Lemma 6.1.3.a

Let $\mathcal{H}$ be closed under affine transformations. Suppose f is calibrated, and $\exists$ $h \in \mathcal{H}$ s.t.

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

Then $\exists$ $h' \in \mathcal{H}$ s.t.

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

and also if B, then A w. different constants.

Def | Fix $D \in \Delta \mathcal{Z}$ & 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. $\Pr[x \in S] > 0$,

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

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

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

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?

<u>Theorem[34]</u> | 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

[assume $\mathcal{X}$, $\mathcal{Y}$ countable]

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

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

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

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

$\Rightarrow \exists v$ s.t. $\_\_\_ > 0$

that is a subset of $\mathcal{X}$ (where $f(x)=v$), and since f is calibrated it's prediction is right on average there

$$\mathbb{E}[(\underbrace{v}_{f(x)} - y)^2 \mid x \in S] = \min_{c \in \mathbb{R}} [(c-y)^2 \mid x \in S]$$

$\Rightarrow$ $\mathbb{E}[(f^*(x) - y)^2 \mid x \in S] < \_\_\_$

So, weak learning implies $\exists h \in \mathcal{H}$ s.t.

$$\mathbb{E}[(h(x)-y)^2 \mid x \in S] < \min_{c \in \mathbb{R}} \mathbb{E}[(\underbrace{c}_{f(x) \text{ or } v} - y)^2 \mid x \in S]$$

[Lemma ?] $\Rightarrow \exists h' \in \mathcal{H}$ s.t. $\mathbb{E}[h'(x)(y - \underbrace{v}_{f(x)}) \mid f(x) = v] > 0$

A violation of MC wrt $\mathcal{H}$ $\times$ Contradiction

$\Leftarrow$ (For any $\mathcal{H}$ not satisfying WLC wrt $\mathcal{D}$, MC wrt $\mathcal{H}$ & $\mathcal{D}$ doesn't imply Bayes-opt over $\mathcal{D}$.

In particular $\exists$ f MC wrt $\mathcal{H}$ & $\mathcal{D}$ which isn't Bayes-OPT: $\mathbb{E}[(f(x)-y)^2] > \mathbb{E}[(f^*(x)-y)^2]$

Since $\mathcal{H}$ doesn't satisfy WLC over $\mathcal{D}$, $\exists$ $S \subseteq \mathcal{X}$ w $\Pr[x \in S] > 0$ s.t.

$$\min_h \mathbb{E}[(h(x)-y)^2 \mid x \in S] > \min_{c \in \mathbb{R}} [(c-y)^2 \mid x \in S]$$

Let $c(S) = \mathbb{E}[y \mid x \in S]$

Define $f(x) = \begin{cases} f^*(x) & x \notin S \\ c(S) & x \in S \end{cases}$

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

$$= \Pr[x \in S]\,\mathbb{E}[(c(S)-y)^2 \mid x \in S] + \Pr[x \notin S]\,\mathbb{E}[(f^*(x)-y)^2 \mid x \notin S]$$

$$> \Pr[x \in S]\,\mathbb{E}[(f^*(x)-y)^2 \mid x \in S] + \Pr[x \notin S]\,\mathbb{E}[(f^*(x)-y)^2 \mid x \notin S]$$

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

So f isn't Bayes OPT. Is it MC wrt $\mathcal{H}$, $\mathcal{D}$?

Informally, it should be: it's Bayes OPT everywhere but S, and on S, it's better than the best ~~h∈H~~.

Suppose f isn't MC wrt $\mathcal{H}$. then $\exists\, h \in \mathcal{H}$ and $v \in R(f)$ s.t.

$$\mathbb{E}[h(x)(y-v) \mid f(x)=v] > 0$$

So $\boxed{\text{Lemma}} \Rightarrow \exists\, h'$:

$$\ast \quad \mathbb{E}[(h'(x)-y)^2 - (f(x)-y)^2 \mid f(x)=v] > 0$$

Notice v must be $c(S)$, since $f(x)$ is Bayes OPT elsewhere.

