1. Adversarial Attacks

Targeted FGSM: $x' = x - \eta, \quad \eta = \varepsilon \cdot \operatorname{sign}(\nabla_x \operatorname{loss}_t(x))$
Guarantees $\eta \in [-\varepsilon, \varepsilon]$, η not minimized

Untargeted FGSM: $x' = x + \eta, \quad \eta = \varepsilon \cdot \operatorname{sign}(\nabla_x \operatorname{loss}_s(x))$
Guarantees $\eta \in [-\varepsilon, \varepsilon]$, η not minimized

Carlini & Wagner: Find targeted adv. sample $x' = x + \eta$ *and* minimize $\|\eta\|_p$ via minimizing $\|\eta\|_p + c \cdot \operatorname{obj}_t(x')$, where obj_t is s.t. $\operatorname{obj}_t(x') \leq 0 \Rightarrow f(x') = t$, e.g. $\operatorname{CE}(x', t) - 1$; $\max\left(0, 0.5 - p_f(x')_t\right)$
Prior e.g. $x + \eta \in [0, 1]$: use specialized optimizer (LBFGS-B) or PGD. Optimizing $\|\eta\|_\infty$ is hard, use $\operatorname{ReLU}(\sum |\eta_i| - \tau)$, lower τ gradually.

PGD: Repeat FGSM with $\varepsilon_{\text{step}}$ and proj. to $x \pm \varepsilon$.

2. Adversarial Defenses

Defense as Optimization: $\min_\theta \mathbb{E}_{(x,y)\sim D}\left[\max_{x' \in S(x)} L(\theta, x', y)\right]$
usually $S(x) = \mathbb{B}_\varepsilon^\infty$, $\mathbb{E} \approx$ empirical risk.

PGD Defense algorithm: Run PGD on every batch and use $\nabla_\theta \mathcal{L}(x_{\text{adv}})$ for backprop.

TRADES defense: $\min_\theta \mathbb{E}_{(x,y)\sim D}\left[L(\theta, x, y) + \lambda \max_{x' \in \mathbb{B}_\varepsilon(x)} L(\theta, x', y)\right]$

3. Certification of Neural Networks

Given NN N, precond. φ, postcond. ψ prove: $\forall i \quad i \vDash \varphi \Rightarrow N(i) \vDash \psi$ or return a violation.

3.1 Complete Methods (always return result)

MILP Encoding: Encode NN as MILP instance. Doesn't scale well.
- Affine: $y = Wx + b$ is a direct MILP constraint. $Wx + b \leq y \leq Wx + b$.
- $\operatorname{ReLU}(x)$: $y \leq x - l_x \cdot (1 - a), y \geq x, y \leq u_x \cdot a, y \geq 0, a \in \{0, 1\}$, for box bound $x \in [l, u]$.
- $a = 0$: $y = 0, x \in [l, 0]$
- $a = 1$: $y = x, y \in [0, u]$

To check an encoding for f, plot constraint regions for all cases of int. variables. They should match plot of f. Can't use $a \cdot x$.
$\varphi = \mathbb{B}_\varepsilon^\infty(x)$: $x_i - \varepsilon \leq x_{i'} \leq x_i + \varepsilon, \forall i$
precomp. Box bounds: $l_i \leq x_i^p \leq u_i$
$\psi = o_0 > o_1$: MILP objective $\min o_0 - o_1$.

3.2 Incomplete Methods (may abstain)

Box: $\mathcal{O}(n^2 L)$: Bounds are l_∞ balls.
$[a, b] +^\# [c, d] = [a + b, c + d]$, $\quad -^\#[a, b] = [-b, -a]$;
$\operatorname{ReLU}^\#[a, b] = [\operatorname{ReLU}(a), \operatorname{ReLU}(b)]$;
$\lambda \cdot^\# [a, b] = [\lambda a, \lambda b]$ $(\lambda \geq 0)$

DeepPoly: $\mathcal{O}(n^3 L^2)$: For each x_i keep constraints:
interval $l_i \leq x_i, x_i \leq u_i$;
relational $a_i^\leq \leq x_i, x_i \leq a_i^\geq$ where $a_i^\leq, a_i^\geq$ are of the form $\sum_j w_j \cdot x_j + \nu$
- $x_j = \operatorname{ReLU}^\#(x_i)$: interval constr. $x_i \in [l_i, u_i]$:

$u_i \leq 0$: $a_j^\leq = a_j^\geq = 0, l_j = u_j = 0$;
$l_i \geq 0$: $a_j^\leq = a_j^\geq = x_i, l_j = l_i, u_j = u_i$;
$l_i < 0, u_i > 0$: $\quad \lambda := u_i / (u_i - l_i), x_j \leq \lambda(x_i - l_i)$, $\alpha \in [0, 1], \alpha x_i \leq x_j, l_j = 0, u_j = u_i$.
Min area: if $u \leq -l$, $\alpha = 0$, otherwise 1.
When proving $y_2 > y_1$, add a layer that computes $y_2 - y_1$ and prove $l_{y_2 - y_1} > 0$.

Branch & Bound: Split ReLU based on $x_i \leq 0$, resulting bound is the worst of two cases. Naive split still covers extra space, need constraints. KKT: $(\max f(x) \texttt{ mid } g(x) \leq 0) \leq \max_x \min_\beta f(x) - \beta g(x)$
- $(\max_x \vec{a}\vec{x} + c \text{ s.t. } -x_i \leq 0) \leq \max_x \min_\beta \vec{a}\vec{x} + c + \beta x_i$
- $(\max_x \vec{a}\vec{x} + c \text{ s.t. } x_i \leq 0) \leq \max_x \min_\beta \vec{a}\vec{x} + c - \beta x_i$

Usually you use the weak duality after this. β is found by GD, and on each step you do full backsubstitution after the split, as the sign in front of symbolic variables can change when β changes.

4. Certified Defenses

Produces models that are easier to certify.

4.1 DiffAI

PGD: $\min \mathbb{E}_{(x,y)\sim D}\left[\max_{z \in \gamma(\mathrm{NN}^\#(S(x)))} L(\theta, z, y)\right]$
Can use any abstract transformer (Box, DeepPoly).
To find max loss, use abstract loss $L^\#(\vec{z}, y)$, where $y =$ target label, $\vec{z} =$ vector of logits:
- $L(z, y) = \max_{q \neq y}(z_q - z_y)$: Compute $d_c = z_c - z_y \ \forall c \in \mathcal{C}$, where z_c the abstract logit shape of class i. Then compute box bounds of d_c and compute max upper bound: $\max_{c \in \mathcal{C}}(\max(\operatorname{box}(d_c)))$
- $L(z, y) = \operatorname{CE}(z, y)$: Compute box bounds $[l_c, u_c]$ of z_c. $\forall c \in \mathcal{C}$ pick u_c if $c \neq y$, pick l_c if $c = y$, hence $v = \left[u_1, .., l_c, .., u_{|\mathcal{C}|}\right]$.

Compute $\operatorname{CE}(\operatorname{softmax}(v), y)$.

4.2 COLT

COLT: Run relaxation up to some layer: $S' = \mathrm{NN}_{1 \ldots i}^\#(S(x))$, then run PGD on the region to train layers $i + 1 \ldots n$. For PGD we need to project back to S', which is not efficient for DeepPoly.

5. Randomized Smoothing for Robustness

Smoothed Classifier: Given any classifier f, make a smoothed classifier $g(x) := \arg\max_{c_A \in Y} \mathbb{P}_\varepsilon(f(x + \varepsilon) = c_A)$, where $\varepsilon \sim \mathcal{N}(0, \sigma I)$, $p_A(x)$ is the probability under argmax.
If $\exists \underline{p_{A,x}}, \overline{p_{B,x}} \in [0, 1]$ s.t. $p_A(x) \geq \underline{p_{A,x}} \geq \overline{p_{B,x}} \geq \max_{B \neq A} p_B(x)$, then $g(x + \delta) = c_A \quad \forall \|\delta\|_2 < R_x$ aka certification radius $= \delta/2\left(\Phi^{-1}\left(\underline{p_{A,x}}\right) - \Phi^{-1}\left(\overline{p_{B,x}}\right)\right)$.
Calculating $p_{A,x}, p_{B,x}$ directly is hard, so we use bounds. Calculating $\overline{p_{B,x}}$ is also hard, so let's assume $\underline{p_{A,x}} > 0.5$, then $\overline{p_{B,x}} = 1 - \underline{p_{A,x}}$.

```
Certification: $\widehat{c_A} \leftarrow$ guess_top_class($f, \sigma, x, n_0$)
$\underline{p_A} \leftarrow$ lower_bound_p($\widehat{c_A}, f, \sigma, x, n, \alpha$)
If $\underline{p_A} > 0.5$:
    $R \leftarrow \sigma\Phi^{-1}\left(\underline{p_A}\right)$
    Return $\widehat{c_A}, R$
Else:
    Return ABSTAIN
```

