Decision-Making Under Miscalibration

Guy N. Rothblum[1] ✉ 🏠
Weizmann Institute, Rehovot, Israel

Gal Yona ✉ 🏠
Weizmann Institute, Rehovot, Israel

Abstract

How should we use ML-based predictions (e.g., risk of heart attack) to inform downstream binary classification decisions (e.g., undergoing a medical procedure)? When the risk estimates are perfectly calibrated, the answer is well understood: a classification problem's cost structure induces an optimal treatment threshold $j^\star$. In practice, however, predictors are often miscalibrated, and this can lead to harmful decisions. This raises a fundamental question: how should one use potentially miscalibrated predictions to inform binary decisions?

In this work, we study this question from the perspective of algorithmic fairness. Specifically, we focus on the impact of decisions on protected demographic subgroups, when we are only given a bound on the predictor's anticipated degree of subgroup-miscalibration. We formalize a natural (distribution-free) solution concept for translating predictions into decisions: given anticipated miscalibration of α, we propose using the threshold j that minimizes the worst-case regret over all α-miscalibrated predictors, where the regret is the difference in clinical utility between using the threshold in question and using the optimal threshold in hindsight. We provide closed form expressions for j when miscalibration is measured using both expected and maximum calibration error which reveal that it indeed differs from $j^\star$ (the optimal threshold under perfect calibration).

2012 ACM Subject Classification Theory of computation → Models of learning

Keywords and phrases risk prediction, calibration, algorithmic fairness, multi-group fairness

Digital Object Identifier 10.4230/LIPIcs.ITCS.2023.92

Funding *Guy N. Rothblum*: This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17), from the U.S.-Israel Binational Science Foundation (grant 2018102), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Part of this work was done while the author was visiting Microsoft Research
Gal Yona: Additionally supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, by a research grant from the Estate of Tully and Michele Plesser, and by a Google PhD fellowship. Part of this work was done while the author was visiting Google.

1 Introduction

ML-based predictions are used to inform increasingly consequential decisions about individuals. We consider a setup in which a single risk predictor (e.g., estimating the risk of a cardiovascular event in a 10-year followup period) is used to inform multiple downstream binary classification decisions (e.g., whether to prescribe a certain medication, or have the individual undergo a medical procedure). Naturally, these problems may differ in their *cost structure* (the relative costs of correct and incorrect predictions). For example, a false positive can be quite costly in the context of a potentially dangerous medical procedure, but less costly in the context

[1] Currently at Apple.

14th Innovations in Theoretical Computer Science Conference (ITCS 2023).
Editor: Yael Tauman Kalai; Article No. 92; pp. 92:1–92:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of prescribing a medication without significant side effects[2]. The cost structure determines the *utility* of a proposed classifier; specifically, we consider the Net Benefit [35], defined as a weighted combination of the fraction of true positive and false positive predictions, with the weights determined by the cost structure.

With this setup in mind, how should predictions be translated into binary treatment decisions? When the risk predictor is *calibrated*, this question is well understood: we should treat individuals whose predicted risk exceeds the *therapeutic threshold*[3], $j^\star$ (a function of the cost structure of the classification problem) [27]. Here, (perfect) calibration is the requirement that for each prediction "category" $v \in [0,1]$, of the individuals for whom the predicted risk was v, exactly a v-fraction actually receive positive outcomes. Previous work demonstrated, both in real-world scenarios and through simulations, that applying $j^\star$ to *miscalibrated* risk predictions negatively effects the utility of downstream decisions [6, 33]; see Section 2.2 for an extended discussion.

1.1 From predictions to decisions under miscalibration

Our work is motivated by concerns about the impact of decisions on protected demographic subgroups of the population. While the above discussion stresses the importance of calibration, risk predictors are often miscalibrated on demographic subgroups [4]. Starting with [14], a series of works aim to address this concern by learning *multi-calibrated* risk predictors, whose calibration errors are bounded for a rich collection of subgroups $\mathcal{C}$. Crucially, however, we do not expect to achieve perfect or near-perfect subgroup calibration: assuming the number of training samples is bounded, and the collection $\mathcal{C}$ is rich, calibration errors – while bounded – are inevitable. This raises the concern that the utility experienced by protected subgroups (whose predictions are not perfectly calibrated) will be negatively impacted.

Our starting point is the following question: if $j^\star$ is the optimal threshold to apply to a predictor p that is *perfectly* calibrated (irrespective of both the distribution on inputs and the specifics of the predictor p), is there an analogously optimal threshold $\hat{j}$ to apply to an imperfectly-calibrated predictor? This raises many natural questions: which solution concept should we use to determine optimality? What should the threshold be? Is it actually different from $j^\star$? We focus on the setting where we know an upper bound α on the predictor's calibration error (and this bound applies to all subgroups in the collection).

To build intuition, consider a classification task for which the therapeutic threshold is $j^\star = 0.05$ (this means the benefit from a true positive prediction is 19 times larger than the harm from a false positive prediction; see the discussion in Section 2.2). Should we treat individuals whose *predicted* risk is 0.04? Under perfect calibration we would be guaranteed that on average, these individuals would not benefit from treatment; but what if we anticipate some amount of miscalibration? If we use $j^\star$, but our predictor is under-estimating the risk (the probability of positive outcomes w.r.t the target distribution), we will regret not treating. Perhaps we should use a lower threshold, e.g. $\hat{j} = 0.03$? But if we use $\hat{j}$ and our predictor is *over*-estimating the risk, we will regret treating! Balancing these competing scenarios is tricky, since the non-symmetric errors within a certain miscalibration bound α (over- or under-estimation) may have different consequences in our context. Given these complex considerations, what treatment threshold should we use?

[2] For the remainder of this paper we have this example in mind and therefore refer to the binary decisions as treat/don't treat decisions.

[3] The therapeutic threshold is the smallest *true* risk level (true probability of a positive outcome) where treatment gives non-negative utility. If a predictor is calibrated, then applying the therapeutic threshold to the *predicted* risks maximizes the utility. The well-known fact that thresholding the predictions at 0.5 minimizes the ℓ_1 (i.e., 0/1) loss (symmetric costs) is a special case.

We give an overview of our main contributions:

A new solution concept: minimizing the worst-case regret under miscalibration

We formalize the above intuition by proposing that decisions should be made using a *regret minimization* approach. We define the regret from thresholding a predictor p at a threshold $j \in [0,1]$ rather than at a threshold $j' \in [0,1]$ w.r.t a distribution D', as the difference between the expected utility obtained by using j' and the expected utility obtained by using j (the expectation is over D').

Naturally, the eventual form of such a regret-minimizing threshold $\hat{j}$ will depend on what information is available about the predictor, the true distribution, and the collection of groups $\mathcal{C}$. For example: are the predictor and the collection $\mathcal{C}$ fixed in advance, and can the threshold be tailored to them? Can we obtain labeled or unlabeled samples from the true distribution and use these to inform the threshold? In this work, our emphasis is on the setting where all we know is a bound on the predictor's calibration error on the subgroups in $\mathcal{C}$: the predictor p and the distribution D are unknown (as are the restrictions of D to subgroups in $\mathcal{C}$). In particular, the treatment threshold should be independent of the true distribution. We therefore study the minimization of the *worst-case* regret, defined as the maximal regret that can be experienced, over all the possible alternative thresholds j', and over all possible predictors p and distributions D' (that satisfy that p is at most α-miscalibrated on D' w.r.t. the collection $\mathcal{C}$).

We emphasize several points regarding this solution concept:

- The regret-minimizing threshold is, by design (and similarly to the therapeutic threshold $j^\star$), a *distribution free* notion: it is only a function of the anticipated miscalibration α and of the cost structure (as manifested in the treatment threshold $j^\star$). In particular, the regret-minimizing threshold does not vary between groups or individuals. This is important, since the groups may intersect: if the threshold varied by group, which threshold would we use for an individual who is a member of several groups?
- The worst-case regret does not depend on the collection $\mathcal{C}$: since we account for the worst-case predictor p and distribution D' (of bounded calibration error), the regret experienced by any sub-group is automatically bounded. Indeed, the regret is also bounded for sub-groups outside of the collection $\mathcal{C}$, so long as they experience bounded miscalibration error.
- Looking ahead, the therapeutic threshold $j^\star$ does not always minimize the regret: it diverges from the regret-minimizing threshold when we anticipate errors ($\alpha > 0$) and when the problem's cost structure is asymmetrical.

Minimax-regret as a solution concept for robust decision-making has roots a variety of disciplines, including econometrics [23, 20], statistics [2] and operations [16]. Specifically, our notion of regret is related to the notion of *Clinical Harm*, a performance measure from the decision-analytic literature [32]. The clinical harm of a classification model is the difference between its utility and the utility of (the better of) two naive baselines: treating all individuals (treatment threshold 0) and treating none (threshold 1). We measure regret with respect to *all* possible alternative threshold (not just 0 and 1) and thus the regret for using a threshold is always at least as large as its clinical harm. Bounding the worst-case regret also implies bounds on the clinical harm.

