Improving Screening Processes via Calibrated Subset Selection

Lequn Wang [1 2] **Thorsten Joachims** [1] **Manuel Gomez Rodriguez** [3]

Abstract

Many selection processes such as finding patients qualifying for a medical trial or retrieval pipelines in search engines consist of multiple stages, where an initial screening stage focuses the resources on shortlisting the most promising candidates. In this paper, we investigate what guarantees a screening classifier can provide, independently of whether it is constructed manually or trained. We find that current solutions do not enjoy distribution-free theoretical guarantees and we show that, in general, even for a perfectly calibrated classifier, there always exist specific pools of candidates for which its shortlist is suboptimal. Then, we develop a distribution-free screening algorithm—called Calibrated Subset Selection (CSS)—that, given any classifier and some amount of calibration data, finds near-optimal shortlists of candidates that contain a desired number of qualified candidates in expectation. Moreover, we show that a variant of CSS that calibrates a given classifier multiple times across specific groups can create shortlists with provable diversity guarantees. Experiments on US Census survey data validate our theoretical results and show that the shortlists provided by our algorithm are superior to those provided by several competitive baselines.

1. Introduction

Screening is an essential part of many selection processes where an often intractable number of candidates is reduced to a shortlist of the most promising candidates for detailed—and more resource intensive—evaluation. Screening thus enables an allocation of resources that improves the overall quality of the decisions under limited resources. Examples of such screening problems are: finding patients in a large database of electronic health records to manually evaluate for qualification to take part in a medical trial (Liu et al., 2021); the first stage of a multi-stage retrieval pipeline of a search engine (Covington et al., 2016; Geyik et al., 2019); or which people to reach out to with a personalized invitation to apply to a specific job posting (Raghavan et al., 2020). In each of these examples, there is significant pressure to make high-quality, unbiased screening decisions quickly and efficiently, often about thousands or even millions of candidates under limited resources and additional diversity requirements (Bendick Jr et al., 1997; Bertrand & Mullainathan, 2004; Johnson et al., 2016; Covington et al., 2016). While these screening decisions have been made manually or through manually constructed rules in the past, automated predictive tools for optimizing screening decisions are becoming more prevalent (Cowgill, 2018; Chamorro-Premuzic & Akhtar, 2019; Raghavan et al., 2020; Sánchez-Monedero et al., 2020).

Algorithmic screening has been typically studied together with other high-stakes decision making problems as a supervised learning problem (Corbett-Davies et al., 2017; Kilbertus et al., 2020; Sahoo et al., 2021). Under this perspective, algorithmic screening reduces to: (i) training a classifier that estimates the probability that a candidate is *qualified* given a set of observable features; (ii) designing a deterministic threshold rule that shortlists candidates by thresholding the candidates' probability values estimated by the classifier. Here, the classifier and the threshold rule aim to maximize a measure of average accuracy and average utility, respectively, possibly subject to diversity constraints. Only very recently, it has been argued that the classifier must also satisfy threshold calibration, a specific type of group calibration, to be able to accurately estimate the average utility of the threshold rule (Sahoo et al., 2021).

Unfortunately, the above screening algorithms do not enjoy distribution-free guarantees[1] on the quality of the shortlisted candidates. As a result, they may not always hold their promise of increasing the efficiency of a selection

[1]Department of Computer Science, Cornell University [2]Most of the work was done during Wang's internship at the Max Planck Institute for Software Systems. [3]Max Planck Institute for Software Systems. Correspondence to: Lequn Wang <lw633@cornell.edu>, Thorsten Joachims <tj@cs.cornell.edu>, Manuel Gomez Rodriguez <manuelgr@mpi-sws.org>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]We refer to distribution-free guarantees as finite-sample distribution-free guarantees, since we never have infinite amount of data in practice.

process without decreasing the quality of the screening decisions. The results in this paper bridge this gap. In particular, we focus on developing screening algorithms that, without making any distributional assumptions on the candidates, provide the smallest shortlists of candidates, among those in a given pool, containing a desired expected number of qualified candidates with high probability.[2]

Our Contributions. We first show that, even if a classifier is perfectly calibrated, in general, there always exist specific pools of candidates for which the shortlist created by a policy that makes decisions based on the probability predictions from the classifier will be significantly suboptimal, both in terms of size and expected number of qualified candidates. Then, we develop a distribution-free screening algorithm—called Calibrated Subset Selection (CSS)—that calibrates any given classifier using calibration data in a way such that the shortlist of candidates created by thresholding the candidates' probability values estimated by the calibrated classifier is near-optimal in terms of expected size and it provably contains, in expectation across all potential pools of candidates, a desired number of qualified candidates with high probability. Moreover, we theoretically characterize how the accuracy of the classifier and the amount of calibration data affect the expected size of the shortlists CSS provides. In addition, motivated by the Rooney rule (Collins, 2007), which requires that, when hiring for a given position, at least one candidate from the underrepresented group be interviewed, we demonstrate that a variant of CSS that calibrates the given classifier multiple times across different groups can be used to create a shortlist with provable diversity guarantees—namely that this shortlist contains, in expectation across all potential pools of candidates, a desired number of qualified candidates from each group with high probability.

Finally, we validate CSS on simulated screening processes created using US Census survey data (Ding et al., 2021). The results show that, compared to several competitive baselines, CSS consistently selects the shortest shortlists of candidates among those methods that select enough qualified candidates. Moreover, the results also demonstrate that the amount of human effort CSS helps reduce—the difference in size between the pools of candidates and the shortlists it provides—depends on the accuracy of the classifier it relies upon as well as the amount of calibration data. However, the expected quality of the shortlists—the expected number of qualified candidates in the shortlists—never decreases below the user-specified requirement. Our code is accessible at https://github.com/LequnWang/Improve-Screening-via-Calibrated-Subset-Selection.

[2]If the overall pool of candidates does not contain the desired number of qualified candidates with high probability, the algorithms should also determine that.

Further Related Work. Our work builds upon prior literature on distribution-free uncertainty quantification, which includes calibration and conformal prediction.

Calibration (Brier et al., 1950; Dawid, 1982; Platt et al., 1999; Zadrozny & Elkan, 2001; Gneiting et al., 2007; Gupta et al., 2020) measures the accuracy of the probability outputs from predictive models. Many notions of calibration have been proposed to measure the accuracy of the probability outputs from a classifier (Platt et al., 1999; Guo et al., 2017) or a regression model (Gneiting et al., 2007; Kuleshov et al., 2018). Arguably, the most commonly used notion is marginal (or average) calibration (Gupta et al., 2020; Zhao et al., 2020; Gneiting et al., 2007; Kuleshov et al., 2018) for both classifiers and regression models, which requires the probability outputs be marginally calibrated on the whole population. Our definition of perfectly calibrated classifier inherits from the marginal calibration definition for classifiers in prior works (Gupta et al., 2020). We also show how CSS produces an approximately calibrated regression model under average calibration definition in regression (Gneiting et al., 2007; Kuleshov et al., 2018). In the other extreme, individual calibration (Zhao et al., 2020) refers to a classifier that predicts the probability distribution for each example. We refer to classifiers with such properties the omniscient classifiers in this paper. Many recent works (Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017) discuss the relationship between calibration and fairness in binary classification, and show that marginal calibration is incompatible with many fairness definitions in machine learning. Motivated by multicalibration (Hébert-Johnson et al., 2018; Jung et al., 2021), which requires the predictive models be calibrated on multiple groups of candidates, we propose a variant of CSS that selects calibrated subsets within each group to ensure diversity in the final selected shortlists. Very recently, some works (Sahoo et al., 2021; Zhao et al., 2021; Straitouri et al., 2022; Wang & Joachims, 2022) realize that we can design calibration algorithms (and calibration definitions) well-suited for different downstream tasks to achieve better performance. CSS is specifically designed for screening processes so that it provides near-optimal shortlists in terms of the expected size among those having distribution-free guarantees on the expected number of qualified candidates.

Conformal prediction (Vovk et al., 1999; 2005; Shafer & Vovk, 2008; Romano et al., 2019; Balasubramanian et al., 2014; Gupta et al., 2020; Angelopoulos & Bates, 2021; Chzhen et al., 2021) aims to build confidence intervals on the probability outputs from predictive models. Within this literature, the work most closely related to ours is arguably the work by Bates et al. (2021), which has focused on generating set-valued predictions from a black-box predictor that controls the expected loss on future test points at a user-specified level. While one can view our problem from the

perspective of set-valued predictions, applying their methodology to find near-optimal solutions in our problem is not straightforward, and one would need to assume to have access to qualification labels for all the candidates in multiple pools, something we view as rather impractical.

Moreover, our work also builds on work in budget-constrained decision making, where one needs to first select a set of candidates to screen and then, given the result of that screening process, determine to whom to allocate the resources (Cai et al., 2020; Bakker et al., 2021). This contrasts with our work where all candidates undergo screening.

Many large-scale recommender systems adopt multi-stage pipelines (Bendick Jr & Nunes, 2013). Existing works on multi-stage recommender systems (Ma et al., 2020; Hron et al., 2021) focus on learning accurate classifiers in different stages. Complementary to these works, we assume the classifiers are already given and provide distribution-free and finite-sample guarantees on the quality of the shortlists.

2. Problem Formulation

Given a candidate with a feature vector $x \in \mathcal{X}$, we assume the candidate can be either qualified ($y = 1$) or unqualified ($y = 0$) for the selection objective[3]. Moreover, let $f : \mathcal{X} \to [0, 1]$ be a classifier that maps a candidate's feature vector x to a quality score[4] $f(x)$. The higher the quality score $f(x)$, the more the classifier believes the candidate is qualified. Given a pool of m candidates with feature vectors $\boldsymbol{x} = \{x_i\}_{i\in[m]}$, an algorithmic screening policy $\pi : [0, 1]^m \to \mathcal{P}(\{0, 1\}^m)$ maps the candidates' quality scores $\{f(x_i)\}_{i\in[m]}$ to a probability distribution over shortlisting decisions $\boldsymbol{s} = \{s_i\}_{i\in[m]}$. Here, each shortlisting decision s_i specifies whether the corresponding candidate is shortlisted ($s_i = 1$) or is not shortlisted ($s_i = 0$). For brevity, we will write $\boldsymbol{S} \sim \pi$ whenever there is no ambiguity[5]. Furthermore, we use π_f to indicate that a policy makes shortlisting decisions based on the quality scores from classifier f. This implies that for any pool of candidates $\boldsymbol{x}$ and any two candidates $i, j \in [m]$ in the pool, if $f(x_i) = f(x_j)$, then $\Pr(S_i = 1) = \Pr(S_j = 1)$. We denote the set of all possible policies π_f as Π_f.

For any screening process, we would ideally like a screening policy π that shortlists *only* candidates who are qualified ($y = 1$). Unfortunately, as long as there is no deterministic mapping between x and y, such a *perfect* screening policy π does not exist in general. Instead, our goal is to find a screening policy π that shortlists a small set of candidates that provably contains *enough* qualified candidates, without making any assumptions about the data distribution. These shortlisted candidates will then move forward in the selection process and will be evaluated in detail, possibly multiple times, until one or more qualified candidates are selected.

More specifically, let each candidate's feature vector x and label y be sampled from a data distribution $P_{X,Y} = P_X \times P_{Y \mid X}$. Then, for a pool of m candidates[6] with feature vectors $\boldsymbol{X} = \{X_i\}_{i\in[m]}$ and unobserved labels $\boldsymbol{Y} = \{Y_i\}_{i\in[m]}$, where $X_i \sim P_X$, $Y_i \sim P_{Y|X_i}$ for all $i \in [m]$, we will investigate to what extent it is possible to find screening policies π with near-optimal guarantees with respect to two different oracle policies. In particular:

(i) an oracle policy $\pi^\star$ that, for any set of candidates $\boldsymbol{x} \in \mathcal{X}^m$, shortlists the smallest set of candidates that contains, in expectation with respect to $P_{Y \mid X}$, more than k qualified candidates, *i.e.*, $\forall \boldsymbol{x} \in \mathcal{X}^m$,

$$\pi^\star \in \operatorname*{argmin}_{\pi\in\Pi} \mathbb{E}_{\boldsymbol{S}\sim\pi}\left[\sum_{i\in[m]} S_i\right], \tag{1}$$

where

$$\Pi = \left\{\pi \,\middle|\, \mathbb{E}_{\boldsymbol{S}\sim\pi,\boldsymbol{Y}\sim P}\left[\sum_{i\in[m]} S_i Y_i\right] \geq k\right\}.$$

(ii) an oracle policy $\pi^{\star\star}$ that shortlists the smallest set of candidates that contains, in expectation with respect to $P_{X,Y}$, more than k qualified candidates, *i.e.*,

$$\pi^{\star\star} \in \operatorname*{argmin}_{\pi\in\Pi} \mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi}\left[\sum_{i\in[m]} S_i\right], \tag{2}$$

where

$$\Pi = \left\{\pi \,\middle|\, \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi}\left[\sum_{i\in[m]} S_i Y_i\right] \geq k\right\}.$$

In the language of distribution-free uncertainty quantification (Balasubramanian et al., 2014), the optimality guarantees with respect to the first and the second oracle policy can be viewed as individual and marginal guarantees respectively.

