4.4 Group Conditional Calibration

Group conditional bias asks that the average prediction be correct within each group, but it does not require the individual prediction values to be meaningful within that group. To obtain both guarantees at once, we can ask for calibration conditional on group membership. For a predictor f with finite range, a group g, and a prediction p, define the intersection group

$$g_{f,p}(x) := g(x)\mathbb{1}[f(x) = p].$$

Thus $g_{f,p}$ contains the members of g that receive prediction p, and $\beta_{\mathcal{D}}(f, g_{f,p})$ is precisely the calibration bias at p within g.

The definition below directly mirrors the definition of ECE from Class 1. ECE averages the absolute calibration bias over the prediction values p; here we restrict attention to members of a particular group g and average the *squared* calibration bias instead:

Definition 4.9 (Calibration conditional on a group)**.** For a group g with $\mu_{\mathcal{D}}(g) > 0$, a finite-range predictor f is *calibrated conditional on* g if, for every p for which $\mu_{\mathcal{D}}(g_{f,p}) > 0$,

$$\beta_{\mathcal{D}}(f, g_{f,p}) = p - \mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid g(X) = 1, f(X) = p] = 0.$$

As with group conditional bias, we want to control this property simultaneously over a collection of groups. We use the same mass-dependent normalization so that smaller groups receive more tolerance.

Definition 4.10 (Uniform group conditional calibration)**.** Given a collection $\mathcal{G}$ of group indicators and a parameter $\alpha \geq 0$, a finite-range predictor f has *α-uniform group conditional calibration* with respect to $\mathcal{G}$ if, for every $g \in \mathcal{G}$ with $\mu_{\mathcal{D}}(g) > 0$,

$$\sum_{p \in \text{range}(f)} \mathop{\mathbb{P}}_{\mathcal{D}}[f(X) = p \mid g(X) = 1]\beta_{\mathcal{D}}(f, g_{f,p})^2 \leq \frac{\alpha^2}{\mu_{\mathcal{D}}(g)},$$

where terms corresponding to intersection groups of mass zero are omitted. When $\alpha = 0$, this is equivalent to f being calibrated conditional on every group $g \in \mathcal{G}$ of positive mass.

Taking $g(x) = 1$ recovers ordinary calibration, while averaging the equalities over p shows that calibration conditional on g implies zero bias conditional on g. In our weather example, the humidity-only predictor is calibrated but is not calibrated conditional on the group of cloudy days: among humid and cloudy days it predicts $5/8$ even though the probability of rain is $3/4$.

The same MSE argument used for group conditional bias gives an iterative algorithm. The only complication is that correcting the predictions on $g_{f,p}$ can introduce a new prediction value, so the range could keep growing. Following Roth, we avoid this by restricting every prediction to the grid

$$[1/m] := \{0, 1/m, 2/m, \ldots, 1\},$$

and writing $\text{Round}(q; m)$ for the closest grid point to q. For simplicity, assume that $1/\alpha^2$ is an integer.

Algorithm 3: Group conditional calibration

Input: A predictor $f : \mathcal{X} \to [0,1]$, a collection of groups $\mathcal{G}$, and a parameter $\alpha \in (0,1]$.

1. Set $m = 1/\alpha^2$, let $f_0(x) = \mathrm{Round}(f(x); m)$, and set $t = 0$.
2. While f_t does not have α-uniform group conditional calibration with respect to $\mathcal{G}$, choose a group g_t that violates the bound in Definition 4.10, and choose $p_t \in \mathrm{range}(f_t)$ maximizing

$$\mu_{\mathcal{D}}((g_t)_{f_t,p})\beta_{\mathcal{D}}(f_t, (g_t)_{f_t,p})^2.$$

   Let

$$q_t = \mathrm{Round}\left(\mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid g_t(X) = 1, f_t(X) = p_t]; m\right),$$

   set

$$f_{t+1}(x) = \begin{cases} q_t & \text{if } g_t(x) = 1 \text{ and } f_t(x) = p_t, \\ f_t(x) & \text{otherwise,} \end{cases}$$

   and increment t.

Output: The predictor f_t.

