Certifiable Distributional Robustness with Principled Adversarial Training

Aman Sinha* **Hongseok Namkoong*** **John Duchi**

{amans,hnamk,jduchi}@stanford.edu

Abstract

Neural networks are vulnerable to adversarial examples and researchers have proposed many heuristic attack and defense mechanisms. We take the principled view of distributionally robust optimization, which guarantees performance under adversarial input perturbations. By considering a Lagrangian penalty formulation of perturbation of 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. We match or outperform heuristic approaches on supervised learning tasks.

1 Introduction

Consider the classical supervised learning problem, where we minimize the expected loss $\mathbb{E}_{P_0}[\ell(\theta; Z)]$ over a parameter $\theta \in \Theta$, where ℓ is a loss and $Z \sim P_0$ is a distribution on a space $\mathcal{Z}$. In many systems, we desire robustness to changes in P_0, either from covariate shifts, changes in the underlying domain [3], or adversarial attacks [12, 18]. For deep networks in performance-critical systems (e.g. perception for self-driving cars or automated detection of tumors), model failure leads to life-threatening situations; in these systems, it is irresponsible to deploy models whose robustness we cannot certify.

However, recent works have shown that neural networks are vulnerable to adversarial examples; seemingly imperceptible perturbations to data can lead to misbehavior of the model, such as misclassifications of the output [12, 18, 21, 22]. Subsequently, many researchers have proposed adversarial attack and defense mechanisms [27, 23, 24, 25, 28, 8, 20, 13]. While these works provide an initial foundation for adversarial training, there are no guarantees on whether proposed heuristic attacks can find the most adversarial perturbation and whether there is a class of attacks such defenses can successfully prevent. On the other hand, verification of deep networks using SMT solvers [16, 17, 14] provides formal guarantees on robustness but is NP-hard in general; this approach requires prohibitive computational expense even on small networks.

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 $Z \sim P_0$ and consider the problem

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

The choice of $\mathcal{P}$ influences robustness guarantees and computability; we develop robustness sets $\mathcal{P}$ with computationally efficient relaxations that apply even when 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 for the worst-case population loss $\sup_{P \in \mathcal{P}} \mathbb{E}_P[\ell(\theta; Z)]$. This smoothness can be obtained in standard deep architectures with exponential linear units (ELU's) [10]. A simple Tensorflow implementation of our method takes 5–10$\times$ as long as stochastic gradient methods for empirical risk minimization (ERM).

Let us overview our approach briefly. Let $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$, where $c(z, z_0)$ represents the "cost" for an adversary to perturb z_0 to z (we typically use $c(z, z_0) = \|z - z_0\|_p^2$ for $p \geq 1$). We consider the region $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$, a ρ-neighborhood of the distribution P_0 under the Wasserstein metric $W_c(\cdot, \cdot)$ (formally defined in Appendix A). The formulation (1), however, is still intractable for arbitrary robustness ρ—at least for deep networks or other complex models. Instead, we consider its Lagrangian relaxation for a fixed penalty $\gamma \geq 0$, giving the following reformulation whose proof we defer to Appendix A:

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

*Equal contribution. Full paper available at https://arxiv.org/abs/1710.10571

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 $\ell(\theta^t; z) - \gamma c(z, z^t)$
  $\theta^{t+1} \leftarrow \mathsf{Proj}_\Theta(\theta^t - \alpha_t \nabla_\theta \ell(\theta^t; \widehat{z}^t))$

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

That is, we have replaced the usual $\ell(\theta; Z)$ by the robust surrogate $\phi_\gamma(\theta; z)$, which adversarially perturbs data z modulated by γ. We typically solve the problem (2) with P_0 replaced by the empirical distribution $\widehat{P}_n$.

The key feature of the penalty problem (2) is that moderate levels of robustness 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. As a consequence, the stochastic gradient method applied to problem (2) has similar convergence guarantees as for non-robust methods (ERM). In Section 3, we give a *certificate of robustness* showing that we are approximately protected against all distributional perturbations satisfying $W_c(P, P_0) \le \widehat{\rho}_n$, where $\widehat{\rho}_n$ is the achieved robustness for the empirical objective. We upper-bound the *population* worst-case scenario $\sup_{P: W_c(P,P_0) \le \widehat{\rho}_n} \mathbb{E}_P[\ell(\theta; Z)]$ by an efficiently computable empirical counterpart. These results suggest advantages of networks with smooth activations rather than ReLUs. 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.

2 Proposed approach

Our approach is motivated by the following insight: assume $z \mapsto \ell(\theta; z)$ is smooth, i.e. $\nabla_z \ell(\theta; z)$ is L-Lipschitz for some L. For $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ strongly convex in its first argument, a Taylor expansion yields

$$\ell(\theta; z') - \gamma c(z', z_0) \le \ell(\theta; z) - \gamma c(z, z_0) + \left\langle \nabla_z(\ell(\theta; z) - \gamma c(z, z_0)), z' - z \right\rangle + \frac{L - \gamma}{2} \left\| z - z' \right\|_2^2. \tag{3}$$

For $\gamma \ge L$ this is the first-order condition for concavity of $z \mapsto (\ell(\theta; z) - \gamma c(z, z_0))$. When ℓ is smooth and γ is large enough, the surrogate (2b) is a strongly-concave optimization. Leveraging this insight, instead of prescribing the amount ρ of robustness, we focus on the (empirical counterpart to the) penalty problem (2).

We now develop stochastic gradient-type methods for the relaxed robust problem (2), making clear the computational benefits of relaxing the strict robustness requirements of formulation (1). We begin with assumptions we require, which roughly quantify the amounts of robustness we can provide.

