CS182/282A: Designing, Visualizing and Understanding Deep Neural Networks

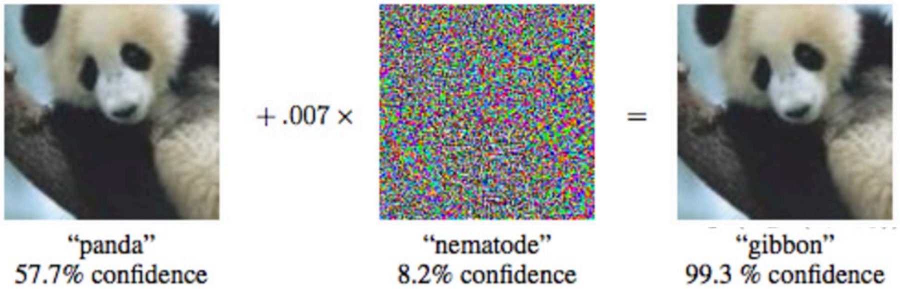
John Canny

Spring 2020

Lecture 19: Fairness in Deep Learning

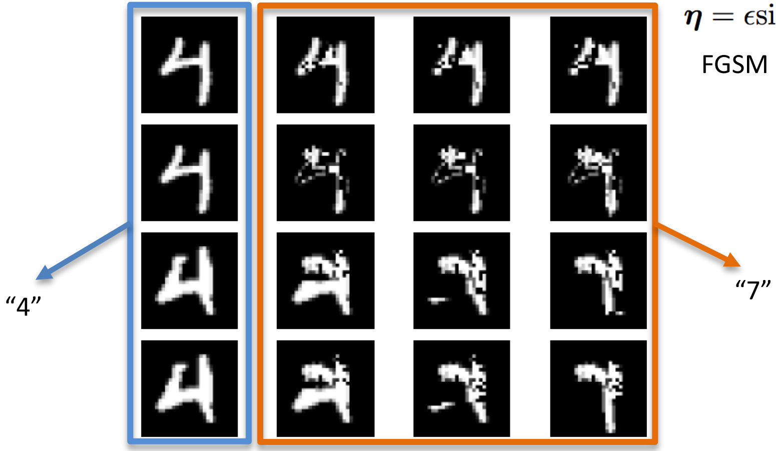
John Canny

Last Time: Adversarial Examples



$$\boldsymbol{\eta} = \epsilon \text{sign}(\nabla_x J(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}))$$

FGSM attack



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *ICLR 2015.*

Bo Li, Yevgeniy Vorobeychik, and Xinyun Chen. "A General Retraining Framework for Scalable Adversarial Classification." ICLR. (2016).

Last Time: Optimizing an Adversarial Objective

- Fast approaches
  - Fast gradient sign ($d = ||\cdot||_\infty$): $x^* = x + B\,\mathrm{sgn}\big(\nabla_x \ell(f_\theta(x), y)\big)$
  - Fast gradient ($d = ||\cdot||_2$): $x^* = x + B\left(\frac{\nabla_x \ell(f_\theta(x), y)}{||\nabla_x \ell(f_\theta(x), y)||_2}\right)$
- Iterative approaches
  - E.g., use a SGD optimizer, such as Adam, to optimize
- $\max_{x^*} \ell(f_\theta(x^*), y) + \lambda d(x, x^*)$

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

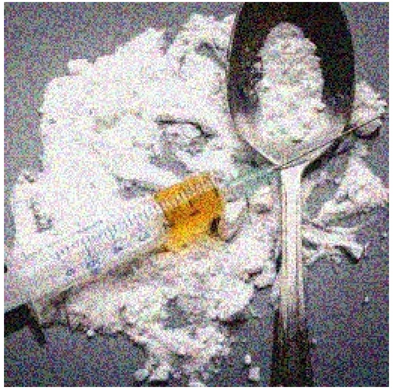
- Optimization
- **Need to know model f_θ**

Last Time: Black-box Attacks

Black-box attacks are possible on deep neural networks with query access.

