days are equally likely, so the probability of rain is

$$\frac{1}{2}\left(\frac{3}{4}+\frac{1}{2}\right)=\frac{5}{8}.$$

Similarly, conditional on the prediction $f(x)=3/8$, the probability of rain is

$$\frac{1}{2}\left(\frac{1}{2}+\frac{1}{4}\right)=\frac{3}{8}.$$

Thus f is calibrated even though it ignores the cloud-cover feature.

Our theory of calibrated decision loss says that if we only base our decisions on the predictions, then we should treat these predictions as if they are ground truth probabilities. But should we only base our decisions on the predictions? If it's cloudy or sunny, and incorporate that information into our decsion making. If we do that, then on a day that is humid and cloudy, the prediction is $p=5/8$ but we know that the true probability is actually $3/4$, and can do better than simply trusting the prediction. Thus the predictions can be biased after we condition on some of the observable features that factor into our decision.

4.1 Group conditional bias

To this end, we can revisit the definition of average bias and add in the notion of groups that capture the information we are using to make a decision.

As before, let's start with the concept of average bias. Recall that the average bias of a predictor is the expected difference between its prediction and the outcome.

Definition 4.2 (Average bias)**.** For a distribution $\mathcal{D}$ and a predictor f, the *average bias* is

$$\beta_{\mathcal{D}}(f) := \mathop{\mathbb{E}}_{\mathcal{D}}[f(X)-Y].$$

The predictor is *unbiased on average* if $\beta_{\mathcal{D}}(f)=0$.

As we saw in Class 1, we can remove the average bias without increasing the mean squared error. In fact, the improvement in MSE is exactly the squared bias.

Lemma 4.3 (Unbiasing reduces MSE)**.** *Let $f_{\beta\to 0}$ be the unbiasing of f, defined by*

$$f_{\beta\to 0}(x) := f(x)-\beta_{\mathcal{D}}(f).$$

Then $f_{\beta\to 0}$ is unbiased on average and

$$\mathrm{MSE}_{\mathcal{D}}(f_{\beta\to 0})=\mathrm{MSE}_{\mathcal{D}}(f)-\beta_{\mathcal{D}}(f)^2.$$

Proof. First,

$$\mathop{\mathbb{E}}_{\mathcal{D}}[f_{\beta\to 0}(X) - Y] = \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y] - \beta_{\mathcal{D}}(f) = 0,$$

so $f_{\beta\to 0}$ is unbiased on average. Moreover,

$$\begin{aligned}\mathrm{MSE}_{\mathcal{D}}(f_{\beta\to 0}) &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f(X) - Y - \beta_{\mathcal{D}}(f))^2\right] \\ &= \mathrm{MSE}_{\mathcal{D}}(f) - 2\beta_{\mathcal{D}}(f)\mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y] + \beta_{\mathcal{D}}(f)^2 \\ &= \mathrm{MSE}_{\mathcal{D}}(f) - \beta_{\mathcal{D}}(f)^2.\end{aligned}$$

□

Now, we want to capture the idea that the bias is small *conditioned on some property of the feature vector*. We will eventually combine this with calibration, where we condition on the prediction itself as well, but for now what we do will be orthogonal to calibration.

We can define a *group* $G \subseteq \mathcal{X}$ to represent feature vectors with some property. The group is represented by a function $g\colon \mathcal{X} \to \{0, 1\}$, where $g(x) = 1$ means that x is in the group and $g(x) = 0$ means x is not in the group. For example, if G is the group of cloudy days then

$$G = \{(x_1, x_2) : x_2 = \text{cloudy}\}$$

and

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

We can ask whether a predictor is biased when we restrict attention to a particular group. Let $\mathcal{G}$ be a collection of group indicators $g\colon \mathcal{X} \to \{0, 1\}$, and let all expectations below be over $(X, Y) \sim \mathcal{D}$. Define the probability mass of a group by

$$\mu_{\mathcal{D}}(g) := \mathop{\mathbb{P}}_{\mathcal{D}}[g(X) = 1].$$

Group conditional bias is the average bias from Class 1, computed under the distribution restricted to the group. This definition is especailly interesting when we consider a large family $\mathcal{G} \subseteq 2^{\mathcal{X}}$ of groups simultaneously.

Definition 4.4 (Group conditional bias)**.** For a predictor $f\colon \mathcal{X} \to \mathbb{R}$ and a group g with $\mu_{\mathcal{D}}(g) > 0$, the *group conditional bias* is

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

The predictor is *unbiased conditional on* g if $\beta_{\mathcal{D}}(f, g) = 0$. Given a collection of groups $\mathcal{G}$ and a parameter $\alpha \geq 0$, we require the following group conditional bias bound for every $g \in \mathcal{G}$ with $\mu_{\mathcal{D}}(g) > 0$:

$$\beta_{\mathcal{D}}(f, g)^2 \leq \frac{\alpha^2}{\mu_{\mathcal{D}}(g)}.$$

When $\alpha = 0$, this requires zero group conditional bias on every group of positive mass.

