NPSL: TRACTABLE AND ROBUST CLASSIFICATION UNDER SELECTIVE LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many binary classification problems involve learning from data with partially missing labels. In the selective label setting, a prior decision determines which labels are observed. For example, banks observe repayment outcomes only for applicants who were granted a loan. Error control is then difficult because the observations with labels may not be representative of all observations. We study classification in this setting using the Neyman–Pearson (NP) framework, which seeks to control the prioritized type I error while minimizing the type II error. We propose NPSL (Neyman–Pearson under Selective Labels), using structural assumptions about the unlabeled population to construct a set of plausible labelings. Given an independently trained scoring function, NPSL selects a threshold that ensures the empirical type I error is at most a user-specified target across all plausible labelings simultaneously. Although the number of possible labelings grows exponentially, we show that this can be done in polynomial time. Under additional conditions, NPSL converges to the optimal robust threshold — the smallest threshold that still satisfies the type I error constraint across all plausible distributions. Synthetic experiments empirically show that NPSL provides reliable type I error control while achieving lower type II error than benchmark methods that satisfy the NP constraint. A case study using loan application data illustrates risk control of NPSL with artificially hidden outcomes.

1 INTRODUCTION

The goal of binary classification is to predict class labels $Y \in \{0, 1\}$ from observed covariates $X \in \mathbb{R}^p$. In this setting, two types of errors may be made: predicting class 1 when the actual class is 0 (a type I error) or predicting class 0 when the actual class is 1 (a type II error). In practice, one error can be more consequential than the other, and we will consider the type I error to be the more severe error. One technique to address this asymmetry is Neyman–Pearson (NP) classification, which seeks to keep the probability of making a type I error at most α while minimizing the probability of making a type II error (Cannon et al., 2002; Scott & Nowak, 2005). The NP-umbrella algorithm (Tong et al., 2018) provides a general framework for such guarantees by splitting the data into a training set and calibration set. The training set is used to learn a scoring function, and the calibration set is used to select a threshold based on the empirical type I error, adjusting for sampling uncertainty.

In many applications, however, outcome labels are available only for a subset of observations. In this setting, each observation includes X and $A \in \{0, 1\}$, where A records whether Y is missing ($A = 0$) or observed ($A = 1$). Missing labels prevent direct computation of the empirical type I error, but calibration on only the labeled data may be biased. This concern is particularly relevant in the selective label setting (Lakkaraju et al., 2017), where A is determined by a historical decision, such as in medical diagnosis, bail decisions, and loan approval (Begg & Greenes, 1983; Lakkaraju et al., 2017; Ehrhardt et al., 2021; Balachandar et al., 2024). In these cases, label observation can depend on X, Y, or both. Consequently, the labeled observations may have different distributions from the full population, potentially biasing calibration using only this set.

A common starting point for analysis when labels are missing is to assume that Y and A are conditionally independent given X. Under this assumption, the conditional distribution of Y given X is the same for the labeled and unlabeled observations. Thus, under suitable regularity conditions, this distribution can be estimated using a range of existing techniques (Bang & Robins, 2005; Cher-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nozhukov et al., 2018; Chakrabortty et al., 2019; Wang & Zubizarreta, 2020; Zhang et al., 2023). However, this assumption can be unrealistic in practice, as label observation may depend on features not present in the data (Ehrhardt et al., 2021). In this setting, the conditional distribution is generally not identifiable without further assumptions (Little, 1993; Sun et al., 2018). Approaches to this problem include semiparametric methods, using auxiliary variables to enable identification (Sun et al., 2018; Miao et al., 2024) or sensitivity analyses, which assess how risk estimation changes as the independence assumption is relaxed (Zhao et al., 2019; Rambachan et al., 2026).

Rather than attempting to identify the distribution of the missing labels, we take an approach motivated by distributionally robust optimization (DRO) (e.g., Mohajerin Esfahani & Kuhn, 2018). We consider a set of plausible labelings, then require that type I error be at most α across the entire set. NPSL (Neyman–Pearson under Selective Labels) implements this strategy, calibrating any independently learned scoring function. At each candidate threshold t, NPSL computes the worst-case empirical type I risk curve $\widehat{U}_0(t)$ over all plausible labelings, selecting the smallest candidate threshold satisfying $\widehat{U}_0(t) \leq \alpha$. Unlike imputation or pseudo-labeling, no observation is assigned a single label or probability.

The primary challenge is then defining restrictions to construct a plausible set. Three considerations guide the construction of these restrictions. First, they should correspond to mild, interpretable assumptions that hold across a broad range of settings. Second, the resulting set should include enough labelings to protect against underestimation of the risk, while excluding implausible labelings that lead to overly conservative classifiers. Finally, the restrictions should permit efficient computation of the worst-case empirical type I error, despite the exponentially many possible labelings. We address these considerations by specifying restrictions on class proportions and score distributions (Section 2.3). For example, we may assume that the score is informative for the unlabeled observations, requiring that the observations assigned to class 1 have generally higher scores than those assigned to class 0.

At the population level, the overall score distribution of the unlabeled observations G is identifiable, but its decomposition into class 0 and class 1 components is not. We can thus consider plausible decompositions of G, replacing empirical class proportions and score distributions in our restrictions with their population counterparts. The corresponding objective is then to minimize the type II risk while keeping type I risk at most α across all plausible distributions. We define the optimal robust threshold, t^*_α, as the smallest threshold satisfying the type I constraint.

We make two primary contributions.

(i) *Practical restrictions with exact, tractable calibration.* We use restrictions on plausible labelings to enable computation of the empirical type I error for Neyman–Pearson classification. We show that the structure of these restrictions allows exact computation of $\widehat{U}_0(t)$ without enumeration over all possible labelings, returning the estimated threshold in $O(N^2 \log^2 N)$ time. Here, N is the number of points in the calibration set (Theorem 3.1).

(ii) *Convergence to the optimal robust threshold.* We show that, under suitable regularity conditions and assumptions on the plausible set, the NPSL threshold converges in probability to t^*_α (Proposition 4.2).

In practice, the set of plausible labelings depends on several parameters. We provide data-dependent default choices for these parameters and evaluate their performance empirically. In synthetic-data experiments, NPSL with these defaults shows low population type I error violation rates and lower average type II error than methods that use only labeled data or consider all possible labelings. A case study on loan application data studies error control using artificially hidden outcomes.

1.1 Related Work

We adapt the split-sample framework of the NP-umbrella algorithm (Tong et al., 2018) to selective labels by calibrating against the worst-case empirical type I error across a set of plausible labelings. Weighted conformal prediction methods use the same training-calibration split setup to address risk under covariate shift, which in our setting corresponds to the assumption that label observation is independent of the label conditional on the covariates (Tibshirani et al., 2019; Lei & Candès, 2021). This assumption can be violated in the general selective label setting.

Semi-supervised learning also uses labeled and unlabeled data to learn classifiers, often using predicted labels or class probabilities during training (Bennett & Demiriz, 1998; Berthelot et al., 2019). The predicted labels could also be used to estimate risk for calibration, but this does not necessarily account for the bias inherent to the selective label problem. NPSL addresses uncertainty by considering a set of plausible labelings and evaluating the worst-case empirical type I risk over this set. In general, NPSL can calibrate any independently trained scoring function, including one learned using semi-supervised learning.

Related work addresses the selective label problem within semi-supervised learning. For example, Sportisse et al. (2023) estimate the label selection mechanism under the assumption that labeled and unlabeled observations have the same feature distribution within each class. Our restrictions allow these distributions to differ, without attempting to estimate the label selection mechanism. Another technique bounds risks by applying sensitivity models that impose pointwise bounds on departures from conditional independence between label observation and the label given the covariates (Rambachan et al., 2026). We consider plausible score distributions induced by the missing labels as a whole, which allow us to rule out labelings that satisfy the pointwise bounds but violate our restrictions. The structure of our restrictions also permits exact computation of the worst-case empirical risk. Appendix A provides a more detailed comparison of the connections to the literature.

2 Model Setup

2.1 Selective labels and the Neyman–Pearson objective

We consider i.i.d. observations of the form (X, A, Y), where $X \in \mathbb{R}^p$ represents the covariates, $Y \in \{0, 1\}$ represents the class label, and $A \in \{0, 1\}$ records whether the label is observed. The label is observed when $A = 1$ and missing when $A = 0$. We split the data into two independent sets, consisting of a training set used to learn a scoring function $S = s(X) \in \mathbb{R}$ and a calibration set used to select the threshold for that score. Larger scores provide greater evidence for class 1. The scoring function may be learned using any method, such as an SVM, logistic regression, or a random forest. We write N for the size of the calibration set, consisting of n_0 and n_1 labeled observations in class 0 and class 1, respectively, and m unlabeled observations, with $N = n_0 + n_1 + m$.

For $t \in \mathbb{R}$, we consider threshold classifiers of the form $\phi_t(X) = \mathbb{1}\{S \geq t\}$. Their population type I and type II risks over the full population are $R_0(t) := \mathbb{P}(S \geq t \mid Y = 0)$ and $R_1(t) := \mathbb{P}(S < t \mid Y = 1)$. The population Neyman–Pearson objective minimizes $R_1(t)$ over $t \in \mathbb{R}$ subject to $R_0(t) \leq \alpha$, where $\alpha \in (0, 1)$ is the target population type I risk. Because $R_0(t)$ and $R_1(t)$ are monotone in t, the optimal rule uses the smallest threshold satisfying the constraint, provided one exists. In practice, the population risks are unknown, so the threshold must be learned from the data.

It is instructive to describe the standard NP-umbrella algorithm when all labels are observed. On the calibration set, let $X_0(t)$ be the number of class 0 observations at or above t and N_0 be the total number of class 0 observations, so that the empirical type I risk curve is given by $\widehat{R}_0(t) = X_0(t)/N_0$. The NP-umbrella algorithm can be viewed as constructing a conservative risk curve $V(t)$ that accounts for sampling uncertainty. The NP-umbrella algorithm then selects the threshold

$$\widehat{t} = \min\{t \in \mathcal{T} : V(t) \leq \alpha\}, \tag{2.1}$$

where $\mathcal{T}$ is the set of candidate thresholds.

NPSL uses the same framework, replacing $V(t)$ with $\widehat{U}_0(t)$, the worst-case empirical type I risk over a set of hypothetical labelings. Here, neither N_0 nor X_0 is known, but given a particular assignment of the unlabeled data, both quantities can easily be computed. Thus, we may select the smallest threshold such that $\widehat{U}_0(t) \leq \alpha$, enforcing the empirical constraint simultaneously over all plausible labelings. At the population level, an additional adjustment is required to account for sampling uncertainty (Section 4).

2.2 Notation

Throughout, we condition on the training data, and thus treat the scoring function as fixed. We write $\mathcal{S}_0$ and $\mathcal{S}_1$ for the observed labeled scores of classes 0 and 1, respectively, with sample sizes $n_0, n_1 >$

0 and $U = (U_1, \ldots, U_m)$ for the unlabeled scores. For NPSL, we take the candidate thresholds to be the distinct observed scores in $\mathcal{S}_0 \cup \mathcal{S}_1 \cup \{U_1, \ldots, U_m\}$, because the relevant empirical risks can only change at these values.

We consider hypothetical labelings of U, written $z = (z_1, \ldots, z_m) \in \{0,1\}^m$. Then $k(z) := \sum_{i=1}^m z_i$ is the number of unlabeled observations set to class 1 under labeling z. For a hypothetical labeling z, $N_{0,z} = n_0 + m - k(z)$ is the total number of class 0 observations, and $X_{0,z}(t)$ is the number of type I errors made if the threshold were set to t. The corresponding empirical type I risk is $\widehat{R}_{0,z}(t) := X_{0,z}(t)/N_{0,z}$, and the worst-case empirical type I risk over a set of labelings $\mathcal{Z}$ is

$$\widehat{U}_0(t) = \max_{z \in \mathcal{Z}} \widehat{R}_{0,z}(t), \tag{2.2}$$

with the convention that $\widehat{U}_0(t) = 1$ if $\mathcal{Z}$ is empty. The corresponding analogue of Equation (2.1) is then

$$\mathcal{T}_\alpha = \{t \in \mathcal{T} : \widehat{U}_0(t) \le \alpha\}, \qquad \widehat{t} = \min \mathcal{T}_\alpha. \tag{2.3}$$

If $\mathcal{T}_\alpha = \varnothing$, we set $\widehat{t} = +\infty$, by convention.

We define $\widehat{F}_y$ to be the empirical CDF of the labeled scores in $\mathcal{S}_y$, for $y \in \{0,1\}$. Given a labeling z, with $1 \le k(z) \le m-1$, we define $\widehat{G}_{y,z}$ to be the empirical CDF of the unlabeled scores U_i with $z_i = y$. Finally, we will restrict the possible labelings using assumptions about how these distributions vary across groups. For example, we may expect labeled observations within a class to have higher scores than unlabeled ones. To that end, for two CDFs G and F, we define the quantity $D_+(G,F) := \int_{\mathbb{R}} (G(u) - F(u))_+ \, du$, where $(a)_+ := \max\{a, 0\}$. Intuitively, this quantity measures how much lower scores from G can be relative to the scores from F.