The number of queries needed can be reduced

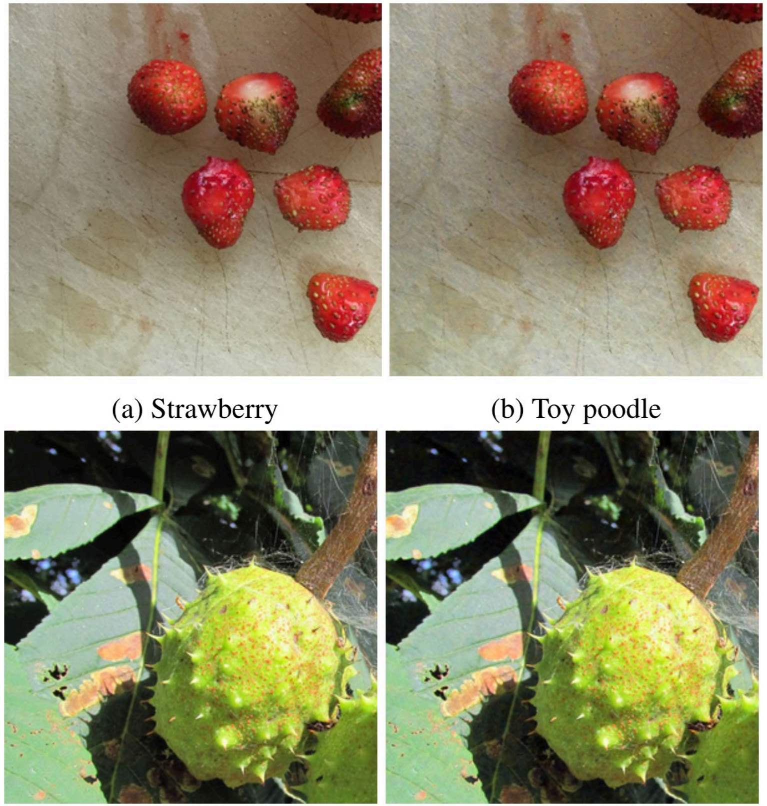
Original image, classified as “drug” with a confidence of 0.99



Adversarial example, classified as “safe” with a confidence of 0.96

The Gradient Estimation black-box attack on Clarifai’s Content Moderation Model

Last Time: Adversarial Examples from Adversarial Nets



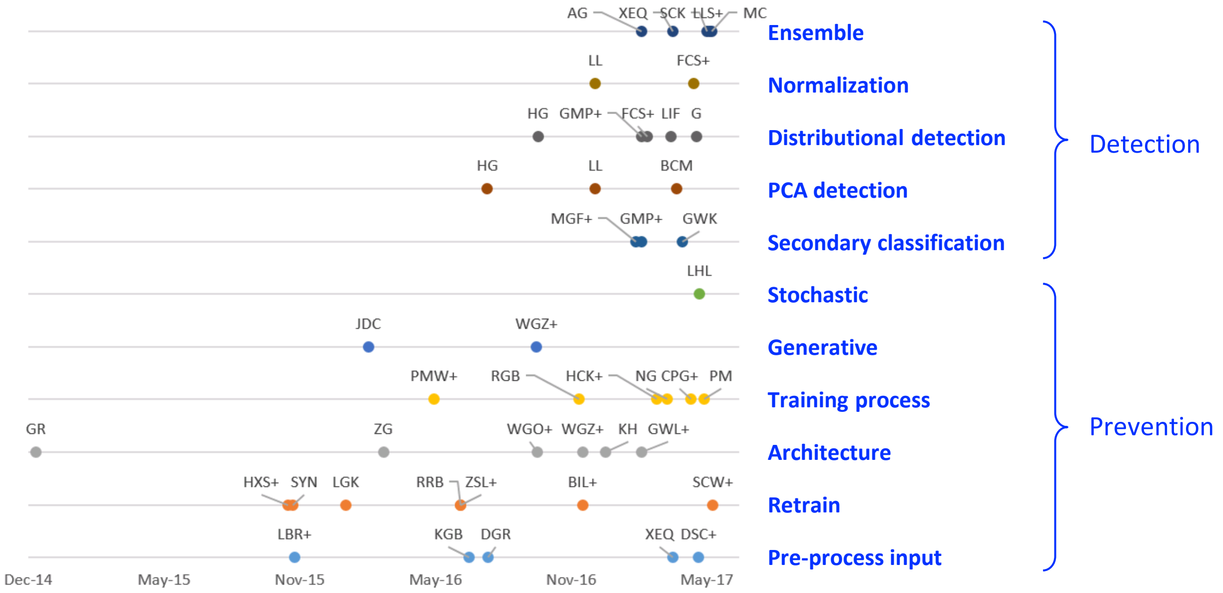
(a) Strawberry

(b) Toy poodle

(c) Buckeye

(d) Toy poodle

Last Time: Detecting/Defending Against Adversarial Examples



Updates

First Quiz is next week Mon-Thursday.

Open-book, designed to take about 90 minutes: Its due for everyone at the same time.

Covers material from first Midterm through Spring Break.

JFC offices hour this week will be Thursday 5-6pm.

This Time: Fairness in Deep Learning

From the United Nations' *Universal Declaration of Human Rights:*

> **Article 2:** Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status…

Algorithms and Human Rights

Housing: Algorithmic credit scoring

Employment: AI-mediated talent acquisition and job search

Sentencing and Parole: Recidivism prediction

Safety: Insurance risk and pricing

Education: Standardized testing evaluation

Legally Protected Characteristics:

Race

Color

Sex

Religion

National Origin

Citizenship

Age

Pregnancy

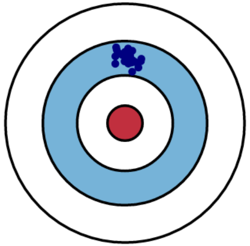
Familial Status

Disability Status

Veteran Status

Genetic Information

Deep Networks and Bias

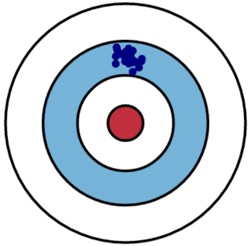


Deep nets, like other machine learning methods, are subject to bias.