[3] In practice, one needs to measure qualification using proxy variables and these proxy variables need to be chosen carefully to not perpetuate historical biases (Bogen & Rieke, 2018; Garr & Jackson, 2019; Tambe et al., 2019).

[4] Our theory and algorithms apply to any classifier with a bounded range, by scaling the scores to $[0, 1]$.

[5] We use upper case letters to denote random variables, and lower case letters to denote realizations of random variables.

[6] For ease of presentation, we assume a constant pool size m in the main paper. However, all the theoretical results and algorithms can be easily adapted to settings where the pool size changes across selection processes. Refer to Appendix A.1 for more details.

3. Impossibility of Screening with Individual Guarantees

If we had access to an omniscient classifier $f^\star(x) = \Pr(Y = 1 \,|\, X = x)$ for all $x \in \mathcal{X}$, then we could recover the oracle policy $\pi^\star$ defined by Eq. 1 just by thresholding the quality scores of each candidate in the pool. More specifically, we have the following theorem[7]:

Theorem 3.1. *The screening policy $\pi^\star_{f^\star}$ that, given any pool of m candidates with feature vectors $\boldsymbol{x} \in \mathcal{X}^m$, takes shortlisting decisions as*

$$s_i = \begin{cases} 1 & \text{if } f^\star(x_i) > t^\star, \\ \text{Bernoulli}(\theta^\star) & \text{if } f^\star(x_i) = t^\star, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

where

$$t^\star = \sup \left\{ t \in [0,1] \,\middle|\, \sum_{i \in [m]} \mathbb{I}\{f^\star(x_i) \geq t\} f^\star(x_i) \geq k \right\}$$

and

$$\theta^\star = \frac{k - \sum_{i \in [m]} \mathbb{I}\{f^\star(x_i) > t^\star\} f^\star(x_i)}{\sum_{i \in [m]} \mathbb{I}\{f^\star(x_i) = t^\star\} f^\star(x_i)},$$

is a solution (if there is one) to the constrained minimization problem defined in Eq. 1.

Unfortunately, without distributional assumptions about $P_{X,Y}$, finding the omniscient classifier $f^\star$ from data is impossible, even asymptotically, if the distribution $P_{f^\star(X)}$ induced by $P_{X,Y}$ is nonatomic, as recently shown by Barber (2020) and Gupta et al. (2020). Alternatively, one may think whether there exist other classifiers h allowing for near-optimal screening policies π_h with individual guarantees. In this context, a natural alternative is a perfectly calibrated classifier h other than the omniscient classifier $f^\star$. A classifier h is perfectly calibrated if and only if

$$\Pr(Y = 1 \,|\, h(X) = a) = a \quad \forall a \in \text{Range}(h). \tag{4}$$

However, the following theorem shows that there exist data distributions $P_{X,Y}$ for which there is no perfectly calibrated classifier $h \neq f^\star$ allowing for a screening policy π_h with individual guarantees of optimality.

Proposition 3.2. *Let $\mathcal{X} = \{a, b\}$, $\Pr(Y = 1 \,|\, X = a) = 1$, $\Pr(Y = 1 \,|\, X = b) = \frac{k}{m}$, and $f^\star$ be the omniscient classifier. Then, for any screening policy π_h using any perfectly calibrated classifier $h \neq f^\star$, there exists a pool of candidates $\boldsymbol{x} = \{x_i\}_{i \in [m]} \in \mathcal{X}^m$ such that*

$$\left| \mathbb{E}_{\boldsymbol{S} \sim \pi_h, \boldsymbol{S}^\star \sim \pi^\star_{f^\star}} \left[\sum_{i \in [m]} (S_i - S_i^\star) \right] \right| \geq \frac{m - k}{2};$$

and a pool of candidates $\boldsymbol{x}' = \{x'_i\}_{i \in [m]} \in \mathcal{X}^m$ such that

$$\left| \mathbb{E}_{\boldsymbol{Y} \sim P, \boldsymbol{S} \sim \pi_h, \boldsymbol{S}^\star \sim \pi^\star_{f^\star}} \left[\sum_{i \in [m]} (S_i - S_i^\star) Y_i \right] \right| \geq \frac{k}{2} \left(1 - \frac{k}{m} \right).$$

The above result reveals that, if $m \gg k$, for any screening policy π_h using a calibrated classifier $h \neq f^\star$, there are always scenarios in which π_h provides shortlists that are $\frac{m}{2}$ apart from those provided by the oracle policy $\pi^\star$ in size or $\frac{k}{2}$ apart in terms of the number of qualified candidates. In particular, the second part of Proposition 3.2 directly implies that there is no policy $\pi \in \Pi_h$ that satisfies the individual guarantee. This negative result motivates us to look for screening policies with marginal guarantees of optimality.

4. Screening Algorithms with Marginal Guarantees

To investigate to what extent it is possible to find screening policies with near-optimal guarantees with respect to the oracle policy $\pi^{\star\star}$ defined in Eq. 2, we focus on perfectly calibrated classifiers h with finite range, *i.e.*, $|\text{Range}(h)| < \infty$. The reason is that, similarly as in the case of the omniscient classifier $f^\star$, it is impossible to find nonatomic perfectly calibrated classifiers h from data, even asymptotically (Barber, 2020; Gupta et al., 2020). We will first introduce the optimal algorithm assuming we have access to a perfectly calibrated classifier, and discuss how the accuracy of the classifier affects the performance of the optimal algorithm. Then, we will introduce the proposed CSS algorithm that, given a classifier and some amount of calibration data, selects shortlists that are near-optimal in terms of the expected size, while ensuring that the expected number of qualified candidates in the shortlists is above the desired threshold.

An Optimal Calibration Algorithm with Perfectly Calibrated Classifiers. Let h be a perfectly calibrated classifier with $\text{Range}(h) = \{\mu_b\}_{b \in [B]}$, and denote $\rho_b = \mathbb{E}_{X \sim P}[\mathbb{I}\{h(X) = \mu_b\}]$. First, note that the classifier h induces a sample-space *partition* of $\mathcal{X}$ into B regions or bins $\{\mathcal{X}_b\}_{b \in [B]}$, where $\mu_b = \Pr(Y = 1 \,|\, X \in \mathcal{X}_b)$ and $\rho_b = \Pr(X \in \mathcal{X}_b)$. Then, the following theorem shows that the optimal screening policy $\pi^\star_h$ that shortlists the smallest set of candidates within the set of policies Π_h, which contain, in expectation with respect to $P_{X,Y}$, more than k qualified candidates, is given by a threshold decision rule.

Theorem 4.1. *The screening policy $\pi^\star_h$ that takes shortlisting decisions as*

$$s_i = \begin{cases} 1 & \text{if } h(x_i) > t_h, \\ \text{Bernoulli}(\theta_h) & \text{if } h(x_i) = t_h, \\ 0 & \text{otherwise}, \end{cases}$$

[7] All proofs can be found in Appendix C.

where

$$t_h = \sup \left\{ \mu \in \{\mu_b\}_{b\in[B]} \,\middle|\, \sum_{b\in[B]} \mu_b \mathbb{I}\{\mu_b \geq \mu\}\rho_b \geq \frac{k}{m} \right\}$$

and

$$\theta_h = \frac{k/m - \sum_{b\in[B]} \mu_b \mathbb{I}\{\mu_b > t_h\}\rho_b}{\sum_{b\in[B]} \mu_b \mathbb{I}\{\mu_b = t_h\}\rho_b},$$

is a solution (if there is one) to the constrained minimization problem defined in Eq. 2 over the set of policies Π_h.

However, it is important to realize that the expected size of the shortlists provided by $\pi_h^\star$ for different perfectly calibrated classifiers h may differ. To put it differently, not all screening policies $\pi_h^\star$ (and classifiers h) will help reduce the downstream effort by the same amount. To characterize this difference, we introduce a notion of dominance between perfectly calibrated classifiers.

Definition 4.2. *Let h and h' be perfectly calibrated classifiers. We say that h dominates h' if, for any $x_1, x_2 \in \mathcal{X}$ such that $h(x_1) = h(x_2)$, it holds that $h'(x_1) = h'(x_2)$.*

Equipped with this notion, we now characterize the difference in size between shortlists provided by different perfectly calibrated classifiers using the following corollary, which follows from Theorem 4.1.

Corollary 4.3. *Let h and h' be perfectly calibrated classifiers. If h dominates h', then*

$$\mathbb{E}_{\boldsymbol{X}\sim P, \boldsymbol{S}\sim\pi_h^\star, \boldsymbol{S}'\sim\pi_{h'}^\star} \left[\sum_{i\in[m]} (S_i - S_i') \right] \leq 0.$$

This notion of dominance relates to the notion of sharpness (Gneiting et al., 2007; Kuleshov et al., 2018), which links the accuracy of a calibrated classifier to how fine-grained the calibration is within the sample-space. In particular, if h dominates h', it can be readily shown that h is sharper than h'. However, existing works have not studied the effect of the sharpness of a classifier on its performance on screening tasks.

A Near-Optimal Screening Algorithm with Calibration Data. Until here, we have assumed that a perfectly calibrated classifier h with finite range, as well as the size of each bin ρ_b are given. However, using finite amounts of calibration data, we can only hope to find *approximately* calibrated classifiers. Next, we will develop an algorithm that, rather than training an approximately calibrated classifier from scratch, it *approximately* calibrates a given classifier f, *e.g.*, a deep neural network, using a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i^c, y_i^c)\}_{i\in[n]}$, which are independently sampled from $P_{X,Y}$[8]. In doing so, the algorithm will find the optimal sample-space partition and decision rule that minimize the expected size of the provided shortlists among those ensuring that an empirical lower bound on the expected number of qualified candidates is greater than k.

From now on, we will assume that the given classifier f is nonatomic[9] and satisfies a natural monotonicity property[10] with respect to the data distribution $P_{X,Y}$, a significantly weaker assumption than calibration.

Definition 4.4. *A classifier f is monotone with respect to a data distribution $P_{X,Y}$ if, for any $a, b \in \text{Range}(f)$ such that $a < b$, it holds that*

$$\Pr\{Y = 1 \mid f(X) = a\} \leq \Pr\{Y = 1 \mid f(X) = b\}.$$

Under this assumption[11], we can first show that the solution (if there is one) to the constrained minimization problem in Eq. 2 over Π_f is a threshold decision rule $\pi_{f,t_f^\star}$ with some threshold $t_f^\star \in [0, 1]$ that takes shortlisting decisions as

$$s_i = \begin{cases} 1 & \text{if } f(x_i) \geq t_f^\star, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

More specifically, we have the following theorem.

Theorem 4.5. *Let f be a monotone classifier with respect to $P_{X,Y}$ and the distribution $P_{f(X)}$ induced by $P_{X,Y}$ is nonatomic. Then, for any $\pi_f \in \Pi_f$, there always exists a threshold decision rule $\pi_{f,t} \in \Pi_f$ with some threshold $t \in [0, 1]$ such that*

$$\mathbb{E}_{\boldsymbol{X}\sim P, \boldsymbol{S}\sim\pi_f, \boldsymbol{S}'\sim\pi_{f,t}} \left[\sum_{i\in[m]} (S_i - S_i') \right] \geq 0$$

and

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_f, \boldsymbol{S}'\sim\pi_{f,t}} \left[\sum_{i\in[m]} Y_i(S_i - S_i') \right] \leq 0.$$

The theorem directly implies there exists a threshold decision policy $\pi_{f,t_f^\star}$ that is optimal in the constraint minimization problem defined in Eq. 2 (if there is a solution).