2.3 Admissible labelings

In this section, we define plausible labelings through four restrictions on class proportions and score distributions. These restrictions encode substantive assumptions about the structure of the scores, which we motivate below. Let $\mathcal{K} \subseteq \{1, \ldots, m-1\}$. For $\delta_0, \delta_1, \gamma \ge 0$, we define the set of admissible labelings to be

$$\begin{aligned} \widehat{\mathcal{Z}} := \Big\{ z \in \{0,1\}^m : k(z) \in \mathcal{K}, \quad D_+(\widehat{F}_0, \widehat{G}_{0,z}) \le \delta_0, \quad D_+(\widehat{G}_{1,z}, \widehat{F}_1) \le \delta_1, \\ \widehat{G}_{1,z}(u) \le \widehat{G}_{0,z}(u) + \gamma \quad \text{for every } u \in \mathbb{R} \Big\}. \end{aligned} \tag{2.4}$$

We may then take $\mathcal{Z} = \widehat{\mathcal{Z}}$ in Equation (2.2) for NPSL. If $\widehat{\mathcal{Z}}$ is empty, we return a conservative but uninformative rule, $\widehat{\phi}(x) \equiv 0$, assigning every observation to class 0. Each of the parameters δ_0, δ_1, and γ must be chosen or tuned, with larger values of δ_0, δ_1, and γ allowing for more labelings and corresponding to more conservative thresholds. Below, we explain the restrictions and provide recommended, empirically motivated defaults. We study sensitivity to these parameters in Appendix D.2.

To make the exposition concrete, we use the loan approval setting, though the framework can apply more generally to many selective label settings when these restrictions are defensible. Thus, X corresponds to information in a loan application, $A = 1$ denotes an approved application and $A = 0$ a declined application, while $Y = 1$ denotes repayment and $Y = 0$ default if a loan was approved. Here, the selective label problem is evident, as we generally only observe repayment information for applicants who were actually granted a loan.

Class proportions among unlabeled observations should be limited. The first restriction requires that any admissible labeling contain $k(z) \in \mathcal{K}$ class 1 observations in the unlabeled data. This allows us to exclude, for example, a labeling that simply sets all unlabeled observations to class 1. As a default, we use $\mathcal{K} = \{1, \ldots, k^*\}$, where $k^* = \min\{m-1, \lfloor m n_1/(n_0+n_1) \rfloor\}$. This states that the proportion of class 1 points may not be higher in the declined applicants than in the approved applicants, reflecting the intuition that loan approval decisions should favor applicants more likely to repay.

Within class 0, labeled observations tend to have higher scores. The second restriction bounds how far labeled class 0 scores can fall below unlabeled class 0 scores, as measured by D_+. A value of $\delta_0 = 0$ reflects the intuition that approved defaulters should have higher scores than declined applicants who would default. To account for finite-sample variability, we set $\delta_0 = \widehat{\sigma} N^{-1/3}$, where $\widehat{\sigma}$ is the sample standard deviation of the scores in the calibration set and N is the size of the calibration set.

Within class 1, unlabeled scores cannot be too much lower than labeled scores. The third restriction bounds how far unlabeled class 1 scores can fall below labeled class 1 scores, as measured by D_+. Because the true score distribution of the class 1 labels is unknown, we choose the tolerance relative to an observable benchmark, writing $\delta_1 = \lambda D_+(\widehat{G}, \widehat{F}_1)$, where $\widehat{G}$ is the empirical CDF of all unlabeled scores and $\lambda \in [0, 1]$. When $\lambda = 1$, this quantity measures the one-sided difference between the labeled class 1 scores and *all* of the unlabeled scores. If class 1 scores tend to be higher than class 0 scores among unlabeled observations, the gap should be smaller when restricted to class 1, motivating a value of λ at most 1. In the loan approval setting, this corresponds to the intuition that among declined applicants, defaulters should tend to have lower scores than repayers.

Domain knowledge about the historical decisions can inform a choice of λ, with larger values corresponding to more conservative rules. We recommend $\lambda = 0.8$ as an empirically motivated default (see Appendices D.2 and D.3 for experiments that examine its behavior). Appendix D.2 also provides a simple sufficient condition under which the population version of this restriction holds for this choice of λ.

Among unlabeled observations, class 1 scores tend to be higher than class 0 scores. The final restriction corresponds to the intuition that class 1 scores tend to be higher than class 0 scores among unlabeled observations. A value of $\gamma = 0$ reflects the intuition that at any score, the proportion of defaulters at or below this score should be at least as large as the corresponding proportion of repayers. Again, to account for finite-sample variability, we set $\gamma = N^{-1/3}$.

We note that the true labeling need not necessarily lie in the admissible set, as this is a sufficient but not necessary condition for the worst-case empirical type I error to be a conservative estimate. At the selected threshold, it suffices that the admissible set contain at least one labeling whose empirical type I risk is at least as large as that of the true labeling. Thus, the admissible set is constructed to give a conservative risk bound, not necessarily to contain the true labeling with high probability.

3 The NPSL Algorithm

At a high level, NPSL computes $\widehat{U}_0(t)$ (Equation (2.2)), then selects the smallest threshold at which this curve is at most α. Computation of $\widehat{U}_0(t)$ is conceptually simple, but direct evaluation can require searching over all 2^m possible labelings to determine if they are admissible. We now outline a practical way to compute Equation (2.2). We first observe that we may reparameterize the problem. Given a labeling z, we may define a function $C_z(u) = \sum_{i=1}^{m} z_i \mathbb{1}\{U_i < u\}$, so that C_z is a cumulative count of the number of unlabeled observations set to class 1. For each threshold $t \in \mathcal{T}$, we would like to compute the largest possible empirical type I error over all admissible labelings. We first fix k, the total number of unlabeled observations assigned to class 1. The empirical type I risk can then be explicitly written as

$$\widehat{R}_{0,z}(t) = \frac{X_{0,z}(t)}{N_{0,z}} = \frac{\sum_{S_i \in \mathcal{S}_0} \mathbb{1}\{S_i \geq t\} + \sum_{i=1}^{m} \mathbb{1}\{U_i \geq t\} - (k - C_z(t))}{n_0 + m - k}. \tag{3.1}$$

Crucially, the only dependence on z is through $C_z(t)$. We may thus optimize over the possible functions C_z rather than over the labelings themselves. We next observe that, when k and t are fixed, $\widehat{R}_{0,z}(t)$ is maximized when $C_z(t)$ is maximized. It thus suffices to find the largest attainable value of $C_z(t)$ for the fixed k and t, provided the value corresponds to an admissible labeling.

The key insight is that to check whether a proposed $C_z(t)$ is attainable, it suffices to check a single, explicitly constructable labeling. Furthermore, attainability is monotone in $C_z(t)$. If a value is

Algorithm 1 High level overview of NPSL

Ensure: Scoring function; calibration scores and observed labels; target error α; $\mathcal{K}, \delta_0, \delta_1, \gamma$.

1. Take the candidate thresholds to be the distinct calibration scores.
2. For each t and $k \in \mathcal{K}$, use binary search to find the largest attainable value of $C_z(t)$.
3. Compute the empirical type I error (Equation (3.1)) and take the maximum to obtain $\widehat{U}_0(t)$.
4. Select the smallest threshold $\widehat{t}$ satisfying $\widehat{U}_0(t) \leq \alpha$. If no admissible labeling or no threshold exists, return the trivial classifier, setting every observation to class 0.
5. Otherwise, return the classifier $\widehat{\phi}(x) = \mathbb{1}\{s(x) \geq \widehat{t}\}$.

attainable, every smaller feasible value is also attainable. Consequently, we may find the largest feasible value using binary search. Maximizing the resulting empirical risks over $k \in \mathcal{K}$ gives $\widehat{U}_0(t)$. Algorithm 1 summarizes this process, and Appendix B.1 provides the detailed version.

The following theorem states that NPSL computes $\widehat{U}_0(t)$ and $\widehat{t}$ given in Equations (2.2)–(2.3) exactly and gives the time complexity. As stated, Algorithm 1 computes the entire curve $\widehat{U}_0(t)$ for all t. If only $\widehat{t}$ is required, we may exploit the monotonicity of $\widehat{U}_0(t)$ to compute $\widehat{t}$ more efficiently. Appendix D.4 reports runtimes for both approaches, and the proof of the theorem is given in Appendix B.1.

Theorem 3.1 (Exactness and complexity)**.** *Let $N = n_0 + n_1 + m$ and suppose $\widehat{\mathcal{Z}} \neq \varnothing$. Algorithm 2, summarized in Algorithm 1, computes $\widehat{U}_0(t)$ for all t and returns the threshold in Equation* (2.3)*. The algorithm runs in $O(N^3 \log N)$ time. If only $\widehat{t}$ is required, it runs in $O(N^2 \log^2 N)$ time.*

4 Theoretical Results

The goal of this section is to establish that Algorithm 1 controls the type I error uniformly over all "plausible" completions of G, in a sense we make precise below. We first condition on the training data and any randomness that arises during training. We may thus treat the scoring function as fixed, and all probabilities and limits in this section are conditioned on this information. If the assumptions hold over almost all draws of the training data, the unconditional bounds in the statements follow by averaging over training randomness.

For each $y \in \{0, 1\}$, let F_y denote the distribution of scores conditional on $(A = 1, Y = y)$, and let G denote the distribution of scores conditional on $A = 0$. Additionally, we write $a = \mathbb{P}(A = 1, Y = 0), b = \mathbb{P}(A = 1, Y = 1)$, and $c = \mathbb{P}(A = 0)$. For simplicity, in this section, we treat the group sizes as fixed at $n_0 = aN$, $n_1 = bN$, and $m = cN$, ignoring rounding.

Fix constants p_- and p_+ with $0 < p_- \leq p_+ < 1$, and define the population counterpart of $\widehat{\mathcal{Z}}$ to be

$$\mathcal{G} := \left\{ (p, G_0, G_1) : p \in [p_-, p_+], \quad G = pG_0 + (1-p)G_1, \quad D_+(F_0, G_0) \leq \delta_0, \right. \\ \left. D_+(G_1, F_1) \leq \delta_1, \quad G_1(u) \leq G_0(u) \text{ for every } u \in \mathbb{R} \right\}. \tag{4.1}$$

It collects all triples $g = (p, G_0, G_1)$ that decompose G and satisfy the population versions of the constraints defining $\widehat{\mathcal{Z}}$. Throughout this section, we assume that $\mathcal{G}$ is nonempty. The bounds p_- and p_+ on the mixing fraction p exclude completions in which the declined class 0 (equivalently, class 1) fraction is too extreme. For each $g = (p, G_0, G_1) \in \mathcal{G}$, the associated type I error at threshold t is given by $R_0(t; g) = w(p)(1 - F_0(t)) + (1 - w(p))(1 - G_0(t))$ where the CDF evaluations are understood as left limits and $w(p) := \mathbb{P}(A = 1 \mid Y = 0) = a/(a + cp)$.

Our aim is therefore to show that the threshold $\widehat{t}$ produced by Algorithm 1 asymptotically controls $U_0\left(\widehat{t}\right) := \sup_{g \in \mathcal{G}} R_0\left(\widehat{t}; g\right)$, the worst-case type I error. Here, to account for sample randomness, we assume that Algorithm 1 is applied to an enlarged admissible label set, defined as

$$\widehat{\mathcal{Z}}(\varepsilon, \gamma) := \Big\{ z \in \{0,1\}^m : 1 \leq k(z) \leq m-1, \; p_- - \frac{1}{m} \leq \frac{m - k(z)}{m} \leq p_+ + \frac{1}{m}, \\ D_+(\widehat{F}_0, \widehat{G}_{0,z}) \leq \delta_0 + \varepsilon, \quad D_+(\widehat{G}_{1,z}, \widehat{F}_1) \leq \delta_1 + \varepsilon, \\ \widehat{G}_{1,z}(u) \leq \widehat{G}_{0,z}(u) + \gamma \quad \text{for every } u \in \mathbb{R} \Big\}, \tag{4.2}$$

which relaxes each constraint in Equation (2.4) by a small slack.

Theorem 4.1. *Assume that F_0, F_1, and G are subgaussian. Then,*

$$\mathbb{P}\left(U_0\left(\widehat{t}\right) \leq \alpha + \varepsilon + \frac{1}{N}\right) \geq 1 - C\left(e^{-c_1 n_0 \varepsilon^2} + e^{-c_2 m \gamma^2} + m e^{-c_3 m^2 \varepsilon^2}\right),$$

where C, c_1, c_2, c_3 are independent of N, ε, γ and $g \in \mathcal{G}$.

The proof is deferred to Appendix B. We remark that the subgaussian assumption on F_0, F_1, and G is imposed purely to obtain explicit exponential rates of concentration. It can be relaxed to weaker tail or moment conditions by truncation (see, for example, the techniques used in Tong et al. (2020)). In this case, the conclusion would still hold asymptotically but without an explicit rate. We consider this outside of the scope of the current paper and leave it for future work.

Theorem 4.1 shows that the threshold $\widehat{t}$ selected by Algorithm 1 applied to $\widehat{\mathcal{Z}}(\varepsilon, \gamma)$ controls the worst-case type I error $U_0(\widehat{t})$ at level α, up to a slack that vanishes as the sample size grows. Indeed, taking $\varepsilon = \varepsilon_N \asymp \frac{\log n_0}{\sqrt{n_0}}$ and $\gamma = \gamma_N \asymp \frac{\log m}{\sqrt{m}}$, so that the slack in $\widehat{\mathcal{Z}}(\varepsilon_N, \gamma_N)$ is just large enough to absorb sampling fluctuations while remaining asymptotically negligible, we obtain $U_0(\widehat{t}) \leq \alpha + o_p(1)$ as the sample size grows. In other words, the type I error is asymptotically controlled uniformly over all plausible completions in $\mathcal{G}$.

