Additional Materials

The discrimination scores computed by the Discrimination Test of our algorithm quantify the distance from the statistical parity condition $\forall s \in SA. P(Y = +|SA = s) = P(Y = +|SA \neq s)$, which is equivalent to independence of Y and SA, as seen below.

Lemma 1. *Let Y be binary, and SA discrete random variables. $Y \perp\!\!\!\perp SA$ iff $\forall s \in SA. P(Y|SA = s) = P(Y|SA \neq s)$.*

Proof. $Y \perp\!\!\!\perp SA$ occurs by definition if $\forall s \in SA. P(Y|SA = s) = P(Y)$.

The *if* part follows by observing $P(Y) = P(Y|SA = s)P(SA = s) + P(Y|SA \neq s)P(SA \neq s)$, and, since $P(Y|SA = s) = P(Y|SA \neq s)$, that $P(Y) = P(Y|SA = s)(P(SA = s) + P(SA \neq s)) = P(Y|SA = s)$, for every s.

The *only-if* part follows by observing that $P(Y) = P(Y|SA = s)P(SA = s) + P(Y|SA \neq s)(1 - P(SA = s))$, and then by independence, $P(Y)(1 - P(SA = s)) = P(Y|SA \neq s)(1 - P(SA = s))$, which yields $P(Y|SA \neq s) = P(Y) = P(Y|SA = s)$ for every s.