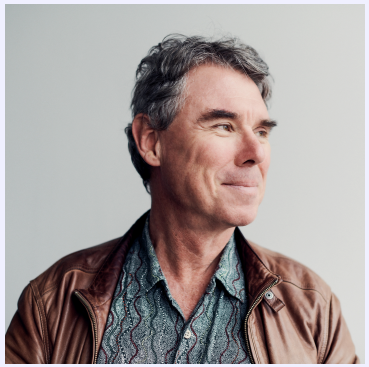
Hierarchies of calibration for multiclass settings

Rabanus Derr[a], Jessie Finocchiaro[b], Bob Williamson[a]

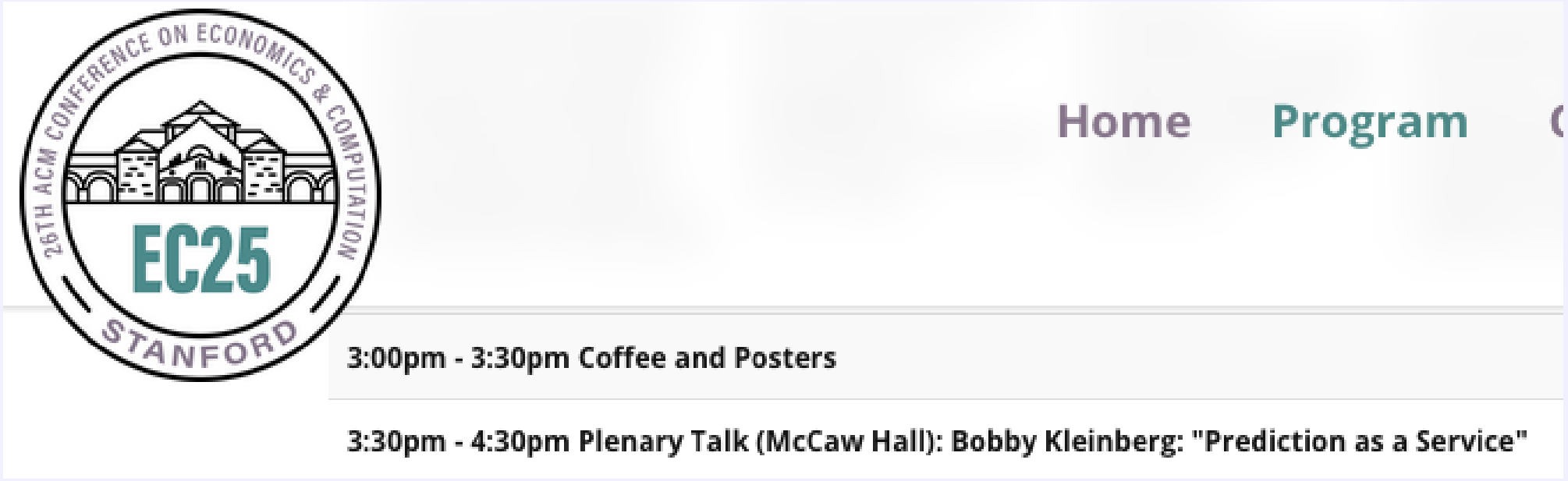


[a] Tubingen AI Center and University of Tubingen

[b] Boston College

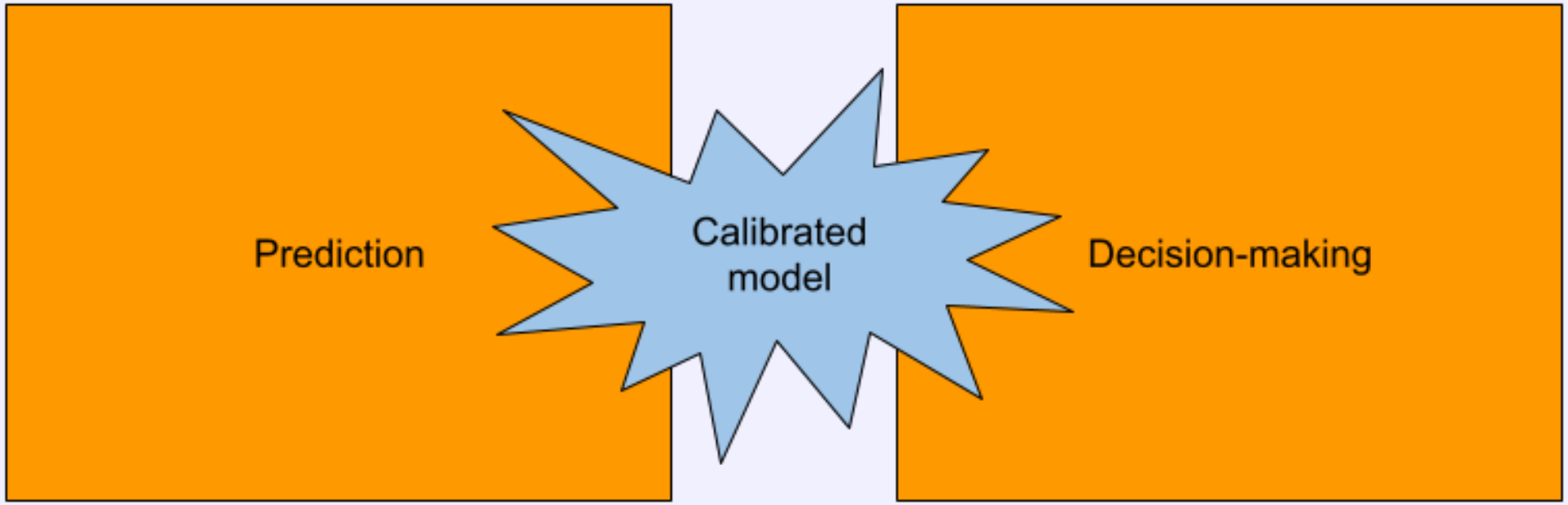
Beauty of calibration

Bobby's EC talk "lobotomizing" prediction from decision-making



Beauty of calibration

(My hasty replication of Bobby's motivating figure)