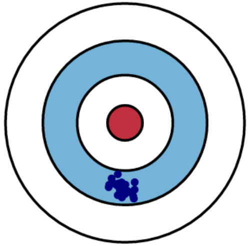
i.e. predictions systematically depart from the exact, population values.

So it seems very possible that algorithms could discriminate if average bias (e.g. for predicting credit-worthiness) were higher for one group than another:

The Universal Approximation Theorem for deep nets implies that we can reduce bias to arbitrarily small values. Is this enough?



Prediction for Americans



Prediction for Australians

Is Universal Approximation Enough?

No, for several reasons: the data we train on can be flawed:

- **Sampling bias:** How the training data were collected. Often caused when certain populations, e.g. Australians, are under-represented in the sample.

- **Selection bias:** Includes and extends sampling bias. May include working with subsets of the data, choosing subsets that confirm a hypothesis etc.

- **Reporting bias:** Subjects may not reveal protected characteristics, or distort them.

Is Universal Approximation Enough?

No, for several reasons: the model we train can be imperfect:

- **Universal Approximation is a Theoretical Ideal:** We don’t want to drive bias to zero because variance would explode. Instead we balance bias and variance on finite training data which means we live with a certain amount of bias in our models.
- **Inductive Bias:** We deliberately design models with characteristics that allow them to fit the data well/train quickly: latent dimensions, residual links, attention links.

Toward Fairness:

It seems that we have to live with a certain amount of bias in ML algorithms.

But can we at least minimize the effects on protected characteristics like race, gender etc?

i.e. Can we define, measure and maximize the ***Fairness*** of Deep Network learning algorithms?

A First Cut at Fairness: Unawareness

Let X be a set of features of individuals, A a protected attribute, and Y an outcome variable, and $\hat{Y}$ a predictor for Y. e.g.

X = Owns home in US (or not), income, debt, number of cars.

A = National Origin.

Y = Received a loan (from US bank) or didn't, $\{1,0\}$.

$\hat{Y} = f(X, A)$ = Predictor of a loan (assume we already trained it).

Unawareness: we force the predictor to be unaware of A, i.e. $f(X, A) = f(X)$.

The problem is that there might still be features in X which are very predictive of A, and f may not need to see A to make predictions which are very unfair.

A Second Cut at Fairness: Demographic Parity

Let X be a set of features of individuals, A a protected attribute, and Y an outcome variable, and $\hat{Y}$ a predictor for Y. e.g.

X = Owns home in US (or not), income, debt, number of cars.

A = National Origin.

Y = Received a loan (from US bank) or didn't, $\{1,0\}$.

$\hat{Y} = f(X, A)$ = Predictor of a loan (assume we already trained it).

We can define **demographic parity** as

$$\Pr\left(\hat{Y} = 1 \middle| A = \text{Australian}\right) = \Pr\left(\hat{Y} = 1 \middle| A = \text{American}\right)$$

But is this fair?

Demographic Parity and Independence

When we define **demographic parity** as

$$\Pr\left(\hat{Y} = 1 \middle| A = \text{Australian}\right) = \Pr\left(\hat{Y} = 1 \middle| A = \text{American}\right)$$

It implies

$$\Pr\left(\hat{Y} = 0 \middle| A = \text{Australian}\right) = \Pr\left(\hat{Y} = 0 \middle| A = \text{American}\right)$$

So in fact $\Pr(\hat{Y})$ is *independent* of the conditioning on A, written $\hat{Y} \perp\!\!\!\perp A$

or

$$\Pr\left(\hat{Y}, A\right) = \Pr\left(\hat{Y}\right) \Pr(A)$$

Demographic Parity is Intuitively Unfair

Demographic parity:

$$\Pr\left(\hat{Y} = 1 \middle| A = \text{Australian}\right) = \Pr\left(\hat{Y} = 1 \middle| A = \text{American}\right)$$

Equalizes the probability of getting a loan among the two groups. But it doesn't involve the admissible attributes.

Its likely that among any group of Australians seeking a loan, many do not reside in the US (no home), do not earn income in the US, may not own a car in the US.

An American who did not have a US home, did not have any US income, and did not have a car in the US, would be low in the ranking of Americans and therefore unlikely to get a loan. An Australian with those same traits would be average in the rankings of Australians, and much more likely to get a loan.

So in fact giving loans to Australians for the sake of Demographic parity would be *unfair*.

A Fairness Principle: “All Other Things Being Equal”

We want to treat people with *similar acceptable attributes* the same.

e.g. two individuals of different race or gender, *but who have achieved a similar level of performance in a company*, *should be equally likely to receive a promotion and the same salary*.

So we can think about hypothetical people who did not occur in our dataset, and who had a different value of a protected attribute as a real individual, and force the same outcome.

This is called *counter-factual* or *potential outcome* reasoning.