Assumption A. *$c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_+$ is continuous and for each $z \in \mathcal{Z}$, $c(\cdot, z)$ is 1-strongly convex w.r.t. $\|\cdot\|$.*

To guarantee that the robust surrogate (2b) is tractably computable, we require a few smoothness assumptions. Let $\|\cdot\|_*$ be the dual norm to $\|\cdot\|$; we overload notation with $\|\cdot\|$ 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*

$$\left\|\nabla_\theta \ell(\theta; z) - \nabla_\theta \ell(\theta'; z)\right\|_* \le L_{\theta\theta} \left\|\theta - \theta'\right\|, \quad \left\|\nabla_z \ell(\theta; z) - \nabla_z \ell(\theta; z')\right\|_* \le L_{zz} \left\|z - z'\right\|,$$
$$\left\|\nabla_\theta \ell(\theta; z) - \nabla_\theta \ell(\theta; z')\right\|_* \le L_{\theta z} \left\|z - z'\right\|, \quad \left\|\nabla_z \ell(\theta; z) - \nabla_z \ell(\theta'; z)\right\|_* \le L_{z\theta} \left\|\theta - \theta'\right\|.$$

These properties guarantee both (i) the smoothness of the robust surrogate ϕ_γ (see Lemma 1 in Appendix B) and (ii) its efficient computability. Since we only perturb features (and not labels) in supervised learning settings, we can easily modify these assumptions to work with cost c_x defined on features (see Appendix F). The well-behavedness of ϕ_γ motivates Algorithm 1, a stochastic-gradient approach for the penalty problem (2). 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.

When ℓ is nonconvex in θ[2], the following theorem guarantees convergence to a stationary point of problem (2) at rate $1/\sqrt{T}$ when $\gamma \ge L_{zz}$. Recall that $F(\theta) = \mathbb{E}_{P_0}[\phi_\gamma(\theta; Z)]$ is the robust surrogate objective for the Lagrangian relaxation (2). We prove the theorem in Section B.

Theorem 1 (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 the constant stepsize $\alpha = \sqrt{\frac{2\Delta_F}{L\sigma^2 T}}$, with $L = L_{\theta\theta} + \frac{L_{\theta z} L_{z\theta}}{\gamma - L_{zz}}$. Algorithm 1 satisfies $\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - \frac{2L_{\theta z}^2}{\gamma - L_{zz}}\epsilon \le \sigma\sqrt{8\frac{L\Delta_F}{T}}$.*

[2]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 [15, 9] at rate $1/\sqrt{T}$.

The condition $\mathbb{E}[\|\nabla F(\theta) - \nabla_\theta \phi_\gamma(\theta, Z)\|_2^2] \leq \sigma^2$ holds (to within a constant factor) whenever $\|\nabla_\theta \ell(\theta, z)\|_2 \leq \sigma$ for all θ, z. Theorem 1 shows that Algorithm 1 achieves the same rates of convergence on the penalty problem (2) as in standard smooth non-convex optimization [11]. Key to this result is that ℓ is smooth in z: the inner supremum (2b) is NP-hard to compute for non-smooth deep networks (see Lemma 2 in Section C for a proof with ReLU's). Replacing ReLU's with sigmoids or ELU's [10] allows us to apply Theorem 1, making distributionally robust optimization tractable for deep learning.

3 Robustness Certificate

From results in the previous section, Algorithm 1 provably learns to protect against adversarial perturbations on the training dataset. Now, we show that this procedure generalizes, 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. Our main result gives a data-dependent upper bound on the population worst-case objective $\sup_{P: W_c(P, P_0) \leq \rho} \mathbb{E}_P[\ell(\theta; Z)] \leq \gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + O(1/\sqrt{n})$ for all $\theta \in \Theta$ and 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 (2). Our bound is efficiently computable and hence *certifies* a level of robustness for the worst-case population objective.

To make this rigorous, fix $\gamma > 0$, and consider the worst-case perturbation, typically called the *transportation map* or Monge map [30], $T_\gamma(\theta; z_0) := \operatorname{argmax}_{z \in \mathcal{Z}} \{\ell(\theta; z) - \gamma c(z, z_0)\}$. Under our assumptions, it is easy to compute T_γ whenever $\gamma \geq L_{\mathbf{zz}}$. Letting δ_z denote the point mass at z, Proposition 3 (or Kantorovich duality [30, Chs. 9–10]) shows the empirical maximizers of the Lagrangian formulation (11) are attained by $P_n^*(\theta) := \operatorname{argmax}_P \left\{ \mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, \widehat{P}_n) \right\}$ and

$$P_n^*(\theta) = \frac{1}{n} \sum_{i=1}^n \delta_{T_\gamma(\theta, Z_i)} \quad \text{and} \quad \widehat{\rho}_n(\theta) := W_c(P_n^*(\theta), \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta; Z), Z)]. \tag{4}$$

The equalities (4) show that $\mathbb{E}_{P_n^*(\theta)}[\ell(\theta; Z)]$ is efficiently computable, thereby providing a data-dependent performance guarantee for the worst-case population loss.

