DISCUSSION PAPER SERIES

DP15418

BINARY CHOICE WITH ASYMMETRIC LOSS IN A DATA-RICH ENVIRONMENT: THEORY AND AN APPLICATION TO RACIAL JUSTICE

Eric Ghysels, Andrii Babii, Xi Chen and Rohit Kumar

LABOUR ECONOMICS

ISSN 0265-8003

BINARY CHOICE WITH ASYMMETRIC LOSS IN A DATA-RICH ENVIRONMENT: THEORY AND AN APPLICATION TO RACIAL JUSTICE

Eric Ghysels, Andrii Babii, Xi Chen and Rohit Kumar

Discussion Paper DP15418
Published 31 October 2020
Submitted 27 October 2020

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

BINARY CHOICE WITH ASYMMETRIC LOSS IN A DATA-RICH ENVIRONMENT: THEORY AND AN APPLICATION TO RACIAL JUSTICE

Abstract

The importance of asymmetries in prediction problems arising in economics has been recognized for a long time. In this paper, we focus on binary choice problems in a data-rich environment with general loss functions. In contrast to the asymmetric regression problems, the binary choice with general loss functions and high-dimensional datasets is challenging and not well understood. Econometricians have studied binary choice problems for a long time, but the literature does not offer computationally attractive solutions in data-rich environments. In contrast, the machine learning literature has many computationally attractive algorithms that form the basis for much of the automated procedures that are implemented in practice, but it is focused on symmetric loss functions that are independent of individual characteristics. One of the main contributions of our paper is to show that the theoretically valid predictions of binary outcomes with arbitrary loss functions can be achieved via a very simple reweighting of the logistic regression, or other state-of-the-art machine learning techniques, such as boosting or (deep) neural networks. We apply our analysis to racial justice in pretrial detention.

JEL Classification: N/A

Keywords: N/A

Eric Ghysels - eghysels@unc.edu
University of North Carolina and CEPR

Andrii Babii - andrii@email.unc.edu
University of North Carolina Chapel Hill

Xi Chen - xich@live.unc.edu
University of North Carolina Chapel Hill

Rohit Kumar - sirohi.rohit@gmail.com
Indian Statistical Institute

Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice

Andrii Babii*
UNC Chapel Hill

Xi Chen†
UNC Chapel Hill

Eric Ghysels‡
UNC Chapel Hill

Rohit Kumar§
ISI, Delhi

October 27, 2020

Abstract

The importance of asymmetries in prediction problems arising in economics has been recognized for a long time. In this paper, we focus on binary choice problems in a data-rich environment with general loss functions. In contrast to the asymmetric regression problems, the binary choice with general loss functions and high-dimensional datasets is challenging and not well understood. Econometricians have studied binary choice problems for a long time, but the literature does not offer computationally attractive solutions in data-rich environments. In contrast, the machine learning literature has many computationally attractive algorithms that form the basis for much of the automated procedures that are implemented in practice, but it is focused on symmetric loss functions that are independent of individual characteristics. One of the main contributions of our paper is to show that the theoretically valid predictions of binary outcomes with arbitrary loss functions can be achieved via a very simple reweighting of the logistic regression, or other state-of-the-art machine learning techniques, such as boosting or (deep) neural networks. We apply our analysis to racial justice in pretrial detention.

*Department of Economics, University of North Carolina at Chapel Hill. Email: babii.andrii@gmail.com.

†Department of Economics, University of North Carolina at Chapel Hill. Email: xich@live.unc.edu.

‡Department of Economics and Department of Finance, Kenan-Flagler Business School, University of North Carolina at Chapel Hill and CEPR. Email: eghysels@unc.edu.