More generally, this is part of *Causal Inference*, which studies the effects of *changes to a probabilistic model*, such as manipulating race or gender.

A Third Try: Separation

Hardt et al., “Equality of Opportunity in Supervised Learning” NIPS 2016

Idea is to make $\hat{Y}$ (loan prediction) independent of A (national origin), while “keeping all other things equal”.

Actually “keeping all other things equal” is too strong. What we really mean is: “keeping things equal that directly affect Y, the true loan choice”, or “among people who are equally worthy of a loan”. The simplest way to do this is to condition on Y.

Separation (also called Equalized Odds): $(\hat{Y} \perp\!\!\!\perp A) \mid Y$ i.e. $\hat{Y}$ and A are *conditionally independent given* Y or

$$\Pr(\hat{Y}, A \middle| Y) = \Pr(\hat{Y} \middle| Y) \Pr(A|Y)$$

Which means

$$\Pr(\hat{Y}, A \middle| Y = y) = \Pr(\hat{Y} \middle| Y = y) \Pr(A|Y = y) \quad \text{for all} \quad y \in Y$$

Separation

Hardt et al., "Equality of Opportunity in Supervised Learning" NIPS 2016

Separation: $(\hat{Y} \perp\!\!\!\perp A) \mid Y$ i.e. $\hat{Y}$ and A are *conditionally independent given* Y or

$$\Pr\left(\hat{Y}, A \middle| Y\right) = \Pr\left(\hat{Y} \middle| Y\right) \Pr(A|Y)$$

- But what happened to dependence on the acceptable attributes X ?
- The "true" outcome Y is already assumed to depend (fairly) on X, so that's all we need.

Separation

Hardt et al., “Equality of Opportunity in Supervised Learning” NIPS 2016

Separation: $(\hat{Y} \perp\!\!\!\perp A) \mid Y$ i.e. $\hat{Y}$ and A are *conditionally independent given* Y or

$$\Pr\left(\hat{Y}, A \middle| Y\right) = \Pr\left(\hat{Y} \middle| Y\right) \Pr(A|Y)$$

- Note that we are acknowledging that our predictor $\hat{Y} = f(X, A)$ has some bias. If it didn’t, the condition would be satisfied automatically. Why?
- If we had a perfect predictor, $\hat{Y} = Y$, so $\Pr\left(\hat{Y} \middle| Y = y\right)$ would be a point distribution, and the product (independence) formula holds automatically.

Separation

Hardt et al., “Equality of Opportunity in Supervised Learning” NIPS 2016

Separation: $(\hat{Y} \perp\!\!\!\perp A) \mid Y$ i.e. $\hat{Y}$ and A are *conditionally independent given* Y or

$$\Pr(\hat{Y}, A | Y) = \Pr(\hat{Y} | Y) \Pr(A|Y)$$

- This criterion groups together people according to the value of Y, the loan assignment from the training data. Couldn’t that already be biased?
- Yes, but we don’t have control over that, nor another ground truth against which to measure bias in Y. We instead want to make sure that our prediction algorithm $\hat{Y} = f(X, A)$ doesn’t add bias against protected groups.

A Variation: Equal Opportunity

Hardt et al., “Equality of Opportunity in Supervised Learning” NIPS 2016

Equal Opportunity (for binary Y): $(\hat{Y} \perp\!\!\!\perp A) \mid Y = 1$ i.e. $\hat{Y}$ and A are *conditionally independent given* $Y = 1$ or

$$\Pr\left(\hat{Y}, A \middle| Y = 1\right) = \Pr\left(\hat{Y} \middle| Y = 1\right) \Pr(A|Y = 1)$$

- We want our predictor to be fair on the set of people worthy of loans.
- We don’t enforce independence on the group not worthy of loans, and this would normally allow the model to be fairer to the group who are worthy of loans.

Fourth Try: Sufficiency

Sufficiency means $(Y \perp\!\!\!\perp A) \mid \hat{Y}$ i.e. Y and A are *conditionally independent given* $\hat{Y}$ or

$$\Pr(Y, A|\hat{Y}) = \Pr(Y|\hat{Y}) \Pr(A|\hat{Y})$$

- So its really the dual of equalized odds (flip the roles of Y and $\hat{Y}$).
- i.e. the actual label and the protected attribute are independent, given the predicted label.

- Sounds reasonable, but you couldn’t be blamed at this point if your head is spinning from all these measures of fairness.

Summary: Four Notions of Fairness

Acceptable attributes X, protected attribute A, actual outcome Y, and predicted outcome $\hat{Y} = f(X, A)$.

1. **Unawareness:** $\hat{Y} = f(X, A) = f(X)$
2. **Demographic Parity:** $\hat{Y} \perp\!\!\!\perp A$
3. **Separation:** $\left(\hat{Y} \perp\!\!\!\perp A\right) \mid Y$
4. **Sufficiency:** $(Y \perp\!\!\!\perp A) \mid \hat{Y}$

Summary: Four Notions of Fairness

1. **Unawareness:** $\hat{Y} = f(X, A) = f(X)$

2. **Demographic Parity:** $\hat{Y} \perp\!\!\!\perp A$

3. **Separation:** $\left(\hat{Y} \perp\!\!\!\perp A\right) \mid Y$

4. **Sufficiency:** $(Y \perp\!\!\!\perp A) \mid \hat{Y}$

We don’t like the first two. We do like the last two, but its not clear which is better. Could we try to prove both?