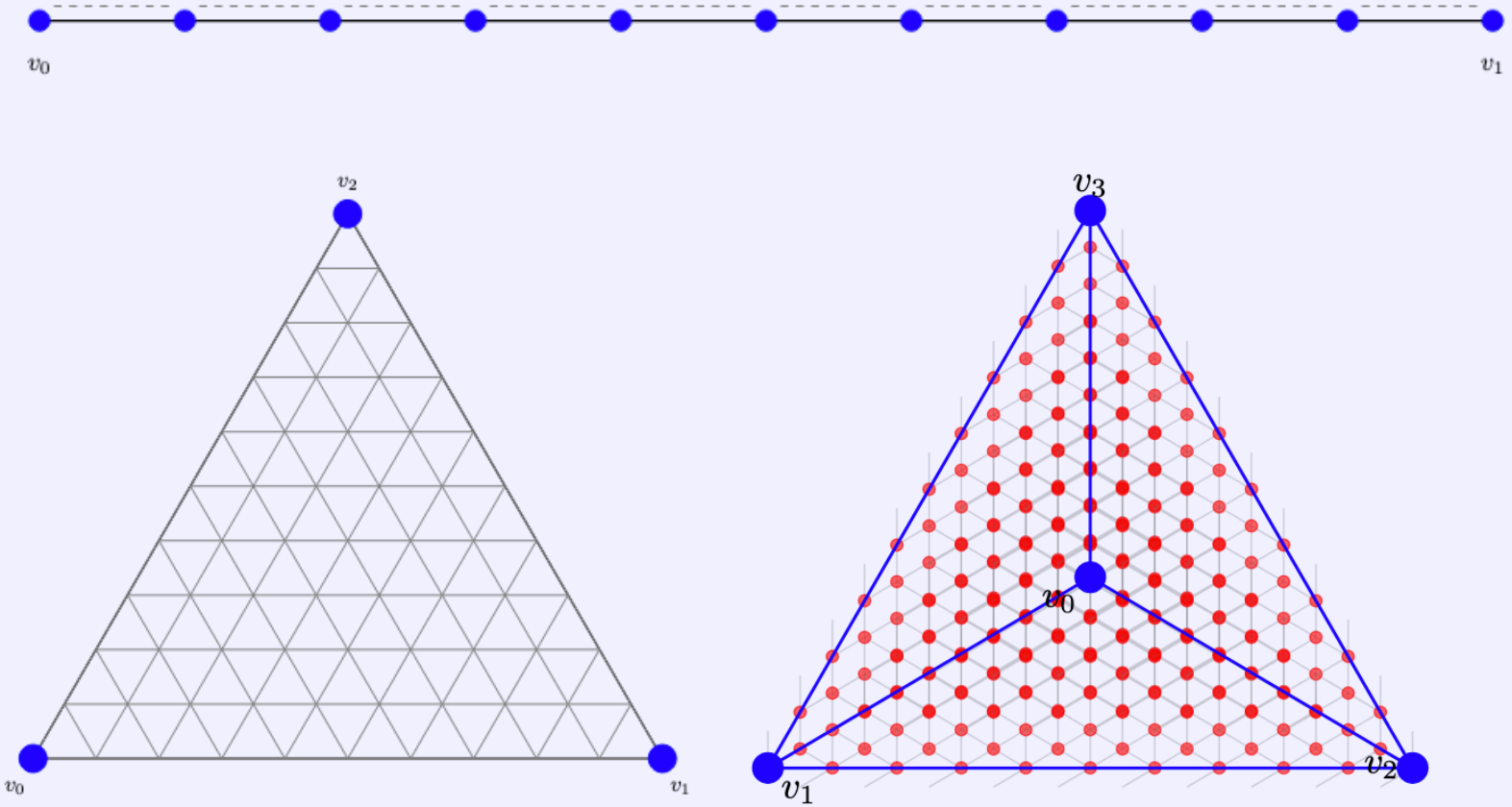


Challenges of calibration (especially for multiclass)

Complexity grows in number of bins... which is exponential in number of outcomes

Challenges of calibration (especially for multiclass)

Complexity grows in number of bins... which is exponential in number of outcomes



The core question

How to define calibration in muticlass settings without fully losing all of this beauty?