Proposition 4.2 (Convergence to the optimal robust threshold)**.** *Assume the conditions of Theorem 4.1 and that F_0, F_1, G are continuous. Suppose $U_0(t) = \alpha$ has a unique solution t^*_α, and that $\varepsilon = \gamma = N^{-1/3}$. Then $\widehat{t} \xrightarrow{\mathbb{P}} t^*_\alpha$.*

The theoretical results in this section motivate the recommended parameter settings in Section 2.3, but do not by themselves guarantee population type I error control, as the true parameters are generally unknown. Section 5.3 addresses this question by fixing the default values and evaluating their type I error risks across settings, including those that violate our assumptions.

5 Numerical Experiments

In this section, we evaluate the performance of NPSL using numerical experiments. For the first two experiments, we fit a logistic regression model using the labeled data in the training set. We independently generate the calibration set and labels, then determine whether their labels are observed. The first two experiments use $\alpha = 0.15$, and the label selection mechanism is *not* independent of the label conditional on the covariates. That is, the labels are missing not at random (MNAR). Many methods require this selection to be independent, corresponding to a missing at random (MAR) assumption. Full details are provided in Appendix C.1.

5.1 Analysis of Type I and Type II Risks

In the first experiment, we conduct 500 replications with total sample size $n \in \{1000, 10000\}$ of which $2/3$ are reserved for training and $1/3$ for calibration. We compare NPSL to five benchmarks. The *full label oracle* is the standard NP-umbrella algorithm with full access to all labels in the calibration set, setting a target violation rate of 10%. The *all labelings* rule is similar to NPSL, but considers every possible labeling rather than using our restricted set of plausible labelings. The *observed labels only* rule calibrates only on observed class 0 observations. The final two rules are commonly studied in the reject inference literature. Both implementations of these methods require the MAR assumption to be consistent, which is not satisfied in the current setting. *IPW* reweights the labeled observations to account for differences between the labeled and full population distributions. Finally, *fuzzy augmentation* takes the fitted score among labeled observations and uses this score to estimate the missing labels.

Figure 1(a–b) shows the distribution of risks. With the default parameters (Section 2.3), the results provide empirical evidence of population type I error control in this setting. For NPSL, the population type I error did not exceed 0.15 in any of the 500 replications at either sample size. While the observed labels only and all labelings rules also demonstrate evidence of type I error control, they achieve this control by being overly conservative, with much larger average type II errors. In

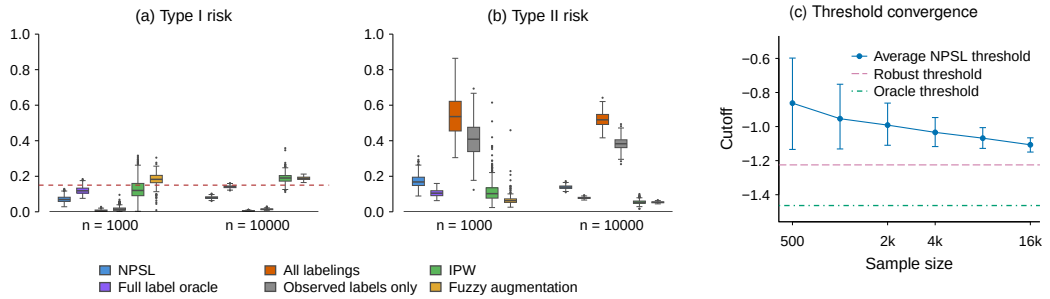

Figure 1: Results for synthetic data experiments. (a–b) Population type I and type II risks over 500 replications. The dashed line in (a) corresponds to $\alpha = 0.15$. (c) Average selected threshold over 500 replications, with error bars representing the 10th–90th percentiles. The two dashed lines show the optimal robust and oracle threshold corresponding to $\alpha = 0.15$ under the true distribution.

contrast, IPW and fuzzy augmentation have much higher violation rates, likely due to the violation of the MAR assumption. The corresponding MAR experiment is reported in Appendix D.1.

5.2 Convergence to the Optimal Robust Threshold

Proposition 4.2 states that under the given assumptions on parameters, the threshold returned by NPSL converges to the optimal robust threshold t^*_α. In this experiment, we investigate whether the same behavior occurs when using our suggested default parameters. For each replication, we refit the score and run NPSL on the calibration set. We compare the returned thresholds with the numerically computed $t^*_\alpha \approx -1.23$ and the oracle threshold computed from the population score (Appendix C.1). We conduct 500 replications for $n \in \{500, 1000, 2000, 4000, 8000, 16000\}$. Figure 1(c) shows the selected thresholds approaching t^*_α as the sample size increases. Additionally, we note that in this setting, the oracle threshold lies below t^*_α. This gap reflects the cost incurred by the missing labels.

5.3 Robustness of Recommended Default Parameters

In the third experiment, the goal is to evaluate whether the recommended parameter settings are robust across a range of settings. The default $\lambda = 0.8$ is intended as a practical compromise between type I error protection and power, and we examine the behavior of all the recommended defaults further in Appendix D.

Here, we simulate the scores directly and fix the NPSL parameters at the defaults, because the aim is to isolate the effect of the parameters on the calibration step. For each setting, we set $N = 1000$ and run 500 replications. As a baseline, we set $S \mid Y = 1 \sim N(0, 1)$, $S \mid Y = 0 \sim N(-1.5, 1)$, prevalence $\mathbb{P}(Y = 1) = 0.20$, and the same label selection mechanism as in the previous two experiments. Table 1 provides selected representative examples of settings that vary prevalence, class separation, score shape, missingness mechanism, and α. The full results are found in Appendix D.3.

Table 1: Representative results for robustness experiment under the recommended default parameters. Rows use $\alpha = 0.15$ except the final row, which uses $\alpha = 0.05$. Best R_1 is the minimum type II risk among threshold classifiers satisfying $R_0 \leq \alpha$. Violation rate is the percentage of replications with $R_0(\widehat{t}) > \alpha$.

Setting	Average R_0	Average R_1	Best Possible R_1	Violation Rate (%)
Baseline	0.089	0.439	0.321	0.0
Prevalence 0.8	0.053	0.552	0.321	0.0
Mean separation 1	0.111	0.588	0.515	0.0
Mean separation 2	0.067	0.310	0.168	0.0
Bimodal	0.074	0.587	0.321	0.0
Reversed selection	0.006	0.982	0.321	2.2
Target $\alpha = 0.05$	0.028	0.660	0.558	0.0

Table 2: Results for real-data illustration. Average type I risk, approval rate (%), and violation rate (%) with $\alpha = 0.10$, over 500 replications.

	MAR			MNAR		
Method	R_0	Approval (%)	Viol. Rate (%)	R_0	Approval (%)	Viol. Rate (%)
Observed Labels Only	0.021	6.2	0.0	0.027	7.6	0.0
NPSL	0.036	9.0	0.0	0.037	9.1	0.0

The primary observation is that the violation rate is low in each of these settings. Across all settings, the largest violation rate was 2.2%, indicating that the recommended parameters offer reasonable protection in a wide range of settings. We may also compare the average type II error to the best possible type II error, interpreting the increased type II error as a price to pay for the missing labels. Finally, we highlight the reversed selection row, which only achieves a violation rate of 2.2% by classifying nearly all points as class 0, resulting in an average type II error of 98.2%. In this scenario, we reverse the selection mechanism, making it more likely we observe labels with *low* scores and violating one of our structural assumptions. In this setting, the admissible set is usually empty, which causes NPSL to return a classifier that sets all observations to class 0. This demonstrates that violations of the selection assumptions can result in incompatibility with our restrictions, allowing NPSL to respond conservatively. However, we emphasize that when the admissible set is not empty, the returned threshold is too aggressive, and the type I error constraint is violated. Thus, this only provides partial protection against misspecification. NPSL should thus be used only when there is justification that the restrictions are reasonable for the application.

5.4 Real-data illustration: LendingClub

Studying the Neyman–Pearson framework is difficult on real data, because the population type I risk is unobservable. This problem is compounded by selective labels, as many of the outcomes are not observed. We thus illustrate NPSL using loan approval data from LendingClub, evaluating outcomes on the approved applicants. We first take the $78,435$ approved loans and reserve half of them for evaluation. Of the remaining loans, half are used to fit a logistic regression model using FICO and debt-to-income ratio. The other half is used for calibration, hiding labels artificially under MAR and MNAR settings. To reduce the number of distinct candidate thresholds, calibration and evaluation scores are rounded to three decimal places. Finally, we run 500 replications with target $\alpha = 0.10$, and further details are provided in Appendix C.2.

Based on Table 2, in this setting NPSL shows empirical evidence of type I risk control, with no violations of the error constraint. Additionally, NPSL approves more loans while maintaining this error control, increasing the approval by about 1.5 percentage points in the MNAR setting and 2.8 percentage points in the MAR setting when compared to using only observed labels for calibration.

6 Conclusion

We have introduced NPSL, which turns explicit assumptions on the score distributions of the unlabeled observations into a worst-case empirical type I risk curve, and returns a threshold that controls the worst-case empirical type I error. We have shown that this computation can be done in polynomial time, and that under the additional conditions of Section 4, the threshold approaches the optimal robust threshold. Synthetic experiments support the recommended default parameters in the settings studied, with low observed type I violation rates and lower type II risk than other benchmark methods that satisfy the population type I error constraint. An illustration using LendingClub data with artificially hidden labels on approved loans provides additional evidence for type I risk control.

There are several directions for future work. We first recall that our empirical risk bound can remain valid even when the true labeling lies outside the admissible set. Characterizing when this occurs, and when this corresponds to population risk control, could provide additional theoretical justification for the default parameters. Another direction is to identify broader classes of restrictions on the admissible set that permit efficient computation of the worst-case empirical type I risk.

AI STATEMENT

In this work, we used generative AI tools for help developing theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, implementing methods, cleaning and reformatting datasets, writing software code (including those to generate plots and synthetic data sets), creating drafts, proposing or refining hypotheses, supporting qualitative and thematic data analysis, interpreting results, designing and providing feedback on methodology or experiments, and editing the paper. We have not used generative AI tools for generating synthetic data sets directly or for formulating mathematical claims. Assisting with translation is not applicable to this work. We have reviewed all AI-assisted work, manually reviewing and testing the LLM-generated code for correctness and performing line-by-line audits of each proof. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Appendices C and D give the details of their corresponding experiments. R code contained in the supplementary material is provided to reproduce all figures and tables.

REFERENCES

Sidhika Balachandar, Nikhil Garg, and Emma Pierson. Domain constraints improve risk prediction when outcome data is missing. In *International Conference on Learning Representations*, 2024.

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Colin B. Begg and Robert A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215, 1983.

Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. MixMatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Cannon, J. Howse, D. Hush, and C. Scovel. Learning with the Neyman–Pearson and min-max criteria. Technical Report LA-UR-02-2951, Los Alamos National Laboratory, 2002.

Abhishek Chakrabortty, Jiarui Lu, T. Tony Cai, and Hongzhe Li. High dimensional M-estimation with missing outcomes: A semi-parametric framework, 2019. arXiv preprint arXiv:1911.11345.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning with selectively labeled data from multiple decision-makers. In *International Conference on Machine Learning*, pp. 8480 – 8519. PMLR, 2025.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Adrien Ehrhardt, Christophe Biernacki, Vincent Vandewalle, Philippe Heinrich, and Sébastien Beben. Reject inference methods in credit scoring. *Journal of Applied Statistics*, 48(13-15): 2734–2754, 2021.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

Charlie Frogner, Sebastian Claici, Edward Chien, and Justin Solomon. Incorporating unlabeled data into distributionally robust learning. *Journal of Machine Learning Research*, 22(56):1–46, 2021.

Nathan George. All Lending Club loan data. Kaggle dataset, 2019. URL `https://www.kaggle.com/datasets/wordsforthewise/lending-club`. Accepted and rejected applications, 2007–2018 Q4. Accessed September 21, 2026.

Josh Givens, Song Liu, and Henry W. J. Reeve. Density ratio estimation and Neyman Pearson classification with missing data. In *International Conference on Artificial Intelligence and Statistics*, pp. 8645–8681. PMLR, 2023.

Mohammadreza Mousavi Kalan, Yuyang Deng, Eitan J. Neugut, and Samory Kpotufe. Neyman–Pearson classification under both null and alternative distributions shift. In *International Conference on Learning Representations*, 2026.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Charles F. Manski. Anatomy of the selection problem. *The Journal of Human Resources*, 24(3): 343–360, 1989.

Wang Miao, Lan Liu, Yilin Li, Eric J. Tchetgen Tchetgen, and Zhi Geng. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM / IMS Journal of Data Science*, 1(2):5, 2024.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

Jiaming Qiu, Ying-Qi Zhao, and Yingye Zheng. Shift-adjusted Neyman–Pearson classifiers via single index modeling (SACSIM). *Biometrics*, 82(3):ujag138, 2026.

Ashesh Rambachan, Amanda Coston, and Edward H Kennedy. Robust design and evaluation of predictive algorithms under unobserved confounding. *Review of Economics and Statistics*, 2026.