Theorem 4.11. *For every $\alpha \in (0,1]$ such that $1/\alpha^2$ is an integer, Algorithm 3 stops after fewer than $4/\alpha^4$ updates and outputs a predictor with α-uniform group conditional calibration with respect to $\mathcal{G}$. If it performs T updates, then*

$$\mathrm{MSE}_{\mathcal{D}}(f_T) < \mathrm{MSE}_{\mathcal{D}}(f_0) - \frac{T\alpha^4}{4}.$$

Proof. Fix an iteration and abbreviate g_t by g. Multiplying the violated bound by $\mu_{\mathcal{D}}(g)$ shows that

$$\sum_{p \in \mathrm{range}(f_t)} \mu_{\mathcal{D}}(g_{f_t,p})\beta_{\mathcal{D}}(f_t, g_{f_t,p})^2 > \alpha^2.$$

The range of f_t contains at most $m+1$ grid points, so the maximizing p_t satisfies

$$\mu_{\mathcal{D}}(g_{f_t,p_t})\beta_{\mathcal{D}}(f_t, g_{f_t,p_t})^2 > \frac{\alpha^2}{m+1}.$$

The unrounded group update would decrease MSE by the left-hand side. Rounding its new value to the nearest grid point loses at most $1/(4m^2)$ in squared error, so

$$\mathrm{MSE}_{\mathcal{D}}(f_t) - \mathrm{MSE}_{\mathcal{D}}(f_{t+1}) > \frac{\alpha^2}{m+1} - \frac{1}{4m^2} \geq \frac{\alpha^4}{4},$$

where the final inequality uses $m = 1/\alpha^2$ and $\alpha \leq 1$. Since f_0 and Y take values in $[0,1]$, $\mathrm{MSE}_{\mathcal{D}}(f_0) \leq 1$, while MSE is always nonnegative. There can therefore be fewer than $4/\alpha^4$ updates. When the algorithm stops, no group violates the bound in Definition 4.10. □

Acknowledgments and AI Use

The overall structure of this lecture is inspired by Aaron Roth's lecture notes [Rot26]. The content of the notes is more or less AI-free, but I used AI throughout to help write the LaTeX code for examples, definitions, and calculations.

5 Omniprediction

5.1 Omniprediction relative to predictors

At a very abstract level, the idea of multicalibration is that the predictions are unbiased even if we restrict attention to a rich set of groups. We also saw that calibration has decision theoretic properties based on the idea that we can treat calibrated predictions as if they were real probabilities when we try to maximize our utility or minimize our loss. The idea of *omniprediction* combines these two ideas into one.

Suppose we want to choose an action $a \in A$ to minimize a loss function $\ell\colon A \times \{0,1\} \to \mathbb{R}$. This is the same as the decision tasks from Class 3, with utility $U(a,y) = -\ell(a,y)$. If we treat a prediction p as the true probability that $Y = 1$, then we should choose the best response

$$\mathrm{BR}_\ell(p) := \operatorname*{argmin}_{a \in A} \mathop{\mathbb{E}}_{Y \sim \mathrm{Ber}(p)} [\ell(a, Y)] .$$

Given a predictor f, this gives the decision rule $x \mapsto \mathrm{BR}_\ell(f(x))$. Once we have trained f, computing this best response only requires the prediction and the loss function, without any additional labeled data.

Suppose we already know how to learn a good predictor $h\colon \mathcal{X} \to [0,1]$ for one particular loss. If the loss changes, however, we might need to learn a different predictor. The most direct goal of omniprediction is to avoid this retraining: learn one predictor f whose best response is at least as good as the best response to any benchmark predictor h, simultaneously for many different losses [GKRSW22].

Definition 5.1 (Omniprediction relative to a class of predictors)**.** Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, an action set A, a collection $\mathcal{L}$ of loss functions $\ell\colon A \times \{0,1\} \to \mathbb{R}$, and a collection of predictors $\mathcal{H} \subseteq \{h\colon \mathcal{X} \to [0,1]\}$. For $\alpha \geq 0$, a predictor $f\colon \mathcal{X} \to [0,1]$ is an *α-approximate omnipredictor relative to* $(\mathcal{H}, \mathcal{L})$ if, for every $\ell \in \mathcal{L}$ and every $h \in \mathcal{H}$,

$$\mathop{\mathbb{E}}_{\mathcal{D}}\big[\ell(\mathrm{BR}_\ell(f(X)), Y)\big] \leq \mathop{\mathbb{E}}_{\mathcal{D}}\big[\ell(\mathrm{BR}_\ell(h(X)), Y)\big] + \alpha.$$

When $\alpha = 0$, we simply call f an *omnipredictor relative to* $(\mathcal{H}, \mathcal{L})$.