Notes: Top class is estimated via Monte-Carlo. Lower bound is estimated by CLT, Chebyshev's inequality or binomial confidence bounds. The two function calls involve sampling, the samples should be separate, and $n \gg n_0$. If the algorithm returns ABSTAIN, one of the following is true:
- $\widehat{c_A}$ is wrong, fixed by increasing n_0
- True $p_A \leq 0.5$, unfixable
- Lower bound is too low, fixed by increasing n

```
Inference:
$\widehat{c_A}, n_A, \widehat{c_B}, n_B \leftarrow$ top_two_classes($f, \sigma, x, n$)
Return BinomPValue($n_A, n_A + n_B, =, 0.5$) $\leq \alpha$ ? $\widehat{c_A}$ : ABSTAIN
```

Hypothesis Testing:
- NH: true $p(\text{success})$ of f returning $\hat{c}_A$ is 0.5
- BinomPValue returns p-value of null hypothesis, evaluated on n iid samples with i successes.
- Accept NH if p-value is $> \alpha$, reject otherwise.
- α small: often accept null hypothesis and ABSTAIN, but more confident in predictions.
- α large: more predictions but more mistakes.
- Returns wrong class with probability at most α

6. Privacy

Common attacks: Model Stealing, Model Extraction (representative inputs), Data Extraction (exact training samples), Membership Inference (find out if a sample was used for training).
Black-Box MI: Attacker trains many models on the same data distribution, some with entry x, some without. If logits are given, then attacker trains a classifier to distinguish between the two cases. If not, then do the same with robustness scores.

6.1 Federated Learning

FedSGD: Entities do training steps on minibatches $\{x^k, y^k\}$ from private data $\mathcal{D}_k$ and return gradients $g_k := \nabla_\theta \mathcal{L}\left(f_{\theta_t}(x^k), y^k\right)$, average on server and update the global model $\theta_{t+1} := \theta_t - \gamma g_c$. But sent data still contains information about private data.

Honest but curious server: Server does not manipulate sent weights. For batch size 1 and piecewise linear activation functions, the server can learn the data exactly. For batch size > 1 and some assumptions, a linear combination of some true inputs can be found. The general approach is: $\arg\min_{x^*} d\left(g_k, \nabla_\theta \mathcal{L}\left(f_{\theta_t}(x^*), y^*\right)\right) + \alpha_{\text{reg}} \cdot \mathcal{R}(x^*)$
- d is distance, typically l_1, l_2 or cosine.
- $\mathcal{R}$ is a prior based on domain-specific knowledge.
- Optimization is done via GD.
- y^* is recovered separately (out of scope).
- For each categorical feature create an N-dim. variable that gets put into x^*

through softmax.
For tables, we can use entropy over many randomly initialized reconstructions as a prior, because correct cells are robust to random initializations.

FedAVG: Client runs E epochs of SGD, sends new weights to server. Final weights depend on order of batches, the server does not know it. Attack simulates training. Prior: the average of samples in one epoch is equal to that in another epoch.

6.2 Differential Privacy

MI protection: $\mathbb{P}(M(\mathcal{D}) \in S) \approx \mathbb{P}(M(\mathcal{D}') \in S)$

ε-DP: M is ε-DP if for all "neighboring" (a, a') and for any attack S $p(a) := \mathbb{P}(M(a) \in S) \leq e^\varepsilon \mathbb{P}(M(a') \in S)$.
As $e^\varepsilon \approx 1 + \varepsilon$, $(1 - \varepsilon)p(a') \approx p(a) \approx (1 + \varepsilon)p(a')$.
By a theorem, $f(a) + \operatorname{Lap}(0, \Delta_1/\varepsilon)$ is ε-DP, where $\Delta_p := \max_{(a,a') \in \text{ Neigh}} \|f(a) - f(a')\|_p$.

