Minimax Theory and its Applications
Volume 11 (2026), No. 3s, 1-14

A Minimax Optimization Framework for Robust Deep Learning under Adversarial Perturbations

Ali Sert*

School of Business and Management, University of Sunderland in Hong Kong, Hong Kong, China. E-mail: alisert2028@gmail.com

Abstract

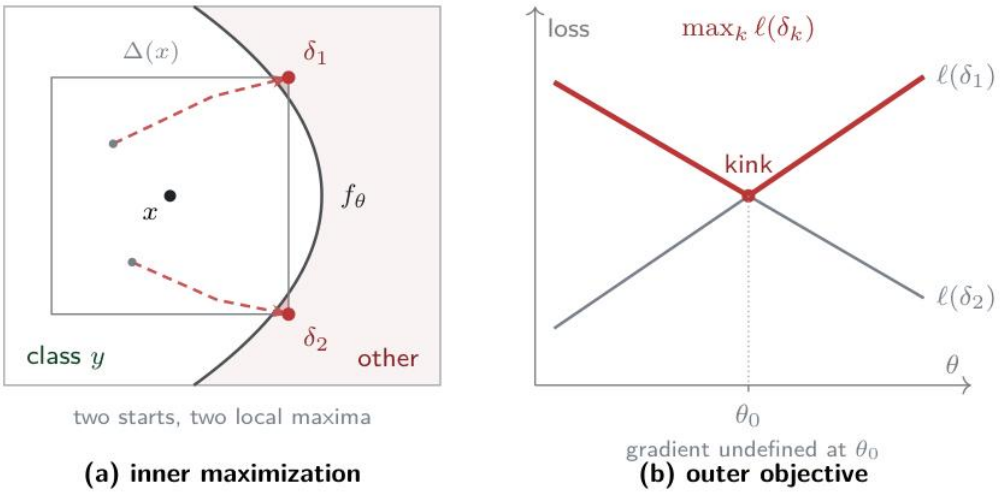
Adversarial training defends deep networks by solving a minimax problem. The learner minimizes the worst loss that an adversary can reach near each input. The worst-case loss has kinks in the network weights. We regularize the adversary with relative entropy and a temperature. The regularized inner problem then has a closed-form Gibbs solution over a set of candidate perturbations. Its value lies between the mean and the maximum of the candidate losses. Its distance below the best candidate is at most the temperature times the logarithm of the number of candidates. We prove that its gradient is Lipschitz with a constant that grows inversely with the temperature, and that this constant is attained. We prove that near-stationary points of the smoothed loss are near-stationary for the original min-max loss. In the zero-temperature limit they are Clarke stationary, and the temperature enters the iteration complexity through a logarithmic factor. We compare it with three defenses on MNIST and Fashion-MNIST at a matched budget of eight ascent steps. The method did not improve on projected gradient adversarial training. Worst-case accuracy was 58.5% for both on Fashion-MNIST, and 32.3% against 34.4% on MNIST, within the test-sampling error. The theory explains why the temperature had little effect: two restarts ended on average only 0.047 nats apart on MNIST. On trained networks, the measured curvature added by the adversary was three to four orders of magnitude below the bound. Splitting the budget into four two-step chains reduced MNIST worst-case accuracy to 5.6%.

***Keywords*:** minimax optimization; adversarial training; entropic regularization; Danskin's theorem; nonconvex stochastic optimization; robust deep learning.

1 Introduction

A deep classifier can change its prediction under an input perturbation too small for a human to see (Goodfellow et al., 2015; Szegedy et al., 2014). The standard defense, adversarial training, is a minimax problem (Madry et al., 2018). The learner minimizes over network weights the worst loss that an adversary can reach inside a norm ball around each input. In practice the inner maximum is approximated by a few steps of projected gradient ascent (PGD), often restarted from several random points.

This minimax problem lacks the structure that classical saddle-point theory needs. The inner loss is not concave in the perturbation, and the outer loss is not convex in the weights. Figure 1 shows two consequences. In panel (a), two PGD chains from different starts can reach two different local maxima. In panel (b), the maximum over such candidates has a kink in the weights, where the gradient is undefined. Danskin's theorem (Danskin, 1967) still gives a directional derivative, and weak convexity still yields subgradient rates (Davis & Drusvyatskiy, 2019), but the objective is not differentiable.



Local maxima and the kink of the worst-case loss

A classical remedy for such kinks is entropic smoothing (Beck & Teboulle, 2012; Nesterov, 2005). Meunier et al. (2021) applied entropic regularization to the adversarial-examples game with mixtures of classifiers. We apply it to a single network trained with PGD restarts. The adversary may randomize over perturbations but pays a temperature-weighted relative-entropy penalty for leaving a base measure. The regularized value is then a log-mean-exp of the candidate losses, between their mean and their maximum. We call the resulting algorithm smoothed minimax adversarial training (SMAT). We ask whether this smoothing helps when the number of ascent steps is fixed, and why. Our contributions are as follows.

- We formulate adversarial training with an entropy-regularized adversary and solve its inner problem in closed form for any base measure.
- We prove the value is within $\tau \log K$ of the best candidate, and a uniform adversary loses order $\tau d \log(1/\tau)$.
- We prove a $(\beta + G^2/\tau)$-Lipschitz gradient and transfer its near-stationary points to the min-max loss, with explicit complexity.
- We evaluate SMAT against three defenses: it ties PGD-AT on Fashion-MNIST (58.5%) and trails on MNIST (32.3% against 34.4%).
- We explain this: restarts end 0.047 nats apart, the bound is loose by three orders, and short chains collapse to 5.6%.

The first three items are proven results; the last two are empirical. The proofs use classical tools, and Section 4.6 states the limits of the small CPU-only study. Section 2 reviews related work, Section 3 develops the theory and protocol, Section 4 reports the results, and Section 5 concludes.

2 Related Work

This section places the framework within three lines of work. Section 2.1 covers attacks and adversarial training, Section 2.2 covers entropic objectives, and Section 2.3 covers minimax and smoothing theory.

2.1 Attacks and adversarial training

Small perturbations change the predictions of deep classifiers (Szegedy et al., 2014), and the fast gradient sign method exploits their local linearity (Goodfellow et al., 2015). Optimization attacks (Carlini & Wagner, 2017) and the analysis of obfuscated gradients (Athalye et al., 2018) showed that many defenses were weaker than reported. AutoAttack (Croce & Hein, 2020) and the gradient-free Square attack (Andriushchenko et al., 2020) are now standard checks, and we use both kinds. Madry et al. (2018) cast robust learning as the minimax problem (2). TRADES (H. Zhang et al., 2019) penalizes the boundary error with a Kullback–Leibler term, and MART (Y. Wang et al., 2020) up-weights misclassified examples. Fast and free training (Shafahi et al., 2019; Wong et al., 2020) cut the inner cost, weight perturbation (Wu et al., 2020) flattens the loss in weight space, and early stopping counters robust overfitting (Rice et al., 2020). Tramèr & Boneh (2019) and Maini et al. (2020) aggregate several threat models by the maximum or the average; our objective interpolates between these rules over restarts. Awasthi et al. (2023) build smooth surrogate losses, and randomized smoothing (Cohen et al., 2019) smooths the classifier; we smooth the adversary instead.

2.2 Entropic and distributional objectives

The closest work is Meunier et al. (2021). They prove that the adversarial-examples game with randomized classifiers and attackers has no duality gap. They also regularize the inner suprema entropically, obtain a log-sum-exp objective, and bound its error. Our Proposition 1 and Theorem 3 are of the same type. We differ in using one

network instead of a mixture, PGD restarts as the base measure, and in studying outer smoothness and stationarity rather than equilibria. Adversarial distributional training (Dong et al., 2020) learns a parametric perturbation distribution, and ATENT (Jagatap et al., 2022) samples from a high-loss Gibbs-like distribution. Sinkhorn distributionally robust optimization (J. Wang et al., 2026) regularizes a Wasserstein ambiguity set, and Sinha et al. (2018) use a Lagrangian Wasserstein penalty. Tilted risk (Li et al., 2021, 2023) applies the same log-mean-exp operator across training examples, and probabilistic robustness (Robey et al., 2022) interpolates between average and worst case differently.