Incompatibility of Separation and Sufficiency

Theorem: *If A and Y are not statistically independent*, then *separation and sufficiency cannot both hold*.

So any dependence of the true value Y on A (which would usually be the case), then only one of Separation/Sufficiency condition can hold.

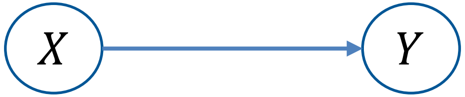
Beyond Observational Measures…

All the measures so far are *observational*, i.e. they depend on data observed from the model without changing it.

But truly measuring the affect of protected attributes would involve *changing those attributes* to see if it changes the outcome. E.g. doing *counter-factual (causal) reasoning* by changing the sex, race, age etc. of a particular individual.

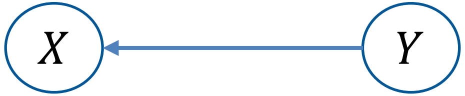
A Very Brief Overview of Causal Reasoning

In probabilistic modeling, we use graphical models to represent dependencies. If we draw this diagram, we imply a factorization:



$$P(X, Y) = P(Y|X)P(X)$$

Or we could have a model:

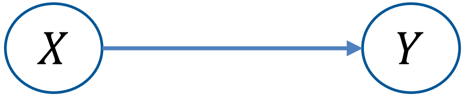


$$P(X, Y) = P(X|Y)P(Y)$$

But both models can represent any joint distribution $P(X, Y)$.

Causal Graphical Models

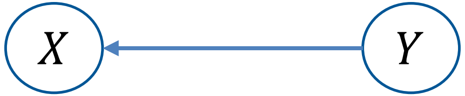
We can interpret the same diagrams as *causal models* by thinking of them as *algorithms* for generating the data:



Sample X first, then Y conditioned on X

$$P(X,Y) = P(Y|X)P(X)$$

Or we could have a model:



Sample Y first, then X conditioned on Y

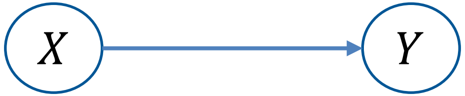
$$P(X,Y) = P(X|Y)P(Y)$$

Both models can still represent any joint distribution $P(X,Y)$.

So we still can't tell between them based on observations of (X,Y).

Causal Graphical Models and do(X=x)

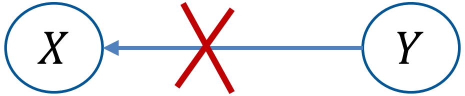
The point of causal models is to study *interventions* (changes to a model) such as by forcing the value of a variable with the $\mathrm{do}(X = x)$ operator.



Suppose we force $X = 0$, then we have

$$P(X,Y) = P(Y|X = 0)\,\delta(X)$$

Or we could have a model:



Sample Y first, then force $X = 0$

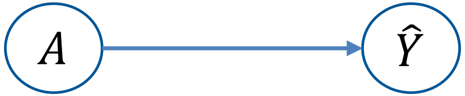
$$P(X,Y) = P(Y)\,\delta(X)$$

The models now represent different distributions from each other and from the original joint distribution $P(X,Y)$.

$\delta(X)$ is the Dirac delta function, which is a probability distribution which is zero for all $X \neq 0$, and "very large" at $X = 0$, such that $\int \delta(X)dX = 1$

Causal Graphical Models and do(X=x)

Its tempting to examine the effect on our model's prediction of manipulating a protected attribute:

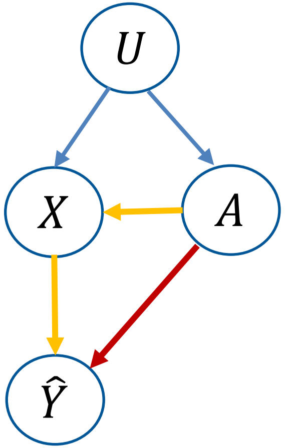


We hope that forcing $A = a$ has no effect

$$P\left(\hat{Y} \mid \mathrm{do}(A = a)\right) = P(\hat{Y} \mid \mathrm{do}(A = a'))$$

But the real graph is not that simple.

Realistically we have something more like:



And we are concerned about direct effects of A on $\hat{Y}$ (red arrow), but also on the indirect effects (orange arrows).

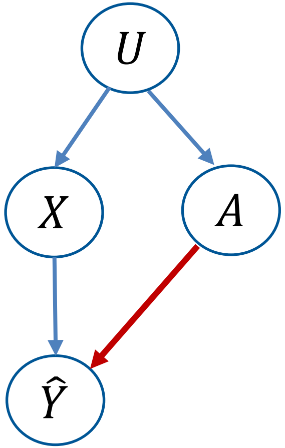
Causal Inference

For the loan/national origin discussion we can simplify the graph: which makes inference easier. Recall that $\hat{Y} = f(X, A)$

Define $\hat{Y}_0 = f(X, 0)$ with X sampled according to its marginal.