ε, δ-DP: M is ε, δ-DP iff $\mathbb{P}(M(a) \in S) \leq e^\varepsilon \mathbb{P}(M(a') \in S) + \delta \ \forall (a, a') \in \text{Neigh}, \forall S$. This allows absolute differences (not only relative). If $p(a') = 0, p(a) \neq 0$, no ε-DP mechanism exists, but ε, δ-DP might.
If output set is discrete, singleton attacks are enough. $f(a) + \mathcal{N}(0, \sigma^2 I)$ is ε, δ-DP, where $\sigma = \sqrt{2 \log(1.25)/\delta} \cdot \Delta_2/\varepsilon$.

Composition: If M_1, M_2 are ε_1, δ_1-DP and ε_2, δ_2-DP, then (M_1, M_2) and $M_1 \circ M_2$ are $\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2$-DP. In particular, if f is a plain function $(0, 0)$-DP, then $f \circ M$ is ε, δ-DP. If A_i has user data and M_i is $(\varepsilon_i, \delta_i)$-DP, $M_1(a_1) \ldots M_k(a_k)$ is $(\max_i \varepsilon_i, \max_i \delta_i)$ -DP.

DP-SGD: Project gradients for each point onto l_2-ball of size C and sum them up. Add $\mathcal{N}(0, \sigma^2 I)$ to the batch gradient, where $\sigma = \sqrt{2 \log(1.25)/\delta} \cdot C/L/\varepsilon$ The resulting model is private, even against a white-box attacker with any number of queries. Clipping is necessary to bound the sensitivity of the gradient.

Privacy Amplification: Applying an (ε, δ)-DP mechanism on a random fraction $q = L/N$ subset yields a $(\tilde{q}\varepsilon, q\delta)$-DP mechanism, where $\tilde{q} \approx q$.
Due to clipping, sensitivity of the gradient for any point is C. If $T = 1$ and no subsampling is used, adding/removing a datapoint changes total gradient by at most C/L. Then by the gaussian mechanism the resulting model is ε, δ-DP. If subsampling is used, by privacy amplification, the model is $(\tilde{q}\varepsilon, q\delta)$-DP. If $T \neq 1$, by the composition theorem, the model is $(\tilde{q}T\varepsilon, qT\delta)$-DP. By out of scope theorems, this is $\left(\mathcal{O}\left(q\varepsilon\sqrt{T \log \frac{1}{\delta}}\right), \mathcal{O}(qT\delta)\right)$ and $\left(\mathcal{O}\left(q\varepsilon\sqrt{T}\right), \delta\right)$ -DP.

PATE: Private Aggregation of Teacher Models: Split data into disjoint partitions and train a model for each. Agreggate models via noisy voting into a teacher, which labels public unlabeled data, on which we train the final model.
T are teachers, $n_j(x) := |\{t(x) = j \texttt{ mid } t \in T\}|$. $\arg\max(n_j(x)) + \operatorname{Lap}(0, \sigma)$ is bad, better $\arg\max(n_j(x) + \operatorname{Lap}(0, 2/\varepsilon))$. $\Delta_1 = 2 \Rightarrow$ model is $(\varepsilon, 0)$-DP for one query. Labeling T data points yields $(\varepsilon T, 0)$-DP. But there are better bounds.

FedSGD/FedAVG with Noise: clip the gradients/ weights and add noise.
DP is closely related to randomized smoothing. We add noise to data, then forward is ε-DP.

7. AI Regulation

Key issues: fairness, explainability, data minimization, unlearning (right to be forgotten), copyright.

8. Private synthetic data

Data is private, make DP synthetic proxy.
1. **Select** marginal queries we want to measure
2. **Measure** marginal queries using DP
3. **Generate** synthetic data

Marginal: **Marginal** on $C \subseteq \mathcal{A}$ (attrs.) is a vector $\mu \in \mathbb{R}^{n_C}$, indexed by $t \in \Omega_C$, where $\Omega_C = \prod_{i \in C} \Omega_i$ and $n_C = |\Omega_C|$. Each entry μ_t is a count $\sum_{x \in D}[x_C = t]$. $M_C : \mathcal{D} \to \mathbb{R}^{n_C}, D \mapsto \mu$ computes the marginal.
$\Delta_2(M_C) = 1$ because adding a row in a dataset can only change one element of the vector. 1-way marginals ($n_C = 1$) are histograms, 2-way marginals are heatmaps.