§Indian Statistical Institute, Delhi. Email: sirohi.rohit@gmail.com

1 Introduction

The downside risk and upside gains of many economic decisions are not symmetric. The importance of asymmetries in prediction problems arising in economics has been recognized for a long time; see Granger (1969), Granger and Pesaran (2000), Pesaran and Skouras (2002), and the textbook treatment in Elliott and Timmermann (2016), among many others. In this paper, we focus on binary choice problems in a data-rich environment with general loss functions. Hence, our analysis covers many life-changing decisions such as college admission, job hiring, pre-trial release from jail, medical testing and treatment, which have become increasingly driven by automated algorithmic processes based on vast data inputs. It also covers many routine tasks such as fraud detection, spam filters, credit risk, etc. The topic has gained interest from a diverse set of fields, ranging from economics, computer science, to machine learning, among others, and depending on the discipline is also known as classification or screening problems.

The combination of high-dimensional data and general loss functions in the binary prediction problems is challenging and not well understood.[1] Econometricians have studied the binary choice problems for a long time, but the literature does not offer computationally attractive solutions for high-dimensional datasets. Numerous attempts to relax the strong parametric distributional assumptions resulted in *non-smooth* combinatorial optimization problems, also known as NP-hard problems. Indeed, the maximum score method of Manski (1975), or the extension to asymmetric loss functions proposed by Elliott and Lieli (2013), leads to such NP-hard optimization problems which makes them computationally challenging in data-rich environments. In contrast, the machine learning (ML) literature has many computationally attractive algorithms that form the basis for much of the automated procedures that are implemented in practice, but it is focused on symmetric loss functions that are independent of individual characteristics (sometimes called features in ML). In particular, the ML literature emphasizes the importance of smooth optimization and made significant progress on coming up with smooth convex and computationally attractive relaxations for the binary classification problem.[2]

[1]In contrast, it is known that the asymmetric prediction in the regression setting can be treated, e.g., with the quantile regression, cf., Koenker and Bassett Jr (1978), the asymmetric least-squares, cf., Aigner, Amemiya, and Poirier (1976), or more generally with M-estimators based on a suitable asymmetric loss function; see Elliott and Timmermann (2016) for more examples.

[2]The importance of smooth and convex surrogate objective functions in the binary classification problem is emphasized in Vapnik (1995), and more recently in Zhang (2004) and Bartlett, Jordan, and McAuliffe (2006); see also Boucheron, Bousquet, and Lugosi (2005) for an excellent survey.

Concerns about Algorithmic Decision-making

As more and more decisions affecting our daily lives have become digitized and automated by data-driven algorithms, concerns have been raised regarding so-called algorithmic biases. Gender and race are two leading examples. Cowgill and Tucker (2019) discuss the case of computer scientists at Amazon who developed powerful new technology to screen resumes and discovered the algorithm placed a negative coefficient on terms associated with women and as a result appeared to be amplifying and entrenching male dominance in the technology industry. Another example of gender discrimination is discussed by Datta, Tschantz, and Datta (2015) who study AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. They found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male.

Another example related to the empirical application in our paper is pre-trial detention. Journalists at the news website ProPublica reported on commercial software used by judges in Broward County, Florida, that helps to decide whether a person charged with a crime should be released from jail before their trial. They found that the software tool called COMPAS resulted in a disproportionate number of false positives for black defendants who were classified as high risk but subsequently not charged with another crime.[3]

Practical implementation: A quartet of functions and an econometrician's toolbox

The potential for ML algorithmic outcomes to reproduce and reinforce existing discrimination against legally protected groups has been of great concern lately. In response, there is a burgeoning literature in computer science dealing with fairness-aware classification decision rules. Much of the literature in computer science approaches the problem of algorithmic fairness by first introducing a definition of a fair prediction function.[4] Economists have argued that defining fairness in terms of

[3]Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) note that pretrial detention decisions provide an attractive template for when and how machine learning might be used to improve on decisions made by judges and does not pertain to uncovering causal relationships, cf., Athey and Wager (2020) and references therein for the causal treatment assignment problem; see also Belloni, Chernozhukov, and Hansen (2014), Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018), Athey and Imbens (2019), and Babii, Ghysels, and Striaukas (2020), regarding machine learning and causal inference.