[8] Superscript c is used to differentiate between candidates in the calibration set and candidates in the pool at test time.

[9] We can add arbitrarily small noise to break atoms.

[10] The monotonicity property we consider is different from that considered in the literature on monotonic classification (Cano et al., 2019).

[11] There is empirical evidence that well-performing classifiers learned from data are (approximately) monotone (Guo et al., 2017; Kuleshov et al., 2018; Gupta et al., 2020). In this context, we would like to emphasize that the monotone property is only necessary to prove that the expected size of the shortlists provided by our algorithm is near-optimal. However, the distribution-free guarantees on the expected number of qualified shortlisted candidates holds even for non-monotone classifiers.

As a result, we focus our attention on finding a near-optimal threshold decision policy. We first notice that each threshold decision policy $\pi_{f,t}$ induces a sample space partition of $\mathcal{X}$ into two bins $\mathcal{X}_{t,1} = \{x \in \mathcal{X} \,|\, f(x) \geq t\}$ and $\mathcal{X}_{t,2} = \{x \in \mathcal{X} \,|\, f(x) < t\}$. Thus, it is sufficient to analyze the calibration errors of the family of approximately calibrated classifiers h_t with the two bins. In Appendix A.3, we show that using the calibration errors of calibrated classifiers that partition the sample-space into more bins will only worsen the performance guarantees of threshold policies. At this point, one may think of applying conventional distribution-free calibration methods to bound the calibration errors of each classifier h_t with high probability. However, prior distribution-free calibration methods can only bound calibration errors on the mean values μ_b for each bin b, but not the bin sizes ρ_b. To make things worse, to find the threshold value $\hat{t}_f$ with optimal guarantees, we need to bound the calibration errors of all classifiers h_t, simultaneously with high probability. Unfortunately, since $t \in [0, 1]$, there are infinitely many of them and we cannot naively apply a union bound on the bounds derived separately for each classifier.

To overcome the above issues, we will now leverage the Dvoretzky–Kiefer–Wolfowitz–Massart (DKWM) inequality (Dvoretzky et al., 1956; Massart, 1990), which bounds how close an empirical cumulative distribution function (CDF) is to the cumulative distribution function of the distribution from which the empirical samples are sampled. More specifically, let $\delta_{t,1} := \mathbb{E}_{(X,Y)\sim P}[\mathbb{I}\{f(X) \geq t\}\, Y]$ and

$$\hat{\delta}_{t,1} = \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{f(x_i^c) \geq t\}\, y_i^c$$

be an empirical estimator of $\delta_{t,1}$ using samples from the calibration set $\mathcal{D}_{\text{cal}}$. Then, we can use the DKWM inequality to bound the calibration errors $|\hat{\delta}_{t,1} - \delta_{t,1}|$ across all approximately calibrated classifiers h_t with high probability:

Proposition 4.6. *For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ (in f and $\mathcal{D}_{\text{cal}}$), it holds that*

$$\left|\delta_{t,1} - \hat{\delta}_{t,1}\right| \leq \sqrt{\ln(2/\alpha)/(2n)} := \epsilon(\alpha, n)$$

simultaneously for all $t \in [0, 1]$.

In Appendix B, we show that based on the above error guarantees, we can build a regression model that achieves average calibration in regression (Gneiting et al., 2007; Kuleshov et al., 2018), if we regard the binary classification problem as a regression problem. Further, building on the above proposition, we can derive an empirical lower bound on the expected number of qualified candidates in the shortlists provided by all the threshold decision rules $\pi_{f,t}$.

Corollary 4.7. *For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$ (in f and $\mathcal{D}_{\text{cal}}$), it holds that*

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]} Y_i S_i\right] \leq m\left(\hat{\delta}_{t,1} + \epsilon(\alpha, n)\right)$$

simultaneously for all $t \in [0, 1]$.

Now, to find the decision threshold rule $\pi_{f,\hat{t}_f}$ that provides, in expectation, the smallest shortlists of candidates subject to a constraint on the lower bound on the expected number of qualified candidates in the provided shortlists, *i.e.*,

$$\hat{t}_f = \operatorname*{argmin}_{t\in\mathcal{T}} \mathbb{E}_{\boldsymbol{X}\sim P, \boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]} S_i\right] \tag{6}$$

where $\mathcal{T} = \{t \in [0, 1] \,|\, m\left(\hat{\delta}_{t,1} - \epsilon(\alpha, n)\right) \geq k\}$, we resort to the following theorem.

Theorem 4.8. *The threshold value*

$$\hat{t}_f = \sup\left\{t \in [0, 1] \,\middle|\, \hat{\delta}_{t,1} \geq k/m + \epsilon(\alpha, n)\right\}. \tag{7}$$

is a solution (if there is one) to the constrained minimization problem defined in Eq. 6.

Algorithm 1 Calibrated Subset Selection (CSS)

input: k, m, $\mathcal{D}_{\text{cal}}$, f, α, $\boldsymbol{x}$
initialize: $\boldsymbol{s} = \boldsymbol{0}$
$\hat{t}_f = \sup\left\{t \in [0, 1] \,\middle|\, \hat{\delta}_{t,1} \geq k/m + \epsilon(\alpha, n)\right\}$
for $i = 1, \ldots, m$ **do**
  if $f(x_i) \geq \hat{t}_f$ **then**
    $s_i = 1$
  end if
end for
return $\boldsymbol{s}$

Algorithm 1 summarizes our resulting CSS algorithm, whose complexity is $\mathcal{O}(n \log(n))$.

Finally, we can show that the expected size of the shortlists provided by CSS is near-optimal and we can also derive a lower bound on the worst-case size of the provided shortlists. More specifically, the following propositions show that the difference in expected number of qualified candidates between the shortlists provided by $\pi_{f,\hat{t}_f}$ and $\pi_{f,t_f^\star}$ decreases at a rate $1/\sqrt{n}$ and the worst-case size is on the order of $\sqrt{k}$.

Proposition 4.9. *Let f be a monotone classifier with respect to $P_{X,Y}$ and assume that the distribution $P_{f(X)}$ induced by $P_{X,Y}$ is nonatomic. Then, for any $\alpha \in (0, 1)$, with*

probability at least $1-\alpha$*, it holds that*

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,t_f^\star},\boldsymbol{S}'\sim\pi_{f,\hat{t}_f}}\left[\sum_{i\in[m]}(S_i'-S_i)Y_i\right] \leq \frac{m}{n}+m\sqrt{2\ln(2/\alpha)/n}.$$

Proposition 4.10. *For any* $\alpha_1 \in (0,1), \alpha_2 \in (e^{-\frac{9}{4}k},1)$ *such that* $\alpha_1+\alpha_2<1$*, the CSS algorithm with parameter* $\alpha=\alpha_1$ *guarantees that, with probability at least* $1-\alpha_1-\alpha_2$ *(in* f*,* $\mathcal{D}_{\text{cal}}$*, and* $\boldsymbol{X},\boldsymbol{Y}$*),*

$$\sum_{i\in[m]}Y_iS_i \geq k-\frac{1}{3}\ln(1/\alpha_2)-\frac{1}{3}\sqrt{\ln^2(1/\alpha_2)+18k\ln(1/\alpha_2)}.$$

Finally, note that we can use the above worse-case guarantee to ensure that the worst-case size of the shortlists provided by CSS is greater than a target t_{worst} by setting k to be slightly larger than k_{worst}

$$k = k_{\text{worst}}+\frac{1}{3}\ln(1/\alpha_2)-\frac{1}{3}\sqrt{\ln^2(1/\alpha_2)+18k\ln(1/\alpha_2)}.$$

By doing so, CSS will satisfy the constraints defined in Eq. 1 with high probability with respect to the test pool of candidates. However, CSS might not be optimal in terms of expected or worst-case shortlist size among those satisfying the above constrains. How to design screening algorithms that are optimal in terms of expected or worst-case shortlist size while satisfying that each individual shortlist contains enough qualified candidates with high probability is an interesting problem to explore in future work.

5. Increasing the Diversity of Screening

Our theoretical results have shown that CSS is robust to the accuracy of the classifier and the amount of calibration data. However, CSS does not account for the potential differences in accuracy or in the amount of calibration data across demographic groups. As a result, qualified candidates in minority groups may be unfairly underrepresented in the shortlists provided by CSS.

To tackle the above problem and increase the diversity of the shortlists, we design a variant of CSS, which we name CSS (Diversity), motivated by the Rooney rule (Collins, 2007) and multicalibration (Hébert-Johnson et al., 2018) (refer to Appendix A.2). CSS (Diversity) quantifies the uncertainty in the estimation of the number of qualified candidates separately for each demographic group, so that we have distribution-free guarantees for each demographic group. This means that CSS (Diversity) will include more candidates in the shortlist for groups with higher uncertainty to ensure a sufficient number of qualified candidates from each group. Effectively, CSS (Diversity) shifts the cost of uncertainty from minority candidates to the decision maker, since it creates a potentially longer shortlist. This provides an economic incentive for the decision maker to both build classifiers that perform well across demographic groups and collect more calibration data about minority groups.

6. Empirical Evaluation Using Survey Data

In this section, we compare CSS against several competitive baselines on multiple instances of a simulated screening process created using US Census survey data.

Experiment Setup. We create a simulated screening process using a dataset comprised of employment information for ∼3.2 million individuals from the US Census (Ding et al., 2021). For each individual, we have sixteen features $x \in \mathbb{R}^{16}$ (*e.g.*, education, marital status) and a label $y \in \{0,1\}$ that indicates whether the individual is employed ($y=1$) or unemployed ($y=0$). Race is among the features, which we use as the protected attribute as suggested by Ding et al. (2021). We treat white as the majority group g_{maj} and all other races as the minority group g_{min}. To ensure that there is a limited number of qualified candidates (∼20%) in each of the simulated screening processes, we randomly downsample the dataset[12]. After downsampling, the dataset contains ∼2.2 million individuals.

For the experiments, we randomly split the dataset into two equally-sized and disjoint subsets. We use one of the subsets to create a training set of 10,000 individuals, as well as calibration sets with varying sizes n. We use the other subset to simulate pools of candidates for testing. To get a screening classifier, we train a logistic regression f_{LR} on the training set to predict the probability $f_{\text{LR}}(x)$ that a candidate is qualified. Here, we vary the accuracy of the classifier by replacing, with probability r_{noise}, each of its predictions with some noise $\beta \sim \text{Beta}(1,4)$, *i.e.*, $f=\gamma\beta+(1-\gamma)f_{\text{LR}}$, where $\gamma \sim \text{Bernoulli}(r_{\text{noise}})$ and r_{noise} is the *classifier noise ratio*.

In each simulated screening process, we set the size of the test pool of candidates to $m=100$, the desired expected number of qualified candidates to $k=5$, and the success probability to $1-\alpha=0.9$. For the diversity experiments, we set the desired expected number of qualified candidates k_{maj} and k_{min} so that the equal opportunity constraint,*i.e.*, $k_{\text{maj}}/(m\mathbb{E}_{(X,Y)}[Y\mathbb{I}\{X\in g_{\text{maj}}\}]) = k_{\text{min}}/(m\mathbb{E}_{(X,Y}[Y\mathbb{I}\{X\in g_{\text{min}}\}])$ is satisfied (Hardt et al., 2016) subject to $k_{\text{maj}}+k_{\text{min}}=5$.

Methods. In our experiments, we compare CSS with several baselines. Since no prior screening algorithms with distribution-free guarantees exist, we introduce a simple screening algorithm based on uniform mass binning (Zadrozny & Elkan, 2001) (UBM B Bins), which

[12]For the diversity experiments, we downsample the majority and minority groups independently.

