Decision from Suboptimal Classifiers: Excess Risk Pre- and Post-Calibration

Alexandre Perez-Lebel
Soda, Inria Saclay, France
Stanford University, USA
Fundamental Technologies*, USA

Gael Varoquaux
Soda, Inria Saclay, France

Sanmi Koyejo
Stanford University, USA

Matthieu Doutreligne
Haute Autorité de Santé, France

Marine Le Morvan
Soda, Inria Saclay, France

Abstract

Probabilistic classifiers are central for making informed decisions under uncertainty. Based on the maximum expected utility principle, optimal decision rules can be derived using the posterior class probabilities and misclassification costs. Yet, in practice only learned approximations of the oracle posterior probabilities are available. In this work, we quantify the excess risk (a.k.a. regret) incurred using approximate posterior probabilities in batch binary decision-making. We provide analytical expressions for miscalibration-induced regret (R^{CL}), as well as tight and informative upper and lower bounds on the regret of calibrated classifiers (R^{GL}). These expressions allow us to identify regimes where recalibration alone addresses most of the regret, and regimes where the regret is dominated by the grouping loss, which calls for post-training beyond recalibration. Crucially, both R^{CL} and R^{GL} can be estimated in practice using a calibration curve and a recent grouping loss estimator. On NLP experiments, we show that these quantities identify when the expected gain of more advanced post-training is worth the operational cost. Finally, we highlight the potential of multicalibration approaches as efficient alternatives to costlier fine-tuning approaches.

1 Introduction

Whether it's marking a financial transaction as fraudulent, or deciding if a suspected cancer warrants a biopsy, making a decision involves carefully weighting the inconvenience of false positives (e.g. an invasive biopsy on a healthy patient) with that of false negatives (e.g. delaying cancer treatment due to a missed diagnosis). Often, the true outcome cannot be deterministically characterized (e.g. in medicine, Sox et al., 2013, sec 3.1.1). A rational decision-maker thus seeks the decision that offers the best harm–benefit tradeoff according to its preferences and the probabilities of each outcome.

In decision theory (Peterson, 2017; Kochenderfer, 2015), the Maximum Expected Utility Principle (or similarly, the Minimum Expected Cost Principle) offers a framework for optimal decisions. By using the class-conditional probabilities $P(Y|X)$ of the outcome Y given the input data X to quantify the uncertainty, along with the utilities (or costs) associated with decisions, one can derive the decision that maximizes utility. In a binary decision setting, the optimal decision for a record x is 1 whenever the probability of the corresponding outcome $\mathbb{P}(Y=1|X=x)$ is above a threshold $t^\star$ function of the utilities (Elkan, 2001).

The optimal decision depends on both the utilities and the *oracle* class-conditional probabilities $P(Y|X)$. In practice, the utilities can be defined by an expert, it is a task in itself (Sox et al., 2013, chap. 8), (Kochenderfer, 2015, sec. 3.1.4-5). However, the oracle probabilities are unknown and must be estimated *e.g.* using a learned probabilistic classifier within the machine learning framework. As relying on approximate probabilities affects the resulting decisions, it is essential to choose a model that leads to the best possible decisions.

Models are often selected based on common metrics

[*] Current affiliation

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

such as accuracy, AUC or Brier score. Yet, a model's high accuracy is no guarantee of its ability to improve subsequent decisions, and nor are the AUC or Brier score (Localio and Goodman, 2012). Model calibration is also known to be desirable and histogram binning—a well-known recalibration method—was introduced specifically to enhance decision-making by calibrating class-conditional probabilities (Zadrozny and Elkan, 2001a). Van Calster et al. (2019) highlighted the importance of assessing calibration when using estimated probabilities for clinical decision-making. Nonetheless, it remains unclear what degree of calibration is necessary for a model to be suitable for production or for preferring one model over another.

Validating decision-rules in practice is crucial to ensure that AI does more good than harm in production. For this purpose, decision-analytic measures such as Expected Utility and Net Benefit (Vickers et al., 2016) can be used, with the latter gaining traction in medical communities. In this work, we *investigate how inaccuracies in the estimated class-conditional probabilities translate into regret*, i.e., into an expected utility lower than the best possible expected utility for the given task. The interplay between decision-analytic measures and quantifications of class-conditional inaccuracies has not been thoroughly studied. Most research has focused on mitigating the detrimental effects of miscalibration on resulting decisions (Zhao et al., 2021; Rothblum and Yona, 2022), while Van Calster and Vickers (2015) evaluated its impact on Net Benefit through simulations.

This work We study how errors in estimated probabilities affect the optimality of decisions derived from these probabilities. Our theoretical results address practical questions such as: Is a pre-trained model suited to a new task? What is the simplest way to correct decisions based on a sub-optimal probabilistic classifier? How much will a model benefit from post-training? This approach is particularly valuable in the current trend of applying large pre-trained models to new tasks, rather than training models from scratch—*e.g.* using foundation models. Our contributions are:

- We formally describe how discrepancies between estimated probabilities and the underlying distribution $P(Y|X)$ affect the decision regret. In particular, we give an analytical expression for the regret induced by miscalibration, as well as *tight* and *informative* upper and lower bounds on the regret of a re-calibrated classifier. These bounds are distribution-agnostic, model-agnostic, and solely controlled by the grouping loss, decision threshold, and re-calibrated probabilities.
- Using these bounds, we describe two regimes: one where recalibration is a cheap and effective post-training strategy and another where recalibration fails to mitigate the regret.
- We show that these scenarios can be successfully identified in practice. The bounds identify the cases where fine-tuning improves utility upon recalibration alone on NLP tasks. This enables a new model validation procedure to guide post-training.
- We investigate the potential of multicalibration and show it is a cost-effective and controllable alternative to fine-tuning, while concurrently diminishing regret compared to calibration alone.

2 Background

2.1 Related work

Cost-sensitive learning Cost-sensitive learning (Ling and Sheng, 2008; Fernández et al., 2018, chap. 4) focuses on minimizing expected costs rather than misclassification rates. This can be achieved through three main approaches: direct methods, such as modifying tree-based splitting criteria (Ling et al., 2004; Petrides and Verbeke, 2022; Fernández et al., 2018, sec. 4.4) or training losses to embed cost information (Chung et al., 2015); pre-processing methods, which alter the training set to account for the cost (Zadrozny et al., 2003; Ting, 1998); and post-processing approaches, including threshold adjustment and refining class-conditional probabilities. For threshold adjustment, Sheng and Ling (2006) select the threshold that minimizes expected costs on the training set, while recent work seeks to adapt the decision threshold to miscalibrated classifiers (Rothblum and Yona, 2022).

Unlike direct and pre-processing approaches, post-processing methods do not need model retraining when the costs change. This makes them appealing in a many settings: when the costs are not known at training time, change over time, when the model is too expensive to retrain (*e.g.* large pre-trained models), or when the model is not accessible (*e.g.* using an API). In this work, we focus on post-training methods in batch decision-making, and in particular on post-training the obtained class-conditional probabilities.

Calibration Refining class-conditional probabilities often involves recalibration. Calibration ensures that, on average, the predicted probabilities match the positive rate within groups of the same estimated probability. For instance, if a classifier estimates an 80% probability, then 80% of those predictions should be actual positive outcomes. Learned classifiers are often miscalibrated; boosted trees tend to be under-confident (Niculescu-Mizil and Caruana, 2005), whereas naive Bayes classifiers or modern neural networks tend to be over-confident (Guo et al., 2017). To address these issues, many recalibration techniques have been developed, including Platt scaling (Platt, 1999), histogram binning (Zadrozny and Elkan, 2001b), isotonic regression (Zadrozny and Elkan, 2002), or temperature scaling (Guo et al., 2017). Recalibrating is advocated for better decisions (Van Calster et al., 2019), and some recalibration methods are specifically framed in decision settings, *e.g.* multiclass (Zhao et al., 2021).

Beyond calibration: post-training Calibration, being a control on averages, does not control individual probabilities. A complete characterization of predicted probabilities is given by decomposing the expected loss, and thus prediction errors (Kull and Flach, 2015, 3.1 and 5.1):

$$\begin{array}{c}\text{Expected}\\\text{Loss}\end{array} = \underbrace{\begin{array}{c}\text{Calibration}\\\text{Loss}\end{array} + \begin{array}{c}\text{Grouping}\\\text{Loss}\end{array}}_{\text{Epistemic Uncertainty}} + \underbrace{\begin{array}{c}\text{Irreducible}\\\text{Loss}\end{array}}_{\text{Aleatoric Uncertainty}}, \quad (1)$$

Aleatoric uncertainty, stems from the randomness of the outcome Y and cannot be reduced even with an optimal model and infinite data. On the opposite, *epistemic uncertainty* is due to model imperfection and can be reduced with a better model (Hüllermeier and Waegeman, 2021); it is a good indicator of whether a model can be improved (Lahlou et al., 2021). The recalibration methods listed above only reduce the calibration loss in eq. (1). Multicalibration recently pushed further, notably for fairness (Hébert-Johnson et al., 2018), as well as calibration within groups (Kleinberg et al., 2016; Pfohl et al., 2022). More general post-training methods tackle the full error in eq. (1), *e.g.* stacking (Pavlyshenko, 2018), which learns a model on top of the output of another model, or fine-tuning, particularly useful with the advent of large pretrained models.

Measuring the grouping loss Miscalibration is well characterized (as with expected calibration error, ECE, Naeini et al., 2015), however it is only part of the epistemic error. Recently, Perez-Lebel et al. (2023) gave an estimator for the remainder, the grouping loss, showing that modern classifiers often exhibit grouping loss in real-world settings. The grouping loss can be seen as the loss of grouping together entities having different odds. Concretely, it measures the variance of the true individual probabilities within groups of same estimated probabilities. The estimator uses a partitioning of the feature space to estimate local averages of the true probabilities. This enables detecting dissimilar entities that were grouped together by the model, thus approximating the grouping loss.

2.2 Decision-making under uncertainty: definitions and notations

Decision theory In this article, we consider the classic setting of batch supervised machine learning and focus on the binary setting. Let (X, Y) a pair of jointly distributed random variables on $\mathcal{X} \times \{0,1\}$. Binary decision rules map each point in $\mathcal{X}$ to a decision in $\{0,1\}$. Let $U \in \mathbb{R}^{2\times 2}$ be a matrix of utilities, where U_{ij} is the utility of predicting i when the true outcome is $Y = j$.[*] The expected utility of a decision rule $\delta : \mathcal{X} \to \{0,1\}$ is defined for $x \in \operatorname{supp} X$ as:[*]

$$\text{Pointwise} \quad \mathrm{EU}(\delta, x) \triangleq \mathbb{E}\left[U_{\delta(X),Y} \middle| X = x\right] \quad (2)$$

$$\text{Overall} \quad \mathrm{EU}(\delta) \triangleq \mathbb{E}_X[\mathrm{EU}(\delta, X)]\,. \quad (3)$$

Let $U_{\triangle} = U_{00} - U_{10} + U_{11} - U_{01}$. Decision theory (Elkan, 2001) states that the best decision rule, in the sense that it maximizes the (pointwise) expected utility,[*] is:

$$\delta^\star : x \mapsto \mathbb{1}_{\mathbb{P}(Y=1|X=x)\geq t^\star} \quad \text{where} \quad t^\star \triangleq \tfrac{U_{00}-U_{10}}{U_{\triangle}}. \quad (4)$$

The optimal decision thus amounts to assigning class 1 to x whenever $\mathbb{P}(Y=1|X=x) \geq t^\star$ and 0 otherwise. When misclassification costs are equal, *i.e.* $U_{00} = U_{11}$, and $U_{10} = U_{01}$, the optimal threshold is $t^\star = 0.5$, similar to regular cost-insensitive classification. We will denote by $f^\star : x \mapsto \mathbb{P}(Y=1|X=x)$ the (unknown) conditional probability of the positive class, and $f : \mathcal{X} \to [0,1]$ the probabilistic predictor estimating the probabilities $f^\star$. Without loss of generality, the set of binary decision rules $\delta : \mathcal{X} \to \{0,1\}$ can be parametrized with estimated class-conditional probabilities $f : \mathcal{X} \to [0,1]$ and a threshold $t \in [0,1]$ (Lem. B.1) as:

$$\delta_{f,t} : x \mapsto \mathbb{1}_{f(x)\geq t}. \quad (5)$$

With these notations, the optimal decision rule thus writes $\delta_{f^\star,t^\star}$. In the following, we consider a candidate decision rule $\delta_{f,t}$ with $t \in [0,1]$.

Calibration and grouping loss The recalibrated predictor is defined as $c \circ f$, where c is the calibration curve given by:

$$\text{Calibration curve} \qquad c : p \mapsto \mathbb{E}[Y | f(X) = p]\,. \quad (6)$$

A binary classifier f is calibrated when $c(p) = p$ for all $p \in \operatorname{supp} f(X)$. Finally, writing $\mathbb{V}$ the variance, the grouping loss associated to the squared loss is defined as:

$$\text{Grouping loss} \quad \mathrm{GL} : p \mapsto \mathbb{V}[f^\star(X) | f(X) = p]\,. \quad (7)$$

It can be thought of as the variance of the unknown individual probabilities around their mean $c(p)$ in the bin p of a reliability diagram (Perez-Lebel et al., 2023).

[*] Utility in the sense of von Neumann and Morgenstern (1944), which can be viewed as the opposite of a cost.

[*] We note supp the support of a random variable, *i.e.* the values for which the probability density function is nonzero.

[*] Equivalently, minimizes the expected cost.

3 Theory: regret on a decision

In this section, we establish a connection between the errors in the estimated probabilities (a form of uncertainty quantification) and the suboptimality (or regret) of the resulting decisions.

3.1 Regret decomposition of suboptimal decision rules

While the best decision rule is given by $\delta_{f^\star,t^\star}$, only imperfect estimates f of $f^\star$ are available in practice. It is thus of interest to characterize the optimal decision rule that can be achieved from f.

Proposition 3.1 (Best decision given estimated probabilities, B.4)**.** *Let $\mathcal{D}_f$ be the set of decision rules function of the estimated probabilities f. Then the calibrated probabilities thresholded at $t^\star$,*

$$\delta_{c\circ f,t^\star} : x \mapsto \mathbb{1}_{(c\circ f)(x)\geq t^\star}, \tag{8}$$

maximize the conditional expected utility over $\mathcal{D}_f$, i.e.,

$$\delta_{c\circ f,t^\star} \in \operatorname*{argmax}_{\delta\in\mathcal{D}_f} \mathrm{EU}(\delta|p) \qquad \textit{for all } p \in \operatorname{supp} f(X)$$

$$\begin{aligned} &\textit{with} \quad \mathrm{EU}(\delta|p) \triangleq \mathbb{E}[\mathrm{EU}(\delta, X)|f(X) = p] \\ &\textit{and} \quad \mathcal{D}_f = \{\delta_{g\circ f,t} : \quad g : [0,1] \to [0,1],\ t \in [0,1]\}\,. \end{aligned}$$

Prop. 3.1 shows that given an approximation f of the oracle probabilities $f^\star$, recalibrating the classifier and using $t^\star$ as threshold achieves the highest utility among decisions based on f only. While it is commonly admitted that recalibration is desirable to improve decisions, to the best of our knowledge, the optimality of $\delta_{c\circ f,t^\star}$ in batch binary decision-making has not been previously demonstrated. Note that $\mathcal{D}_f$ contains all possible decision rules taking only f as input (Lem. B.2). This includes recalibration but excludes decision rules based on both $f(x)$ and x. This prevents adjusting for the grouping loss, thereby incurring regret relative to $\delta_{f^\star,t^\star}$.

Since $\delta_{c\circ f,t^\star}$ is the best decision over a subset $\mathcal{D}_f$ of all possible decision rules, it holds for $p \in \operatorname{supp} f(X)$ that:

$$\underbrace{\mathrm{EU}(\delta_{f,t}|p)}_{\text{Naive}} \leq \underbrace{\mathrm{EU}(\delta_{c\circ f,t^\star}|p)}_{\text{Recalibrated}} \leq \underbrace{\mathrm{EU}(\delta_{f^\star,t^\star}|p)}_{\text{Oracle}}. \tag{9}$$

Ranking (9) leads us to define the conditional *calibration regret* $R^{\mathrm{CL}}_{f,t}(p)$ as the expected utility gap between the best decision given f and the candidate decision: $R^{\mathrm{CL}}_{f,t}(p) \triangleq \mathrm{EU}(\delta_{c\circ f,t^\star}|p) - \mathrm{EU}(\delta_{f,t}|p)$. $R^{\mathrm{CL}}_{f,t}$ quantifies the regret of using a miscalibrated classifier rather than a calibrated one. Similarly, we define the conditional *grouping regret* $R^{\mathrm{GL}}_{f}(p)$ as the expected utility gap between the oracle decision and the best decision given f: $R^{\mathrm{GL}}_{f}(p) \triangleq \mathrm{EU}(\delta_{f^\star,t^\star}|p) - \mathrm{EU}(\delta_{c\circ f,t^\star}|p)$. R^{GL}_{f} quantifies the regret of using the recalibrated classifier $c \circ f$ instead of oracle class-conditional probabilities $f^\star$. The conditional *regret* $R_{f,t}(p)$ of the candidate decision to the oracle decision $\delta_{f^\star,t^\star}$ is defined as $R_{f,t}(p) \triangleq \mathrm{EU}(\delta_{f^\star,t^\star}|p) - \mathrm{EU}(\delta_{f,t}|p)$ and can naturally be decomposed as:

$$R_{f,t}(p) = \underbrace{R^{\mathrm{CL}}_{f,t}(p)}_{\geq 0} + \underbrace{R^{\mathrm{GL}}_{f}(p)}_{\geq 0}. \tag{10}$$

The overall regrets can then be obtained by integrating over p. Ranking (9) and decomposition (10) trivially hold marginally. We note $R_{f,t}$, $R^{\mathrm{CL}}_{f,t}$ and R^{GL}_{f} the marginal counterparts. Importantly, the literature often focuses on the calibration regret $R^{\mathrm{CL}}_{f,t}$ (*e.g.*, Zhao et al., 2021; Noarov et al., 2023, "type regret"), which is blind to the grouping regret. In this work, we consider the full regret $R_{f,t}$ between the candidate and oracle decisions. In general, R^{GL}_{f} is nonzero due to the grouping loss of the estimated probabilities f. In the following sections 3.2–3.3 we provide an analytical expression for the calibration regret $R^{\mathrm{CL}}_{f,t}(p)$ as well as bounds on the grouping loss regret $R^{\mathrm{GL}}_{f}(p)$.

3.2 Expression of the regret stemming from miscalibration

The calibration regret can be estimated using the calibration curve as described in Prop. 3.2.

Proposition 3.2 (Expression of the calibration regret, B.5.1)**.** *For all $p \in \operatorname{supp} f(X)$,*

$$R^{\mathrm{CL}}_{f,t}(p) = \begin{cases} U_\Delta |c(p) - t^\star| & \textit{if } \mathbb{1}_{c(p)\geq t^\star} \neq \mathbb{1}_{p\geq t} \\ 0 & \textit{otherwise} \end{cases}. \tag{11}$$

$R^{\mathrm{CL}}_{f,t}$ scales as the distance between $t^\star$ and the calibration curve c in areas of disagreements between $\delta_{f,t}$ and $\delta_{c\circ f,t^\star}$ (red area in Fig. 1a). Whenever $\delta_{f,t}$ and $\delta_{c\circ f,t^\star}$ agree everywhere, $R^{\mathrm{CL}}_{f,t} = 0$. This is in particular the case for calibrated probabilities with threshold $t^\star$ (Fig. 1c). Yet, it is not necessary for f to be calibrated to imply $R^{\mathrm{CL}}_{f,t} = 0$, as the decision from a miscalibrated classifier $\delta_{f,t}$ can agree everywhere with $\delta_{c\circ f,t^\star}$ (Fig. 1b). Prop. 3.3 shows how to achieve this by threshold adjustment.