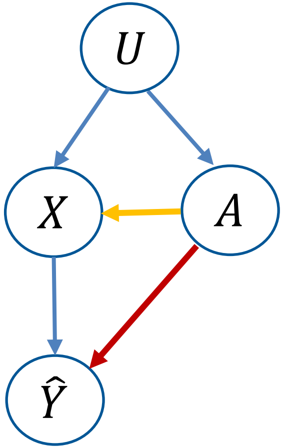
Define $\hat{Y}_1 = f(X, 1)$ with X sampled according to its marginal.

We want $p(\hat{Y}_0) = p(\hat{Y}_1)$.



Domains and Appropriate Attributes

But in general its hard to figure out the dependency of X on A so causal modeling is rarely used.



e.g. Not every **Domain** (e.g. Health Care) is compatible with every **protected attribute** (e.g. sex) because admissible attributes, such as susceptibility to particular diseases or pregnancy, are specific to sex.

GAN-like Optimization

Let's start with a problem we know how to solve. $X \perp\!\!\!\perp Y$ implies

$$P(X,Y) = P(X)P(Y)$$

And we can turn that into a quantitative measure of independence with

$$J(X;Y) = D_{JS}(P(X,Y)||P(X)P(Y))$$

where D_{JS} is the Jensen-Shannon Divergence.

If $X \perp\!\!\!\perp Y$, then $J(X;Y) = 0$, while any dependence will cause $J(X;Y) > 0$.

GAN-like Optimization

Now for demographic parity:

$$J(\hat{Y};A) = D_{JS}\left(P(\hat{Y},A)||P(A)P(\hat{Y})\right)$$

Which suggests a GAN that compares “real” pairs $(\hat{Y},A)$ from the joint distribution $P(\hat{Y},A)$ and a “fake” distribution $(\hat{Y},A')$ where $\hat{Y}$ and A' are (independently) drawn from their marginal distributions $p(\hat{Y})$ and $p(A')$.

The “generator” is just the classifier $\hat{Y} = f(X,A)$ which we’re trying to learn.

The data for the discriminator are real pairs $(\hat{Y},A)$ and fake pairs $(\hat{Y},A')$.

GAN-like Optimization

The “generator” is just the classifier $\hat{Y} = f(X, A)$ which we’re trying to learn.

The data for the discriminator are real pairs $(\hat{Y}, A)$ and fake pairs $(\hat{Y}, A')$.

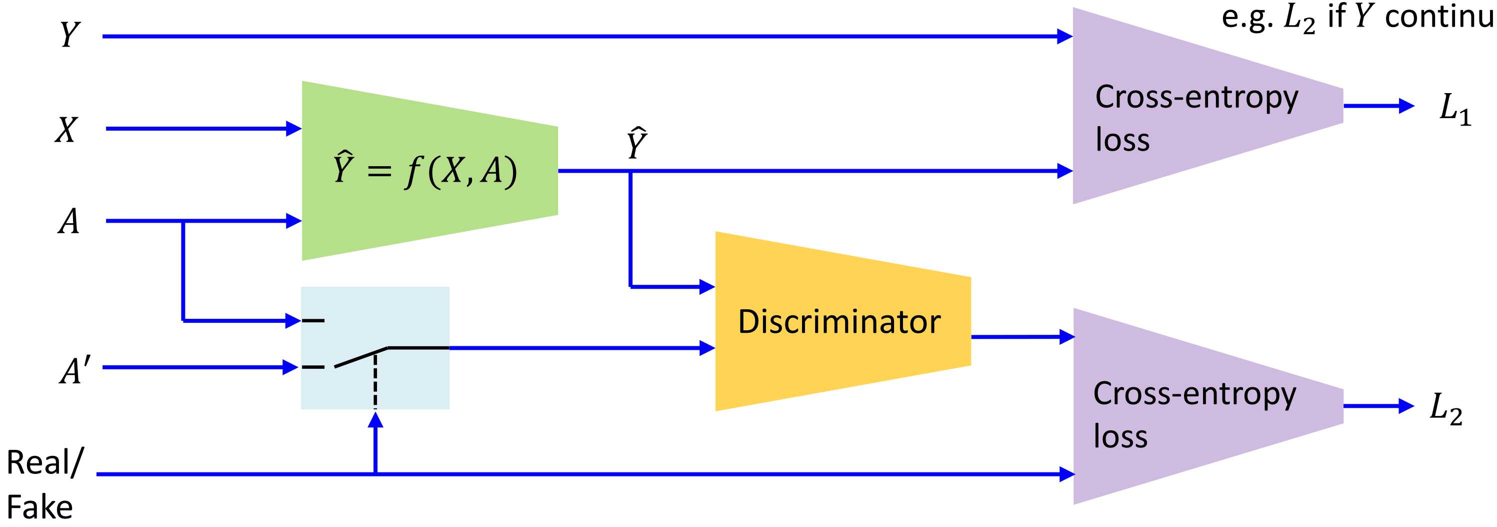
We train the discriminator to classify the $(\hat{Y}, A)$ pairs as accurately as possible (minimize cross-entropy loss to a label of 1 = real, 0 = fake).