Don't need to propose more definitions... but can we unify currently existing definitions?

Calibration definitions evaluated

Distribution calibration w.r.t. a decision γ

- NRRX23, RS24, GPSW17, GR21

Property calibration (aka Γ-calibration)

- V84, NR23, GR23

Decision calibration w.r.t. a set of loss functions $\mathcal{L}$

- ZKSME21, GHKRW22

> An incomplete list, but some of the most prevalent definitions

But first: properties

A property is a function mapping probability distributions (over outcomes) $\mathcal{P} \subseteq \Delta(\mathcal{Y})$ to decisions

- Differentiate between a *optimization-level* property $\Gamma : \mathcal{P} \to \mathbb{R}^d$ and a *decision-level* property $\gamma : \mathcal{P} \to \mathcal{R}$
- If $\mathcal{R}$ is discrete, direct optimization intractible, but often considering only a finite set of decisions.

Observe: properties appear in first two definitions

Distribution calibration w.r.t. a decision γ

> Emphasis on decision-level properties

Property calibration (aka Γ-calibration)

> Focuses on both optimization and decision-level properties

Decision calibration w.r.t. a set of loss functions $\mathcal{L}$

> Properties also connected to losses, more on that offline.

An upfront(ish) caveat

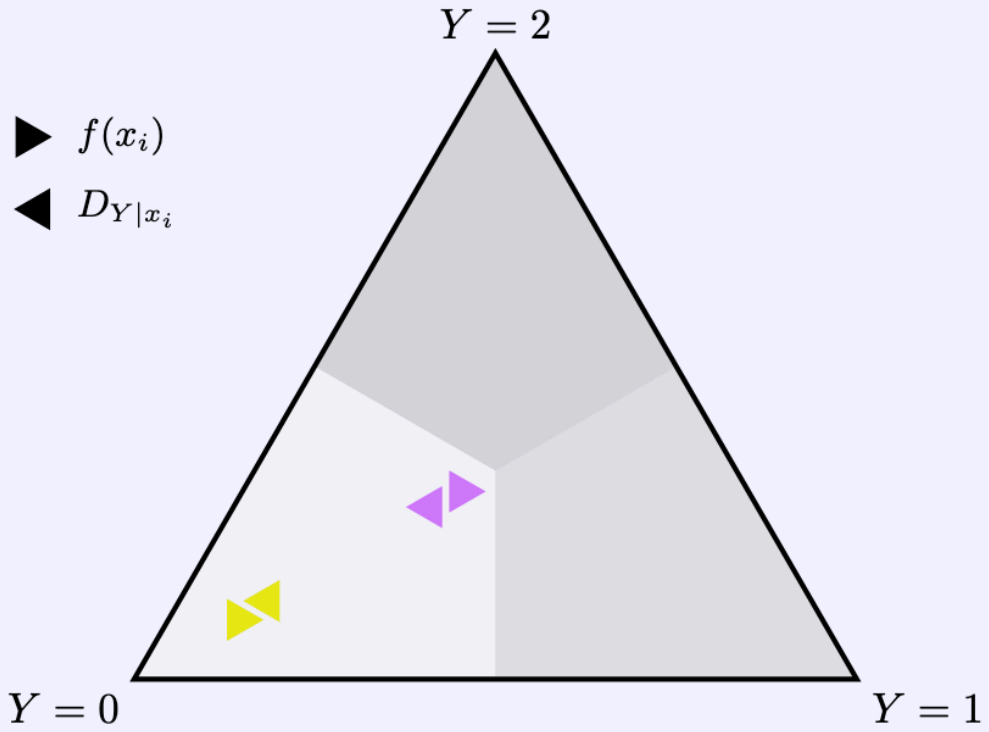
For simplicity, I will just discuss definitions of exact calibration.
Results for relationships in the approximate case (for continuous properties) are found in the paper.

However, if properties are discrete (e.g., mode, ranking), approximate calibration is non-trivial to define.