Simple closed-form expressions for the regret-minimizing thresholds

We provide simple and efficiently-computable closed-form solutions for computing the regret-minimizing threshold, when the miscalibration is quantified using the standard measures of maximum and expected calibration errors (Definitions 2 and 1). This is our main technical contribution. We note that the regret-minimizing threshold is the solution to a min-max problem over an exponential space of possibilities (in particular, all possible predictors and distributions that conform to the miscalibration bound), so it is infeasible to compute it by enumerating over all possibilities.

Our derivation highlights that the optimal threshold naturally follows a "conservatism in the face of uncertainty" approach. Recall the example above, where $j^\star = 0.05$ and the predicted risk is 0.04: intuitively, it made sense to hedge our bets and treat these individuals, since the potential benefit if their risk was underestimated is much larger than the potential harm if the risk was over-estimated. The regret-minimizing threshold gives a rigorous explanation and formalization for this intuition.

On a technical level, we obtain the closed-form expressions in two steps. First, we show that the regret can always be maximized by a predictor that comes from a very restricted class, where all individuals have one of (at most) two possible risk levels. We then derive a closed-form expression for computing a regret-minimizing threshold over this restricted class. See Section 5.

Discussion: Beyond worst-case regret

In this work we have initiated the study of decision-making under miscalibration. Our work shows that even slight miscalibration may require taking the miscalibration into account – and changing the method by which we "translate" risk predictions to binary decisions. We demonstrated this behaviour for the distribution-free setting, in which p, D and $\mathcal{C}$ are assumed to be unknown, but it's natural to consider what happens when we relax some of these assumptions. For example, we may have a fixed predictor and collection of sets in mind. Perhaps we can also obtain unlabeled examples from D at the time of thresholding p. Technically, our framework can be adapted to such settings by modifying the notion of regret: instead of worst-case regret (in which we optimize over all possible predictors and distributions subject to the miscalibration bound), consider the regret w.r.t a more restricted collection of predictors and distributions. The resulting optimization problem (which may be more computationally challenging) will give rise to thresholds that can be distribution-dependent, predictor-dependent, and dependent on the collection of subgroups. An interesting direction for future exploration is whether (and how) the optimal treatment threshold under miscalibration changes when we make these additional assumptions.

Organization

The rest of the manuscript is organized as follows. In Section 2 we define our setup and the notions of calibration and utility we will use. In Section 3 we define our notion of regret. In Section 4 we show that under perfect calibration, treating at $j^\star$ guarantees non-negative regret, but that under miscalibration, the regret can be negative (and significant, depending on the cost structure). We study regret-minimizing thresholds under miscalibration in Section 5. We conclude with additional related work in Section 6.

2 Preliminaries

Let $\mathcal{X}$ denote the space of covariates and $Y \in \{0,1\}$ the modeled event (e.g. heart attack within 10 years). Let $\mathcal{D}$ denote an unknown distribution on $\mathcal{X}$. We distinguish between *risk predictors* (whose goal is to predict the probability of the target event given covariates) and *classifiers* (denoting the binary decisions that are informed by the risk estimate).

Risk predictors. Throughout, we assume the $[0,1]$ risk prediction interval is discretized to a grid of width $1/m$, $G_m = \{\frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}\}$, where $m \in \mathbb{N}$ is an external parameter controlling the fine-ness of the grid. Hence, predictors are a mapping $p : \mathcal{X} \to G_m$. As a convention, we use $p^\star$ to denote predictors that determine the data-generation process (specifying the probability of positive outcome given covariates) and p to denote estimates. For a predictor p, $\mu_{p,i}$ denotes the probability mass of individuals whose prediction is i/m: $\mu_{p,i} = \mathbf{Pr}_{x\sim\mathcal{D}}[p(x) = i/m]$. In Section 5 we use Ξ to denote all the set of all "legal" level-sets, i.e. $\{\mu_{p,i}\}_{i=1}^{m}$ such that each $\mu_{p,i}$ is non-negative and all the $\mu_{p,i}$'s sum to 1.

Classifiers. A classifier is a mapping $h : \mathcal{X} \to \{0,1\}$. Given that the outcomes are generated according to $p^\star$, we use $\mathtt{TP}_{p^\star}(h)$ and $\mathtt{FP}_{p^\star}(h)$ to denote the fraction of *true positive* and *false positive* classifications in the target population[4]: $\mathtt{TP}_{p^\star}(h) = \mathbf{Pr}_{x\sim\mathcal{D},y\sim p^\star(x)}[h(x) = 1 \wedge y = 1]$ and $\mathtt{FP}_{p^\star}(h) = \mathbf{Pr}_{x\sim\mathcal{D},y\sim p^\star(x)}[h(x) = 1 \wedge y = 0]$.

In this work, our focus is on classifiers derived from the risk predictions. I.e., in which $h(x) \in \{0,1\}$ is some function of $p(x)$ (and possibly x itself). We restrict our attention to family of transformations that determine treatment by thresholding the risk predictor at a *fixed threshold*; we use $h_{p,j}$ to denote the binary decisions obtained by thresholding p at threshold j/m: $h_{p,j}(x) := \mathbf{1}[p(x) > j/m]$.

In this section we review two different ways to quantify the performance of risk predictors. *Clinical harm* (Section 2.2) explicitly takes into account the downstream decisions informed by the predictions via a simplified utility-based analysis, and measures the potential decrease in performance relative to two naive benchmarks. *Calibration* (Section 2.1), on the other hand, does not take into account the downstream decisions and instead asks that the risk predictor outputs scores that can meaningfully be interpreted as probabilities.

2.1 Calibration

Perfect calibration requires that for every value $v \in [0,1]$, the true expectation of the outcomes among those who receive prediction v is exactly v. In practice, as mentioned, we don't expect predictors to be perfectly calibrated. We now discuss two different ways to measure *approximate* calibration. We use notions from the literature but our presentation will be slightly different: Usually, the data generating process (for us, denoted by $p^\star$) is considered as fixed and so calibration is only a property of the predictor in question. Looking ahead, we will instead think of calibration as a joint property of both p and $p^\star$ (formally, as a *relation* over pairs of predictors). In our setup, p is perfectly calibrated w.r.t $p^\star$ if for every $i \in [m]$ for which $\mu_{p,i} > 0$, $\underbrace{\mathop{\mathbf{Pr}}_{x\sim D,y\sim p^\star(x)}[y|p(x) = i/m]}_{\triangleq \tilde{y}_{p,p^\star,i}} = i/m$.

[4] Note that these are different from the true positive and false positive *rates* which measure the conditional probability, e.g. $\mathbf{Pr}_{x\sim\mathcal{D},y\sim p^\star(x)}[h(x) = 1|y = 0]$.

The calibration error on the i-th bin is therefore $|\tilde{y}_{p,p^\star,i} - i/m|$. *Expected calibration error* bounds the weighted sum of calibration errors, where the $i-$th bin is weighted according to the fractional mass of $\mathcal{D}$ that "lands" in this bin according to the predictor in question. *Maximum calibration error* instead bounds each $|\tilde{y}_{p,p^\star,i} - i|$ separately.

▶ **Definition 1** (Expected Calibration Error (ECE) [25]). *Fix $\gamma > 0$. We say that a predictor p has an expected calibration error of γ w.r.t $p^\star$ if $\sum_{i=1}^m \mu_{p,i} \cdot |\tilde{y}_{p,p^\star,i} - i/m| \leq \gamma$.*

▶ **Definition 2** (Maximum Calibration Error (MCE) [25]). *Fix $\gamma > 0$. We say that a predictor p has a maximum calibration error of γ w.r.t $p^\star$ if for every $i \in [m]$ for which $\mu_{p,i} > 0$, we have $|\tilde{y}_{p,p^\star,i} - i/m| \leq \gamma$.*

We note that both ECE and MCE depend on the underlying distribution via the level sets $\{\mu_{p,i}\}_{i=1}^m$. We thus use $(p, p^\star) \in R_{ECE}(\gamma)$ and $(p, p^\star) \in R_{MCE}(\gamma)$ as shorthand notation for the ECE and MCE relations, when the relevant $\{\mu_{p,i}\}_{i=1}^m$'s are clear from context.

We also note that MCE is a stronger requirement than ECE: keeping $\{\mu_{p,i}\}_{i=1}^m$ fixed, $(p, p^\star) \in R_{MCE}(\gamma)$ implies $(p, p^\star) \in R_{ECE}(\gamma)$.

2.2 Utility-based metrics

Background. Predictive models are often evaluated using notions of accuracy, such as calibration and various measures of discrimination (sensitivity, specificity, AUC). However, these methods often have little clinical relevance: e.g., how high an AUC is high enough to justify clinical use of a prediction model? [3] Naturally, the answer depends on the consequences of the particular clinical decisions informed by the predictive model. Decision analytic methods explicitly consider the clinical consequences of decisions. Let U_{TP}, U_{FP}, U_{TN} and U_{FN} denote the utilities for all combinations of treatment decisions and disease outcomes[5].