Clayton Scott and Robert Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

Matteo Sesia, YX Rachel Wang, and Xin Tong. Adaptive conformal classification with noisy labels. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):796–815, 2025.

Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, and Pierre-Alexandre Mattei. Are labels informative in semi-supervised learning? Estimating and leveraging the missing-data mechanism. In *International Conference on Machine Learning*, pp. 32521–32539. PMLR, 2023.

BaoLuo Sun, Lan Liu, Wang Miao, Kathleen Wirth, James Robins, and Eric J. Tchetgen Tchetgen. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965–1983, 2018.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman–Pearson classification algorithms and NP receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018.

Xin Tong, Lucy Xia, Jiacheng Wang, and Yang Feng. Neyman-Pearson classification: parametrics and sample size requirement. *Journal of Machine Learning Research*, 21(12):1–48, 2020.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, Heidelberg, 2009.

Yixin Wang and José R. Zubizarreta. Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Dennis Wei. Decision-making under selective labels: Optimal finite-domain policies and beyond. In *International Conference on Machine Learning*, pp. 11035–11046. PMLR, 2021.

Shunan Yao, Bradley Rava, Xin Tong, and Gareth James. Asymmetric error control under imperfect supervision: a label-noise-adjusted Neyman–Pearson umbrella algorithm. *Journal of the American Statistical Association*, 118(543):1824–1836, 2023.

Yuqian Zhang, Abhishek Chakrabortty, and Jelena Bradic. Double robust semi-supervised inference for the mean: Selection bias under MAR labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, 12(3):2066–2159, 2023.

Qingyuan Zhao, Dylan S. Small, and Bhaswar B. Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

A Extended Related Work

Selective labels and semi-supervised learning. We follow the selective label model, also called reject inference when the application is in credit scoring. This literature studies how historical decisions can determine which labels are observed. Lakkaraju et al. (2017) formalize this problem and exploit variation across judges to evaluate classifiers. In loan application and credit scoring, Ehrhardt et al. (2021) analyze the assumptions underlying many common reject-inference methods. Chen et al. (2025) use multiple judges to bound classification risk and develop a cost-sensitive learning procedure. We focus on the Neyman–Pearson objective, studying what guarantees on the type I error can still be provided when some labels are unobserved. We additionally study two of the methods analyzed in Ehrhardt et al. (2021), inverse probability weighting and fuzzy augmentation, adapting them to the NP framework and comparing them to NPSL in numerical simulations. Wei (2021) studies the missing label problem in the online setting, constructing classifiers that balance the cost of observing labels against future utility.

Error bounds under ambiguity. Partial identification characterizes what can be learned when missing information results in non-identifiability of the quantities of interest (Manski, 1989). In selective label settings, Rambachan et al. (2026) derive sharp bounds on error rates under a sensitivity model. We instead impose joint restrictions on the score distributions directly. Thus, hypothetical labelings that may be compatible under a pointwise sensitivity model may be excluded from the admissible set under our assumptions. In this way, the domain knowledge used to construct our admissible set can permit more aggressive classifiers by ruling out these implausible labelings.

Distributionally robust optimization also considers worst-case distributions within an admissible set. Frogner et al. (2021) also use unlabeled data to restrict this set and obtain better guarantees on the error risks. We instead optimize over possible labelings directly, choosing the admissible set specifically to balance informative bounds with computational tractability. Under suitable assumptions, the selective label setting can induce a covariate shift between the labeled and unlabeled data. In this setting, Tibshirani et al. (2019) develop weighted conformal prediction to provide guarantees on prediction set coverage. The setting used in NPSL bypasses the specific missing at random assumptions required for covariate shift. Related work in conformal prediction considers label noise, constructing prediction sets with coverage guarantees (Sesia et al., 2025).

Neyman–Pearson classification. Our methodology uses the split-sample calibration framework from the NP-umbrella algorithm (Tong et al., 2018) and risk-controlling prediction sets (Bates et al., 2021). Recent extensions involve imperfect or incomplete data. Yao et al. (2023) consider a corrupted-label model, while Givens et al. (2023) study density-ratio estimation and NP classification with missing features. Qiu et al. (2026) also develop NP classifiers under distribution shift using a semiparametric model, which our analysis does not require. Similarly, Kalan et al. (2026) study NP transfer learning with distributional shift, though their setting does not incorporate missing labels.

B Proofs

B.1 Proof of Theorem 3.1

In this section, we prove Theorem 3.1 in three steps. First, we give a precise version of Algorithm 1. Second, we prove that it computes the curve and rule in Equations (2.2)–(2.3) exactly. Finally, we establish its time complexity.

Algorithm 1 outlines the high-level details of NPSL. In this section, we provide an explicit algorithm that provides the details of each computation. For each $k \in \mathcal{K}$, define $\widehat{\mathcal{Z}}_k := \{z \in \widehat{\mathcal{Z}} : k(z) = k\}$. For a fixed k and t, write

$$C(t, z) = \sum_{i=1}^{m} z_i \mathbb{1}\{U_i < t\},$$

the number of unlabeled observations assigned to class 1 below t. As described in Section 3, for fixed k with $\widehat{\mathcal{Z}}_k \neq \varnothing$, maximizing the type I error over admissible labelings with $k(z) = k$ corresponds to finding $C^*(t) = \max_{z \in \widehat{\mathcal{Z}}_k} C(t, z)$, with the dependence on k suppressed.

We now introduce some useful notation. Let $q_1 < \cdots < q_L$ be the L distinct scores in $\mathcal{S}_0 \cup \mathcal{S}_1 \cup \{U_1, \ldots, U_m\}$. Define

$$N_\ell := \sum_{i=1}^{m} \mathbb{1}\{U_i \le q_\ell\} \qquad \text{and} \qquad C_\ell(z) := \sum_{i=1}^{m} z_i \mathbb{1}\{U_i \le q_\ell\},$$

with $N_0 = C_0 = 0$. Thus, N_ℓ represents the total number of unlabeled scores at or below q_ℓ and $C_\ell(z)$ represents the number of these scores set to class 1 under labeling z. Note that $C(t, z)$ cannot change arbitrarily, but can change only at the observed scores q_ℓ. It is easier to work with the relative position of these change points rather than the continuous threshold t. To that end, we define $b(t) := |\{\ell \mid q_\ell < t\}|$, the number of distinct scores below the threshold t, and note that for any threshold t, $C(t, z) = C_{b(t)}(z)$. Finally, write $d_\ell = N_\ell - N_{\ell-1}$, so d_ℓ represents the number of unlabeled observations with a score of q_ℓ.

Let $C := (C_0(z), \ldots, C_L(z))$ be a vector consisting of the cumulative counts of class 1 observations induced by a labeling z, where we suppress the dependence on z for clarity. Once k is fixed, the relevant empirical risks depend on z only through C. We may therefore replace optimization over admissible labelings z by optimization over their corresponding feasible cumulative counts C. The following elementary lemma makes the correspondence between labelings z and vectors precise.

Lemma B.1. *Fix $k \in \{0, \ldots, m\}$. A vector of integers $C = (C_0, \ldots, C_L)$ corresponds to some $z \in \{0, 1\}^m$ with $k(z) = k$ if and only if, for every $\ell = 1, \ldots, L$,*

$$C_0 = 0, \qquad C_L = k \qquad \textit{and} \qquad 0 \le C_\ell - C_{\ell-1} \le d_\ell.$$

Proof. If $C = C(z)$ for a labeling z with $k(z) = k$, then

$$C_\ell - C_{\ell-1} = \sum_{i=1}^{m} z_i \mathbb{1}\{U_i = q_\ell\},$$

which is an integer between 0 and d_ℓ. Additionally, $C_0 = 0$ and $C_L = \sum_{i=1}^{m} z_i = k$.

Conversely, suppose C satisfies the stated conditions. For each $\ell = 1, \ldots, L$, assign class 1 to exactly $C_\ell - C_{\ell-1}$ of the d_ℓ unlabeled observations with score q_ℓ, and class 0 to the remaining observations. The fact that $0 \le C_\ell - C_{\ell-1} \le d_\ell$ ensures that these assignments are possible. The resulting labeling satisfies

$$C_\ell(z) = \sum_{j=1}^{\ell} (C_j - C_{j-1}) = C_\ell$$

for every ℓ, and thus $k(z) = C_L = k$. □

We now fix a k and a threshold t, and write $b = b(t)$. To compute C^*, we first determine which values of C_b are possible. To that end, let v denote a candidate value of $C_b(z)$. A simple counting argument gives that, with no other restrictions, v can be any integer satisfying $a_{k,b} \le v \le f_{k,b}$, where

$$a_{k,b} = \max\{0, k - (m - N_b)\} \qquad \text{and} \qquad f_{k,b} = \min\{k, N_b\}.$$

Of course, the restrictions in Equation (2.4) further limit the possible values of v. We thus rewrite Equation (2.4) in terms of the cumulative counts C. Write

$$D_0(z) := \sum_{\ell=1}^{L-1} (q_{\ell+1} - q_\ell) \left(\widehat{F}_0(q_\ell) - \frac{N_\ell - C_\ell}{m - k} \right)_+,$$
$$D_1(z) := \sum_{\ell=1}^{L-1} (q_{\ell+1} - q_\ell) \left(\frac{C_\ell}{k} - \widehat{F}_1(q_\ell) \right)_+,$$

noting that these quantities also depend on z only through C. We also write

$$h'_\ell := \left\lfloor \frac{kN_\ell}{m} + \frac{k(m-k)}{m}\gamma \right\rfloor \qquad \text{and} \qquad h_\ell := \min\{N_\ell, k, h'_\ell\},$$

with $h_0 = 0$. Then it is easy to check that $z \in \widehat{\mathcal{Z}}$ if and only if $C(z)$ satisfies

$$C_\ell(z) \le h_\ell \text{ for each } \ell = 1, \dots, L, \qquad D_0(z) \le \delta_0, \qquad \text{and} \qquad D_1(z) \le \delta_1. \tag{B.1}$$

Finally, for an integer v satisfying $a_{k,b} \le v \le f_{k,b}$, we must determine if there exists a C such that $C_b = v$, $C_L = k$, and C satisfies each of the restrictions in Equation (B.1). Two reductions make determining this easier. First, this property is monotone, in the sense that if such a C exists for v, there also exists one for every integer v' with $a_{k,b} \le v' \le v$. Second, it suffices to check the restrictions for a specifically constructed C, rather than all possible C. Define the coordinate-wise smallest C satisfying $C_b = v$, $C_L = k$ to be

$$C_\ell^{\min}(v) = \begin{cases} \max\{0,\ v - (N_b - N_\ell)\}, & \ell \le b, \\ \max\{v,\ k - (m - N_\ell)\}, & \ell > b \end{cases}. \tag{B.2}$$

Strictly speaking, $C_\ell^{\min}(v)$ depends on k and b, but we suppress this dependence when the variables are fixed. The following theorem makes the previous two statements precise.

Theorem B.2. *Fix $k \in \mathcal{K}$ and $b \in \{0, \dots, L\}$. For every integer v satisfying $a_{k,b} \le v \le f_{k,b}$, there exists a $z \in \widehat{\mathcal{Z}}_k$ with $C_b(z) = v$ if and only if $C^{\min}(v)$ satisfies Equation* (B.1)*. Additionally, if $C^{\min}(v)$ satisfies Equation* (B.1)*, then $C^{\min}(v')$ also satisfies Equation* (B.1) *for every integer $a_{k,b} \le v' \le v$.*

We now have the preliminaries required to describe the NPSL algorithm in full detail. For each k and b, find the largest possible value of v such that there exists a C such that $C_b = v$, $C_L = k$, and C satisfies each of the restrictions in Equation (B.1). By Theorem B.2, we may use a binary search over $a_{k,b} \le v \le f_{k,b}$. Once v is found, compute the corresponding empirical risk $X_0(t)/N_{0,k}$ and maximize over feasible k to obtain $\widehat{U}_0(t)$ in Equation (2.2).

Algorithm 2 NPSL

Require: A score function s; $\mathcal{S}_0, \mathcal{S}_1, U$; target type I error α; $\mathcal{K}$; parameters $\delta_0, \delta_1, \gamma$.
Ensure: Exact empirical risk curve and classifier.

$(q_1, \dots, q_L) \leftarrow \text{sort}(\text{unique}(\mathcal{S}_0 \cup \mathcal{S}_1 \cup \{U_1, \dots, U_m\}))$; $\mathcal{T} \leftarrow \{q_1, \dots, q_L\}$
$N_\ell \leftarrow \sum_{i=1}^m \mathbb{1}\{U_i \le q_\ell\}$, $\ell = 1, \dots, L$; $N_0 \leftarrow 0$
$b(t) \leftarrow \#\{\ell : q_\ell < t\}$
Initialize $\widehat{U}_0(t) \leftarrow -\infty$ for all $t \in \mathcal{T}$.
for each $k \in \mathcal{K}$ **do**
    $N_{0,k} \leftarrow n_0 + m - k$
    for each $t \in \mathcal{T}$ **do**
        $b \leftarrow b(t)$
        $v^* \leftarrow$ largest integer $v \in [a_{k,b}, f_{k,b}]$ such that $C^{\min}(v)$ satisfies Equation (B.1); if none exists, continue to the next t.
        $X_0(t) \leftarrow \sum_{S_i \in \mathcal{S}_0} \mathbb{1}\{S_i \ge t\} + \sum_{i=1}^m \mathbb{1}\{U_i \ge t\} - (k - v^*)$
        $\widehat{U}_0(t) \leftarrow \max\{\widehat{U}_0(t), X_0(t)/N_{0,k}\}$
    end for
end for
if $\widehat{U}_0(t) = -\infty$ for all $t \in \mathcal{T}$ **then**
    $\widehat{U}_0(t) \leftarrow 1$ for all $t \in \mathcal{T}$
end if
$\mathcal{T}_\alpha \leftarrow \{t \in \mathcal{T} : \widehat{U}_0(t) \le \alpha\}$
if $\mathcal{T}_\alpha = \varnothing$ **then**
    $\widehat{t} \leftarrow +\infty$
    return $\widehat{\phi}(x) \equiv 0$ and computed curves.
else
    $\widehat{t} \leftarrow \min \mathcal{T}_\alpha$ and **return** $\widehat{\phi}(x) = \mathbb{1}\{s(x) \ge \widehat{t}\}$ together with the computed curves.
end if

Threshold-only computation. The above algorithm computes the full curve in $O(N^3 \log N)$ time. Using binary search over t gives the $O(N^2 \log^2 N)$ variant found in Remark B.5.