We train the generator $\hat{Y} = f(X, A)$ to fool the discriminator, which means that the joint distribution of $(\hat{Y}, A)$ should be as close as possible to the fake (product) distribution $(\hat{Y}, A')$.

GAN-like Optimization for Demographic Parity

Using neural networks, we have this picture:

Assumes Y is discrete.
Can use any loss appropriate,
e.g. L_2 if Y continuous.



The Discriminator is trained to minimize L_2.
The predictor f is trained to minimize $L_1 - \beta L_2$ for $\beta > 0$

Mutual Information: A Quantitative Measure of Independence

A more commonly used measure of dependence is *mutual information*.

For perfect independence between X and Y we have $P(X,Y) = P(X)P(Y)$, and we define

$$I(X;Y) = D_{KL}(P(X,Y)||P(X)P(Y))$$

Since this quantity is always ≥ 0, and 0 iff $P(X,Y) = P(X)P(Y)$, minimizing this quantity is maximizing independence or fairness.

$I(X;Y)$ is called the **Mutual Information** between X and Y.

MINE: Mutual Information Neural Estimation

A lower bound on the KL-divergence is given by the Donsker-Varadhan formula:

$$D_{KL}(P||Q) = \sup_{T:\Omega \to R} E_p\,[T] - \log E_q\,[\exp T]$$

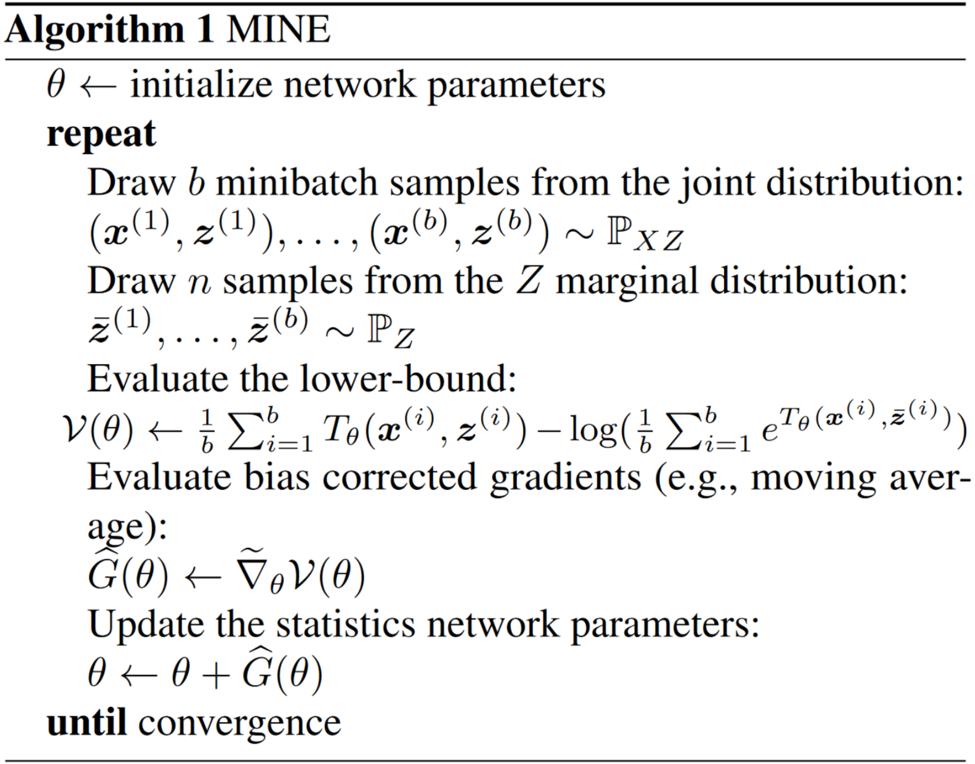
Which looks rather cryptic, unless we interpret T as the log of a density ratio estimator (close to a discriminator). The optimal T satisfies:

$$T = \log \frac{P}{Q} + c$$

So minimizing mutual information based on the Donsker-Varadhan formula suggests a GAN-like alternation:

- Minimize the KL-divergence estimate over the prediction $\hat{Y} = f(X, A)$
- Maximize the KL-divergence estimate over the function T.