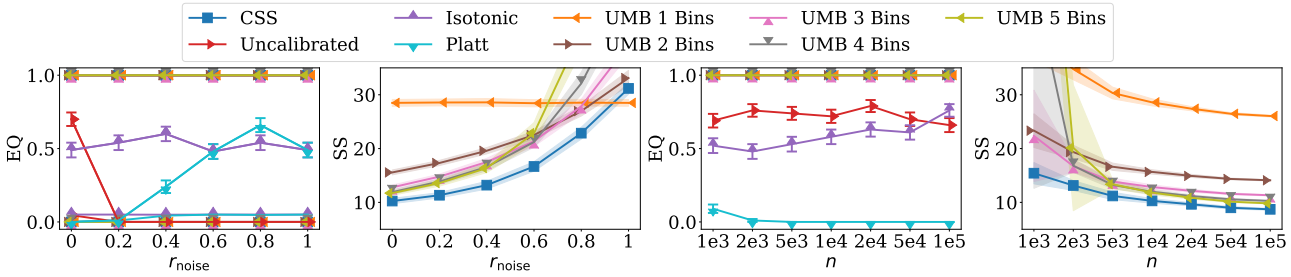

Figure 1. Analysis of CSS and baselines when varying the classifier noise ratio r_{noise} (*i.e.*, accuracy) and calibration set sizes n. The first and third plots show the empirical probability that each algorithm provides enough qualified candidates (EQ) from 100 runs with standard error bars (higher is better). The second and fourth plots show the average, among 100 runs with one standard deviation as shaded regions, expected size (SS) of the shortlists (lower is better). We do not plot SS for algorithms that fail the quality requirement in terms of EQ. In the left two plots, the size of the calibration set is $n = 10,000$. In the right two plots, the classifier noise ratio is $r_{\text{noise}} = 0$.

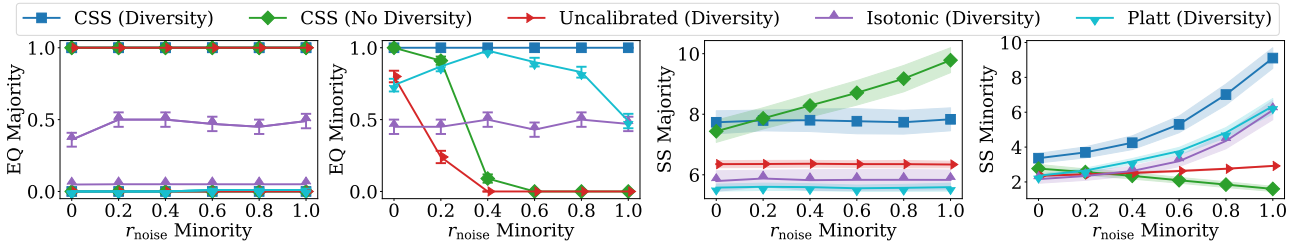

Figure 2. Analysis of diversity guarantees of CSS (Diversity) and several baselines when varying the classifier noise ratio r_{noise} (*i.e.*, accuracy) for individuals in the minority group. For individuals in the majority group, the classifier noise ratio is $r_{\text{noise}} = 0$. The first and third plots focus on the majority group and the second and fourth plots focus on the minority group.

also enjoys distribution-free guarantees on the expected number of qualified candidates. The algorithm bounds the calibration error of a classifier that is calibrated on B bins of roughly equal size and selects the candidates from top-scored bins to low-scored bins (and possibly at random from the last bin it selects candidates from) until a lower bound on the expected number of qualified candidates is no smaller than k. Refer to Appendix A.3 for more details about this baseline algorithm.

We also compare CSS with three other baselines that do not provide distribution-free guarantees. The first is called *Uncalibrated*, and it applies the optimal decision rule for omniscient classifiers as if f were the omniscient classifier defined in Eq. 3. The second is called *Platt*, since it first calibrates f using Platt scaling (Platt et al., 1999) and then proceeds like Uncalibrated with the calibrated classifier. The third is called *Isotonic*. It treats the classifier f as a regression model from x to y, and then employs Isotonic regression calibration (Kuleshov et al., 2018) to produce a calibrated regression model h that estimates $h(t) \approx \mathbb{E}[(Y\mathbb{I}\{f(x) \geq t\}]$. It then selects the largest threshold t such that $mh(t) \geq k$ for shortlisting.

Metrics. To compare the screening algorithms, we run the experiments 100 times for each algorithm and setting. For each run, we estimate whether each algorithm provides shortlists that contain a large enough expected number of qualified candidates $\text{EQ} = \mathbb{I}\left\{\hat{\mathbb{E}}\left[\sum_{i\in[m]} Y_i S_i\right] \geq k\right\}$, and its expected shortlist size $\text{SS} = \hat{\mathbb{E}}\left[\sum_{i\in[m]} S_i\right]$, using $1,000$ independent pools of candidates sampled at random from the test set. We then compare the algorithms in terms of the percentage of times, along with standard errors, they provide enough qualified candidates. We also compare the average shortlist size, along with standard deviations.

How does the accuracy of the classifier affect different screening algorithms? The left two plots in Figure 1 show how CSS compares to the baselines for classifiers of varying accuracy (*i.e.*, varying r_{noise}). The results show that the shortlists provided by the algorithms with distribution-free guarantees (i.e., CSS and UMB) do contain enough qualified candidates, while the others fail. Moreover, CSS and UMB are robust to the accuracy of the classifier they use, as they compensate for decreased accuracy with longer shortlists to maintain their guarantees. In addition, we find that the shortlists provided by these algorithms contain enough qualified candidates more frequently than suggested by the worst-case theoretical guarantee of $1 - \alpha = 0.9$. Compared to all UMB variants, CSS provides smaller shortlists except for the pure-noise case. Note that this exception does not

violate the optimality of CSS among deterministic threshold policies using the empirical lower bound as shown in Theorem 5, since UMB algorithms are allowed to randomly select candidates in the last bin they select candidates from, as discussed previously and in Appendix A.3 in detail.

What is the effect of different amounts of calibration data on the screening algorithms? The right two plots in Figure 1 show how the screening algorithms perform for increasing amounts of calibration data n. We see that CSS and UMB are robust to the amount of calibration data, and can effectively account for less data by increasing the shortlist size to maintain their guarantees. In terms of shortlist size, CSS is more effective over the whole range of n, since it provides smaller shortlists than all of the UMB variants.

How does the accuracy of the classifier affect different groups of candidates? Figure 2 shows how CSS (Diversity) and the baselines perform on the majority and minority groups as we decrease the accuracy of the classifier for individuals in the minority group (*i.e.*, we increase r_{noise} only for individuals in the minority group). Here, CSS (No Diversity) refers to naively applying CSS on the entire pool of candidates, without diversity requirements. We allow all other algorithms to select candidates from the majority group and the minority group separately. The results show that, as the accuracy of the classifier for individuals in the minority group decreases, the shortlists provided by CSS (No Diversity) contain more and more (fewer and fewer) candidates from the majority (minority) group. This suggests that we should explicitly account for diversity in the screening process, especially when the accuracy of the classifier differs across groups. While Uncalibrated, Platt and Isotonic select candidates across groups separately, the shortlists they provide contain fewer qualified candidates from the minority group as the accuracy worsens (Uncalibrated) or do not contain enough qualified candidates from both groups (Platt, Isotonic). In contrast, CSS (Diversity) adapts to the loss in accuracy of the classifier for individuals in the minority group and the shortlists it provides contain enough qualified candidates from both groups. We found similar results also when we varied the number of individuals from each group in the calibration data, instead of the classifier accuracy.

7. Conclusions

In this work, we initiated the development of screening algorithms that provide distribution-free guarantees on the quality of the shortlisted candidates. In particular, we proposed the CSS algorithm, which can be applied to any screening classifier. We show that for any amount of calibration data, CSS selects near-optimal shortlists of candidates, in terms of the expected shortlist size, while ensuring that the expected number of qualified candidates is above a desired target with high probability. Moreover, we showed how the CSS (Diversity) variant, which shortlists different groups of candidates separately, can ensure diversity of the shortlists with similar guarantees. Both the theoretical analysis and the empirical evaluation confirm that CSS and CSS (Diversity) are robust to the accuracy of the given classifier and the amount of calibration data.

Our work opens up many interesting avenues for future work. For example, we have assumed that candidates do not present themselves strategically to the screening algorithms and that the data distribution at calibration and test time is the same. It would thus be interesting to develop screening algorithms that are robust to strategic behaviors (Tsirtsis & Gomez Rodriguez, 2020) and distribution shifts (Podkopaev & Ramdas, 2021). Moreover, it is crucial to fully understand how screening algorithms can most effectively and most fairly augment (biased) human decision making in real selection processes.

Acknowledgements

We thank Eleni Straitouri and Nastaran Okati for discussions and feedback in the initial stage of this work. Gomez-Rodriguez acknowledges the support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719). Wang and Joachims acknowledge the support from NSF Awards IIS-1901168 and IIS-2008139. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Bakker, M. A., Tu, D. P., Gummadi, K. P., Pentland, A. S., Varshney, K. R., and Weller, A. Beyond reasonable doubt: Improving fairness in budget-constrained decision making using confidence thresholds. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 346–356, 2021.

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Barber, R. F. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2): 3487–3524, 2020.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Bendick Jr, M. and Nunes, A. P. Developing the research

basis for controlling bias in hiring. *Journal of Social Issues*, 68:238–262, 2013.

Bendick Jr, M., Jackson, C. W., and Romero, J. H. Employment discrimination against older workers: An experimental study of hiring practices. *Journal of Aging & Social Policy*, 8(4):25–46, 1997.

Bertrand, M. and Mullainathan, S. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Bogen, M. and Rieke, A. Help wanted: An examination of hiring algorithms, equity, and bias. 2018.

Brier, G. W. et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Cai, W., Gaebler, J., Garg, N., and Goel, S. Fair allocation through selective information acquisition. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 22–28, 2020.

Cano, J.-R., Gutiérrez, P. A., Krawczyk, B., Woźniak, M., and García, S. Monotonic classification: An overview on algorithms, performance measures and data sets. *Neurocomputing*, 341:168–182, 2019.

Chamorro-Premuzic, T. and Akhtar, R. Should companies use al to assess job candidates? *Harvard Business Review*, 17, 2019.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chzhen, E., Denis, C., Hebiri, M., and Lorieul, T. Set-valued classification–overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

Collins, B. W. Tackling unconscious bias in hiring practices: The plight of the rooney rule. *NYUL Rev.*, 82:870, 2007.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In *ACM conference on recommender systems*, pp. 191–198, 2016.

Cowgill, B. Bias and productivity in humans and algorithms: Theory and evidence from resume screening. *Columbia Business School, Columbia University*, 29, 2018.

Dawid, A. P. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In *Advances on Neural Information Processing Systems*, 2021.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Garr, S. S. and Jackson, C. Diversity & inclusion technology: The rise of a transformative market. *Red Thread Research and Mercer*, 2019.

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *ACM International Conference on Knowledge Discovery and Data Mining*, pp. 2221–2231, 2019.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Gupta, C. and Ramdas, A. K. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, 2021.

Gupta, C., Podkopaev, A., and Ramdas, A. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, volume 29, pp. 3315–3323, 2016.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948, 2018.

Hron, J., Krauth, K., Jordan, M., and Kilbertus, N. On component interactions in two-stage recommender systems. *Advances in Neural Information Processing Systems*, 34, 2021.

Johnson, S. K., Hekman, D. R., and Chan, E. T. If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, 26(04), 2016.

Jung, C., Lee, C., Pai, M., Roth, A., and Vohra, R. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pp. 2634–2678, 2021.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 277–287, 2020.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*, 2017.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804, 2018.

Liu, R., Rizzo, S., Whipple, S., Pal, N., Pineda, A. L., Lu, M., Arnieri, B., Lu, Y., Capra, W., Copping, R., et al. Evaluating eligibility criteria of oncology trials using real-world data and ai. *Nature*, 592(7855):629–633, 2021.

Ma, J., Zhao, Z., Yi, X., Yang, J., Chen, M., Tang, J., Hong, L., and Chi, E. H. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*, pp. 463–473, 2020.

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, 2017.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In *Conference on Uncertainty in Artificial Intelligence*, 2021.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Conference on Fairness, Accountability, and Transparency*, pp. 469–481, 2020.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, 2019.

Sahoo, R., Zhao, S., Chen, A., and Ermon, S. Reliable decisions with threshold calibration. In *Advances in Neural Information Processing Systems*, 2021.

Sánchez-Monedero, J., Dencik, L., and Edwards, L. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Conference on fairness, accountability, and transparency*, pp. 458–468, 2020.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. Provably improving expert predictions with prediction sets. *arXiv preprint arXiv:2201.12006*, 2022.

Tambe, P., Cappelli, P., and Yakubovich, V. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review*, 61(4): 15–42, 2019.

Tsirtsis, S. and Gomez Rodriguez, M. Decisions, counterfactual explanations and strategic behavior. In *Advances in Neural Information Processing Systems*, 2020.

Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. In *International Conference on Machine Learning*, 1999.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Wang, L. and Joachims, T. Fairness in the first stage of two-stage recommender systems. *arXiv preprint arXiv:2205.15436*, 2022.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, volume 1, pp. 609–616, 2001.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pp. 11387–11397, 2020.

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

Algorithm 2 CSS (Dynamic Pool Size)

input: $k, \mathbb{E}[M], \mathcal{D}_{\text{cal}}, f, \alpha, \boldsymbol{x}$
initialize: $\boldsymbol{s} = \emptyset$
$\hat{t}_f = \sup \left\{ t \in [0,1] \,\middle|\, \hat{\delta}_{t,1} \geq k/\mathbb{E}[M] + \epsilon(\alpha, n) \right\}$
for $x \in \boldsymbol{x}$ **do**
  if $f(x) \geq \hat{t}_f$ **then**
    add x to $\boldsymbol{s}$
  end if
end for
return $\boldsymbol{s}$

A. Additional Algorithms

A.1. Calibrated Screening Algorithm under the Dynamic Pool Size Setting

In reality, the size of the pool of candidates might be different across times. Here, we derive an algorithm CSS (Dynamic Pool Size) illustrated in Algorithm 2 that allows for dynamic size of the pool of applicants. we overload the use of notation $\boldsymbol{s}$ to denote the set of the selected candidates rather than a vector, for ease of presentation, especially for the screening algorithm to ensure diversity as we will discuss in the next subsection. CSS (Dynamic Pool Size) requires that the expected size $\mathbb{E}[M]$ is given, which can be estimated from past pools. We show that all our theoretical results naturally generalize to the dynamic pool size setting.

In the dynamic pool size setting, instead of assuming that the size of the pool of candidates m is constant, we assume the size is a random variable $M \sim P_M$ that follows some distribution P_M, with mean $\mathbb{E}[M] = \mathbb{E}_{M \sim P_M}[M]$. Thus, the data generation process for the pool of candidates becomes $M, \boldsymbol{X}, \boldsymbol{Y} \sim P_M \times P_{X,Y}^M$.

For the individual guarantees, Theorem 3.1 naturally holds for all $m \in \mathbb{R}^+$, and thus for the dynamic size setting; the impossibility results in Proposition 3.2 still hold in the dynamic pool size setting, since the constant size setting is a special case of the dynamic pool size setting.

So we focus on showing that the theoretical results in the marginal guarantees still hold here, by showing that there is a family of policies that are oblivious to the size of the pools while representative enough. We call this family of policies pool-independent policies.

Definition A.1 (Pool-Independent Policies)**.** *Given a classifier f, a policy $\pi_f \in \Pi_f$ is pool-independent if there is a mapping $p_{\pi_f} : \text{Range}(f) \to [0,1]$, such that for any candidate x_i in any pool $\boldsymbol{x}$ of any size m, the probability that the candidate is selected under π_f is $\Pr(S_i = 1) = p_{\pi_f}(f(x_i))$.*

This family of policies have the nice property that their expected number of qualified candidates in the shortlist and the expected shortlist size depend on P_M only through $\mathbb{E}[M]$. Formally speaking, for a pool-independent policy π_f and its selection probability mapping p_{π_f},

$$\mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y}) \sim P, \boldsymbol{S} \sim \pi_f} \left[\sum_{i \in [M]} S_i Y_i \right] = \mathbb{E}[M] \, \mathbb{E}_{(f(X),Y) \sim P_{f(X),Y}} \left[p_{\pi_f}(f(X)) Y \right]$$

and

$$\mathbb{E}_{(M,\boldsymbol{X}) \sim P, \boldsymbol{S} \sim \pi_f} \left[\sum_{i \in [M]} S_i \right] = \mathbb{E}[M] \, \mathbb{E}_{f(X) \sim P_{f(X)}} \left[p_{\pi_f}(f(X)) \right].$$

We show in the following proposition that the family of pool-independent policies are representative enough in the dynamic pool size setting, in a sense that for any given classifier f and any policy $\pi_f \in \Pi_f$, we can always find a pool-independent policy in Π_f that achieves the same expected number of qualified candidates in the shortlists and the same expected size of the shortlists as π_f.

Proposition A.2. *Given a classifier f, for any policy $\pi_f \in \Pi_f$, there exists a pool-independent policy $\pi'_f \in \Pi_f$ such that*

Algorithm 3 CSS (Diversity)

input: $\mathcal{G}$, $\left\{\mathbb{E}\left[M\right]_g\right\}_{g\in\mathcal{G}}$, $\{k_g\}_{g\in\mathcal{G}}$, $\{\mathcal{D}^g_{\text{cal}}\}_{g\in\mathcal{G}}$, f, α, $\{\boldsymbol{x}_g\}_{g\in\mathcal{G}}$
for $g \in \mathcal{G}$ **do**
  $\boldsymbol{s}_g =$ CSS (Dynamic Pool Size) $(k_g, \mathbb{E}\left[M\right]_g, \mathcal{D}^g_{\text{cal}}, f, \alpha/|\mathcal{G}|, \boldsymbol{x}_g)$
end for
return $\cup_{g\in\mathcal{G}} \boldsymbol{s}_g$

they have the same expected number of qualified candidates

$$\mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi}\left[\sum_{i\in[M]} Y_i S_i\right] = \mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi'}\left[\sum_{i\in[M]} Y_i S_i\right]$$

and the same expected shortlist size

$$\mathbb{E}_{(M,\boldsymbol{X})\sim P,\boldsymbol{S}\sim\pi}\left[\sum_{i\in[M]} S_i\right] = \mathbb{E}_{(M,\boldsymbol{X})\sim P,\boldsymbol{S}\sim\pi'}\left[\sum_{i\in[M]} S_i\right].$$

This implies that it is sufficient to find optimal/near-optimal solutions within this family of pool-independent policies to achieve global optimal/near-optimal guarantees. In fact, all the optimal/near-optimal policies in all the theorems, propositions, corollaries are all pool-independent policies. They all can be translated to the dynamic pool size setting by setting $m = \mathbb{E}\left[M\right]$ in the policy design.

A.2. Calibrated Subset Selection Algorithm to Ensure Diversity

CSS (Diversity) (Algorithm 3) selects calibrated shortlists of candidates for each demographic group independently. CSS (Diversity) requires inputs of the groups $\mathcal{G}$, the expected pool size of each group $\left\{\mathbb{E}\left[M\right]_g\right\}_{g\in\mathcal{G}}$, the target expected number of qualified candidates for each group $\{k_g\}_{g\in\mathcal{G}}$, the calibration data for each group $\{\mathcal{D}^g_{\text{cal}}\}_{g\in\mathcal{G}}$, and the pool of candidates from each group $\{\boldsymbol{x}_g\}_{g\in\mathcal{G}}$. Note that the groups can have overlap, *i.e.*, each candidate might belong to multiple groups. The guarantees on the expected number of qualified candidates from CSS (Dynamic Pool Size) directly apply to each group. The near-optimality guarantees of CSS (Dynamic Pool Size) only apply for every group when each candidate belongs to exactly one group, since otherwise it is possible that we select more-than-enough candidates from a particular group to satisfy the guarantees for the other groups.

A.3. Calibrated Screening Algorithms with Multiple Bins

We develop algorithms that can ensure distribution-free guarantees on the expected number of qualified candidates in the shortlists, when we partition the sample-space into multiple bins by the prediction scores from a given classifier f. For ease of notation and presentation of the proof, we illustrate the theory and the algorithm in the constant pool size setting. But they can be similarly adapted to the dynamic pool size setting as described in Appendix A.1.

Let $t_0 = 1, \ldots, t_B = 0$ be the thresholds on the prediction scores from f in decreasing order, which partition the sample-space $\mathcal{X}$ into B bins: $\{x \in \mathcal{X} \,|\, t_0 \leq f(x) \leq t_1\}$, $\{x \in \mathcal{X} \,|\, t_1 < f(x) \leq t_2\}$, $\ldots$, $\{x \in \mathcal{X} \,|\, t_{B-1} < f(x) \leq t_B\}$. Let $\mathcal{B} : \mathcal{X} \to [B]$ be the bin identity function, which maps a candidate x to the bin it belongs to $\mathcal{B}(x)$. Let $\delta_b = \mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[\mathbb{I}\left\{\mathcal{B}(X) = b\right\} Y\right]$, and $\hat{\delta}_b = \frac{1}{n}\sum_{i\in[n]} Y_i^c \mathbb{I}\left\{\mathcal{B}(X_i^c)\right\}$ be an empirical estimate of it from the calibration data $\mathcal{D}_{\text{cal}}$. There exist many concentration inequalities to bound the calibration errors. But they generally require sample-splitting, *i.e.*, to split the calibration data into two sets, one set for selecting the sample-space partition, and the other set for estimating the calibration error. A recent work (Gupta & Ramdas, 2021) on distribution-free calibration without sample-splitting does not apply here, since their proposed method can only bound the calibration errors of $\mu_b = \mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y_i \,|\, \mathcal{B}(X) = b\right]$, rather than δ_b. To avoid sample-splitting while still bounding the calibration errors of δ_b on sample-space partitions with thresholds $\{t_b\}_{b\in[B-1]}$ chosen in a data dependent way (*e.g.*, uniform mass binning (Zadrozny & Elkan, 2001)), we again use the DKWM inequality to bound the calibration error of δ_b on any bin b that corresponds to an interval on the scores predict by f, by the following proposition.

Algorithm 4 Calibrated Subset Selection with Multiple Bins

input: k, m, B, $\mathcal{B}$, $\left\{\hat{\delta}_b\right\}_{b\in[B]}$, $\epsilon(\alpha, n)$, $\boldsymbol{x}$
initialize: $\boldsymbol{s} = \emptyset$
$\hat{b} = \inf\left\{a \in [B] \,\middle|\, \sum_{b\in[a]}\left(\hat{\delta}_b - 2\epsilon(\alpha,n)\right) \geq k/m\right\}$
$\hat{\theta} = \frac{k/m - \sum_{b\in[\hat{b}-1]}\left(\hat{\delta}_b - 2\epsilon(\alpha,n)\right)}{\hat{\delta}_{\hat{b}} - 2\epsilon(\alpha,n)}$
for $x \in \boldsymbol{x}$ **do**
  if $\mathcal{B}(x) < \hat{b}$ **then**
    add x to $\boldsymbol{s}$
  else if $\mathcal{B}(x) = \hat{b}$ **then**
    add x to $\boldsymbol{s}$ with probability $\hat{\theta}$
  end if
end for
return $\boldsymbol{s}$

Proposition A.3. *For any $\alpha \in (0,1)$, with probability at least $1-\alpha$ (in f and $\mathcal{D}_{\text{cal}}$), it holds that for any bin $\{x \in \mathcal{X} \,|\, t_{b-1} \leq f(x) \leq t_b\}$ indexed by b and parameterized by $0 \leq t_{b-1} < t_b \leq 1$,*

$$\left|\delta_b - \hat{\delta}_b\right| \leq \sqrt{2\ln(2/\alpha)/n} = 2\epsilon(\alpha, n).$$

With this calibration error guarantee that holds for any bin, and motivated by the monotone property of f, we derive an empirical lower bound on the expected number of qualified candidates similarly as Corollary 4.7 using a family of threshold decision policies that select candidates from the top-scored bins to lower-scored bins, possibly at random for the candidates in the last bin it selects candidates from. Any policy $\pi_{f,a,\theta}$ given a classifier f in this family makes decisions based on the following decision rule

$$s_i = \begin{cases} 1 & \text{if } \mathcal{B}(x_i) < a, \\ \text{Bernoulli}(\theta) & \text{if } \mathcal{B}(x_i) = a, \\ 0 & \text{otherwise.} \end{cases}$$

An empirical lower bound on the expected number of qualified candidates in the shortlists can be derived as follows

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_{f,a,\theta}}\left[\sum_{i\in[m]} S_i Y_i\right] &= m\mathbb{E}_{(\mathcal{B}(X),Y)\sim P_{\mathcal{B}(X),Y}}\left[Y\left(\mathbb{I}\left\{\mathcal{B}(X) < a\right\} + \theta\mathbb{I}\left\{\mathcal{B}(X) = a\right\}\right)\right] \\
&= m\sum_{b\in[B]} \delta_b \left(\mathbb{I}\left\{b < a\right\} + \theta\mathbb{I}\left\{b = a\right\}\right) \\
&\geq m\sum_{b\in[B]} \left(\hat{\delta}_b - 2\epsilon(\alpha,n)\right)\left(\mathbb{I}\left\{b < a\right\} + \theta\mathbb{I}\left\{b = a\right\}\right).
\end{aligned}$$

We solve the optimization problem with this empirical lower bound similarly as in Eq. 7:

$$\hat{b}, \hat{\theta} = \operatorname*{argmin}_{b,\theta\in\Theta} \mathbb{E}_{\boldsymbol{X}\sim P, \boldsymbol{S}\sim\pi_{f,b,\theta}}\left[\sum_{i\in[m]} S_i\right] \tag{8}$$

where $\Theta = \{a \in [B], \theta \in [0,1) \,|\, m\sum_{b\in[B]}\left(\hat{\delta}_b - 2\epsilon(\alpha,n)\right)\left(\mathbb{I}\left\{b < a\right\} + \theta\mathbb{I}\left\{b = a\right\}\right) \geq k\}$.