Given a predictor p and costs $\mathcal{U} = \{U_{TP}, U_{FP}, U_{FN}, U_{TN}\}$, how should treatment decisions be made? Consider an individual that receives a prediction of i/m. Recalling the short-hand notation $\tilde{y}_{p,p^\star,i} \triangleq \mathbf{Pr}_{x\sim\mathcal{D},y\sim p^\star(x)}[y = 1|p(x) = v]$, their expected utility from opting to receive treatment is $\tilde{y}_{p,p^\star,i} \cdot U_{TP} + (1 - \tilde{y}_{p,p^\star,i}) \cdot U_{FP}$; similarly, their expected utility from opting to not receive treatment is $\tilde{y}_{p,p^\star,i} \cdot U_{FN} + (1 - \tilde{y}_{p,p^\star,i}) \cdot U_{TN}$. Thus, from a utility theory perspective, the optimal threshold, which we denote $j^\star$ and refer to as the *therapeutic threshold*, is the point for which these two utilities are equal: $\tilde{y}_{p,p^\star,j^\star} \cdot U_{TP} + (1 - \tilde{y}_{p,p^\star,j^\star}) \cdot U_{FP} = \tilde{y}_{p,p^\star,j^\star} \cdot U_{FN} + (1 - \tilde{y}_{p,p^\star,j^\star}) \cdot U_{TN}$. Under *perfect calibration* $\tilde{y}_{p,p^\star,j^\star} = j^\star/m$; plugging this in and re-organizing, we obtain $\frac{m-j^\star}{j^\star} = \frac{U_{TP}-U_{FN}}{U_{TN}-U_{FP}}$. In other words, under perfect calibration, we have the following relationship between the therapeutic threshold $j^\star$ and the derived costs:

$$\frac{m - j^\star}{j^\star} = \frac{P}{L} \tag{1}$$

where the *profit* $P \triangleq U_{TP} - U_{FN}$ is the difference in utilities from making a positive instead of a negative prediction for disease among those *with* the disease, and the *loss* $L \triangleq U_{TN} - U_{FP}$ is the negative of the difference in utilities from making a positive instead of a negative prediction for disease among those *without* the disease.

[5] For a concrete example, [11] give the following numeric example for colorectal cancer: $U_{FN} = -100$, for the possibility of death and morbidity due to failing to detect colorectal cancer; $U_{FP} = -1$, for the risk of bleeding or perforation of the colon; $U_{TP} = -11$, for the risk of bleeding or perforation of the colon and the lowered chance of death or morbidity from colorectal cancer due to early detection; and finally, U_{TN} is set to zero as a reference value.

Net Benefit, Clinical Utility and Clinical Harm. Decision curve analysis [35] is a simplified utility-based analysis: instead of specifying the full costs $\mathcal{U} = \{U_{TP}, U_{FP}, U_{FN}, U_{TN}\}$, it only relies on the relative values P and L[6]. By fixing the profit at $P = 1$, the relative harm of a False Positive prediction, following Equation (1), is given by $L = -\frac{j^\star}{m-j^\star}$. [35] proceed to define the *Net Benefit* of the predictor p as the weighted average of False Positives and True Positives of binary classifier $h_{p,j^\star}$ (obtained by thresholding the risk scores at $j^\star$): $\texttt{NetBenefit}(p) = P \cdot \texttt{TP}_{p^\star}(h_{p,j^\star}) - L \cdot \texttt{FP}_{p^\star}(h_{p,j^\star}) = \texttt{TP}_{p^\star}(h_{p,j^\star}) - \frac{j^\star}{m-j^\star} \cdot \texttt{FP}_{p^\star}(h_{p,j^\star})$.

This is the called Net Benefit of the model as it reflects the "effective" fraction of True Positives: the fraction of True Positive predictions, minus the number of False Positive predictions – as "valued" in terms of True Positives.

Suppose we calculated the Net Benefit of a risk model p at some threshold to be, say, 0.151. How "good" is this value? Is it high enough to justify treatment using the model? One simple "sanity check" is that the Net Benefit of the model is better than the Net Benefit obtained by two naive alternatives, that could have been carried out *without* the risk prediction model: (i) to treat everyone, and (ii) to treat no one. The *Clinical Utility* of a prediction model p is the excess Net Benefit over the better of these two simple baselines, and a model is said to be *Clinically harmful* if there exists a treatment threshold for which the clinical utility is negative [32]. When the risk threshold is varied, the Net Benefit of the model, the Net Benefit of the default strategies and the model's Clinical Utility can be plotted by risk threshold to obtain a *decision curve*; see [17, 36] for an overview of reading such decision curves from a practitioner's viewpoint.

3 Defining Regret

Our work extends classic decision-curve analysis by applying ideas from online learning and regret minimization. Our notion of regret requires two extensions of the classic Net Benefit: the first makes the Net Benefit symmetric and the second generalizes it to classifiers derived by thresholds *other* than $j^\star$. We refer to the result as the *generalized Net Benefit*, and derive a useful closed-form expression for computing it.

Symmetric Net Benefit

The Net Benefit as originally defined in [35] is not symmetric: it takes values in the range $(-\infty, \beta)$, where β is the base rate in the target population. This is an artifact of the (arbitrary) choice to fix the profit (rather than the loss) at 1. To address this, we will consider the Net Benefit w.r.t new *symmetric* relative costs P' and L', ensuring that the ratio remains unchanged, $P/L = P'/L'$ (recall, from Equation 1, only the ratio of the costs matters in this analysis). [7]

Disentangled Net Benefit

It's crucial to note that in the above discussion, the therapeutic threshold $j^\star$ served **two distinct roles**: on one hand, it defined the transformation from predictions (p) to binary decisions ($h_{p,j^\star}$); on the other hand, it also implicitly captured the costs in question, and

[6] For example, in the example of colorectal cancer discussed above [11], the profit is $P = 89$ the loss is $L = 1$.

[7] Specifically, we will ensure that the *lower* of the two costs is set to 1: *(i)* If $j^\star < m/2$, then $P > L$, so we set $L' = 1$ and obtain $P' = \frac{m-j^\star}{j^\star}$; *(ii)* If $j^\star > m/2$, then $P < L$, so we set $P' = 1$ and obtain $L' = \frac{j*}{m-j*}$.

thus determined how the derived binary classifier $h_{p,j^\star}$ was evaluated. In other words, the existing literature on Net Benefit implicitly assumes that the transformation from predictions to decisions will use the therapeutic threshold, $j^\star$. As discussed, an important component of our work is considering making binary decisions using thresholds *other than* $j^\star$. This is a natural perspective when miscalibration is involved, since the derivation of $j^\star$ as the "optimal" threshold was under the assumption of perfect calibration. It is therefore crucial for us to disentangle these two roles of $j^\star$. We will work with the following more general notion of Net Benefit, that quantifies the Net Benefit of making predictions using a threshold j when the *actual costs* are implied by $j^\star$:

▶ **Definition 3** ((Disentangled) Net Benefit). *Fix a predictor p. The Net Benefit of making predictions using a threshold j when the therapeutic threshold is $j^\star$ is*

$$\Lambda(p, j; p^\star, j^\star) = P' \cdot \mathtt{TP}_{p^\star}(h_{p,j}) - L' \cdot \mathtt{FP}_{p^\star}(h_{p,j}) \tag{2}$$

We can combine Equation 2 with the definitions of the symmetric costs P' and L' from the previous section to obtain a useful expression for the disentangled Net Benefit (see Appendix A for the proof):

▶ **Lemma 4.** *For every $\{\mu_{p,i}\}_{i=1}^m$, $p, p^\star$ and $j, j^\star$,*

$$\Lambda(p, j; p^\star, j^\star) = \sum_{i>j} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m - j^\star,\ j^\star\}} \tag{3}$$

Regret

Recall that the Clinical Harm from making decisions using a predictor p and threshold j is the Net Benefit of the better of the two naive strategies (treat all, and treat none) minus the Net Benefit of the model. Since *treat all* is like thresholding p at 0 and and *treat none* is like thresholding p at 1, we can write the clinical harm succinctly as $\max\{\Lambda(p, 0; p^\star, j^\star),\ \Lambda(p, 1; p^\star, j^\star)\} - \Lambda(p, j; p^\star, j^\star)$.

With this in mind, it is natural to "measure" j not only against the two specific thresholds $\{0, 1\}$, but also against *every* constant threshold. We refer to this as the *regret* incurred by making decisions using a predictor p and threshold j, akin to the notion of regret from online learning (where the "benchmark" class we compete against is the class of all constant thresholds):

$$\mathtt{regret}(p, j; p^\star, j^\star) = \max_{j'} \Lambda(p, j'; p^\star, j^\star) - \Lambda(p, j; p^\star, j^\star) \tag{4}$$

In Section 5 we study the question of minimizing the worst-case regret under miscalibration.

4 Treating at the therapeutic threshold

In this section, we formally study the guarantees obtained by thresholding p at the therapeutic threshold $j^\star$. In particular, we show that: (i) when p is *perfectly calibrated*, using $j^\star$ guarantees *no regret* (in that the regret from Equation 4 is non-positive); (ii) when p is α-miscalibrated, using $j^\star$ can lead to regret that scales like $\alpha \cdot m$.