2.3 Minimax theory and smoothing

The theorem of Sion (1958) needs quasi-convexity, quasi-concavity, and a compact convex set, and these fail for deep networks. Danskin's theorem (Danskin, 1967) gives the derivative of a pointwise maximum used by Madry et al. (2018), and Liu et al. (2020) showed that the adversarial loss loses second-order smoothness in the weights. Entropic smoothing of convex max-type functions (Beck & Teboulle, 2012; Nesterov, 2005) gives the known constants of our Theorems 3 and 6 for finitely many candidates. For nonconvex minimax problems, Lin et al. (2020), Nouiehed et al. (2019), and Jin et al. (2020) study gradient descent ascent and local optimality. J. Zhang et al. (2020) analyse smoothed descent for a finite maximum of nonconvex functions, and Davis & Drusvyatskiy (2019) give rates for weakly convex functions. Ghadimi & Lan (2013) give the stochastic rate that our Lemma 7 extends to a biased oracle. For adversarial training, Y. Wang et al. (2019) measure the quality of the inner maximum, and Gao et al. (2019) prove convergence for wide networks. We did not find a stationarity transfer like Theorem 9 for a regularized adversary over PGD candidates in these works.

3 Methodology

This section states the problem (Section 3.1), solves the regularized inner problem (Section 3.2), and bounds its error (Section 3.3). It then proves smoothness (Section 3.4) and a stationarity transfer (Section 3.5). Finally, it gives the algorithm (Section 3.6) and the experimental protocol (Section 3.7).

3.1 Problem setting

A classifier $f_\theta\colon [0,1]^d \to \mathbb{R}^C$ with parameters θ in a compact convex set $\Theta \subset \mathbb{R}^p$ maps an input of d pixels to C class scores. Let $\ell \geq 0$ denote the cross-entropy loss and $\mathcal{D}$ the data distribution over pairs (x, y). For radius $\varepsilon \in (0, 1/2]$, the admissible perturbations of x form the box

$$\Delta(x) = \{\delta \in \mathbb{R}^d \colon \| \delta \|_\infty \leq \varepsilon,\ x + \delta \in [0,1]^d\}.$$

We write $\ell(\theta, \delta)$ for $\ell(f_\theta(x + \delta), y)$ when (x, y) is fixed. Adversarial training (Madry et al., 2018) solves the minimax problem

$$\min_{\theta \in \Theta} R(\theta), \qquad R(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\Phi(\theta)], \qquad \Phi(\theta) = \max_{\delta \in \Delta(x)} \ell(\theta, \delta).$$

For a deep network, ℓ is in general neither quasi-concave in δ nor quasi-convex in θ. The hypotheses of the Sion (1958) theorem therefore fail, and we treat (2) as a min-max problem with the inner maximum taken first. The maximum Φ is continuous, and Danskin's theorem (Danskin, 1967) gives its directional derivative. When several maximizers have different θ-gradients, Φ has a kink. Under Assumption 1 below, Φ is β-weakly convex, so subgradient methods have Moreau-envelope rates (Davis & Drusvyatskiy, 2019). Smoothing offers a different trade: a differentiable surrogate with an explicit gradient-Lipschitz constant, at the price of a controlled bias.

3.2 Entropic regularization of the inner maximization

We lift the inner problem from points to probability measures. Let ν be a Borel probability measure on $\Delta(x)$, called the base measure. Let $\mathcal{P}_\nu$ be the set of probability measures q that are absolutely continuous with respect to ν. For a temperature $\tau > 0$, define the regularized inner value

$$\Phi_\tau^\nu(\theta) = \sup_{q \in \mathcal{P}_\nu} \{\mathbb{E}_{\delta\sim q}\, \ell(\theta, \delta) - \tau\, \mathrm{KL}(q \parallel \nu)\}.$$

Here $\mathrm{KL}(q \parallel \nu) = \mathbb{E}_q[\log(\mathrm{d}q/\mathrm{d}\nu)]$ is the relative entropy. The adversary may now randomize, and it pays τ per unit of relative entropy for moving away from ν. The maximization over q is a concave problem, because the objective is linear in q minus a convex penalty. The next proposition solves it.

Proposition 1 (Gibbs variational principle). Assume $\ell(\theta, \cdot)$ is bounded and measurable on $\Delta(x)$. Then the supremum in (3) is attained by the unique Gibbs measure $q_\theta^\star$ with density $\mathrm{d}q_\theta^\star/\mathrm{d}\nu = \exp(\ell(\theta, \cdot)/\tau)/Z_\tau(\theta)$, where $Z_\tau(\theta) = \mathbb{E}_\nu \exp(\ell(\theta, \delta)/\tau)$. Its value is

$$\Phi_\tau^\nu(\theta) = \tau \log \mathbb{E}_{\delta\sim\nu} \exp(\ell(\theta, \delta)/\tau).$$

Proof. Boundedness and $\ell \geq 0$ give $1 \leq Z_\tau(\theta) < \infty$, so $q_\theta^\star$ is a probability measure equivalent to ν. For $q \in \mathcal{P}_\nu$ with $\mathrm{KL}(q \parallel \nu) < \infty$, the chain rule for densities gives $\mathrm{KL}(q \parallel q_\theta^\star) = \mathrm{KL}(q \parallel \nu) - \tau^{-1}\mathbb{E}_q \ell(\theta, \delta) + \log Z_\tau(\theta)$. Multiplying by τ and rearranging, $\mathbb{E}_q \ell - \tau \mathrm{KL}(q \parallel \nu) = \tau \log Z_\tau(\theta) - \tau \mathrm{KL}(q \parallel q_\theta^\star)$. Relative entropy is nonnegative and vanishes only at $q = q_\theta^\star$. Measures with infinite relative entropy give the value $-\infty$. ▫

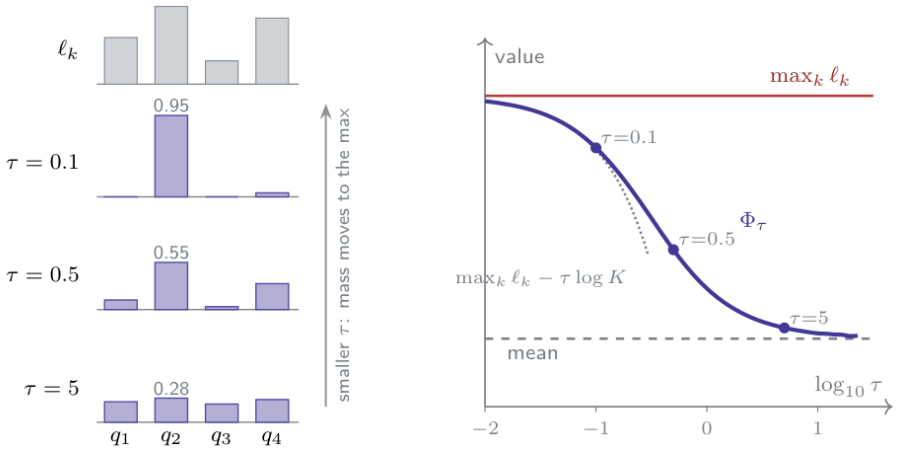
Proposition 1 is the Gibbs variational principle, also known as the Donsker–Varadhan variational formula (Donsker & Varadhan, 1975), applied to the adversary. Equation (4) is the tilted loss of Li et al. (2023), applied over

perturbations of one input instead of over training examples. The temperature interpolates between the average-case loss and the worst-case loss, as the next proposition makes exact.

Proposition 2 (Monotonicity and limits in τ). Under the assumption of Proposition 1, the map $\tau \mapsto \Phi_\tau^\nu(\theta)$ is nonincreasing on $(0,\infty)$, with derivative $-\mathrm{KL}(q_\theta^\star \parallel \nu)$. Moreover, $\Phi_\tau^\nu(\theta) \to \mathbb{E}_\nu \ell(\theta,\delta)$ as $\tau \to \infty$, and $\Phi_\tau^\nu(\theta) \to \operatorname{ess\,sup}_\nu \ell(\theta,\cdot)$ as $\tau \to 0$.