Defining calibration in multiclass settings

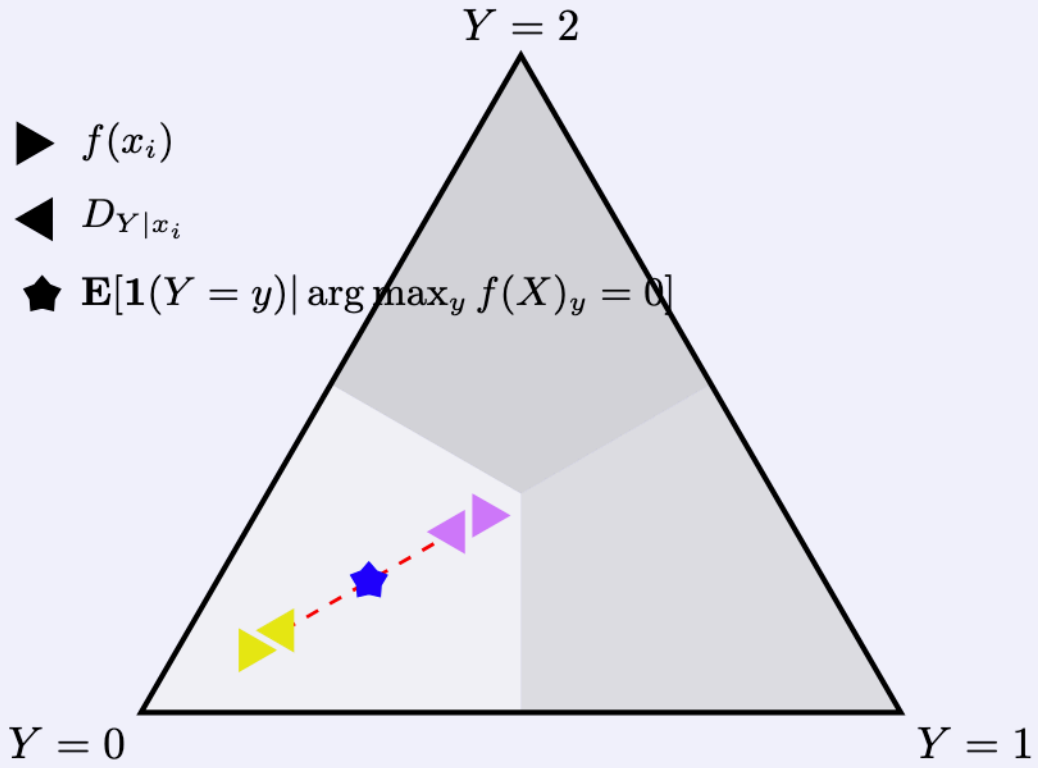
Distribution calibration with respect to a decision property γ

> Let $f : \mathcal{X} \to \mathcal{P}$ be a distribtuional predictor. Then f is *distribution calibrated* with respect to a decision property γ on distribution D if, for all $r \in \mathbf{im}(\gamma \circ f)$,
> $\mathbb{E}_D[\mathbf{1}(Y = y)|\gamma \circ f(X) = r] = \mathbb{E}_D[f_y(X)|\gamma \circ f(X) = r]$



Distribution calibration with respect to a decision property γ

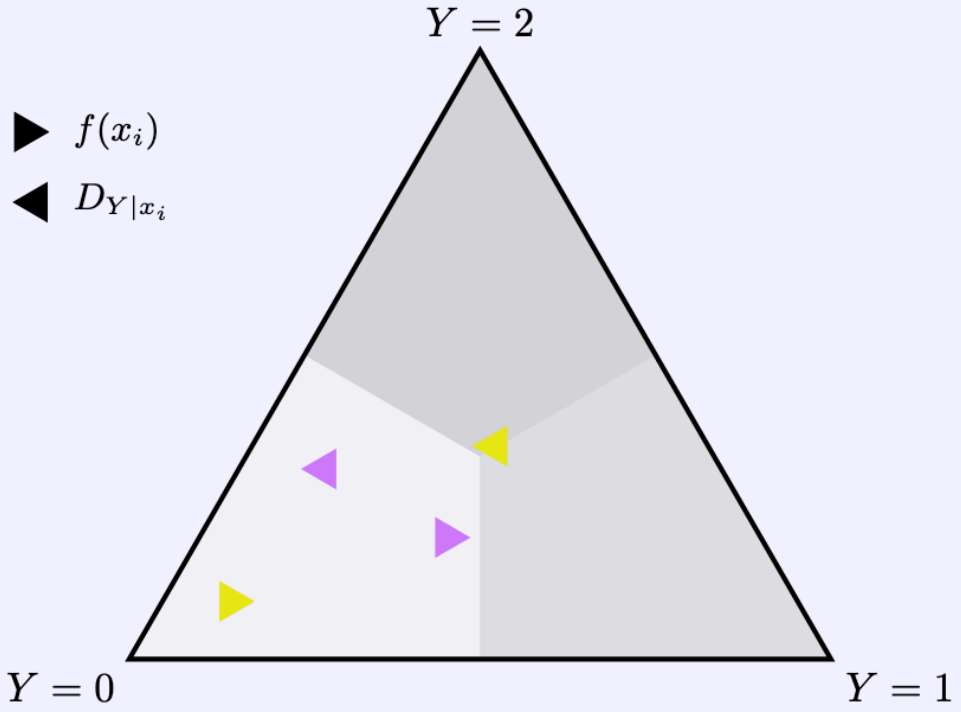
> Let $f : \mathcal{X} \to \mathcal{P}$ be a distribtuional predictor. Then f is *distribution calibrated* with respect to a decision property γ if, for all $r \in \mathbf{im}(\gamma \circ f)$,
> $\mathbb{E}_D[\mathbf{1}(Y = y)|\gamma \circ f(X) = r] = \mathbb{E}_D[f_y(X)|\gamma \circ f(X) = r]$



Property calibration (Γ-calibration)