▶ **Lemma 5** (No regret under perfect calibration.).

$$(p, p^\star) \in R_{ECE}(0) \implies \mathtt{regret}(p, j^\star; p^\star, j^\star) \leq 0$$

See Appendix B for the proof.

▶ **Example** (Regret under α-miscalibration)**.** *Consider the following two constant predictors: $p^\star \equiv 1-\alpha$ and $p \equiv 1$. Clearly, $(p,p^\star) \in R_{ECE}(\alpha)$. Consider the Clinical Utility of treating at the therapeutic threshold when the latter is equal to $j^\star = m-1$. Note that since $p=1$, the Net Benefit of the model is essentially the Net Benefit of the "treat all" strategy, which is $((1-\alpha)\cdot 1) - (\alpha \cdot \frac{j^\star}{m-j^\star}) = 1-\alpha\cdot m$. On the other hand, the Net Benefit of the "treat none" strategy is 0. Hence the Clinical Utility of the model is $1-\alpha\cdot m$, and the regret of using p with threshold $j^\star$ is lower bounded by $\alpha\cdot m - 1$.*

5 Decision-making under miscalibration

The example above shows that under miscalibration, there are extreme cases in which using the therapeutic threshold $j^\star$ "as is" could be very costly. It also highlights two natural ways to reduce this "worst case" clinical harm: the first is to obtain predictors with stronger calibration guarantees to begin with, and the second is to use coarser predictions (effectively limiting the largest expressible cost of a misclassification in this framework). In practice, some amount of miscalibration is unavoidable, and having the ability to make granular predictions is important. We therefore turn our attention to a third "knob": the mechanism by which we translate predictions to decisions. Specifically, we ask:

Given $\alpha > 0$ and the therepeutic threhsold $j^\star$, which threhsold minimizes the worst-case regret under α-miscalibration?

Giving rise to the following min-max optimization problem:

$$\hat{j}(\alpha,m,j^\star) \in \arg\min_{j} \underbrace{\max_{\substack{\{\mu_{p,i}\}_{i=1}^{m}\in\Xi \\ (p,p^\star)\in R(\alpha)}} \texttt{regret}(p,j;p^\star,j^\star)}_{\triangleq c(j;j^\star,R(\alpha))} \tag{5}$$

Here, $R(\alpha)$ is a miscalibration relation (e.g. ECE or MCE) and we define the "cost" of a threshold, $c(j;j^\star,R(\alpha))$, as the maximal regret that we might experience by using it. The maximum is effectively over any distribution D on the domain $\mathcal{X}$ and every predictors p and $p^\star$ that are within the allowed miscalibration level; in practice, since the only effect the distribution D has on this expression is via the level sets of the predictor p, we replace D with the masses of the level sets, $\{\mu_{p,i}\}_{i=1}^{m}$. Recall that Ξ ensures these choices are legal (non-negative and sum to 1).

We refer to a solution of Equation (5) as the *conservative therapeutic threshold*. We note that the fact we have shown that using $j^\star$ can cause $\alpha\cdot m$ clinical harm does not yet imply that $j^\star$ itself is *not* a solution to (5); in principle, we anticipate that in this setup every choice of threshold j would give rise to some regret, and so using $j^\star$ itself still could be the best threshold, in a *regret minimization* sense.

Our main technical contribution is a closed-form expression for the optimal threshold (a solution to Equation (5)), when miscalibration is measured using either maximum calibration error:

▶ **Theorem 6** (Optimal thresholds under miscalibration)**.** *Fix $m \in \mathbb{N}$ and $\alpha > 0$. Then $\hat{j}_{MCE}(\alpha,m,j^\star)$ (the optimal threshold when the MCE can be as large as α) is given by*

$$\begin{cases} 0 & 0 \le j^\star \le \frac{\alpha}{2}\cdot m \\ 2j^\star - \alpha m & \frac{\alpha}{2}\cdot m \le j^\star \le \alpha m \\ j^\star & \alpha m \le j^\star \le (1-\alpha)m \\ 2j^\star - (1-\alpha)m & (1-\alpha)m \le j^\star \le (1-\frac{\alpha}{2})\cdot m \\ m & (1-\frac{\alpha}{2})\cdot m \le j^\star \le m \end{cases}$$

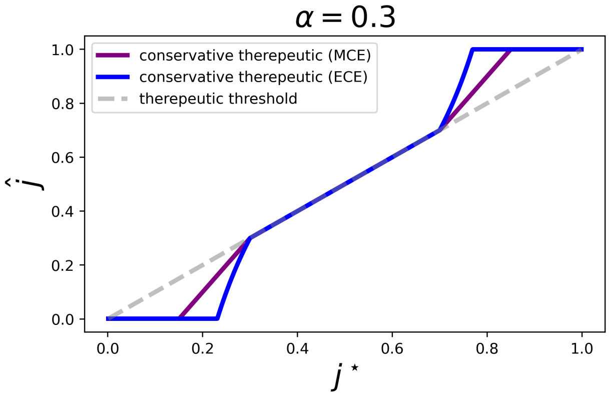

Figure 1 Plotting the conservative therapeutic thresholds $\hat{\jmath}_{MCE}$ and $\hat{\jmath}_{ECE}$ (see Theorem 6) vs the standard therapeutic threshold.

and $\hat{\jmath}_{ECE}(\alpha, m, j^\star)$ *(the optimal threshold when the ECE can be as large as* α*) is given by*

$$\begin{cases} \max\left\{1,\ (1+\alpha)\cdot m - \frac{\alpha m^2}{j^\star}\right\} & j^\star < \alpha m \\ j^\star & \alpha m < j^\star < (1-\alpha)m \\ \min\left\{m,\ \alpha m \cdot \frac{j^\star}{m - j^\star}\right\} & j^\star > (1-\alpha)m \end{cases}$$

Theorem 6 reveals that irrespective of how we measure miscalibration and for every $\alpha > 0$, $j^\star$ is indeed *not* the regret-minimizing threshold (at least not everywhere). Intuitively, the expression for $\hat{\jmath}$ reveals a natural "conservatism in the face of uncertainty" behaviour: the conservative threshold "clips" the decisions in an α-region around the edges of the prediction interval, where the relative costs of false positive or false negative predictions are highest, and α-miscalibration could be most detrimental. Figure 1 demonstrates visually how the optimal thresholds differ from $j^\star$, and also how they differ from one another. Intuitively, we see that $\hat{\jmath}_{ECE}$ is even more "conservative" than $\hat{\jmath}_{MCE}$. Intuitively, this makes sense – bounding the expected calibration error of a predictor is a weaker guarantee than bounding the maximum calibration error.

See Appendix C for the proof; we sketch the main idea below. For both notions, the proof consists of three parts:

1. Reducing the original optimization problem (Equation 5) into a simplified one, where p is additionally constrained to be "simple". Formally, simple depends on how we measure miscalibration: for MCE, this is a constant predictor (i.e., that is supported on only a single value); and for ECE, this is an "almost" constant predictor (i.e., supported on at most two values).
2. Showing that under the constraint that p is "simple", we can carefully derive a closed-form expression for the maximal regret under miscalibration (cost).
3. Showing that the minimizer of the above cost follows the expressions defined in the theorem statement.

6 Further Related Work

Clinical decision-making. The derivation of the therapeutic threshold dates back to [27]. Decision curve analysis was introduced by Vickers and Elkin in [35], and has since gained popularity in the area of clinical risk prediction and decision making, see e.g. [22, 24, 21, 1]. Other decision-analytic measures such as Relative Utility [3] and Net Reclassification

Improvement [18] are also common, but are all simple transformations of the Net Benefit [34]. The effect of discrimination on the Net Benefit was explored in [34, 35]. Recently, [29, 28] consider the impact of different fairness interventions on clinical utility.

Calibration: clinical decision-making perspective. Calibration is an important requirement when using risk prediction models to support medical decision making. Several studies demonstrate that miscalibration can negatively effect a model's clinical utility. For example, Collins and Altman [6] evaluated prognostic models for predicting the 10-year risk of cardiovascular disease on a large cohort of general practice patients in the UK. Their results demonstrate that the Framingham risk score overestimates the risk in women and in men, resulting in a harmful model for risk thresholds of around 20% and higher. Additional examples in this vein are discussed in [33], which like us, study the effect of miscalibration on Net Benefit. They use simulation studies to demonstrate that inducing miscalibration can make a model clinically harmful, stressing the importance of calibration as a desiderata for risk prediction models. Our theoretical results and experiments complement theirs: motivated by the observation that miscalibration can never be eliminated entirely, we shift the focus to the mapping from predictions to decisions as a means to combat anticipated miscalibration. Finally, [32] define a hierarchy of calibration measures, proving that "moderate calibration", which corresponds to the standard notion of (exact) calibration, guarantees no clinical harm. We generalize this result to the stronger notion of no regret and depart from their analysis by considering the effect of miscalibration.