MINE: Mutual Information Neural Estimation



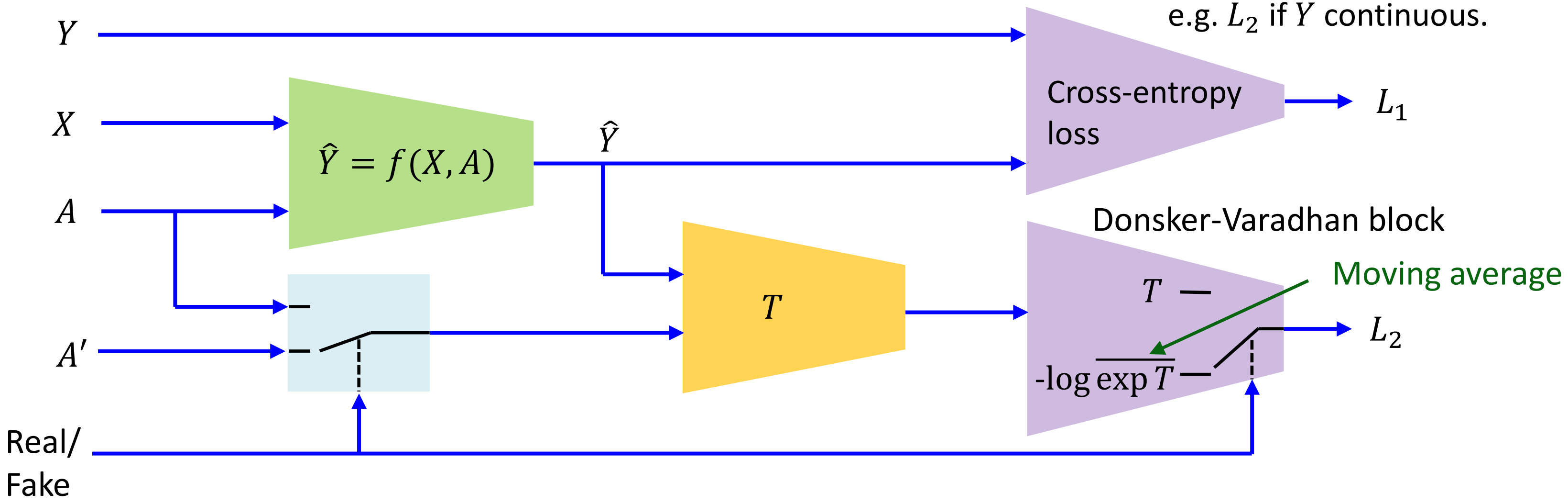
Algorithm 1 MINE

```
$\theta \leftarrow$ initialize network parameters
repeat
    Draw $b$ minibatch samples from the joint distribution:
    $(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \ldots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$
    Draw $n$ samples from the $Z$ marginal distribution:
    $\bar{\boldsymbol{z}}^{(1)}, \ldots, \bar{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$
    Evaluate the lower-bound:
    $\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{z}}^{(i)})})$
    Evaluate bias corrected gradients (e.g., moving average):
    $\widehat{G}(\theta) \leftarrow \widetilde{\nabla}_\theta \mathcal{V}(\theta)$
    Update the statistics network parameters:
    $\theta \leftarrow \theta + \widehat{G}(\theta)$
until convergence
```

MINE for Demographic Parity

Using neural networks, we have this picture:

Assumes Y is discrete.
Can use any loss appropriate, e.g. L_2 if Y continuous.



The T estimator is trained to maximize L_2.
The predictor f is trained to minimize $L_1 + \beta L_2$ for $\beta > 0$

Conditional Mutual Information

The same idea works for conditional independence. Lets looks at separation: we want $\left(\hat{Y} \perp\!\!\!\perp A\right) \mid Y$ or $P\left(\hat{Y}, A \mid Y\right) = P\left(\hat{Y} \mid Y\right)P(A \mid Y)$, so here we minimize

$$I\left(\hat{Y}; A \mid Y\right) = D_{KL}(P\left(\hat{Y}, A \mid Y\right)||P\left(\hat{Y} \mid Y\right)P(A \mid Y))$$

Which is zero iff $P\left(\hat{Y}, A \mid Y\right) = P\left(\hat{Y} \mid Y\right)P(A \mid Y)$

And $I(\hat{Y}; A \mid Y)$ is called the **Conditional Mutual Information**.

Conditional Mutual Information

Conditional Mutual Information can be expressed as a difference between mutual information terms:

$$I\left(\hat{Y};A\middle|Y\right)=I(\hat{Y};A,Y)-I(\hat{Y};Y)$$

And we can estimate both of the RHS terms using MINE.

So we can still effectively optimize Separation or Sufficiency Fairness in neural networks.

Takeaways

Acceptable attributes X, protected attribute A,
actual outcome Y, and predicted outcome $\hat{Y} = f(X, A)$.

- **Unawareness:** $\hat{Y} = f(X, A) = f(X)$
- **Demographic Parity:** $\hat{Y} \perp\!\!\!\perp A$
- **Separation:** $\left(\hat{Y} \perp\!\!\!\perp A\right) \mid Y$
- **Sufficiency:** $(Y \perp\!\!\!\perp A) \mid \hat{Y}$
- **Causality:** ideally force A and see if distribution of $\hat{Y}$ stays the same.
- **GAN-like optimization:** minimize $J(\hat{Y}; A)$
- **MINE:** minimize $I(\hat{Y}; A)$