Proposition 3.3 (Adjusting the threshold t_f, B.5.3)**.** *For all $t^\star \in [0,1]$ let $t_f \in c^{-1}(\{t^\star\})$ if it exists, otherwise let $t_f \triangleq \inf\{t : c(t) \geq t^\star\}$. If c is monotonic non-decreasing, then thresholding f at t_f,* i.e. *δ_{f,t_f}, achieves zero miscalibration regret: $R^{\mathrm{CL}}_{f,t_f} = 0$.*

When the calibration curve is non-decreasing (which is assumed by isotonic regression), Prop. 3.3 shows that instead of recalibrating the classifier f, one can achieve

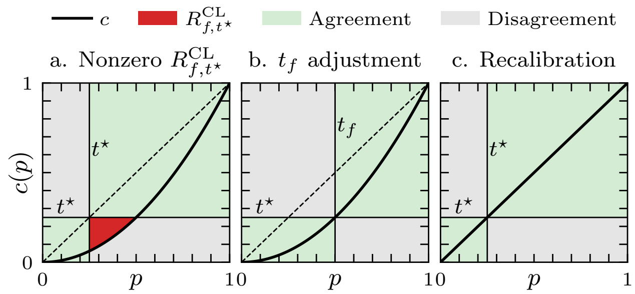


Figure 1: **Impact of miscalibration on the regret $R^{\mathrm{CL}}_{f,t}$.** (a) The oracle decision $p \mapsto \mathbb{1}_{p \geq t^\star}$ applied on miscalibrated estimated probabilities f, that is $\delta_{f,t^\star}$, yields a non zero regret $R^{\mathrm{CL}}_{f,t^\star}$ within areas of disagreement with amount $|c - t^\star|$ (red area). The regret $R^{\mathrm{CL}}_{f,t}$ can be reduced to 0 either by adapting the decision to a new threshold $t_f = c^{-1}(t^\star)$, that is δ_{f,t_f} (b), or by recalibrating the estimated probabilities and using $t^\star$ as threshold, that is $\delta_{c \circ f,t^\star}$ (c) (Prop. 3.3).

zero miscalibration regret by adjusting the threshold: $\delta_{c \circ f,t^\star} = \delta_{f,c^{-1}(t^\star)}$ (Fig. 1b). Note that these two approaches are equally costly as calibrated probabilities must be estimated in both cases. Calibration however has the advantage of being usable even if the calibration curve is not monotonic. The dual formulation between optimal threshold on f and calibration brings another insight: if the calibrated probabilities still hide over- or underconfident subgroups (i.e., nonzero grouping loss), it could lead to suboptimal decisions that cannot be tackled simply by setting a better threshold.

3.3 Bounding the grouping-loss-induced regret of the calibrated classifier

Calibration does not directly control individual probabilities $f^\star$: within a bin p, the individual probabilities $f^\star(x)$ may vary around their mean $c(p)$ with a variance $\mathrm{GL}(p)$. In what follows, we address the regret arising from this variance. Unlike miscalibration, there is no one-to-one correspondence between a specific grouping loss level and the resulting regret R^{GL}_f. Specifically, $\mathrm{GL}(p)$ represents the variance of individual probabilities, while R^{GL}_f depends on their distribution. As shown by the expression of the GL-induced regret (Lem. B.8), this regret depends on the proximity of individual probabilities $f^\star(x)$ to the threshold $t^\star$, as well as on the region of agreement between $\delta_{f^\star,t^\star}$ and $\delta_{c(p),t^\star}$. Hence for the same GL, the regret R^{GL}_f can vary. This is why we provide lower and upper bounds on R^{GL}_f. Our bounds are derived by identifying, among all distributions $\mathbb{P}(f^\star(X)|f(X) = p)$ with a fixed mean $c(p)$ and variance $\mathrm{GL}(p)$, the distributions that result in the lowest and highest regrets.

Theorem 3.4 (Grouping regret lower bound, B.8)**.** *The conditional grouping regret is lower bounded for all $p \in \operatorname{supp} f(X)$ as $R^{\mathrm{GL}}_f(p) \geq L^{\mathrm{GL}}_f(p)$, by:*

$$L^{\mathrm{GL}}_f(p) \triangleq U_{\Delta} \left[\mathrm{GL}(p) - V_{\min}(p)\right]_+ \tag{12}$$

with: $[\cdot]_+ = \max\{\cdot, 0\}$

$$\textit{and:}\;\; V_{\min}(p) \triangleq \begin{cases} (1-c(p))\,(c(p)-t^\star) & \textit{if } c(p) \geq t^\star \\ c(p)\,(t^\star - c(p)) & \textit{otherwise} \end{cases}.$$

Tightness. *The lower bound is tight. For any $p \in \operatorname{supp} f(X)$ for which f admits at least 3 antecedent values, and for all admissible mean $c(p) \in [0, 1]$ and variance $\mathrm{GL}(p) \in [0, c(p)(1 - c(p))]$, there exists a distribution of (X, Y) such that the conditional distribution $\mathbb{P}(f^\star(X)|f(X) = p)$ has mean $c(p)$ and variance $\mathrm{GL}(p)$, and the grouping regret attains its lower bound: $R^{\mathrm{GL}}_f(p) = L^{\mathrm{GL}}_f(p)$.*

Here $V_{\min}(p)$ represents the largest possible variance GL with zero regret. This is achieved by a distribution of the true probabilities $\mathbb{P}(f^\star(X)|f(X) = p)$ having all its mass on 0 and $t^\star$, or $t^\star$ and 1 depending on the side of $c(p)$ to $t^\star$. Such a distribution ensures that $\delta_{f^\star,t^\star}$ and $\delta_{c(p),t^\star}$ always agree, as individual probabilities always lie on the same side of the threshold as $c(p)$, thus guaranteeing zero regret.

Measuring a GL in the range $[0, V_{\min}]$ does not provide informative insights on the regret (Fig. 2), as it could correspond to a distribution with zero regret, like the

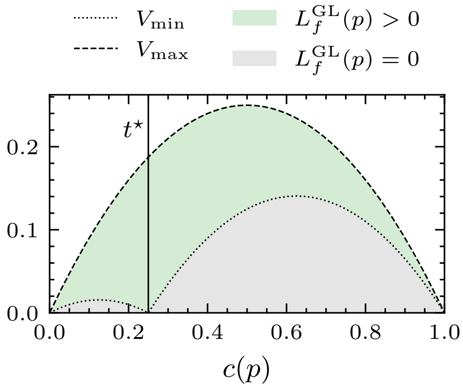


Figure 2: **Impact of the grouping loss on the minimal regret L^{GL}_f of a recalibrated classifier.** In a bin $p \in \operatorname{supp} f(X)$, the grouping loss exceeding $V_{\min}(p)$ incurs to the recalibrated classifier a nonzero regret $R^{\mathrm{GL}}_f(p)$ of at least $U_{\Delta} \left[\mathrm{GL}(p) - V_{\min}(p)\right]_+$ (Th. 3.4). Measuring a variance smaller than $V_{\min}(p)$ is not informative with respect to the grouping regret as there exists $(f^\star, f)$ where $\mathrm{GL}(p) = V_{\min}(p)$ and $R^{\mathrm{GL}}_f(p) = 0$. The variance cannot exceed $V_{\max}(p) \triangleq c(p)(1 - c(p))$. The informative area is highlighted in green.

one described above. Yet it is an "adversarial" distribution in the sense that most distributions with lower variance are still likely to incur some regret. Conversely, all distributions with a GL higher than $V_{\min}$ will necessarily incur regret proportional to $\mathrm{GL}(p) - V_{\min}(p)$: their high variance implies that some $f^\star$ values exist on the wrong side of the threshold, implying disagreements between $\delta_{f^\star,t^\star}$ and $\delta_{c(p),t^\star}$, and thus regret.

Interestingly, $V_{\min}$ equals 0 for $c = t^\star$ and is small for values close to $t^\star$ (Fig. 2). On the threshold, the presence of grouping loss necessarily incurs some regret. This highlights that the more the grouping loss occurs when the calibrated probability c is close to the decision threshold $t^\star$, the more likely it will incur regret. The grouping loss matters the most in regions of high uncertainty, *e.g.*, when the calibrated probabilities c are close to 0.5 with a threshold at $t^\star = 0.5$.

Theorem 3.5 (Grouping regret upper bound, B.9)**.** *The conditional grouping regret is upper bounded for all $p \in \operatorname{supp} f(X)$ as $R_f^{\mathrm{GL}}(p) \le U_f^{\mathrm{GL}}(p)$, by:*

$$U_f^{\mathrm{GL}}(p) \triangleq \tfrac{1}{2}U_\Delta\Big(\sqrt{\mathrm{GL}(p)+(c(p)-t^\star)^2}-|c(p)-t^\star|\Big)\,. \tag{13}$$

Tightness. *The upper bound is tight when $t^\star = \frac{1}{2}$. For any $p \in \operatorname{supp} f(X)$, and for all admissible mean $c(p) \in [0,1]$ and variance $\mathrm{GL}(p) \in [0, c(p)(1-c(p))]$, there exists a distribution of (X, Y) such that the conditional distribution $\mathbb{P}(f^\star(X)|f(X) = p)$ has mean $c(p)$ and variance $\mathrm{GL}(p)$, and the grouping regret attains its upper bound: $R_f^{\mathrm{GL}}(p) = U_f^{\mathrm{GL}}(p)$.*

This upper bound was obtained by finding the distribution $\mathbb{P}(f^\star(X)|f(X) = p)$, with fixed mean $c(p)$ and variance $\mathrm{GL}(p)$, that leads to the largest regret (details in the proof B.9). A more compact upper-bound immediately follows since $U_f^{\mathrm{GL}}(p) \le \frac{1}{2}U_\Delta\sqrt{\mathrm{GL}(p)}$, with equality when $c(p) = t^\star$. This upper bound scales as a squared root of the grouping loss, meaning that large GL opens the door to large R_f^{GL}, and small GL implies small GL-induced regret. In particular $\mathrm{GL}(p) = 0$ implies $R_f^{\mathrm{GL}}(p) = 0$.

L_f^{GL} and U_f^{GL} are informative. The upper and lower bounds are entirely defined by the calibration curve $c(p)$, the grouping loss $\mathrm{GL}(p)$, as well as the threshold $t^\star$. Fig. 3 plots the bounds for different values of these quantities. Importantly, it shows that the bounds are informative as the gap between them is not too large. It also illustrates that the GL-induced regret is larger when c is close to the decision threshold, and that it increases when GL increases.

We underline that these bounds are *distribution-agnostic* and *model-agnostic*, requiring no assumptions on either the data distribution, $f^\star$, or the probabilistic classifier f. Moreover, thanks to the GL estimator proposed by Perez-Lebel et al. (2023), all quantities involved in these bounds can be evaluated in practice.

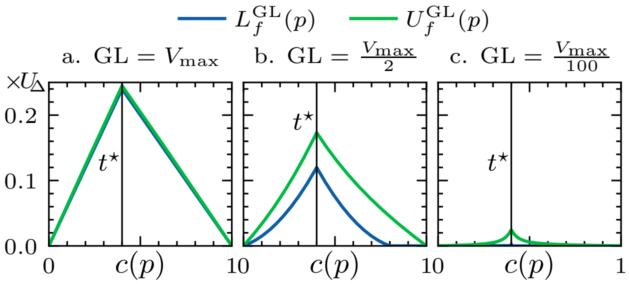


Figure 3: **Impact of the grouping loss on the bounds of the regret of the recalibrated classifier.** Lower and upper bounds $L_f^{\mathrm{GL}}(p)$ and $U_f^{\mathrm{GL}}(p)$ as a function of the calibrated probabilities $c(p) \in [0,1]$ for a bin $p \in \operatorname{supp} f(X)$, in three settings of grouping loss: maximal (a), intermediate (b) and small (c). The gap between the lower and upper bounds reduces when the grouping loss is high or low. $V_{\max} \triangleq c(p)(1 - c(p))$.

Definition 3.6 (Regret estimators)**.** Let $\hat{c} : [0,1] \to [0,1]$ and $\widehat{\mathrm{GL}} : [0,1] \to [0, \frac{1}{4}]$ be the estimates of c and GL. We define the plugin estimators of the conditional regrets and bounds for $p \in [0,1]$ as:

$$\hat{L}_f^{\mathrm{GL}}(p) \triangleq U_\Delta\left[\widehat{\mathrm{GL}}(p) - V_{\min}(p)\right]_+$$
$$\hat{U}_f^{\mathrm{GL}}(p) \triangleq \tfrac{1}{2}U_\Delta\Big(\sqrt{\widehat{\mathrm{GL}}(p)+(\hat{c}(p)-t^\star)^2}-|\hat{c}(p)-t^\star|\Big)$$
$$\hat{R}_f^{\mathrm{GL}}(p) \triangleq \tfrac{1}{2}(\hat{L}_f^{\mathrm{GL}}(p) + \hat{U}_f^{\mathrm{GL}}(p))$$
$$\hat{R}_{f,t}^{\mathrm{CL}}(p) \triangleq U_\Delta|\hat{c}(p) - t^\star|\mathbb{1}_{\mathbb{1}_{\hat{c}(p)\ge t^\star} \neq \mathbb{1}_{p \ge t}}$$
$$\hat{R}_{f,t}(p) \triangleq \hat{R}_{f,t}^{\mathrm{CL}}(p) + \hat{R}_f^{\mathrm{GL}}(p).$$

We note $\hat{L}_f^{\mathrm{GL}}$, $\hat{U}_f^{\mathrm{GL}}$, $\hat{R}_f^{\mathrm{GL}}$, $\hat{R}_{f,t}^{\mathrm{CL}}$, and $\hat{R}_{f,t}$ their counterpart obtained by averaging over the bins of $f(X)$.

Two regimes Comparing the conditional calibration and grouping regrets highlights two different regimes. Within a bin p, when $c(p)$ is close to $t^\star$, grouping loss matters more than miscalibration in terms of regret. When $c(p) = t^\star$, the effect of the grouping loss on the regret is maximal: $\mathrm{GL}(p) \le R_f^{\mathrm{GL}}(p) \le \frac{1}{2}\sqrt{\mathrm{GL}(p)}$ (Fig. 3). Conversely when $c(p)$ is further from $t^\star$, miscalibration leading to disagreement between $\delta_{f,t}$ and $\delta_{c\circ f,t^\star}$, if any, typically matters more (Eq. 11). On the overall population, when averaging across bins, these effects blend according to the distribution of $f(X)$. Conditional estimators (Def. 3.6) enables detecting these regimes.

3.4 Grouping Loss Adaptative Recalibration

Recalibration only addresses part of the overall regret $R_{f,t}$, leaving unchanged the grouping-loss part. The estimation of $\mathrm{GL}(p)$ from Perez-Lebel et al. (2023) involves finding regions in the input space $\mathcal{X}$ that explain the variance of the true probabilities $f^\star(X)$ whithin a bin $f(X) = p$. This work naturally motivates a multicalibration method to reduce grouping loss by using the same estimated region probabilities.

Definition 3.7 (GLAR)**.** Let $\mathcal{P} : \mathcal{X} \to \mathbb{R}$ be a partition of the feature space $\mathcal{X}$. We define the Grouping Loss Adaptative Recalibration (GLAR) of a function $f : \mathcal{X} \to \mathbb{R}$ relative to a partition $\mathcal{P}$ as:

$$\begin{aligned} f_{\mathcal{P}} \colon & \mathcal{X} \to [0,1] \\ & x \mapsto \mathbb{E}[Y | f(X) = f(x), \mathcal{P}(X) = \mathcal{P}(x)] \,. \end{aligned} \tag{14}$$

The GLAR correction consists of replacing the output of f with that of $f_{\mathcal{P}}$. GLAR provides a calibrated classifier $f_{\mathcal{P}}$ that has a lower grouping loss than the original classifier (Prop. B.10). Moreover, the decision based on the GLAR-corrected estimator $f_{\mathcal{P}}$, *i.e.* $\delta_{f_{\mathcal{P}},t^\star}$, yields a better expected utility than both the decision based on the initial classifier $\delta_{f,t}$ and the recalibrated classifier $\delta_{c\circ f,t^\star}$ (Prop. B.12). The choice of partition $\mathcal{P}$ is important. A trivial partition in one region would give histogram binning recalibration. A too fined-grained partition would give bad estimations because of a low number of samples per partition. GLAR has the notable advantage of reusing the partitions and local probabilities computed for the estimation of $\hat{R}^{\mathrm{GL}}_f$, and thus does not incur any additional cost once the regret is estimated. The implementation details of the method are given in Sec. C.3.

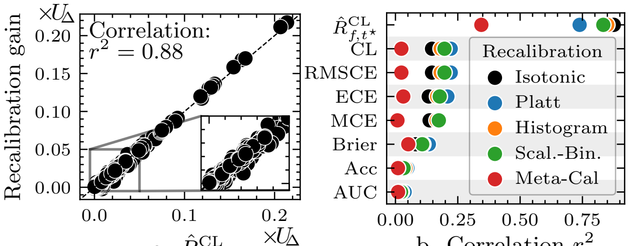
4 Experiments: validation of regret bounds and link with post-training

Settings For evaluation on practical scenarios, we consider a hate-speech detection task on real-world language datasets. From the perspective of a platform (*e.g.* a social network), failing to identify hate speech will incur a reputation cost, while wrongly identifying a text as hate speech will incur an opportunity cost (*e.g.* loss of content and users in the long term). Formally, these costs can be gathered into a 2×2 utility matrix U to be determined by the platform. The goal is then to solve a binary cost-sensitive decision-making problem with classes *hate speech* ($Y = 1$) and *no hate speech* ($Y = 0$). We investigate post-training of pretrained models on hate-speech datasets, with potential distribution shift. We benchmark 6 pre-trained models (Tab. 1) on 14 real-world datasets (Tab. 2), and 9 post-training methods. These include calibration methods, both classical (isotonic regression (Zadrozny and Elkan, 2002), Platt scaling (Platt, 1999), histogram binning (Zadrozny and Elkan, 2001b)) and more recent (Scaling-Binning (Kumar et al., 2019), Meta-Cal (Ma and Blaschko, 2021)); finetuning, where only the last layer is fine-tuned; stacking, where `scikit-learn`'s (Pedregosa et al., 2011) Random Forests or Gradient Boosted Trees are trained on the concatenation of the inputs with the probabilistic outputs of the pretrained model; and finally the GLAR correction (Sec. 3.4).

For each model, we extract the embedding space (usually the penultimate layer) and consider post-training from this representation to the class probability space. We draw utility matrices corresponding to 11 values of $t^\star$ in the range $[0.01, 0.99]$. We consider decision rules formed by thresholding estimated probabilities at $t^\star$, *i.e.* $\delta_{f,t^\star}$. Experimental details are given in Sec. C.3. Our experimental question is whether our regret estimators (Def. 3.6) are better than classical performance metrics (Brier score, AUC, accuracy, calibration errors defined in Sec. C.3) at identifying post-training gains.

$R^{\mathrm{CL}}_{f,t^\star}$ captures the gain of recalibration First, we estimate the calibration regret $R^{\mathrm{CL}}_{f,t^\star}$ of each model f using the estimator $\hat{R}^{\mathrm{CL}}_{f,t^\star}$ from Def. 3.6. We estimate c using an histogram binning with 15 equal-mass bins. We compare the expected gain of recalibration, $\hat{R}^{\mathrm{CL}}_{f,t^\star}$, to the gain obtained by recalibrating the pre-trained models using Isotonic Regression. Fig. 4a shows a near perfect identity relation between the estimated calibration regret $R^{\mathrm{CL}}_{f,t^\star}$, and the gain obtained by recalibrating the model (Pearson's correlation: $r^2 = 0.88$). Fig. 4b shows the correlation between $R^{\mathrm{CL}}_{f,t^\star}$ and the gain obtained with 4 other recalibration methods: Platt Scaling, Histogram Binning, Scaling-Binning, and Meta-Cal.