B.2 Equivalence with exhaustive labeling optimization

Proof of Theorem B.2. We first show that $C^{\min}(v)$ corresponds to a valid labeling. First, fix any integer v satisfying $a_{k,b} \le v \le f_{k,b}$. By definition, for $\ell \le b$, $C_\ell^{\min} = \max\{0, v - N_b + N_\ell\}$. Noting that $N_\ell = N_{\ell-1} + d_\ell$ and considering cases, we thus have

$$0 \le C_\ell^{\min} - C_{\ell-1}^{\min} \le N_\ell - N_{\ell-1} = d_\ell. \tag{B.3}$$

For $\ell > b$, by definition, $C_\ell^{\min} = \max\{v, k - m + N_\ell\}$. A similar case analysis shows that Equation (B.3) is also satisfied when $\ell > b$. Additionally, since $N_L = m$ and $v \le k$, we have $C_L^{\min} = k$. Finally, the fact that $v \le N_b$ gives $C_0^{\min} = 0$, and by construction, $C_b^{\min} = v$. By Lemma B.1, $C^{\min}$ corresponds to a valid labeling.

Now let C be any cumulative count with $C_b = v$ and $C_L = k$. For $\ell \le b$, the total increase from ℓ through b is at most $N_b - N_\ell$, so

$$\begin{aligned} v - C_\ell = C_b - C_\ell &\le N_b - N_\ell \\ C_\ell &\ge v - (N_b - N_\ell) \end{aligned}$$

and $C_\ell \ge 0$. Thus $C_\ell \ge C_\ell^{\min}$ for $\ell \le b$. Similarly, for $\ell > b$, monotonicity gives $C_\ell \ge v$, while the total increase from ℓ through L is at most $N_L - N_\ell$, so

$$\begin{aligned} k - C_\ell = C_L - C_\ell &\le m - N_\ell \\ C_\ell &\ge k - (m - N_\ell). \end{aligned}$$

Thus, $C_\ell \ge C_\ell^{\min}$ for $\ell > b$, and we have

$$C_\ell \ge C_\ell^{\min} \text{ for all } \ell. \tag{B.4}$$

We now prove the first statement in the theorem, that there exists a feasible cumulative count C with $C_b = v$ if and only if $C^{\min}(v)$ satisfies Equation (B.1). The backward direction is trivial. For the forward direction, suppose some $z \in \widehat{\mathcal{Z}}_k$ has $C_b = v$. By Equation (B.4), $C_\ell^{\min} \le C_\ell$ for every ℓ. Since $C_\ell \le h_\ell$ for every ℓ,

$$C_\ell^{\min}(v) \le C_\ell \le h_\ell.$$

Noting that D_0 is nondecreasing with respect to C_ℓ for each ℓ, we have

$$D_0\left(C^{\min}(v)\right) \le D_0(C) \le \delta_0.$$

By a similar argument, $D_1\left(C^{\min}(v)\right) \le \delta_1$. Thus, $C^{\min}(v)$ satisfies Equation (B.1).

We now prove the second statement in the theorem, that $C^{\min}(v)$ satisfying Equation (B.1) implies $C^{\min}(v')$ also satisfies Equation (B.1) for any $a_{k,b} \le v' \le v$. First, note that $C_\ell^{\min}(v)$ is nondecreasing in v for each ℓ. By the monotonicity of the restrictions in Equation (B.1) established above, we have that if $C^{\min}(v)$ satisfies the restrictions, then $C^{\min}(v')$ must also satisfy the restrictions. □

Proposition B.3 (Exactness of NPSL)**.** *If $\widehat{\mathcal{Z}} \neq \varnothing$, Algorithm 2 computes Equation* (2.2) *exactly for each $t \in \mathcal{T}$.*

Proof. By Lemma B.1, for each $k \in \mathcal{K}$, it suffices to consider cumulative paths C that satisfy Equation (B.1) and have $C_L = k$. By Theorem B.2, when $\widehat{\mathcal{Z}}_k \neq \varnothing$, the possible values of v are exactly $\{a_{k,b}, \dots, C^*(t)\}$; otherwise, no value of v is feasible. As shown in Section 3, for fixed k and t, the empirical type I error $\widehat{R}_{0,z}(t)$ is increasing in v. Thus, its maximum over a nonempty $\widehat{\mathcal{Z}}_k$ is attained at the largest feasible value of v. Maximizing these empirical risks over feasible $k \in \mathcal{K}$ gives $\widehat{U}_0(t)$ in Equation (2.2). □

B.3 Computational complexity

Let

$$N = n_0 + n_1 + m, \qquad K = |\mathcal{K}|.$$

Proposition B.4 (Detailed operation count)**.** *Algorithm 2 runs in $O(N^3 \log N)$ time.*

Proof. We first sort the N scores in $O(N \log N)$ time. Computing all N_ℓ takes $O(N)$ time using a single scan through the sorted scores. Now fix $k \in \mathcal{K}$. For each t, the interval $[a_{k,b}, f_{k,b}]$ contains at most $m+1$ values. By Theorem B.2, we may use a binary search to find v^*, by testing $O\{1 + \log(m+1)\}$ values. Testing a single value of v requires constructing $C^{\min}(v)$ and checking whether it corresponds to an admissible labeling, which takes $O(L)$ time. Repeating this for each of the at most KN pairs (k,t) takes $O\{KNL[1+\log(m+1)]\}$ time.

Given v^*, computing $X_0(t)$ from the precomputed counts and updating $\widehat{U}_0(t)$ takes $O(1)$ time per (k,t) pair. These updates therefore take $O(KN)$ time overall. Since $K \le m \le N$ and $L \le N$, the total cost is $O(N \log N + KNL[1+\log(m+1)] + KN + N) = O(N^3 \log N)$. □

Remark B.5 (Threshold-only computation). If only the selected threshold is required, we may replace the t-loops in Algorithms 1 and 2 with binary search over the L candidate thresholds. At each t, evaluate $\widehat{U}_0(t)$ by maximizing over all feasible $k \in \mathcal{K}$. If no feasible k is found, return $\widehat{\phi}(x) \equiv 0$ because $\widehat{\mathcal{Z}} = \varnothing$. Otherwise, $\widehat{U}_0(t)$ is nonincreasing in t, so binary search finds the smallest candidate satisfying $\widehat{U}_0(t) \le \alpha$. After $O(N \log N + KL)$ preprocessing, each evaluation costs $O(KL[1+\log(m+1)])$, and the search requires $O(\log(L+1))$ evaluations. The total cost is therefore

$$O\big(N \log N + KL + KL[1+\log(m+1)] \log(L+1)\big) = O(N^2 \log^2 N),$$

where we use $K \le m \le N$ and $L \le N$.

B.4 Proof of Theorem 4.1

In this section, we present the proof of Theorem 4.1. This proof relies on an important lemma which bridges the completion g and labeling z.

Lemma B.6. *Let $p \in (0,1)$ be a number such that $\min(p, 1-p)\, m \ge 2$, and (G_0, G_1) be any pair of distributions on $\mathbb{R}$ satisfying*

$$G = pG_0 + (1-p)G_1. \tag{B.5}$$

Moreover, let $U_1, \dots, U_m$ be i.i.d. samples generated from G, and $\widehat{G} = m^{-1}\sum_{i=1}^m \delta_{U_i}$ be the associated empirical distribution. For a labeling $z \in \{0,1\}^m$, set $m_{j,z} = \sum_{i=1}^m \mathbb{1}\{z_i = j\}$ and $\widehat{G}_{j,z} := m_{j,z}^{-1}\sum_{i:z_i=j}\delta_{U_i}$ for any $j \in \{0,1\}$. Then there exists a labeling $z \in \{0,1\}^m$ with $m_{0,z} = \lceil pm \rceil$ such that the following hold.

(i) With $\widehat{p} := m_{0,z}/m$,

$$\widehat{p}\,\widehat{G}_{0,z} + (1-\widehat{p})\,\widehat{G}_{1,z} = \widehat{G}, \qquad |\widehat{p} - p| \le \frac{1}{m}.$$

(ii) For every $\tau \ge 1$,

$$D(\widehat{G}_{0,z}, G_0) + D(\widehat{G}_{1,z}, G_1) \le \frac{1}{p(1-p)}\left[D(G, \widehat{G}) + \frac{8}{m^{1-1/\tau}}\left(\frac{\sum_{i=1}^m |U_i|^\tau}{m}\right)^{1/\tau}\right]. \tag{B.6}$$

(iii) Writing $\|F - H\|_\infty := \sup_{u\in\mathbb{R}} |F(u) - H(u)|$ for the Kolmogorov distance,

$$\|\widehat{G}_{0,z} - G_0\|_\infty + \|\widehat{G}_{1,z} - G_1\|_\infty \le \frac{1}{p(1-p)}\left[\|G - \widehat{G}\|_\infty + \frac{4}{m}\right]. \tag{B.7}$$

The proof of this lemma relies on a fundamental result in transportation theory stated below.

Fact B.1 (Villani (2009), Theorem 4.1)**.** *Let P, Q be two distributions defined on $\mathbb{R}$, and $D(P,Q)$ be their Wasserstein distance, i.e.,*

$$D(P,Q) = \int_{\mathbb{R}} |P(u) - Q(u)| du.$$

Moreover, denote $\Xi(P, Q)$ as the set of all couplings, i.e., joint measures defined on $\mathbb{R}^2$ whose marginal distributions are P and Q, respectively. Then,

$$D(P, Q) = \inf_{\xi \in \Xi(P,Q)} \int |x - y| \, d\xi(x, y), \tag{B.8}$$

Moreover, the infimum can be attained.

Next, we proceed to present the proof of the lemma.

Proof of Lemma B.6. The proof proceeds in three steps. We first construct discrete distributions $(\widetilde{G}_0, \widetilde{G}_1)$ supported on the sample that are close to (G_0, G_1) in Wasserstein distance (Step 1). We then round discrete $(\widetilde{G}_0, \widetilde{G}_1)$ to a hard labeling z with induced distributions $(\widehat{G}_{0,z}, \widehat{G}_{1,z})$ (Step 2). Then, the theorem follows by the triangle inequality (Step 3).

Step 1. We will first construct weights $\omega_1, \ldots, \omega_m$ and then define

$$\widetilde{G}_0 := \frac{1}{pm} \sum_{i=1}^{m} \omega_i \delta_{U_i}, \qquad \widetilde{G}_1 := \frac{1}{(1-p)m} \sum_{i=1}^{m} (1 - \omega_i) \delta_{U_i}. \tag{B.9}$$

Note that for every Borel set $B \subset \mathbb{R}$ with $G(B) = 0$, we get $G_0(B) = 0$. So $G_0 \ll G$. So, we can define

$$\eta = p \frac{dG_0}{dG}, \tag{B.10}$$

where $\frac{dG_0}{dG}$ is the Radon-Nikodym derivative. Moreover, we claim that $\eta \leq 1$ G-almost surely. To see this, assume $\eta > 1$ on a set B with $G(B) > 0$. Then, $pG_0(B) = \int_B \eta \, dG > G(B)$, contradicting $pG_0 \leq G$. Next, in view of Equation (B.5), we conclude that

$$1 - \eta = (1 - p) \frac{dG_1}{dG}.$$

Now, let $V \sim \mathrm{Unif}(0, 1)$ and let $\xi \in \Xi(\widehat{G}, G)$ be the quantile coupling, that is, the joint distribution of $\big(\widehat{G}^{-1}(V), G^{-1}(V)\big)$, where $F^{-1}(s) := \inf\{x : F(x) \geq s\}$ denotes the quantile function. The quantile coupling attains the minimum in the definition of the Wasserstein distance (Fact B.1), so $\int |u - v| \, d\xi(u, v) = D(\widehat{G}, G)$. It is clear that ξ is supported on $\{U_1, \ldots, U_m\} \times \mathbb{R}$. Thus, for each $i \in [m]$, we define the conditional distribution

$$\xi_i(B) := \xi(\{U_i\} \times B) / \widehat{G}(\{U_i\}) \tag{B.11}$$

for every Borel set B and, finally,

$$\omega_i := \int_{\mathbb{R}} \eta(v) \, d\xi_i(v) \in [0, 1], \tag{B.12}$$

where the containment in $[0, 1]$ holds because $0 \leq \eta \leq 1$. Furthermore, it is easy to see that

$$\frac{1}{m} \sum_{i=1}^{m} \omega_i = \int \eta(v) d\xi(u, v) = \int \eta dG = p. \tag{B.13}$$