Our bound relies on the usual covering numbers for the model class $\mathcal{F} = \{\ell(\theta; \cdot) : \theta \in \Theta\}$ as the notion of complexity [e.g. 29], so, despite the infinite-dimensional problem (2), 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 $\|v - v_i\| \leq \epsilon$. The *covering number* of V with respect to $\|\cdot\|$ is $N(V, \epsilon, \|\cdot\|) := \inf\{N \in \mathbb{N} \mid \text{there is an } \epsilon\text{-cover of } V \text{ with respect to } \|\cdot\|\}$. 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 whose proof we given in Section D.1. We first show from the duality result (11) 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 2. *Assume that $|\ell(\theta; z)| \leq M_\ell$ 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 $\theta \in \Theta$, and $\rho \geq 0$,*

$$\sup_{P: W_c(P, P_0) \leq \rho} \mathbb{E}_P[\ell(\theta; Z)] \leq \gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + \epsilon_n(t). \tag{5}$$

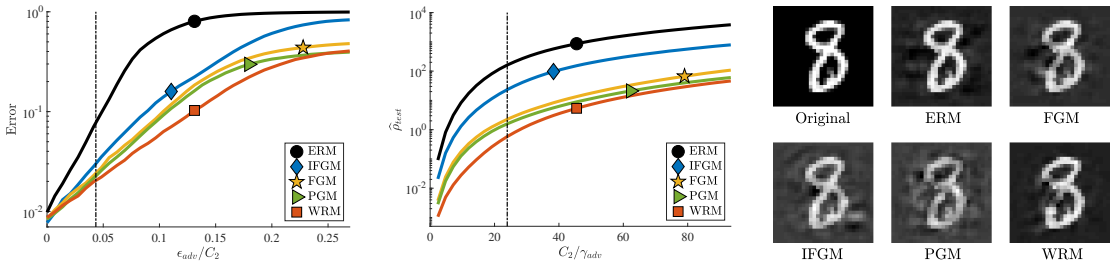
In particular, if $\rho = \widehat{\rho}_n(\theta)$ then with probability at least $1 - e^{-t}$, for all $\theta \in \Theta$

$$\sup_{P: W_c(P, P_0) \leq \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] \leq \sup_{P: W_c(P, \widehat{P}_n) \leq \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] + \epsilon_n(t). \tag{6}$$

A key consequence of the bound (5) is that $\gamma\rho + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]$ *certifies* robustness for the worst-case population objective at any level ρ. Letting $\theta = \theta_{\mathrm{WRM}}$, we expect the certificate (5) to be tight since θ_{WRM} was choosen to be close to the minimizer of $\mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]$. In particular, when $\rho = \widehat{\rho}_n(\theta)$, duality shows that $\mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + \gamma\widehat{\rho}_n(\theta) = \sup_{P: W_c(P, \widehat{P}_n) \leq \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] = \mathbb{E}_{P_n^*(\theta)}[\ell(\theta; Z)]$. (See Section D.1 for a proof of these equalities.) This certificate is easy to compute via expression (4): the transportation mappings $T(\theta, Z_i)$ are efficiently computable for large enough γ, as noted in Section 2, and $\widehat{\rho}_n = W_c(P_n^*, \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T(\theta, Z), Z)]$. See Corollary 1 for a concrete adaptation of Theorem 2 to Lipschitz functions. In Appendix D we also show that adversarial perturbations of the training data generalize.

4 Experiments

We consider a standard benchmark: training a neural network classifier on the MNIST dataset. We compare performance of our method (WRM) with ERM and models trained with other heuristic adversarial training



(a) Test error vs. ϵ_{adv} for $\|\cdot\|_2$ attack (b) $\widehat{\rho}_{test}$ vs. $1/\gamma_{\text{adv}}$ (c) Perturbations on a test datapoint

Figure 1. PGM and Wasserstein attacks on the MNIST dataset ($C_2 = 9.21$). (a) shows test misclassification error vs. adversarial perturbation level ϵ_{adv} for the 2-norm PGM attack. (b) and (c) shows stability of the loss surface. (b) plots average distance of the perturbed distribution $\widehat{\rho}_{\text{test}}$ for a given γ_{adv}. The vertical bar in (b) indicates the γ we use for training WRM, the bar in (a) is the corresponding ϵ. (c) visualizes the smallest WRM perturbation (largest γ_{adv}) necessary to make a model misclassify a datapoint. More experiments in Appendix E.

procedures: the fast-gradient method (FGM) [12], its iterated variant (IFGM) [18], and the projected-gradient method (PGM) [20]. 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)

$$\Delta x_i^{t+1}(\theta) := \underset{\|\eta\|_p \le \epsilon}{\text{argmax}} \{\nabla_x \ell(\theta; x_i^t, y_i)^T \eta\} \text{ and } x_i^{t+1} := \Pi_{\mathcal{B}_{\epsilon,p}(x_i^t)} \left\{x_i^t + \alpha_t \Delta x_i^t(\theta)\right\} \tag{7}$$

for $t = 1, \ldots, T_{\text{adv}}$, where Π denotes projection onto $\mathcal{B}_{\epsilon,p}(x_i) := \{x : \|x - x_i\|_p \le \epsilon\}$. We use the squared Euclidean cost for the feature vectors $c_x(x, x') := \|x - x'\|_2^2$ (and total cost $c(z, z') := \|x - x'\|_2^2 + \infty \cdot \mathbf{1}\{y \ne y'\}$) for WRM and $p = 2$ for FGM, IFGM, PGM training in all experiments; we test against adversarial perturbations with respect to the norms $p = 2$. We use $T_{\text{adv}} = 15$ iterations for all iterative methods (IFGM, PGM, and WRM) in training and attacks. Larger adversarial budgets correspond to smaller γ for WRM and larger ϵ for other models. Our network consists of 8×8, 6×6, 5×5 convolutional filter layers with ELU activations followed by a fully connected layer and softmax output. We train our method with $\gamma = .04\mathbb{E}_{\widehat{P}_n} \|X\|_2$, and for other methods we choose ϵ as the achieved level of robustness by WRM:[3]

$$\epsilon^2 = \widehat{\rho}_n(\theta_{\text{WRM}}) = W_c(P_n^*(\theta_{\text{WRM}}), \widehat{P}_n) = \mathbb{E}_{\widehat{P}_n}[c(T(\theta_{\text{WRM}}, Z), Z)]. \tag{8}$$

In the figures, we scale the budgets $1/\gamma_{\text{adv}}$ and ϵ_{adv} for the adversary with $C_2 := \mathbb{E}_{\widehat{P}_n} \|X\|_2 = 9.21$. All methods achieve at least 99% test-time accuracy on natural examples, implying there is little penalty for the level of robustness (ϵ and γ) used for training the adversarial models.