The order of the quantifiers is the point: we train one predictor f, and then choose the loss function ℓ. The best-response map can change with the loss, but the predictor stays the same. Both $f(X)$ and $h(X)$ are interpreted as probabilities and converted into actions using the same best-response map.

For example, take $A = [0,1]$. Under squared loss $\ell(a,y) = (a-y)^2$, the best response is $\mathrm{BR}_\ell(p) = p$, so the definition compares f directly with each $h \in \mathcal{H}$. Under absolute loss $\ell(a,y) = |a-y|$, a best response is $\mathrm{BR}_\ell(p) = \mathbb{1}[p \geq 1/2]$, so the definition compares the thresholded classifiers induced by f and h. The same f must support both comparisons.

5.2 Omniprediction relative to policies

There is also a stronger, more decision-theoretic variant. Instead of comparing with the best response to another purported probability, we can compare with an arbitrary decision rule that uses the features directly.

Definition 5.2 (Omniprediction relative to a class of policies)**.** Fix a distribution $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$, an action set A, a collection $\mathcal{L}$ of loss functions $\ell\colon A\times\{0,1\}\to\mathbb{R}$, and a collection of policies $\mathcal{P}\subseteq\{\pi\colon\mathcal{X}\to A\}$. For $\alpha\geq 0$, a predictor $f\colon\mathcal{X}\to[0,1]$ is an *α-approximate omnipredictor relative to* $(\mathcal{P},\mathcal{L})$ if, for every $\ell\in\mathcal{L}$ and every $\pi\in\mathcal{P}$,

$$\mathop{\mathbb{E}}_{\mathcal{D}}\big[\ell(\mathrm{BR}_\ell(f(X)),Y)\big]\leq\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(\pi(X),Y)]+\alpha.$$

When $\alpha=0$, we call f an *omnipredictor relative to* $(\mathcal{P},\mathcal{L})$.

The advantage of Definition 5.2 is that the benchmark need not have probability semantics: it can be any treatment rule, classifier, or other policy. It also keeps the benchmark class fixed as the loss varies. The disadvantage is that it asks for a stronger guarantee than we may need when our only goal is to compete with a class of probability predictors. If the policy class contains $\mathrm{BR}_\ell\circ h$ for every $h\in\mathcal{H}$ and $\ell\in\mathcal{L}$, this stronger definition immediately implies Definition 5.1. We will otherwise focus on the predictor-relative definition.

5.3 From group conditional calibration to omniprediction

The connection to Class 4 is especially clean. For each benchmark predictor $h\in\mathcal{H}$ and loss $\ell\in\mathcal{L}$, the induced rule $\mathrm{BR}_\ell\circ h$ partitions the feature space according to the action it takes. If our probabilities remain calibrated after conditioning on every part of this partition, then best responding to those probabilities must outperform the induced rule.

For a predictor $h\in\mathcal{H}$, a loss $\ell\in\mathcal{L}$, and an action $a\in\mathrm{range}(\mathrm{BR}_\ell\circ h)$, define the *action group*

$$g_{h,\ell,a}(x):=\mathbb{1}[\mathrm{BR}_\ell(h(x))=a].$$

Let

$$\mathcal{G}_{\mathcal{H},\mathcal{L}}:=\{g_{h,\ell,a}:h\in\mathcal{H},\ \ell\in\mathcal{L},\ a\in\mathrm{range}(\mathrm{BR}_\ell\circ h)\}.$$

Thus the groups do not represent demographic categories in this application. They represent the regions on which a benchmark predictor induces a particular action for a particular loss.

