Fairness Metrics (4):

Average Precision (AP, Scikit-Learn):
Summarizes precision-recall curve as the weighted mean of precisions achieved at each threshold (increase in recall from previous threshold used as weight):

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

- Measure accuracy of classifiers
- Threshold invariant accuracy metric

- Does not capture behavior on different protected classes (expecting slight dip after fairness adjustment)

(Training Note: "trained classifiers using different number of synthetic pairs for 4 different attributes, and found that AP stabilizes after 160,000 pairs, which is what we used to train our classifiers.")

Difference in Equality of Opportunity (DEO, Lokhande et al.):
Absolute Difference between false negative rates for both gender expressions
- Threshold Variant

(Definition from paper)
A classifier *h* satisfies Equality of Opportunity (EO) if *h(x)* is independent of the protected attribute *s* for $y \in \{0, 1\}$.

Equivalently, *h* satisfies the EO if $d_h^y = 0$ where we set $\mu_h^{s_i} = e_h^{s_i} \,|\, (y \in \{0, 1\}) =: e_h^{s_i, y_i}$, conditioning on both *s* and *y*.

Depending on choice of *y* in $\mu_h^{s_i}$, two different metrics:

1. *y* = 0 corresponds to *h* with equal *False Positive Rate (FPR)* across s_i
2. *y* = 1 corresponds to *h* with equal *False Negative Rate (FNR)* across s_i

(These observations observed in this paper from NeurIPS 2016)

h satisfies *Equality of Odds* if $d_h^0 + d_h^1 = 0$ (i.e. *h* equalizes both TPR and RPR across *s*)

Bias Amplification (BA, Wang and Russakovsky):

Measures how much more often a target attribute is predicted with a protected attribute than the ground truth value. (So measuring correlation, in layman's)

- Threshold variant

Let,

$P_{t\,|\,g}$ the fraction of images with protected attribute g that have target attribute t.

$P_{\hat{t}\,|\,g}$ the fraction of images with protected attribute g that have predicted target attribute $\hat{t}$

$P_{t,\,g}$ the fraction of images with target *t* and protected attribute g

P_t and P_g - the fraction of images with attribute *t* and *g* respectively

To Compute:

```
For each target, protected attribute:
    if P_{t,g} > P_t P_g:
        add (P_{t|g} − P_{t̂|g})
    else:
        add − (P_{t|g} − P_{t̂|g})
```

(A negative value implies that the bias now exists in a different direction than in the training data)

Divergence Between Score Distributions (KL, Chen and Wu)

$s_{g,\,t}$ represents a smoothed histogram of classifier scores of a certain protected attribute label and a target label, appropriately normalized as a probability distribution.

For each target attribute label *t*, measure

$$KL\left[s_{g=-1,t} \| s_{g=1,t}\right] + KL\left[s_{g=1,t} \| s_{g=-1,t}\right]$$

Measuring the divergence of g = -1 and g = 1 score distributions, separately for positive and negative attribute samples

(Stricter notion of *equalized odds*)