The allowed squared bias scales as $1/\mu_{\mathcal{D}}(g)$. This adjustment reflects the fact that a sample contains fewer examples from a smaller group, making its mean harder to estimate.

For example, let g indicate the cloudy days in our weather example. Then $\mu_{\mathcal{D}}(g) = 1/2$, and

$$\mathop{\mathbb{E}}_{\mathcal{D}}[f(X) \mid g(X) = 1] = \frac{1}{2}\left(\frac{5}{8} + \frac{3}{8}\right) = \frac{1}{2},$$
$$\mathop{\mathbb{E}}_{\mathcal{D}}[Y \mid g(X) = 1] = \frac{1}{2}\left(\frac{3}{4} + \frac{1}{2}\right) = \frac{5}{8}.$$

Thus $\beta_{\mathcal{D}}(f, g) = -1/8$. The predictor systematically underestimates rain on cloudy days, even though it is calibrated. This group satisfies the group conditional bias bound exactly when $\alpha \geq 1/(8\sqrt{2})$.

4.2 Reducing group conditional bias

How can we reduce this bias? As in Class 1, we can shift the predictions, but now we only change predictions inside the group. For a group g of positive mass, define the updated predictor by

$$f_{\beta(g)\to 0}(x) := f(x) - \beta_{\mathcal{D}}(f, g)g(x).$$

The shift subtracts the group's conditional bias from each prediction in the group, making the group exactly unbiased. It also gives the following exact decrease in MSE.

Lemma 4.5 (A group update reduces MSE)**.** *For any predictor $f : \mathcal{X} \to \mathbb{R}$ and group g with $\mu_{\mathcal{D}}(g) > 0$, the update above satisfies*

$$\mathrm{MSE}_{\mathcal{D}}(f_{\beta(g)\to 0}) = \mathrm{MSE}_{\mathcal{D}}(f) - \mu_{\mathcal{D}}(g)\beta_{\mathcal{D}}(f, g)^2.$$

Proof. Write $\beta = \beta_{\mathcal{D}}(f, g)$ and $\mu = \mu_{\mathcal{D}}(g)$. The predictions only change inside the group, so

$$\begin{aligned}\mathrm{MSE}_{\mathcal{D}}(f) - \mathrm{MSE}_{\mathcal{D}}(f_{\beta(g)\to 0}) &= \mu \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f(X) - Y)^2 - (f(X) - Y - \beta)^2 \mid g(X) = 1\right] \\ &= \mu\left(2\beta \mathop{\mathbb{E}}_{\mathcal{D}}[f(X) - Y \mid g(X) = 1] - \beta^2\right) \\ &= \mu\beta^2.\end{aligned}$$

□

Now what to do if we have a set of groups $\mathcal{G}$? If the groups are disjoint then we can simply apply this operation to each group. But if the groups overlap then applying the operation to fix one group might break another group? Nevertheless, since each update decreases the MSE, we have to eventually make progress towards having perfect predictions. We can capture this idea using the following iterative procedure.

Algorithm 1: Group conditional unbiasing

Input: A predictor $f : \mathcal{X} \to [0, 1]$, a collection of groups $\mathcal{G}$, and a parameter $\alpha > 0$.
1. Set $f_0 = f$ and $t = 0$.

2. While there is a group $g_t \in \mathcal{G}$ of positive mass such that

$$\beta_{\mathcal{D}}(f_t, g_t)^2 > \frac{\alpha^2}{\mu_{\mathcal{D}}(g_t)},$$

choose any such group, set

$$f_{t+1}(x) = f_t(x) - \beta_{\mathcal{D}}(f_t, g_t) g_t(x),$$

and increment t.

Output: The predictor f_t.

We only evaluate conditional expectations on groups of positive mass. As in Roth's population-level algorithm, we assume access to exact expectations and a way to find a violating group whenever one exists. The bound below counts updates; it does not bound the cost of finding a group or the number of samples needed to estimate expectations.

Theorem 4.6. *For every $\alpha > 0$, Algorithm 1 stops after T updates, where*

$$T \le \frac{\mathrm{MSE}_{\mathcal{D}}(f_0)}{\alpha^2} \le \frac{1}{\alpha^2}.$$

Its output $f_T : \mathcal{X} \to \mathbb{R}$ satisfies the group conditional bias bounds in Definition 4.4 with parameter α, and

$$\mathrm{MSE}_{\mathcal{D}}(f_T) \le \mathrm{MSE}_{\mathcal{D}}(f_0) - T\alpha^2.$$

Proof. Whenever an update is made, the chosen group satisfies $\mu_{\mathcal{D}}(g_t)\beta_{\mathcal{D}}(f_t, g_t)^2 > \alpha^2$. By Lemma 4.5, the MSE decreases by more than α^2. After any k updates it is therefore at most $\mathrm{MSE}_{\mathcal{D}}(f_0) - k\alpha^2$. MSE is nonnegative, and $\mathrm{MSE}_{\mathcal{D}}(f_0) \le 1$ because both predictions and outcomes lie in $[0, 1]$. Consequently the procedure must stop within the stated number of updates. At termination no group violates the defining inequality. □

