Adversarial Attacks (White-Box)

Targeted FGSM: $f(x - \varepsilon \cdot \text{sgn}(\nabla_x \text{loss}_t(x)))$ approaches t

Untargeted FGSM: $f(x + \varepsilon \cdot \text{sgn}(\nabla_x \text{loss}_s(x)))$ strays from s

↳ both guarantee $x' \in [x-\varepsilon, x+\varepsilon]$ box

$\min_\eta \|\eta\|_p + c\, \text{obj}_t(x+\eta)$ s.t. $x+\eta \in [0,1]^n$

with $\text{obj}_t(x+\eta) \le 0$ if $f(x+\eta) = t$, e.g. $\text{obj}_t(x) = \max(0, 0.5 - p_f(x)_t)$

CW: Use LBFGS-B optimizer or $\text{obj}_t(x) = -\log_2(p_f(x)_t)$

PGD: Repeat FGSM, projecting to $[x-\varepsilon, x+\varepsilon]$

Adversarial Defenses

Adversarial Accuracy $= \frac{\text{correctly classified, even after PGD-attack}}{\text{tested samples}}$

Training aims to solve $\min_\theta \mathbb{E}_{(x,y)\sim D}[\max_{x' \in S(x)} L(\theta, x', y)]$

Approximate inner maximization with PGD

GCG attacks on LLMs: Use "Sure" as next-token target x_{n+2}.

$x'_{n+1} := \arg\min_{x_{n+1} \in S} L(x_{n+1})$ for $S := \text{argTopK}(\nabla_{\text{onehot}(x_{n+1})} L(x_{n+1}))$

NN-Certification

Rice's Theorem: no sound & complete general certification algorithms

Box ($O(n^2L)$): $-^\#[a,b] = [-b,-a]$, $\text{ReLU}^\#[a,b] = [\text{ReLU}(a), \text{ReLU}(b)]$,

$[a,b] \cdot^\# [c,d] = [\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)]$, + and λ trivial

DeepPoly ($O(n^3L^2)$):

Affine: encode exactly, being careful with signs for l_y and u_y !!

ReLU: if $u_x \le 0$: $0 \le y \le 0$, $l_y = 0$, $u_y = 0$

if $l_x \ge 0$: $x \le y \le x$, $l_y = l_x$, $u_y = u_x$

otherwise: $\lambda := \frac{u_x}{u_x - l_x}$, $\alpha x \le y \le \lambda(x - l_x)$, $l_y = 0$, $u_y = u_x$

with $\alpha \in [0,1]$, min. area says $\alpha = 0$ iff $u_x \le -l_x$

ψ: encode using single output neuron o, proving $l_o > 0$ etc.

Branch and Bound: Split ReLU at $x=0$ and use input constraints via KKT:

$(\max_x f(x)$ st. $g(x) \le 0) \le \max_x \min_\beta f(x) - \beta g(x)$.

	$\min_x$	$\max_x$
$g(x)<0$	$\max_\beta$ +	$\min_\beta$ −
$g(x)>0$	$\max_\beta$ −	$\min_\beta$ +

MILP (NP-complete):

Affine: $Wx + b \le y \le Wx + a$

ReLU: $y \ge 0$, $y \ge x$, $y \le u_x a$, $y \le x - l_x(1-a)$, $a \in \{0,1\}$

ϕ: $x_i - \varepsilon \le x'_i \le x_i + \varepsilon$, ψ: objective $\min o_0 - o_1$

Box-constraints (to accelerate): $l_i \le x_i^p \le u_i$

PRIMA: Abstract Neurons jointly, computing convex hull via dual problem of intersecting halfspaces (under-approximate)

Certified Defenses

Training aims to solve $\min_\theta \mathbb{E}_{(x,y)\sim D}[\max_{z \in \gamma(NN^\#(S(x)))} L^\#(\theta, z, y)]$

with $L(z,y) = \max_{q \ne y}(z_q - z_y)$, compute $\max_c(\max(\text{box}(z_c -^\# z_y)))$

$L(z,y) = CE(z,y)$, compute $CE(\text{softmax}([u_1, \dots, l_y, \dots, u_{|c|}]), y)$

COLT: Only compute symbolically up to nth layer, then PGD.