And it turns out that there is a closed form solution

$$\hat{b} = \inf\left\{a \in [B] \,\middle|\, \sum_{b\in[a]}\left(\hat{\delta}_b - 2\epsilon(\alpha,n)\right) \geq k/m\right\}$$

and

$$\hat{\theta} = \frac{k/m - \sum_{b\in[\hat{b}-1]} \left(\hat{\delta}_b - 2\epsilon(\alpha, n)\right)}{\hat{\delta}_{\hat{b}} - 2\epsilon(\alpha, n)}.$$

We summarize the algorithm using this decision policy in Algorithm 4. It guarantees that the expected number of qualified candidates is greater than k with probability $1-\alpha$ as shown in the empirical lower bound. And, it also satisfies the near-optimality guarantees on the expected size of the shortlist among this extended family of threshold decision policies given a fixed sample-space partition similarly as Calibrated Subset Selection does in Proposition 4.9. We omit the proof since it can be simply adapted from the proof of Proposition 4.9.

The disadvantage of Algorithm 4 is that it requires a given sample-space partition and finds near-optimal policy only among the policies using this fixed sample-space partition, while CSS directly optimizes the sample-space partition and the policy jointly. To also optimize over sample-space partitions, *i.e.*, the number of bins and the thresholds, is computationally intractable and hard to identify which one is actually optimal. CSS solves this problem by considering a smaller policy space (no random selection from the last bin) and the sample-space partition space (only two bins), so that it is easy to find the optimal sample-space partition and the optimal policy simultaneously. We now show that among deterministic threshold decision rules, sample-space partitions with two bins are optimal.

Calibration on More Bins Worsens the Performance of Threshold Policies. When we consider deterministic threshold decision rules, Algorithm 4 becomes selecting the candidates in bin $\hat{b}$ with probability 1 instead of with probability $\hat{\theta}$, to ensure distribution-free guarantees on the expected number of qualified candidates in the shortlists. This is equivalent to finding the largest threshold $\hat{t}$ among $\{t_b\}_{b\in[B]}$ to ensure enough qualified candidates

$$\hat{t} = \sup\left\{ t \in \{t_b\}_{b\in[B]} \;\middle|\; \sum_{b\in[B]} \mathbb{I}\{t_b \geq t\}\,(\hat{\delta}_b - 2\epsilon(\delta, n)) \geq k \right\}.$$

This is essentially solving the optimization problem using the empirical lower bound on the expected number of qualified candidates

$$\sum_{b\in[B]} \mathbb{I}\{t_b \geq t\}\,(\hat{\delta}_b - 2\epsilon(\delta, n)),$$

over a subset of thresholds in $[0, 1]$. This empirical lower bound is looser compared to the bound used in CSS, which also optimizes over the entire threshold space $[0, 1]$, as shown in the following

$$\sum_{b\in[B]} \mathbb{I}\{t_b \geq t\}\,(\hat{\delta}_b - 2\epsilon(\delta, n)) = \delta_{t_b,1} - 2\sum_{b\in[B]} \mathbb{I}\{t_b \geq t\}\,\epsilon(\alpha, n) > \delta_{t_b,1} - \epsilon(\alpha, n).$$

So it is clear that CSS which uses a tighter bound over a larger space of thresholds achieves shorter shortlists in expectation.

UMB B Bins Algorithms. In the empirical evaluation, the UBM B Bins algorithms correspond to selecting the thresholds such that the calibration data $\mathcal{D}_{\text{cal}}$ are partitioned evenly into B bins, and running Algorithm 4 with the sample-space partition characterized by these thresholds.

B. Calibration Error Bounds Imply Distribution-Free Average Calibration in Regression

If we regard the given classifier f as a regression model from x to y, then the calibration error bounds in Proposition 4.6 directly imply a classifier that (approximately) achieves average calibration for any interval (Gneiting et al., 2007; Kuleshov et al., 2018). Translating average calibration in regression to our setting, where the range of the regression model has only two values $\{0, 1\}$, a classifier h is average-calibrated if for any $0 \leq t \leq 1$

$$\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\{h(X) \geq t\}\right] = 1 - t.$$

From the calibration error guarantees in Proposition 4.6, we know that for any $\delta \in (0, 1)$, with probability $1-\delta$, for any $t \in [0, 1]$,

$$\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\{f(X) \geq t\}\right] - \hat{\delta}_{t,1}\right| \leq \epsilon(\alpha, n).$$

Let $g(t) = \hat{\delta}_{t,1}$, which is a monotonically decreasing function. Let $t = g^{-1}(1-t')$ in the above inequalities, we get for any $t' \in [0,1]$,

$$\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{f(X) \geq g^{-1}(1-t')\right\}\right] - (1-t')\right| \leq \epsilon(\alpha, n).$$

Since g is monotone, we can get that for any $t \in [0,1]$,

$$\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{(g\circ f)(X) \geq 1-t\right\}\right] - (1-t)\right| \leq \epsilon(\alpha, n).$$

Let $\hat{h} = g \circ f$, $\hat{h}$ is approximately calibrated in a sense that with probability $1-\alpha$, for any $t \in [0,1]$,

$$1 - t - \epsilon(\alpha,\delta) \leq \mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{\hat{h}(X) \geq t\right\}\right] \leq 1 - t + \epsilon(\alpha,\delta).$$

C. Proofs

C.1. Proof of Theorem 3.1

We first show that for any $\boldsymbol{x} \in \mathcal{X}^m$, the constraint in the minimization problem in Eq. 1 is satisfied with equality under $\pi^\star_{f^\star}$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{S}\sim\pi^\star_{f^\star},\boldsymbol{Y}\sim P}\left[\sum_{i\in[m]} S_iY_i\right] &= \mathbb{E}_{\{Y_i\}_{i\in[m]}\sim\prod_{i\in[m]}P_{Y|X=x_i},\boldsymbol{S}\sim\pi(\{f(x_i)\}_{i\in[m]})}\left[\sum_{i\in[m]} S_iY_i\right] \\
&= \sum_{i\in[m]} \Pr\left(Y=1 \mid X=x_i\right)\left(\mathbb{I}\left\{f(x_i) > t^\star\right\} + \mathbb{I}\left\{f(x_i) = t^\star\right\}\theta^\star\right) \\
&= \sum_{i\in[m]} f(x_i)\left(\mathbb{I}\left\{f(x_i) > t^\star\right\} + \mathbb{I}\left\{f(x_i) = t^\star\right\}\theta^\star\right) \\
&= k.
\end{aligned}$$

The last equality is by the definition of $t^\star$ and $\theta^\star$. Next, we will show that it is an optimal solution for any $\boldsymbol{x} \in \mathcal{X}^m$ by contradiction.

The objective in Eq. 1 using $\pi^\star_{f^\star}$ is

$$\mathbb{E}_{\boldsymbol{S}\sim\pi^\star_{f^\star}}\left[\sum_{i\in[m]} S_i\right] = \sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) > t^\star\right\} + \theta^\star\mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) = t^\star\right\}.$$

Suppose a policy π' achieves smaller objective than $\pi^\star_{f^\star}$ for some $\boldsymbol{x} \in \mathcal{X}^m$, *i.e.*,

$$\mathbb{E}_{\boldsymbol{S}\sim\pi'(\{f(x_i)\}_{i\in[m]})}\left[\sum_{i\in[m]} S_i\right] := R_{\pi'}(\boldsymbol{x}) < \sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) > t^\star\right\} + \theta^\star\mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) = t^\star\right\}.$$

We will show that the constraint in the optimization problem in Eq. 1 for $\boldsymbol{x}$ is not satisfied using π'. Let

$$t' := \sup\left\{t \in [0,1] : \sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) \geq t\right\} \geq R_{\pi'}(\boldsymbol{x})\right\}$$

and

$$\theta' := \frac{R_{\pi'}(\boldsymbol{x}) - \sum_{i\in[m]}\mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) > t'\right\}}{\sum_{i\in[m]}\mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) = t'\right\}}.$$

Note that $t^\star \leq t'$ since

$$\begin{aligned}
&\sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) \geq t^\star\right\} \\
&\qquad\geq \sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) > t^\star\right\} + \theta^\star\mathbb{I}\left\{\Pr\left(Y=1 \mid X=x_i\right) = t^\star\right\} > R_{\pi'}(\boldsymbol{x}).
\end{aligned}$$

Now, we can show that the constraint is not satisfied using π'. The left hand side of the constraint using π' can be bounded as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{S}\sim\pi'}\left[\sum_{i\in[m]} Y_i S_i\right] &= \mathbb{E}_{\{Y_i\}_{i\in[m]}\sim\prod_{i\in[m]} P_{Y|X=x_i},\boldsymbol{S}\sim\pi'}\left[\sum_{i\in[m]} Y_i S_i\right] \\ &= \mathbb{E}_{\boldsymbol{S}\sim\pi'}\left[\sum_{i\in[m]} \Pr(Y=1\mid X=x_i) S_i\right] \\ &\le \sum_{i\in[m]} \Pr(Y=1\mid X=x_i)\left(\mathbb{I}\left\{\Pr(Y=1\mid X=x_i)>t'\right\}+\mathbb{I}\left\{\Pr(Y=1\mid X=x_i)=t'\right\}\theta'\right),\end{aligned}$$

where the last inequality is by the definition of $R_{\pi'}(\boldsymbol{x})$, t' and θ'. When $t^\star < t'$,

$$\begin{aligned}&\sum_{i\in[m]} \Pr\left(Y=1\mid X=x_i\right)\left(\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t'\right\}+\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)=t'\right\}\theta'\right) \\ &\qquad < \sum_{i\in[m]} \Pr\left(Y=1\mid X=x_i\right)\left(\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t^\star\right\}+\mathbb{I}\left\{\Pr[Y=1\mid X=x_i]=t^\star\right\}\theta^\star\right)=k.\end{aligned}$$

When $t^\star = t'$,

$$\theta' = \frac{R_{\pi'} - \sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t^\star\right\}}{\sum_{i\in[m]} \mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)=t^\star\right\}} < \theta^\star.$$

Thus

$$\begin{aligned}&\sum_{i\in[m]} \Pr\left(Y=1\mid X=x_i\right)\left(\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t'\right\}+\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)=t'\right\}\theta'\right) \\ &\quad = \sum_{i\in[m]} \Pr\left(Y=1\mid X=x_i\right)\left(\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t^\star\right\}+\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)=t^\star\right\}\theta'\right) \\ &\quad < \sum_{i\in[m]} \Pr\left(Y=1\mid X=x_i\right)\left(\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)>t^\star\right\}+\mathbb{I}\left\{\Pr\left(Y=1\mid X=x_i\right)=t^\star\right\}\theta^\star\right)=k.\end{aligned}$$

This shows that for any $\boldsymbol{x}\in\mathcal{X}^m$, for any policy π' that achieves smaller objective than $\pi^\star_{f^\star}$, π' does not satisfy the constraint, which concludes the proof.

C.2. Proof of Proposition 3.2

We construct two pools of candidates such that any policy will fail on at least one them.

The two pools of candidates are all as and all bs, $\boldsymbol{x}=\{a,a,\ldots,a\}$ and $\boldsymbol{x}'=\{b,b,\ldots,b\}$. And also note that in this kind of candidate distribution, the only perfectly calibrated predictor that is not the omniscient predictor is the predictor that predicts $h(x)=\Pr(Y=1)$ for all $x\in\mathcal{X}$, since there are only two possible candidate feature vectors. For any policy π_h that makes decisions S based purely on the quality scores by predicted by h, the distribution of $\boldsymbol{S}$ will be the same for the two pools of candidates, since they have the same predictions from h.