Chow-Liu: Mutual information of two variables X, Y is $\mathrm{I}(X,Y) = \sum_{x,y} \frac{p(x,y)}{p(x)p(y)}$. Chow-Liu algorithm makes a complete graph of features, edge weigths $\mathrm{I}(X,Y)$. Find MST, the optimal 2nd-order approximation. Generate by sampling from MST, each node is conditioned on its parent, i.e. $p(F_1 = f_1, F_2 = f_2, F_3 = f_3) = p(F_1 = f_1)p(F_2 = f_2 \texttt{ mid } F_1 = f_1)p(F_3 = f_3 \texttt{ mid } F_1 = f_1)$, if F_1 is parent of F_2 and F_3.
Add DP, i.e. add noise to every step of the algorithm. MST is done with the exponential mechanism, marginals are measured with Gaussian noise.

ProgSyn:

- Sample random noise $z \sim \mathcal{N}(0, I_p)$
- Pass z through a generative model g_θ
- Get synthetic dataset $g_\theta(z)$
- Adapt θ to make $g_\theta(z)$ close to original X
- Fine-tune g_θ to make $g_\theta(z)$ satisfy constraints

9. Logic and Deep Learning (DL2)

9.1 Querying Neural Networks

Standard Logic: Use standard logic ($\forall, \exists, \wedge, \vee, f : \mathbb{R}^m \to \mathbb{R}^n, ..$) and high-level queries to impose constraints.
$(\mathrm{class}(\mathrm{NN}(i)) = 9) = \bigwedge_{j=1, j \neq 9}^{k} \mathrm{NN}(i)[j] < \mathrm{NN}(i)[9]$
Use translation T of logical formulas into differentiable loss function $T(\varphi)$ to be solved with gradient-based optimization to minimize $T(\varphi)$. Regular SAT solvers can't handle non-small NNs.

Theorem: $\forall x, T(\varphi)(x) = 0 \Leftrightarrow x \vDash \varphi$

Logical Formula to Loss:

Logical Term	Loss
$t_1 \leq t_2$	$\max(0, t_1 - t_2)$
$t_1 \neq t_2$	$[t_1 = t_2]$
$t_1 = t_2$	$T(t_1 \leq t_2 \wedge t_2 \leq t_1)$
$t_1 < t_2$	$T(t_1 \leq t_2 \wedge t_1 \neq t_2)$
$\varphi \vee \psi$	$T(\varphi) \cdot T(\psi)$
$\varphi \wedge \psi$	$T(\varphi) + T(\psi)$

By construction $T(\varphi)(x) \geq 0, \forall x, \varphi$. Negation can be implemented by using de Morgan's laws.

Box constraints: hard to enforce in GD. Use L-BFGS-B and give box constraints to optimizer.

9.2 Training NN with Background Knowledge

Problem statement: Enforce logical property φ when training NN. find θ that maximizes the expected value of property $\mathbb{E}_{s \sim D}[\forall z. \varphi(z, s, \theta)]$.
BUT: Universal quantifiers are difficult.

Reformulation: get the worst violation of φ and minimize its effect, i.e. $\mathbb{E}_{s \sim D}[\max_z \neg\varphi(z, s, \theta)]$.
Reform. 2: minimize $\mathbb{E}_{s \sim D}[T(\varphi)(bz, s, \theta)]$, where $bz = \mathrm{argmin}(T(\neg\varphi)(z, s, \theta))$. This is an adv. attack. $\exists$ different bz which minim. $T(\neg\varphi)$ which can produce different $T(\varphi)$. $bz \neq$ worst example.
Restrict z to a convex set with efficient projs., i.e. L_∞ -balls. Remove the constraint from φ that restricts z on the convex set and do PGD while projecting z onto the convex set.