Thus, $\widetilde{G}_0$ and $\widetilde{G}_1$ in Equation (B.9) are indeed probability distributions defined on $\{U_1, \ldots, U_m\}$ and

$$p\widetilde{G}_0 + (1 - p)\widetilde{G}_1 = m^{-1} \sum_i \delta_{U_i} = \widehat{G}.$$

Next, we show that

$$D(G_0, \widetilde{G}_0) \leq \frac{1}{p} D(G, \widehat{G}), \qquad D(G_1, \widetilde{G}_1) \leq \frac{1}{1 - p} D(G, \widehat{G}). \tag{B.14}$$

Define the probability measure $\widetilde{\xi}$ on $\mathbb{R}^2$ by

$$d\widetilde{\xi}(u,v) := \frac{1}{p}\eta(v)\, d\xi(u,v).$$

It is easy to verify that, for every Borel set $B \subset \mathbb{R}$,

$$\widetilde{\xi}(\mathbb{R}\times B) = \frac{1}{p}\int_B \eta dG = \int_B dG_0 = G_0(B),$$

where the second equality is given by Equation (B.10), and

$$\widetilde{\xi}(B\times\mathbb{R}) = \frac{1}{pm}\sum_{i:U_i\in B}\int \eta\, d\xi_i = \frac{1}{pm}\sum_{i:U_i\in B}\omega_i = \widetilde{G}_0(B),$$

where we use the definitions of ξ_i in Equation (B.11), ω_i in Equation (B.12), and $\widetilde{G}_0$ in Equation (B.9) in the above transitions. Thus, $\widetilde{\xi} \in \Xi(\widetilde{G}_0, G_0)$. So, given $\eta \le 1$,

$$\begin{aligned} D(\widetilde{G}_0, G_0) &\le \int |u-v|\, d\widetilde{\xi}(u,v) = \frac{1}{p}\int |u-v|\eta(v) d\xi(u,v) \\ &\le \frac{1}{p}\int |u-v| d\xi(u,v) = \frac{D(G,\widehat{G})}{p}. \end{aligned}$$

Similar reasoning yields $D(G_1, \widetilde{G}_1) \le \frac{1}{1-p} D(G, \widehat{G})$, establishing Equation (B.14).

In addition, we show a Kolmogorov-type analogue of Equation (B.14). Note that for any t,

$$\widetilde{G}_0(t) = \frac{1}{pm}\sum_{i:U_i\le t}\omega_i = \frac{1}{p}\int \mathbb{1}\{u\le t\}\eta(v) d\xi(u,v),$$

and

$$G_0(t) = \frac{1}{p}\int \mathbb{1}\{v\le t\}\eta(v) d\xi(u,v).$$

Thus,

$$\left|\widetilde{G}_0(t) - G_0(t)\right| \le \frac{1}{p}\int |\mathbb{1}\{u\le t\} - \mathbb{1}\{v\le t\}|\, d\xi(u,v).$$

On the other hand, using the fact that $|u-v| = \int_{\mathbb{R}} |\mathbb{1}\{u\le t\} - \mathbb{1}\{v\le t\}|\, dt$ and invoking Fubini's theorem, we obtain

$$\int \left|G(t) - \widehat{G}(t)\right| dt = \int |v-u| d\xi(u,v) = \int_{\mathbb{R}}\int |\mathbb{1}\{u\le t\} - \mathbb{1}\{v\le t\}|\, d\xi(u,v) dt, \quad \text{(B.15)}$$

where the first equality is implied by the definition of ξ and Fact B.1. Next, since

$$\left|G(t) - \widehat{G}(t)\right| = \left|\int \mathbb{1}\{u\le t\} - \mathbb{1}\{v\le t\} d\xi(u,v)\right| \le \int |\mathbb{1}\{u\le t\} - \mathbb{1}\{v\le t\}|\, d\xi(u,v),$$

for every t, the above equation implies the above inequality can be taken as equality. Thus,

$$\left|\widetilde{G}_0(t) - G_0(t)\right| \le \frac{1}{p}\left|G(t) - \widehat{G}(t)\right|. \quad \text{(B.16)}$$

Similarly, $\left|\widetilde{G}_1(t) - G_1(t)\right| \le \frac{1}{1-p}\left|G(t) - \widehat{G}(t)\right|$.

Step 2. We now round the weights $\omega_1, \ldots, \omega_m$ constructed in Step 1 to a labeling $z \in \{0,1\}^m$. Let $U_{(1)} \le \cdots \le U_{(m)}$ denote the order statistics of the sample and $\omega_{(k)}$ be the weight associated with $U_{(k)}$. Define

$$S_0 := 0, \qquad S_k := \sum_{i=1}^{k}\omega_{(i)}, \quad k = 1, \ldots, m.$$

Since $\sum_{i=1}^{m} \omega_i = pm$ by Equation (B.13), it is immediate that $S_m = pm$. Since $\omega_{(k)} \in [0,1]$, we have $\lceil S_k \rceil - \lceil S_{k-1} \rceil \in \{0,1\}$ for every k. Assign the label to $U_{(k)}$ by setting

$$z_{(k)} := 0 \quad \text{if and only if} \quad \lceil S_k \rceil = \lceil S_{k-1} \rceil + 1.$$

Now, we can verify Part (i). Writing $M_k := \sum_{i=1}^{k} \mathbb{1}\{z_{(i)} = 0\}$, the telescoping sum $\lceil S_k \rceil = \sum_{i=1}^{k} (\lceil S_i \rceil - \lceil S_{i-1} \rceil)$ yields $M_k = \lceil S_k \rceil$, and thus $0 \le M_k - S_k < 1$ for all k. In particular $m_0 = M_m = \lceil pm \rceil$. So $|\widehat{p} - p| \le 1/m$, and $m_1 = m - \lceil pm \rceil \ge (1-p)m - 1 \ge 1$ by the assumption $(1-p)m \ge 2$. Hence, Part (i) is immediate.

Next, we claim that for every $t \in \mathbb{R}$,

$$\left|\widehat{G}_{0,z}(t) - \widetilde{G}_0(t)\right| \le \frac{2}{m_0}, \qquad \left|\widehat{G}_{1,z}(t) - \widetilde{G}_1(t)\right| \le \frac{2}{m_1}. \tag{B.17}$$

Indeed, for an arbitrary $t \in \mathbb{R}$, we can write

$$\widehat{G}_{0,z}(t) = \frac{M_{m\widehat{G}(t)}}{m_0}, \qquad \widetilde{G}_0(t) = \frac{S_{m\widehat{G}(t)}}{pm}.$$

Therefore,

$$\left|\widehat{G}_{0,z}(t) - \widetilde{G}_0(t)\right| \le \frac{\left|M_{m\widehat{G}(t)} - S_{m\widehat{G}(t)}\right|}{m_0} + S_{m\widehat{G}(t)} \left|\frac{1}{m_0} - \frac{1}{pm}\right| \le \frac{1}{m_0} + \frac{S_{m\widehat{G}(t)}|m_0 - pm|}{m_0\, pm} \le \frac{2}{m_0},$$

where the second inequality is implied by $0 \le M_k - S_k < 1$ for all k, and the last inequality is given by $m_0 = \lceil pm \rceil$. Similar reasoning yields the second half of Equation (B.17). Since $\widehat{G}_{0,z}, \widehat{G}_{1,z}, \widetilde{G}_0$ and $\widetilde{G}_1$ are all supported on $\{U_1, \ldots, U_m\}$, the bound in Equation (B.17) implies

$$D(\widehat{G}_{0,z}, \widetilde{G}_0) + D(\widehat{G}_{1,z}, \widetilde{G}_1) \le \left(\frac{2}{m_0} + \frac{2}{m_1}\right)\left(U_{(m)} - U_{(1)}\right) \le \frac{4}{p(1-p)} \frac{U_{(m)} - U_{(1)}}{m},$$

where the last step uses $m_0 = \lceil pm \rceil$ and thus $m_1 \ge (1-p)m - 1 \ge (1-p)m/2$. It remains to bound the sample range. For any $\tau \ge 1$, the relations

$$U_{(m)} - U_{(1)} \le 2 \left(\max_{1 \le i \le m} |U_i|^\tau\right)^{1/\tau} \le 2 \left(\sum_{i=1}^{m} |U_i|^\tau\right)^{1/\tau} = 2m^{1/\tau} \left(\frac{\sum_{i=1}^{m} |U_i|^\tau}{m}\right)^{1/\tau}$$

yield

$$D(\widehat{G}_{0,z}, \widetilde{G}_0) + D(\widehat{G}_{1,z}, \widetilde{G}_1) \le \frac{8}{m^{1-1/\tau} p(1-p)} \left(\frac{\sum_{i=1}^{m} |U_i|^\tau}{m}\right)^{1/\tau}. \tag{B.18}$$

Step 3: combining the two steps. By the triangle inequality for the Wasserstein distance,

$$\begin{aligned} D(\widehat{G}_{0,z}, G_0) + D(\widehat{G}_{1,z}, G_1) &\le \left[D(\widehat{G}_{0,z}, \widetilde{G}_0) + D(\widehat{G}_{1,z}, \widetilde{G}_1)\right] + \left[D(\widetilde{G}_0, G_0) + D(\widetilde{G}_1, G_1)\right] \\ &\le \frac{8}{m^{1-1/\tau} p(1-p)} \left(\frac{\sum_{i=1}^{m} |U_i|^\tau}{m}\right)^{1/\tau} + \frac{D(G, \widehat{G})}{p(1-p)}, \end{aligned}$$

where the second inequality applies Equation (B.18) to the first bracket, and Equation (B.14) together with $\frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$ to the second. This proves Equation (B.6) in Part (ii).

Similarly, by the triangle inequality for the Kolmogorov distance, together with Equation (B.17), Equation (B.16), together with its analogue to $\left|\widetilde{G}_1(t) - G_1(t)\right|$,

$$\|\widehat{G}_{0,z} - G_0\|_\infty + \|\widehat{G}_{1,z} - G_1\|_\infty \le \frac{2}{m_0} + \frac{2}{m_1} + \frac{\|G - \widehat{G}\|_\infty}{p(1-p)}.$$

Since $m_0 = \lceil pm \rceil \ge pm$ and $m_1 \ge (1-p)m/2$ as shown in Step 2,

$$\frac{2}{m_0} + \frac{2}{m_1} \le \frac{2}{pm} + \frac{4}{(1-p)m} = \frac{2+2p}{p(1-p)m} \le \frac{4}{p(1-p)m},$$

which proves Equation (B.7) in Part (iii). □

We now prove Theorem 4.1.

Proof. Let $U_1, \dots, U_m$ be the i.i.d. random variables generated from distribution G, and $\widehat{G}$ be the induced empirical distribution. We focus on the event

$$\mathcal{E} = \left\{ \rho \left[\max_{y\in\{0,1\}} D(F_y, \widehat{F}_y) + D(G, \widehat{G}) + \frac{4\left(U_{(m)} - U_{(1)}\right)}{m} \right] \le \varepsilon \quad \text{and} \quad \rho \left[\|G - \widehat{G}\|_\infty + \frac{4}{m} \right] \le \gamma \right\}.$$

Here, $\rho = \frac{1}{p_-(1-p_-)\wedge p_+(1-p_+)}$. Since $0 < p_- \le p_+ < 1$, ρ is bounded and thus treated as a constant. Concentration result for Wasserstein distance between subgaussian distributions and their empirical counterparts (see Theorem 2 in Fournier & Guillin (2015)) shows that

$$\mathbb{P}\left(D\left(F_1, \widehat{F}_1\right) > c_0\varepsilon\right) + \mathbb{P}\left(D\left(F_0, \widehat{F}_0\right) > c_0\varepsilon\right) + \mathbb{P}\left(D\left(G, \widehat{G}\right) > c_0\varepsilon\right) \le C\left(e^{-c_1 n_0 \varepsilon^2} + e^{-c_3 n_1 \varepsilon^2} + e^{-c_2 m \varepsilon^2}\right)$$

Moreover, using the fact that $U_{(m)} - U_{(1)} \le 2\max_{j\in[m]}|U_j|$

$$\mathbb{P}\left(\frac{U_{(m)} - U_{(1)}}{m} > c_0\varepsilon\right) \le m\mathbb{P}\left(|U_1| \ge \frac{c_0}{2} m\varepsilon\right) \le Cme^{-c' m^2 \varepsilon^2},$$

where we use the fact that U_1 is subgaussian in the last inequality. Furthermore, the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality implies that

$$\mathbb{P}\left(\|G - \widehat{G}\|_\infty > c_0\gamma\right) \le 2e^{-Cm\gamma^2}.$$

Combining these estimates and using that n_1 and m are fixed positive multiples of n_0, we obtain, after adjusting constants to absorb the $4\rho/m$ term,

$$\mathbb{P}\left(\mathcal{E}^c\right) \le C\left(e^{-c_1 n_0 \varepsilon^2} + e^{-c_2 m\gamma^2} + me^{-c_3 m^2 \varepsilon^2}\right).$$