Proof. Write $\psi(\tau) = \tau \log Z_\tau$. Differentiation under the integral, allowed since ℓ is bounded, gives $\psi'(\tau) = \log Z_\tau - \tau^{-1}\mathbb{E}_{q^\star}\ell$. By the density formula, this equals $-\mathbb{E}_{q^\star}[\log(\mathrm{d}q^\star/\mathrm{d}\nu)] = -\mathrm{KL}(q^\star \parallel \nu) \le 0$. For $\tau \to \infty$, a first-order expansion gives $\log \mathbb{E}_\nu e^{\ell/\tau} = \tau^{-1}\mathbb{E}_\nu \ell + O(\tau^{-2})$. For $\tau \to 0$, set $p = 1/\tau$; then $\exp(\psi) = \| e^\ell \|_{L^p(\nu)}$, and the L^p norm tends to the essential supremum as $p \to \infty$. ▫

Figure 2 shows both effects on four fixed losses. In panel (a), lowering τ moves the adversary's mass onto the largest loss. In panel (b), the value Φ_τ falls from the maximum to the mean as τ grows, and it stays above the lower bound of Theorem 3 below.



(a) adversary's weights **(b) regularized value**

Gibbs weights and regularized value versus temperature

3.3 How far the regularized value is from the maximum

Regularization only underestimates the worst-case loss. We bound the underestimate for the empirical measure on K candidates, which the algorithm uses, and for the uniform measure on the box.

Theorem 3 (Finite candidate sandwich). Let $\nu_K = K^{-1}\sum_{k=1}^K \delta_{\delta_k}$ for points $\delta_1, \dots, \delta_K \in \Delta(x)$, and write $\ell_k = \ell(\theta, \delta_k)$. Then

$$\max\left\{1/K \sum_k \ell_k\,,\ \max_k \ell_k - \tau \log K\right\} \le \Phi_\tau^{\nu_K}(\theta) \le \max_k \ell_k.$$

Proof. The upper bound holds because each $e^{\ell_k/\tau} \le e^{\max_j \ell_j/\tau}$. Keeping only the largest term in the sum gives $\Phi_\tau^{\nu_K} \ge \tau \log(K^{-1} e^{\max_k \ell_k/\tau})$. Jensen's inequality for the convex exponential gives the mean lower bound. ▫

The bound (5) measures the gap against the best candidate, not against the true maximum $\Phi(\theta)$. How close $\max_k \ell_k$ is to $\Phi(\theta)$ is the usual question of attack strength, which no smoothing argument settles. The next theorem removes the candidates and pays a price in the dimension d.

Theorem 4 (Uniform base measure on the box). Let μ be the uniform probability measure on $\Delta(x)$. Assume $\delta \mapsto \ell(\theta,\delta)$ is L-Lipschitz on $\Delta(x)$ with respect to $\|\cdot\|_\infty$. Then $0 \le \Phi(\theta) - \Phi_\tau^\mu(\theta) \le L\varepsilon + \tau d \log 2$ for every $\tau > 0$. If in addition $\tau \le \varepsilon L/d$, then

$$0 \le \Phi(\theta) - \Phi_\tau^\mu(\theta) \le \tau d\left(1 + \log \frac{2\varepsilon L}{\tau d}\right).$$

Proof. Coordinate i of $\Delta(x)$ is an interval $[a_i, b_i]$ of width $w_i = b_i - a_i$. Since $x_i \in [0,1]$ and $\varepsilon \le 1/2$, we have $\varepsilon \le w_i \le 2\varepsilon$. Let $\delta^\star$ maximize the continuous function $\ell(\theta,\cdot)$ on the compact box. Fix $r \in (0,\varepsilon]$ and let $S_r = \{\delta \in \Delta(x) : \| \delta - \delta^\star \|_\infty \le r\}$. In each coordinate, $[\delta_i^\star - r, \delta_i^\star + r] \cap [a_i, b_i]$ has length at least r, because $r \le w_i$. Hence $\mu(S_r) \ge \prod_i r / w_i \ge (r/2\varepsilon)^d$. On S_r the Lipschitz property gives $\ell(\theta,\delta) \ge \Phi(\theta) - Lr$. Keeping only S_r in the expectation, $\Phi_\tau^\mu \ge \tau \log\big(\mu(S_r)\, e^{(\Phi - Lr)/\tau}\big) \ge \Phi - Lr - \tau d \log(2\varepsilon/r)$. The choice $r = \varepsilon$ gives the first bound. The choice $r = \tau d/L$, admissible when $\tau \le \varepsilon L/d$, gives (6). The lower bound 0 follows from Proposition 2. ▫

The factor d in (6) is not an artefact of the proof. The next proposition computes the gap exactly for a loss that is linear in the perturbation.

Proposition 5 (Matching lower bound). Let $\ell(\theta,\delta) = c + \langle g,\delta\rangle$ on $\Delta(x)$, with $c \geq \varepsilon \parallel g \parallel_1$ so that $\ell \geq 0$, and let μ be uniform on $\Delta(x)$. Write $w_i \in [\varepsilon, 2\varepsilon]$ for the side lengths of the box and $\psi(s) = \log(s/(1-e^{-s}))$ for $s > 0$. Then

$$\Phi(\theta) - \Phi_\tau^\mu(\theta) = \tau \sum_{i:\, g_i \neq 0} \psi\left(|g_i| w_i/\tau\right) \;\geq\; \tau \sum_{i:\, g_i \neq 0} \log \frac{|g_i| w_i}{\tau}.$$

If $|g_i| = L/d$ for every i, the loss is L-Lipschitz with respect to $\|\cdot\|_\infty$, and for $\tau \leq \varepsilon L/d$ the gap is at least $\tau d \log(\varepsilon L/(\tau d))$. This differs from the upper bound (6) by at most $\tau d(1+\log 2)$.

Proof. Under μ the coordinates of δ are independent and uniform on $[a_i, b_i]$, and the exponent separates over coordinates. Hence $\Phi_\tau^\mu = c + \sum_i \tau \log\left(w_i^{-1} \int_{a_i}^{b_i} e^{g_i t/\tau}\, \mathrm{d}t\right)$. Take $g_i > 0$; the case $g_i < 0$ is symmetric with a_i in place of b_i, and $g_i = 0$ contributes nothing. The integral equals $\frac{\tau}{g_i} e^{g_i b_i/\tau}\left(1 - e^{-g_i w_i/\tau}\right)$, so the coordinate term is $g_i b_i - \tau\psi(g_i w_i/\tau)$. The maximum of ℓ is $c + \sum_i \max(g_i a_i, g_i b_i)$, which gives the equality in (7). The inequality holds because $1 - e^{-s} < 1$. With $|g_i| = L/d$ and $w_i \geq \varepsilon$, each term is at least $\tau \log(\varepsilon L/(\tau d))$. Subtracting this from (6) leaves $\tau d(1+\log 2)$. ▫

Remark 1 (The dimension factor). Theorem 4 and Proposition 5 together show that the gap of a uniform adversary is of order $\tau d \log(1/\tau)$, with the factor d attained. The constant $L = \sup \parallel \nabla_\delta \ell \parallel_1$ may itself grow with d. For MNIST, $d = 784$ and $\varepsilon = 0.3$; if L is of order one, a gap of 0.1 nats requires τ of order 10^{-5}. At such temperatures the smoothness bound of Theorem 6 is enormous. For this reason the algorithm does not use μ. It uses a base measure concentrated near high-loss points, namely the empirical measure on projected gradient ascent endpoints. Proposition 1 and Theorem 3 hold for any base measure, so they cover this choice.

3.4 Differentiability and smoothness of the outer objective

We now study $\theta \mapsto \Phi_\tau^\nu(\theta)$ for a fixed base measure ν. The following assumption is the only regularity we need.

Assumption 1. There is an open convex set $U \supset \Theta$ such that the following hold. The map $(\theta,\delta) \mapsto \ell(\theta,\delta)$ is jointly measurable and bounded on $U \times \Delta(x)$. For ν-almost every δ, the map $\theta \mapsto \ell(\theta,\delta)$ is continuously differentiable on U. There are constants G, β such that, for all $\theta, \theta' \in U$ and ν-almost every δ, $\parallel \nabla_\theta \ell(\theta,\delta) \parallel \leq G$ and $\parallel \nabla_\theta \ell(\theta,\delta) - \nabla_\theta \ell(\theta',\delta) \parallel \leq \beta \parallel \theta - \theta' \parallel$. When we average over data, these constants hold uniformly in (x,y).

A network with smooth activations, such as Softplus, and average pooling is infinitely differentiable in θ and in its input. On a bounded open set U it therefore satisfies Assumption 1 with finite constants. Rectified linear units and max pooling violate it. The constants grow with the size of U, and they do not hold on all of $\mathbb{R}^p$, because gradients grow with the weight norms. We do not prove that the iterates of an optimizer stay in a given bounded set.