Calibration: additional perspectives. Calibration is an important concept that has been studied in different contexts (fairness, safety, robustness, etc). As discussed, most relevant to our work is the multi-calibration requirement proposed in [14]. Technically, it sits between the notions of "moderate calibration" and "strong calibration" from [31, 32]. Multi-calibration has since been extended [10] and studied in a variety of additional contexts: ranking [9], multi-group learning with general loss functions [30], online decision-making [13], and the prediction of higher order moments [15]. Most related to this work is [4], which applies the methodology of [14, 19] to post-process risk models, such as the Framingham risk score mentioned above, to satisfy subgroup calibration w.r.t a large collection of demographic subgroups of interest. The connection between calibration and eliminating clinical harm serves as another motivation for guaranteeing predictions are also calibrated on a subgroup-level. Finally, calibration has also been studied in the context of domain generalization, see e.g. [38, 37].

Post-processing predictors. Our focus on translating predictions to binary decisions is conceptually related to a fundamental question in learning theory, which is whether predictors that are optimal for one loss function (e.g., ℓ_2 loss, or calibration loss) can be post-processed into predictors that are optimal for other objectives. [7, 5] study the power of post-processing calibrated scores into decisions when the objective is to equalize certain statistical fairness notions across subgroups, which is not our objective. More related to our work are results in [26, 8] showing that for accuracy metrics including F-scores and AUC, applying the optimal post-processing transformation to the ℓ_2-loss minimizing predictor in a class $\mathcal{H}$ yields a classifier competitive with the best classifier in the class of binary classifiers derived by thresholding models from $\mathcal{H}$. Recently, [12] proposed the notion of an *omnipredictor*, which takes this idea one step further and seeks a single predictor that can be post-processed to be competitive with a large collection of loss functions and w.r.t arbitrary classes of functions.

Their work leverages connections to the notion of multicalibration to prove that this objective is computationally feasible: there exists a predictor p with the guarantee that for every convex loss function, applying the optimal post-processing transformation for $p^\star$ to p yields an optimal classifier for this loss, albeit with a degradation in performance that depends on the Lipchitzness of the loss function in question. Our point about the clinical harm scaling like $O(\alpha \cdot m)$ under α-miscalibration can be viewed as a lower bound showing that there indeed exist interesting and natural cases in which this is unavoidable (here, the granularity parameter m acts as the Lipchitz constant). Unlike the above works, we don't assume that the distribution on which the predictor was trained is the same distribution on which performance is evaluated after post-processing. Rather, our only knowledge of the target distribution is that the predictor's miscalibration error is bounded, and we aim to minimize the worst-case regret over all target distributions and all possible alternative treatment thresholds. We further remark that the post-processing transformation we apply (the conservative threshold) is in fact *not* what we would apply to $p^\star$ (which has 0 miscalibration error).

References

1 Todd A Alonzo. Clinical prediction models: a practical approach to development, validation, and updating: by ewout w. steyerberg, 2009.

2 Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

3 Stuart G Baker, Nancy R Cook, Andrew Vickers, and Barnett S Kramer. Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(4):729–748, 2009.

4 Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558, 2021.

5 Ran Canetti, Aloni Cohen, Nishanth Dikkala, Govind Ramnarayan, Sarah Scheffler, and Adam Smith. From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency*, pages 309–318, 2019.

6 Gary S Collins and Douglas G Altman. Predicting the 10 year risk of cardiovascular disease in the united kingdom: independent and external validation of an updated version of qrisk2. *Bmj*, 344, 2012.

7 Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

8 Krzysztof Dembczyński, Wojciech Kotłowski, Oluwasanmi Koyejo, and Nagarajan Natarajan. Consistency analysis for binary classification revisited. In *International Conference on Machine Learning*, pages 961–969. PMLR, 2017.

9 Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

10 Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

11 Mitchell H Gail and Ruth M Pfeiffer. On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239, 2005.

12 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint*, 2021. `arXiv:2109.05389`.

13 Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint*, 2021. `arXiv:2101.01739`.

14 Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948, 2018.

15 Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

16 Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

17 Kathleen F Kerr, Marshall D Brown, Kehao Zhu, and Holly Janes. Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use. *Journal of Clinical Oncology*, 34(21):2534, 2016.

18 Kathleen F Kerr, Zheyu Wang, Holly Janes, Robyn L McClelland, Bruce M Psaty, and Margaret S Pepe. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology (Cambridge, Mass.)*, 25(1):114, 2014.

19 Michael P. Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

20 Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

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

22 Susan Mallett, Steve Halligan, Matthew Thompson, Gary S Collins, and Douglas G Altman. Interpreting diagnostic accuracy studies for patient care. *Bmj*, 345, 2012.

23 Charles F Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246, 2004.

24 Karel GM Moons, Joris AH de Groot, Kristian Linnet, Johannes B Reitsma, and Patrick MM Bossuyt. Quantifying the added value of a diagnostic test or marker. *Clinical chemistry*, 58(10):1408–1417, 2012.

25 Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

26 Nagarajan Natarajan, Oluwasanmi Koyejo, Pradeep Ravikumar, and Inderjit S Dhillon. Optimal decision-theoretic classification using non-decomposable performance metrics. *arXiv preprint*, 2015. `arXiv:1505.01802`.

27 Stephen G Pauker and Jerome P Kassirer. Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine*, 293(5):229–234, 1975.

28 Stephen R Pfohl, Yizhe Xu, Agata Foryciarz, Nikolaos Ignatiadis, Julian Genkins, and Nigam H Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. *arXiv preprint*, 2022. `arXiv:2202.01906`.

29 Stephen R Pfohl, Haoran Zhang, Yizhe Xu, Agata Foryciarz, Marzyeh Ghassemi, and Nigam H Shah. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. *Scientific reports*, 12(1):1–13, 2022.

30 Guy N Rothblum and Gal Yona. Multi-group agnostic pac learnability. *ICML*, 2021.

31 Werner Vach. Calibration of clinical prediction rules does not just assess bias. *Journal of clinical epidemiology*, 66(11):1296–1301, 2013.

32 Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

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

34 Ben Van Calster, Andrew J Vickers, Michael J Pencina, Stuart G Baker, Dirk Timmerman, and Ewout W Steyerberg. Evaluation of markers and risk prediction models: overview of relationships between nri and decision-analytic measures. *Medical Decision Making*, 33(4):490–501, 2013.

35 Andrew J Vickers and Elena B Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006.

36 Andrew J Vickers, Ben van Calster, and Ewout W Steyerberg. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and prognostic research*, 3(1):1–8, 2019.

37 Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. On calibration and out-of-domain generalization. *arXiv preprint*, 2021. arXiv:2102.10395.

38 Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable calibration with lower bias and variance in domain adaptation. *arXiv preprint*, 2020. arXiv:2007.08259.

A Proof of Lemma 4

Note that by definition, we can write the fraction of True Positives and False Positives as

$$\mathtt{TP}_{p^\star}(h_{p,j}) = \sum_{i>j} \mu_{p,i} \cdot \tilde{y}_{p,p^\star,i}, \quad \mathtt{FP}_{p^\star}(h_{p,j}) = \sum_{i>j} \mu_{p,i} \cdot (1 - \tilde{y}_{p,p^\star,i})$$

So, by definition of the symmetric Net Benefit, we have:

$$\begin{aligned}
\Lambda(p,j;p^\star,j^\star) &= P' \cdot \mathtt{TP}_{p^\star}(h_{p,j}) - L' \cdot \mathtt{FP}_{p^\star}(h_{p,j}) \\
&= \sum_{i>j} \mu_{p,i} \left[P' \cdot \tilde{y}_{p,p^\star,i} - L' \cdot (1 - \tilde{y}_{p,p^\star,i})\right] \\
&= \begin{cases} \sum_{i>j} \mu_{p,i} \left[\frac{m-j^\star}{j^\star} \cdot \tilde{y}_{p,p^\star,i} - (1 - \tilde{y}_{p,p^\star,i})\right] & j^\star \le m/2 \\ \sum_{i>j} \mu_{p,i} \left[\tilde{y}_{p,p^\star,i} - \frac{j^\star}{m-j^\star}(1 - \tilde{y}_{p,p^\star,i})\right] & j^\star > m/2 \end{cases} \\
&= \begin{cases} \sum_{i>j} \mu_{p,i} \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{j^\star} & j^\star \le m/2 \\ \sum_{i>j} \mu_{p,i} \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{m-j^\star} & j^\star > m/2 \end{cases} \\
&= \sum_{i>j} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m-j^\star,\ j^\star\}}
\end{aligned}$$

B Proof of Lemma 5

Consider $p, p^\star$ such that $(p, p^\star) \in R_{ECE}(0)$. By definition, this guarantees that for every level set i, $\tilde{y}_{p,p^\star,i} = i/m$. Using Lemma 4, we have that for every threshold j',

$$\begin{aligned}
\Lambda(p,j';p^\star,j^\star) - \Lambda(p,j^\star;p^\star,j^\star) &= \sum_{i>j'} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m-j^\star,\ j^\star\}} - \sum_{i>j^\star} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m-j^\star,\ j^\star\}} \\
&= \sum_{j'<i\le j^\star} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m-j^\star,\ j^\star\}} - \sum_{j^\star<i\le j'} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m-j^\star,\ j^\star\}} \\
&= \sum_{j'<i\le j^\star} \mu_{p,i} \underbrace{\frac{i - j^\star}{\min\{m-j^\star,\ j^\star\}}}_{\le 0} - \sum_{j^\star<i\le j'} \mu_{p,i} \underbrace{\frac{i - j^\star}{\min\{m-j^\star,\ j^\star\}}}_{\ge 0} \\
&\le 0
\end{aligned}$$

Since this holds for every threshold j', it also holds for the maximum, hence the required.