It is immediate that on event $\mathcal{E}$, there exists a labeling z depending on p, G_0, G_1 such that

$$\max_{y\in\{0,1\}} D(F_y, \widehat{F}_y) + D(G_0, \widehat{G}_{0,z}) + D(G_1, \widehat{G}_{1,z}) \le \varepsilon \quad \text{and} \quad \|\widehat{G}_{0,z} - G_0\|_\infty + \|\widehat{G}_{1,z} - G_1\|_\infty \le \gamma,$$

uniformly over $\{g = (p, G_0, G_1) : p \in [p_-, p_+],\ G = pG_0 + (1-p)G_1\}$. So, for $g \in \mathcal{G}$, the labeling z belongs to $\widehat{\mathcal{Z}}(\varepsilon, \gamma)$ defined in Equation (4.2), which implies $\widehat{R}_{0,z}(\widehat{t}) \le \alpha$. Here, recall that $\widehat{t}$ is the threshold given by Algorithm 2. Moreover

$$U_0(\widehat{t}) \le \sup_{g\in\mathcal{G}} \left| R_0(\widehat{t}; g) - \widehat{R}_{0,z}(\widehat{t}) \right| + \sup_{g\in\mathcal{G}} \widehat{R}_{0,z}(\widehat{t}),$$

where the subscript z is understood to denote the labeling corresponding to the triple $g = (p, G_0, G_1)$. It is clear that NPSL guarantees the second term on the right to be at most α. Thus, it suffices to bound the first term on the right. Define $\widehat{w}_z = \frac{a}{a+\widehat{p}c}$ to be the empirical version of $w(p)$. We have

$$\begin{aligned}
&\left| R_0(\widehat{t}; g) - \widehat{R}_{0,z}(\widehat{t}) \right| \\
&= \left| \left(w(p)(1 - F_0(\widehat{t})) + (1 - w(p))(1 - G_0(\widehat{t}))\right) - \left(\widehat{w}_z(1 - \widehat{F}_0(\widehat{t})) + (1 - \widehat{w}_z)(1 - \widehat{G}_{0,z}(\widehat{t}))\right) \right| \\
&\le 2\left|w(p) - \widehat{w}_z\right| + \left\| F_0 - \widehat{F}_0 \right\|_\infty + \left\| G_0 - \widehat{G}_{0,z} \right\|_\infty \\
&\le 2\left|w(p) - \widehat{w}_z\right| + \left\| F_0 - \widehat{F}_0 \right\|_\infty + \rho\left(\left\| G - \widehat{G} \right\|_\infty + \frac{4}{m} \right),
\end{aligned}$$

where the last inequality is implied by Lemma B.6. Here, we note that the CDF evaluations in the previous equation should be understood as left limits. For clarity, we suppress this notation, noting that the subsequent bounds still apply. Invoking the DKW inequality twice, the last two terms can be bounded by $\frac{\varepsilon}{2} + 4\rho/m$ with probability at least $1 - C\left(e^{-c_1 n_0 \varepsilon^2} + e^{-c_2 m\varepsilon^2}\right)$. Moreover, since $w(p) = \frac{a}{a+pc}$, we have $|w(p) - \widehat{w}_z| \le \frac{c}{a}|\widehat{p} - p| \le \frac{1}{n_0}$. Since $n_0 = aN$ and $m = cN$, we have $\frac{4\rho}{m} + \frac{2}{n_0} \le \frac{C_0}{N}$ for some constant $C_0 > 0$. For $N\varepsilon$ sufficiently large, the remaining $\varepsilon/2$ absorbs C_0/N. Otherwise, increasing C makes the claimed probability lower bound nonpositive, so the conclusion holds trivially. □

B.5 Proof of Proposition 4.2

In this section, we prove Proposition 4.2. Let $\widehat{U}_0$ denote the empirical risk envelope over $\widehat{\mathcal{Z}}(\varepsilon, \gamma)$, with the convention that $\widehat{U}_0(t) = 1$ when this set is empty.

Lemma B.7. *Under the assumptions of Proposition 4.2, for every fixed threshold t,*

$$[\widehat{U}_0(t) - U_0(t)]_+ \xrightarrow{\mathbb{P}} 0. \tag{B.19}$$

Proof. The event $\mathcal{E}$ in Appendix B.4 has probability tending to one when $\varepsilon = \gamma = N^{-1/3}$. On that event, Lemma B.6 implies the empirical admissible set is nonempty. The observed CDFs converge uniformly to F_0, F_1, G by DKW.

First, consider a sequence of samples on which these convergences hold and the empirical admissible sets are eventually nonempty. Given an empirical maximizing labeling z_N at t, write $p_N = m_{0,z_N}/m$ and $\widehat{G}_{y,N} = \widehat{G}_{y,z_N}$. Then $p_N \widehat{G}_{0,N} + (1 - p_N)\widehat{G}_{1,N} = \widehat{G}$. By Equation (4.2) and using the fact that $\widehat{G} \Rightarrow G$, we may find a subsequence along which $p_N \to p \in [p_-, p_+]$ and $\widehat{G}_{y,N} \Rightarrow G_y$ for $y = 0, 1$, so $G = pG_0 + (1 - p)G_1$.

Using Equation (4.2), an application of Fatou's lemma gives

$$D_+(F_0, G_0) \le \liminf_N D_+(\widehat{F}_0, \widehat{G}_{0,N}) \le \delta_0,$$
$$D_+(G_1, F_1) \le \liminf_N D_+(\widehat{G}_{1,N}, \widehat{F}_1) \le \delta_1,$$

because $\varepsilon \to 0$. Similarly, since $\widehat{G}_{1,N} \le \widehat{G}_{0,N} + \gamma$ and $\gamma \to 0$, we have $G_1 \le G_0$. Thus, $g = (p, G_0, G_1) \in \mathcal{G}$.

Continuity of G at t and the fact that $G = pG_0 + (1 - p)G_1$ imply that G_0 and G_1 are continuous at t. Additionally, since $n_0 = aN$, $m = cN$, and $p_N \to p$, we have

$$\widehat{w}_{z_N} = \frac{n_0}{n_0 + m_{0,z_N}} = \frac{a}{a + cp_N} \to \frac{a}{a + cp} = w(p).$$

Substituting this into the definition of the empirical type I risk gives

$$\widehat{U}_0(t) = \widehat{R}_{0,z_N}(t) \to R_0(t; g) \le U_0(t).$$

Since the same argument applies to any subsequence, $\limsup_N \widehat{U}_0(t) \le U_0(t)$. Finally, the previous statement holds for a sequence on which the CDFs converge and the empirical admissible set is nonempty. Since we may choose a subsequence where this occurs almost surely, a standard subsequence argument gives Equation (B.19). □

We now prove Proposition 4.2.

Proof. For a sufficiently small $h > 0$, since U_0 is nonincreasing and $U_0(t) = \alpha$ has a unique solution, we have $\Delta_h := U_0(t^*_\alpha - h) - \alpha > 0$. Theorem 4.1 gives $U_0(\widehat{t}) \le \alpha + o_{\mathbb{P}}(1)$ for $\varepsilon = \gamma = N^{-1/3}$. Thus,

$$\mathbb{P}\{\widehat{t} \le t^*_\alpha - h\} \le \mathbb{P}\{U_0(\widehat{t}) > \alpha + \Delta_h/2\} \to 0.$$

For the other direction, since $G = pG_0 + (1 - p)G_1$ and $pG_0 \le G$ as measures, we have for $s < t$,

$$|U_0(s) - U_0(t)| \le F_0([s, t)) + (c/a)G([s, t)). \tag{B.20}$$

Thus, U_0 is continuous. Since $U_0(t^*_\alpha) > U_0(t^*_\alpha + h)$, $(t^*_\alpha, t^*_\alpha + h)$ has positive probability on $M = aF_0 + bF_1 + cG$. Choose $r \in (0, h)$ such that the interval $I = (t^*_\alpha + r, t^*_\alpha + h)$ has positive probability. At least one of $F_0(I), F_1(I)$ and $G(I)$ must be positive, so we have

$$\mathbb{P}\{\mathcal{T} \cap I = \varnothing\} = (1 - F_0(I))^{n_0}(1 - F_1(I))^{n_1}(1 - G(I))^m \to 0.$$

Uniqueness gives $U_0(t^*_\alpha + r) < \alpha$, so Lemma B.7 implies $\widehat{U}_0(t^*_\alpha + r) < \alpha$ with probability tending to one. By monotonicity, the smallest feasible candidate must satisfy

$$\mathbb{P}\{\widehat{t} > t^*_\alpha + h\} \to 0.$$

Together, these bounds prove $\widehat{t} \to t^*_\alpha$ in probability and $\mathbb{P}\{\widehat{t} = +\infty\} \to 0$. By continuity, $U_0(\widehat{t}) \to \alpha$ in probability. □

C Details of experiments

C.1 Synthetic experiments

For the first two experiments in Section 5, we set $\mathbb{P}(Y = 1) = 0.20$ and $X \mid Y = y \sim N_5(\mu_y, \Sigma)$, with $\mu_1 = \mathbf{0}$, $\mu_0 = -(\sqrt{11}/2)\mathbf{1}$, and $\Sigma = 0.70I_5 + 0.30\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1}$ is the five-dimensional vector of ones. These choices give unit marginal variances, covariate correlation 0.30, and population class separation 2.50. The fitted logistic regression score is standardized so that, conditional on the trained score, $S \mid Y = 1 \sim N(0, 1)$ and $S \mid Y = 0 \sim N(-d, 1)$, where d is the separation achieved by the fit. The label-selection mechanism is given by

$$\mathbb{P}(A = 1 \mid S, Y) = \text{logit}^{-1}\left[1 + 2S + 4Y \exp\left\{-\frac{(S + 0.5)^2}{2(0.5)^2}\right\}\right]. \tag{C.1}$$

The standardization of the scores was done to make the score units and preceding equation more interpretable; this step could be omitted without changing the resulting risks, provided the label-selection mechanism was transformed appropriately. For training, the same label selection mechanism is used with $S = \mathbf{1}^\top X/\sqrt{11}$.

IPW and fuzzy augmentation. Let S_i denote the fixed score, A_i indicate whether the calibration label is observed, and $\mathcal{I}_0 = \{i : A_i = 1, Y_i = 0\}$. For IPW, we fit a binomial logistic regression of A_i on a natural cubic spline of S_i with four degrees of freedom, using the entire calibration set. For stability, we clip fitted probabilities to $[10^{-6}, 1 - 10^{-6}]$ and assign observed class 0 observations weights $w_i = 1/\widehat{e}(S_i)$. The weighted type I criterion is

$$\widehat{R}_{0,\text{IPW}}(t) = \frac{\sum_{i \in \mathcal{I}_0} w_i \mathbb{1}\{S_i \geq t\}}{\sum_{i \in \mathcal{I}_0} w_i}.$$

For fuzzy augmentation, we fit a separate binomial logistic regression of Y_i on a natural cubic spline of S_i with four degrees of freedom, using the labeled observations in the calibration set. For each unlabeled observation, let $\nu_i = 1 - \widehat{\mathbb{P}}(Y_i = 1 \mid S_i, A_i = 1)$. We then set

$$\widehat{R}_{0,\text{fuzzy}}(t) = \frac{\sum_{i \in \mathcal{I}_0} \mathbb{1}\{S_i \geq t\} + \sum_{i: A_i = 0} \nu_i \mathbb{1}\{S_i \geq t\}}{|\mathcal{I}_0| + \sum_{i: A_i = 0} \nu_i}.$$

Both methods select the smallest threshold whose respective empirical criterion is at most α.

Computation of population thresholds. This paragraph describes the computation of the population thresholds used in the convergence experiment. Let $\widehat{\beta}$ be the logistic regression slope vector. The score used for calibration and the population score used for dashed lines in Figure 1(c) are, respectively,

$$s_{\widehat{\beta}}(x) = \frac{\widehat{\beta}^\top (x - \mu_1)}{\sqrt{\widehat{\beta}^\top \Sigma \widehat{\beta}}}, \qquad s_\star(x) = \frac{\mathbf{1}^\top (x - \mu_1)}{\sqrt{\mathbf{1}^\top \Sigma \mathbf{1}}} = \frac{\mathbf{1}^\top x}{\sqrt{11}}.$$

For the computation of the optimal robust threshold, let F_y be the CDF of $s_\star(X)$ given $(A = 1, Y = y)$ and let G be its CDF given $A = 0$, under Equation (C.1). Write $a = \mathbb{P}(A = 1, Y = 0)$, $b = \mathbb{P}(A = 1, Y = 1)$, and $c = \mathbb{P}(A = 0)$. We use the limiting population counterparts of the defaults in Section 2.3, with $0 < \pi (:= 1 - p) \leq \pi_{\max} := b/(a + b)$. Then we may numerically compute $\pi_{\max} \approx 0.760$ and $\delta_{1,\text{pop}} \approx 2.201$. The optimal robust threshold is the smallest threshold at which the supremum, over completions satisfying the limiting restrictions described, of

$$\frac{a\{1 - F_0(t)\} + c(1 - \pi)\{1 - G_0(t)\}}{a + c(1 - \pi)}$$

is at most 0.15.

C.2 LendingClub illustration

Data and score. We use LendingClub data (George, 2019), combining approved applicants with 36-month repayment outcomes and declined applications with credit scores, from January–October 2013. A random split reserves 50% of applicants for evaluation and divides the remainder equally between training and calibration. We use approved loans with status "Fully Paid," "Charged Off," or "Default," assigning $Y = 1$ to "Fully Paid" and $Y = 0$ to the other two statuses.