[4]Barocas, Hardt, and Narayanan (2020) provide a textbook (in progress) treatment of fairness in machine learning from a computer science perspective.

properties of the underlying prediction function may not be appealing. One reason is that many commonly used definitions of fairness in the computer science literature cannot be simultaneously satisfied (the so-called impossibility theorem, see Chouldechova (2017), Kleinberg, Mullainathan, and Raghavan (2016) among others). Economists instead emphasize that one should focus on preferences (of a social planner) regarding treatment for protected groups. The fact that economists emphasize the importance of general loss functions also reinforces the importance of the contributions of our paper regarding the ongoing discussions of fairness in automated binary decision/classification problems. To the best of our knowledge, it is not known how to implement ML methods in such economic environments.

The practical implementation of our procedure is remarkably simple, which we why it can be discussed here. The first ingredient comes from a policy/decision maker who has to provide a quartet of loss functions pertaining to (a) true positives (denoted by $\ell_{1,1}(x)$), (b) true negatives ($\ell_{-1,-1}(x)$), (c) false positives ($\ell_{-1,1}(x)$) and last but not least (d) false negatives ($\ell_{1,-1}(x)$). Note that this not only implies the selection of functional forms, but also the selection of inputs x, i.e. individual covariates. It is important to emphasize that the functions $\ell_{i,j}(x)$ i, j = -1, 1 are *not* estimated in general. They are determined by either a utility function of the social planner (policy/decision maker), or are obtained through some type of cost-benefit analysis.[5]

Given these inputs, the task of the econometrician is to estimate the binary choice model, given the asymmetries reflected in the quartet of functions. We show that the logistic regression, boosting, and (deep) neural networks with a very simple reweighing for asymmetries of the loss function lead to theoretically valid binary predictions *without* requiring strong distributional assumptions. The econometrician can decide which estimation procedure to implement. The major point and key contribution of our paper is that the standard procedures, whichever one is selected, are applied to a reweighted objective function with the weights determined by the quartet of functions supplied by the policy/decision maker expressing concerns regarding the treatment for protected groups. Hence, all the standard procedures in the econometrician's toolbox apply once the reweighting is applied.

The optimal decisions with asymmetric loss functions typically yield covariate-driven threshold rules. This is the case for the general social planner setting studied by Rambachan, Kleinberg, Ludwig, and Mullainathan (2020) as well as the special cases studied by Corbett-Davies, Pierson, Feller, Goel, and Huq (2017) and Klein-

[5]In the empirical section 6 we will rely on a cost-benefit study to determine this quartet of functions.

berg, Ludwig, Mullainathan, and Rambachan (2018), among others. To the best of our knowledge, there is no theory that supports empirical implementation under general conditions involving either parametric predictors, boosting, shallow and deep learning. We provide such theory for all these types of estimators.[6] In particular, one of the main contributions of our paper is to show that predicting binary outcomes with arbitrary loss functions can be achieved via a very simple reweighting of the logistic regression, or other state-of-the-art machine learning techniques, such as boosting or (deep) neural networks. Despite the popularity of boosting algorithms and neural networks in practice, to the best of our knowledge, there is not much theoretical work done in econometrics on these methods in the asymmetric binary choice setting. We establish the theoretical guarantees for both methods for arbitrary loss functions, and in particular, we show that a carefully crafted fully connected feedforward (deep) neural network achieves the minimax optimal convergence rates (up to the $\log n$ factor) of the excess risk.