10. Fairness

Individual Fairness: A mapping $M : \mathcal{X} \to \Delta(\mathcal{Y})$ is (D, d)-**Lipschitz**, if for every $x_1, x_2 \in \mathcal{X}$ $D(M(x_1), M(x_2)) \leq d(x_1, x_2)$. If M is a model, it's **individually fair** wrt. D and d. d is a distance in feature space, D is a metric on probability distributions. Choosing metrics is hard.
Lemma: 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 in x.

Lipschitz Property: Let $L \in \mathbb{R}$ be s.t. $D(M(x), M(x')) \leq L d(x, x')$ (smaller value is stronger). Let $d(x, x') := (x - x')^\top S(x - x')$, where S is a symmetric positive definite covariance matrix. Let $D(M(x), M(x')) := [M(x) \neq M(x')]$. Then the Lipschitz property is equivalent to $\forall \delta \in \mathbb{B}_S(0, 1/L) \quad M(x) = M(x + \delta)$, where $\|x\|_S := \sqrt{x^\top S x}$. We have reformulated individual fairness as robustness.

10.1 Fair Representation Learning

FRL: FRL is often more efficient (reuse fair data) and simplifies audits. But it has less precise control of the fairness/performance tradeoff, is susceptible to adv. attacks by the consumer, can be expensive and provides no certification.

10.2 Learning Certified Individually Fair Representations

LCIFR: Keep pros of FRL, but also allow the regulator to certify the fairness of the E2E model and allows to define D and d via logical constraints that are accepted by MILP and DL2. Example: $d(x, x') = \bigwedge_{i \in \text{ Cat } \setminus\{\text{race, gender}\}}(x_i = x_{i'}) \bigwedge_{j \in \text{ Num}} |x_j - x_{j'}| \leq \alpha$. Logic captures cat. features exactly, norms don't.
Let $S_d(x)$ denote the set of all points similar to x and assume $D(M(x), M(x')) = [M(x) \neq M(x')]$.
The encoder $f_\theta : \mathbb{R}^n \to \mathbb{R}^k$ is trained using DL2 s.t. $\forall x' \in S_d(x) \quad \|f_\theta(x) - f_\theta(x')\|_\infty \leq \delta$. $S_d(x)$ is a complicated set, which we bound by a box in latent space. The producer encodes $S_d(x)$ and f_θ as MILP to compute ε s.t. $f_\theta(S_d(x)) \subseteq \{z' \texttt{ mid } \|f_\theta(x) - z'\|_\infty \leq \varepsilon\}$, which gives the consumer a simple robustness problem.
Train encoder using training with background knowledge with classifier to keep latent space useful. Train decoder via randomized smoothing.

10.3 Latent Space Smoothing for individually Fair Representations

LSS: Use semantic feature space from a good gen. model encoder for similarity formulas for images etc. Center smoothing produces a bound on the radius of the ball in latent space. The E2E model is individually fair with probability $1 - \alpha_{\text{rs}} - \alpha_{\text{cs}}$.

10.4 Group Fairness

Definitions: **Demographic parity**: $\mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } G = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } G = 1\right)$, where G is a group feature.
Equal opportunity: $\mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } Y = 1, G = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } Y = 1, G = 1\right)$
Equalized odds: Equal opportunity and $\mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } Y = 0, G = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } Y = 0, G = 1\right)$

Postprocessing: Example of **postprocessing**: for a binary classifier with output probability $h(x)$. Use separate thresholds for each group, tuned to achieve group fairness.

In-training: Example of **in-training**: add relaxed fairness constraints that are solved with DL2, i.e. $-\varepsilon \leq \mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } s = 0\right) - \mathbb{P}\left(\hat{Y} = 1 \texttt{ mid } s = 1\right) \leq \varepsilon$

Preprocessing: FRL: Notation: data $(x, s) \in \mathbb{R}^d \times \{0,1\}$, encoder $f : \mathbb{R}^d \times \{0,1\} \to \mathbb{R}^{d'}, z = f(x, s)$, classifier $g : \mathbb{R}^{d'} \to \{0,1\}$, adversary $h : \mathbb{R}^{d'} \to \{0,1\}$ is a classifier that tries to predict the sensitive attribute from data in the latent space, $Z_i := \{z \texttt{ mid } s = i\}, p_i(z) := \mathbb{P}(z \texttt{ mid } s = i)$.

