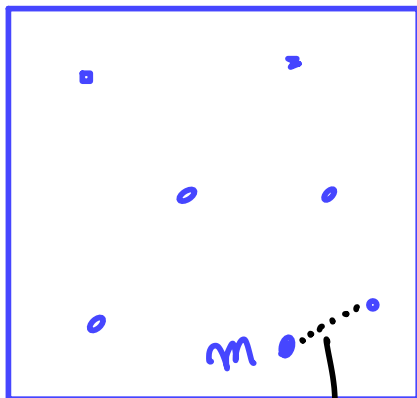


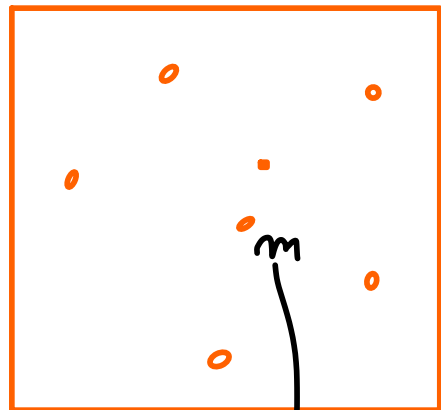
Adversarial Accuracy Indicator

S , data

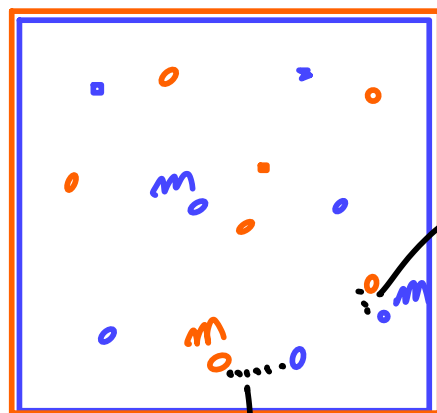


$d_{SS}(m)$ → average
over m

T , model



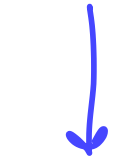
$d_{TT}(m)$
↓ average



$d_{ST}(m)$

$d_{TS}(m)$

$$A_S = \frac{1}{N_S} \sum_m \mathbb{1}(d_{SS}(m) < d_{ST}(m))$$



$\frac{1}{2}$ ideally

$$A_T = \frac{1}{N_T} \sum_m \mathbb{1}(d_{TT}(m) < d_{TS}(m))$$



$\frac{1}{2}$ ideally

$$\mathcal{E}^{AAI} = \left(A_S - \frac{1}{2}\right)^2 + \left(A_T - \frac{1}{2}\right)^2$$

distance $d_{mm'} = \sum_{i=1}^L \mathbb{1}(m_i \neq m'_i)$

$L = A \cdot G$ (number of different bits)

$$d_{SS}(m) \stackrel{?}{=} d_{SY}(m)$$

$$\Downarrow$$

$$\mathbb{1}(\dots) \Rightarrow \frac{1}{2}$$

$$A=4 \quad h_0: \boxed{0001} \quad \boxed{-1-1-11}$$

$$h_1=2 \quad \boxed{1}$$

$$h_2=1 \quad \boxed{1}$$

$$h_3=1 \quad \boxed{1}$$

$$\boxed{1100}$$

$k \in \mathbb{Z}$

$h_m = \pm 1, [0,1]$

$$E_q = \sum_{k \in \mathbb{Z}} \left[a_k + \underbrace{\sum_{\mu} W_{k\mu} h_{\mu}}_{\text{local field}} \right] \psi_k^z$$

$$E_1, E_2, E_3, E_4$$

$$B_n = e^{-E_1}, \dots$$

$$P_i = \frac{B_i}{B_1 + B_2 + B_3 + B_4 + \dots}$$

$$\sum_i P_i = 1$$

$$C_i = \sum_{j \leq i} P_j$$