It is thus important to distinguish the methods' abilities to combat attacks. We test performance of the five methods under PGM attacks (7) with respect to the 2-norm. In Figure 1(a), 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 attacks, which we show in Appendix E.1.

Next we study stability of the loss surface with respect to perturbations to inputs. First, consider the distance to adversarial examples under the models $\theta = \theta_{\text{ERM}}, \theta_{\text{FGM}}, \theta_{\text{IFGM}}, \theta_{\text{PGM}}, \theta_{\text{WRM}}$,

$$\widehat{\rho}_{\text{test}}(\theta) := \mathbb{E}_{\widehat{P}_{\text{test}}}[c(T_{\gamma_{\text{adv}}}(\theta, Z), Z)], \tag{9}$$

where $\widehat{P}_{\text{test}}$ is the test distribution, $c(z, z') := \|x - x'\|_2^2 + \infty \cdot \mathbf{1}\{y \ne y'\}$ as before, and $T_{\gamma_{\text{adv}}}(\theta, Z) = \text{argmax}_z \{\ell(\theta; z) - \gamma_{\text{adv}} c(z, Z)\}$ is the adversarial perturbation of Z (Monge map) for the model θ. We note that small values of $\widehat{\rho}_{\text{test}}(\theta)$ correspond to small magnitudes of $\nabla_z \ell(\theta; z)$ in a neighborhood of the nominal input, which ensures stability of the model. Figure 1(b) shows that $\widehat{\rho}_{\text{test}}$ differs by orders of magnitude between the training methods; the trend is nearly uniform over all γ_{adv}, with θ_{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. The vertical bars indicate the perturbation level used for training the FGM, IFGM, and PGM models as well as the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$.

In Figure 1(c) 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 γ_{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 E.2 for examples of other digits); WRM defends against gradient-based attacks by learning a representation that makes gradients point towards inputs of other classes. Overall, Figure 1 depicts 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.

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

Acknowledgments

AS, HN, and JCD 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. JCD was also partially supported by the National Science Foundation award NSF-CAREER-1553086. We thank Jacob Steinhardt for valuable feedback.

References

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

[2] T. Başar and P. Bernhard. *H-infinity optimal control and related minimax design problems: a dynamic game approach*. Springer Science & Business Media, 2008.

[3] 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.

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

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

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

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

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

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

[10] 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.

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

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

[13] 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.

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

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

[16] 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]*, 2017.

[17] 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]*, 2017.

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

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

[20] 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.

[21] 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*, pages 2574–2582, 2016.

[22] 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*, pages 427–436, 2015.

[23] 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]*, 2016.

[24] 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*, pages 372–387. IEEE, 2016.

[25] 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*, pages 582–597. IEEE, 2016.

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

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

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

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

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

A Wasserstein Robustness and Duality

Wasserstein distances define a notion of closeness between distributions. Let $\mathcal{Z} \subset \mathbb{R}^m$ be convex, 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, continuous, and convex in its first argument 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 data generating distribution P_0, we consider the Wasserstein form of the robust problem (1), with $\mathcal{P} = \{P : W_c(P, P_0) \leq \rho\}$, and its Lagrangian relaxation (2) with $\gamma \geq 0$.

The following duality result, which we prove in Appendix A.1, gives the equality (2) and an analogous result for the worst-case problem (1).

Proposition 3. *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 $\rho > 0$,*

$$\sup_{P : W_c(P,Q) \leq \rho} \mathbb{E}_P[\ell(\theta; Z)] = \inf_{\gamma \geq 0} \left\{\gamma\rho + \mathbb{E}_Q[\phi_\gamma(\theta; Z)]\right\}, \tag{10}$$

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

$$\sup_P \left\{\mathbb{E}_P[\ell(\theta; Z)] - \gamma W_c(P, Q)\right\} = \mathbb{E}_Q[\phi_\gamma(\theta; Z)]. \tag{11}$$

Leveraging the insight (3), 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

$$\underset{\theta \in \Theta}{\text{minimize}} \left\{F_n(\theta) := \sup_P \left\{\mathbb{E}[\ell(\theta; Z)] - \gamma W_c(P, \widehat{P}_n)\right\} = \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)]\right\}. \tag{12}$$

A.1 Proof of Proposition 3

For completeness, we provide an alternative proof to that given in Blanchet and Murthy [5] 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 3 as an immediate special case. Recalling Rockafellar and Wets [26, 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

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

is closed-valued and measurable. We recall that if g is continuous, then g is a normal integrand [26, Cor. 14.34]; therefore, $g(x, z) = \gamma c(x, z) - \ell(\theta; x)$ is a normal integrand. We have the following theorem.

Theorem 4. *Let f, c be such that for any $\gamma \geq 0$, 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*

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

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 [19, Thm. 8.7.1] to obtain

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

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

$$\sup_{P : W_c(P,Q)} \int f(x) dP(x) = \inf_{\gamma \geq 0} \sup_{P, M \in \Pi(P,Q)} \left\{\int [f(x) - \gamma c(x, z)] dM(x, z) + \gamma\rho\right\}. \tag{13}$$

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 (13), 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 and Wets [26, 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 (11) and completes the proof. □

B Optimization

First, the following lemma shows (more generically) that if γ is large enough and Assumptions A and B hold, the surrogate ϕ_γ is still smooth.

