OK | We'd discussed the alg

Multigroup Calibrate V1$(f, G, \alpha)$:

while ( $f^t$ isn't $\alpha$-MGC for $G$):

$(v^t, g^t) \in \operatorname{argmax} \Pr[f(x)=v, g(x)=1]\,(v - E[y \mid f(x)=v, g(x)=1])^2$

$f^{t+1} = \text{Patch}^{\text{Value/shift}}(f^t, g^t, v^t \Rightarrow ?)$

And while each patch will reduce $\ell_2$ error

by $(v^t - v^{t'})^2 \cdot \Pr[f^t(x)=v^t, g^t(x)=1] \geq \alpha \cdot \Pr(f^t(x)=v, g^t(x)=1]$

$\geq$ ?

In the argument about (nongroup) multicalibration,

We argued $(v^t - v^{t'})^2 \Pr[f^t(x)=v^t] \geq \frac{\alpha}{m} \; \forall t$

We kept replacing $v^t$ by $v^{t'}$, so still $m$ predicted values after the patch. The above doesn't have that structure. we might add new values of prediction

$? \geq \alpha \cdot \frac{\Pr[g^t(x)=1]}{m+t}$, So analyzing convergence $\rightarrow$ bounding the # of rounds T st

$$\alpha \sum_{t=0}^{T} \frac{\Pr[g^t(x)=1]}{m+t} \geq 1$$

We can analyze, but it's simpler to round.

Consider $\left[\frac{1}{m}\right]$, we previously assumed $|R(f)| = m$, now we'll enforce it on this grid.

$$\text{Round}(v, m) = \underset{v' \in [\frac{1}{m}]}{\operatorname{argmin}} |v' - v|$$

$$\text{Round}(f, m) \Rightarrow f'(x) = \text{Round}(f(x), m)$$

---

Then,

RandMulticalibrate $(f, \alpha, G, D)$:

Let $m = \frac{1}{\alpha}$

$f_0 = \text{Round}(f, m)$, $t = 0$

while ($f_t$ is not $\alpha$-MGC wrt $G$)

$(v^t, g^t) \in \underset{v, g}{\operatorname{argmax}} \Pr(f^t(x) = v, g(x) = 1)(v - E[y \mid f^t(x) = v, g(x) = 1])^2$

$\tilde{v}^t = E[y \mid f(x) = v^t, g^t(x) = 1]$

$\bar{v}^t = \text{Round}(\tilde{v}^t, m)$

$f^{t+1} = \text{ValuePatch}(f^t, v^t \Rightarrow \bar{v}^t, g^t)$

$t = t + 1$

---

Then

Lemma

$$B[f_t] - B[f_{t+1}] \geq \mu_t(v_t, g_t)\left((v^t - \tilde{v}^t)^2 - \frac{1}{4m^2}\right)$$

[Since we can analyze red in 2ed error $\Rightarrow \tilde{v}^t$ then rounding]

---

So, you can argue a violation means

$K_2(f^t, g^t) \geq \frac{\alpha}{\mu(g_t)}$ + one value has $\geq \frac{1}{m+1}$ mass

$$\sum_v \mu_t(g_t, v)[v - E[y \mid f_t(x) = v, g_t(x) = 1]]^2 \geq \alpha$$

$\exists\ v \in [\frac{1}{m}]$ st $\nearrow$ $\geq \frac{\alpha}{m+1}$

So $\mu_t(v_t, g_t)(v_t - \tilde{v}_t) \geq \frac{\alpha}{m+1}$

$$\text{Lemma}$$

$p$ $\quad B(f_t) - B(f_{t+1}) \geq \mu_t(v_t, g_t)\left((v_t - \hat{v}_t)^2 - \frac{1}{4m^2}\right)$

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

$$\geq \frac{\alpha}{\frac{1}{\alpha} + 1} - \frac{\alpha^2}{4}$$

$$\geq \frac{\alpha^2}{4}.$$

So, $T \cdot \frac{\alpha^2}{4} \geq 1$ if $T \geq \frac{4}{\alpha^2}$ rounds.

Groups :

Mean
Quantile consistency →

We talked
about using
value ~~patches~~
$g(x)=1 \rightarrow \hat{f}(x)=E[y \mid g(x)=1]$

Even for group multicalibration we were using these value patches.

---

While we certainly can do this, "Shift" patches often work abit better in practice:
for an $f$, on subset defined by $g$, add $\Delta$:

$$h(f,g,\Delta,x) = \begin{cases} f(x) & \text{if } g(x)=0 \\ f(x)+\Delta & \text{if } g(x)=1 \end{cases}$$

$g$ can be a group, or a value ~~+~~ group.

---

Why? Well, start w group $g$ mean consistency
A value patch → one const pred/group

If $f$ had some good behavior on $g$, we've lost it.