> Suppose D is a data distribution over $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathcal{X} \to \mathcal{R}$ be a Γ-predictor for the property $\Gamma : \mathcal{P} \to \mathcal{R}$. The predictor f is Γ-calibrated on D if, for every $r \in \mathbf{im}(f)$, we have $\Gamma(D_{Y|f(X)=r}) = r$

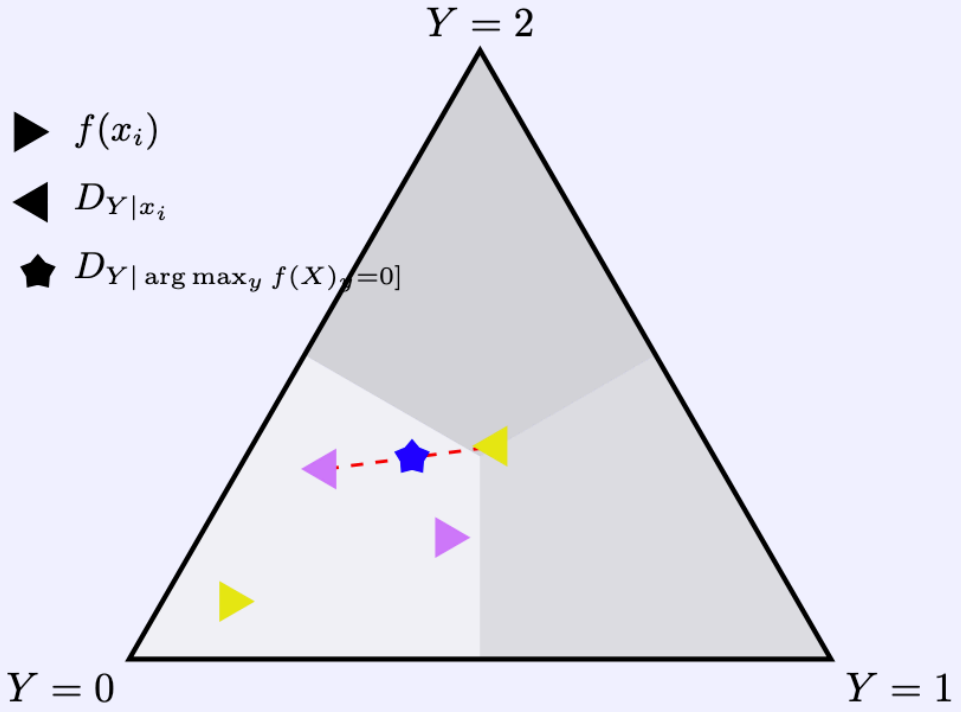
- Need to be careful with approximate calibration here-- open work!



Property calibration (Γ-calibration)

> Suppose D is a data distribution over $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathcal{X} \to \mathcal{R}$ be a Γ-predictor for the property $\Gamma : \mathcal{P} \to \mathcal{R}$. The predictor f is Γ-calibrated on D if, for every $r \in \mathbf{im}(f)$, we have $\Gamma(D_{Y|f(X)=r}) = r$

- Need to be careful with approximate calibration here-- open work!



Decision calibration

> Let $\mathcal{L}$ be a set of loss functions whose minimizer is the same property Γ. Moreover, let $f : \mathcal{X} \to \mathcal{P}$ be a distribution predictor. Then f is decision calibrated with respect to $\mathcal{L}$ if, for all $\ell \in \mathcal{L}$, we have $\mathbb{E}_{(X,Y)\sim D}\ell(\Gamma(f(X)), Y) = \mathbb{E}_{\hat{Y}\sim f(X)}\ell(\Gamma(f(X)), \hat{Y})$

Connections to loss outcome indistinguishability!

Semantic desiderata of calibration

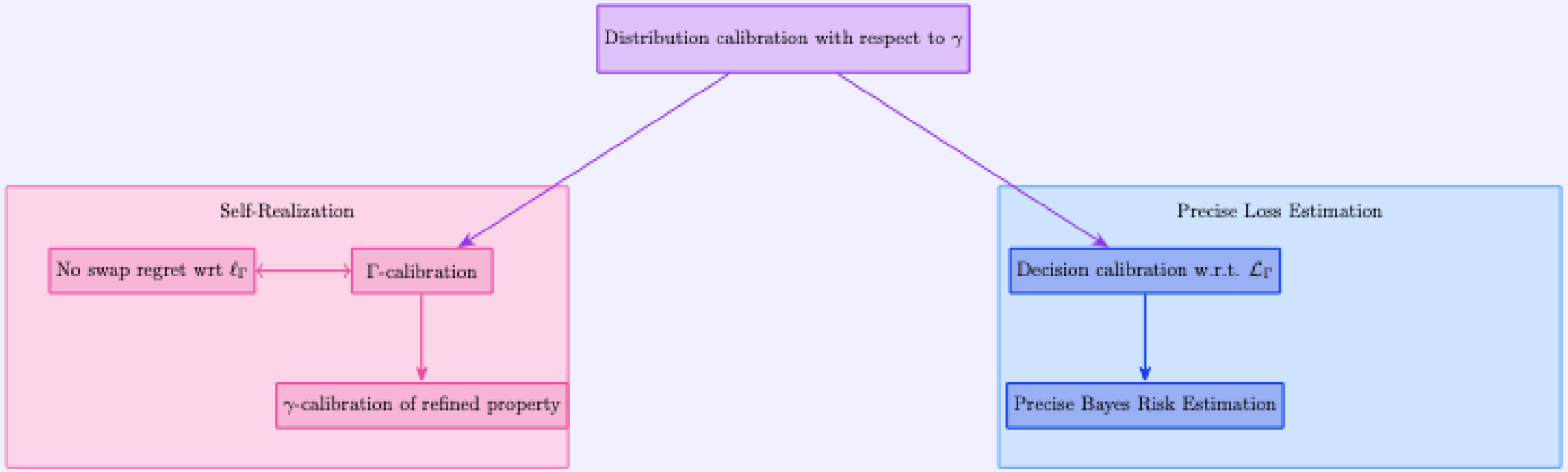
Self-realization:

> For the instances where p is forecasted, the actual values can be summarized as close to p

Precise loss estimation:

> The forecasted values let one provide estimates of incurred losses (for certain loss functions) close to the actual materized losses

Proposed hierarchy



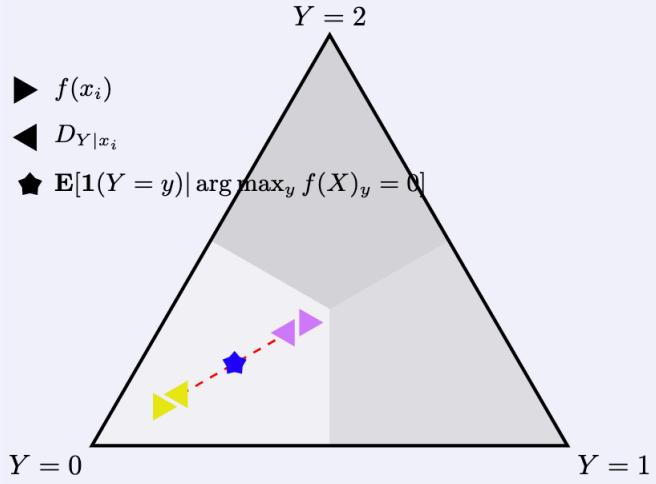
Γ-calibration as a prototypical definition of self-realization

> Intuitively: When I predict that r is the mode, it should be the mode.

Distribution calibration with respect to Γ implies Γ-calibration

Let Γ be a property with convex level sets[1], and D a distribution on $\mathcal{X} \times \mathcal{Y}$. Moreover, assume $f : \mathcal{X} \to \mathcal{P}$ is a distribution predictor which is distribution calibrated with respect to Γ, with $|\mathbf{im}(f)| < \infty$. Then $\Gamma \circ f$ is Γ-calibrated.

> Proof idea: since level sets are convex, $D_{Y|\Gamma \circ f(X)=r}$ is a convex combination of distributions in the level set Γ_r.



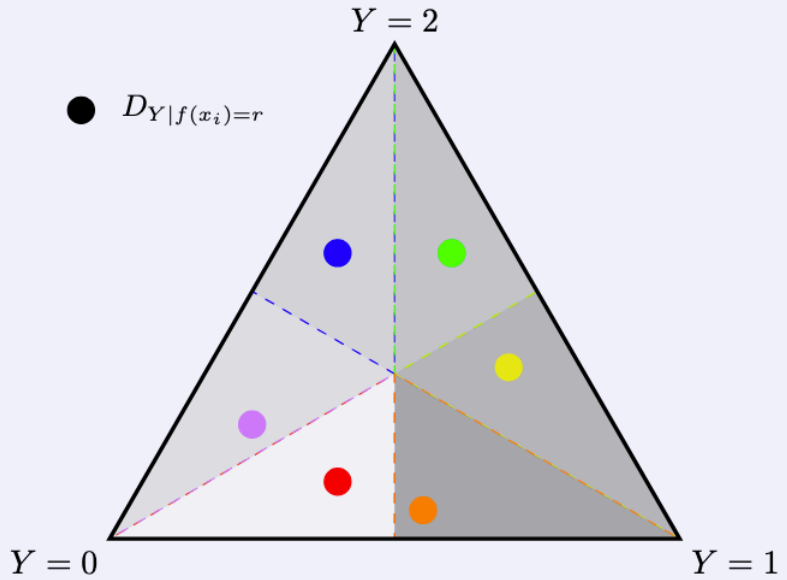
[1] this assumption is implied by the *elicitability* of a property, connecting a property to a loss

More generally: Γ-calibration is inherited by refined properties

Consider the gap between an *optimization-level* property Γ and a *decision-level* property γ.

Let Γ be a property, and f a predictor which is Γ-calibrated. For every elicitable property γ which is *refined* by Γ, the γ-predictor $\psi \circ f$ is γ-calibrated.