Theorem 6 (Smoothness of the regularized robust loss). Under Assumption 1, Φ_τ^ν is differentiable on Θ with

$$\nabla \Phi_\tau^\nu(\theta) = \mathbb{E}_{\delta \sim q_\theta^\star}[\nabla_\theta \ell(\theta,\delta)],$$

and its gradient is Lipschitz with constant $\beta + G^2/\tau$.

Proof. The integrand $e^{\ell/\tau}$ is differentiable in θ, and its gradient $\tau^{-1} e^{\ell/\tau} \nabla_\theta \ell$ is bounded uniformly in δ. Dominated convergence therefore allows differentiation under the integral. The identity $\nabla(\tau \log Z_\tau) = \tau \nabla Z_\tau / Z_\tau$ then gives (8). Fix $\theta, \theta' \in \Theta$. Let $u = \theta - \theta'$, and let $\theta_s = \theta' + su$ for $s \in [0,1]$. Write $w_s = \mathrm{d}q_{\theta_s}^\star/\mathrm{d}\nu$ for the Gibbs density. Write $X_s(\delta) = \langle \nabla_\theta \ell(\theta_s,\delta), u\rangle$ for the directional derivative. The chain rule and (8) give the following derivative.

$$\partial_s w_s = \tau^{-1} w_s \left(X_s - \mathbb{E}_{q_{\theta_s}^\star} X_s\right).$$

Integrate $|\partial_s w_s|$ over $s \in [0,1]$, exchange the integrals by Fubini's theorem (joint measurability), and use $\mathbb{E}|Z| \leq \sqrt{\mathbb{E}Z^2}$ for the centred variable. Since $|X_s| \leq G \parallel u \parallel$, this gives the bound below.

$$\int |w_1 - w_0|\, \mathrm{d}\nu \leq \tau^{-1} \sup_s \sqrt{\mathrm{Var}_{q_{\theta_s}^\star} X_s} \leq G \parallel u \parallel/\tau.$$

Now split the gradient difference as follows.

$$\begin{aligned} \nabla \Phi_\tau^\nu(\theta) - \nabla \Phi_\tau^\nu(\theta') = &\int w_1 (\nabla_\theta \ell(\theta,\cdot) - \nabla_\theta \ell(\theta',\cdot))\, \mathrm{d}\nu \\ &+ \int (w_1 - w_0)\, \nabla_\theta \ell(\theta',\cdot)\, \mathrm{d}\nu. \end{aligned}$$

The first term has norm at most $\beta \parallel u \parallel$, since w_1 is a probability density. The second has norm at most $G \int |w_1 - w_0| \mathrm{d}\nu \leq G^2 \parallel u \parallel/\tau$. ▫

The proof shows where the term G^2/τ comes from. It is the speed at which the adversary's Gibbs weights move as θ moves. The next example shows that the order G^2/τ cannot be improved in general.

Example 1 (Tightness). Take $\Theta = [-1,1]$, ν uniform on two points, and $\ell(\theta,\delta_1) = \theta$, $\ell(\theta,\delta_2) = -\theta$. Then $G = 1$, $\beta = 0$, and the maximum is $|\theta|$, which is not differentiable at 0. The regularized value is $\Phi_\tau(\theta) = \tau \log\cosh(\theta/\tau)$. Its second derivative at 0 equals $1/\tau = \beta + G^2/\tau$.

As $\tau \to 0$ with finitely many candidates and a unique best candidate $k^\star$, the Gibbs weights in (8) converge to the indicator of $k^\star$. The gradient (8) then tends to $\nabla_\theta \ell(\theta, \delta_{k^\star})$, which is the Danskin gradient used by projected gradient descent adversarial training. Smoothing thus replaces a selection of one candidate by a softmax-weighted average of candidates.

3.5 Stochastic gradient guarantee

Let $F_\tau(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Phi_\tau^{\nu_x}(\theta)$, where each input has its own fixed base measure ν_x. By Theorem 6 and linearity of expectation, F_τ has a β_τ-Lipschitz gradient with $\beta_\tau = \beta + G^2/\tau$. We analyse the update $\theta_{t+1} = \theta_t - \eta g_t$ with a possibly biased stochastic gradient g_t. For simplicity we state the unconstrained case, with the constants of Assumption 1 holding on all of $\mathbb{R}^p$. As noted above, this global assumption does not hold for the networks we train. The lemma is a conditional statement: it applies to any run whose iterates stay in a region where the constants hold. A projected method on a compact Θ would remove this condition at the cost of a gradient-mapping criterion; we do not develop it.

Lemma 7 (Biased stochastic gradient descent). Let F have a β_F-Lipschitz gradient and $F^\star = \inf F > -\infty$. Let $\mathcal{F}_t$ be the history up to iteration t. Assume $\mathbb{E}[g_t \mid \mathcal{F}_t] = \nabla F(\theta_t) + b_t$ with $\| b_t \| \le b$, and $\mathbb{E}[\| g_t - \mathbb{E}[g_t \mid \mathcal{F}_t] \|^2 \mid \mathcal{F}_t] \le \sigma^2$. If $\eta \le 1/\beta_F$, then

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E} \| \nabla F(\theta_t) \|^2 \le \frac{2(F(\theta_0) - F^\star)}{\eta T} + \beta_F \eta \sigma^2 + b^2.$$

Proof. The descent lemma for a Lipschitz gradient gives $F(\theta_{t+1}) \le F(\theta_t) - \eta\langle \nabla F, g_t\rangle + \beta_F \eta^2/2 \| g_t \|^2$. Take the conditional expectation and write $m_t = \nabla F(\theta_t) + b_t$. Then $\mathbb{E} \| g_t \|^2 \le \| m_t \|^2 + \sigma^2$, and $\beta_F \eta^2/2 \le \eta/2$. Hence $\mathbb{E}F(\theta_{t+1}) \le F(\theta_t) - \eta \| \nabla F \|^2 - \eta\langle \nabla F, b_t\rangle + \eta/2 \| m_t \|^2 + \beta_F \eta^2 \sigma^2/2$. Expanding $\| m_t \|^2$ cancels the cross term and leaves $-\eta/2 \| \nabla F \|^2 + \eta/2 \| b_t \|^2 + \beta_F \eta^2 \sigma^2/2$. Summing over t, telescoping, and dividing by $\eta T/2$ gives (9). ▫

Lemma 7 is the bias-aware form of the analysis of Ghadimi & Lan (2013). Combining it with Theorem 6 gives the temperature trade-off.

Corollary 8 (Temperature trade-off). Let $D_0 = F_\tau(\theta_0) - \inf F_\tau$ and choose the step size

$$\eta = \min\left\{\frac{1}{\beta_\tau}, \sqrt{\frac{2D_0}{\beta_\tau \sigma^2 T}}\right\}.$$

Then

$$\min_{t<T} \mathbb{E} \| \nabla F_\tau(\theta_t) \|^2 \le \frac{2\beta_\tau D_0}{T} + 2\sqrt{\frac{2\beta_\tau D_0 \sigma^2}{T}} + b^2, \qquad \beta_\tau = \beta + \frac{G^2}{\tau}.$$

With K candidates per input, every θ also satisfies $F_\tau(\theta) \le F_K(\theta) \le F_\tau(\theta) + \tau \log K$, where $F_K(\theta) = \mathbb{E}\max_k \ell(\theta, \delta_k)$. Hence any θ that is γ-suboptimal for F_τ is $(\gamma + \tau \log K)$-suboptimal for F_K.

Proof. The rate follows from Lemma 7 with $\beta_F = \beta_\tau$; the two cases of the minimum each give at most the stated sum. The sandwich follows by taking expectations in Theorem 3. For the transfer, $F_K(\theta) \le F_\tau(\theta) + \tau \log K \le \inf F_\tau + \gamma + \tau \log K \le \inf F_K + \gamma + \tau \log K$. ▫

A small τ makes F_τ close to F_K in value but inflates the smoothness bound. A small gradient of F_τ is useful only if it certifies near-stationarity of the nonsmooth min-max objective F_K; the next theorem supplies this link.

