$$K_2(\pi^{\le s+1}) - K_2(\pi^{\le s}) = \Delta_{s+1}(p_{s+1}, y_{s+1}) \le \frac{2V_s^{p_{s+1}}}{\sqrt{n(\pi^{\le s}, p_{s+1})}} \cdot (y_{s+1} - p_{s+1}) + \frac{1}{n(\pi^{\le s}, p_{s+1})}$$

$$\left(\frac{\sum_{t=1}^{s} \mathbb{1}(p_t = p_{s+1})(y_t - p_t)}{\sqrt{n(\pi^{\le s}, p_{s+1})}} \right)^2$$

Want to find a dist over p_{s+1} to make $\nearrow$ small $\forall y_{s+1}$

$V_s^p \ge 0$ means our predscore weakly too small

≤ 0 (too big)

If $V_s^1(\pi^{\le s}) \ge 0$ [y_t's are $\ge$ p_ts on days where $p_t = 1$]

Let $p_{s+1} = 1$ w.p. 1

$Q^{\ge 0} \cdot (y_{s+1} - p_{s+1})$, $Q \in [0,1]$

setting $p_{s+1} = 1$ makes this as small as possible ≤ 0

If $V_s^0(\pi^{\le s}) \le 0$ [y_t's are $\le$ p_ts on $p_t = 0$ days]

Let $p_{s+1} = 0$ w.p. 1

$Q^{\le 0} \cdot (y_{s+1} - p_{s+1})$

setting $p_{s+1} = 0$ makes $y_{s+1} - p_{s+1}$ max (and $\ge$)

So prod is min (≤ 0)

One dumb obs/reminder

$$V_s^p(\pi) = \frac{\sum \mathbb{1}(p_t = p)(y_t - p_t)}{\sqrt{n(\pi^{\le s}, p)}}$$

$n(\pi, p) = 0$, $= 0$

$\le \frac{|n(\pi^{\le s}, p)|}{\sqrt{n(\pi^{\le s}, p)}}$

So $|V_s^p| \le \sqrt{n(\pi, p)}$

Else ($V_s'(\pi^{\le s}) < 0$ ⇒ $V_s^0(\pi^{\le s}) > 0$)

$\exists\ p \in \{0, \frac{1}{m}, \frac{2}{m}, \dots 1\}$ s.t. $V_s^p(\pi^{\le s}) \ge 0$

$\Rightarrow$ $V_s^{p+\frac{1}{m}}(\pi^{\le s}) \le 0$

Compute q:

$$\frac{q \cdot V_s^p(\pi^{\le s})}{\sqrt{n(\pi^{\le s}, p)}} + \frac{(1-q)\, V_s^{p+\frac{1}{m}}(\pi^{\le s})}{\sqrt{n(\pi^{\le s}, p+\frac{1}{m})}} = 0$$

Play p w.p. q
$p+\frac{1}{m}$ w.p. $1-q$

This means the

$$\mathbb{E}\left[\Delta'_{s+1}(p_{s+1}, y_{s+1})\right] \le q\,\Delta_s^p(\pi^{\le s})(y_s - p) + (1-q)\,\Delta_s^{p+\frac{1}{m}}(\pi^{\le s}) \quad (y_s - p - \tfrac{1}{m})$$

$$= \frac{1}{q} \cdot -\frac{1}{m}\Delta_s^{p+\frac{1}{m}}(\pi^{\le s})$$

Ends up $\le \frac{2}{m}$

So, this strategy makes the 1st term $O(\frac{1}{m})$

Won't show but $\sum_{s=1}^{T} \frac{1}{n(\pi^{\le s}, p_{s+1})} \le O\left(m \log\left(\frac{T}{m}\right)\right)$

So $\mathbb{E}[K_2] = \mathbb{E}[K_2/T] \le O\left(\frac{1}{m} + \frac{m}{T}\log\frac{T}{m}\right)$

Now, we argued one can get online calibration. It's also more meaningful than online MMC since

@ the least it won't allow $\begin{matrix} y & 0&1&0&1&0&1&0&1 \\ p & 1&0&1&0&1&0&1&0 \end{matrix}$

But it still is mostly arguing that y's & p's aren't too anticorrelated, still pretty global if your predictions are constant.

E.G. $p = \bar{y}$ is calibrated.

More meaningful when your predicter has more variability

What about asking for "good" pred on sets defined in a way $\perp$ of predictions?

Eg, are p's similarly intepretable for $x \in$ Tall as $x \in$ Short ?

Also kind of weird that we haven't touched xs yet, other than that predictions might come from them.

- Attempt 1: Marginal mean consistency for groups (quantile)

  2: Calibration I groups

Let $G \in 2^X$ be a collection of group fns,
$g \in G$ is a subset of X, $g(x) \in \{0, 1\}$
$g(x) = 1$ "x belongs to g"

Then, $M(g) = \Pr_{x \sim D_X}[g(x) = 1]$ is g's frequency/mass

---

Def f is α-approx group mmc for G if
$\forall g \in G$

$$M(g)\left[\mathbb{E}[f(x) \mid g(x)=1] - \hat{f}[y \mid g(x)=1]\right]^2 \leq \alpha$$

If f isn't α-approx group MMC for G, patching will decrease ^{2}ed error, by

$$M(g) \cdot \Delta^2 \geq \alpha$$

So $\frac{1}{\alpha}$ rounds will lead to termination.

This is more meaningful than MMC for expressive G: eg $G = 2^X$.

However, this should lead us to wondering how much additional data we need to satisfy it... to estimate MMC for each group, for example.
(Hint: One way to do this is to treat G like 2^X!) $|G|$ or $VC(G)$

Let's talk now about group conditional calibration. We call this multicalibration:

Def Fix $f: \mathcal{X} \to [0,1]$ and group $g: \mathcal{X} \to \{0,1\}$. Avg^2 cal error of f on g is

$$K_2(f,g,D) = \sum_v \Pr[f(x)=v \mid g(x)=1] \left(v - \mathbb{E}_{\substack{g(x)=1 \\ f(x)=v}} y\right)^2$$

Ideally, want α avg^2 multicalibration error $\forall g \in G$, but weighted by group mass

$$\frac{K_2(f,g,D)}{\mu(g)} \leq \alpha.$$

---

well, if your f isn't α-multicalibrated wrt G, we can patch it!

$$h(x, f; v \to v', g) = \begin{cases} v' \text{ if } f(x)=v, g(x)=1 \\ f(x) \text{ o/w} \end{cases}$$

One subtlety here. Originally calibration patching mapped values $v \in R(f)$ to $v' \in (R(f))$, no more than $|R(f)|$ outputs.

w this, $v \in R(f)$ & $g \in G \Rightarrow v'$, could increase to $|R(f)| \cdot |G|$

& even worse sorta since we're adding to $R(f)$]

Our Brier score ↓ is $B(f_{t+1}) - B(f_t) = \mu_t(v_t, g_t)(v_t' - v_t)^2$

A natural fix is to define our grid of prediction values in advance,

$$m = \lceil \tfrac{1}{\alpha} \rceil \quad , \quad R(f) = \{0, \tfrac{1}{m}, \dots, 1\}$$

$$\geq \mu_t(v_t, g_t)(v_t' - v_t)^2 - \frac{1}{4m^2}$$