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 the Lipschitz conditions of Assumption B. 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,*

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

See Section B.1 for the proof. Focusing on the ℓ_2-norm case, an immediate application of Lemma 1 shows that if Assumption B holds, then ϕ_γ has $L = L_{\theta\theta} + \frac{L_{\theta\mathsf{z}} L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{z}\mathsf{z}}]_+}$-Lipschitz gradients, and

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

We are now ready to give a proof of Theorem 1.

Proof of Theorem 1

Our proof is based on that of Ghadimi and Lan [11].

For shorthand, let $f(\theta, z; z_0) = \ell(\theta; z) - \gamma c(z, z_0)$, noting that we perform gradient steps with

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

for $\widehat{z}^t$ an ϵ-approximate maximizer of $f(\theta, z; z^t)$ in z, and $\theta^{t+1} = \theta^t - \alpha_t g^t$. By a Taylor expansion using the L-smoothness of the objective F, we have

$$\begin{aligned}
F(\theta^{t+1}) &\le F(\theta^t) + \left\langle \nabla F(\theta^t), \theta^{t+1} - \theta^t \right\rangle + \frac{L}{2} \left\| \theta^{t+1} - \theta^t \right\|_2^2 \\
&= F(\theta^t) - \alpha_t \left\| \nabla F(\theta^t) \right\|_2^2 + \frac{L\alpha_t^2}{2} \left\| g^t \right\|_2^2 + \alpha_t \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - g^t \right\rangle \\
&= F(\theta^t) - \alpha_t \left(1 - \frac{L\alpha_t}{2}\right) \left\| \nabla F(\theta^t) \right\|_2^2 \\
&\qquad + \alpha_t \left(1 + \frac{L\alpha_t}{2}\right) \left\langle \nabla F(\theta^t), \nabla F(\theta^t) - g^t \right\rangle + \frac{L\alpha_t^2}{2} \left\| g^t - \nabla F(\theta^t) \right\|_2^2 .
\end{aligned} \tag{14}$$

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 $\delta^t = g^t - \nabla_\theta \phi_\gamma(\theta^t; z^t)$. Letting $z_\star^t = \operatorname{argmax}_z f(\theta^t, z; z^t)$, these errors evidently satisfy

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

where the final inequality uses the $\lambda = \gamma - L_{\mathsf{zz}}$ strong-concavity of $z \mapsto f(\theta, z; z_0)$. For shorthand, let $\widehat{\epsilon} = \frac{2L_{\theta \mathsf{z}}^2}{\gamma - L_{\mathsf{zz}}} \epsilon$. Substituting the preceding display into the progress guarantee (14), we have

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

Noting that $\mathbb{E}[\nabla_\theta \phi_\gamma(\theta^t; z^t) \mid \theta^t] = \nabla F(\theta^t)$, we take expectations to find

$$\mathbb{E}[F(\theta^{t+1}) - F(\theta^t) \mid \theta^t] \le -\frac{\alpha_t}{2}(1 - L\alpha_t) \left\| \nabla F(\theta^t) \right\|_2^2 + \left(\frac{\alpha_t}{2} + \frac{5L\alpha_t^2}{4}\right) \widehat{\epsilon} + L\alpha_t^2 \sigma^2, \tag{15}$$

where we have used that $\mathbb{E}[\|\nabla \phi_\gamma(\theta; Z) - \nabla F(\theta)\|_2^2] \le \sigma^2$ by assumption.

The bound (15) gives the theorem essentially immediately for fixed stepsizes α, as we have

$$\frac{\alpha}{2}(1 - L\alpha) \mathbb{E}\left[\sum_{t=1}^T \left\| \nabla F(\theta^t) \right\|_2^2 \right] \le F(\theta^0) - \mathbb{E}[F(\theta^{T+1})] + \frac{T\alpha}{2} \left(1 + \frac{5L\alpha}{4}\right) \widehat{\epsilon} + TL\alpha^2 \sigma^2.$$

Noting that $\inf_\theta F(\theta) \le F(\theta^{T+1})$ gives the final result.

B.1 Proof of Lemma 1

Differentiability is a consequence of one of the many forms of Danskin's Theorem (e.g. Appendix B in [2]). For smoothness, 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 $\frac{\lambda}{2} \left\| z_1^\star - z_2^\star \right\|^2 \le f(\theta_2, z_2^\star) - f(\theta_2, z_1^\star)$ 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,

$$\left\| z_1^\star - z_2^\star \right\| \le \frac{L_{\mathsf{z}\theta}}{\lambda} \left\| \theta_1 - \theta_2 \right\|. \tag{16}$$

Then we have

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

where we have used inequality (16) again. This is the desired result.

C 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. [16]. In the following, we use polynomial time 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. There exist 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. [16] 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. □

D Generalization

First, we give a concrete variant of Theorem 2 for Lipschitz functions. When the parameter set Θ is finite dimensional ($\Theta \subset \mathbb{R}^d$), Theorem 2 provides a robustness guarantee scaling with d in spite of the infinite-dimensional Wasserstein penalty. Assuming there exist $\theta_0 \in \Theta$, $M_{\theta_0} < \infty$ such that such that $|\ell(\theta_0; z)| \le M_{\theta_0}$ for all $z \in \mathcal{Z}$, we have the following corollary (see Section D.3 for a proof).