Filling a gap: Machine learning classification and asymmetries

Our theoretical analysis is agnostic about which econometric method to use. Indeed, as noted earlier we cover logistic regression, boosting, and (deep) neural networks and therefore the typical collection of machine learning tools for supervised classification. Regarding deep learning our paper also relates to early important work on neural networks done by Xiaohong Chen who noted that such models can be viewed as a nonlinear sieve.[7] The recent literature emphasizes the importance of depth and considers deep neural networks with ReLU activation function, see Schmidt-Hieber (2017), Farrell, Liang, and Misra (2019), and Bauer and Kohler (2019). This literature covers regression and semiparametric inference problems, whereas our paper pertains to discrete choice models with general loss functions and we, therefore, establish new and comparable results for a different setting. We noted earlier that our analysis covers a wide range of applications such as college admission, job hiring, pre-trial release from jail, medical testing and treatment, fraud detection, spam filters, credit risk, etc. In almost all these cases asymmetric loss functions are

[6]Our framework also allows for the support vector machines and the ℓ_1 regularized methods. The focus on boosting and (deep) neural networks is motivated by the fact that these methods are frequently the most successful in various classification ML competitions.

[7]Chen (2007) discusses all kinds of sieves, including nonlinear sieves such as neural networks and ridgelets, among others. Her chapter covers many of her early work, such as Chen and Shen (1998) and Chen and White (1999) connecting sieve estimators with single- and multi-layer networks. See also the recent paper by Chen, Chen, and Tamer (2020).

relevant. We, therefore, fill an important gap: machine learning classification with loss functions of interest to economists.

The paper is organized as follows. Section 2 introduces notation and specifies the binary prediction problem in terms of risk payoffs and what an optimal prediction looks like. Examples of such decision/prediction problems are also provided. Section 3 covers the main convexification theorem and provides examples of convexifying functions. In Section 4, we provide risk bounds for the accuracy of binary predictions and discuss several cases, ranging from parametric prediction models, boosting, shallow and deep learning. A Monte Carlo simulation study is reported in Section 5. Finally, we provide an empirical application pertaining to pretrial detention in Section 6 followed by conclusions and appendices with technical details.

2 Binary decisions

Let $Y \in \{-1, 1\}$ be the target variable and let $X \in \mathcal{X} \subset \mathbf{R}^d$ be covariates. A measurable function $f : \mathcal{X} \to \{-1, 1\}$ is called the binary prediction/decision/choice or simply the prediction. Making a binary decision amounts to minimizing a risk function that describes its consequences in different states of the world

$$\mathcal{R}(f) = \mathbb{E}_{Y,X}[\ell(f(X), Y, X)],$$

where $\ell : \{-1, 1\}^2 \times \mathcal{X} \to \mathbf{R}$ is a loss function specified by the decision maker. Note that the decision f is typically random and the expectation is taken with respect to the distribution of (Y, X) only. The loss function can be asymmetric and can also depend on the covariates X, which is economically a more realistic scenario faced by the decision maker than the one provided by the classification setting.

2.1 Some examples

In the Introduction, we alluded to many examples in economic decision making. To motivate the first example, we can think of a credit risk application, where with the false negative mistakes ($f(X) = -1$ and $Y = 1$) the bank suffers losses from the borrower's default, while with the false positive mistakes ($f(X) = 1$ and $Y = -1$), the bank simply foregoes its potential earnings. Moreover, the size of the loan itself may determine the loss:

Example 2.1 (Big vs. small)**.** *Let $x = (s, z) \in \mathbf{R}^d$ be a vector of covariates, where $s \geq 0$ is the size of the loan and z are some other covariates. The "big vs small" loss* *function is*

$$\ell(f(x), y, x) = \tau \mathbb{1}\{f(x) \neq y, s \leq s^*\} + \mathbb{1}\{f(x) \neq y, s > s^*\}$$

where $\tau \in (0,1)$ *measures the relative loss from default for large and small loans defined by the threshold* s^*. *More generally, the decision maker may have a loss function that depends continuously on the size of the loan.*

As noted in the Introduction, an important policy debate pertains to the fairness and the discrimination bias of machine learning algorithms towards, e.g., low income groups, gender, or race. The following example suggests that the algorithmic bias can be reduced introducing the asymmetry of the loss function across groups:[8]