We state it for a finite training set $\{(x_i, y_i)\}_{i=1}^n$ with fixed candidates δ_{ik}, and write $\ell_{ik}(\theta) = \ell(f_\theta(x_i + \delta_{ik}), y_i)$. Then $F_K(\theta) = \frac{1}{n}\sum_i \max_k \ell_{ik}(\theta)$ and $F_\tau(\theta) = \frac{1}{n}\sum_i \tau \log \frac{1}{K}\sum_k e^{\ell_{ik}(\theta)/\tau}$. For $\rho \ge 0$, the ρ-active set of example i is $A_i^\rho(\theta) = \{k: \ell_{ik}(\theta) \ge \max_m \ell_{im}(\theta) - \rho\}$. We define the set of ρ-active gradient averages

$$\mathcal{G}_\rho(\theta) = \frac{1}{n}\sum_{i=1}^n \operatorname{conv}\{\nabla \ell_{ik}(\theta): k \in A_i^\rho(\theta)\}.$$

When every ℓ_{ik} is continuously differentiable, $\mathcal{G}_0(\theta)$ is the Clarke subdifferential $\partial_C F_K(\theta)$. A pointwise maximum of finitely many continuously differentiable functions is Clarke regular, and its subdifferential is the convex hull of the active gradients. For regular functions the sum rule holds with equality (Clarke, 1990, Prop. 2.3.12 and Cor. 3 to Prop. 2.3.3). For $\rho > 0$, $\mathcal{G}_\rho$ enlarges each hull to the nearly active candidates.

Theorem 9 (Stationarity transfer). Let $K \ge 2$. Assume each ℓ_{ik} is continuously differentiable on U, with gradient norm at most G.

1. For every $\theta \in U$, $\tau > 0$, and $\rho > 0$,
$$\operatorname{dist}\big(0, \mathcal{G}_\rho(\theta)\big) \le \| \nabla F_\tau(\theta) \| + 2G(K-1)\, e^{-\rho/\tau}.$$
2. If $\tau_j \downarrow 0$, $\theta_j \to \bar\theta \in U$, and $\nabla F_{\tau_j}(\theta_j) \to 0$, then $0 \in \partial_C F_K(\bar\theta)$.

Proof. For (a), write q_{ik} for the Gibbs weights, so that $\nabla F_\tau(\theta) = \frac{1}{n}\sum_i \sum_k q_{ik} \nabla \ell_{ik}(\theta)$. Dividing numerator and denominator of q_{ik} by the largest term shows $q_{ik} < e^{-\rho/\tau}$ for every $k \notin A_i^\rho$. Hence the off-active mass $m_i = \sum_{k \notin A_i^\rho} q_{ik}$ satisfies $m_i \le (K-1)e^{-\rho/\tau}$. Split the weighted gradient of example i as $(1 - m_i)a_i + m_i c_i$. Here a_i is the normalized average over A_i^ρ, so a_i lies in its hull, and c_i is the normalized average over the rest. Both have norm at

most G, so the weighted gradient differs from a_i by m_i $\| c_i - a_i \| \le 2Gm_i$. The point $\frac{1}{n}\sum_i a_i$ lies in $\mathcal{G}_\rho(\theta)$, and its norm is at most $\| \nabla F_\tau(\theta) \| + \frac{1}{n}\sum_i 2\,Gm_i$.

For (b), fix i and $k \notin A_i^0(\bar\theta)$, and let $\gamma > 0$ be its gap below the maximum at $\bar\theta$. By continuity, the gap exceeds $\gamma/2$ at θ_j for large j, so $q_{ik}(\theta_j, \tau_j) < e^{-\gamma/(2\tau_j)} \to 0$. The weight vectors lie in a compact simplex. Along a subsequence they converge to some λ_i supported on $A_i^0(\bar\theta)$. By continuity of the gradients, $\nabla F_{\tau_j}(\theta_j) \to \frac{1}{n}\sum_i \sum_{k \in A_i^0(\bar\theta)} \lambda_{ik} \nabla \ell_{ik}(\bar\theta)$. This limit is 0 and lies in $\mathcal{G}_0(\bar\theta) = \partial_C F_K(\bar\theta)$. ▫

Theorem 9 turns the temperature trade-off into a single complexity statement.

Corollary 10 (Complexity of approximate stationarity). In the setting of Theorem 9, suppose additionally that the constants of Assumption 1 hold globally and that minibatch gradients of F_τ are unbiased, so $b = 0$. Fix $\zeta \in (0, 2G(K-1))$ and $\rho > 0$, and set $\tau = \rho / \log(2G(K-1)/\zeta)$. Run the method of Corollary 8 for

$$T \;\ge\; \max\left\{ \frac{4\beta_\tau D_0}{\zeta^2},\; \frac{32\beta_\tau D_0 \sigma^2}{\zeta^4} \right\} \quad \text{iterations, where} \quad \beta_\tau = \beta + \frac{G^2}{\rho} \log \frac{2G(K-1)}{\zeta}.$$

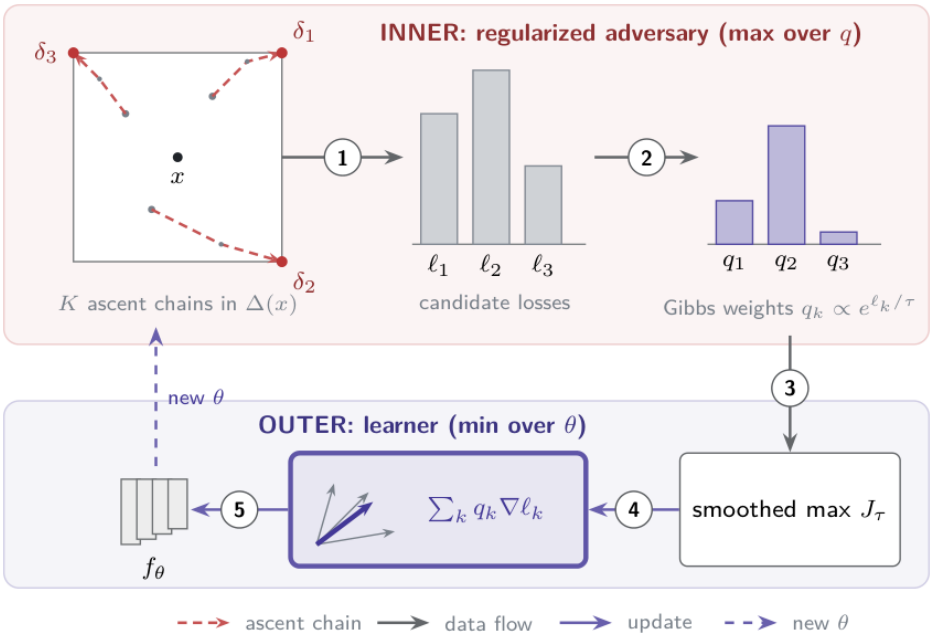
Then some iterate θ_t with $t < T$ satisfies $\mathbb{E}\operatorname{dist}\big(0, \mathcal{G}_\rho(\theta_t)\big) \le 2\zeta$.

Proof. The two lower bounds on T make each term of (10) at most $\zeta^2/2$. Hence $\mathbb{E} \| \nabla F_\tau(\theta_t) \|^2 \le \zeta^2$ for some t, and Jensen's inequality gives $\mathbb{E} \| \nabla F_\tau(\theta_t) \| \le \zeta$. The choice of τ makes $2G(K-1)e^{-\rho/\tau} = \zeta$, so (12) gives the claim after taking expectations. ▫

Corollary 10 needs $O((\beta + G^2 \log(GK/\zeta)/\rho)\, D_0 \sigma^2 / \zeta^4)$ iterations for $(\rho, 2\zeta)$-stationarity of the min-max objective F_K. The temperature enters only through the ratio $\rho / \log(GK/\zeta)$. The price of smoothing is a logarithmic factor, a factor $1/\rho$, and a stationarity notion in which candidates within ρ of the maximum also count. Subgradient methods on the weakly convex F_K reach a ζ-small Moreau-envelope gradient in $O(\zeta^{-4})$ iterations (Davis & Drusvyatskiy, 2019); the two measures differ, so neither rate dominates. With fixed candidates the minibatch gradient is unbiased ($b = 0$); if the candidates are samples from the law of PGD endpoints, the log-mean-exp is biased downward, and we do not bound that bias.

3.6 Algorithm and the scope of the theory

Algorithm 1 (SMAT) runs K PGD chains of s steps from uniform random starts for each example. It minimizes the regularized value (4) over the empirical measure on the endpoints, held fixed, so the gradient is the Gibbs-weighted average (8). With $K = 1$ it reduces to PGD adversarial training (Madry et al., 2018), and we fix the ascent budget at $Ks = 8$ gradient evaluations per example. Figure 3 shows one iteration.