> Proof idea: still taking convex combinations of items staying in the level set

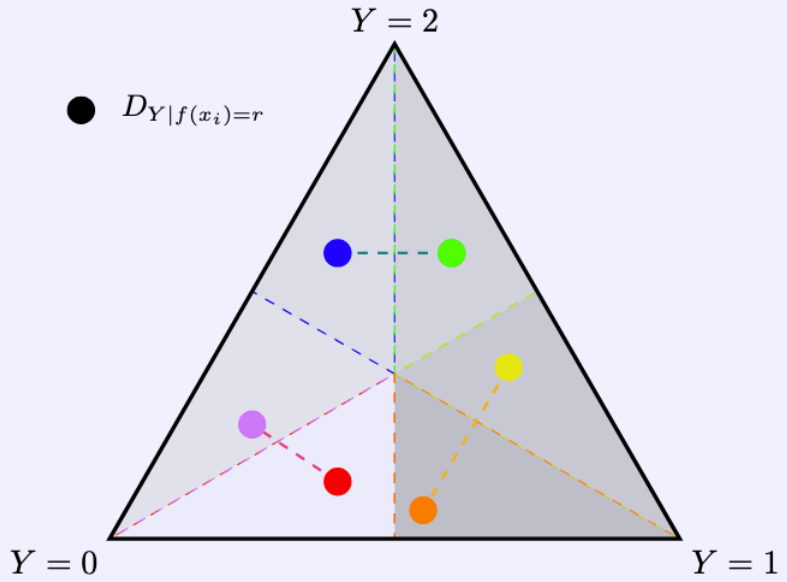


More generally: Γ-calibration is inherited by refined properties

Consider the gap between an *optimization-level* property Γ and a *decision-level* property γ.

Let Γ be a property, and f a predictor which is Γ-calibrated. For every elicitable property γ which is *refined* by Γ, the γ-predictor $\psi \circ f$ is γ-calibrated.

> Proof idea: still taking convex combinations of items staying in the level set



Applications of Γ-calibration inheritance

- Rewrite distribution calibration as Γ calibration, where $\Gamma(p) = p$ is the identity
- Move from mean-calibration to truncated means
- Move from mode to ranking

> Gives us a framework to think about the upshot of lobotomies again! Think of the "lever" of the optimization-level property, which can be used for γ-calibrated decisions for any refined property

Decision calibration as a prototypical definition of precise loss estimation

> $\mathbb{E}_{(X,Y)\sim D}\ell(\Gamma(f(X)),Y) \approx \mathbb{E}_{\hat{Y}\sim f(X)}\ell(\Gamma(f(X)),Y)$, where ℓ is minimized in expectation by Γ.

> Suggests $f(X)$, even if wrong, predicts expected loss.

Distribution calibration (wrt Γ) $\implies$ decision calibration (wrt losses $\mathcal{L}$ eliciting Γ)

Let $\mathcal{Y}$ be finite, and $f : \mathcal{X} \to \mathcal{P}$ be a distribution-calibrated predictor with respect to Γ. Then f is decision calibrated with respect to the set $\mathcal{L} = \{\ell : \ell \text{ is minimized by } \Gamma(p) \text{ for all } p \in \mathcal{P}\}$.

> Proof idea: $D_{Y|\Gamma(f(X))=r}$ is the correct conditional distribution, and $\Gamma(D_{Y|\Gamma(f(X))=r})$ is plugged into the loss

Connections to actuarial definitions

Actuarial fairness FW24:

> The forecaster should offer actuarially fair insurance for the uncertain loss [...]. Actuarial fairness here means that in the long run, the forecaster neither loses nor profits from offering insurance under the data model.

Can be considered on scales ranging from *individual* to *average* precise loss estimation similar to multicalibration.

Self-realization implies precise loss estimation

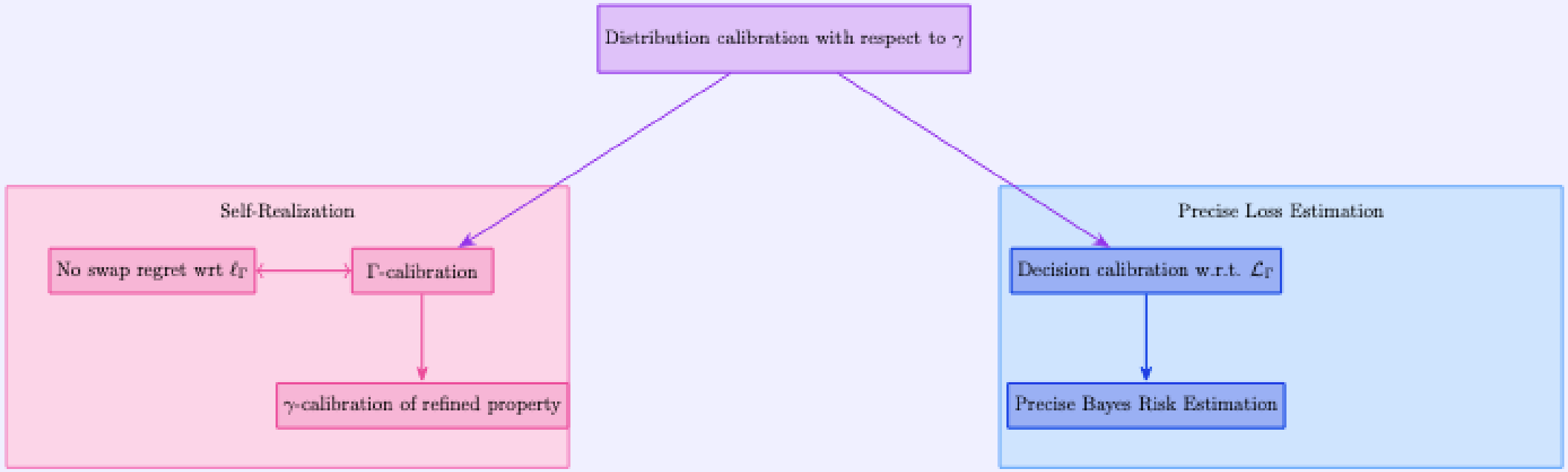
Given a loss ℓ, the Bayes pair $(\Gamma_\ell, \Theta_\ell)$ is defined by