---

For group (multi)calibration, in the case that our predictions were in $[0, \frac{1}{m}, \dots 1]$, and $\Delta = \frac{1}{m}$, these are equiv. If our preds were instead $\in \mathbb{R}$, and we "bucketed" into $\sqcup_{0} \sqcup_{\frac{1}{m}} \dots \sqcup_{1-\frac{1}{m}}$
then shifting also keeps info

---

Part of why this is relevant is that shift patches, as additive moves, can be done in any order and give the same behavior,

Even when groups aren't disjoint. [Sometimes algorithmic improvements]

---

To argue about finite sample behavior, there's some additional work. [See Chap 4.2 in uncertainty notes]

Ok. Now we're going to further interrogate the relationship between accuracy (loss) & multicalibration

[already know improving multicalibration → improved loss]

What other relationships?

[Today: Multicalibration can be used to <u>boost</u> regression accuracy]

[Others, if this is a topic folks like]

---

Recall perfect multicalibration as an eqn:

$$\mathop{\mathbb{E}}_{x,y\sim D}[(y-f(x)) \mid f(x)=v, g(x)=1] = 0 \quad \forall v, g$$

Equivalently since $g$ was binary,

$$\mathbb{E}[g(x)[y-f(x)] \mid f(x)=v]$$

Makes sense to study even for real valued $g$'s.

Def | Multicalibration w.r.t. R.V. functions |

Fix dist $D$ over $\Delta Z$, a model $f: \mathcal{X} \to [0,1]$

Let $\mathcal{H}$ be any set of functions : $\mathcal{X} \to \mathbb{R}$.

We call $f$ $\alpha$-apx MC wrt $\mathcal{H}$ if

$\forall h \in \mathcal{H}$

$$K_2(f, h, D) = \sum_{v \in R(f)} \Pr[f(x) = v] \mathbb{E}[h(x)(y-v) \mid f(x) = v]^2 \leq \alpha$$

In other words, $\mathcal{H}$ is good at finding weak parts of $f$.

Lemma | Fix a calibrated $f$. Suppose $\exists v \in R(f)$, $h \in \mathcal{H}$ s.t.

$$\mathbb{E}[(f(x) - y)^2 - (h(x) - y)^2 \mid f(x) = v] \geq \alpha$$

then it must be that

$$\mathbb{E}[h(x)(y-v) \mid f(x) = v] \geq \frac{\alpha}{2}$$

Squared errors in level sets @ $v$

So, $h(x)$ having lower 2ed error @ $v$ means a MC violation.

Proof | $\mathbb{E}[h(x)(y-v) \mid f(x)=v]$

$$= \mathbb{E}[h(x)\cdot y \mid f(x)=v] - v\,\mathbb{E}[h(x) \mid f(x)=v]$$

$$= \frac{1}{2}\left[2\mathbb{E}[h(x)\cdot y \mid f(x)=v] - 2v\,\mathbb{E}[h(x) \mid f(x)=v]\right)$$

$$\geq \frac{1}{2}\left[ \quad '' \quad\quad '' \quad - \mathbb{E}[(h(x)-v)^2 \mid f(x)=v]\right)$$

$$= \frac{1}{2}\left[\mathbb{E}[2h(x)y - h^2(x) - v^2 \mid f(x)=v]\right]$$

[Since $f$ is calibrated, $E[y \mid f(x)=v] = v$

$v \cdot ( \quad ) = v^2$

$$= \frac{1}{2}\mathbb{E}\left[2h(x)y - h^2(x) - \underline{v^2} + v^2 - \underline{vy} \mid f(x)=v\right]$$

$-2vy + v^2$

$$= \frac{1}{2}E\left[(v-y)^2 - (h(x)-y)^2 \mid f(x=v)\right]$$

$$\geq \frac{\alpha}{2}$$

---

OK, so if $\exists\, h$ which is better than $f$ on one of $f$'s level sets, $h$ will be a MC violation for $f$.

Lemma 2 | Fix $f: \mathcal{X} \to [0,1]$. Suppose $\exists v \in R(f)$

$\nexists h \in \mathcal{H}$ s.t.

$$\mathbb{E}[h(x)(y-v) \mid f(x)=v] \geq \alpha$$

(h is an mc viol)

Let

$h' = v + \eta h(x)$

$$\eta = \frac{\alpha}{\mathbb{E}[h(x)^2 \mid f(x)=v]}$$

then,

$$\mathbb{E}[(f(x)-y)^2 - (h'(x)-y)^2 \mid f(x)=v] \geq \frac{\alpha^2}{\mathbb{E}[h(x)^2 \mid f(x)=v]}$$

Proof |

$$= \mathbb{E}[(v-y)^2 - (v+\eta h(x)-y)^2 \mid f(x)=v]$$

$$= \mathbb{E}[v^2 - 2\cancel{v}y + y^2 - (v+\eta h(x))^2 + 2y(v+\eta h(x)) - y^2 \mid f(x)=v]$$

$$= \mathbb{E}[2y\eta h(x) - 2v\eta h(x) - \eta^2 h(x)^2 \mid f(x)=v]$$

$$= \mathbb{E}[2\eta h(x)[y-v] - \eta^2 h(x^2) \mid f(x)=v]$$

$$\geq 2\eta\alpha - \eta^2 \mathbb{E}[h(x)^2 \mid f(x)=v]$$

$$\geq 2\frac{\alpha^2}{\mathbb{E}[h^2(x) \mid f(x)=v]} - \frac{\alpha^2}{\mathbb{E}[h^2(x) \mid f(x)=v]}$$