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 $\operatorname{diam}(\Theta) = \sup_{\theta,\theta'\in\Theta} \|\theta - \theta'\| < \infty$. Then, the bounds* (5) *and* (6) *hold with*

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

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

Next, we show that adversarial perturbations on the training set (in a sense) generalize: solving the empirical penalty problem (12) guarantees a similar level of robustness as directly solving its population counterpart (2). Our starting point is Lemma 1, which shows that $T_\gamma(\cdot; z)$ is smooth under Assumptions A and B:

$$\|T_\gamma(\theta_1; z) - T_\gamma(\theta_2; z)\| \le \frac{L_{\mathsf{z}\theta}}{[\gamma - L_{\mathsf{zz}}]_+}\|\theta_1 - \theta_2\| \tag{17}$$

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 5. *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 $1 - e^{-t}$,

$$\sup_{\theta\in\Theta} |\mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta; Z), Z)] - \mathbb{E}_{P_0}[c(T_\gamma(\theta; Z), Z)]| \le 4D\sqrt{\frac{1}{n}\left(t + \log N\left(\Theta, \frac{[\gamma - L_{\mathsf{zz}}]_+ t}{4L_{\mathsf{c}}L_{\mathsf{z}\theta}}, \|\cdot\|\right)\right)}. \tag{18}$$

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

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

See Section D.2 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 (21) gives the usual $\sqrt{d/n}$ generalization rate for the distance between adversarial perturbations and natural examples. Another consequence of Theorem 5 is that $\widehat{\rho}_n(\theta_{\mathrm{WRM}})$ in the certificate (6) 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 (21).

D.1 Proof of Theorem 2

We first show the bound (5). From the duality result (10), 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 [6].

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 [7, Thm. 6.2], and applying standard results on Rademacher complexity [1] and entropy integrals [29, Ch. 2.2] gives the result.

To see the second result (6), we substitute $\rho = \widehat{\rho}_n$ in the bound (5). Then, with probability at least $1 - e^{-t}$, we have

$$\sup_{P: W_c(P,P_0) \le \widehat{\rho}_n(\theta)} \mathbb{E}_P[\ell(\theta; Z)] \le \gamma\widehat{\rho}_n(\theta) + \mathbb{E}_{\widehat{P}_n}[\phi_\gamma(\theta; Z)] + \epsilon_{n,1}(t).$$

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 3, our second result follows.

D.2 Proof of Theorem 5

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 [4, 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 [30, Corollary 6.11] then imply that

$$\lim_{k \to \infty} W_c(P^{1/k}, P_0) = W_c(P^*, P_0). \tag{19}$$

Combining (19) 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 (21). Recall the transportation mapping

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

which is unique and well-defined under our strong concavity assumption that $\gamma > L_{zz}$, and smooth (recall Eq. (17)) in θ. Then by Proposition 3 (or by using a variant of Kantorovich duality [30, 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

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

under both cases (i), that c is Lipschitz, and (ii), that ℓ is Lipschitz in z, using a covering argument on Θ. Recall inequality (17) (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 3. *Assume the conditions of Theorem 5. 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 (17) 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 3 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 $\frac{[\gamma - L_{\mathsf{zz}}]_+ t}{4L_{\mathsf{c}} L_{\mathsf{z}\theta}}$-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 $|c(T(\theta; Z), Z)| \leq 2L_{\mathsf{c}} M_{\mathsf{z}}$. Applying Hoefdding's inequality, for any fixed $\theta \in \Theta$

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

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

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

which was our desired result (21).

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

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

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

D.3 Proof of Corollary 1

The result is essentially standard [29], 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 $\min_i |\ell(\theta; z) - \ell(\theta_i; z)| \leq L\epsilon$ for all θ, z, or

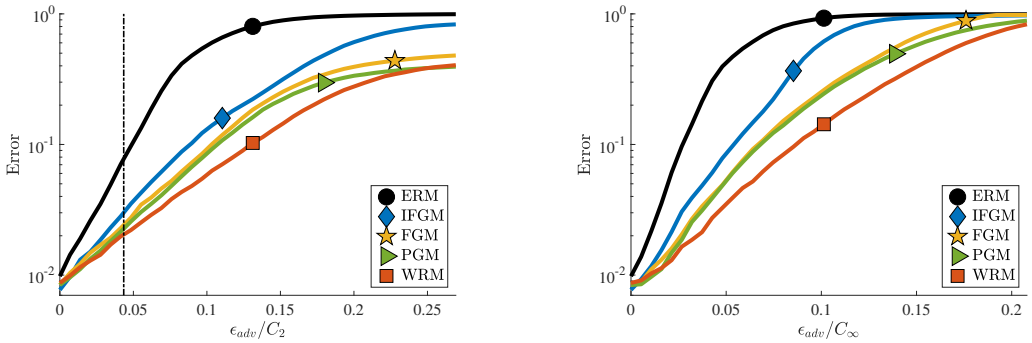
$$N(\mathcal{F}, \epsilon, \|\cdot\|_{L^\infty(\mathcal{Z})}) \leq N(\Theta, \epsilon/L, \|\cdot\|) \leq \left(1 + \frac{\operatorname{diam}(\Theta) L}{\epsilon}\right)^d,$$

where $\operatorname{diam}(\Theta) = \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\|$. Noting that $|\ell(\theta; Z)| \leq L \operatorname{diam}(\Theta) + M_0 =: M_\ell$, we have the result.