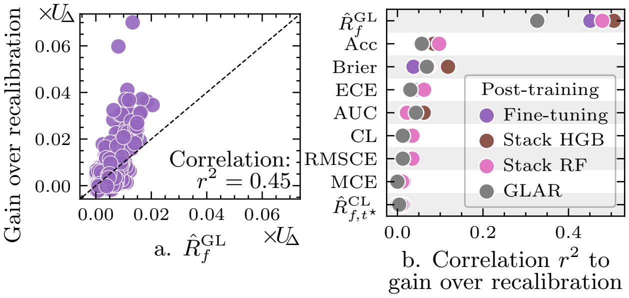
Miscalibrated models do not necessarily incur regret compared to the calibrated classifier. Indeed, the 4 cal-



a. $\hat{R}^{\mathrm{CL}}_{f,t^\star}$

b. Correlation r^2 to recalibration gain

Figure 4: $\hat{R}^{\mathrm{CL}}_{f,t}$ **captures the gain of recalibration.** (a) Gain in utility of isotonic recalibration versus the regret to the recalibrated classifier $\hat{R}^{\mathrm{CL}}_{f,t^\star}$, for each (model, dataset, $t^\star$). (b) Pearson's r^2 correlation of the gain in utility of each recalibration method with $\hat{R}^{\mathrm{CL}}_{f,t^\star}$ and other metrics.



a. $\hat{R}_f^{\mathrm{GL}}$

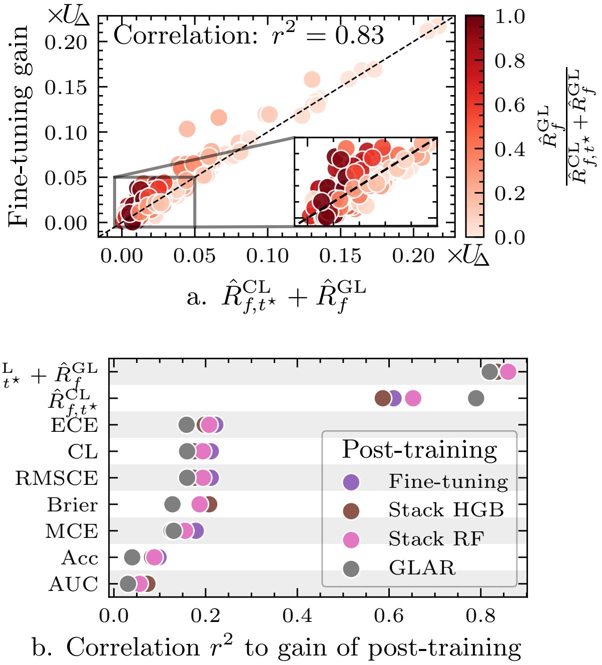
b. Correlation r^2 to gain over recalibration

Figure 5: **Gain of Post-Training on top of recalibration.** (a) Gain in utility of fine-tuning over isotonic recalibration versus $\hat{R}_f^{\mathrm{GL}}$, for each (model, dataset, $t^\star$). (b) Pearson's r^2 correlation of the gain in utility over isotonic recalibration with $\hat{R}_f^{\mathrm{GL}}$, $\hat{R}_{f,t^\star}^{\mathrm{CL}}$, and other metrics.

ibration error metrics (ECE, MCE, RMSCE, and CL) are very poorly correlated to the gain of recalibration for any of the 5 recalibration approaches (Fig. 4b, see Sec. C.4 for detailed figures). This is because for a given utility level $t^\star$, the model does not need to be calibrated to achieve $R^{\mathrm{CL}} = 0$ as shown in Sec. 3.2. The expected gain from recalibration is best measured by $\hat{R}_{f,t}^{\mathrm{CL}}$.

Gain of post-training on top of recalibration The calibrated classifier can be suboptimal due to the grouping loss and R_f^{GL}. We demonstrate this with 4 post-training methods (fine-tuning, stacking with boosted trees or random forest, and GLAR). We compare their gain in utility to the gain of isotonic recalibration (Fig. 5a). $\hat{R}_f^{\mathrm{GL}}$ is by far the metric that best explains the excess gain of post-training among all the other metrics. Across the 4 post-training methods, $\hat{R}_f^{\mathrm{GL}}$ has $r^2 \approx 0.5$ while calibration errors metrics (ECE, MCE, RMSCE, CL) or model-performance metrics (Brier, AUC) all have $r^2 \leq 0.1$ (Fig. 5b). The lower levels of correlation of $\hat{R}_f^{\mathrm{GL}}$ compared to the levels obtained for the calibration regret $\hat{R}_{f,t}^{\mathrm{CL}}$ in Fig. 4 are partly due to the fact that $\hat{R}_f^{\mathrm{GL}}$ is derived from bounds on the regret. For any value of R_f^{GL}, the regret can vary in a range, hence decreasing the correlation. The fact that $y \gtrsim x$ on Fig. 5a means that post-training yields a higher gain than what was given by $\hat{R}_f^{\mathrm{GL}}$. This is expected since the grouping loss estimator can only capture a fraction of the total GL.

Gain of post-training Theory points to using $\hat{R}_{f,t^\star} = \hat{R}_{f,t^\star}^{\mathrm{CL}} + \hat{R}_f^{\mathrm{GL}}$ to measure the potential utility gain from using a good post-training method. Fig. 6a shows that the estimated $\hat{R}_{f,t^\star}$ captures well the gain of fine-tuning ($r^2 = 0.83$ and a relation $y \approx x$). Fig. 6b shows that the gain of stacking (with boosted trees or random forest) and GLAR reaches similar levels of correlation. $\hat{R}_{f,t^\star}^{\mathrm{CL}}$ reaches a lower correlation of $r^2 = 0.6$ which is expected since $R_{f,t}^{\mathrm{CL}}$ is blind to gains on top of recalibration. All other metrics, calibration errors, Brier score and AUC correlate poorly to the gain of post-training ($r^2 \leq 0.2$).



a. $\hat{R}_{f,t^\star}^{\mathrm{CL}} + \hat{R}_f^{\mathrm{GL}}$

b. Correlation r^2 to gain of post-training

Figure 6: **Gain of Post-Training.** (a) Gain in utility of fine-tuning versus $\hat{R}_{f,t^\star} = \hat{R}_{f,t^\star}^{\mathrm{CL}} + \hat{R}_f^{\mathrm{GL}}$, for each (model, dataset, $t^\star$). (b) Pearson's r^2 correlation of the gain in utility of each post-training method with $\hat{R}^{\mathrm{GL}}$, $\hat{R}_{f,t}^{\mathrm{CL}}$, and other performance metrics.

Improving recalibration hits diminishing returns In the binary setting, isotonic regression is one of the simplest recalibration methods and is parameter-free. Fig. 7a shows the utility gain of each of the remaining 8 post-training methods relatively to the gain of isotonic regression. Isotonic regression provides better utility gains that any of the other recalibration methods, even the more advanced Scaling-Binning and MetaCal (see also Fig. 19). Fig. 7b shows the computational time of the post-training methods. Isotonic regression is also the fastest recalibration method, by at least a factor 10. Fine-tuning and stacking provides better gain than recalibration, but their cost can become prohibitive: stacked boosted trees are 10^7 times more expensive than isotonic regression. GLAR provides a cheaper post-training method but with lower gains. These observations along with theoretical results on the regret, suggest that recalibration reaches a ceiling in its ability to improve decision-making. Enhancing recalibration methods often does not lead to better decision-making. Instead, the focus should be on methods to reduce the grouping regret, such as stacking, fine-tuning or

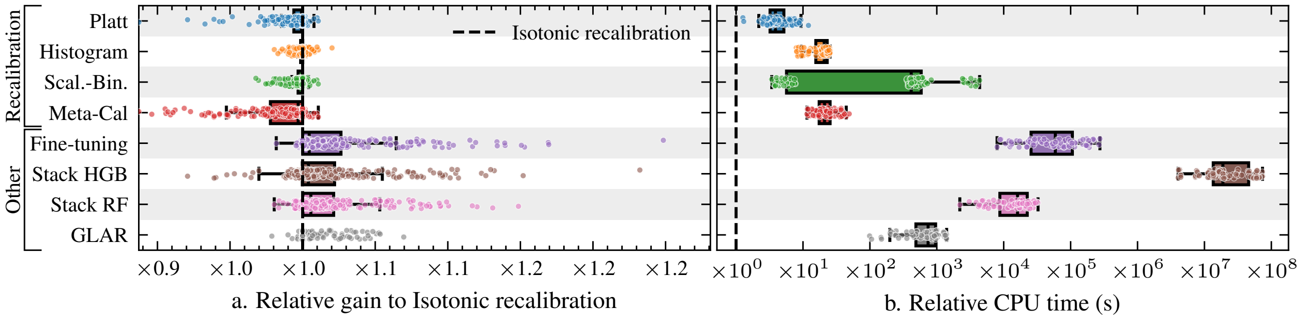


Figure 7: **Comparing various recalibration and advanced post-training.** Gain (a.) or CPU time (b.) of post-training methods relative to isotonic recalibration, for each (model, dataset, $t^\star$).

developping cheaper alternatives in the spirit of GLAR.

5 Conclusion

This work quantifies both theoretically and empirically how imperfect class-conditional probability estimates affect the expected utility of downstream decisions. We provided an analytical expression for the regret in expected utility caused by miscalibration, along with upper and lower bounds on the regret due to grouping loss. Our experiments show that these quantities better capture the potential gains from recalibration and post-training in expected utility compared to common metrics. In the future, it would be of interest to extend theses results to the multiclass setting.

Acknowledgments

We thank Thomas Moreau for valuable discussions that helped refine the derivations of the bounds.

References

Bhatia, R. and Davis, C. (2000). A better bound on the variance. *The American Mathematical Monthly*, 107:353–357.

Chung, Y.-A., Lin, H.-T., and Yang, S.-W. (2015). Cost-aware pre-training for multiclass cost-sensitive deep learning. *arXiv preprint arXiv:1511.09337*.

Collins, G. S. and Altman, D. G. (2012). Predicting the 10 year risk of cardiovascular disease in the united kingdom: independent and external validation of an updated version of qrisk2. *BMJ*, 344.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning from imbalanced data sets*, volume 10. Springer.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *34th International Conference on Machine Learning, ICML 2017*, 3:2130–2143.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110:457–506.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics, LIPIcs*, 67.

Kochenderfer, M. J. (2015). *Decision making under uncertainty: theory and application.* MIT press.

Kull, M. and Flach, P. (2015). Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. volume 9284.

Kumar, A., Liang, P., and Ma, T. (2019). Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Lahlou, S., Jain, M., Nekoei, H., Butoi, V. I., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. (2021). Deup: Direct epistemic uncertainty prediction.

Ling, C. X. and Sheng, V. S. (2008). Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning*, 2011:231–235.

Ling, C. X., Yang, Q., Wang, J., and Zhang, S. (2004). Decision trees with minimal costs. In *Proceedings of the twenty-first international conference on Machine learning*, page 69.

Localio, A. R. and Goodman, S. (2012). Beyond the usual prediction accuracy metrics: reporting results for clinical decision making. *Annals of internal medicine*, 157(4):294–295.

Ma, X. and Blaschko, M. B. (2021). Meta-cal: Well-controlled post-hoc calibration by ranking. In *International Conference on Machine Learning*, pages 7235–7245. PMLR.

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature 2023 616:7956*, 616:259–265.

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Noarov, G., Ramalingam, R., Roth, A., and Xie, S. (2023). High-dimensional prediction for sequential decision making.

Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. *Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018*, pages 255–258.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay (2011). Scikit-learn: Machine learning in python.

Perez-Lebel, A., Morvan, M. L., and Varoquaux, G. (2023). Beyond calibration: estimating the grouping loss of modern neural networks. In *The Eleventh International Conference on Learning Representations*.

Peterson, M. (2017). *An introduction to decision theory*. Cambridge University Press.

Petrides, G. and Verbeke, W. (2022). Cost-sensitive ensemble learning: a unifying framework. *Data Mining and Knowledge Discovery*, 36:1–28.

Pfohl, S. R., Xu, Y., Foryciarz, A., Ignatiadis, N., Genkins, J., and Shah, N. H. (2022). Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. *ACM International Conference Proceeding Series*, 22:1039–1052.

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, pages 61–74.

Rothblum, G. N. and Yona, G. (2022). Decision-making under miscalibration. *arXiv preprint arXiv:2203.09852*.

Sheng, V. S. and Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *Aaai*, volume 6, pages 476–481.

Sox, H. C., Higgins, M. C., and Owens, D. K. (2013). Medical decision making. *Medical Decision Making*, pages 93–142.

Ting, K. M. (1998). Inducing cost-sensitive trees via instance weighting. In *European symposium on principles of data mining and knowledge discovery*, pages 139–147. Springer.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., diagnostic tests, T. G. E., and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230.

Van Calster, B. and Vickers, A. J. (2015). Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making*, 35(2):162–169.

Vickers, A. J., Van Calster, B., and Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, 352.

von Neumann, J. and Morgenstern, O. (1944). Theory of games and economic behavior. Princeton University Press.

Zadrozny, B. and Elkan, C. (2001a). Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 204–213.

Zadrozny, B. and Elkan, C. (2001b). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. pages 609–616.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. ACM Press.

Zadrozny, B., Langford, J., and Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. In *Third IEEE International Conference on Data Mining*, pages 435–442.

Zhao, S., Kim, M. P., Sahoo, R., Ma, T., and Ermon, S. (2021). Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 27:22313–22324.

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., et al. (2023). A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes. For the theoretical part, all assumptions, definitions are specified and proofs given in appendix. For the experimental part, a detailed procedure is given in Sec. C.3 and a python code repository enables reproducing the results (`https://github.com/aperezlebel/decision_suboptimal_classifiers`).
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes. All theorems have their assumptions defined.
   (b) Complete proofs of all theoretical results. Yes. All proofs are given in appendix.
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes. The python code repository is given in the supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes. The training details are given in Sec. C.3.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes. Details given in Sec. C.3.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes. All datasets and pretrained models used are refered in Tab. 2 and Tab. 1.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Appendix

A Proposed evaluation procedure

A procedure to evaluate whether to post-train or not for given utilities Our findings suggest a new model evaluation procedure for batch decision-making under uncertainty. For a given trained model, it is important to determine whether post-training will benefit a decision task, what type of post-training is sufficient (recalibration or more advanced methods), and what the potential gain might be. In practice, with large or regulatory-constrained models (*e.g.* foundation models (Zhou et al., 2023), health models (Collins and Altman, 2012) or both (Moor et al., 2023)), it is prohibitive to apply a post-train-and-find-out strategy. Guiding post-training upfront is thus valuable. Concretely (Fig. 8), we propose to first measure the model's regret $R_{f,t}$ for the specified decision task using $\hat{R}_{f,t} = \hat{R}^{\mathrm{CL}}_{f,t} + \hat{R}^{\mathrm{GL}}_{f}$, and to assess whether the model is utility-suboptimal ($R_{f,t} > 0$). The regret decomposition then informs on whether the regret comes mainly from miscalibration ($R_{f,t} \approx R^{\mathrm{CL}}_{f,t}$). This enables selecting the appropriate post-training method and avoiding unnecessary, costly methods. The strong correlation between the estimated regret $\hat{R}_{f,t}$ and the effective gain of post-training enables a cost-benefit analysis: is the potential utility gain worth the cost of post-training the model? Note that this cost not only includes computational costs, but also the cost of changing the production model and the potential risks of deploying a new model. If the analysis favors post-training, we start-over the procedure to check whether the post-training was effective (post-training can sometimes undermine the original model). If the analysis rejects post-training, or if the model was not suboptimal, the decision-maker assesses whether the achieved utility is satisfying for their task. If not, gathering more informative features may then reduce aleatoric uncertainty and potentially improve the utility.

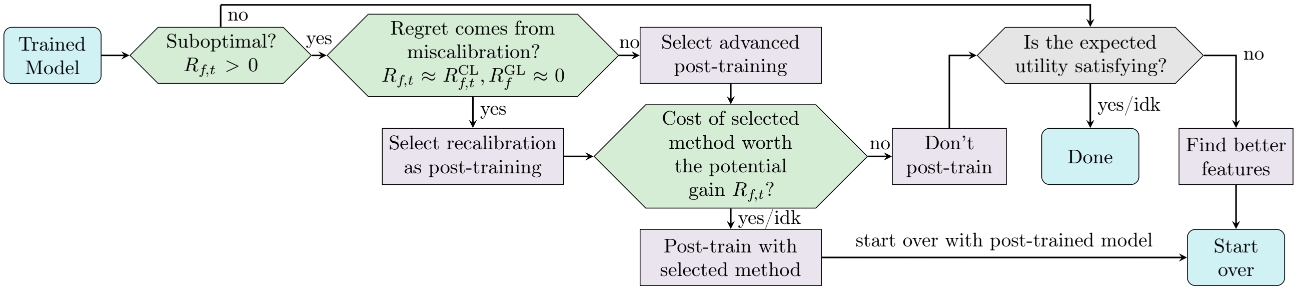

Trained Model
Suboptimal? $R_{f,t} > 0$
no
yes
Regret comes from miscalibration? $R_{f,t} \approx R^{\mathrm{CL}}_{f,t}, R^{\mathrm{GL}}_{f} \approx 0$
no
Select advanced post-training
yes
Select recalibration as post-training
Cost of selected method worth the potential gain $R_{f,t}$?
no
Don't post-train
yes/idk
Post-train with selected method
start over with post-trained model
Is the expected utility satisfying?
no
yes/idk
Done
Find better features
Start over

Figure 8: **Proposed model evaluation procedure.** From a trained model, our work enables three major steps: assessing whether a model is suboptimal for the decision task at hand, quantifying the expected regret, and finding its origin: miscalibration or not (green boxes). This enables disambiguating between when to improve the model (post-train) and when to improve the data (finding better features).

B Theoretical results

B.1 Parametrization of decision rules

Lemma B.1 (Parametrization of decision rules)**.** *The set of binary decision rules $\mathcal{X}^{\{0,1\}}$ can be parametrized with a function $p : \mathcal{X} \to [0,1]$ and a threshold $t \in [0,1]$:*

$$\mathcal{X}^{\{0,1\}} = \left\{\delta_{p,t} : x \mapsto \mathbb{1}_{p(x)\geq t} \text{ where } p : \mathcal{X} \to [0,1], t \in [0,1]\right\} \tag{15}$$

Proof of Lem. B.1. First let's show the inclusion of left in right. Let $\delta \in \mathcal{X}^{\{0,1\}}$. Let's show that there exists a function $p : \mathcal{X} \to [0,1]$ and a threshold $t \in [0,1]$ such that $\delta = \delta_{p,t}$.

Define $p : x \mapsto \delta(x)$ and $t = \frac{1}{2}$. Then for all $x \in \mathcal{X}$, $\delta_{p,t}(x) = \mathbb{1}_{p(x)\geq t} = \mathbb{1}_{\delta(x)\geq \frac{1}{2}} = \delta(x)$.

The inclusion of right in left is trivial because all $\delta_{p,t}$ are decision rules from $\mathcal{X}$ to $\{0,1\}$. □

B.2 $\mathcal{D}_f$ contains all decisions relying on f only

Lemma B.2 ($\mathcal{D}_f$)**.** $\mathcal{D}_f \triangleq \left\{\delta_{g(f),t} : \quad g : [0,1] \to [0,1],\ t \in [0,1]\right\}$ *contains all possible decision rules taking f as input:*

$$\mathcal{D}_f = \left\{x \mapsto d(f(x)) : d \in [0,1]^{\{0,1\}}\right\} \tag{16}$$

Proof of Lem. B.2. Let $\delta \in \mathcal{D}_f$. By definition of $\mathcal{D}_f$, there exists $g : [0,1] \to [0,1]$ and $t \in [0,1]$ such that $\delta(x) = \mathbb{1}_{g(f(x))\geq t}$ for all $x \in \mathcal{X}$. Define $d : p \mapsto \mathbb{1}_{g(p)\geq t}$. Then $d \in [0,1]^{\{0,1\}}$ and $\delta = x \mapsto d(f(x))$.