C Proof of Theorem 6

C.1 Maximum Calibration Error

MCE: Part 1. Recall that we defined $c(j; j^\star, R) \triangleq \max_{\{\mu_{p,i}\}_{i=1}^m \in \Xi,\ (p,p^\star)\in R(\alpha)} \texttt{regret}(p, j; p^\star, j^\star)$. In this section we will use $\tilde{c}(j; j^\star, R)$ to denote the maximal regret when p is additionally constrained to be a constant predictor. Our objective in this part is to show that for every j and $j^\star$, $c(j; j^\star, R_{\mathbf{MCE}}) = \tilde{c}(j; j^\star, R_{\mathbf{MCE}})$. Note that the direction $\tilde{c}(j; j^\star, R_{\mathbf{MCE}}) \leq c(j; j^\star, R_{\mathbf{MCE}})$ is trivial; it's left to prove the other direction.

Fix $j, j^\star$ and $\alpha > 0$. We want to show that for every $\{\mu_{p,i}\}_{i=1}^m \in \Xi$ and predictors $p, p^\star$ such that $(p, p^\star) \in R_{MCE}(\alpha)$, and for every j_R, there exist $\{\mu_{\tilde{p},i}\}_{i=1}^m \in \Xi$ and predictors $\tilde{p}, \tilde{p^\star}$ and $\tilde{j}_R$ such that:

$$(\tilde{p}, \tilde{p}^*) \in R_{MCE}(\alpha) \tag{6}$$

$$\exists i \quad \text{such that} \quad \mu_{\tilde{p},i} = 1 \tag{7}$$

$$\texttt{regret}(j; j^\star, p, p^\star, j_R) \leq \texttt{regret}(j; j^\star, \tilde{p}, \tilde{p^\star}, \tilde{j}_R) \tag{8}$$

Fix $\{\mu_{p,i}\}_{i=1}^m \in \Xi$ and predictors $p, p^\star$ such that $(p, p^\star) \in R_{MCE}(\alpha)$, and fix a threshold j_R. W.l.o.g, assume $j_R < j$. Let $i^* = \arg\max_{j_R<i<j} \frac{m \cdot \tilde{y}_{p,p^*,i} - j^*}{\min\{m-j^*, j^*\}}$, and define the following predictors: $\tilde{p}(x) \equiv i^*/m$ and $\tilde{p}^*(x) \equiv \tilde{y}_{p,p^*,i^*}$. Note that $\tilde{p}$ is a constant prediction, so (7) holds. Also, (6) follows directly by the assumption that $(p, p^\star) \in R_{MCE}(\alpha)$ (but note that here we use the fact that miscalibration is measured using *maximum* calibration error). We will now argue that taking $\tilde{j}_R = j_R$, also (8) holds. Note that by definition of $i^\star$,

$$\begin{aligned}
\texttt{regret}(j; j^\star, p, p^\star, j_R) &= \sum_{i>j_R} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m - j^\star, j^\star\}} - \sum_{i>j} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i} - j^\star}{\min\{m - j^\star, j^\star\}} \\
&\leq \sum_{j_R<i<j} \mu_{p,i} \cdot \frac{m \cdot \tilde{y}_{p,p^\star,i^*} - j^\star}{\min\{m - j^\star, j^\star\}} \\
&= \frac{m \cdot \tilde{y}_{p,p^\star,i^*} - j^\star}{\min\{m - j^\star, j^\star\}} \cdot \underbrace{\sum_{j_R<i<j} \mu_{p,i}}_{\leq 1} \\
&\leq \frac{m \cdot \tilde{y}_{p,p^\star,i^*} - j^\star}{\min\{m - j^\star, j^\star\}}
\end{aligned}$$

On the other hand, by the definition of $\tilde{p}$ and $\tilde{p^\star}$, we have:

$$\begin{aligned}
\texttt{regret}(j; j^\star, \tilde{p}, \tilde{p}^*, \tilde{j}_R) &= \sum_{i>j_R} \mu_{\tilde{p},i} \cdot \frac{m \cdot \tilde{y}_{\tilde{p},\tilde{p}^*,i} - j^*}{\min\{m - j^\star, j^\star\}} - \sum_{i>j} \mu_{\tilde{p},i} \cdot \frac{m \cdot \tilde{y}_{\tilde{p},\tilde{p}^*,i} - j^\star}{\min\{m - j^\star, j^\star\}} \\
&= \sum_{j_R<i<j} \mu_{\tilde{p},i} \cdot \frac{m \cdot \tilde{y}_{\tilde{p},\tilde{p}^*,i} - j^\star}{\min\{m - j^\star, j^\star\}} \\
&= \frac{m \cdot \tilde{y}_{\tilde{p},\tilde{p}^*,i^*} - j^\star}{\min\{m - j^\star, j^\star\}} \\
&= \frac{m \cdot \mathbf{E}_x[\tilde{p}^*(x)] - j^\star}{\min\{m - j^\star, j^\star\}} \\
&= \frac{m \cdot \tilde{y}_{p,p^*,i^*} - j^\star}{\min\{m - j^\star, j^\star\}}
\end{aligned}$$

Combining, we have shown (8), which concludes the claim that $c(j; j^\star, R_{\mathbf{MCE}}) \leq \tilde{c}(j; j^\star, R_{\mathbf{MCE}})$.

MCE: Part 2. Next, we will show that the maximal regret w.r.t $(p, p^\star) \in R_{MCE}(\alpha)$ when p is constant is

$$\frac{\max\left\{\min\left\{m - j^\star, j - j^\star + \alpha m\right\},\quad \min\left\{j^\star, j^\star - j + \alpha m\right\}\right\}}{\min\{m - j^\star, j^\star\}} \tag{9}$$

Fix a constant predictor p and an arbitrary predictor $p^\star$. Let v denote the value p is supported on (i.e., the level set for which that $\mu_{p,v} = 1$). Note that w.l.o.g, we can additionally assume that $p^\star$ is constant too - the regret consists of the difference between Net Benefits, which is only a function of $\tilde{y}_{p,p^\star,v}$, which is the same if we replace $p^\star$ with its expectation. We use $v^\star$ to denote the value $p^\star$ is supported on. In this case, the regret of j w.r.t $j^\star$ is exactly

$$\begin{cases} 0 & v < j, v^\star < j^\star \\ \left|\frac{m \cdot v^\star - j^\star}{\min\{m - j^\star, j^\star\}}\right| & v < j, v^\star \geq j^\star \\ \left|\frac{m \cdot v^\star - j^\star}{\min\{m - j^\star, j^\star\}}\right| & v \geq j, v^\star < j^\star \\ 0 & v \geq j, v^\star \geq j^\star \end{cases}$$

Recall that to determine the cost of j, we would like to choose $v, v^\star$ to maximize this expression. We will consider two choices for v: $v \leq j$ and $v > j$:

1. When $v \leq j$, to maximize the regret we want to choose $v^\star \geq j$. In that case the regret will be $\left|\frac{m \cdot v^\star - j^\star}{\min\{m - j^\star, j^\star\}}\right|$, which is maximized when $v^\star$ is maximized; under the maximum calibration constraint, $m \cdot v^\star$ can be as large as $\min\{m, v + \alpha m\}$, which in turn can be as large as $\min\{m, j + \alpha m\}$. So the regret in this case is

$$\frac{\min\{m, j + \alpha m\} - j^\star}{\min\{m - j^\star, j^\star\}} = \frac{\min\{m - j^\star, j - j^\star + \alpha m\}}{\min\{m - j^\star, j^\star\}}$$

2. When $v > j$, to maximize the regret we want to choose $v^\star < j$. In that case the regret will be $\left|\frac{j^\star - m \cdot v^\star}{\min\{m - j^\star, j^\star\}}\right|$, which is maximized when $v^\star$ is minimized; under the maximum calibration constraint, $m \cdot v^\star$ can be as small as $\max\{0, v - \alpha m\}$, which in turn can be as small as $\max\{0, j - \alpha m\}$. So the regret in this case is

$$\frac{\max\{0, j - \alpha m\} - j^\star}{\min\{m - j^\star, j^\star\}} = \frac{\max\{-j^\star, j - j^\star - \alpha m\}}{\min\{m - j^\star, j^\star\}} = \frac{\min\{j^\star, j^\star - j + \alpha m\}}{\min\{m - j^\star, j^\star\}}$$

Finally, since we can also chose whether $v \leq j$ or $v > j$, we obtain that the overall maximum regret is exactly the expression in Equation (9), which concludes this part.

MCE: Part 3 . It's left to argue why the minimizer of the cost from Equation (9) is $\hat{j}$ defined in the theorem statement. First, we note that since the denominator of Equation (9) is non-negative and independent of j, it suffice to minimize the following cost

$$\max\left\{\min\left\{m - j^\star, j - j^\star + \alpha m\right\},\quad \min\left\{j^\star, j^\star - j + \alpha m\right\}\right\}$$

Note:

- In the first minimum, the first expression "dominates" (i.e., is smallest) iff $j \geq (1 - \alpha)m$;
- In the second minimum, the first expression "dominates" (i.e., is smallest) iff $j \leq \alpha m$.