E Additional Experiments

E.1 MNIST attacks

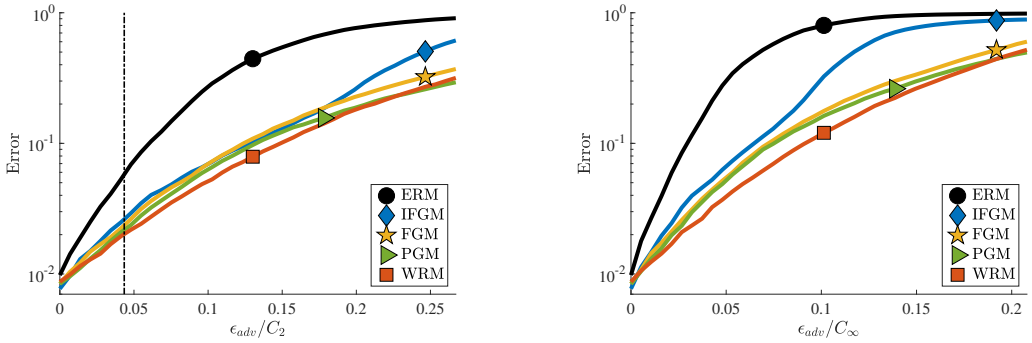
We repeat Figure 1(a) using FGM (Figure 3) and IFGM (Figure 4) attacks. The same trends are evident as in Figure 1(a). We also show PGM attacks again in Figure 2 for comparison. We scale γ in the figures with



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

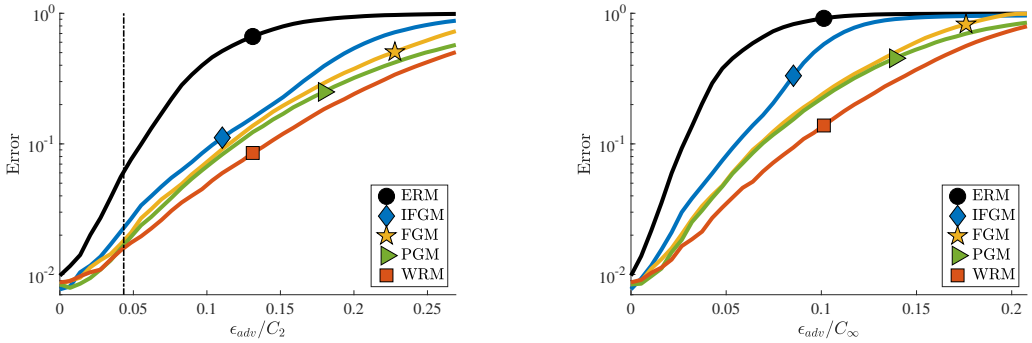
Figure 2. PGM attacks on the MNIST dataset. (a) and (b) show test misclassification error vs. the adversarial perturbation level ϵ_{adv} for the PGM attack with respect to Euclidean and ∞ norms respectively. The vertical bar in (a) indicates the perturbation level used for training the FGM, IFGM, and PGM models as well as the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$. For MNIST, $C_2 = 9.21$ and $C_\infty = 1.00$.

$C_p := \mathbb{E}_{\widehat{P}_n} \|X\|_p$. 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) Test error vs. ϵ_{adv} for $\|\cdot\|_2$ attack (b) Test error vs. ϵ_{adv} for $\|\cdot\|_\infty$ attack

Figure 3. Fast-gradient attacks on the MNIST dataset. (a) and (b) show test misclassification error vs. the adversarial perturbation level ϵ_{adv} for the FGM attack with respect to the Euclidean and ∞ norms respectively. The vertical bar in (a) indicates the perturbation level that was used for training the FGM, IFGM, and PGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$.



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

Figure 4. Iterated fast-gradient attacks on the MNIST dataset. (a) and (b) show test misclassification error vs. the adversarial perturbation level ϵ_{adv} for the IFGM attack with respect to the Euclidean and ∞ norms respectively. The vertical bar in (a) indicates the perturbation level that was used for training the FGM, IFGM, and PGM models and the estimated radius $\sqrt{\widehat{\rho}_n(\theta_{\text{WRM}})}$.