Thus the expectation on the size of the shortlist for the two pools of candidates are the same

$$\mathbb{E}_{\boldsymbol{S}\sim\pi_h\left(\{h(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i\right] = \mathbb{E}_{\boldsymbol{S}\sim\pi_h\left(\left\{h\left(x'_i\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i\right] := c_{\pi_h}.$$

On the other hand, for the optimal policy $\pi^\star_{f^\star}$, we know that

$$\mathbb{E}_{\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\{f^\star(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i\right] = k,$$

and

$$\mathbb{E}_{\boldsymbol{S}\sim\pi^\star\left(\left\{f^\star\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_i\right]=m.$$

The sum of the differences on the two pools of candidates is

$$\left|\mathbb{E}_{\boldsymbol{S}\sim\pi_h\left(\{h(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_i\right]-\mathbb{E}_{\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\{f^\star(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_i\right]\right|$$
$$+\left|\mathbb{E}_{\boldsymbol{S}\sim\pi_h\left(\left\{h\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_i\right]-\mathbb{E}_{\boldsymbol{S}\sim\pi^\star\left(\left\{f^\star\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_i\right]\right|=|c_{\pi_h}-k|+|c_{\pi_h}-m|\geq m-k.$$

So for any policy π_h, there exists a pool of candidates (either $\boldsymbol{x}$ or $\boldsymbol{x}'$) such that the difference is at least $\frac{m-k}{2}$.

For the expected number of qualified candidates in the shortlists, for any policy π_h, we have

$$\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi_h\left(\{h(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]=c_{\pi_h},$$

and

$$\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi_h\left(\left\{h\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]=\frac{k}{m}c_{\pi_h}.$$

On the other hand, for the optimal policy $\pi^\star_{f^\star}$, we know that

$$\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\{f^\star(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]=\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\left\{f^\star\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]=k.$$

The sum of differences on the two pools of candidates is

$$\left|\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi_h\left(\{h(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]-\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\{f^\star(x_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]\right|$$
$$+\left|\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi_h\left(\left\{h\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]-\mathbb{E}_{\boldsymbol{Y}\sim P,\boldsymbol{S}\sim\pi^\star_{f^\star}\left(\left\{f^\star\left(x_i'\right)\right\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}S_iY_i\right]\right|$$
$$=|c_{\pi_h}-k|+\left|\frac{k}{m}c_{\pi_h}-k\right|\geq k\left(1-\frac{k}{m}\right),$$

where the equality is achieved when $c_\pi=k$. So for any policy π_h, there exist a set of candidates (either $\boldsymbol{x}$ or $\boldsymbol{x}'$) such that the difference is at least $\frac{k}{2}\left(1-\frac{k}{m}\right)$.

C.3. Proof of Theorem 4.1

First, we show that the constraint in the minimization problem in (2) is satisfied with equality:

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_h^\star}\left[\sum_{i\in[m]} Y_i S_i\right] &= \mathbb{E}_{\{(h(X_i),Y_i)\}_{i\in[m]}\sim P^m_{h(X),Y},\boldsymbol{S}\sim\pi_h^\star\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} Y_i S_i\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)},\boldsymbol{S}\sim\pi_h^\star\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i \Pr\left(Y=1 \mid h(X_i)\right)\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)},\boldsymbol{S}\sim\pi_h^\star\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i h(X_i)\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)}}\left[\sum_{i\in[m]} \left[\mathbb{I}\left\{h(X_i)>t_h\right\}h(X_i) + \mathbb{I}\left\{h(X_i)=t_h\right\}h(X_i)\theta_h\right]\right] \\
&= m\mathbb{E}_{h(X)\sim P_{h(X)}}\left[\mathbb{I}\left\{h(X)>t_h\right\}h(X) + \mathbb{I}\left\{h(X)=t_h\right\}h(X)\theta_h\right] \\
&= k,
\end{aligned}$$

where the last equality is by the definition of t_h and θ_h.

Next, we will show that it is an optimal solution by contradiction. The objective in (2) using $\pi_h^\star$ is

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_h^\star}\left[\sum_{i\in[m]} S_i\right] &= \mathbb{E}_{\boldsymbol{X}\sim P}\left[\sum_{i\in[m]} \left(\mathbb{I}\left\{h(X_i)>t_h\right\} + \theta_h\mathbb{I}\left\{h(X_i)=t_h\right\}\right)\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)}}\left[\sum_{i\in[m]} \left(\mathbb{I}\left\{h(X_i)>t_h\right\} + \theta_h\mathbb{I}\left\{h(X_i)=t_h\right\}\right)\right] \\
&= m\mathbb{E}_{h(X)\sim P_{h(X)}}\left[\mathbb{I}\left\{h(X)>t_h\right\} + \theta_h\mathbb{I}\left\{h(X)=t_h\right\}\right] \\
&= m\sum_{b\in[B]} \rho_b^\star\left(\mathbb{I}\left\{\mu_b^\star>t_h\right\} + \theta_h\mathbb{I}\left\{\mu_b^\star=t_h\right\}\right).
\end{aligned}$$

For any policy π_h' that achieves smaller objective than $\pi_h^\star$

$$\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_h'}\left[\sum_{i\in[m]} S_i\right] \coloneqq R_{\pi_h'} < m\sum_{b\in[B]} \rho_b^\star\left(\mathbb{I}\left\{\mu_b^\star>t_h\right\} + \theta_h\mathbb{I}\left\{\mu_b^\star=t_h\right\}\right),$$

we will show that the constraint is not satisfied using π_h'. Let

$$t_h' \coloneqq \sup\left\{t\in[0,1] : \sum_{b\in[B]} \rho_b^\star\mathbb{I}\left\{\mu_b^\star\geq t\right\} \geq R_{\pi_h'}/m\right\},$$

and

$$\theta' = \frac{R_{\pi_h'}/m - \sum_{b\in[B]}\rho_b^\star\mathbb{I}\left\{\mu_b^\star>t'\right\}}{\sum_{b\in[B]}\rho_b^\star\mathbb{I}\left\{\mu_b^\star=t'\right\}}.$$

Note that $t_h \leq t'$ since

$$\sum_{b\in[B]} \rho_b^\star\mathbb{I}\left\{\mu_b^\star\geq t_h\right\} \geq \sum_{b\in[B]} \rho_b^\star\left(\mathbb{I}\left\{\mu_b^\star>t_h\right\} + \theta_h\mathbb{I}\left\{\mu_b^\star=t_h\right\}\right) > R_{\pi_h'}/m.$$

Now we can show that the constraint is not satisfied using π'_h. The left hand side of the constraint using π'_h can be bounded as

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi'_h}\left[\sum_{i\in[m]} Y_i S_i\right] &= \mathbb{E}_{\{(h(X_i),Y_i)\}_{i\in[m]}\sim P^m_{h(X),Y},\boldsymbol{S}\sim\pi'_h\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} Y_i S_i\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)},\boldsymbol{S}\sim\pi'_h\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i \Pr\left(Y=1 \mid h(X_i)\right)\right] \\
&= \mathbb{E}_{\{h(X_i)\}_{i\in[m]}\sim P^m_{h(X)},\boldsymbol{S}\sim\pi'_h\left(\{h(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]} S_i h(X_i)\right] \\
&\le m\mathbb{E}_{h(X)\sim P_{h(X)}}\left[h(X)\left(\mathbb{I}\left\{h(X)>t'_h\right\}+\theta'_h\mathbb{I}\left\{h(X)=t'_h\right\}\right)\right] \\
&= m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t'_h\right\}+\theta'_h\mathbb{I}\left\{\mu^\star_b=t'_h\right\}\right).
\end{aligned}$$

The inequality is by the definition of $R_{\pi'_h}$, t'_h and θ'_h. When $t_h < t'_h$,

$$m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t'_h\right\}+\theta'_h\mathbb{I}\left\{\mu^\star_b=t'_h\right\}\right) m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t_h\right\}+\theta_h\mathbb{I}\left\{\mu^\star_b=t_h\right\}\right)=k.$$

When $t_h = t'_h$,

$$\theta'_h=\frac{R_{\pi'_h}/m-\sum_{b\in[B]}\rho^\star_b\mathbb{I}\left\{\mu^\star_b>t_h\right\}}{\sum_{b\in[B]}\rho^\star_b\mathbb{I}\left\{\mu^\star_b=t_h\right\}}<\theta_h.$$

Thus

$$\begin{aligned}
m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t'_h\right\}+\theta'_h\mathbb{I}\left\{\mu^\star_b=t'_h\right\}\right) &= m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t_h\right\}+\theta'_h\mathbb{I}\left\{\mu^\star_b=t_h\right\}\right) \\
&< m\sum_{b\in[B]}\rho^\star_b\mu^\star_b\left(\mathbb{I}\left\{\mu^\star_b>t_h\right\}+\theta_h\mathbb{I}\left\{\mu^\star_b=t_h\right\}\right) \\
&= k.
\end{aligned}$$

This shows that no policy in Π_h can achieve smaller objective than $\pi^\star_h$, while satisfying the constraint. So $\pi^\star_h$ is a solution to the minimization problem among Π_h.

C.4. Proof of Corollary 4.3

By the definition of Π_h, it is easy to see that the policy space of using h' is a subset of that of h $\Pi_{h'}\subseteq\Pi_h$, the corollary directly follows from Theorem 4.1.

C.5. Proof of Theorem 4.5

The proof constructs a policy $\pi_{f,t}$ for any policy π_f such that the two inequalities in the theorem hold. For any policy π_f, let

$$k_{\pi_f}\coloneqq\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[m]} S_i Y_i\right]$$

be the expected number of qualified candidates selected by π_f, $\mu_f(\cdot)=\Pr\left(Y=1\,|\,f(X)=\cdot\right)$ denote the conditional probability that a candidate is qualified given a prediction quality score from f. We construct the the threshold in policy $\pi_{f,t}$ as

$$t=\sup\left\{u\in[0,1]: m\mathbb{E}_{f(X)\sim P_{f(X)}}\left[\mu_f(f(X))\mathbb{I}\left\{f(X)\ge u\right\}\right]\ge k_{\pi_f}\right\}.$$

The expected number of qualified candidates using $\pi_{f,t}$ is

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]}Y_iS_i\right] &= \mathbb{E}_{\{(f(X_i),Y_i)\}_{i\in[m]}\sim P^m_{f(X),Y},\boldsymbol{S}\sim\pi_{f,t}\left(\{f(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}Y_iS_i\right]\\ &= \mathbb{E}_{\{f(X_i)\}_{i\in[m]}\sim P^m_{f(X)}}\left[\sum_{i\in[m]}\mu_f(f(X_i))\mathbb{I}\left\{f(X_i)\geq t\right\}\right]\\ &= m\mathbb{E}_{f(X)\sim P_{f(X)}}\left[\mu_f(f(X))\mathbb{I}\left\{f(X)\geq t\right\}\right]\\ &= k_{\pi_f}.\end{aligned}$$

This shows that the expected number of qualified candidates selected by $\pi_{f,t}$ is no smaller than π_f. Now we only need to show the expected size of the shortlist using $\pi_{f,t}$ is no larger than that of using π_f. The expected size of the shortlist using $\pi_{f,t}$ is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]}S_i\right] = \mathbb{E}_{\{f(X_i)\}_{i\in[m]}\sim P^m_{f(X)}}\left[\sum_{i\in[m]}\mathbb{I}\left\{f(X_i)\geq t\right\}\right] &= m\mathbb{E}_{f(X)\sim P_{f(X)}}\left[\mathbb{I}\left\{f(X)\geq t\right\}\right]\\ &= m\Pr\left(f(X)\geq t\right).\end{aligned}$$

We will show that π_f achieves no smaller expected size of the shortlists by using contradiction. Suppose π_f achieves smaller expected size of the shortlist than $\pi_{f,t}$

$$\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[m]}S_i\right] \coloneqq R_{\pi_f} < m\Pr\left(f(X)\geq t\right),$$

we will show that under this assumption, we will arrive at a contradiction that the expected number of qualified candidates selected by π_f is smaller than k_{π_f}. Let

$$t' = \sup\left\{u\in[0,1] : m\Pr\left(f(X)\geq u\right)\geq R_{\pi_f}\right\}.$$

Note that $t < t'$ since

$$m\Pr\left(f(X)\geq t\right) > R_{\pi_f}.$$

Then, the expected number of qualified candidates using π_f can be bounded as

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[m]}Y_iS_i\right] &= \mathbb{E}_{\{f(X_i)\}_{i\in[m]}\sim P^m_{f(X)},\boldsymbol{S}\sim\pi_f\left(\{f(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}\mu_f(f(X_i))S_i\right]\\ &\leq \mathbb{E}_{\{f(X_i)\}_{i\in[m]}\sim P^m_{f(X)},\boldsymbol{S}\sim\pi_{f,t'}\left(\{f(X_i)\}_{i\in[m]}\right)}\left[\sum_{i\in[m]}\mu_f(f(X_i))S_i\right]\\ &= m\mathbb{E}_{f(X)\sim P_{f(X)}}\left[\mu_f(f(X))\mathbb{I}\left\{f(X)\geq t'\right\}\right]\\ &< m\mathbb{E}_{f(X)\sim P_{f(X)}}\left[\mu_f(f(X))\mathbb{I}\left\{f(X)\geq t\right\}\right]\\ &= k_{\pi_f}.\end{aligned}$$

The inequality is by monotone property of f. This arrives at a contradiction with the definition that the expected size of shortlist using π_f is k_{π_f}. So, we get

$$\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[m]}S_i\right] \geq \mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]}S_i\right],$$

which concludes the proof.