Let $\delta \in \left\{x \mapsto d(f(x)) : d \in [0,1]^{\{0,1\}}\right\}$. Thus, there exists $d \in [0,1]^{\{0,1\}}$ such that $\delta = x \mapsto d(f(x))$. Define $g : p \mapsto d(p)$ and $t = \frac{1}{2}$. Then $\delta(x) = g(f(x)) = \mathbb{1}_{g(f(x))\geq t}$. Hence $\delta \in \mathcal{D}_f$. □

B.3 Expression of expected utility

Lemma B.3 (Expected utility)**.** *Let $\delta \in \mathcal{X} \to \{0,1\}$ and $x \in \mathcal{X}$. Then the expected utility of δ conditional to x is:*

$$\mathrm{EU}(\delta, x) = U_{\triangle}\delta(x)(f^\star(x) - t^\star) + f^\star(x)U_{01} + (1 - f^\star(x))U_{00} \tag{17}$$

Corollary B.4 (Difference in expected utilities)**.** *Let $\delta_1, \delta_2 \in \mathcal{X} \to \{0,1\}$ and $x \in \mathcal{X}$. Then the difference in expected utilities of δ_1 and δ_2 conditional to x is:*

$$\mathrm{EU}(\delta_2, x) - \mathrm{EU}(\delta_1, x) = U_{\triangle}(\delta_2(x) - \delta_1(x))(f^\star(x) - t^\star) \tag{18}$$

Proof of Lem. B.3 and Cor. B.4. Let $\delta \in \mathcal{X} \to \{0,1\}$ and $x \in \mathcal{X}$.

First note that: $U[\delta(x), 1] = \begin{cases} U_{11} & \text{if } \delta(x) = 1 \\ U_{01} & \text{otherwise} \end{cases} = \delta(x)(U_{11} - U_{01}) + U_{01}$.

Similarly, $U[\delta(x), 0] = \begin{cases} U_{10} & \text{if } \delta(x) = 1 \\ U_{00} & \text{otherwise} \end{cases} = \delta(x)(U_{10} - U_{00}) + U_{00}$.

Then the expected utility of δ conditional to x writes:

$$\begin{aligned}
\mathrm{EU}(\delta, x) &= \mathbb{E}\left[U_{\delta(X),Y} \middle| X = x\right] && \text{Eq. 2} \\
&= f^\star(x)U_{\delta(x),1} + (1 - f^\star(x))U_{\delta(x),0} \\
&= f^\star(x)(\delta(x)(U_{11} - U_{01}) + U_{01}) \\
&\quad + (1 - f^\star(x))(\delta(x)(U_{10} - U_{00}) + U_{00}) \\
&= \delta(x)(f^\star(x)(U_{00} - U_{10} + U_{11} - U_{01}) - (U_{00} - U_{01})) \\
&\quad + f^\star(x)U_{01} + (1 - f^\star(x))U_{00} \\
&= U_{\triangle}\delta(x)(f^\star(x) - t^\star) + f^\star(x)U_{01} + (1 - f^\star(x))U_{00} && \text{Def of } U_{\triangle} \text{ and } t^\star \text{ (Eq. 4)}
\end{aligned}$$

Hence for all $\delta_1, \delta_2 \in \mathcal{X}^{\{0,1\}}$ and $x \in \mathcal{X}$:

$$\mathrm{EU}(\delta_2, x) - \mathrm{EU}(\delta_1, x) = U_\Delta(\delta_2(x) - \delta_1(x))(f^\star(x) - t^\star) \tag{19}$$

□

B.4 Best decision over $\mathcal{D}_f$

Proposition 3.1 (Best decision given estimated probabilities, B.4)**.** *Let $\mathcal{D}_f$ be the set of decision rules function of the estimated probabilities f. Then the calibrated probabilities thresholded at $t^\star$,*

$$\delta_{c\circ f,t^\star} : x \mapsto \mathbb{1}_{(c\circ f)(x)\geq t^\star}, \tag{8}$$

maximize the conditional expected utility over $\mathcal{D}_f$, i.e.,

$$\delta_{c\circ f,t^\star} \in \operatorname*{argmax}_{\delta\in\mathcal{D}_f} \mathrm{EU}(\delta|p) \qquad \textit{for all } p \in \operatorname{supp} f(X)$$

$$\begin{aligned} &\textit{with} \quad \mathrm{EU}(\delta|p) \triangleq \mathbb{E}[\mathrm{EU}(\delta, X)|f(X) = p] \\ &\textit{and} \quad \mathcal{D}_f = \{\delta_{g\circ f,t} : \quad g : [0,1] \to [0,1],\ t \in [0,1]\}\,. \end{aligned}$$

Proof of Prop. 3.1. Let $\delta \in \mathcal{D}_f$. Let's show that $\mathrm{EU}(\delta_{c\circ f,t^\star}|p) \geq \mathrm{EU}(\delta|p)$.

By definition of $\mathcal{D}_f$, there exist $g : [0,1] \to [0,1]$ and $t \in [0,1]$ such that $\delta = x \mapsto \mathbb{1}_{g(f(x))\geq t}$.

Let $p \in \operatorname{supp} f(X)$.

$$\begin{aligned}
&\mathrm{EU}(\delta_{c\circ f,t^\star}|p) - \mathrm{EU}(\delta|p) && \tag{20}\\
&= \mathbb{E}[\mathrm{EU}(\delta_{c\circ f,t^\star}, X) - \mathrm{EU}(\delta, X)|f(X) = p] && \text{Definition of EU}(\cdot|p) \tag{21}\\
&= U_\Delta \mathbb{E}[(\delta_{c\circ f,t^\star}(X) - \delta(X))(f^\star(X) - t^\star)|f(X) = p] && \text{Cor. B.4} \tag{22}\\
&= U_\Delta \mathbb{E}\big[(\mathbb{1}_{c(f(X))\geq t^\star} - \mathbb{1}_{g(f(X))\geq t})(f^\star(X) - t^\star)\big|f(X) = p\big] && \text{Def. of } \delta_{c\circ f,t^\star} \text{ and } \delta \tag{23}\\
&= U_\Delta(\mathbb{1}_{c(f(X))\geq t^\star} - \mathbb{1}_{g(f(X))\geq t})(\mathbb{E}[f^\star(X)|f(X)] - t^\star) && \delta_{c\circ f,t^\star} \text{ and } \delta \text{ func. of } f(X) \tag{24}\\
&= U_\Delta(\mathbb{1}_{c(f(X))\geq t^\star} - \mathbb{1}_{g(f(X))\geq t})(c(f(X)) - t^\star) && \text{Def. of } c \text{ (6)} \tag{25}\\
&= U_\Delta \mathbb{1}_{c(f(X))\geq t^\star} \mathbb{1}_{g(f(X))<t}(c(f(X)) - t^\star) && \tag{26}\\
&\quad + U_\Delta \mathbb{1}_{c(f(X))<t^\star} \mathbb{1}_{g(f(X))\geq t}(t^\star - c(f(X))) && \tag{27}\\
&= U_\Delta(\mathbb{1}_{c(f(X))\geq t^\star} \mathbb{1}_{g(f(X))<t} + \mathbb{1}_{c(f(X))<t^\star} \mathbb{1}_{g(f(X))\geq t}) && \tag{28}\\
&\quad \times |c(f(X)) - t^\star| && \tag{29}\\
&= U_\Delta |\mathbb{1}_{c(f(X))\geq t^\star} - \mathbb{1}_{g(f(X))\geq t}||c(f(X)) - t^\star| && \tag{30}\\
&\geq 0 && \tag{31}
\end{aligned}$$

Hence:

$$\delta_{c\circ f,t^\star} \in \operatorname*{argmax}_{\delta\in\mathcal{D}_f} \mathrm{EU}(\delta|p) \tag{32}$$

□

B.5 Calibration regret $R^{\mathrm{CL}}_{f,t}$

B.5.1 Expression

Proposition 3.2 (Expression of the calibration regret, B.5.1)**.** *For all $p \in \operatorname{supp} f(X)$,*

$$R^{\mathrm{CL}}_{f,t}(p) = \begin{cases} U_\Delta |c(p) - t^\star| & \textit{if } \mathbb{1}_{c(p)\geq t^\star} \neq \mathbb{1}_{p\geq t} \\ 0 & \textit{otherwise} \end{cases}. \tag{11}$$

Proof of Prop. 3.2. Let $\delta \in \mathcal{D}_f$. By definition of $\mathcal{D}_f$, there exist $g : [0,1] \to [0,1]$ and $t \in [0,1]$ such that $\delta = x \mapsto \mathbb{1}_{g(f(x)) \geq t}$. Let $p \in \operatorname{supp} f(X)$.

$$\begin{aligned}
R^{\mathrm{CL}}_{f,t}(p) &= \mathrm{EU}(\delta_{c\circ f,t^\star}|p) - \mathrm{EU}(\delta_{f,t}|p) && \text{Definition of } R^{\mathrm{CL}}_{f,t}(p) && (33)\\
&= \mathbb{E}\big[\mathrm{EU}(\delta_{c(f),t^\star}, X) - \mathrm{EU}(\delta_{f,t}, X) \big| f(X) = p\big] && \text{Definition of } \mathrm{EU}(\cdot|p) && (34)\\
&= U_{\triangle}(\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{g(p)\geq t})(c(p) - t^\star) && \text{Eq. 25} && (35)\\
&= U_{\triangle}|\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{g(p)\geq t}||c(p) - t^\star| && \text{Eq. 30} && (36)
\end{aligned}$$

□

B.5.2 Zero calibration regret for all utilities implies calibration

Proposition B.5. *For all $p \in \operatorname{supp} f(X)$,*

$$R^{\mathrm{CL}}_{f,t^\star}(p) = 0, \quad \forall t^\star \in [0,1] \iff c(p) = p. \tag{37}$$

Proof of Prop. B.5. Let $p \in \operatorname{supp} f(X)$. Suppose for all $t^\star \in [0,1], R^{\mathrm{CL}}_{f,t^\star}(p) = 0$. Let's show that $c(p) = p$. By contradiction, suppose $c(p) \neq p$. Let's show that there exists $t^\star \in [0,1]$ such that $R^{\mathrm{CL}}_{f,t^\star}(p) > 0$. Take $t^\star = \frac{1}{2}(c(p) + p)$. Then $t^\star \in [0,1]$ and $|\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{p\geq t^\star}| = 1$ and $|c(p) - t^\star| = \frac{1}{2}|c(p) - p| > 0$. Hence $R^{\mathrm{CL}}_{f,t^\star}(p) > 0$ which proves the contradiction. Now suppose $c(p) = p$. Then $|\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{p\geq t^\star}| = |\mathbb{1}_{p\geq t^\star} - \mathbb{1}_{p\geq t^\star}| = 0$ and $R^{\mathrm{CL}}_{f,t^\star}(p) = U_{\triangle}|\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{p\geq t^\star}||c(p) - t^\star| = 0$ (Prop. 3.2). □

B.5.3 Adjusting the threshold to address the calibration regret

Proposition 3.3 (Adjusting the threshold t_f, B.5.3)**.** *For all $t^\star \in [0,1]$ let $t_f \in c^{-1}(\{t^\star\})$ if it exists, otherwise let $t_f \triangleq \inf\{t : c(t) \geq t^\star\}$. If c is monotonic non-decreasing, then thresholding f at t_f,* i.e. *δ_{f,t_f}, achieves zero miscalibration regret: $R^{\mathrm{CL}}_{f,t_f} = 0$.*

Proof of Prop. 3.3. Let $t^\star \in [0,1]$. Suppose c is monotonic non-decreasing. Let $p \in \operatorname{supp} f(X)$. Then $p \geq t_f \Leftrightarrow c(p) \geq c(t_f)$ because c is non-decreasing. When $t_f \in c^{-1}(\{t^\star\})$, then $c(t_f) = t^\star$ and $p \geq t_f \Leftrightarrow c(p) \geq t^\star$ hence $\mathbb{1}_{c(p)\geq t^\star} = \mathbb{1}_{p\geq t_f}$. When $t_f = \inf\{t : c(t) \geq t^\star\}$, then $c(t_f) \geq t^\star$ by definition. Let's show that we still have $p \geq t_f \Leftrightarrow c(p) \geq t^\star$. Suppose $p \geq t_f$. Then $c(p) \geq c(t_f) \geq t^\star$. Suppose $c(p) \geq t^\star$. Then $p \in \{t : c(t) \geq t^\star\}$. But t_f is the smallest element of this set. So $p \geq t_f$. Hence $\mathbb{1}_{c(p)\geq t^\star} = \mathbb{1}_{p\geq t_f}$. Hence:

$$R^{\mathrm{CL}}_{f,t^\star}(p) = U_{\triangle}(\mathbb{1}_{c(p)\geq t^\star} - \mathbb{1}_{p\geq t_f})(c(p) - t^\star) = 0 \qquad \text{Eq. 35} \tag{38}$$

□

B.6 Lemma on the variance

Lemma B.6 (Bhatia-Davis inequality (Bhatia and Davis, 2000))**.** *Let $m, M \in \mathbb{R}$ such that $m \leq M$. Let Z a random variable valued in $[m, M]$. Then the variance of Z is upper bounded by:*

$$\mathbb{V}[Z] \leq (M - \mathbb{E}[Z])(\mathbb{E}[Z] - m)$$

Lemma B.7 (Bounds on the variance of a random variable in [0, 1])**.** *Let $t \in [0,1]$. Let Z be a random variable valued in $[0,1]$. Define $\mu^+ \triangleq \mathbb{E}[Z|Z \geq t]$, $\mu^- \triangleq \mathbb{E}[Z|Z < t]$ and $w \triangleq \mathbb{P}[Z \geq t]$.*

When both μ^+ and μ^- are defined, the expectation of Z satisfies:

$$\mathbb{E}[Z] = w\mu^+ + (1 - w)\mu^- \tag{39}$$

and the variance of Z is bounded by:

$$\begin{aligned}
\mathbb{V}[Z] &\geq w(1 - w)(\mu^+ - \mu^-)^2 && \textit{Variance lower bound} && (40)\\
\mathbb{V}[Z] &\leq w(\mu^+ - t) + \mathbb{E}[Z]\,(t - \mathbb{E}[Z]) && \textit{Variance upper bound} && (41)
\end{aligned}$$

When only one of μ^+ and μ^- is defined, the variance of Z is bounded by:

$$\mathbb{V}[Z] \geq 0 \qquad \textit{Variance lower bound} \tag{42}$$

$$\mathbb{V}[Z] \leq \begin{cases} \mathbb{E}[Z]\,(t - \mathbb{E}[Z]) & \textit{if } \mathbb{E}[Z] < t \\ (1-\mathbb{E}[Z])(\mathbb{E}[Z] - t) & \textit{if } \mathbb{E}[Z] \geq t \end{cases} \qquad \textit{Variance upper bound} \tag{43}$$

Proof of Lem. B.7. Let $t \in [0,1]$. Let Z be a random variable valued in $[0,1]$. Define $\mu^+ \triangleq \mathbb{E}[Z|Z \geq t]$, $\mu^- \triangleq \mathbb{E}[Z|Z < t]$ and $w \triangleq \mathbb{P}[Z \geq t]$.

The case where only one of μ^+ and μ^- is defined amounts to the Bhatia-Davis inequality (Lem. B.6) on Z which is valued in $[0,t]$ if μ^+ is undefined and in $[t,1]$ if μ^- is undefined. This proves Eq. 43.

From now we suppose that both μ^+ and μ^- are defined.

Expectation. The expectation of Z satisfies:

$$\begin{aligned} \mathbb{E}[Z] &= \mathbb{E}[\mathbb{E}[Z|\mathbb{1}_{Z\geq t}]] && \text{Law of total expectation} \\ &= \mathbb{P}[Z \geq t]\,\mathbb{E}[Z|Z \geq t] + \mathbb{P}[Z < t]\,\mathbb{E}[Z|Z < t] \\ &= w\mu^+ + (1-w)\mu^- && \text{Def. of } \mu^+, \mu^- \text{ and } w \end{aligned}$$

This proves Eq. 39. Next we show the bounds on the variance.

Variance. For convinience, we note: $\mu \triangleq \mathbb{E}[Z]$, $v^+ \triangleq \mathbb{V}[Z \geq t]$ and $v^- \triangleq \mathbb{V}[Z\,|\,Z < t]$. The variance of Z satisfies:

$$\begin{aligned} \mathbb{V}[Z] &= \mathbb{E}[\mathbb{V}[Z\,|\,\mathbb{1}_{Z\geq t}]] + \mathbb{V}[\mathbb{E}[Z|\mathbb{1}_{Z\geq t}]] && \text{Law of total variance} \\ &= wv^+ + (1-w)v^- + w(\mu^+ - \mu)^2 + (1-w)(\mu^- - \mu)^2 \end{aligned} \tag{44}$$

From this expression, we derive both the upper and lower bounds, using either the positivity of v^+ and v^- or the Bhatia-Davis inequality (Lem. B.6) on v^+ and v^-.

Upper bound. First, we show the upper bound. Using the fact that $Z|Z \geq t$ is in $[t,1]$ and $Z|Z < t$ is in $[0,t]$, the Bhatia-Davis inequality gives:

$$v^+ \leq (1-\mu^+)(\mu^+ - t) \tag{45}$$

$$v^- \leq (t - \mu^-)\mu^- \tag{46}$$

Hence, Eq. 44 gives:

$$\begin{aligned} \mathbb{V}[Z] \leq\; & w(1-\mu^+)(\mu^+ - t) + (1-w)(t-\mu^-)\mu^- \\ & + w(\mu^+ - \mu)^2 + (1-w)(\mu^- - \mu)^2 \\ =\; & w(\mu^+ - t + t\mu^+ - 2\mu^+\mu) + (1-w)\mu^-(t - 2\mu) + \mu^2 \\ =\; & w(\mu^+ - t + t\mu^+ - 2\mu^+\mu) + (\mu - w\mu^+)(t - 2\mu) + \mu^2 && \text{Using Eq. 39} \\ =\; & w(\mu^+ - t) + \mu(t - 2\mu) + \mu^2 \\ =\; & w(\mu^+ - t) + \mu(t - \mu) \end{aligned}$$

This proves the upper bound (Eq. 41)

Lower bound. Now, we prove the lower bound. First notice that Eq. 39 gives:

$$\mu^- - \mu = w(\mu^- - \mu^+) \tag{47}$$

$$\mu^+ - \mu = (1-w)(\mu^+ - \mu^-) \tag{48}$$

Using the positivity of v^+ and v^- in Eq. 44, we have:

$$\begin{aligned}
\mathbb{V}[Z] &\geq w(\mu^+ - \mu)^2 + (1-w)(\mu^- - \mu)^2 \\
&= w(1-w)^2(\mu^+ - \mu^-)^2 + (1-w)w^2(\mu^+ - \mu^-)^2 && \text{Using Eq. 47 and 48} \\
&= w(1-w)(\mu^+ - \mu^-)^2 && (49)
\end{aligned}$$

This proves the lower bound (Eq. 40).

Note that the results and proof holds for $Z|V$ for any random variable V, by conditioning all expectations and variance by V. □

B.7 R^{GL} regret reformulation

Lemma B.8 (Grouping loss induced regret)**.** *For all $p \in [0,1]$ where R_f^{GL} is defined, then:*

$$R_f^{\mathrm{GL}}(p) = U_\Delta\, \mathbb{E}\left[(\mathbb{1}_{f^\star(X)\geq t^\star} - \mathbb{1}_{c(p)\geq t^\star})(f^\star(X) - t^\star)\middle| h(X){=}p\right]. \tag{50}$$

Proof of Lem. B.8. Let $p \in [0,1]$ where R_f^{GL} is defined.

$$\begin{aligned}
R_f^{\mathrm{GL}}(p) &= \mathbb{E}\left[\mathrm{EU}(\delta_{f^\star, t^\star}, X) - \mathrm{EU}(\delta_{c_f(f), t^\star}, X)\middle| f(X){=}p\right] && (51) \\
&= U_\Delta \mathbb{E}\left[(\delta_{f^\star, t^\star}(X) - \delta_{c_f(f), t^\star}(X))(f^\star(X) - t^\star)\middle| f(X) = p\right] && \text{Cor. B.4} \quad (52) \\
&= U_\Delta \mathbb{E}\left[(\mathbb{1}_{f^\star(X)\geq t^\star} - \mathbb{1}_{c_f(p)\geq t^\star}(X))(f^\star(X) - t^\star)\middle| f(X) = p\right] && \text{Eq. 5} \quad (53)
\end{aligned}$$

□

Lemma B.9 (Regret reformulation)**.** *Let $p \in \operatorname{supp} f(X)$. Define $w(p) \triangleq \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p)$ and $c^+(p) \triangleq \mathbb{E}[f^\star(X) | f^\star(X) \geq t^\star, f(X) = p]$. Then, the grouping loss induced regret R_f^{GL} can be written as:*

$$R_f^{\mathrm{GL}}(p) = \begin{cases} U_\Delta \left(w(p)(c^+(p) - t^\star) - \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star)\right) & \text{if } 0 < w(p) < 1. \\ 0 & \text{otherwise.} \end{cases} \tag{54}$$

Proof of Lem. B.9. Let $p \in \operatorname{supp} f(X)$.