Problem: Projection onto DeepPoly shape is hard.

Randomized Smoothing for Robustness

Given classifier f, make $g(x) := \arg\max_{c_A} \mathbb{P}_\varepsilon[f(x+\varepsilon) = c_A]$ with

with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $p_A(x) := \mathbb{P}_\varepsilon[f(x+\varepsilon) = g(x)]$, $p_B(x) := \max_{c_B \ne g(x)} \dots$ (2nd highest)

Robustness guarantee: $p_A(x) \ge \underline{p_{A,x}} \ge \overline{p_{B,x}} \ge p_B(x) \Rightarrow g(x+\delta) = c_A$ for all $\|\delta\|_2 < R_x := \sigma/2\,(\Phi^{-1}(\underline{p_{A,x}}) - \Phi^{-1}(\overline{p_{B,x}}))$ certification radius

If $\underline{p_A} > \frac{1}{2}$, then $R_x \ge \sigma\Phi^{-1}(\underline{p_A})$, allows for efficient certification.

↳ Guess $\hat{c}_A$ via Monte-Carlo Integration $p_A(x) = \int_\varepsilon \mathbb{I}_{f(x+\varepsilon)=c_A} \text{PDF}_{\mathcal{N}(0,\sigma^2 I)}(\varepsilon)\, d\varepsilon$

Repeat Monte-Carlo with large n, counting k hits

Estimate $\underline{p_A} := \beta^{-1}(\frac{\alpha}{2}, k, n-k+1)$ via Copper-Pearson

Then, with probability $> 1-\alpha$, $g(x+\delta) = \hat{c}_A$ for all $\delta < R_x$

If ABSTAIN, true $p_A \le 0.5$ OR guess was wrong OR lower bound too lose.

At inference, get top-2-classes via Monte-Carlo Integration and do Binomial Pvalue test on $n_A, n_A + n_B, =, 0.5$.

Privacy

Attacks: Model Stealing, Inversion (exact / representative), Membership Inference

Black-Box MI: Train n models on the same data distribution, some with, some without x. If logits given, train classifier mapping logits to x-membership. Otherwise, train classifier mapping adversarial robustness to x-membership.

Federated Learning (FedSGD):

Client k computes $g_k := \nabla_\theta L(\theta, x_k, y_k)$ for minibatch $(x_k, y_k) \sim D_k$

Server Updates model $\theta_{t+1} := \theta_t - \gamma g_k$

Attack: batch size $1 \Rightarrow$ exact reconstruction by gradient inversion for piecewise-linear NN (ReLU-based networks)

batch size $>1 \Rightarrow$ reconstruct linear combination of inputs

$\arg\min_{x^*} \text{dist}(\nabla_\theta L(\theta, x^*, y^*), g_k) + \alpha_{reg} \mathcal{R}(x^*)$

Regularization prior, e.g. text perplexity, image variation

FedAVG:

Client runs E epochs of SGD. Server updates $\theta_{t+1} := \frac{1}{k}\sum_{k=1}^{k} \theta_{t,k}$.

Attack: Simulate client to $\tilde{\theta}_k$, compute $\nabla_{\tilde{x}_k} \text{dist}(\tilde{\theta}_k, \theta_k)$.

$\arg\min_{x^k} \text{dist}(\tilde{\theta}_k, \theta_k) + \alpha_{reg} \underbrace{\frac{1}{E^2}\sum_{e_1, e_2} \mathcal{R}(g(\{\tilde{x}_{e_1,b}\}), g(\{\tilde{x}_{e_2,b}\}))}$

Regularization over average distance between average images of epoch pairs

Differential Privacy

Mechanism M is ε-DP iff $\forall (a,a') \in \text{Neigh}.\ \forall S.\ p(a) := \mathbb{P}[M(a) \in S] \le e^\varepsilon p(a')$

Laplace Mechanism: $f(a) + \text{Lap}(0, \frac{\Delta_1}{\varepsilon})$ with $\Delta_1 := \max_{(a,a') \in \text{Neigh}} \|f(a) - f(a')\|_1$

M is (ε,δ)-DP iff $\forall (a,a') \in \text{Neigh}.\ \forall S.\ p(a) \le e^\varepsilon p(a') + \delta$