C.6. Proof of Proposition 4.6

Before introducing the proof of the proposition, we first introduce the DKWM inequality (Dvoretzky et al., 1956; Massart, 1990) as a lemma.

Lemma C.1 (DKWM inequality (Dvoretzky et al., 1956; Massart, 1990))**.** *Given a natural number n, let Z_1, Z_2,..., Z_n be real-valued independent and identically distributed random variables with cumulative distribution function $F(\cdot)$. Let F_n denote the associated empirical distribution function defined by*

$$F_n(z) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{Z_i \le z\} \quad z \in \mathbb{R}.$$

For any $\alpha \in (0,1)$, with probability $1-\alpha$

$$\sup_{z\in\mathbb{R}} |F(x) - F_n(z)| \le \sqrt{\frac{\ln(2/\alpha)}{2n}}.$$

Equipped with this lemma, we provide the proof for Proposition 4.6 as follows.

Let

$$Z_i^\delta = -Y_i^c f(X_i^c) \quad \forall i \in [n].$$

They are real-valued independent and identically distributed random variables. Let $F^\delta(\cdot)$ denote the cumulative distribution function of Z^δ and

$$F_n^\delta(z) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{-Y_i^c f(X_i^c) \le z\} \quad \forall z \in \mathbb{R}.$$

Applying the DKWM inequality, we can get that for any $\alpha \in (0,1)$, with probability at least $1-\alpha$, for any $z \in \mathbb{R}$,

$$\left|F_n^\delta(z) - F^\delta(z)\right| \le \sqrt{\frac{\ln(2/\alpha)}{2n}}.$$

Note that

$$F_n^\delta(t) = -\hat{\delta}_{t,1},$$

and

$$F^\delta(t) = -\delta_{t,1},$$

which concludes the proof.

C.7. Proof of Corollary 4.7

When the event in Proposition 4.6 holds,

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]} S_i Y_i\right] &= \mathbb{E}_{\{(f(X_i),Y_i)\}_{i\in[m]}\sim P^m_{f(X),Y}, \boldsymbol{S}\sim\pi_{f,t}(\{f(X_i)\}_{i\in[m]})}\left[\sum_{i\in[m]} S_i Y_i\right] \\
&= \mathbb{E}_{\{(f(X_i),Y_i)\}_{i\in[m]}\sim P^m_{f(X),Y}}\left[\sum_{i\in[m]} Y_i \mathbb{I}\{f(X_i) \ge t\}\right] \\
&= m\mathbb{E}_{(f(X),Y)\sim P_{f(X),Y}}\left[Y\mathbb{I}\{f(X)\ge t\}\right] \\
&= m\delta_{t,1}.
\end{aligned}$$

By Proposition 4.6, this corollary holds, which concludes the proof.

C.8. Proof of Theorem 4.8

We can see that

$$\mathbb{E}_{\boldsymbol{X}\sim P,\boldsymbol{S}\sim\pi_{f,t}}\left[\sum_{i\in[m]} S_i\right]$$

is a non-increasing function of t. So to solve the constraint minimization problem, we only need to select the largest t that satisfies the constraint, which is Eq. 7. This concludes the proof.

C.9. Proof of Proposition 4.9

For the optimal policy $\pi_{f,t^\star_f}$, the constraint is active (otherwise we can randomly select the candidates selected by $\pi_{f,t^\star_f}$ to derive a better policy, which contradicts that $\pi_{f,t^\star_f}$ is optimal)

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,t^\star_f}}\left[\sum_{i\in[m]} S_iY_i\right] = k.$$

Since the distribution of $f(X)$ is nonatomic, all the candidates have different prediction scores almost surely. Thus, by the definition of $\hat{t}_f$, we know that

$$\hat{\delta}_{\hat{t}_f,1} - \epsilon(\alpha, n) < k/m + 1/n$$

almost surely. By Proposition 4.6 and the proof in Corollary 4.7, we have that for any $\alpha \in (0,1)$, with probability at least $1-\alpha$,

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,\hat{t}_f}}\left[\sum_{i\in[m]} Y_iS_i\right] = m\delta_{\hat{t}_f,1} \le m\left(\hat{\delta}_{\hat{t}_f,1} + \epsilon(\alpha,n)\right) < k + 2m\epsilon(\alpha,n) + m/n.$$

This concludes the proof.

C.10. Proof of Proposition 4.10

Let $Q_i = S_iY_i$ for all $i \in [m]$ be the random variables of whether a candidate i is qualified and selected in the pool of candidates. Note that

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,\hat{t}_f}}[Q_i] = \delta_{\hat{t}_f,1}.$$

By Theorem 4.8, we know that for any $\alpha_1 \in (0,1)$ with probability at least $1-\alpha_1$,

$$\mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})\sim P,\boldsymbol{S}\sim\pi_{f,\hat{t}_f}}\left[\sum_{i\in[m]} Q_i\right] = m\delta_{\hat{t}_f,1} \ge k.$$

By applying Bernstein's inequality to these m independent and identically distributed random variables, we have that, for any $\alpha_2 \in (0,1)$, with probability at least $1-\alpha_2$

$$\sum_{i\in[m]} Q_i \ge m\delta_{\hat{t}_f,1} - \frac{1}{3}\ln(1/\alpha_2) - \sqrt{1/9\ln^2(1/\alpha_2) + 2m\delta_{\hat{t}_f,1}\ln(1/\alpha_2)}.$$

The right hand side of the above inequality is an increasing function of $m\delta_{\hat{t}_f,1}$ when $m\delta_{\hat{t}_f,1} \ge 4/9\ln(1/\alpha_2)$. When $k \ge 4/9\ln(1/\alpha_2)$, we can apply the union bound to the above two events to get that, with probability at least $1-\alpha_1-\alpha_2$,

$$\sum_{i\in[m]} Q_i \ge m\delta_{\hat{t}_f,1} - \frac{1}{3}\ln(1/\alpha_2) - \sqrt{1/9\ln^2(1/\alpha_2) + 2m\delta_{\hat{t}_f,1}\ln(1/\alpha_2)} \ge k - \frac{1}{3}\ln(1/\alpha_2) - \frac{1}{3}\sqrt{\ln^2(1/\alpha_2) + 18k\ln(1/\alpha_2)}.$$

This concludes the proof.

C.11. Proof of Proposition A.2

We prove the proposition for f that has a discrete range, and the proof for classifiers with continuous range can be derived similarly. For any π_f, we construct a pool-independent policy π'_f by constructing $p_{\pi'_f} \in \mathrm{Range}(f) \to [0,1]$ as

$$p_{\pi'_f}(v) = \frac{\mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[M]} \mathbb{I}\left\{f(X_i)=v\right\} S_i\right]}{\mathbb{E}\left[M\right]\mathbb{E}_{f(X)\sim P_{f(X)}}\left\{\mathbb{I}\left\{f(X)=v\right\}\right\}}.$$

Let π'_f be the policy using $p_{\pi'_f}$. The expected number of qualified candidates for policy π'_f equals that of π_f

$$\begin{aligned}
&\mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi'_f}\left[\sum_{i\in[M]} Y_i S_i\right] \\
&= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi'_f\left(\{f(X_i)\}_{i\in[M]}\right)}\left[\sum_{i\in[M]} S_i \Pr\left(Y=1 \mid f(X_i)\right)\right] \\
&= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}}\left[\sum_{i\in[M]} p_{\pi'_f}(f(X_i)) \Pr\left(Y=1 \mid f(X_i)\right)\right] \\
&= \mathbb{E}\left[M\right]\mathbb{E}_{v\sim P_{f(X)}}\left[p_{\pi'_f}(v)\Pr\left(Y=1\mid v\right)\right] \\
&= \mathbb{E}_{v\sim P_{f(X)}}\left[\frac{\mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[M]} \mathbb{I}\left\{f(X_i)=v\right\} S_i\right]\Pr\left(Y=1\mid v\right)}{\mathbb{E}_{f(X)\sim P_{f(X)}}\left\{\mathbb{I}\left\{f(X)=v\right\}\right\}}\right] \\
&= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi_f\left(\{f(X_i)\}_{i\in[M]}\right)}\left[\sum_{i\in[M]} S_i \Pr\left(Y=1 \mid f(X)\right)\right] \\
&= \mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[M]} Y_i S_i\right],
\end{aligned}$$

and the expected size of the shortlist equals that of π_f

$$\begin{aligned}
\mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi'_f}\left[\sum_{i\in[M]} Y_i S_i\right] &= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi'_f\left(\{f(X_i)\}_{i\in[M]}\right)}\left[\sum_{i\in[M]} S_i\right] \\
&= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}}\left[\sum_{i\in[M]} p_{\pi'_f}(f(X_i))\right] \\
&= \mathbb{E}\left[M\right]\mathbb{E}_{v\sim P_{f(X)}}\left[p_{\pi'_f}(v)\right] \\
&= \mathbb{E}_{v\sim P_{f(X)}}\left[\frac{\mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[M]} \mathbb{I}\left\{f(X_i)=v\right\} S_i\right]}{\mathbb{E}_{f(X)\sim P_{f(X)}}\left\{\mathbb{I}\left\{f(X)=v\right\}\right\}}\right] \\
&= \mathbb{E}_{M\sim P_M, \{f(X_i)\}_{i\in[M]}\sim P^M_{f(X)}, \boldsymbol{S}\sim\pi_f\left(\{f(X_i)\}_{i\in[M]}\right)}\left[\sum_{i\in[M]} S_i\right] \\
&= \mathbb{E}_{(M,\boldsymbol{X},\boldsymbol{Y})\sim P, \boldsymbol{S}\sim\pi_f}\left[\sum_{i\in[M]} Y_i S_i\right].
\end{aligned}$$

C.12. Proof of Proposition A.3

From the proof of Proposition 4.6, we know that with probability at least $1-\alpha$

$$\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{f(X)\geq t\right\}\right]-\frac{1}{n}\sum_{i\in[n]}Y_i^c\mathbb{I}\left\{f(X_i^c)\geq t\right\}\right|\leq\epsilon(\alpha,n)$$

for all $t\in[0,1]$. So, with probability at least $1-\alpha$, for any $0\leq t_a<t_b\leq 1$,

$$\begin{aligned}
&\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{t_a\leq f(X)\leq t_b\right\}\right]-\frac{1}{n}\sum_{i\in[n]}Y_i^c\mathbb{I}\left\{t_a\leq f(X_i^c)\leq t_b\right\}\right| \\
&\qquad\leq\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{f(x)\geq t_a\right\}\right]-\frac{1}{n}\sum_{i\in[n]}Y_i^c\mathbb{I}\left\{f(X_i^c)\geq t_a\right\}\right| \\
&\qquad\qquad+\left|\mathbb{E}_{(X,Y)\sim P_{X,Y}}\left[Y\mathbb{I}\left\{f(X)\geq t_b\right\}\right]-\frac{1}{n}\sum_{i\in[n]}Y_i^c\mathbb{I}\left\{f(X_i^c)\geq t_b\right\}\right|\leq 2\epsilon(\alpha,n).
\end{aligned}$$

This concludes the proof.