Using only approved training loans, we fit a logistic regression model with FICO and debt-to-income ratio (DTI) as predictors. Because approved loans' FICO scores are given in bins of five points, we use the lower endpoint of the bin, and round down the scores of the declined applicants to a multiple of five. FICO scores outside $[300, 850]$ are excluded. For DTI, any values that are missing, invalid, or outside $[0, 100]$ are replaced by the median among approved loans in the training set. We set $\alpha = 0.1$.

We retain calibration labels independently with probability given by a logistic function of the unrounded repayment score, standardized using its mean and standard deviation across training applicants. We estimate the score coefficient from historical funding indicators in the training population and multiply it by $\kappa = 1.5$ to strengthen score-dependent label retention. The outcome-dependent (MNAR) mechanism additionally uses a coefficient of -1 on the repayment indicator, so repayers' labels are retained less often at a fixed score; this coefficient is zero for score-only (MAR) selection. We set the intercept so that about 11.72% of the approved training loans have observed labels in expectation, matching the historical approval rate in the training data. After selecting which labels are observed, we round the scores in the calibration and evaluation sets to three decimal places, giving 238 distinct scores in the calibration set. The split and the fitted score are fixed across the 500 replications.

D Additional experiments

D.1 Comparison under Missing at Random

We repeat the first experiment in Section 5 under MAR. Thus, in this experiment, the label selection mechanism is given by $\mathbb{P}(A = 1 \mid S, Y) = \text{logit}^{-1}(1 + 2S)$.

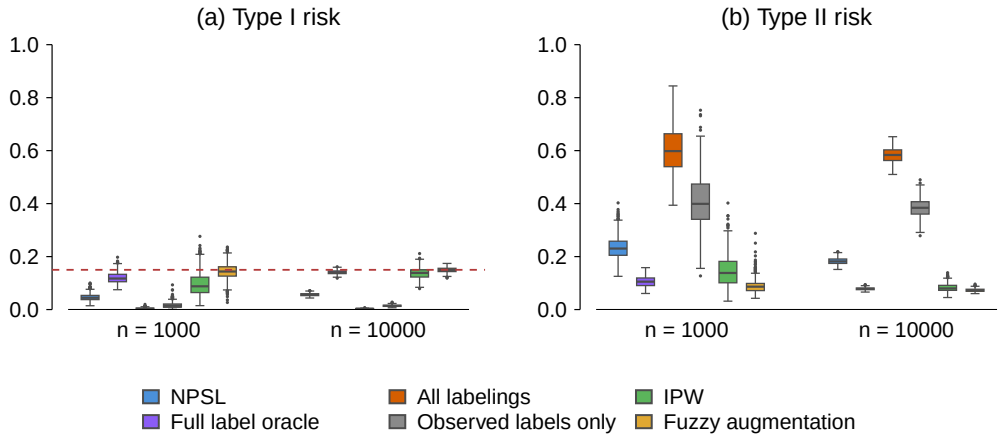

Figure 2: MAR counterpart of Figure 1(a–b). Population type I and type II risks over 500 replications at each sample size. The dashed line marks the type I target $\alpha = 0.15$.

Figure 2 show no observed type I violations for NPSL, observed labels only, or the all labelings rule at either sample size. Among these methods, NPSL has the lowest mean type II risk. IPW and fuzzy augmentation achieve lower type II risk, but their violation rates can still be high.

D.2 SENSITIVITY TO PARAMETERS

An interpretation of λ. We first give a simple restriction on the population distributions that relates λ to score separation. Suppose F_1, G_0, G_1 have finite first moments and $F_1(u) \leq G_1(u) \leq G_0(u)$ for every u. Let $\mu_{F_1}, \mu_{G_1}, \mu_{G_0}$ denote their means, and define

$$\Delta_s = \mu_{F_1} - \mu_{G_1}, \qquad \Delta_c = \mu_{G_1} - \mu_{G_0}.$$

Under the stated assumptions, the displacement integrals equal differences in means. Since $G = pG_0 + (1-p)G_1$, where p is the class 0 proportion among unlabeled observations, we have

$$D_+(G_1, F_1) = \Delta_s, \qquad D_+(G, F_1) = \Delta_s + p\Delta_c.$$

Consequently, $D_+(G_1, F_1) \leq \lambda D_+(G, F_1)$ holds if and only if

$$(1-\lambda)\Delta_s \leq \lambda p \Delta_c.$$

For $\lambda = 0.8$, this corresponds to $\Delta_s \leq 4p\Delta_c$. This does not introduce additional constraints into NPSL, but does not provide any finite-sample guarantees on the algorithm.

Empirical sensitivity analysis. We now examine the sensitivity of NPSL to δ_1, γ, and δ_0 using the same setup as the first experiment in Section 5, with $n = 1000$ observations split into 667 observations for training and $N = 333$ for calibration, with target $\alpha = 0.15$. We use 500 replications, refitting the score in each replication. We recall that $\widehat{\sigma}$ is the sample standard deviation of all scores in the calibration set and $\widehat{G}$ is the empirical CDF of unlabeled observations in the calibration set. Then using $\delta_1 = \lambda D_+(\widehat{G}, \widehat{F}_1)$ and the admissible set from Section 2.3, we vary

$$\lambda \in \{0.5, 0.55, \ldots, 0.95, 1\}, \qquad \gamma \in \{0, N^{-1/3}, 1\}, \qquad \frac{\delta_0}{\widehat{\sigma}} \in \{0, N^{-1/3}, 1\},$$

one parameter at a time, fixing the others at the defaults, $\lambda = 0.8$, $\gamma = N^{-1/3}$, and $\delta_0 = \widehat{\sigma} N^{-1/3}$. Here, $N^{-1/3} \approx 0.144$.

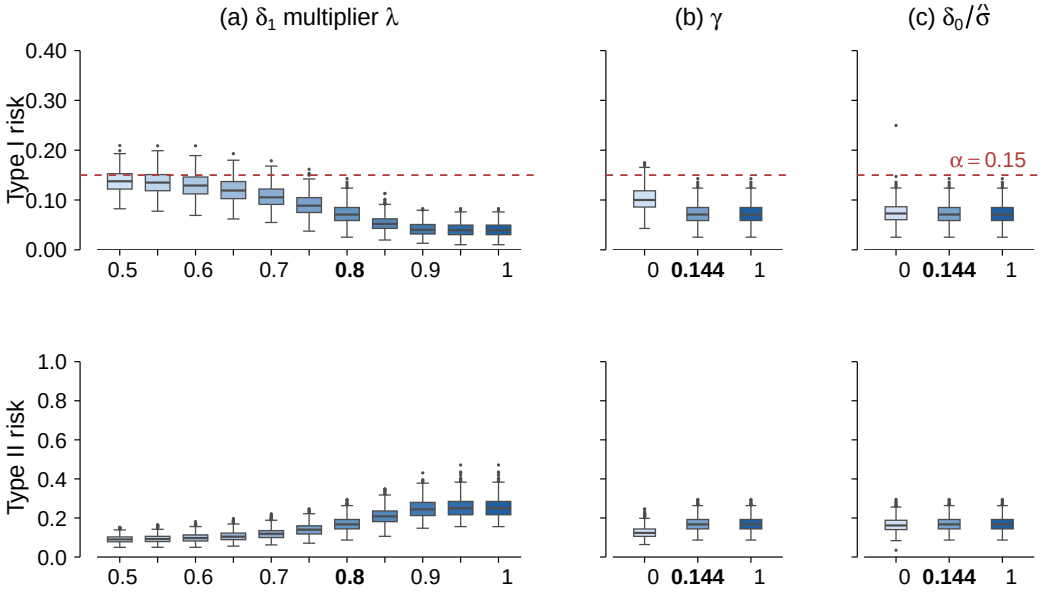

Figure 3: Population type I (top) and type II (bottom) risks over 500 replications. Panels (a–c) vary λ, γ, and $\delta_0/\widehat{\sigma}$, respectively. Bold indicates the recommended default. The dashed line is $\alpha = 0.15$.

Figure 3 shows the clearest tradeoff for δ_1. Decreasing λ from 1 to 0.8 lowers average type II risk, and both have violation rates of 0%. At lower values of λ, the violation rates increase substantially. Increasing γ from zero to $N^{-1/3}$ decreases the violation rate, but increases the average type II risk. Increasing γ further to one has little additional effect. The class 0 parameter δ_0 has an even smaller effect, with $\delta_0 = 0$ having a violation rate of 0.2% (with one violation out of 500 replications), while the other two settings have a violation rate of 0%. Increasing $\delta_0/\widehat{\sigma}$ from zero to $N^{-1/3}$ increases the average type II risk, with little further change at one.

D.3 Robustness of Recommended Defaults

Complementing Appendix D.2, we evaluate type I risk protection under the recommended defaults $\delta_1 = 0.8D_+(\widehat{G}, \widehat{F}_1)$, $\delta_0 = \widehat{\sigma}N^{-1/3}$, and $\gamma = N^{-1/3}$. We consider the violation rate, the percentage of replications in which the population type I risk $R_0(\widehat{t})$ exceeds the target α.

For each of the settings, we run 500 replications, with $N = 1000$. The baseline is $S \mid Y = 1 \sim N(0, 1)$ and $S \mid Y = 0 \sim N(-d, 1)$, with separation $d = 1.5$ and prevalence $\mathbb{P}(Y = 1) = 0.20$. The label selection mechanism follows Equation (C.1), with default intercept $a_0 = 1$, score coefficient 2, and outcome-dependent coefficient 4. The localized term has width 0.5 and center $-a_0/2$. MNAR(θ) denotes outcome-dependent coefficient θ, with $\theta = 0$ giving MAR. The default α is set to 0.15, with additional evaluations at $\alpha = 0.05$ and $\alpha = 0.1$.

We vary one quantity at a time. The score-shape setting either increases the class 0 standard deviation to 1.5, or changes the shape of the score, using a bimodal mixture with weights 0.7 and 0.3 for the lower and higher-mean components respectively, unit variances, and means four units apart. The class 0 means are chosen to match the true distribution oracle threshold at $\alpha = 0.15$, with class 1 scores unchanged. The class 0 selection setting replaces its approval score coefficient by 0 (flat), -1 (weak reversal), or -2 (strong reversal), adjusting the intercept to preserve the overall approval probability within class 0. Note that we report "weak reversal" as "reversed selection" in the main text for ease of exposition.

Table 3: Performance of the recommended defaults over 500 replications. Best possible R_1 is the minimum population type II risk among score thresholds satisfying $R_0 \leq \alpha$. Violation rate is the percentage with $R_0(\widehat{t}) > \alpha$. Rows use $\alpha = 0.15$ except the final two, which vary α at the default configuration.

Varied quantity	Value	Mean R_0	Mean R_1	Best R_1	Viol. Rate (%)
Default	–	0.089	0.439	0.321	0.0
Prevalence	0.3	0.072	0.487	0.321	0.0
	0.4	0.064	0.510	0.321	0.0
	0.5	0.062	0.518	0.321	0.0
	0.6	0.060	0.525	0.321	0.0
	0.8	0.053	0.552	0.321	0.0
Approval intercept	0	0.072	0.486	0.321	0.0
	0.5	0.081	0.460	0.321	0.0
Class separation	1	0.111	0.588	0.515	0.0
	2	0.067	0.310	0.168	0.0
Score shape	unequal variance	0.082	0.527	0.321	0.0
	bimodal	0.074	0.587	0.321	0.0
Label selection mechanism	MAR	0.091	0.435	0.321	0.0
	MNAR(2)	0.089	0.440	0.321	0.0
	MNAR(2.5)	0.089	0.439	0.321	0.0
	MNAR(3)	0.089	0.441	0.321	0.0
Class 0 selection	flat	0.116	0.381	0.321	0.2
	weak reversal	0.006	0.982	0.321	2.2
	strong reversal	0.000	1.000	0.321	0.0
Target level α	0.05	0.028	0.660	0.558	0.0
	0.1	0.058	0.530	0.414	0.0

At $\alpha = 0.15$, every setting except the flat and weak reversal settings has zero violation rate. These results support the defaults as reasonable choices for type I protection in the settings studied. The reversed selection settings show the importance of the class 0 selection restriction. At $\alpha = 0.15$, weak reversal gives 2.2% violation rate and 97.8% decline-all outcomes. The low violation rates in these cases therefore reflect almost universally uninformative rules.

D.4 COMPUTATIONAL RUNTIME

Table 4 compares full-curve computation with threshold only search using the setup of the first experiment, with $N \in \{1000, 2000, 4000, 8000\}$. Timings were obtained in R 4.5.3 on a Mac Studio with an Apple M1 Ultra chip (20 CPU cores: 16 performance and 4 efficiency), 64 GB of unified memory, and macOS 15.2. Both implementations parallelize over k using 20 workers. Times are medians of three runs on each set, excluding score fitting and data generation.

Table 4: Median wall-clock time in seconds by calibration-set size N. Dashes indicate timings not measured.

N	Full curve (s)	Threshold search (s)
1000	12.08	1.23
2000	109.38	2.44
4000	–	3.83
8000	–	14.65

In this experiment, all scores are distinct, but tied scores can reduce the number of distinct scores used for computation. For example, the credit experiment uses $19,609$ observations in the calibration set but only 238 distinct scores after rounding to three decimal places (Appendix C.2), which effectively reduces the size of the calibration set.