LAFTR: jointly train f, g and h. No guarantees. $\min_{f,g} \max_h(\mathcal{L}_{\text{clf}}(f(x,s), g) - \gamma \mathcal{L}_{\text{adv}}(f(x,s), h))$
Use adversary to add guarantees by computing an upper bound on unfairness of any g. Convert hard constraint (DP, EO) into a soft measure, e.g. for demographic parity: $\Delta_{Z_0,Z_1}(g) := |\mathbb{E}_{z \sim Z_0} g(z) - \mathbb{E}_{z \sim Z_1} g(z)|$, lower is better. Balanced accuracy is $\mathrm{BA}_{Z_0,Z_1}(h) = \frac{1}{2}\left(\mathbb{E}_{z \sim Z_0}(1 - h(z)) + \mathbb{E}_{z \sim Z_1} h(z)\right) = \frac{1}{2}\int_Z (p_0(z)(1 - h(z)) + p_1(z)h(z))$, h chooses p_0 or p_1. The optimal adversary is $h^*(z) := [p_1(z) \geq p_0(z)]$. Theorem: $\Delta_{Z_0,Z_1}(g) \leq 2 \cdot \mathrm{BA}_{Z_0,Z_1}(h^*) - 1$. We can't find neither BA nor h^* exactly.

Fair Normalizing Flows: sample x from a known distribution q, apply an invertible encoder $z = f(x)$, find density of the new distribution by $\log p(z) = \log q(f^{-1}(z)) + \log|\det \frac{\partial f^{-1}(z)}{\partial z}|$. Learn normalizing flows f_0 and f_1 as encoders for Z_0 and Z_1. This lets us find $p_0(z)$ and $p_1(z)$, given $q_0(x), q_1(x)$. They can be estimated with density estimation, e.g. Gaussian Mixture Model. Given $p_0(z), p_1(z)$, we estimate an UB of BA with probability $1 - \varepsilon$ by Hoeffding's inequality, and then apply the theorem for UB of Δ.
For good bounds, need low accuracy of $h^* \Rightarrow$ low dist. between Z_0 and Z_1. Add KL divergence between p_0 and p_1 (and $\mathrm{KL}(p_1, p_0)$) to loss of g. g will be thrown away after training, as it exists only to increase utility of the flows.

The bound holds only when the q estimates are accurate, which is a major limitation.

Fairness with Restricted Encoders: restrict the space of representations to be finite. This allows to get the distribution of sensitive attributes at each z, hence we have $p_i(z)$. First, we bound $P(s = i)$ using binom. conf. intervals, then per-cell balanced accuracy, then BA. This is done on different datasets to achieve independence.

Appendix

De Morgan: $\neg(\varphi \wedge \psi) = \neg\varphi \vee \neg\psi$; $\neg(\varphi \vee \psi) = \neg\varphi \wedge \neg\psi$

Ball Relations: $\mathbb{B}_\varepsilon^1 \subseteq \mathbb{B}_\varepsilon^2 \subseteq \mathbb{B}_\varepsilon^\infty \subseteq \mathbb{B}_{\varepsilon\sqrt{d}}^2 \subseteq \mathbb{B}_{\varepsilon \cdot d}^1$

Jensen: g convex: $g(E[X]) \leq E[g(X)]$

Bayes: $P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y|X)P(X)}{P(Y)}$

Matrix Inverse: $A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

Norms: $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{1}{p}}$ $\|x\|_\infty = \max_{i \in \{1,..,d\}} |x_i|$

Softmax: $\sigma(z)_i = e^{z_i} / \sum_{j=1}^D e^{z_j}$

CE loss: $\mathrm{CE}(\vec{z}, y) = -\sum_{c=1}^K \mathbb{1}[c = y] \cdot \log z_c$

Implication: $\varphi \Rightarrow \psi \Leftrightarrow \neg\varphi \vee \psi$

Gauss: $\mathcal{N} = \frac{1}{\sqrt{(2\pi)^d\ |\Sigma|}} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu))$ **CDF**: $\Phi(v; \mu, \sigma^2) = \int_{-\infty}^v \mathcal{N}(y; \mu, \sigma^2) dy = \Phi\left(\frac{v - \mu}{\sqrt{\sigma^2}}; 0, 1\right)$