One iteration of smoothed minimax adversarial training

Algorithm 1: Smoothed minimax adversarial training

Input: data, radius ε, candidates K, steps s, step size $\alpha = 2.5\varepsilon/s$, temperature τ.

1. For each minibatch $\{(x_i, y_i)\}_{i=1}^B$:
2. for $i = 1, \dots, B$ and $k = 1, \dots, K$: draw δ_{ik} uniformly on $[-\varepsilon, \varepsilon]^d$ and project it onto $\Delta(x_i)$;
3. repeat s times: $\delta_{ik} \leftarrow \Pi_{\Delta(x_i)}(\delta_{ik} + \alpha \operatorname{sign} \nabla_\delta \ell(\theta, \delta_{ik}))$;

4. $J(\theta) \leftarrow B^{-1}\sum_i \tau \log(K^{-1}\sum_k \exp(\ell(\theta, \delta_{ik})/\tau))$ with each δ_{ik} held fixed;
5. $\theta \leftarrow$ optimizer step on $\nabla_\theta J(\theta)$.

The theory covers Algorithm 1 only in part. The results of Sections 3.2 to 3.4 hold exactly for the per-step objective J, whose candidates are fixed during the step. The rate results need a fixed objective, while the algorithm regenerates its candidates at each iterate; PGD adversarial training has the same gap. The next proposition shows when holding the candidates fixed loses nothing.

Proposition 11 (Envelope property of fixed candidates). Suppose that near θ_0 each candidate is a differentiable map $\theta \mapsto \delta_k(\theta)$. Suppose also that each $\delta_k(\theta)$ is a local maximizer of $\ell(\theta,\cdot)$ on $\Delta(x)$ with strict complementarity and a locally constant set of active box constraints. Then the total derivative of $\theta \mapsto \tau\log K^{-1}\sum_k \exp(\ell(\theta, \delta_k(\theta))/\tau)$ at θ_0 equals the fixed-candidate gradient (8) over ν_K.

Proof. The chain rule gives the extra term $(\partial_\theta \delta_k)^\top \nabla_\delta \ell(\theta, \delta_k)$ for each k. Let A be the set of coordinates at a bound. At a local maximizer on a box, $\partial_{\delta_i}\ell = 0$ for $i \notin A$. For $i \in A$ the coordinate stays at its bound near θ_0, because the active set is locally constant, so $\partial_\theta \delta_{k,i} = 0$. Every product $\partial_\theta \delta_{k,i}\ \partial_{\delta_i}\ell$ therefore vanishes. ▫

If ℓ is twice continuously differentiable and δ_k satisfies second-order sufficient conditions with strict complementarity, the sensitivity theory of Fiacco (1983) makes $\delta_k(\theta)$ differentiable; linear independence holds automatically for boxes. A four-step PGD endpoint is not an exact maximizer, and we do not bound the resulting error. Our runs also use Adam on an unconstrained network, so they test the objective, not the convergence statements.

3.7 Experimental protocol

The experiments ask whether smoothing changes robustness at a fixed ascent budget, and how tight the smoothness bound is on trained networks.

3.7.0.1 Data and model.

We use MNIST (LeCun et al., 1998) with $\varepsilon = 0.3$ and Fashion-MNIST with $\varepsilon = 0.1$ (28×28 grey-scale images, ten classes, $d = 784$). All runs used PyTorch (Paszke et al., 2019) on one laptop processor with two threads and no graphics card. We train on a fixed subset of 20,000 training images (seed 1234) and evaluate on the first 500 test images, where the binomial standard error is about 2 points. The network has two 5×5 convolutions (16 and 32 channels) with 2×2 average pooling, a 128-unit dense layer, and Softplus activations of sharpness 10. Every method trains for 6 epochs with Adam (learning rate 10^{-3}, cosine decay, batch 128), and the radius grows linearly over the first third of training.

3.7.0.2 Methods and attacks.

We compare standard training, PGD-AT (Madry et al., 2018), TRADES and MART with trade-off weight 6 (Y. Wang et al., 2020; H. Zhang et al., 2019), and SMAT. Every adversarial method spends 8 input-gradient steps per example: one 8-step chain, or for SMAT $K = 2$ chains of 4 steps with $\tau = 0.5$, at step size 2.5ε divided by the chain length. The ablations use $\tau \in \{0.01, 0.1, 0.5, 2, 100\}$ at $K = 2$ and $K = 4$ chains of 2 steps, with step 1.25ε or capped at ε. We attack with FGSM, PGD-20, and 100-iteration APGD-CE and targeted APGD-T with three targets (Croce & Hein, 2020), and report the per-example worst case. To detect gradient masking (Athalye et al., 2018), we add the Square attack (Andriushchenko et al., 2020) with 1,000 queries on the first 200 test images. The main comparison and the $K = 4$ runs use seeds 0, 1, and 2 (mean and sample standard deviation); the temperature ablation uses seed 0.

3.7.0.3 Smoothness check.

For fixed candidates, the adversary adds $\tau^{-1}C_\tau$ to the Hessian of J, where C_τ is the Gibbs covariance of the candidate gradients; the proof bounds its effect by G^2/τ. On the seed-0 PGD-AT and SMAT models, with 16 frozen four-step candidates on 8 test images, we compute $\lambda_{\max}(C_\tau)$ exactly and estimate G from below. We measure $\| \nabla J(\theta + hu) - \nabla J(\theta) \|/h$ with $h = 10^{-3}$ along the top eigenvector u of C_τ, and along three random directions on 64 images.

4 Results and Discussion

This section compares the methods (Section 4.1), explains the role of the temperature (Section 4.2) and of chain length (Section 4.3), and tests the smoothness bound (Section 4.4) and gradient masking (Section 4.5). Section 4.6 states the limitations.

4.1 Comparison under a matched ascent budget

Table 1 gives clean and robust accuracy over three seeds; Square uses the first 200 test images and every other column the first 500. Standard training reaches 98.0% clean accuracy on MNIST and 88.1% on Fashion-MNIST, with worst-case accuracy 0.0% and 9.4%.

Clean and robust test accuracy (%), three seeds

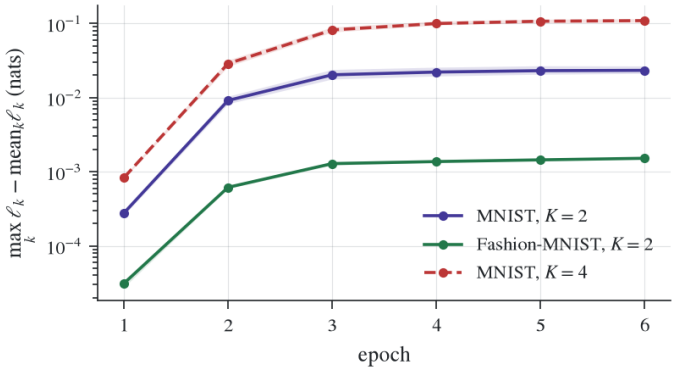
Method	Clean	FGSM	PGD-20	APGD-CE	APGD-T	Worst	Square

Method	Clean	FGSM	PGD-20	APGD-CE	APGD-T	Worst	Square
MNIST, $\varepsilon = 0.3$							
Standard	98.0 ± 0.2	0.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
PGD-AT	93.9 ± 0.3	66.9 ± 2.7	44.4 ± 0.9	41.7 ± 1.1	35.1 ± 1.2	34.4 ± 1.1	41.3 ± 1.0
TRADES	93.1 ± 0.1	66.0 ± 1.6	42.9 ± 0.7	41.3 ± 0.5	29.5 ± 0.4	29.5 ± 0.4	32.5 ± 3.5
MART	87.3 ± 2.7	67.5 ± 0.7	50.5 ± 0.1	45.3 ± 1.9	29.3 ± 1.7	29.3 ± 1.7	35.0 ± 3.1
SMAT ($\tau = 0.5$)	94.0 ± 0.4	69.5 ± 2.4	42.9 ± 0.2	40.1 ± 1.8	32.9 ± 1.4	32.3 ± 1.3	39.2 ± 1.4
Fashion-MNIST, $\varepsilon = 0.1$							
Standard	88.1 ± 0.2	11.1 ± 1.0	9.7 ± 0.9	9.4 ± 0.9	9.4 ± 0.9	9.4 ± 0.9	15.0 ± 0.9
PGD-AT	79.1 ± 0.6	64.5 ± 0.7	62.5 ± 1.3	61.4 ± 1.4	58.6 ± 1.0	58.5 ± 1.0	60.8 ± 0.8
TRADES	79.9 ± 0.5	62.1 ± 0.6	60.7 ± 0.9	60.4 ± 0.9	57.4 ± 0.7	57.4 ± 0.7	58.2 ± 1.0
MART	80.4 ± 0.4	67.5 ± 0.9	65.7 ± 1.2	64.3 ± 1.6	59.9 ± 1.5	59.7 ± 1.7	62.3 ± 1.0
SMAT ($\tau = 0.5$)	79.3 ± 0.7	64.4 ± 1.1	62.5 ± 1.1	61.5 ± 1.1	58.8 ± 0.5	58.5 ± 0.6	61.3 ± 1.3