Gaussian Mechanism: $f(a) + \mathcal{N}(0, \sigma^2 I)$ with $\sigma := \sqrt{2\log(1.25)/\delta}\,\frac{\Delta_2}{\varepsilon}$

↳ useful if $\exists (a,a') \in \text{Neigh}.\ p(a') = 0 \wedge p(a) \ne 0$

M_1 is (ε_1,δ_1)-DP and M_2 is (ε_2,δ_2)-DP $\Rightarrow (M_1, M_2)$ and $M_1 \circ M_2$ are $(\varepsilon_1+\varepsilon_2, \delta_1+\delta_2)$-DP

M is (ε,δ)-DP $\Rightarrow f \circ M$ is (ε,δ)-DP

M_i is (ε_i,δ_i)-DP $\Rightarrow \prod_i M_i(a_i)$ is $(\max_i \varepsilon_i, \max \delta_i)$-DP

DP-SGD: Project within-batch gradients onto l_2-ball of radius C. For batch size L, add $\mathcal{N}(0, \sigma^2 I)$ to aggregate gradient, where $\sigma = \sqrt{2\log(1.25)/\delta}\,\frac{C}{\varepsilon}$.

Privacy Amplification: Applying an (ε,δ)-DP mechanism on a random fraction $q = L/N$ of data yields a $(\tilde{q}\varepsilon, q\delta)$-DP mechanism, where $\tilde{q} \propto q$.

PATE: Split data and train teachers T. Label public dataset by noisy voting: $\arg\max_j(|\{t(x) = j \mid t \in T\}| + \text{Lap}(0, \frac{2}{\varepsilon}))$. Method is $(\varepsilon, 0)$-DP for one label, $(\varepsilon T, 0)$-DP for T labels.

AI Regulation

Fairness, Explainability, Data Minimization, Unlearning, Copyright, Emergency Response.

Private Synthetic Data

Select Marginal Queries, Measure using DP, Generate Data

Marginal on attributes C is $M_C(D) := \mu \in \mathbb{R}^{\prod_{i \in C} \Omega_i}$ where $\mu_i = \sum_{x \in D} \mathbb{I}_{x_C = i}$. 1-way marginals are histograms. $\Delta_2 = 1$.

Mutual Information is $I(X,Y) = \sum_{x,y} \frac{p(x,y)}{p(x)p(y)}$ and is used as edge weight in Chow-Liu algorithm. MST is the optimal 2nd-order approximation. Compute via belief propagation.

Logic in Deep Learning

Test satisfiability of logical formulas over an NN's output.

$(\text{class}(NN(i)) = c^*) \equiv \bigwedge_{c \ne c^*} NN(i)[c] < NN(i)[c^*]$

SAT solvers time out on large networks, so translate:

For all x, $T(\phi)(x) = 0 \iff x \models \phi$ with differentiable T:

$t_1 \le t_2 \iff \max(0, t_1 - t_2)$ $t_1 \ne t_2 \iff \mathbb{I}_{t_1 = t_2}$

$t_1 = t_2 \iff T(t_1 \le t_2 \wedge t_2 \le t_1)$ $\varphi \vee \psi \iff T(\varphi) \cdot T(\psi)$

$t_1 < t_2 \iff T(t_1 \le t_2 \wedge t_1 \ne t_2)$ $\varphi \wedge \psi \iff T(\varphi) + T(\psi)$

Translate negation with deMorgan $\neg(\varphi \wedge \psi)\ \neg\varphi \vee \neg\psi$.

By construction, $T(\varphi)(x) \ge 0$ for all φ, x.

Box constraints hard to optimize, so use L-BFGS-B solver.

Training with logic as maximization

$\max_\theta \mathbb{E}_{s \sim D}[\forall z.\ \phi(z, s, \theta)]$

Generalized adversarial training beyond robustness

$\min_\theta \mathbb{E}_{s \sim D}[T(\phi)(\arg\min_z T(\neg\phi(z, s, \theta)), s, \theta)]$

Restrict z to a convex set with efficient projections.

Fairness

Fairness through unawareness (algorithm does not explicitly use sensitive data) does not work.