What does the set $\mathcal{G}$ mean? This definition lets us interpolate between average bias where $\mathcal{G}$ contains only the constant function $g(x) = 1$, and perfect predictions where $\mathcal{G}$ contains all functions $g(x)$. At zero error, the definition simply requires $\mathbb{E}_{\mathcal{D}}[f(X) \mid g(X) = 1] = \mathbb{E}_{\mathcal{D}}[Y \mid g(X) = 1]$ for all $g \in \mathcal{G}$ of positive mass. If $\mathcal{G}$ only contains the constant function $g(x) = 1$, this requires zero average bias, as in Class 1. At the other extreme, if $\mathcal{G}$ contains all subsets of $\mathcal{X}$, then zero group conditional bias forces $f = f^*$, where f^* is the optimal predictor from Class 1. Ideally $\mathcal{G}$ contains any group that we think might influence our decision.

Algorithmic aspects. Step 2 in Algorithm 1 might require enumerating all groups. That can be prohibitively expensive for very rich classes of groups, often exponential in the number of features. We won't get to it in this class, but we can define the problem solved in Step 2 as an optimization problem to find the group that maximizes the violation, and can solve it using whatever optimization techniques we like to get heuristic guarantees. The specific optimization problem often looks very similar to the type of learning problems that we are good at solving in practice, even if they are hard in theory.

Group conditional bias with a finite sample. As in Class 1, we are pretending that we have direct access to the distribution. However, as in Class 1 we are only using a fairly limited set of properties of the distribution. In particular, if we have a finite sample of data, what we need is just that the squared group conditional bias on the sample is within about $\alpha^2/4\mu(g)$ of its expectation for every $g \in \mathcal{G}$. We won't talk so much about this kind of *uniform convergence* analysis during this unit, but we can typically achieve this with a sample of size $O(\log|\mathcal{G}|/\alpha^2)$.

4.3 Reducing group conditional bias in one shot

Every update in Algorithm 1 adds a constant to the predictions on one group. Since these additions commute, every predictor obtained after any number of updates has the form

$$f_\lambda(x) := f(x) + \sum_{g \in \mathcal{G}} \lambda_g g(x)$$

for some vector of coefficients λ. This suggests choosing all of the group corrections simultaneously by minimizing MSE over this family.

Algorithm 2: One-shot group conditional unbiasing

Input: A predictor $f : \mathcal{X} \to \mathbb{R}$ and a finite collection of groups $\mathcal{G}$. Choose

$$\lambda^* \in \operatorname*{argmin}_{\lambda \in \mathbb{R}^{\mathcal{G}}} \mathop{\mathbb{E}}_{\mathcal{D}}\left[(f_\lambda(X) - Y)^2\right].$$

Output: The predictor f_{λ^*}.

Theorem 4.7. *The predictor f_{λ^*} output by Algorithm 2 is unbiased conditional on every group $g \in \mathcal{G}$ of positive mass. Moreover,*

$$\mathrm{MSE}_{\mathcal{D}}(f_{\lambda^*}) \le \mathrm{MSE}_{\mathcal{D}}(f),$$

and its MSE is no larger than that of any predictor obtained from f by a finite sequence of updates from Algorithm 1.

Proof. Suppose there were some $g \in \mathcal{G}$ of positive mass with $\beta_{\mathcal{D}}(f_{\lambda^*}, g) \neq 0$. Applying the group update from Lemma 4.5 to f_{λ^*} produces another predictor in the same family: it replaces λ^*_g by $\lambda^*_g - \beta_{\mathcal{D}}(f_{\lambda^*}, g)$ and leaves the other coefficients unchanged. By Lemma 4.5, this decreases MSE by the strictly positive quantity $\mu_{\mathcal{D}}(g)\beta_{\mathcal{D}}(f_{\lambda^*}, g)^2$, contradicting the optimality of λ^*. Thus every group of positive mass has zero group conditional bias. Finally, $f = f_0$ belongs to the family by taking $\lambda = 0$, and every finite sequence of group updates also produces a predictor in this family, so optimality gives both MSE comparisons. □

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 4.8 (Group conditional calibration)**.** A predictor f is *group conditionally calibrated* with respect to $\mathcal{G}$ if, for every $g \in \mathcal{G}$ and every p for which $\mu_{\mathcal{D}}(g_{f,p}) > 0$,

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

For $\alpha \geq 0$, we say that f is *α-approximately group conditionally calibrated* 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.

Taking $g(x) = 1$ recovers ordinary calibration, while averaging the equalities over p shows that group conditional calibration implies zero group conditional bias for every $g \in \mathcal{G}$. 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 $f_0(x) = \text{Round}(f(x); m)$, and set $t = 0$.
2. While f_t is not α-approximately group conditionally calibrated, choose a group g_t that violates the bound in Definition 4.8, and choose $p_t \in \text{range}(f_t)$ maximizing

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

Let

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

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 4.9. *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 an α-approximately group conditionally calibrated predictor. 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 4.8. □

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.