SMAT does not improve on PGD-AT. On MNIST its worst-case accuracy is 32.3 ± 1.3%, against 34.4 ± 1.1%; the 2.1-point difference equals the binomial standard error on 500 images. On Fashion-MNIST both reach 58.5%, with similar clean accuracy (79.3% against 79.1%). MART has the highest mean worst case on Fashion-MNIST, 59.7 ± 1.7%, within the noise of PGD-AT and SMAT. The attack columns show why the worst case must be reported. MART reaches the highest PGD-20 accuracy on MNIST, 50.5%, but only 29.3% under targeted APGD, a gap of 21.2 points against 9.3 for PGD-AT and 9.9 for SMAT. A plausible cause is that the MART loss flattens the cross-entropy that PGD ascends while leaving small margins that the targeted margin loss finds. Under the worst case, MART ties TRADES (29.3% against 29.5%) below PGD-AT and SMAT.

4.2 Why the temperature barely matters here

Theorem 3 confines the regularized value between the mean and the maximum of the candidate losses, so the temperature can change the objective by at most their spread. Figure 4 plots this spread during training. At the last epoch it is 0.021 to 0.027 nats on MNIST and 0.0015 to 0.0016 nats on Fashion-MNIST. With $K = 2$ the spread is $|\ell_1 - \ell_2|/2$, so the two restarts end on average 0.047 nats apart on MNIST and 0.003 nats on Fashion-MNIST. The larger Gibbs weight $(1 + e^{-|\ell_1 - \ell_2|/\tau})^{-1}$ is then 0.61 at $\tau = 0.1$ and 0.52 at $\tau = 0.5$ on MNIST, and only 0.58 even at $\tau = 0.01$ on Fashion-MNIST.



Spread of candidate losses during SMAT training

The temperature ablation in Table 2 agrees. On MNIST the worst case varies only from 33.0% to 34.4% across $\tau \in \{0.01, \ldots, 100\}$, within the test-sampling error, and clean accuracy is 93.6% throughout. On Fashion-MNIST the three temperatures give identical clean (79.2%), worst-case (59.0%), and Square (61.5%) accuracy. The largest weight difference between the MNIST models at $\tau = 0.01$ and $\tau = 100$ is 0.033, against 0.123 between SMAT and PGD-AT (0.029 and 0.110 on Fashion-MNIST). Smoothing changes the training signal only when candidates disagree by about τ, and PGD restarts here reach nearly the same loss. The theory thus explains why the temperature had no measurable effect; the 2.1-point MNIST difference is within noise.

Temperature and candidate-count ablation

Data	τ	K	Step	Seeds	Clean	Worst white-box	Square
MNIST	0.01	2	0.625ε	1	93.6	34.4	40.5
MNIST	0.1	2	0.625ε	1	93.6	33.4	40.5
MNIST	0.5	2	0.625ε	1	93.6	33.6	40.0

Data	τ	K	Step	Seeds	Clean	Worst white-box	Square
MNIST	2	2	0.625ε	1	93.6	33.2	40.0
MNIST	100	2	0.625ε	1	93.6	33.0	40.0
MNIST	0.5	4	1.25ε	3	93.5 ± 0.4	5.6 ± 1.2	14.5 ± 4.4
MNIST	0.5	4	ε	3	93.9 ± 0.2	7.3 ± 1.9	15.7 ± 4.5
F-MNIST	0.01	2	0.625ε	1	79.2	59.0	61.5
F-MNIST	0.5	2	0.625ε	1	79.2	59.0	61.5
F-MNIST	100	2	0.625ε	1	79.2	59.0	61.5

4.3 Splitting the budget into short chains

The $K = 4$ runs split the eight steps into four chains of two, on three seeds. With $\alpha = 2.5\varepsilon/s$ the step becomes 1.25ε, and large steps can cause catastrophic overfitting (Wong et al., 2020), so we also cap the step at ε. Both variants collapse: MNIST worst-case accuracy is $5.6 \pm 1.2\%$ uncapped and $7.3 \pm 1.9\%$ capped, against $32.3 \pm 1.3\%$ for two chains of four steps at similar clean accuracy. Step sizes above the radius are therefore not the main cause, although we did not test steps below ε. FGSM accuracy stays at 71% to 77% in all six runs while targeted APGD accuracy falls to 4.8% to 9.6%, the signature of a weak inner adversary. Their candidate spread is 0.094 to 0.116 nats, a maximum minus a mean of four, so it compares only roughly with the $K = 2$ spread. At a fixed budget, chain length matters more than the number of candidates, as Theorem 3 suggests: every guarantee is relative to the best chain.

4.4 How tight is the smoothness bound

With 16 frozen candidates on the seed-0 MNIST PGD-AT model, the maximum is 1.257 nats and the mean 1.194. At $\tau = 0.01$ the value is 1.235, between the bound $1.257 - 0.01\log 16 = 1.229$ and the maximum, as Theorem 3 requires. Table 3 tests Theorem 6 where its proof is tight. The adversary adds curvature $\tau^{-1}\lambda_{\max}(C_\tau)$ along the worst direction, and the proof bounds it by G^2/τ. On the four checked models G^2 is 333 to 553, while $\lambda_{\max}(C_\tau)$ is only 0.021 to 0.357. At $\tau = 0.01$ the curvature term is 2.1 to 28.7, against a bound of 3.3×10^4 to 5.5×10^4. The slack comes from the proof treating each candidate gradient as anywhere in a ball of radius G, whereas here they nearly agree.

The finite-difference ratio along the top eigenvector is 26.7 to 52.0, while the median along random directions is only 0.15 to 0.25. At $\tau = 100$ the ratio of 27.5 to 39.1 reflects the curvature of the losses themselves. The $1/\tau$ term is visible on the two models with the largest covariance term, where the ratio rises from 39.1 to 52.0 (MNIST SMAT) and from 35.2 to 42.4 (Fashion-MNIST PGD-AT). On the other two models it falls slightly, by 3.1 and 0.8. The bound is valid but overstates the adversary's curvature by three to four orders of magnitude on these networks.

4.5 Gradient masking check

A defense that only masks gradients would resist white-box attacks but fall to the gradient-free Square attack. On the same 200 test images, Square never lowers robust accuracy below the white-box worst case on Fashion-MNIST. On MNIST, adding Square lowers robust accuracy by at most 3.5 points for any model. Square alone beats the white-box worst case only for TRADES seeds 1 and 2, by 1.0 and 1.5 points, so we see no evidence that SMAT or PGD-AT masks gradients. With all five attacks combined, PGD-AT and SMAT reach 38.0% and 35.0% on these images.

4.6 Limitations

We trained small networks for 6 epochs on 20,000 images on one laptop processor, so all accuracies are far below full-scale results; MNIST PGD-AT reaches 34.4% worst case. With 500 test images, differences below about two points are within the binomial standard error. Our attacks are lighter than standard AutoAttack: three APGD-T targets, no FAB, and 1,000 Square queries on 200 images. The temperature ablation uses one seed. The rate results assume global constants and fixed candidates, while our runs use Adam with regenerated candidates. Only where restarts disagree could the temperature matter, and that may require larger networks or longer chains.