Suppose $0 < \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) < 1$. Then $c^+(p)$ is defined. We have:

$$\begin{aligned}
& R_f^{\mathrm{GL}}(p) && (55) \\
&= U_\Delta\, \mathbb{E}\left[(\mathbb{1}_{f^\star(X)\geq t^\star} - \mathbb{1}_{c(p)\geq t^\star})(f^\star(X) - t^\star)\middle| h(X){=}p\right] && \text{Lem. B.9} \quad (56) \\
&= U_\Delta\, \mathbb{E}\left[\mathbb{1}_{f^\star(X)\geq t^\star}(f^\star(X) - t^\star)\middle| h(X){=}p\right] && (57) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(\mathbb{E}[f^\star(X)|f(X) = p] - t^\star) && (58) \\
&= U_\Delta\, \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{f^\star(X)\geq t^\star}(f^\star(X) - t^\star)\middle| f^\star(X){\geq} t^\star, f(X)\right]\middle| h(X){=}p\right] && \text{Total exp.} \quad (59) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star) && \text{Def. of } c \text{ (Eq. 6)} \quad (60) \\
&= U_\Delta\, \mathbb{E}\left[\mathbb{1}_{f^\star(X)\geq t^\star}(\mathbb{E}[f^\star(X)|f^\star(X){\geq} t^\star, f(X)] - t^\star)\middle| h(X){=}p\right] && (61) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star) && (62) \\
&= U_\Delta\, \mathbb{E}\left[\mathbb{1}_{f^\star(X)\geq t^\star}(c^+(p) - t^\star)\middle| h(X){=}p\right] && \text{Def. of } c^+ \quad (63) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star) && (64) \\
&= U_\Delta\, \mathbb{E}\left[\mathbb{1}_{f^\star(X)\geq t^\star}\middle| h(X){=}p\right](c^+(p) - t^\star) && (65) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star) && (66) \\
&= U_\Delta\, w(p)(c^+(p) - t^\star) && \text{Def. of } w \quad (67) \\
&\quad - U_\Delta\, \mathbb{1}_{c(p)\geq t^\star}(c(p) - t^\star) && (68)
\end{aligned}$$

Suppose $\mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) = 1$. Then the above proof is still valid and $w(p) = 1$, $c^+(p) = c(p)$ and $c(p) \geq t^\star$. Hence $R_f^{\mathrm{GL}}(p) = 0$.

Suppose $\mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) = 0$. Then $\mathbb{E}\left[\mathbb{1}_{f^\star(X) \geq t^\star}(f^\star(X) - t^\star) \middle| h(X){=}p\right] = 0$ and $\mathbb{1}_{c(p) \geq t^\star} = 0$. Using Eq. 57 we have $R_f^{\mathrm{GL}}(p) = 0$.

□

B.8 Grouping regret lower bound

Theorem 3.4 (Grouping regret lower bound, B.8)**.** *The conditional grouping regret is lower bounded for all $p \in \operatorname{supp} f(X)$ as $R_f^{\mathrm{GL}}(p) \geq L_f^{\mathrm{GL}}(p)$, by:*

$$L_f^{\mathrm{GL}}(p) \triangleq U_\triangle \left[\mathrm{GL}(p) - V_{\min}(p)\right]_+ \tag{12}$$

$$\textit{with:} [\cdot]_+ = \max\{\cdot, 0\}$$

$$\textit{and: } V_{\min}(p) \triangleq \begin{cases} (1{-}c(p))\,(c(p){-}t^\star) & \textit{if } c(p) \geq t^\star \\ c(p)\,(t^\star{-}c(p)) & \textit{otherwise} \end{cases}.$$

Tightness. *The lower bound is tight. For any $p \in \operatorname{supp} f(X)$ for which f admits at least 3 antecedent values, and for all admissible mean $c(p) \in [0,1]$ and variance $\mathrm{GL}(p) \in [0, c(p)(1-c(p))]$, there exists a distribution of (X,Y) such that the conditional distribution $\mathbb{P}(f^\star(X)|f(X) = p)$ has mean $c(p)$ and variance $\mathrm{GL}(p)$, and the grouping regret attains its lower bound: $R_f^{\mathrm{GL}}(p) = L_f^{\mathrm{GL}}(p)$.*

Proof of Th. 3.4. Let $p \in \operatorname{supp} f(X)$.
Suppose $0 < \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) < 1$.

Define $w(p) \triangleq \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p)$ and $c^+(p) \triangleq \mathbb{E}[f^\star(X) | f^\star(X) \geq t^\star, f(X) = p]$. Lem. B.9 gives:

$$R_f^{\mathrm{GL}}(p) = U_\triangle \left(w(p)(c^+(p) - t^\star) - \mathbb{1}_{c(p) \geq t^\star}(c(p) - t^\star) \right) \tag{69}$$

We apply the upper bound of Lem. B.7 to $Z \triangleq f^\star(X)|f(X)$. Hence:

$$\begin{aligned} \mathbb{V}[f^\star(X) \,|\, f(X) = p] &\leq w(p)(c^+(p) - t^\star) + c(p)(t^\star - c(p)) && (70)\\ &= U_\triangle R_f^{\mathrm{GL}}(p) + \mathbb{1}_{c(p) \geq t^\star}(c(p) - t^\star) + c(p)(t^\star - c(p)) && (71)\\ &= U_\triangle R_f^{\mathrm{GL}}(p) + \mathbb{1}_{c(p) \geq t^\star}(c(p) - t^\star)(1 - c(p)) && (72)\\ &\quad + \mathbb{1}_{c(p) < t^\star} c(p)(t^\star - c(p)) && (73)\\ &= U_\triangle R_f^{\mathrm{GL}}(p) + V_{\min}(p) && (74) \end{aligned}$$

Hence:

$$R_f^{\mathrm{GL}}(p) \geq U_\triangle (\mathbb{V}[f^\star(X) \,|\, f(X){=}p] - V_{\min}(p)) \tag{75}$$

Since $R_f^{\mathrm{GL}}(p) \geq 0$, we have:

$$R_f^{\mathrm{GL}}(p) \geq U_\triangle \left[\mathbb{V}[f^\star(X) \,|\, f(X){=}p] - V_{\min}(p)\right]_+ \tag{76}$$

Suppose $\mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) \in \{0,1\}$. Then $R_f^{\mathrm{GL}}(p) = 0$ (Lem. B.9). Let's show that $L_f^{\mathrm{GL}}(p) = 0$. According to Lem. B.7, we have $\mathbb{V}[f^\star(X) \,|\, f(X){=}p] \leq V_{\min}(p)$ (Eq. 43). Then $L_f^{\mathrm{GL}}(p) = 0$. Then $R_f^{\mathrm{GL}}(p) = L_f^{\mathrm{GL}}(p)$.

Tightness of the regret lower bound L_f^{GL}. Following the proof of the regret lower bound, we see that the bound is obtained using the upper bound of the lemma on the variance (Lem. B.7). The proof of Lem. B.7 gives an idea on how to achieve the equality. Taking a distribution where v^+ and v^- saturates the Bhatia-Davis inequality is a good candidate. That is, a distribution of diracs in $\{0, t, 1\}$ with appropriate weights.

Below we explicit distributions achieving the equality between the regret R_f^{GL} and the lower bound L_f^{GL}. For convenience, we work with a random variable Z valued in $[0,1]$ which we will later link to the true probabilities $\mathbb{P}(f^\star(X)|f(X)=p)$.

Let $t \in (0,1)$. Let $c \in [0,1]$ and $v \in [0, c(1-c)]$. The Bhatia-Davis inequality shows that these represent the sets of admissible values of mean and variances of a random variable valued in $[0,1]$. Note $R_Z \triangleq U_{\Delta}\mathbb{E}\big[(\mathbb{1}_{Z\geq t} - \mathbb{1}_{\mathbb{E}[Z]\geq t})(Z-t)\big]$ and $L_Z \triangleq U_{\Delta}[\mathbb{V}[Z] - v_{\min}]_+$ with:

$$v_{\min} \triangleq \begin{cases} c(t-c) & \text{if } c < t \\ (1-c)(c-t) & \text{if } c \geq t \end{cases}. \tag{77}$$

R_Z represents the regret R_f^{GL} (Lem. B.8) and L_Z its lower bound L_f^{GL} (Th. 3.4) when $c = \mathbb{E}[Z]$.

We now show that there exist distributions $\mathbb{P}_Z$ of Z such that $\mathbb{E}[Z] = c$, $\mathbb{V}[Z] = v$ and $R_Z = L_Z$. Depending of the value of c and v, we explicit 4 distributions of Z that achieve the equality.

Case 1: $c < t$ **and** $v < v_{\min}$. Define Z_1 distributed as: $\mathbb{P}_{Z_1} = w_0\delta_0 + w_c\delta_c + w_t\delta_t$ with $w_c = 1 - \frac{v}{v_{\min}}$, $w_t = \frac{c}{t}\frac{v}{v_{\min}}$ and $w_0 = 1 - w_c - w_t$.

$$\begin{aligned}
\mathbb{E}[Z_1] &= w_c c + w_t t && (78)\\
&= c && (79)\\
\mathbb{V}[Z_1] &= w_0 c^2 + w_t(t-c)^2 && (80)\\
&= (1-w_1)c^2 - w_t t(t-2c) && (81)\\
&= \tfrac{v}{v_{\min}}c^2 - \tfrac{v}{v_{\min}}c(t-2c) && (82)\\
&= \tfrac{v}{v_{\min}}c(t-c) && (83)\\
&= v && (84)\\
R_{Z_1} &= U_{\Delta}\mathbb{E}[\mathbb{1}_{Z_1\geq t}(Z_1 - t)] && (85)\\
&= 0 && (86)\\
&= L_{Z_1} && (87)
\end{aligned}$$

Case 2: $c \geq t$ **and** $v < v_{\min}$. Define Z_2 distributed as: $\mathbb{P}_{Z_2} = w_t\delta_t + w_c\delta_c + w_1\delta_1$ with $w_c = 1 - \frac{v}{v_{\min}}$, $w_1 = \frac{c-t}{1-t}\frac{v}{v_{\min}}$ and $w_t = 1 - w_c - w_t$.

$$\begin{aligned}
\mathbb{E}[Z_2] &= w_t t + w_c c + w_1 && (88)\\
&= t + w_c(c-t) + w_1(1-t) && (89)\\
&= t + (1 - \tfrac{v}{v_{\min}})(c-t) + \tfrac{c-t}{1-t}\tfrac{v}{v_{\min}}(1-t) && (90)\\
&= c && (91)\\
\mathbb{V}[Z_2] &= w_t(t-c)^2 + w_1(1-c)^2 && (92)\\
&= (1-w_1)c^2 - w_t t(t-2c) && (93)\\
&= (t-c)^2 - w_c(t-c)^2 + w_1(1-t)(1+t-2c) && (94)\\
&= \tfrac{v}{v_{\min}}(c-t)(1-c) && (95)\\
&= v && (96)\\
R_{Z_2} &= U_{\Delta}\mathbb{E}[(\mathbb{1}_{Z_2\geq t} - 1)(Z_2 - t)] && (97)\\
&= U_{\Delta}\mathbb{E}[\mathbb{1}_{Z_2<t}(t - Z_2)] && (98)\\
&= 0 && (99)\\
&= L_{Z_2} && (100)
\end{aligned}$$

Case 3: $c < t$ **and** $v \geq v_{\min}$**.** Define Z_3 distributed as: $\mathbb{P}_{Z_3} = w_0\delta_0 + w_t\delta_t + w_1\delta_1$ with $w_1 = \frac{1}{1-t}(v - v_{\min})$, $w_t = \frac{c-w_1}{t}$ and $w_0 = 1 - w_t - w_1$.

$$\begin{aligned}
\mathbb{E}[Z_3] &= w_t t + w_1 \qquad &(101)\\
&= c \qquad &(102)\\
\mathbb{V}[Z_3] &= w_0 c^2 + w_t(t-c)^2 + w_1(1-c)^2 \qquad &(103)\\
&= c^2 + w_t t(t-2c) + w_1(1-2c) \qquad &(104)\\
&= c^2 + (c - w_1)(t-2c) + w_1(1-2c) \qquad &(105)\\
&= c(t-c) + w_1(1-t) \qquad &(106)\\
&= c(t-c) + v - v_{\min} \qquad &(107)\\
&= v \qquad &(108)\\
R_{Z_3} &= U_{\Delta}\mathbb{E}[\mathbb{1}_{Z_3 \geq t}(Z_3 - t)] \qquad &(109)\\
&= w_1(1-t) \qquad &(110)\\
&= v - v_{\min} \qquad &(111)\\
&= L_{Z_3} \qquad &(112)
\end{aligned}$$

Case 4: $c \geq t$ **and** $v \geq v_{\min}$**.** Define Z_4 distributed as: $\mathbb{P}_{Z_4} = w_0\delta_0 + w_t\delta_t + w_1\delta_1$ with $w_0 = \frac{1}{t}(v - v_{\min})$, $w_t = \frac{1-c-w_0}{1-t}$ and $w_1 = 1 - w_t - w_0$.

$$\begin{aligned}
\mathbb{E}[Z_4] &= w_t t + w_1 \qquad &(113)\\
&= 1 - w_0 + w_t(t-1) \qquad &(114)\\
&= 1 - w_0 - (1 - c - w_0) \qquad &(115)\\
&= c \qquad &(116)\\
\mathbb{V}[Z_4] &= w_0 c^2 + w_t(t-c)^2 + w_1(1-c)^2 \qquad &(117)\\
&= (1-c)^2 + w_0(2c-1) - w_t(t-1)(t+1-2c) \qquad &(118)\\
&= (1-c)^2 + w_0(2c-1) - (1-c-w_0)(t+1-2c) \qquad &(119)\\
&= v - v_{\min} - (1-c)(c-t) \qquad &(120)\\
&= v \qquad &(121)\\
R_{Z_4} &= U_{\Delta}\mathbb{E}[(\mathbb{1}_{Z_4 \geq t} - 1)(Z_4 - t)] \qquad &(122)\\
&= U_{\Delta}\mathbb{E}[\mathbb{1}_{Z_4 < t}(t - Z_4)] \qquad &(123)\\
&= w_0 t \qquad &(124)\\
&= v - v_{\min} \qquad &(125)\\
&= L_{Z_4} \qquad &(126)
\end{aligned}$$

Note that in the cases where $v < v_{\min}$, the Bhatia-Davis inequality cannot be saturated and an extra dirac δ_c is added in c.

For each of the above distributions, we have: $\mathbb{E}[Z_i] = c$, $\mathbb{V}[Z_i] = v$ and the regret equals the lower bound $R_{Z_i} = L_{Z_i}$.

Now we link Z back to the true probabilities $\mathbb{P}(f^{\star}(X)|f(X) = p)$. Let $p \in [0,1]$ such that f has at least 3 antecedant values that we note x_1, x_2 and x_3. We can thus create the joint distribution $\mathbb{P}_{(X,Y)}$ as a discrete distribution with $\mathbb{P}(x_i) = z_i$ and $\mathbb{P}(Y = 1|X = x_i) = w_i$ for all $i \in \{1,2,3\}$, where triplets (w_1, w_2, w_3) are taken from each above cases depending on the values of c and v, and (z_1, z_2, z_3) their associated positions. Then, the distribution $\mathbb{P}(f^{\star}(X)|f(X) = p)$ has mean c and variance v and $R_f^{\mathrm{GL}}(p) = L_f^{\mathrm{GL}}(p)$. □

B.9 Grouping regret upper bound

Theorem 3.5 (Grouping regret upper bound, B.9)**.** *The conditional grouping regret is upper bounded for all $p \in \operatorname{supp} f(X)$ as $R_f^{\mathrm{GL}}(p) \leq U_f^{\mathrm{GL}}(p)$, by:*

$$U_f^{\mathrm{GL}}(p) \triangleq \tfrac{1}{2} U_\Delta \Big(\sqrt{\mathrm{GL}(p) + (c(p) - t^\star)^2} - |c(p) - t^\star| \Big) . \tag{13}$$

Tightness. *The upper bound is tight when $t^\star = \frac{1}{2}$. For any $p \in \operatorname{supp} f(X)$, and for all admissible mean $c(p) \in [0,1]$ and variance $\mathrm{GL}(p) \in [0, c(p)(1-c(p))]$, there exists a distribution of (X, Y) such that the conditional distribution $\mathbb{P}(f^\star(X)|f(X) = p)$ has mean $c(p)$ and variance $\mathrm{GL}(p)$, and the grouping regret attains its upper bound: $R_f^{\mathrm{GL}}(p) = U_f^{\mathrm{GL}}(p)$.*

Proof of Th. 3.5. Let $p \in \operatorname{supp} f(X)$.

Suppose $0 < \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p) < 1$.

Define $w(p) \triangleq \mathbb{P}(f^\star(X) \geq t^\star | f(X) = p)$ and $c^+(p) \triangleq \mathbb{E}[f^\star(X) | f^\star(X) \geq t^\star, f(X) = p]$. Lem. B.9 gives:

$$R_f^{\mathrm{GL}}(p) = U_\Delta \left(w(p)(c^+(p) - t^\star) - \mathbb{1}_{c(p) \geq t^\star} (c(p) - t^\star) \right) \tag{127}$$

We apply the lower bound of Lem. B.7 to $Z \triangleq f^\star(X)|f(X)$, which gives:

$$\begin{aligned}
\sqrt{\mathbb{V}[f^\star(X)|f(X){=}p]} &\geq \sqrt{w(p)(1-w(p))(c^+(p) - c^-(p))^2} && \text{Eq. 40} && (128)\\
&= \sqrt{w(p)(1-w(p))}(c^+(p) - c^-(p)) && c^+(p) \geq c^-(p) && (129)\\
&= \sqrt{\frac{w(p)}{1-w(p)}}(c^+(p) - c(p)) && \text{Using Eq. 48} && (130)
\end{aligned}$$

Hence:

$$w(p)(c^+(p) - t^\star) \leq \sqrt{w(p)(1-w(p))\mathbb{V}[f^\star(X)|f(X){=}p]} + w(p)(c(p) - t^\star) \tag{131}$$

Using Eq. 127, we have:

$$R_f^{\mathrm{GL}}(p) \leq U_\Delta \left(\sqrt{w(p)(1-w(p))\mathbb{V}[f^\star(X)|f(X){=}p]} + (w(p) - \mathbb{1}_{c(p) \geq t^\star})(c(p) - t^\star) \right) \tag{132}$$

We showed Eq. 132 for all $w(p) \in (0,1)$. It still holds for $w(p) \in \{0,1\}$. Indeed, Lem. B.9 shows that $R_f^{\mathrm{GL}}(p) = 0$ in this case. Also, when $w(p) \in \{0,1\}$, then $w(p) - \mathbb{1}_{c(p) \geq t^\star} = 0$. Hence the right-hand side of Eq. 132 is also 0 for $w(p) \in \{0,1\}$, hence equal to $R_f^{\mathrm{GL}}(p)$.

Since Eq. 132 holds for all $w(p) \in [0,1]$, we now find the worst case. That is, the $w(p) \in [0,1]$ that maximizes the right-hand side of Eq. 132.