So we can start by simplifying according to the following regions: Left (meaning $j \leq \alpha \cdot m$), Middle (meaning $\alpha m \leq j \leq (1 - \alpha)m$) and Right (meaning $j \geq (1 - \alpha)m$).

- **Middle**: The expression becomes $\max\{j - j^\star + \alpha m, \quad j^\star - j + \alpha m\}$. The first term increases in j and the second decreases in j, so the optimal choice is $j - j^\star + \alpha m = j^\star - j + \alpha m \Rightarrow j = j^\star$.
- **Left**: The expression becomes $\max\{j - j^\star + \alpha m, \quad j^\star\}$. Similarly, the optimal choice in this range is $j = 2j^\star - \alpha m$.
- **Right**: The expression becomes $\max\{m - j^\star, \quad j^\star - j + \alpha m\}$, so the optimal choice is $j = 2j^\star - (1-\alpha)m$.

To summarize, given $j^\star$, we have several options: choose $j = j^\star$ incurring a cost of αm; choose $j = 2j^\star - \alpha m$ incurring a cost of $j^\star$; or choose $j = 2j^\star - (1-\alpha)m$ incurring a cost of $m - j^\star$. From this, we can derive the optimal threshold $\hat{j}$, depending on $j^\star$:

- if $\alpha m \leq j^\star \leq (1-\alpha)m$, we have that $j^\star \geq \alpha m$ and $m - j^\star \geq \alpha m$, so the optimal choice is $j = j^\star$.
- if $j^\star \leq \alpha m$, then the optimal choice is $j = \max\{0, \ 2j^\star - \alpha m\}$.
- If $j^\star \geq (1-\alpha)m$, then the optimal choice is $j = \min\{1, \ 2j^\star - (1-\alpha)m\}$.

Which is precisely the expression in the theorem statement, concluding the proof of the optimal threshold under miscalibration when using maximum calibration error.

C.2 Expected Calibration Error

We now follow the same logic but when miscalibration is measured using *expected* calibration error (ECE). Here, the reduction in the first part is more subtle: we show that we also need to take into account slightly more complex predictors, that can be supported on two values (not one).

ECE: Part 1. Fix $j, j^\star$ and $\alpha > 0$. This time, we want to show that for every $\{\mu_{p,i}\}_{i=1}^m \in \Xi$ and predictors $p, p^\star$ such that $(p, p^\star) \in R_{ECE}(\alpha)$, and for every j_R, there exist $\{\mu_{\tilde{p},i}\}_{i=1}^m \in \Xi$ and predictors $\tilde{p}, \tilde{p^\star}$ and $\tilde{j}_R$ such that:

$$(\tilde{p}, \tilde{p^*}) \in R_{ECE}(\alpha) \tag{10}$$

$$\exists i_1, i_2 \quad \text{such that} \quad \mu_{\tilde{p},i_1} + \mu_{\tilde{p},i_2} = 1 \tag{11}$$

$$\texttt{regret}(j; j^\star, p, p^\star, j_R) \leq \texttt{regret}(j; j^\star, \tilde{p}, \tilde{p^\star}, \tilde{j}_R) \tag{12}$$

Fix $\{\mu_{p,i}\}_{i=1}^m \in \Xi$ and predictors $p, p^\star$ such that $(p, p^\star) \in R_{ECE}(\alpha)$, and fix a threshold j_R. W.l.o.g, assume $j_R < j$. Intuitively, our construction of the "simple" $\tilde{p}$ will collect all the level sets outside $[j_R, j]$ into a single level set, whose value is arbitrary (say 0), and all the level sets inside $[j_R, j]$ into a single level set, whose value is the mean. We will construct $\tilde{p^\star}$ accordingly to ensure that the resulting pair $(\tilde{p}, \tilde{p^\star})$ is still in the ECE relation.

- Consider the predictor $\tilde{p}$ that puts μ_1 mass on a value v_1 and μ_2 mass on a value v_2, defined as follows: $v_1 = 0$; $v_2 = \frac{\sum_{i \in [j_R, j]} \mu_{p,i} \cdot (i/m)}{\sum_{i \in [j_R, j]} \mu_{p,i}}$; $\mu_1 = \sum_{i \notin [j_R, j]} \mu_{p,i}$ and $\mu_2 = \sum_{i \in [j_R, j]} \mu_{p,i}$.
- For $\tilde{p^\star}$, we define it such that $\tilde{y}_{\tilde{p},\tilde{p^\star},v_1} = v_1^\star$ and $\tilde{y}_{\tilde{p},\tilde{p^\star},v_2} = v_2^\star$, where $v_1^\star, v_2^\star$ are defined as follows: $v_1^\star = 0$; $v_2^\star = \frac{\sum_{i \in [j_R, j]} \mu_{p,i} \cdot \tilde{y}_{p,p^\star,i}}{\sum_{i \in [j_R, j]} \mu_{p,i}}$.

Note that Equation (11) holds since by construction $\mu_1 + \mu_2 = 1$. For (10), we bound the ECE of $\tilde{p}$ w.r.t $\tilde{p^\star}$:

$$\begin{aligned}\sum_i \mu_{\tilde{p},i} \cdot \left|i/m - \tilde{y}_{\tilde{p},\tilde{p^\star},i}\right| &= \mu_1 \cdot |v_1 - v_1^\star| + \mu_2 \cdot |v_2 - v_2^\star| \\ &= \mu_2 \cdot |v_2 - v_2^\star| \\ &= \left| \sum_{i \in [j_R, j]} \mu_{p,i} \cdot (i/m) - \sum_{i \in [j_R, j]} \mu_{p,i} \cdot \tilde{y}_{p,p^\star,i} \right| \\ &\le \sum_{i \in [j_R, j]} \mu_{p,i} \cdot \left|i/m - \tilde{y}_{p,p^\star,i}\right| \\ &\le \alpha\end{aligned}$$

where the last transition is by the fact that $(p, p^\star) \in R_{ECE}(\alpha)$. Finally, for Eq. (12):

$$\begin{aligned}\texttt{regret}(j; j^\star, \tilde{p}, \tilde{p^\star}, \tilde{j}_R) &= \sum_{j_R < i < j} \mu_{\tilde{p},i} \cdot \frac{m \cdot \tilde{y}_{\tilde{p},\tilde{p^\star},i} - j^\star}{\min\{m - j^\star, j^\star\}} \\ &= \mu_2 \cdot \frac{m \cdot v_2^\star - j^\star}{\min\{m - j^\star, j^\star\}} \\ &= \sum_{i \in [j_R, j]} \mu_{p,i} \cdot \frac{m \cdot \frac{\sum_{i \in [j_R, j]} \mu_{p,i} \cdot \tilde{y}_{p,p^\star,i}}{\sum_{i \in [j_R, j]} \mu_{p,i}} - j^\star}{\min\{m - j^\star, j^\star\}} \\ &= \sum_{i \in [j_R, j]} \mu_{p,i} \cdot \frac{m \cdot \frac{\sum_{i \in [j_R, j]} \mu_{p,i} \cdot \tilde{y}_{p,p^\star,i}}{\sum_{i \in [j_R, j]} \mu_{p,i}} - \frac{j^\star \cdot \sum_{i \in [j_R, j]} \cdot \mu_{p,i}}{\sum_{i \in [j_R, j]} \cdot \mu_{p,i}}}{\min\{m - j^*, j^*\}} \\ &= \frac{\sum_{i \in [j_R, j]} \mu_{p,i} \cdot (m \cdot \tilde{y}_{p,p^\star,i} - j^\star)}{\min\{m - j^\star, j^\star\}} \\ &= \texttt{regret}(j; j^\star, p, p^\star, j_R)\end{aligned}$$

This concludes the proof of Part 1.

ECE: Part 2. Next, we will show that the maximal regret incurred by a threshold j w.r.t $j^\star$, under $(p, p^\star) \in R_{ECE}(\alpha)$ when p is additionally constrained to be "almost constant" (in the sense formalized above) is

$$\frac{1}{\min\{m - j^\star, j^\star\}} \cdot \begin{cases} \max\left\{j^\star, \quad \alpha m \cdot \frac{m - j^\star}{m - j}\right\} & j < j^\star, \quad j \le \alpha m \\ \max\left\{j^\star - j + \alpha m, \quad \alpha m \cdot \frac{m - j^\star}{m - j}\right\} & \alpha m < j \le j^\star \\ \max\left\{j - j^\star + \alpha m, \quad \alpha m \cdot \frac{j^\star}{j}\right\} & j^\star < j \le (1 - \alpha) m \\ \max\left\{m - j^\star, \quad \alpha m \cdot \frac{j^\star}{j}\right\} & j \ge j^\star, \quad j > (1 - \alpha) m \end{cases} \tag{13}$$

Fix $j, j^\star$. Consider some choice of threshold j_R and predictors $p, p^\star \in R_{ECE}(\alpha)$. Since p is assumed to be "simple", we will assume it is supported on two values v_1, v_2 (where v_1 is between j and j_R, and v_2 is not), with $v_1^\star, v_2^\star$ denoting the respective conditional expectation of $p^\star$, similar to the notation used in Part 1. We will split into two cases: $j_R \le j$ and $j_R > j$.