Individual Fairness: For every $x, y \in \mathcal{X}$, $D(M(x), M(y)) \leq d(x,y)$ where (D, d) are two similarity metrics. M is (D, d)-lipschitz. The challange is finding suitable d and D.

Fairness as Robustness:
For $h: \mathbb{R}^d \to [0,1]$, $x \mapsto \Phi^{-1}\left(\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,I)}[h(x+\varepsilon)]\right)$ is 1-Lipschitz.
Chosing $d(x,x') := (x-x')^T S (x-x')$ for symmetric positive definite covariance matrix S and $D(M(x), M(x')) := \mathbb{I}_{M(x) \neq M(x')}$ allows reformulation of lipschitzness for all $\|\delta\|_S < 1/\sqrt{L}$ $M(x) = M(x+\delta)$, where $\|x\|_S := \sqrt{x^T S x}$ (Mahalanobis dist.)

Fair Representation learning:
Data Regulator: Define fairness and data sources
Data Producer: Generate fair encoder $f_\theta: \mathbb{R}^n \to \mathbb{R}^k$.
Data Consumer: Use encoded data to build consumer
Modularity comes at the cost of no certification and unclear fairness-performance tradeoff.

LCIFR:
Define D and d with logic accepted by MILP or DL2, e.g.
$d(x,x') = \bigwedge_{i \in Cat \setminus \{gender\}} (x_i = x_i') \bigwedge_{j \in Num} |x_j - x_j'| \leq \alpha$.
For each x, obtain $S_d(x) = \{x' \mid d(x,x')\}$.
Train encoder using DL2 s.t. $\forall x' \in S_d(x)\ \|f_\theta(x) - f_\theta(x')\|_\infty \leq \delta$ (see above), appending general classifier to preserve utility.
Now, consumers must be δ-robust (e.g. by randomized smoothing) to certify end-to-end fairness.

LaSSI:
For high-dimensional data (e.g. images), use semantic feature space from generative model in similarity formulas. Use center smoothing to produce end-to-end fair model with probability $1 - \alpha_{rs} - \alpha_{cs}$.

Group Fairness:
Demographic parity $\mathbb{P}[\hat{Y}=1 \mid G=0] = \mathbb{P}[\hat{Y}=1 \mid G=1]$
Equal Opportunity $\mathbb{P}[\hat{Y}=1 \mid Y=1, G=0] = \mathbb{P}[\hat{Y}=1 \mid Y=1, G=1]$
Equalized odds and $\mathbb{P}[\hat{Y}=1 \mid Y=0, G=0] = \mathbb{P}[\hat{Y}=1 \mid Y=0, G=1]$

Group Fairness $\not\Rightarrow$ Individual Fairness

Post-Processing Approach: In a binary classifier, use different threshold depending on sensitive attribute.

In-training approach: Add relaxed fairness constraints solved with DL2, e.g. $-\varepsilon \leq \mathbb{P}(\hat{Y}=1 \mid s=0) - \mathbb{P}[\hat{Y}=1 \mid s=1] \leq \varepsilon$.

Preprocessing with guarantees: Jointly train encoder f, classifier g, adversary h trying to predict sensitive attributes from latent data:
$\min_{f,g} \max_h \left(\mathcal{L}_{clf}(f(x,s), g) - \gamma \mathcal{L}_{adv}(f(x,s), h) \right)$ (LAFTR)
Bounding Unfairness via the optimal adversary:
Soft formulation of demographic parity: $\Delta_{z \sim \mathcal{Z}_0, \mathcal{Z}_1} := \left| \mathbb{E}_{z \sim \mathcal{Z}_0} g(z) - \mathbb{E}_{z \sim \mathcal{Z}_1} g(z) \right|$

Balanced accuracy of adversary: $BA_{\mathcal{Z}_0 \mathcal{Z}_1}(h) := \frac{1}{2}\left(\mathbb{E}_{z \sim \mathcal{Z}_0}(1-h(z)) + \mathbb{E}_{z \sim \mathcal{Z}_1} h(z) \right)$
$= \frac{1}{2} \int_z \left(p_0(z)(1-h(z)) + p_1(z)\, h(z) \right) dz$