Example 2.2 (Binary prediction with protected group)**.** *Let* $x = (g, z) \in \mathbf{R}^d$ *be a vector of covariates, where* $g \in \{0, 1\}$ *is a binary variable and* $z \in \mathbf{R}^{d-1}$ *are some other covariates. Consider the loss function*

$$\ell(f(x), y, x) = \psi_g \mathbb{1}\{f(x) \neq y\}$$

placing different weights ψ_g *for losses in the two groups* $g \in \{0, 1\}$. *More generally, we can also place different weights for false positive and false negative losses across groups*

$$\ell(f(x), y, x) = \psi_g \mathbb{1}\{y = 1, f(x) = -1\} + \varphi_g \mathbb{1}\{y = -1, f(x) = 1\}.$$

In a more ambitious setting, Rambachan, Kleinberg, Ludwig, and Mullainathan (2020) consider a social planner with multiple agents of two types, a "disadvantaged group" with G = 1 and the rest with G = 0. In particular, Rambachan, Kleinberg, Ludwig, and Mullainathan (2020) - Appendix, Section A - provide a setting in which the utility of each individual in the (training) sample depends on the classification outcome. The true outcome of interest is, therefore, the change in the utilities of an individual from being selected or not. The social planner's welfare weights may be higher on the disadvantaged group if the utility of an individual from the disadvantaged group is uniformly lower than the utility of an individual from the advantaged group. They interpret this as capturing un-modeled discrimination against the disadvantaged group or existing disparities across groups in a reduced form manner. More formally:

[8] The algorithm is often perceived as unfair if it treats differently the identical individual from two different groups. For instance, Larson, Mattu, Kirchner, and Angwin (2016) report that the COMPAS software tends to make false positive predictions of the recidivism for black individuals two times more frequently compared to white individuals.

Example 2.3 (Social planner with disadvantaged group)**.** *The social planner's welfare function can be written as:*

$$\mathcal{R}(f) = \mathbb{E}_{Y,X,G}[\psi_G \ell(f(X,G), Y, X, G)],$$

where ψ_G >0 are generalized social welfare weights placed upon individuals in group $G \in \{0,1\}$, and for $\psi_1 > \psi_0$ implies that outcomes associated with the disadvantaged group are valued more than outcomes associated with the rest of the population.

Among other things Rambachan, Kleinberg, Ludwig, and Mullainathan (2020) show that the social planner's first-best rule is a threshold rule with group-specific thresholds. Corbett-Davies, Pierson, Feller, Goel, and Huq (2017) and Kleinberg, Ludwig, Mullainathan, and Rambachan (2018), among others, obtain similar results albeit under less general settings.

Computer scientists have focused on defining fairness-aware algorithms by imposing restrictions on f. Often a formal criterion of fairness is defined, and a decision rule is developed to satisfy the criterion. Some of these schemes also yield threshold rules. Two examples are statistical parity and its conditional version. Statistical parity means equal proportions across groups G.[9] Formally, it means $\mathbb{E}_X[f(X,G)|G] = \mathbb{E}_{X,G}[f(X,G)]$ for all G. Another variation is conditional statistical parity: $\mathbb{E}_X[f(X,G)|h(X),G] = \mathbb{E}_{X,G}[f(X,G)|h(X)]$ where $h(X)$ is a projection of the features onto a sub-space.[10] Corbett-Davies, Pierson, Feller, Goel, and Huq (2017) show how (conditional) statistical parity yield covariate-driven threshold rules. We follow the economist's arguments that fairness is naturally defined through losses that we are willing to tolerate when we commit mistakes for different groups. Indeed, the framework of losses is more general than thresholding rules.

In the remainder of the paper, we will continue to work with settings involving a generic vector of covariates $X \in \mathbf{R}^d$ which may contain the binary group membership variable $G \in \{0,1\}$ and some other covariates $Z \in \mathbf{R}^{d-1}$. The point worth keeping in mind, however, is that our framework covers much discussed preference-based notions of fairness characterized by covariate-dependent general loss functions.