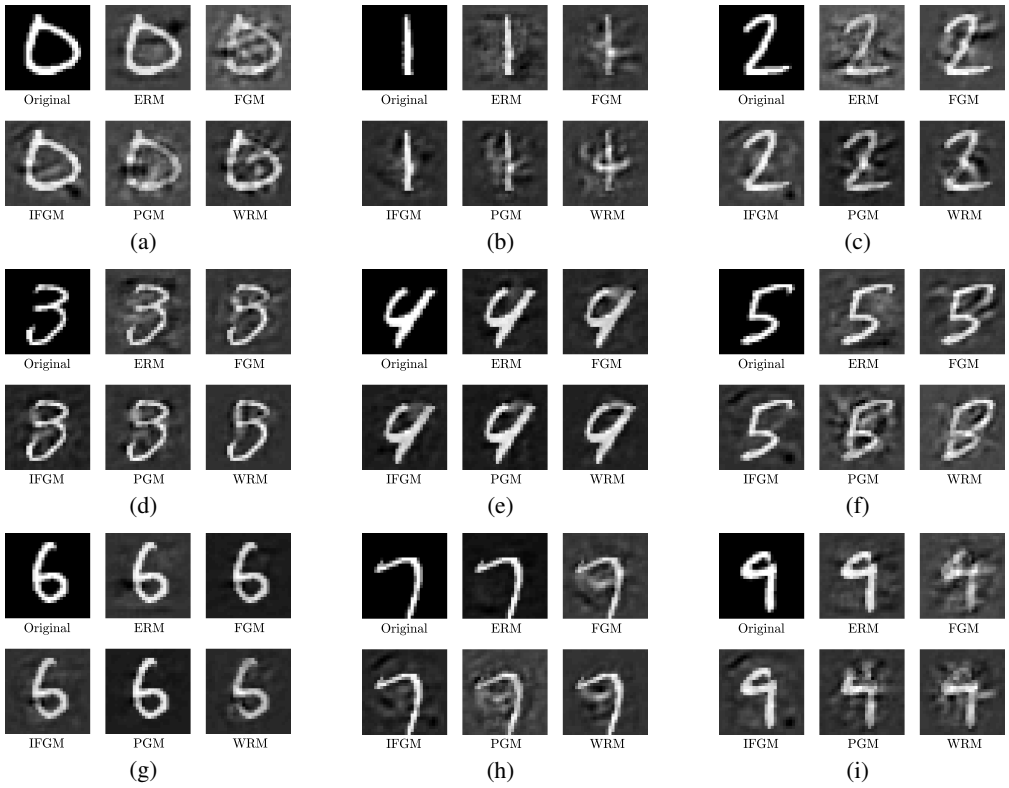


Figure 5. Visualizing stability over inputs. We illustrate the smallest WRM perturbation (largest γ_{adv}) necessary to make a model misclassify a datapoint.

E.2 MNIST stability of loss surface

In Figure 5, we repeat the illustration in Figure 1(c) 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.

E.3 MNIST Experiments with varied γ

In Figure 6, we choose a fixed WRM adversary (fixed γ_{adv}) and perturb WRM models trained with various penalty parameters γ. We note that as the bound (5) with $\eta = \gamma$ suggests, even when the adversary has more budget than that used for training ($1/\gamma < 1/\gamma_{\mathrm{adv}}$), degradation in performance is still *smooth*. Further, as we decrease the penalty γ, we see that the amount of achieved robustness—measured here by test error on adversarial perturbations with γ_{adv}—has diminishing gains; this is again consistent to our theory which says that the inner problem (2b) is not efficiently computable for small values of γ.

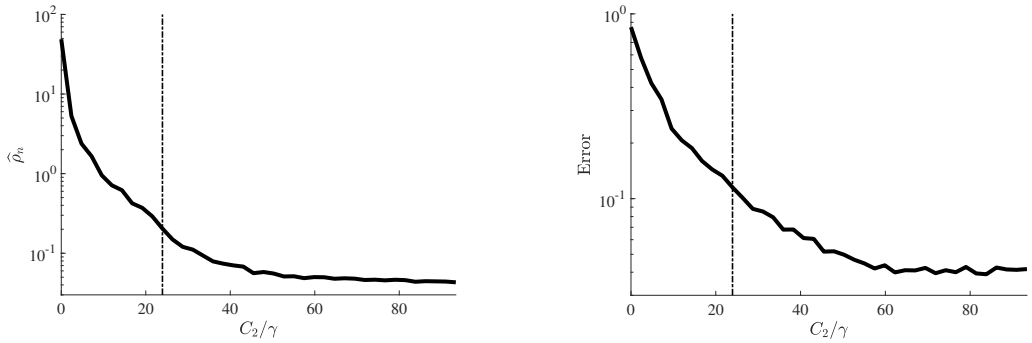
F Supervised Learning

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 B and D (Theorems 1 and 5) 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 $Y \in \{1, \ldots, K\}$. We consider an adversary that can only perturb the feature vector X [12], 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

$$c(z, z') := c_x(x, x') + \infty \cdot \mathbf{1}\{y \neq y'\}, \tag{20}$$

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 $c_x(x, x) = 0$.

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



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

Figure 6. (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 (γ_{adv} indicated by the vertical bar).

Under the cost function (20), the robust surrogate loss in the penalty problem (2) and its empirical counterpart (12) 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 B, 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*
$\|\nabla_\theta\ell(\theta;(x,y)) - \nabla_\theta\ell(\theta';(x,y))\|_* \le L_{\theta\theta}\|\theta-\theta'\|$, $\|\nabla_x\ell(\theta;(x,y)) - \nabla_x\ell(\theta;(x',y))\|_* \le L_{\mathsf{xx}}\|x-x'\|$,
$\|\nabla_\theta\ell(\theta;(x,y)) - \nabla_\theta\ell(\theta;(x',y))\|_* \le L_{\theta\mathsf{x}}\|x-x'\|$, $\|\nabla_x\ell(\theta;(x,y)) - \nabla_x\ell(\theta';(x,y))\|_* \le L_{\mathsf{x}\theta}\|\theta-\theta'\|$.

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 4. *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 $x^\star(\theta) = \operatorname{argmax}_{x\in\mathcal{X}} f(\theta,x)$, we have $\nabla\bar{f}(\theta) = \mathsf{g}_\theta(\theta,x^\star(\theta))$. Moreover,*

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

From Lemma 4, our previous results (Theorems 1 and 5) follow. The following is an analogue of Theorem 1 for the cost function (20).

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

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla F(\theta^t)\right\|_2^2\right] - \frac{2L_{\theta\mathsf{x}}^2}{\gamma - L_{\mathsf{xx}}}\epsilon \le \sigma\sqrt{8\frac{L\Delta_F}{T}}.$$

Similarly, an analogous result to Theorem 5 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 $\mathcal{Z}\subset\{z\in\mathbb{R}^m : \|z\|\le M_\mathsf{z}\}$ so that $\|Z\|\le M_\mathsf{z}$ 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 $1-e^{-t}$,*

$$\sup_{\theta\in\Theta}\left|\mathbb{E}_{\widehat{P}_n}[c(T_\gamma(\theta;Z),Z)] - \mathbb{E}_{P_0}[c(T_\gamma(\theta;Z),Z)]\right| \le 4D\sqrt{\frac{1}{n}\left(t + \log N\left(\Theta, \frac{[\gamma - L_{\mathsf{xx}}]_+ t}{4L_\mathsf{c}L_{\mathsf{x}\theta}}, \|\cdot\|\right)\right)}. \quad (21)$$

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 4 instead of Lemma 1.