First suppose that $c(p) \geq t^\star$. Let $g_{a,b}(w) \triangleq \sqrt{w(1-w)b} + (w-1)a$ with $b \geq 0$. Note that $(132) = U_\Delta\, g_{a,b}(w(p))$ with $a = c(p) - t^\star$ and $b = \mathbb{V}[f^\star(X)|f(X){=}p]$. The maximum of $g_{a,b}$ is reached at $w = \frac{1}{2}(1 - \sqrt{\frac{a^2}{a^2+b}})$ with value $\frac{1}{2}(\sqrt{a^2+b} - a)$. Hence the maximum of (132) is:

$$\frac{U_\Delta}{2} (\sqrt{\mathbb{V}[f^\star(X)|f(X){=}p] + (c(p) - t^\star)^2} - (c(p) - t^\star)). \tag{133}$$

Now suppose that $c(p) < t^\star$. Similarly, let $g_{a,b}(w) \triangleq \sqrt{w(1-w)b} + wa$ with $b \geq 0$. Note that $(132) = U_\Delta\, g_{a,b}(w(p))$ with $a = c(p) - t^\star$ and $b = \mathbb{V}[f^\star(X)|f(X){=}p]$. The maximum of $g_{a,b}$ is reached at $w = \frac{1}{2}(1 - \sqrt{\frac{a^2}{a^2+b}})$ with value $\frac{1}{2}(\sqrt{a^2+b} + a)$. Hence the maximum of (132) is:

$$\frac{U_\Delta}{2} (\sqrt{\mathbb{V}[f^\star(X)|f(X){=}p] + (c(p) - t^\star)^2} - (t^\star - c(p))). \tag{134}$$

In both cases, the maximum of (132) is:

$$\frac{U_\triangle}{2}\big(\sqrt{\mathbb{V}[f^\star(X)\,|\,f(X)=p]+(c(p)-t^\star)^2}-|c(p)-t^\star|\big). \tag{135}$$

which concludes the proof.

Tightness of the regret upper bound U_f^{GL}. Following the proof of the regret upper bound, we see that the bound is obtained using the lower bound of the lemma on the variance (Lem. B.7). The proof of Lem. B.7 gives an idea on how to achieve the equality. Taking a distribution where $v^+=0$ and $v^-=0$ is a good candidate. That is, two diracs in $[0,1]$ with appropriate weights.

Below we explicit distributions achieving the equality between the regret R_f^{GL} and the upper bound R_f^{GL}. For convenience, we work with a random variable Z valued in $[0,1]$ which we will later link to the true probabilities $\mathbb{P}_{f^\star|h=p}$.

Suppose $t^\star\in(0,1)$. Let $c\in[0,1]$ and $v\in[0,c(1-c)]$. The Bhatia-Davis inequality shows that these represent the sets of admissible values of mean and variances of a random variable valued in $[0,1]$. Note $R_Z\triangleq U_\triangle\mathbb{E}\big[(\mathbb{1}_{Z\geq t^\star}-\mathbb{1}_{\mathbb{E}[Z]\geq t^\star})(Z-t^\star)\big]$ and $U_Z\triangleq\frac{1}{2}U_\triangle(\sqrt{\mathbb{V}[Z]+(\mathbb{E}[Z]-t^\star)^2}-|\mathbb{E}[Z]-t^\star|)$. R_Z represents the regret R^{GL} (Lem. B.8) and U_Z its upper bound U^{GL} (Th. 3.5).

We now show that there exist distributions $\mathbb{P}_Z$ of Z such that $\mathbb{E}[Z]=c$, $\mathbb{V}[Z]=v$ and $R_Z=U_Z$. Depending of the value of c, we explicit 2 distributions of Z that achieve the equality.

Define $w=\frac{1}{2}(1-\sqrt{\frac{(c-t^\star)^2}{(c-t^\star)^2+v}})$ and $\mathbb{P}_Z=(1-w)\delta_a+w\delta_b$ with $a,b\in[0,1]$. Suppose $t^\star=\frac{1}{2}$.

Case 1: $c<t^\star$. Take $a=c-\sqrt{v\frac{w}{1-w}}$ and $b=c+\sqrt{v\frac{1-w}{w}}$. Both a and b are in $[0,1]$ when $t^\star=\frac{1}{2}$.

$$\begin{aligned}
\mathbb{E}[Z_1]&=(1-w)a+wb && (136)\\
&=c-\sqrt{vw(1-w)}+\sqrt{vw(1-w)} && (137)\\
&=c && (138)\\
\mathbb{V}[Z_1]&=(1-w)(a-c)^2+w(b-c)^2 && (139)\\
&=vw+v(1-w) && (140)\\
&=v && (141)\\
R_{Z_1}&=U_\triangle\mathbb{E}[\mathbb{1}_{Z_1\geq t^\star}(Z_1-t^\star)] && (142)\\
&=w(c-t^\star+\sqrt{v\tfrac{1-w}{w}}) \qquad b\geq t^\star \text{ and } a\leq c\leq t^\star && (143)\\
&=\sqrt{vw(1-w)}+w(c-t^\star) && (144)\\
&=\frac{1}{2}(\sqrt{v+(c-t^\star)^2}-(c-t^\star))+(c-t^\star) \qquad \text{Eq. 133} && (145)\\
&=\frac{1}{2}(\sqrt{v+(c-t^\star)^2}-(t^\star-c)) && (146)\\
&=U_{Z_1} && (147)
\end{aligned}$$

Case 2: $c\geq t^\star$. Take $a=c-\sqrt{v\frac{1-w}{w}}$ and $b=c+\sqrt{v\frac{w}{1-w}}$. Both a and b are in $[0,1]$ when $t^\star=\frac{1}{2}$.

$$
\begin{aligned}
\mathbb{E}[Z_2] &= (1-w)a + wb && (148)\\
&= c + \sqrt{vw(1-w)} - \sqrt{vw(1-w)} && (149)\\
&= c && (150)\\
\mathbb{V}[Z_2] &= (1-w)(a-c)^2 + w(b-c)^2 && (151)\\
&= vw + v(1-w) && (152)\\
&= v && (153)\\
& && (154)
\end{aligned}
$$

$$
\begin{aligned}
R_{Z_2} &= U_{\Delta}\mathbb{E}[\mathbb{1}_{Z_2<t^\star}(t^\star - Z_2)] && && (155)\\
&= w(t^\star - c + \sqrt{v\tfrac{1-w}{w}}) && b \geq c \geq t^\star \text{ and } a \leq c \leq t^\star && (156)\\
&= \sqrt{vw(1-w)} + w(t^\star - c) && && (157)\\
&= \frac{1}{2}(\sqrt{v + (c-t^\star)^2} - (t^\star - c)) + (t^\star - c) && \text{Eq. 133} && (158)\\
&= \frac{1}{2}(\sqrt{v + (c-t^\star)^2} - (t^\star - c)) && && (159)\\
&= U_{Z_2} && && (160)
\end{aligned}
$$

For each of the above distributions, we have: $\mathbb{E}[Z_i] = c$, $\mathbb{V}[Z_i] = v$ and the regret equals the lower bound $R_{Z_i} = U_{Z_i}$.

Now we link Z back to the true probabilities $\mathbb{P}(f^\star(X)|f(X) = p)$. Let $p \in [0,1]$ such that f has at least 2 antecedant values that we note x_1 and x_2. We can thus create the joint distribution $\mathbb{P}_{(X,Y)}$ as a discrete distribution with $\mathbb{P}(x_1) = a$, $\mathbb{P}(x_2) = b$, $\mathbb{P}(Y=1|X=x_1) = 1-w$ and $\mathbb{P}(Y=1|X=x_2) = w$ Then, the distribution $\mathbb{P}(f^\star(X)|f(X)=p)$ has mean c and variance v and $R_f^{\mathrm{GL}}(p) = U_f^{\mathrm{GL}}(p)$.

If $p \in [0,1]$ is such that f has only one antecedant value, then there is no grouping loss in the level set $f = p$, *i.e.* $\mathrm{GL}(p) = 0$. Hence, $R_f^{\mathrm{GL}}(p) = 0 = U_f^{\mathrm{GL}}(p)$.

□

B.10 GLAR – Grouping Loss Adaptative Recalibration

Proposition B.10 (GLAR reduces grouping loss, B.10)**.** *Let a function $f : \mathcal{X} \to [0,1]$ and a partition $\mathcal{P} : \mathcal{X} \to E$ of the feature space $\mathcal{X}$. Then the GLAR-estimator $f_{\mathcal{P}}$ defined in Eq. 14 satisfies:*

$$
\begin{aligned}
\mathbb{E}[Y|f_{\mathcal{P}}(X) = p] &= p && f_{\mathcal{P}} \textit{ is calibrated} && (161)\\
\mathrm{GL}(f_{\mathcal{P}}) &\leq \mathrm{GL}(f) && f_{\mathcal{P}} \textit{ has lower } \mathrm{GL} \textit{ than } f && (162)
\end{aligned}
$$

The grouping loss is reduced by: $\mathrm{GL}(f) - \mathrm{GL}(f_{\mathcal{P}}) = \mathbb{E}[\mathbb{V}[f_{\mathcal{P}}(X) \mid f(X)]]$.
The remaining grouping loss is: $\mathrm{GL}(f_{\mathcal{P}}) = \mathbb{E}[\mathbb{V}[f^\star(X) \mid f(X), \mathcal{P}(X)]]$.

Proof. Let $\mathcal{P}$ a partition of the feature space. Define $P = \mathcal{P}(X)$, $H = f(X)$, $H_P = f_{\mathcal{P}}(X)$ and $\mathrm{GL}(H) = \mathbb{E}[\mathbb{V}[F \mid H]]$. Using definition of H_P, we have: $H_P = \mathbb{E}[Y|H,P] = \mathbb{E}[\mathbb{E}[F|X]|H,P] = \mathbb{E}[F|H,P]$

Following Perez-Lebel et al. (2023, Theorem 4.1), we have $\mathrm{GL}(H) = \mathrm{GL}_{\text{explained}}(H) + \mathrm{GL}_{\text{residual}}(H)$ with $\mathrm{GL}_{\text{explained}}(H) = \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H,P] \mid H]]$ and $\mathrm{GL}_{\text{residual}}(H) = \mathbb{E}[\mathbb{V}[F \mid H, P]]$.

$$
\begin{aligned}
&\mathrm{GL}_{\text{explained}}(H) - \mathrm{GL}_{\text{explained}}(H_P) && && (163)\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H,P] \mid H]] - \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H,P] \mid H_P]] && && (164)\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H,P] \mid H]] - \mathbb{E}[\mathbb{V}[H_P \mid H_P]] && H_P = \mathbb{E}[F|H,P] && (165)\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H,P] \mid H]] && \mathbb{V}[H_P \mid H_P] = 0 && (166)\\
&= \mathbb{E}[\mathbb{V}[H_P \mid H]] && && (167)
\end{aligned}
$$

We apply the law of total variance on the residual term by conditioning on H:

$$\begin{aligned}
\mathrm{GL}_{\mathrm{residual}}(H_P) &= \mathbb{E}[\mathbb{V}[F \mid H_P, P]] \tag{168}\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H, H_P, P] \mid H_P, P] + \mathbb{E}[\mathbb{V}[F \mid H, H_P, P]|H_P, P]] \tag{169}\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H, P] \mid H_P, P] + \mathbb{E}[\mathbb{V}[F \mid H, P]|H_P, P]] \tag{170}\\
&= \mathbb{E}[\mathbb{V}[H_P \mid H_P, P] + \mathbb{E}[\mathbb{V}[F \mid H, P]|H_P, P]] \tag{171}\\
&= \mathbb{E}[\mathbb{E}[\mathbb{V}[F \mid H, P]|H_P, P]] \tag{172}\\
&= \mathbb{E}[\mathbb{V}[F \mid H, P]] \tag{173}\\
&= \mathrm{GL}_{\mathrm{residual}}(H) \tag{174}
\end{aligned}$$

Hence:

$$\begin{aligned}
&\mathrm{GL}(H) - \mathrm{GL}(H_P) \tag{175}\\
&= \mathrm{GL}_{\mathrm{explained}}(H) - \mathrm{GL}_{\mathrm{explained}}(H_P) + \mathrm{GL}_{\mathrm{residual}}(H) - \mathrm{GL}_{\mathrm{residual}}(H_P) \tag{176}\\
&= \mathbb{E}[\mathbb{V}[\mathbb{E}[F|H, P] \mid H]] \tag{177}
\end{aligned}$$

□

Lemma B.11 (Hierarchy of partition-based decision rules)**.** *Let $\mathcal{P}_1, \mathcal{P}_2 : \mathcal{X} \to E$ be two partitions of the feature space such that $\mathcal{P}_1(X)$ is $\mathcal{P}_2(X)$-measurable. Let $g_{\mathcal{P}} : x \mapsto \mathbb{E}[Y|\mathcal{P}(X) = \mathcal{P}(x)]$. Then:*

$$\mathrm{EU}(\delta_{g_{\mathcal{P}_1},t}) \leq \mathrm{EU}(\delta_{g_{\mathcal{P}_2},t^\star}) \qquad \forall t \in [0,1] \tag{178}$$

Lem. B.11 shows that the finer the partition the better the expected utility. Since all of the original, recalibrated and GLAR-corrected can be seen as instance of GLAR with different granularity of partition (the trivial partition, the partition on f, and the partition on $(f, \mathcal{P})$ respectively), this enables to conclude on a hierarchy between these decision rules (Prop. B.12).

Proof. Let $\mathcal{P}_1, \mathcal{P}_2 : \mathcal{X} \to \mathbb{R}^d$ be two partitions of the feature space such that $\mathcal{P}_1(X)$ is $\mathcal{P}_2(X)$-measurable. Let $g_{\mathcal{P}} : x \mapsto \mathbb{E}[Y|\mathcal{P}(X) = \mathcal{P}(x)]$. Let $G_1 \triangleq g_{\mathcal{P}_1}(X)$, $G_2 \triangleq g_{\mathcal{P}_2}(X)$, $P_1 \triangleq \mathcal{P}_1(X)$ and $P_2 \triangleq \mathcal{P}_2(X)$ the associated random variables.

$$\begin{aligned}
&\mathrm{EU}(\delta_{g_{\mathcal{P}_2},t^\star}) - \mathrm{EU}(\delta_{g_{\mathcal{P}_1},t^\star}) && \\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})(\tfrac{F}{t^\star} - 1)\right] && \text{Cor. B.4}\\
&= \mathbb{E}_X\left[\mathbb{E}\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})(\tfrac{F}{t^\star} - 1)\middle|\mathcal{P}_2\right]\right] && \text{Total expectation}\\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})\mathbb{E}\left[\tfrac{F}{t^\star} - 1\middle|\mathcal{P}_2\right]\right] && P_1 \text{ is } P_2\text{-measurable}\\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})\left(\tfrac{\mathbb{E}[F|\mathcal{P}_2]}{t^\star} - 1\right)\right] && \text{Linearity}\\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})\left(\tfrac{\mathbb{E}[\mathbb{E}[Y|X]|\mathcal{P}_2]}{t^\star} - 1\right)\right] && \text{Definition of } F\\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})\left(\tfrac{\mathbb{E}[Y|\mathcal{P}_2]}{t^\star} - 1\right)\right] && \text{Total expectation}\\
&= \mathbb{E}_X\left[(\mathbb{1}_{G_2 \geq t^\star} - \mathbb{1}_{G_1 \geq t^\star})\left(\tfrac{G_2}{t^\star} - 1\right)\right] && \text{Definition of } G_2\\
&= \mathbb{E}_X\left[\mathbb{1}_{G_2 \geq t^\star}\mathbb{1}_{G_1 < t^\star}\left(\tfrac{G_2}{t^\star} - 1\right) + \mathbb{1}_{G_2 < t^\star}\mathbb{1}_{G_1 \geq t^\star}\left(1 - \tfrac{G_2}{t^\star}\right)\right] && \\
&\geq 0 &&
\end{aligned}$$

□

Proposition B.12 (Hierarchy of decision rules, B.10)**.** *Let $\mathcal{P}_1, \mathcal{P}_2 : \mathcal{X} \to \mathbb{R}$ be two partitions of the feature space such that $\mathcal{P}_1$ is constant on regions of $\mathcal{P}_2$ (*i.e.* $\mathcal{P}_1$ is a function of $\mathcal{P}_2$). Then, for all $\delta_{f,t} \in \mathcal{D}_f$:*

$$\mathrm{EU}(\delta_{f,t}) \leq \mathrm{EU}(\delta_{c\circ f,t^\star}) \leq \mathrm{EU}(\delta_{f_{\mathcal{P}_1},t^\star}) \leq \mathrm{EU}(\delta_{f_{\mathcal{P}_2},t^\star}) \leq \mathrm{EU}(\delta_{f^\star,t^\star})$$

Proof. Let $\mathcal{P}_1, \mathcal{P}_2 : \mathcal{X} \to \mathbb{R}^d$ be two partitions of the feature space such that $\mathcal{P}_1(X)$ is $\mathcal{P}_2(X)$-measurable.

Define $P_0 = 0$, $P_c = f(X)$, $P_1 = (P_c, \mathcal{P}_1(X))$ and $P_2 = (P_c, \mathcal{P}_2(X))$. We have: P_0 is P_c-measurable, P_c is P_1-measurable, and P_1 is P_2-measurable. Hence, Lem. B.11 concludes the proof.

□

A threshold view A duality between probability correction and adaptative thresholding also exists for GLAR. Instead of correcting the estimated probabilities on groups with $f_{\mathcal{P}}$, GLAR can also be viewed as using the original estimates f with an adaptative threshold $t_{\mathcal{P}}$ instead (Def. B.13): $\delta_{f,t_{\mathcal{P}}}$ and $\delta_{f_{\mathcal{P}},t^\star}$ have same regret.

Instead of correcting the estimated probabilities on groups with $f_{\mathcal{P}}$, this procedure can be viewed as using the original estimates f with an adaptative threshold $t_{\mathcal{P}}$ instead (Def. B.13):

Definition B.13 (Grouping loss adaptative threshold GLAT)**.**

$$t_{\mathcal{P}} \colon x \mapsto t^\star - (f_{\mathcal{P}}(x) - f(x)) \tag{179}$$

$\delta_{f,t_{\mathcal{P}}}$ and $\delta_{f_{\mathcal{P}},t^\star}$ have same regret: $f(x) \geq t_{\mathcal{P}} \Leftrightarrow f_{\mathcal{P}}(x) \geq t^\star$.

C Experiments

C.1 Model definitions

Table 1: Detailed description of all the models used for the hate speech detection experiment.

Model	HuggingFace	Latent Layer
CNERG Hatexplain	Hate-speech-CNERG/bert-base-uncased-hate...	classifier
CNERG en MuRIL	Hate-speech-CNERG/english-abusive-MuRIL	classifier
CNERG en mono	Hate-speech-CNERG/dehatebert-mono-englis...	classifier
CNERG portuguese	Hate-speech-CNERG/dehatebert-mono-portug...	classifier
CNERG tamil	Hate-speech-CNERG/tamil-codemixed-abusiv...	classifier
FB Roberta	facebook/roberta-hate-speech-dynabench-r...	classifier.out_proj

C.2 Dataset definitions

Table 2: Detailed description of all the datasets used for the hate speech detection experiment.

Dataset	HuggingFace or CSV	Split	Input	Target	Pos. class
Tweets	tweets_hate_speech_detect...	train	tweet	label	
Speech18	hate_speech18	train	text	label	
Offensive	hate_speech_offensive	train	tweet	class	0
Davidson	krishan-CSE/Davidson_Hate...	train+test	text	labels	0
Gender	ctoraman/gender-hate-spee...	train+test	Text	Label	2
FRENK	classla/FRENK-hate-en	train+val+test	text	label	
	limjiayi/hateful_memes_ex...	train+val+test	text	label	
Check	Paul/hatecheck	test	test_case	label_gold	hateful
Tweets 2	thefrankhsu/hate_speech_t...	train+test	tweet	label	
Open	parnoux/hate_speech_open_...	test	tweet	class	0
UCB	ucberkeley-dlab/measuring...	train	text	hate_speec...	y > 0.5
Merged	tweets_hate_speech_detect...	train	tweet	label	
	hate_speech18	train	text	label	
	hate_speech_offensive	train	tweet	class	0
	krishan-CSE/Davidson_Hate...	train+test	text	labels	0
	ctoraman/gender-hate-spee...	train+test	Text	Label	2
Merged 2	classla/FRENK-hate-en	train+val+test	text	label	
	limjiayi/hateful_memes_ex...	train+val+test	text	label	
	Paul/hatecheck	test	test_case	label_gold	hateful
	thefrankhsu/hate_speech_t...	train+test	tweet	label	
	parnoux/hate_speech_open_...	test	tweet	class	0
DynGen	CSV: bvidgen/Dynamic...		text	label	hate
Merged3	tweets_hate_speech_detect...	train	tweet	label	
	hate_speech18	train	text	label	
	hate_speech_offensive	train	tweet	class	0
	krishan-CSE/Davidson_Hate...	train+test	text	labels	0
	ctoraman/gender-hate-spee...	train+test	Text	Label	2
	classla/FRENK-hate-en	train+val+test	text	label	
	limjiayi/hateful_memes_ex...	train+val+test	text	label	
	Paul/hatecheck	test	test_case	label_gold	hateful

Continued on next page

Table 2: Detailed description of all the datasets used for the hate speech detection experiment.

