^0) = \alpha\lambda\beta^\star[-1, 1] - \alpha\lambda(w_i - z_i^0) \ni 0.$$

Hence, $z^\star$ satisfies the optimality condition (36) as desired.