When $j_R \le j$,

$$\min\{m - j^\star, j^\star\} \cdot \texttt{regret}(j; j^\star, p, p^\star, j_R) = \sum_{j_R < i < j} \mu_{p,i} \cdot (m \tilde{y}_{p,p^\star,i} - j^\star) = \mu_1 \cdot (m \cdot v_1^\star - j^\star)$$

Since we want to maximize this expression, we want to choose $v_1^\star$ as large as possible; we can therefore assume w.l.o.g that $v_1^\star \geq v_1$ (and also trivially $v_2^\star \geq v_2$). We can therefore re-parametrize the expressions in terms of v_1, v_2 and $\delta_1 = v_1^\star - v_1 > 0$, $\delta_2 = v_2^\star - v_2 > 0$, yielding the following optimization problem (whose value is the maximal regret we are interested in):

$$\max_{\mu_1, v_1, v_2, \delta_1, \delta_2} \mu_1 \cdot (m \cdot (v_1 + \delta_1) - j^\star) \text{ subject to } \begin{cases} v_1 & \in [j_R, j] \\ v_2 & \notin [j_R, j] \\ \mu_1 & \in [0,1] \\ \delta_1 & \in [0, 1 - v_1] \\ \delta_2 & \in [0, 1 - v_2] \\ \mu_1 \cdot \delta_1 + (1-\mu_1) \cdot \delta_2 & \leq \alpha \end{cases}$$

Note that we can re-write the objective as $m \cdot \mu_1 v_1 + m \cdot \mu_1 \delta_1 - \mu_1 \cdot j^\star$. Since $\mu_1 \delta_1$ will just be α (or as large as it can be), we are left with maximizing $m \cdot \mu_1 (v_1 - j^\star/m)$. Now, the behaviour will depend on the relationship between j and $j^\star$:

- If $j \geq j^\star$, then the optimal thing is to put the most weight on μ_1 as possible; i.e., have $v_1 = j/m$, $\mu_1 = 1$ and $\delta_1 = \alpha$. For the solution to be legal, $v_1 + \delta_1$ must not exceed 1. So the optimal choice is $\delta_1 = \min\{\alpha, 1 - j/m\}$. In particular:
  - If $\alpha < 1 - j/m$, the value of the objective would be $j - j^\star + \alpha m$.
  - If $\alpha > 1 - j/m$, the value of the objective would be $m - j^\star$.
- If $j < j^\star$, then the optimal thing would be to put as little weight on μ_1 as possible; i.e. have $v_1 = j$, $\delta_1 = 1 - j/m$ and $\mu_1 = \frac{\alpha}{1-j/m}$. The value of the objective would be $\frac{\alpha}{1-j/m} \cdot (m(j/m + 1 - j/m) - j^\star) = \alpha m \cdot \frac{m-j^\star}{m-j}$.

To summarize, the maximal regret achievable when $j_R \leq j$ is

$$\frac{1}{\min\{m - j^\star, j^\star\}} \cdot = \begin{cases} j - j^\star + \alpha m & j^\star < j < (1-\alpha)m \\ m - j^\star & j > j^\star, \quad j > (1-\alpha)m \\ \alpha m \cdot \frac{m-j^\star}{m-j} & j < j^\star \end{cases} \tag{14}$$

Let us now repeat this analysis for the case $j_R > j$. This time,

$$\min\{m - j^\star, j^\star\} \cdot \texttt{regret}(j; j^\star, p, p^\star, j_R) = \sum_{j < i < j_R} \mu_{p,i} \cdot (m \tilde{y}_{p,p^\star,i} - j^\star) = \mu_1 \cdot (j^\star - m \cdot v_1^\star)$$

Since we want to maximize this expression, this time we want to choose $v_1^\star$ as small as possible; we can therefore assume w.l.o.g that $v_1^\star \leq v_1$ (and also trivially $v_2^\star \leq v_2$). Re-parametrizing in terms of $\delta_1 = v_1 - v_1^\star \geq 0$ and $\delta_2 = v_2 - v_2^\star \geq 0$, we have the following optimization problem:

$$\max_{\mu_1, v_1, v_2, \delta_1, \delta_2} \mu_1 \cdot (j^\star - m \cdot (v_1 - \delta_1)) \text{ subject to } \begin{cases} v_1 & \in [j, j_R] \\ v_2 & \notin [j, j_R] \\ \mu_1 & \in [0,1] \\ \delta_1 & \in [0, v_1] \\ \delta_2 & \in [0, v_2] \\ \mu_1 \cdot \delta_1 + (1-\mu_1) \cdot \delta_2 & \leq \alpha \end{cases}$$

Since we can re-write the objective as $\mu_1 \cdot j^\star + m \cdot \mu_1 \delta_1 - m \cdot \mu_1 v_1$, and again $\mu_1 \delta_1$ will just be α (or as large as it can be), we are left with maximizing $m \cdot \mu_1 (j^\star/m - v_1)$. The behaviour again depends on the relationship between j and $j^\star$:

- If $j < j^\star$, then the optimal thing is to put the most weight on μ_1 as possible; i.e., have $v_1 = j/m$, $\mu_1 = 1$ and $\delta_1 = \alpha$. For the solution to be legal, $v_1 - \delta_1$ must be non-negative. So the optimal feasible choice is $\delta_1 = \min\{\alpha, j/m\}$. In particular:
  - If $\alpha < j/m$, the value of the objective will be $j^\star - j + \alpha m$.
  - If $\alpha > j/m$, the value of the objective will be $j^\star$.
- If $j > j^\star$, then the optimal thing is to put as little weight on μ_1 as possible; i.e. have $v_1 = j/m$, $\delta_1 = j/m$ and $\mu_1 = \frac{\alpha}{j/m}$. The value of the objective would be $\frac{\alpha}{j/m} \cdot (j^\star - m(j/m - j/m)) = \alpha m \cdot \frac{j^\star}{j}$.

To summarize, the maximal regret achievable when $j_R > j$ is

$$\frac{1}{\min\{m - j^\star, j^\star\}} \cdot = \begin{cases} j^\star - j + \alpha m & \alpha m < j < j^\star \\ j^\star & j < j^\star, \quad j < \alpha m \\ \alpha m \cdot \frac{j^\star}{j} & j > j^\star \end{cases} \tag{15}$$

Finally, since we also maximize over j_R, we combine Equations (14) and (15) to obtain precisely the expression from Equation (13), which concludes the proof of this part.

ECE: Part 3. It's left to argue that the minimizer of the cost in Equation (13) is the expression from the theorem statement. We will ignore the term $\frac{1}{\min\{m-j^\star, j^\star\}}$ since it is non-negative and does not depend on j.

As we did for MCE, we will begin by simplifying the cost according to the following regions: Left (meaning $j \leq \alpha \cdot m$), Left Middle (meaning $\alpha m < j \leq j^\star$), Right Middle (meaning $j^\star \leq j < (1-\alpha)m$), and Right (meaning $j > (1-\alpha)m$).

- **Left**: The expression becomes $\max\{j^\star, \alpha m \cdot \frac{m-j^\star}{m-j}\}$, and the optimal choice in this region is $\max\{0,\ (1+\alpha)m - \frac{\alpha m^2}{j^\star}\}$.
- **Left Middle**: The expression becomes $\max\{j^\star - j + \alpha m,\ \alpha m \cdot \frac{m-j^\star}{m-j}\}$, and $j^\star$ is an optimal choice in this region.[8]
- **Right Middle**: The expression becomes $\max\{j - j^\star + \alpha m,\ \alpha m \cdot \frac{j^\star}{j}\}$, and $j^\star$ is an optimal choice in this region.[9]
- **Right**: The expression becomes $\max\{m - j^\star, \alpha m \cdot \frac{j^\star}{j}\}$, and the optimal choice in this region is $\min\{m,\ \alpha m \cdot \frac{j^\star}{m-j^\star}\}$.

To summarize, given $j^\star$, we have several options: choose $\max\{0,\ (1+\alpha)m - \frac{\alpha m^2}{j^\star}\}$, incurring a cost of $j^\star$; choose $j^\star$, incurring a cost of αm; or choose $\min\{m,\ \alpha m \cdot \frac{j^\star}{m-j^\star}\}$, incurring a cost of $m - j^\star$. From this, we can derive the optimal threshold $\hat{j}$ as a function of $j^\star$, which exactly follows that in the theorem statement, concluding the required.

[8] This follows by observing that $j^\star$ is a solution to the quadratic equation (in j) $j^\star - j + \alpha m = \alpha m \cdot \frac{m-j^\star}{m-j}$

[9] This follows by observing that $j^\star$ is a solution to the quadratic equation (in j) $j - j^\star + \alpha m = \alpha m \cdot \frac{j^\star}{j}$