Laplace: $\mathcal{L} = \frac{1}{2b}\exp\left(-\frac{|x - \mu|}{b}\right)$, $\Phi(x; \mu, b) = 0.5 + 0.5\ \mathrm{sgn}(x - \mu)\left(1 - \exp\left(-\frac{|x - \mu|}{b}\right)\right)$

Subadditivity of $\sqrt{\cdot}$: $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$

Cauchy Schwarz: $\langle x, y \rangle \leq \|x\|_2 \cdot \|y\|_2$

Hölder's: $\|x \cdot y\|_1 \leq \|x\|_p \cdot \|y\|_q$, if $\frac{1}{p} + \frac{1}{q} = 1$

Minmax: $\max_a \min_b f(a,b) \leq \min_b \max_a f(a,b)$

Variance & Covariance: $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\mathrm{Cov}(X, Y)$
$\mathbb{V}(AX) = A\mathbb{V}(X)A^T, \mathbb{V}[\alpha X] = \alpha^2 \mathbb{V}[X]$
$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Distributions: $\mathrm{Exp}(x|\lambda) = \lambda e^{-\lambda x}$, $\mathrm{Ber}(x|\theta) = \theta^x(1 - \theta)^{(1-x)}$
Sigmoid: $\sigma(x) = 1/(1 + e^{-x})$
$a\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(a\mu_1 + \mu_2, a^2\sigma_1^2 + \sigma_2^2)$

Normal CDF: $x \sim \mathcal{N}(0,1) \Rightarrow \mathbb{P}(x \leq z) = \Phi(z), \mathbb{P}(x \leq \Phi^{-1}(z)) = z$. $x \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \mathbb{P}(x \leq z) = \Phi\left(\frac{z - \mu}{\sigma}\right), \mathbb{P}(x \leq \mu + \sigma\Phi^{-1}(z)) = z$

Chebyshev & Consistency: $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{V}[X]}{\varepsilon^2}$, $\lim_{n \to \infty} \mathbb{P}(|\hat{\mu} - \mu\ | > \varepsilon) = 0$

Derivatives: $(fg)' = f'g + fg'$; $(f/g)' = (f'g - fg')/g^2$
$f(g(x))' = f'(g(x))g'(x)$; $\log(x)' = 1/x$
$\partial_x \boldsymbol{b}^\top \boldsymbol{x} = \partial_x \boldsymbol{x}^\top \boldsymbol{b} = \boldsymbol{b}$, $\partial_x \boldsymbol{x}^\top \boldsymbol{x} = \partial_x \|\boldsymbol{x}\|_2^2 = 2\boldsymbol{x}$,

$\partial_x \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}^\top + \boldsymbol{A})\boldsymbol{x}, \partial_x(\boldsymbol{b}^\top \boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{b}$,
$\partial_X(\boldsymbol{c}^\top \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{c}\boldsymbol{b}^\top, \partial_X(\boldsymbol{c}^\top \boldsymbol{X}^\top \boldsymbol{b}) = \boldsymbol{b}\boldsymbol{c}^\top$,
$\partial_x(\| \boldsymbol{x} - \boldsymbol{b} \|_2) = \frac{\boldsymbol{x}-\boldsymbol{b}}{\|\boldsymbol{x}-\boldsymbol{b}\|_2}, \partial_X(\|\boldsymbol{X}\|_F^2) = 2\boldsymbol{X}$,
$\partial_x \|\boldsymbol{x}\|_1 = \frac{\boldsymbol{x}}{|\boldsymbol{x}|}$, $\partial_x \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \boldsymbol{2}(\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^\top \boldsymbol{b})$,

MILP encodings: $y = |x|, l \leq x \leq u$: $y \geq x, y \geq -x$,
$y \leq -x + a \cdot 2u, y \leq x - (1-a) \cdot 2l, a \in \{0,1\}$
$y = \max(x_1, x_2), l_1 \leq x_1 \leq u_1, l_2 \leq x_2 \leq u_2$:
$y \geq x_1, y \geq x_2, y \leq x_1 + a \cdot (u_2 - l_1)$,
$y \leq x_2 + (1-a) \cdot (u_1 - l_2), a \in \{0,1\}$