Theorem 5.3 (Group conditional calibration implies omniprediction)**.** *If f is calibrated conditional on every group in $\mathcal{G}_{\mathcal{H},\mathcal{L}}$ of positive mass, then f is an omnipredictor relative to $(\mathcal{H},\mathcal{L})$.*

Proof. Fix a loss $\ell\in\mathcal{L}$ and a benchmark predictor $h\in\mathcal{H}$. Write $A_h=\mathrm{BR}_\ell(h(X))$, and condition on a pair of values $f(X)=p$ and $A_h=a$ having positive probability. Calibration conditional on $g_{h,\ell,a}$ says

$$\mathbb{P}[Y=1\mid f(X)=p,\mathrm{BR}_\ell(h(X))=a]=p.$$

Consequently,

$$\begin{aligned}\mathbb{E}[\ell(\mathrm{BR}_\ell(p), Y) \mid f(X) = p, \mathrm{BR}_\ell(h(X)) = a] &= \mathop{\mathbb{E}}_{Y\sim\mathrm{Ber}(p)}[\ell(\mathrm{BR}_\ell(p), Y)] \\ &\le \mathop{\mathbb{E}}_{Y\sim\mathrm{Ber}(p)}[\ell(a, Y)] \\ &= \mathbb{E}[\ell(\mathrm{BR}_\ell(h(X)), Y) \mid f(X) = p, \mathrm{BR}_\ell(h(X)) = a].\end{aligned}$$

The inequality is exactly the definition of a best response. Averaging over (p, a) gives

$$\mathbb{E}[\ell(\mathrm{BR}_\ell(f(X)), Y)] \le \mathbb{E}[\ell(\mathrm{BR}_\ell(h(X)), Y)].$$

Since ℓ and h were arbitrary, the claim follows. □

The approximate notion from Class 4 gives an approximate omniprediction guarantee with almost the same proof.

Theorem 5.4 (Approximate omniprediction)**.** *Suppose every loss $\ell \in \mathcal{L}$ takes values in $[0, 1]$, every induced rule $\mathrm{BR}_\ell \circ h$ takes at most k distinct actions, and f has α-uniform group conditional calibration with respect to $\mathcal{G}_{\mathcal{H},\mathcal{L}}$. Then, for every $\ell \in \mathcal{L}$ and $h \in \mathcal{H}$,*

$$\mathbb{E}[\ell(\mathrm{BR}_\ell(f(X)), Y)] \le \mathbb{E}[\ell(\mathrm{BR}_\ell(h(X)), Y)] + 2\alpha\sqrt{k}.$$

Thus f is a $2\alpha\sqrt{k}$-approximate omnipredictor relative to $(\mathcal{H}, \mathcal{L})$.

Proof. Fix $\ell \in \mathcal{L}$ and $h \in \mathcal{H}$, and write $A_h = \mathrm{BR}_\ell(h(X))$. For every positive-mass cell indexed by (p, a), write

$$q_{p,a} := \mathbb{P}[Y = 1 \mid f(X) = p, A_h = a]$$

and let $w_{p,a} = \mathbb{P}[f(X) = p, A_h = a]$. For any action b, define

$$L_r(b) := r\ell(b, 1) + (1 - r)\ell(b, 0).$$

Because ℓ takes values in $[0, 1]$, $|L_q(b) - L_p(b)| \le |q - p|$. Moreover, $L_p(\mathrm{BR}_\ell(p)) \le L_p(a)$. Therefore the excess loss on cell (p, a) is at most

$$L_{q_{p,a}}(\mathrm{BR}_\ell(p)) - L_{q_{p,a}}(a) \le \left|L_{q_{p,a}}(\mathrm{BR}_\ell(p)) - L_p(\mathrm{BR}_\ell(p))\right| + \left|L_p(a) - L_{q_{p,a}}(a)\right| \le 2|q_{p,a} - p|.$$

For each action a in the range of A_h, uniform group conditional calibration on $g_{h,\ell,a}$ gives

$$\sum_p w_{p,a}(q_{p,a} - p)^2 \le \alpha^2.$$

There are at most k such actions, so

$$\sum_{p,a} w_{p,a}(q_{p,a} - p)^2 \le k\alpha^2.$$

Finally, Cauchy–Schwarz and $\sum_{p,a} w_{p,a} = 1$ imply

$$\sum_{p,a} w_{p,a}|q_{p,a} - p| \le \sqrt{\sum_{p,a} w_{p,a}(q_{p,a} - p)^2} \le \alpha\sqrt{k}.$$

Multiplying by the factor of two in the cell-wise bound proves the theorem. □

Acknowledgments and AI Use

The notion of omniprediction is due to Gopalan, Kalai, Reingold, Sharan, and Wieder [GKRSW22]. I used ChatGPT to help refactor some definitions and to sanity check the way I wanted to present it against the more traditional presentation. I also used ChatGPT to help keep track of the calculations in Theorem 5.4.