5 Conclusion

We studied adversarial training as a minimax problem with an entropy-regularized adversary. The regularized inner problem has a closed-form Gibbs solution whose value interpolates between the mean and the maximum of the candidate losses. We proved that this value lies within $\tau \log K$ of the best of K candidates. We proved that its gradient is $(\beta + G^2/\tau)$-Lipschitz, a constant that is attained in general. We proved that its near-stationary points are near-stationary for the min-max loss, with a logarithmic price in the temperature. These guarantees assume global constants and fixed candidates, which our training runs do not satisfy. At a matched budget of eight ascent steps, SMAT did not improve on PGD-AT. Worst-case accuracy was 58.5% for both on Fashion-MNIST, and 32.3% against 34.4% on MNIST, within the test-sampling error. The theory explains why the temperature barely mattered: two PGD restarts ended on average only 0.047 nats apart on MNIST and 0.003 nats on Fashion-MNIST. On trained

networks the adversary's curvature term was three to four orders of magnitude below the bound, because the candidate gradients nearly agreed. Splitting the budget into four two-step chains reduced MNIST worst-case accuracy to 5.6%, and to 7.3% with the step capped at the radius. These results come from small networks trained on one processor, on two grey-scale datasets. Smoothing the adversary can only help where restarts disagree, and testing larger models and longer chains for such disagreement is the next step.

Data availability
Both datasets analysed in this study are public and need no registration. MNIST (LeCun et al., 1998) is distributed at http://yann.lecun.com/exdb/mnist/ and mirrored at https://ossci-datasets.s3.amazonaws.com/mnist/, the source we downloaded from. Fashion-MNIST is available at https://github.com/zalandoresearch/fashion-mnist. We loaded both through the torchvision dataset classes, which download these files. The training subset is the first 20,000 indices of a fixed random permutation (seed 1234) of the 60,000 training images. The evaluation set is the first 500 images of the official test split.

Competing interests
The authors declare no competing interests.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square attack: A query-efficient black-box adversarial attack via random search. *Computer Vision – ECCV 2020, Lecture Notes in Computer Science*, *12368*, 484–501. https://doi.org/10.1007/978-3-030-58592-1_29

2. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *80*, 274–283. https://proceedings.mlr.press/v80/athalye18a.html

3. Awasthi, P., Mao, A., Mohri, M., & Zhong, Y. (2023). Theoretically grounded loss functions and algorithms for adversarial robustness. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS), Proceedings of Machine Learning Research*, *206*, 10077–10094. https://proceedings.mlr.press/v206/awasthi23c.html

4. Beck, A., & Teboulle, M. (2012). Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, *22*(2), 557–580. https://doi.org/10.1137/100818327

5. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. https://doi.org/10.1109/SP.2017.49

6. Clarke, F. H. (1990). *Optimization and nonsmooth analysis* (Vol. 5). Society for Industrial; Applied Mathematics. https://doi.org/10.1137/1.9781611971309

7. Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *97*, 1310–1320. https://proceedings.mlr.press/v97/cohen19c.html

8. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *119*, 2206–2216. https://proceedings.mlr.press/v119/croce20b.html

9. Danskin, J. M. (1967). *The theory of max-min and its application to weapons allocation problems*. Springer. https://doi.org/10.1007/978-3-642-46092-0

10. Davis, D., & Drusvyatskiy, D. (2019). Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, *29*(1), 207–239. https://doi.org/10.1137/18M1178244

11. Dong, Y., Deng, Z., Pang, T., Zhu, J., & Su, H. (2020). Adversarial distributional training for robust deep learning. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*. https://proceedings.neurips.cc/paper/2020/hash/5de8a36008b04a6167761fa19b61aa6c-Abstract.html

12. Donsker, M. D., & Varadhan, S. R. S. (1975). Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, *28*(1), 1–47. https://doi.org/10.1002/cpa.3160280102

13. Fiacco, A. V. (1983). *Introduction to sensitivity and stability analysis in nonlinear programming* (Vol. 165). Academic Press.

14. Gao, R., Cai, T., Li, H., Wang, L., Hsieh, C.-J., & Lee, J. D. (2019). Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 13029–13040. https://proceedings.neurips.cc/paper/2019/hash/348a38cd25abeab0e440f37510e9b1fa-Abstract.html

15. Ghadimi, S., & Lan, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, *23*(4), 2341–2368. https://doi.org/10.1137/120880811

16. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2015/goodfellow2015iclr-explaining/

17. Jagatap, G., Joshi, A., Chowdhury, A. B., Garg, S., & Hegde, C. (2022). Adversarially robust learning via entropic regularization. *Frontiers in Artificial Intelligence*, *4*, 780843. https://doi.org/10.3389/frai.2021.780843

18. Jin, C., Netrapalli, P., & Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? *Proceedings of the 37th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *119*, 4880–4889. https://proceedings.mlr.press/v119/jin20e.html

19. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

20. Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2021). Tilted empirical risk minimization. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2021/li2021iclr-tilted/

21. Li, T., Beirami, A., Sanjabi, M., & Smith, V. (2023). On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, *24*(142), 1–79. https://www.jmlr.org/papers/v24/21-1095.html

22. Lin, T., Jin, C., & Jordan, M. (2020). On gradient descent ascent for nonconvex-concave minimax problems. *Proceedings of the 37th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *119*, 6083–6093. https://proceedings.mlr.press/v119/lin20a.html

23. Liu, C., Salzmann, M., Lin, T., Tomioka, R., & Süsstrunk, S. (2020). On the loss landscape of adversarial training: Identifying challenges and how to overcome them. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 21476–21487. https://proceedings.neurips.cc/paper/2020/hash/f56d8183992b6c54c92c16a8519a6e2b-Abstract.html

24. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2018/madry2018iclr-deep/

25. Maini, P., Wong, E., & Kolter, Z. (2020). Adversarial robustness against the union of multiple perturbation models. *Proceedings of the 37th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *119*, 6640–6650. https://proceedings.mlr.press/v119/maini20a.html

26. Meunier, L., Scetbon, M., Pinot, R. B., Atif, J., & Chevaleyre, Y. (2021). Mixed Nash equilibria in the adversarial examples game. *Proceedings of the 38th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *139*, 7677–7687. https://proceedings.mlr.press/v139/meunier21a.html

27. Nesterov, Yu. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming*, *103*(1), 127–152. https://doi.org/10.1007/s10107-004-0552-5

28. Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., & Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 14934–14942. https://proceedings.neurips.cc/paper/2019/hash/25048eb6a33209cb5a815bff0cf6887c-Abstract.html

29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 8026–8037. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

30. Rice, L., Wong, E., & Kolter, Z. (2020). Overfitting in adversarially robust deep learning. *Proceedings of the 37th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *119*, 8093–8104. https://proceedings.mlr.press/v119/rice20a.html

31. Robey, A., Chamon, L., Pappas, G. J., & Hassani, H. (2022). Probabilistically robust learning: Balancing average and worst-case performance. *Proceedings of the 39th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *162*, 18667–18686. https://proceedings.mlr.press/v162/robey22a.html

32. Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 3358–3369. https://mlanthology.org/neurips/2019/shafahi2019neurips-adversarial/

33. Sinha, A., Namkoong, H., & Duchi, J. (2018). Certifying some distributional robustness with principled adversarial training. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2018/sinha2018iclr-certifying/

34. Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics*, *8*(1), 171–176. https://doi.org/10.2140/pjm.1958.8.171

35. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2014/szegedy2014iclr-intriguing/

36. Tramèr, F., & Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems (NeurIPS)*, *32*, 5866–5876. https://mlanthology.org/neurips/2019/tramer2019neurips-adversarial/

37. Wang, J., Gao, R., & Xie, Y. (2026). Sinkhorn distributionally robust optimization. *Operations Research*, *74*(3), 1581–1603. https://doi.org/10.1287/opre.2023.0294

38. Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., & Gu, Q. (2019). On the convergence and robustness of adversarial training. *Proceedings of the 36th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *97*, 6586–6595. https://proceedings.mlr.press/v97/wang19i.html

39. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., & Gu, Q. (2020). Improving adversarial robustness requires revisiting misclassified examples. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2020/wang2020iclr-improving/

40. Wong, E., Rice, L., & Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. *International Conference on Learning Representations (ICLR)*. https://mlanthology.org/iclr/2020/wong2020iclr-fast/

41. Wu, D., Xia, S.-T., & Wang, Y. (2020). Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*. https://papers.nips.cc/paper/2020/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html

42. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. *Proceedings of the 36th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research*, *97*, 7472–7482. https://proceedings.mlr.press/v97/zhang19p.html

43. Zhang, J., Xiao, P., Sun, R., & Luo, Z. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*. https://proceedings.neurips.cc/paper/2020/hash/52aaa62e71f829d41d74892a18a11d59-Abstract.html