$\Delta_{\mathcal{Z}_0, \mathcal{Z}_1}(g) \leq 2\, BA_{\mathcal{Z}_0, \mathcal{Z}_1}(h^*) - 1$ for optimal $h^*(z) := \mathbb{I}_{p_1(z) \geq p_0(z)}$

LAFTR assumes family of h $\leadsto$ E2E fairness is overestimated

Fair Normalizing Flows: Estimate densities q_1 and q_2, learn invertible encoder $z = f(x)$. Find density of the output distributions by $\log p(x) = \log q(f^{-1}(z)) + \log \left| \det \frac{\partial f^{-1}(z)}{\partial z} \right|$. Optimize for low KL-divergence between p_0 and p_1 and utility with downstream classifier g. Hoeffding's Inequality bounds BA with probability $1-\varepsilon$. FNF guarantees only w.r.t. estimated densities.

Fairness with Restricted Encoders: Restrict latent space to finite set (e.g. k-Means clusters), then compute direct bounds on each output.

Basics

l_p-norm $\|x\|_p := \left(\sum_i |x_i|^p \right)^{1/p}$, $\|x\|_\infty := \max_i |x_i|$
$\mathbb{B}^1_\varepsilon \subseteq \mathbb{B}^2_\varepsilon \subseteq \mathbb{B}^\infty_\varepsilon \subseteq \mathbb{B}^2_{\varepsilon\sqrt{d}} \subseteq \mathbb{B}^1_{\varepsilon d}$

CE-loss $CE(x,y) := -\sum_c y \log_2(p_f(x)_t)$
↳ for classification $y := onehot(t)$, so $loss_t(x) = -\log_2(p_f(x)_t)$
softmax $p_f(x)_t := \frac{\exp(x_t)}{\sum_c \exp(x_c)}$

Hölder: $\|u^T v\|_1 \leq \|u\|_p \|v\|_q$ for $\frac{1}{p} + \frac{1}{q} = 1$
Weak Duality: $\max_a \min_b f(a,b) \leq \min_b \max_a f(a,b)$
subadditivity: $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$
Cauchy-Schwarz: $x^T y \leq \|x\|_2 \|y\|_2$

Project z onto $\mathbb{B}^p_\varepsilon(x)$
$p=\infty$: $z_i' := \max(\min(z_i, x_i + \varepsilon), x_i - \varepsilon)$
$p=2$: $z' := x + \frac{z-x}{\max(1, \|z-x\|_2 / \ )}$
$p=1$: Soundly approximate using the $p=2$ approach.

Normal CDF:
$x \sim \mathcal{N}(0,1)$: $\mathbb{P}(x \leq z) = \Phi(z)$, $P(x \leq \Phi^{-1}(z)) = z$
$x \sim \mathcal{N}(\mu, \sigma^2)$: $\mathbb{P}(x \leq z) = \Phi\left(\frac{z-\mu}{\sigma}\right)$, $P(x \leq \mu + \sigma\Phi^{-1}(z)) = z$
$\alpha\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\alpha\mu_1 + \mu_2, \alpha^2\sigma_1^2 + \sigma_2^2)$
Chebyshev: $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$
$\mathbb{E}(X) := \int_{-\infty}^{\infty} x f(x)\, dx$ with PDF f.
$\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$\mathbb{P}[A \mid B] := \frac{P[A \cap B]}{P[B]}$
Bayes $P[A \mid B] = \frac{P[B \mid A]\, P[A]}{P[B]}$

Derivatives:
$(fg)' = f'g + fg'$, $(f/g)' = (f'g - fg')/g^2$
$(f \circ g)' = (f' \circ g) g'$, $\ln'(x) = \frac{1}{x}$
$\nabla_x b^T x = \nabla_x x b^T = b$ $\quad \nabla_x x^T x = \nabla_x \|x\|_2^2 = 2x$
$\nabla_x x^T A x = (A^T + A)x$ $\quad \nabla_x \|x - b\|_2 = \frac{x-b}{\|x-b\|_2}$
$\nabla_x \|x\|_1 = \frac{x}{|x|}$ $\quad \nabla_x \|Ax - b\|_2^2 = 2(A^T A x - A^T b)$