Foundations of Trustworthy AI

Jonathan Ullman

CS7180/7880 · Fall 2026

Class 4: Group conditional calibration
(+ wrap up from last week)

Sep 22, 2026

Calibrated Decision Loss

predictions $p_1 \dots p_T \in [0,1]$

outcomes $y_1 \dots y_T \in \{0,1\}$

for each $p \in [0,1]$: $S_p = \{t : p_t = p\}$ $\quad \bar{y}_p = \frac{1}{|S_p|} \sum_{t \in S_p} y_t$

$$ECE(\vec{p}, \vec{y}) = \frac{1}{T} \sum_p |S_p| \, |p - \bar{y}_p|$$

Expected calibration error

for a decision task $G = (A, U)$ there is a best response fn

actions utility

$$BR_G(p) = \underset{a \in A}{\operatorname{argmax}} \ \mathbb{E}_{y \sim Ber(p)} (U(a, y))$$

Calibrated Decision Loss

for a decision task $G = (A, U)$ there is a best response fn

actions utility

$$BR_G(p) = \operatorname*{argmax}_{a \in A} \mathbb{E}_{y \sim Ber(p)}\left(U(a, y)\right)$$

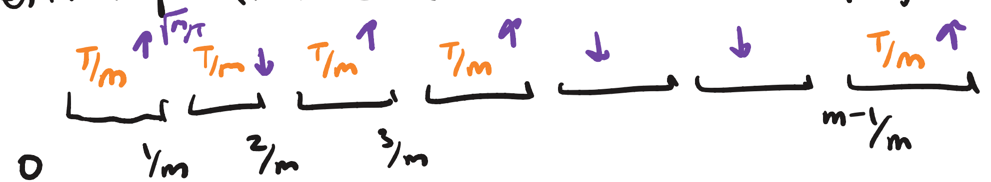
decision loss for $(\vec{p}, \vec{y})$ and task G

$$CDL_G(\vec{p}, \vec{y}) = \frac{1}{T} \sum_{t=1}^{T} \Big[\underbrace{U(BR_G(\bar{y}_{p_t}), y_t)}_{\text{util. if you knew probabilities}} - \underbrace{U(BR_G(p_t), y_t)}_{\text{util. if you trust } p_t} \Big]$$

$$CDL(\vec{p}, \vec{y}) = \max_{\text{tasks } G} CDL_G(\vec{p}, \vec{y})$$

Calibrated Decision Loss vs. ECE

<u>Theorem</u> [FV98]: There is an algorithm that makes sequential predictions such that $ECE(\vec{p},\vec{y}) = O(1/T^{1/3})$



each day nature chooses a bucket p at random

then chooses $y_t \sim Ber(p)$

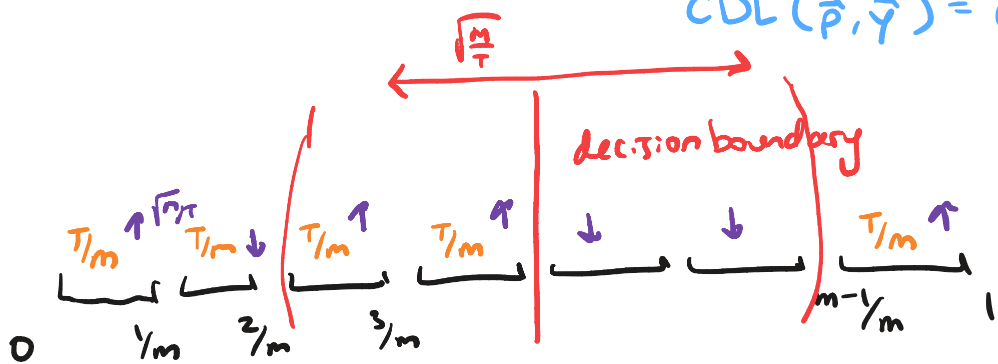
for each bucket $\bar{y}_p = p \pm \sqrt{\frac{1}{T/m}} = p \pm \sqrt{\frac{m}{T}}$

$$ECE(\vec{p},\vec{y}) = \frac{1}{T}\sum_{\text{buckets } p} |S_p| \cdot (\bar{y}_p - p) + \frac{1}{m} = \frac{1}{T} \cdot m \cdot \frac{T}{m} \cdot \sqrt{\frac{m}{T}} + \frac{1}{m} = \sqrt{\frac{m}{T}} + \frac{1}{m}$$

$$= \sqrt{\frac{m}{T}} + \frac{1}{m} \overset{m=T^{1/3}}{=} \frac{1}{T^{1/3}}$$

Calibrated Decision Loss vs. ECE

<u>Theorem</u> [HW24]: There is an algorithm that makes sequential predictions such that ~~$ECE(\vec{p},\vec{y}) = O(1/T^{1/2})$~~
$CDL(\vec{p},\vec{y}) = \tilde{O}(1/T^{1/2})$