$$\mathbb{E}[(h'(x)-y)^2 \mid f(x)=c(S)]$$

$$= \underbrace{\Pr[x \in S]\,\mathbb{E}[(h'(x)-y)^2 \mid f(x)=c(S),\ x \in S]}_{A}$$

$$+ \Pr[x \notin S]\,\mathbb{E}[(h'(x)-y)^2 \mid f(x)=c(S),\ x \notin S]$$

$$\geq A + \Pr[x \notin S]\,\mathbb{E}[(f-y)^2 \mid f(x)=c(S),\ x \notin S]$$

Since $f = f^*$ here

So for $\mathcal{H}$, it must be that

$$\mathbb{E}[(h'(x)-y)^2 \mid x \in S, f(x)=c(S)]$$
$$< \mathbb{E}[(f(x)-y)^2 \mid f(x)=c(S)]$$

$$\Rightarrow \mathbb{E}[(h'(x)-y)^2 \mid x \in S] < \mathbb{E}[(c(S)-y)^2 \mid x \in S]$$

A contradiction to $\mathcal{H}$ violating the WLC.

So, if f is exactly MC wrt $\mathcal{H}, D$,
and $\mathcal{H}$ satisfies WLC

$\Rightarrow f$ is Bayes OPT

(and, if f is Bayes OPT & exactly MC wrt $\mathcal{H}, D$, $\mathcal{H}$ must satisfy the WLC).

Approximate variants hold too.

Conformal prediction

Let's talk about classification!

$\mathcal{X} =$ ____ $\qquad \mathcal{Y} = \{1, \dots, K\}$

$f : \mathcal{X} \to \mathcal{Y}$

How good is a given prediction?

Most of our contemporary models actually produce $f(x) \in \Delta Y$, "probability of labels"

↳ A1: you could train a calibrated regressor to predict Pr[correct]

A2: you could output sets C_x s.t. $y_x \in C_x$ w.p. $\overline{1-\delta}$

[Note, these might feel like confidence intervals, that is a prediction set, but these coverage guarantees generally hold only if the world follows the parametric model class]

$T : \mathcal{X} \to \underline{\underline{2^{\mathcal{Y}}}}$ s.t. $\Pr[y \in T(x)] \geq 1-\delta$

High dimensional.

Idea: learn a nonconformity score fn $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

We have some model f

Given x, $f(x)$, y, how surprising is y?

[example $s(x,y) = |y - f(x)|$ for regression]

Parametrizes a 1-d set $T(x, \tau) = \{y : s(x,y) \leq \tau\}$

Weak Marginal guarantee

Given a sample $D \sim \mathcal{D}^n$, a calibration set

Produce $T(x)$ that have this

coverage guarantee: for new $(x,y) \sim \mathcal{D}$

$$1-\delta \leq \Pr_{\substack{D \sim \mathcal{D}^n \\ (x,y) \sim \mathcal{D}}} [y \in T(x)] \leq 1-\delta + \frac{1}{n+1}$$

[can be made whp D]

Let τ be the smallest val s.t.

$$\sum_{i=1}^{n} \mathbb{1}(s(x_i, y_i) \leq \tau) \geq (1-\delta)(n+1)$$

[the empirical $\frac{(1-\delta)(n+1)}{n}$ quantile of D]

Output $T_D(x) = \{\hat{y} : s(x, \hat{y}) \leq \tau\}$

We might also want guarentees

that aren't just marginal,

eg

Given $G \subseteq 2^{X}$, group conditional covg:

$$\Pr_{(x,y)\sim D}[\, y \in \mathcal{T}(x) \mid g(x)=1 \,] = 1-\delta$$

To get such guarantees is a little trickier, our coverage/pred sets will now look like

was τ

$$\mathcal{T}_D^f(x) = \{ \bar{y} : s(\bar{y}, x) \leq f(x) \}$$

$\mathcal{T}_D^f(x)$ will have group covg

$\iff$ $f(x)$ has <u>group conditional quantile guarantees!</u>

[Which we can obtain]

Unfortunately, these $f(x)$'s instead of a fixed $\hat{\tau}$ add complexities.

w.p. δ, output $\emptyset$ & $1-\delta$ output $\mathcal{Y}$.

↳ This has the right coverage prob. for every x, g, ... but it's completely uniformative.

To fix, we ask for coverage to hold | group & value of thresh, eg $\forall g \in G$, $v \in R(f)$

$$\Pr_{(x,y)}\left[y \in T_D^f(x) \mid g(x)=1, f(x)=v \right] = 1-\delta$$

This will hold $\iff$ $f(x)$ is a multicalibrated quantile predictor for $q = 1-\delta$

This stuff is doable in online settings too ....