$\square$

If $\mathcal{H}$ is closed under affine transformations $h' \in \mathcal{H}$

Let's show now that we can reduce multicalibration to regression "oracles"

**Def** $A_{\mathcal{H}}$ is a squared error regression oracle for a class of real-valued functions $\mathcal{H}$ if $\forall\, D \in \Delta Z$, $A_{\mathcal{H}}(D)$ outputs $h \in \mathcal{H}$

s.t. $$h \in \underset{h' \in \mathcal{H}}{\operatorname{argmin}} \underset{(x,y)\sim D}{\mathbb{E}}\left[(h'(x)-y)^2\right]$$

[Linear & poly regression have these, more complex classes seem to have good heurist

**Alg** RegressionMulticalibrate$(f, \alpha, A_{\mathcal{H}}, D, B)$

Let $m = \frac{2B}{\alpha}$

$f_0 = \text{Round}(f, m)$

$err_0 = \mathbb{E}[(y - f_0(x))^2]$   $err_{-1} = \infty$, $t = 0$

While $(err_t - err_{t-1} > \frac{\alpha}{2B})$ :

for each $v \in [\frac{1}{m}]$

$D_v^{t+1} = D|_{f_t(x)=v}$

Let $h_v^{t+1} = A_{\mathcal{H}}(D_v^{t+1})$

$\tilde{f}_{t+1} = \sum_{v \in [\frac{1}{m}]} \mathbb{1}(f_t(x)=v) \cdot h_v^{t+1}(x)$ , $f_{t+1} = \text{Round}(\tilde{f}_{t+1}, m)$

$err_{t+1} = \mathbb{E}[(y - f_{t+1}(x))^2]$

Output $f_{t-1}$

**Theorem** RegressionMulticalibrate halts after $T \leq \frac{2B}{\alpha}$ rounds, & outputs $f_{t-1}$ which is $\alpha$-MC w.r.t.

$$\mathcal{H}_B = \{ h \in \mathcal{H} \mid h(x)^2 \leq B \}$$ ← norm bounded h

---

Note this takes $(m+1)$ calls of the oracle/iteration

$\frac{2B}{\alpha} + 1$

And so $O\left(\frac{B^2}{\alpha^2}\right)$ calls in total.

---

Ok, so regression oracles can solve MC.
When does MC improve accuracy (more than just not hurting $\ell_2$ error)?
What about when MC on $\mathcal{H}$ implying Bayes optimality?

---

**Def** | Fix $D \in \Delta Z$ & function class $\mathcal{H}$. Let $f^*(x) = \mathbb{E}_{y \sim D(x)}[y]$

We say $\mathcal{H}$ satisfies the <u>weak learner cond.</u> wrt $D$ if $\forall$ $S \subset \mathcal{X}$ w. $\Pr[x \in S] > 0$,

$$\mathbb{E}_{(x,y)\sim D}[(f^*(x)-y)^2 \mid x \in S] < \min_{c \in \mathbb{R}} \mathbb{E}_{(x,y)\sim D}[(c-y)^2 \mid x \in S]$$

then $\exists$ $h \in \mathcal{H}$ s.t.

$$\mathbb{E}_{(x,y)\sim D}[(h(x)-y)^2 \mid x \in S] < \min_{c \in \mathbb{R}} \mathbb{E}_{(x,y)\sim D}[(c-y)^2 \mid x \in S]$$

---

For any subset of $\mathcal{X}$, if $f^*$ is better than the best constant, $\exists$ $h \in \mathcal{H}$ which is better too (though $f^*$ might b $\gg h$)

---

When is it the case that <u>any</u> multicalibrated $f$ wrt $\mathcal{H}$ is Bayes optimal?

---

**Theorem 34** | Fix $D$. Let $\mathcal{H}$ be a set of fns closed under affine transformats: $h \in \mathcal{H} \Rightarrow \alpha h(x) + b \in \mathcal{H}$

MC wrt $\mathcal{H}$ implies Bayes-OPT over $D$ iff $\mathcal{H}$ satisfies the weak learning condition.

$\Rightarrow$ (Show weak learning $\Rightarrow$ (Every MC is Bayes OPT))

if not, $\exists$ MC $f$ which isn't Bayes opt:

$$\mathbb{E}[(y - f^*(x))^2] < \mathbb{E}[(y - f(x))^2]$$

$$\Leftrightarrow \sum_v \Pr[f(x) = v]\, \underbrace{\mathbb{E}[(y - f(x))^2 - (y - f^*(x))^2 \mid f(x) = v]}_{>0}$$

$\Rightarrow \exists v$ st $\underline{\qquad}$ $\underline{\qquad}$ $> 0$

that is a subset of $\mathcal{X}$, and since $f$ is calibrated, it's ~~its~~ prediction is right on average there

$\Rightarrow$ there must exist an $h \in \mathcal{H}$ better than constant $f$

[weak learning cond]