$\Gamma_\ell(P) = \arg\min_r \mathbb{E}_{Y\sim P}\ell(r, Y)$, and
$\Theta_\ell(P) = \min_r \mathbb{E}_{Y\sim P}\ell(r, Y)$.

Let $\Phi := (\Gamma_\ell, \Theta_\ell)$ be a Bayes pair corresponding to loss ℓ. If a Φ- predictor f is (approximately) Φ-calibrated, then it is (approximately) a precise Bayes risk estimator.

> But not the other way around! There exist PLEs that are not property calibrated for any of (a) the pair $(\Gamma_\ell, \Theta_\ell)$, (b) Γ_ℓ alone, or (c) Θ_ℓ alone.

Recall, the hierarchy



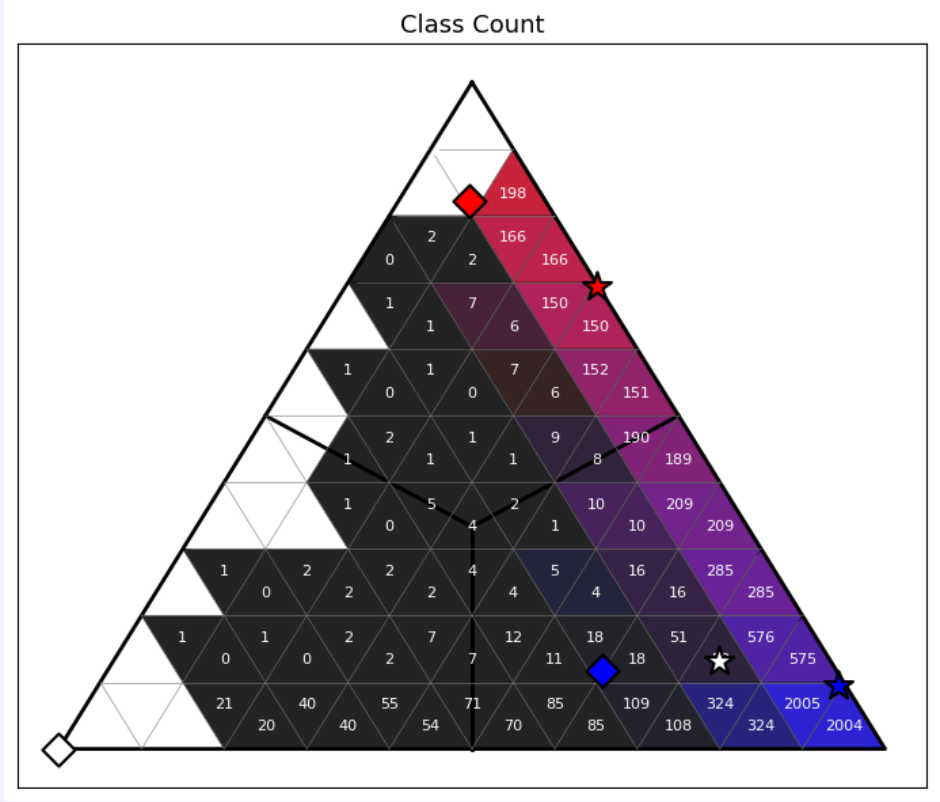
Empirical comparison

Summer work by BC undergrad Pat Lanza 

Home Mortgage Disclosure Act data

- 3 classes:
  - Rejected from loan (red)
  - Granted loan, but application withdrawn (black)
  - Granted and received loan (blue)

Empirical example



$$\star = \mathbb{E}[f_y(X)|\Gamma(f(X)) = r], \diamond = \mathbb{E}[\mathbf{1}(Y = y)|\Gamma(f(X)) = r]$$

Can visibly diagnose non-$\mathbf{mode}$-calibration!

Open questions

- Can we "smooth" discrete decision properties into non-trivial optimization properties?
  - What does this mean for approximate property calibration of the decision property?
- Relationship to CDL/CDR?
- Come chat! or email: finocch@bc.edu