For ECE, the error of every bucket is "relevant"

For CDL, most errors are "not relevant"

↳ only buckets in a small interval cause me to change decision

↳ can set $m \approx T^{1/2}$ and get $CDL(\vec{p},\vec{y}) \approx 1/T^{1/2}$

Group Conditional Bias

For now only consider offline prediction

For now, D is known

Example: predicting weather $y \in \{0,1\}$

$x = (x_1, x_2)$ $\quad$ $x_1 \in \{humid, dry\}$

$x_2 \in \{cloudy, sunny\}$

all four choices of x equally likely

$\mathbb{P}(y \mid x)$	cloudy	sunny
humid	3/4	1/2
dry	1/2	1/4

true probabilities

predictor f

$$f(x_1) = \begin{cases} 5/8 & \text{if } x_1 = \text{humid} \\ 3/8 & \text{if } x_1 = \text{dry} \end{cases}$$

f is perfectly calibrated

Average Bias Recap

Definition: For dist D and pred f the average bias is

$$\beta_D(f) = \mathop{\mathbb{E}}_{D}[f(X) - Y]$$

we will usually drop D

Lemma: If $f_{\beta\to 0}(x) = f(x) - \beta_D(f)$

then $f_{\beta\to 0}$ is unbiased and also $MSE_D(f_{\beta\to 0}) = MSE_D(f) - \beta_D(f)^2$

Proof is an exercise

Group Conditional Bias

A <u>group</u> is a subset $G \subseteq \mathcal{X}$ represented by a function $g : \mathcal{X} \to \{0,1\}$ ($g(x) = 1$ if $x \in G$, $g(x) = 0$ if $x \notin G$)

<u>Example:</u> $G = \{ (x_1, x_2) : x_2 = \text{cloudy} \}$

$$g(x_1, x_2) = \begin{cases} 1 & \text{if } x_2 = \text{cloudy} \\ 0 & \text{if } x_2 = \text{sunny} \end{cases}$$

Bias of f conditioned on G

$$\beta_D(f, g) = \mathop{\mathbb{E}}_{D} [f(x) - Y \mid g(x) = 1]$$

Group Conditional Bias

For now only consider offline prediction

For now, D is known

Example: predicting weather $y \in \{0,1\}$

$x = (x_1, x_2)$ $\quad x_1 \in \{\text{humid}, \text{dry}\}$

$x_2 \in \{\text{cloudy}, \text{sunny}\}$

all four choices of x equally likely

$\mathbb{P}(y \mid x)$	cloudy	sunny
humid	$3/4$	$1/2$
dry	$1/2$	$1/4$

true probabilities

predictor f

$$f(x_1) = \begin{cases} 5/8 & \text{if } x_1 = \text{humid} \\ 3/8 & \text{if } x_1 = \text{dry} \end{cases}$$

f is perfectly calibrated

f is unbiased on average

$$\mathbb{E}[f(X) - Y \mid \text{cloudy}]$$

$$= \underbrace{\frac{1}{2} \cdot \frac{5}{8} + \frac{1}{2} \cdot \frac{3}{8}} - \underbrace{\frac{1}{2} \cdot \frac{3}{4} - \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2} - \frac{5}{8} = \frac{-1}{8}$$

Group Conditional Bias

A group is a subset $G \subseteq \mathcal{X}$ represented by a function $g: \mathcal{X} \to \{0,1\}$ ($g(x)=1$ if $x \in G$, $g(x)=0$ if $x \notin G$)

Bias of f conditioned on G $\quad \beta_D(f,g) = \mathbb{E}_D[f(x) - y \mid g(x)=1]$

A class of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$ (a set of subsets of $\mathcal{X}$)

Define $\mu_D(g) = \mathbb{P}_D[g(x)=1]$

Definition: Bias of f conditioned on groups $\mathcal{G}$ is

$$\alpha^2 = \max_{g \in \mathcal{G}} \beta_D^2(f,g) \cdot \mu_D(g)$$