		Split	Input	Target	Pos. class
Dataset	HuggingFace or CSV				
	thefrankhsu/hate_speech_t...	train+test	tweet	label	
	parnoux/hate_speech_open_...	test	tweet	class	0
	CSV: bvidgen/Dynamic...		text	label	hate

C.3 Experimental Details

Data For each of the pre-trained models for hate-speech detection Tab. 1, we use the embedding space of the pernultimate layers as input space $\mathcal{X}$ for estimating the grouping loss. We note $f : \mathcal{X} \to [0, 1]$ the function mapping the embedding space of the model to the hate-speech probability esitmate. We forward the datasets listed in Tab. 2 through each of the models. We store the ground truth label Y, which equals to 1 if the text is hate speech and 0 otherwise, the embedding X, and the model's predicted probability $f(X)$ that the text is hate-speech. Then we work with each triplet $(X, Y, f(X))$ to investigate the effect of post-training on decision-making. The data is split into a train set on which post-training methods are fit, and a test set on which the post-training methods are applied and the expected utility and regrets are estimated.

Utility We define binary utility matrices $U \in \mathbb{R}^{2\times2}$ of the shape:

$$U \triangleq \begin{bmatrix} 1 & 0 \\ 0 & U_{11} \end{bmatrix} \tag{180}$$

with $U_{11} \in \mathbb{R}^+$. We select values of U_{11} so that the optimal threshold $t^\star$ takes values $[0.01, 0.025, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, 0.975, 0.99]$. Note that the link between U_{11} and $t^\star$ is given by $U_{11} = \frac{1}{t^\star} - 1$ (Eq. 4). Each utilty matrix U represents a different decision-making problem. Following the decision theory result of Elkan (2001), the optimal decision rule is $\delta_{f^\star, t^\star}$. We thus use $\delta_{f,t^\star}$ as decision rule from the probability esimates given by f.

The expected utility of decision rule $\delta_{f,t^\star}$ is given by $\mathrm{EU}(\delta_{f,t^\star}) = \mathbb{E}[U[\delta_{f,t^\star}(X), Y]]$. We estimte the expected utility using the emprirical utility $\widehat{\mathrm{EU}} \triangleq \frac{1}{n} \sum_i U[\delta_{f,t^\star}(X_i), Y_i]$ where (X_i, Y_i) are the samples of (X, Y) and n is the number of samples.

Regret We estimate the regrets using the expression given in Def. 3.6. This necessitates estimates of the calibration curve c and the grouping loss. To estimate c we using an histogram binning with 15 equal-mass bins. The esitmation of the grouping loss is detailed in the next paragarph.

Grouping Loss estimation For the estimation of the grouping loss we follow the estimation procedure given by Perez-Lebel et al. (2023). We use 15 equal-mass bins on the probability space $[0, 1]$. For partitioning the embedding space $\mathcal{X}$, we use a decision tree on the pair (X, Y) constrained to create at most 5 regions per bin. The partition $\mathcal{P}$ is derived from the fitted decision tree using the leaves of the tree: each leave forms a part of the partition. To fit the decision tree we use a train-test split strategy, as recommended by Perez-Lebel et al. (2023), akin to a honest tree. The decision tree is fitted on the first part of the train set and the local averages and grouping loss are estimated on the second part of the train set. This avoids overfitting.

Recalibration We use recalibration methods to correct the estimated probabilities. Below are listed the methods used and implementation details. Recalibration methods are fitted on the train set and applied on the test set.

- *Isotonic Regression.* We use scikit-learn's implementation `IsotonicRegression` from `sklearn.isotonic` and fit it on (S, Y). This method has no hyperparameters.

- *Platt Scaling.* We use scikit-learn's implementation `_SigmoidCalibration` from `sklearn.calibration` and fit it on (S, Y). This method has no hyperparameters.

- *Histogram Binning*. We use the implementation of `HistogramCalibrator` from the `calibration` python package. We use 15 equal-mass bins.
- *Scaling-Binning*. We use the implementation of `PlattBinnerCalibrator` from the `calibration` python package. We use 15 equal-mass bins.
- *Meta-Cal*. We use the implementation of `MetaCalMisCoverage` from the `metacal` python package. We use a miscoverage of 0.05 based on the default value provided in the package's example `metacal/examples/test_metacal.py`.

Post-training Besides recalibration, we also investigate post-training methods. Below are listed the methods used and implementation details. Post-training methods are fitted on the train set and applied on the test set.

- *Fine-tuning*. Since the link between each hate-speech model's embedding space and the output space is a sigmoid, we use scikit-learn to learn the sigmoid function with `LogisticRegression` from `sklearn.linear_model` with default parameters. In this form of finetuning, the initial model weight is not used. This contrasts with finetuning using pytorch's optimizers starting from the pre-trained model's weights.
- *Stacking*. We use `RandomForestClassifier` and `HistGradientBoostingClassifier` implementation from scikit-learn with default parameters for the stacked models. The stack model is fit on the augmented space $(X, f(X))$ as input, that is $((X, f(X)), Y)$.
- *GLAR*. We implemented GLAR as described in the next paragraph.

Metrics The performance metrics used in the experiments are:

- *Brier score* defined as Brier $\triangleq \mathbb{E}\left[(f(X) - Y)^2\right]$.
- *Expected Calibration Error (ECE)* defined as ECE $\triangleq \mathbb{E}[|\mathbb{E}[Y|f(X)] - f(X)|]$.
- *Calibration Loss (CL)* defined as CL $\triangleq \mathbb{E}\left[(\mathbb{E}[Y|f(X)] - f(X))^2\right]$.
- *Root Mean Square Calibration Error (RMSCE)* defined as RMSCE $\triangleq \sqrt{\mathbb{E}[(\mathbb{E}[Y|f(X)] - f(X))^2]}$.
- *Maximum Calibration Error (MCE)* defined as MCE $\triangleq \mathbb{E}[\max|\mathbb{E}[Y|f(X)] - f(X)|]$.

Compute setting All experiments ran on a single compute node of 256 CPUs.

Code repository The code used for the experiments is available at `https://github.com/aperezlebel/decision_suboptimal_classifiers`.

GLAR GLAR builds on the grouping loss estimator proposed by Perez-Lebel et al. (2023). GLAR uses the partitioning strategy based on a decision tree detailed in Sec. C.3. Once identified the partition $\mathcal{P}$ of the input space, GLAR estimates the local averages of the probabilities $f(X)$ on each region of the partition. The GLAR estimator $f_{\mathcal{P}}$ is then the function mapping each input x to the average of the probabilities on the region of the partition to which x belongs. It is applied on the test set to correct the estimated probabilities similarly to the other post-training methods.

For each model f, we split the probability range $[0, 1]$ into 15 equal-mass bins. In each bin, we use a decision tree constrained to form at most 5 leaves. To regularize the correction by GLAR, we use a threshold strategy. We set a threshold $r \in \mathbb{R}$. GLAR applies the correction if the overall grouping regret $\hat{R}_f^{\mathrm{GL}}$ is larger than r *and* only corrects the bins for which the conditional grouping regret $\hat{R}_f^{\mathrm{GL}}(p)$ is larger than r. We set $r = 0.02$. Otherwise, GLAR applies an isotonic correction. This leverages the grouping regret estimator and avoid correcting the probabilities beyond calibration when it is not needed.

C.4 Post-training gain results

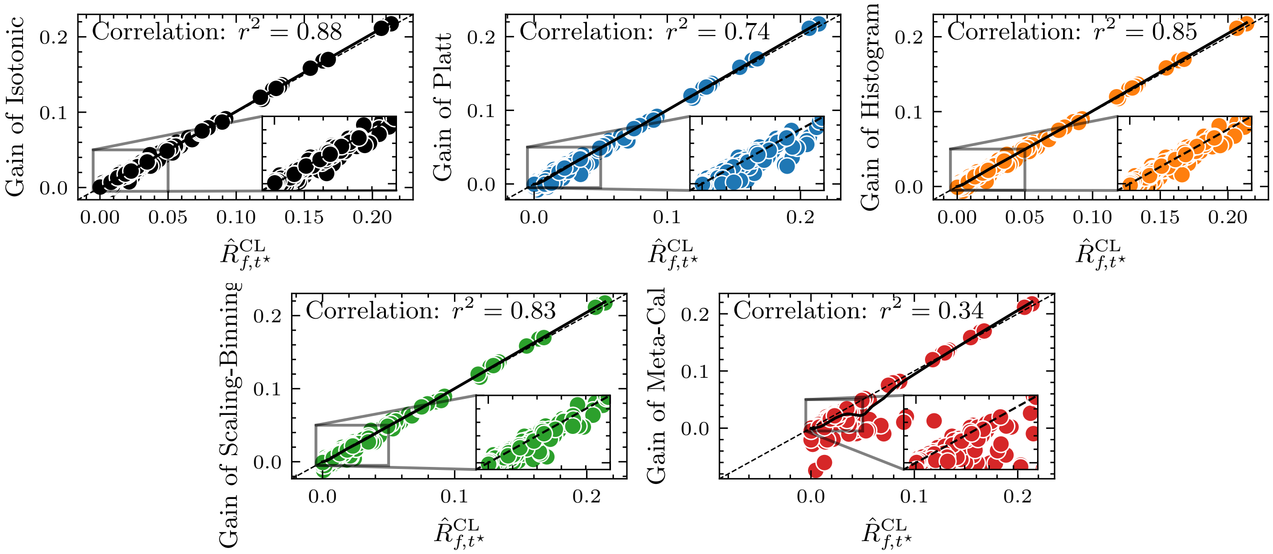


Figure 9: **Recalibration gain vs $\hat{R}^{\mathrm{CL}}_{f,t^\star}$.** Gain of each recalibration method vs $\hat{R}^{\mathrm{CL}}_{f,t^\star}$.

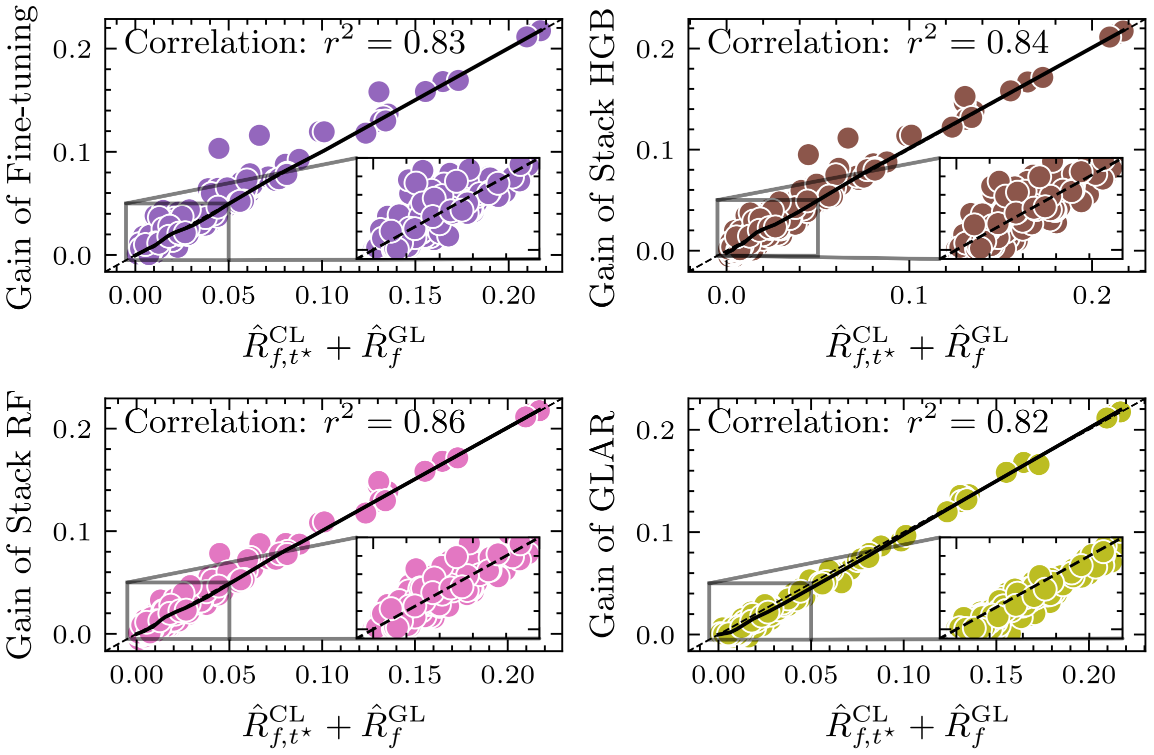


Figure 10: **Post-training gain vs $\hat{R}^{\mathrm{CL}}_{f,t^\star} + \hat{R}^{\mathrm{GL}}_{f}$.** Gain of non-recalibration methods vs $\hat{R}^{\mathrm{CL}}_{f,t^\star} + \hat{R}^{\mathrm{GL}}_{f}$.

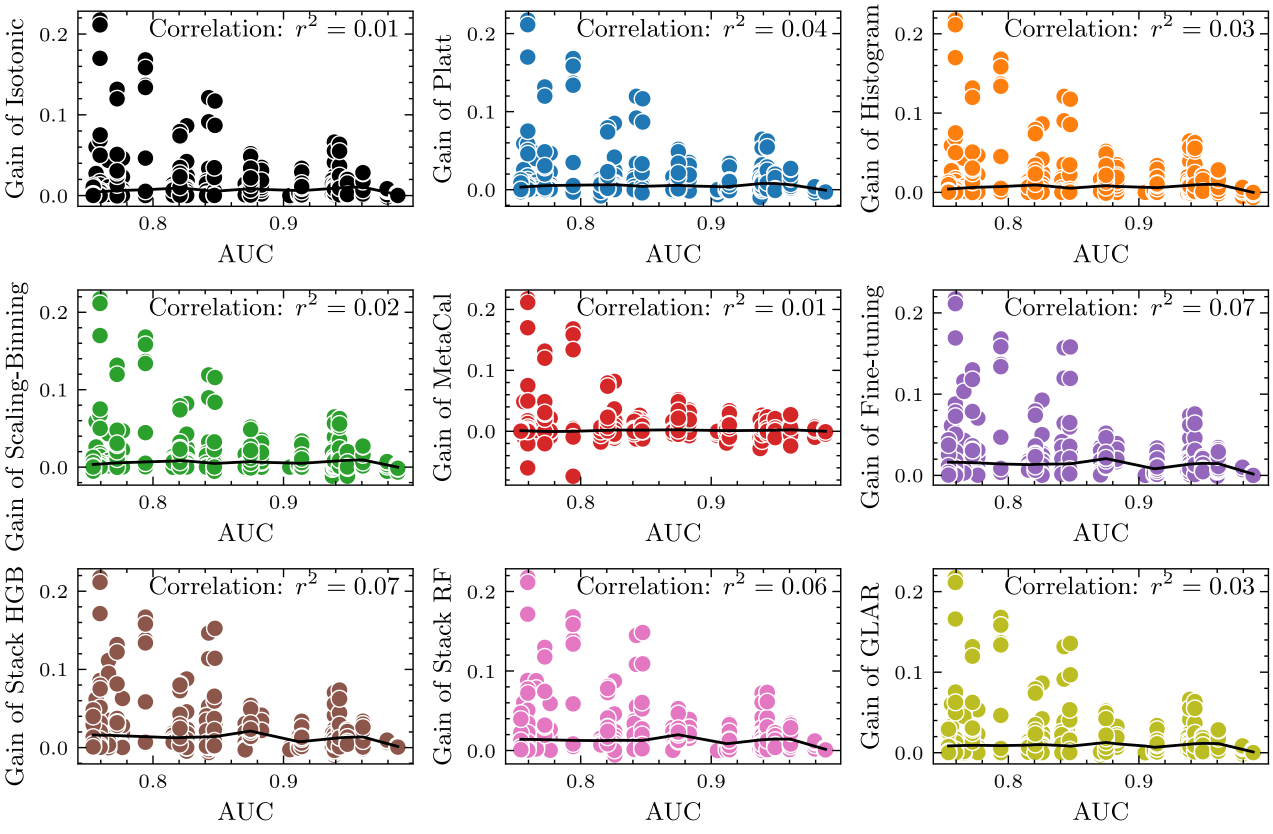


Figure 11: **Gain vs AUC.** Gain of each post-training method vs the AUC of the model.

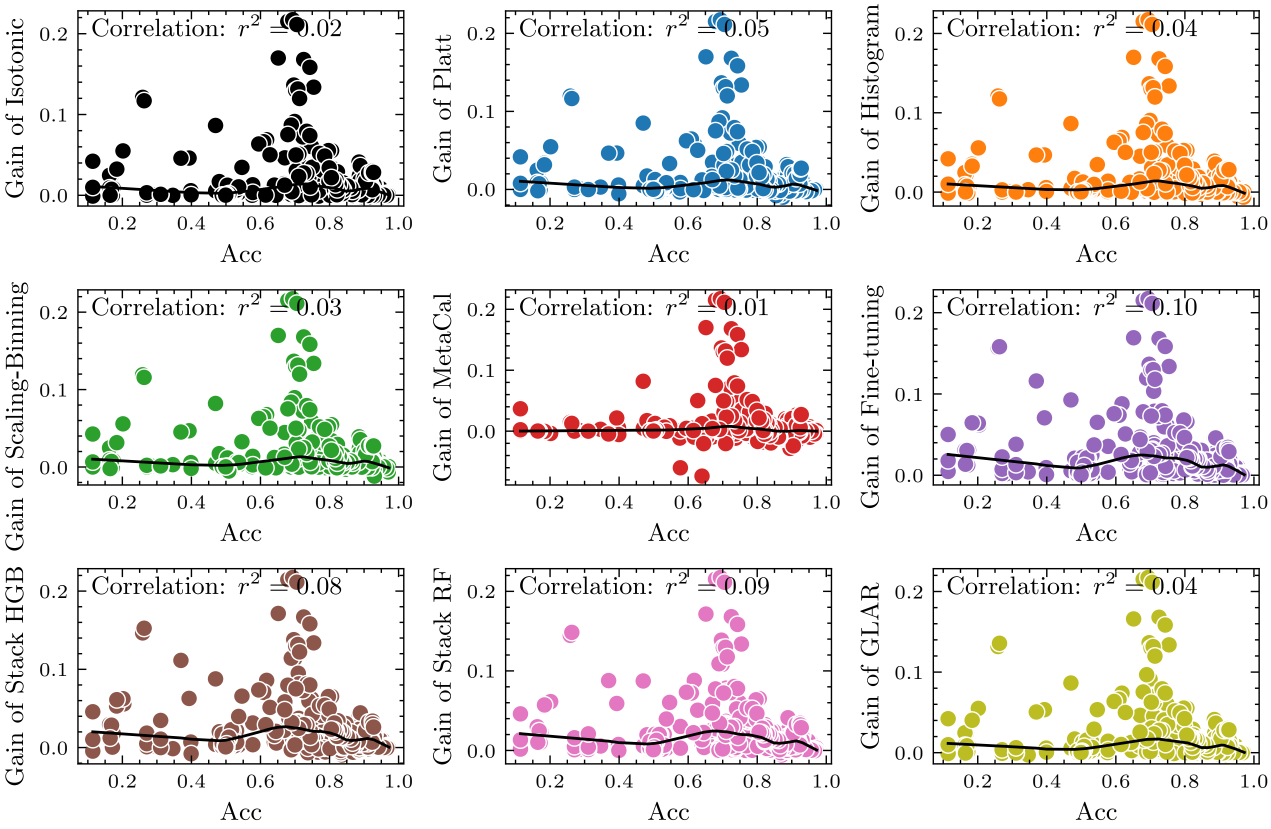


Figure 12: **Gain vs Accuracy.** Gain of each post-training method vs the Accuracy of the model.

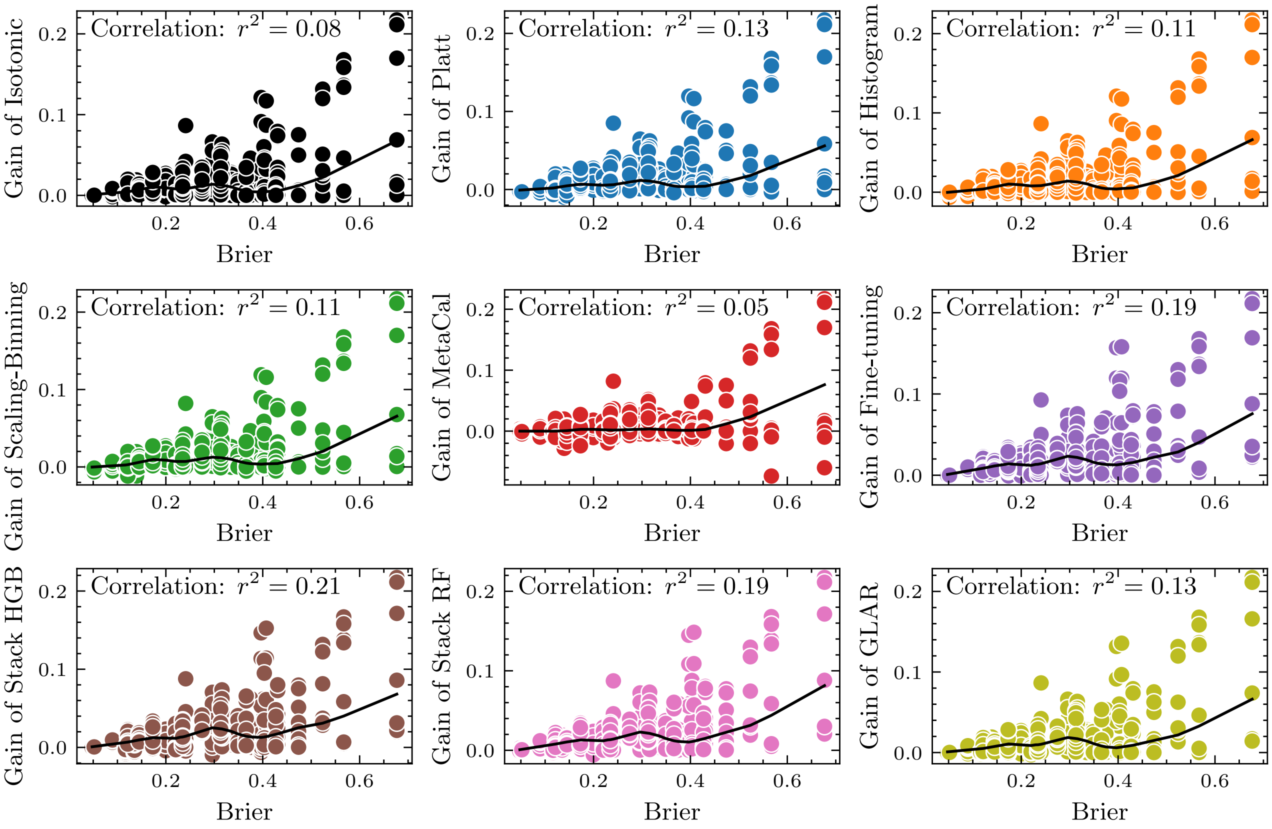


Figure 13: **Gain vs Brier.** Gain of each post-training method vs the Brier score of the model.

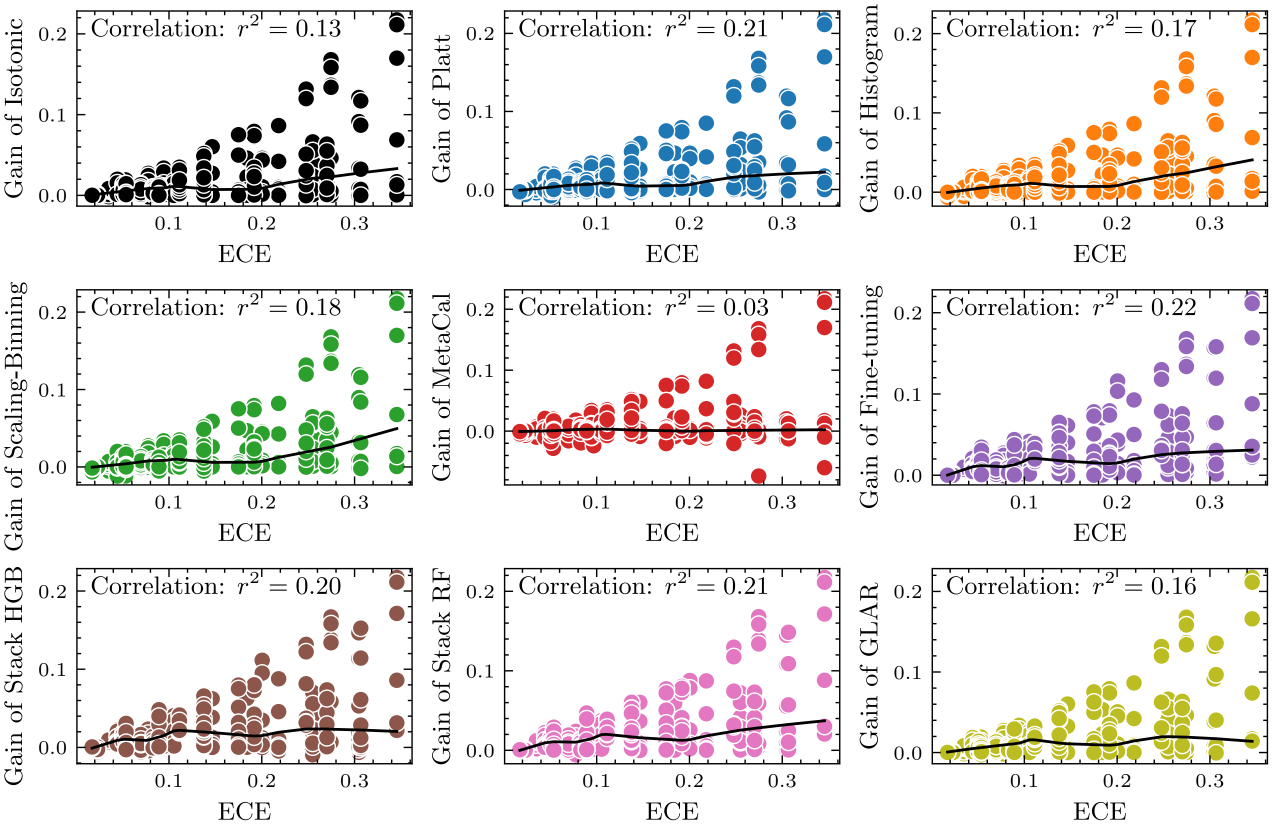


Figure 14: **Gain vs** ECE**.** Gain of each post-training method vs the ECE of the model.

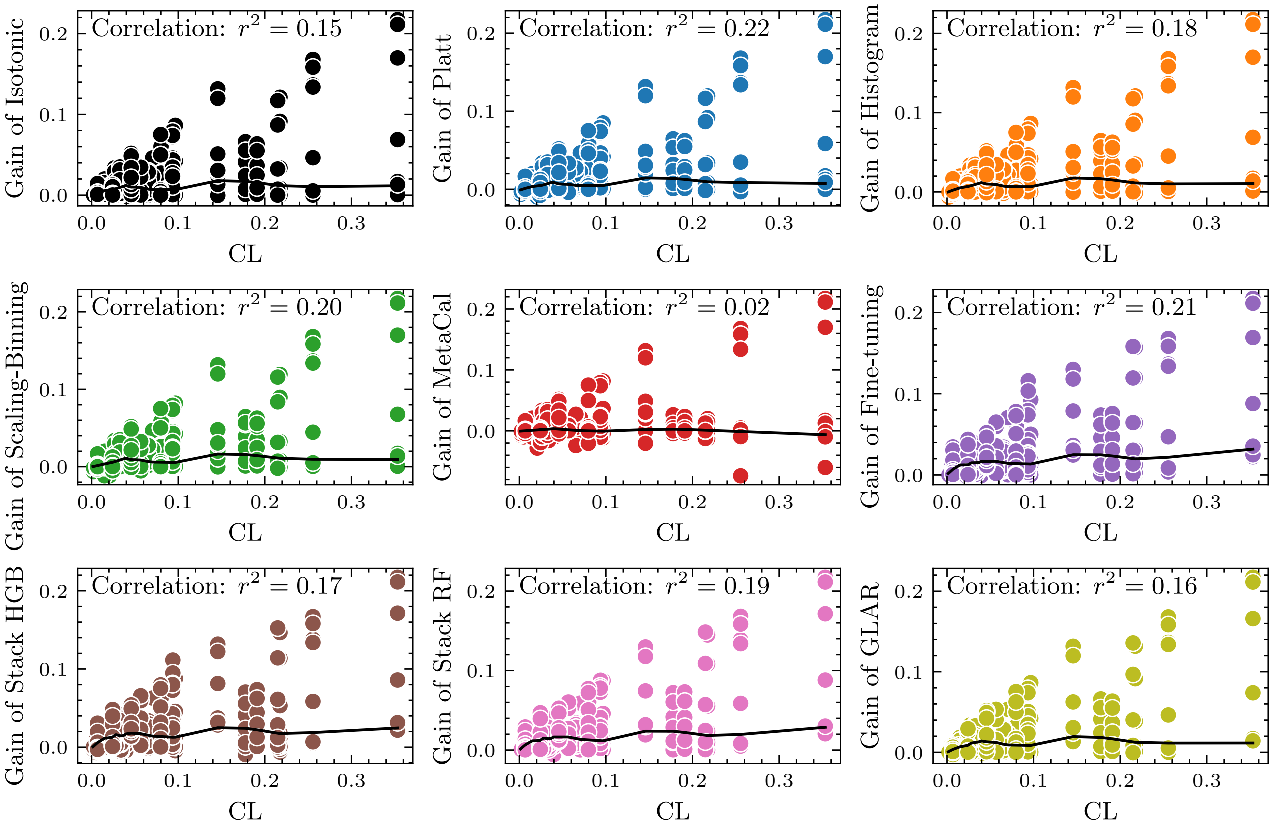


Figure 15: **Gain vs** CL**.** Gain of each post-training method vs the calibration loss of the model.

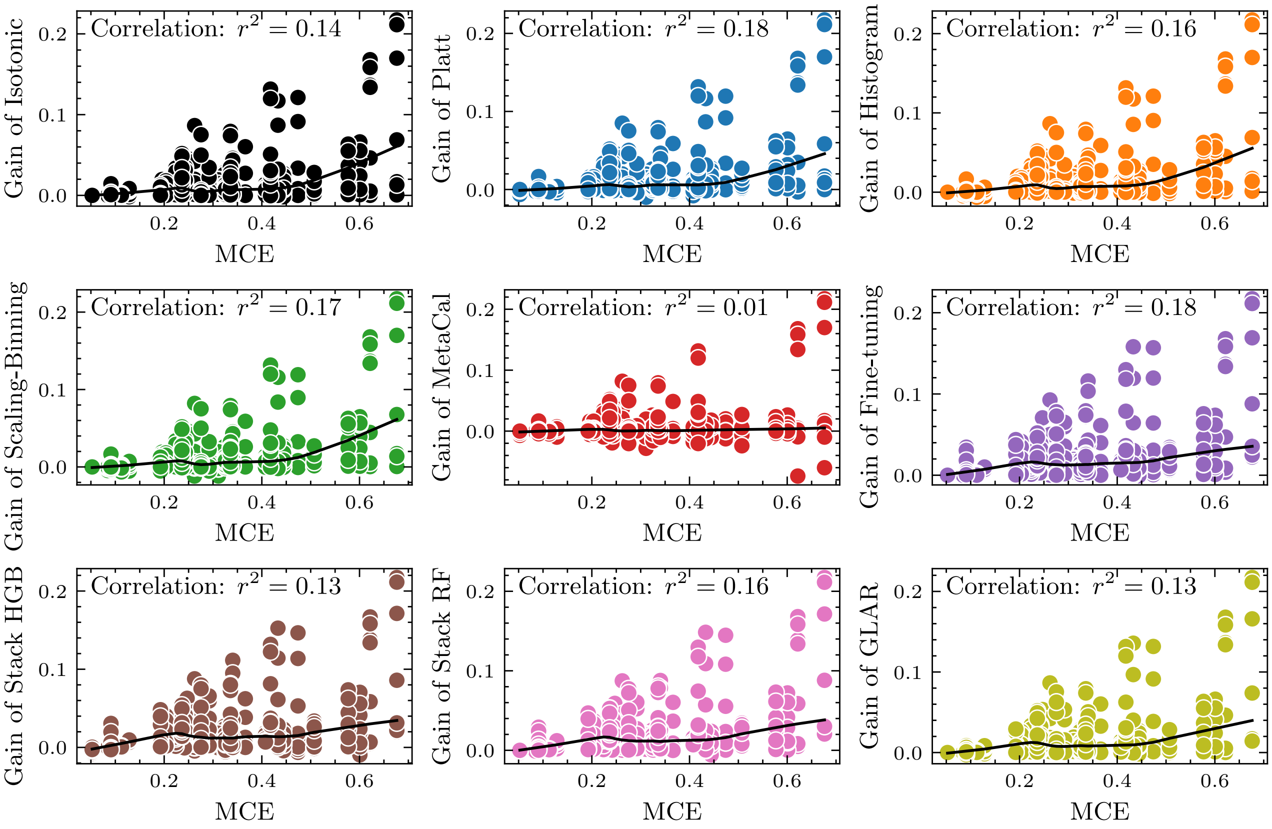


Figure 16: **Gain vs** MCE. Gain of each post-training method vs the MCE of the model.

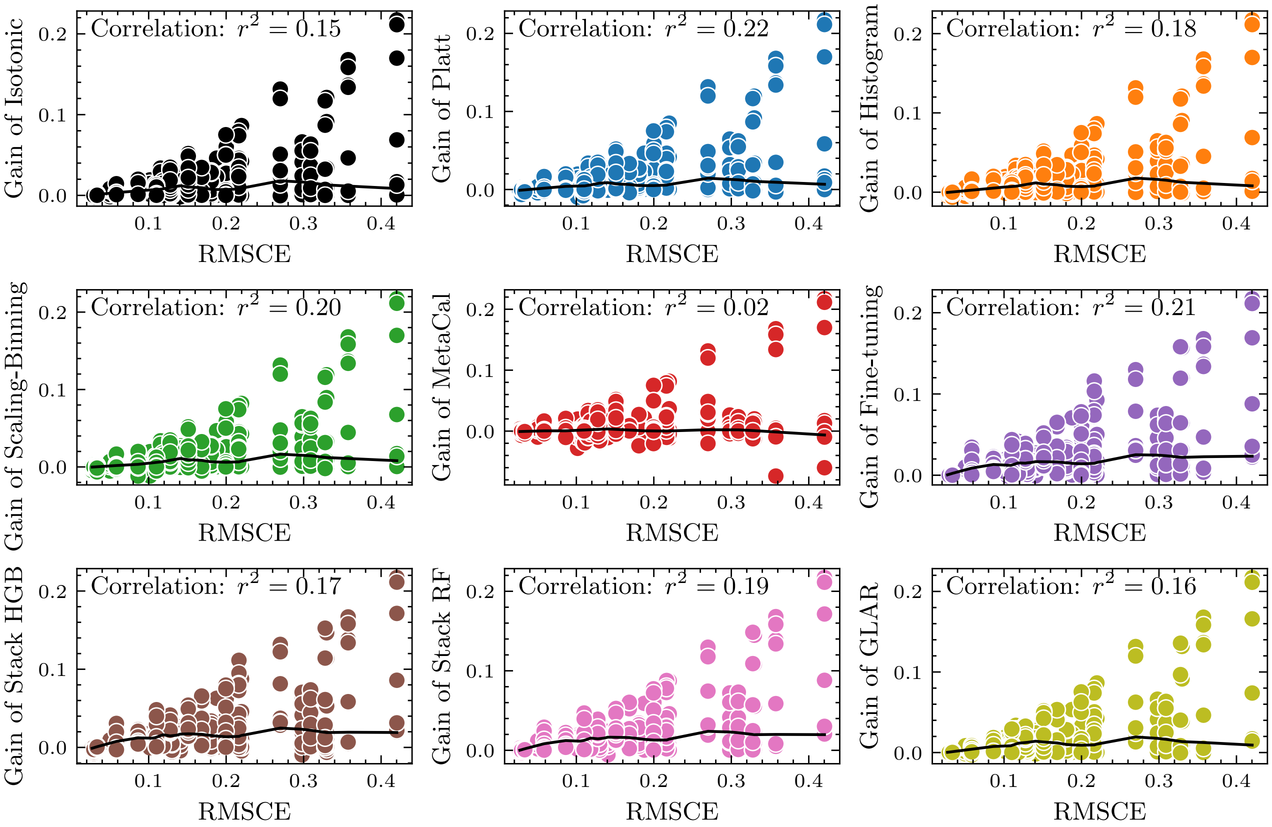


Figure 17: **Gain vs** RMSCE. Gain of each post-training method vs the RMSCE of the model.

D Supplementary results

D.1 Influence of $t^\star$: gains as a function of the utility regime

We investigate the influence of the regime $t^\star$ on the potential gain of post-training. We found that, on average across all models and datasets, values of $t^\star$ in the range $[0.05, 0.95]$ yield higher regrets with a peak around 0.5 (Fig. 18). This corresponds to exchange rates of 1:19 to 1:1, with a peak around 1:1. Similarly to Van Calster and Vickers (2015), we observe that miscalibration is less likely to cause regret when $t^\star$ is close to the event rate $\mathbb{E}[Y]$ (Fig. 20).

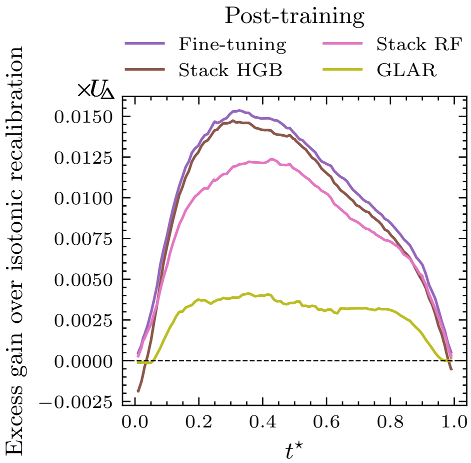


Figure 18: Average gain of post-training over the gain of recalibration as a function of the utility-derived $t^\star$.

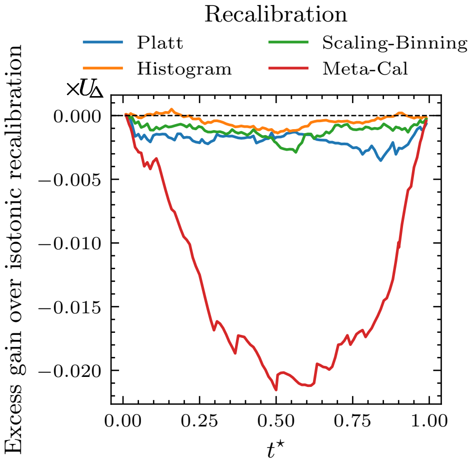


Figure 19: Average gains of recalibration approaches compared to the gain of isotonic recalibration as a function of the utility-derived $t^\star$.

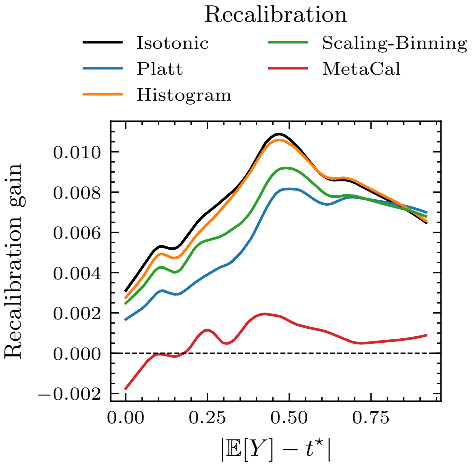


Figure 20: Average gains of recalibration approaches compared to the gain of isotonic recalibration as a function of the distance of the utility-derived $t^{\star}$ to the event rate $\mathbb{E}[Y]$. LOWESS curves were fitted with a width of 0.2.