[9] See e.g. Kamishima, Akaho, Asoh, and Sakuma (2012), Zemel, Wu, Swersky, Pitassi, and Dwork (2013), among others.

[10] See Dwork, Hardt, Pitassi, Reingold, and Zemel (2012), among others.

2.2 Optimal binary decision

Following the decision-theoretic perspective, we define the optimal binary decision as the one achieving the smallest risk

$$\mathcal{R}^* = \inf_{f:\mathcal{X}\to\{-1,1\}} \mathbb{E}[\ell(f(X), Y, X)],$$

where the minimization is done over all measurable functions $f : \mathcal{X} \to \{-1, 1\}$.

Given a sample $(Y_i, X_i)_{i=1}^n$, the goal is to construct a binary decision rule $\hat{f}_n(x) = \hat{f}(x; Y_1, X_1, \ldots, Y_n, X_n)$ such that its excess risk $\mathcal{R}(\hat{f}_n) - \mathcal{R}^*$ is as small as possible. Importantly, this requirement should be satisfied without imposing restrictive parameteric assumptions on the distribution of (Y, X). The empirical risk minimization is a popular method of constructing such predictions; see Vapnik (1995). Assuming that the data are i.i.d. or more generally stationary and ergodic, it consists of minimizing

$$\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^n \ell(f(X_i), Y_i, X_i)$$

with respect to $f \in \mathscr{F}_n$, where $\mathscr{F}_n$ is some class of measurable functions. It is well-known that this problem is NP-hard even in the case of the symmetric X-independent loss function.

The following proposition illustrates the NP-hardness of the problem and provides an equivalent characterization of the optimal binary decision that is amenable to the convexification.[11]

Proposition 2.1. *The optimal binary decision f^* solves*

$$\inf_{f:\mathcal{X}\to\{-1,1\}} \mathbb{E}[(Ya(X) - b(X))\mathbb{1}\{-Yf(X) \geq 0\}]$$

with

$$\begin{aligned} a(x) &= -\ell_{1,1}(x) - \ell_{1,-1}(x) + \ell_{-1,1}(x) + \ell_{-1,-1}(x), \\ b(x) &= \ell_{1,1}(x) - \ell_{1,-1}(x) - \ell_{-1,1}(x) + \ell_{-1,-1}(x), \end{aligned}$$

and $\ell_{f,y}(x) = \ell(f, y, x)$.

Proposition 2.1 shows that the optimal binary decision minimizes the objective function involving the indicator function $z \mapsto \mathbb{1}\{z \geq 0\}$, which is discontinuous and

[11]The proofs of the results appear in Appendix Section A.1.

not convex. This leads to a difficult non-smooth, non-convex, NP-hard empirical risk minimization problem[12]

$$\inf_{f:\mathcal{X}\to\{-1,1\}} \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))\mathbb{1}\{-Y_i f(X_i) \geq 0\}.$$

3 Convexification

3.1 Convexified excess risk

The purpose of this section is to convexify the binary decision problem with a generic loss function making it amenable to modern machine learning algorithms. For a convex function $\phi : \mathbf{R} \to \mathbf{R}$, the convexified risk of a binary prediction $f : \mathcal{X} \to \{-1, 1\}$ is defined as

$$\begin{aligned}\mathcal{R}_\phi(f) = 0.5\mathbb{E}[(Ya(X) - b(X))\phi(-Yf(X))] \\ + 0.25\mathbb{E}[b(X) - Ya(X) + \bar{c}(Y, X)].\end{aligned}$$

Therefore, in the convexified risk, we replace the the indicator function with a convex function $\phi : \mathbf{R} \to \mathbf{R}$; see equation (A.1) in Appendix Section A.1.

Minimizing the convexified risk amounts to solving

$$\inf_{f:\mathcal{X}\to\mathbf{R}} \mathbb{E}[(Ya(X) - b(X))\phi(-Yf(X))].$$

Let f^*_ϕ be a solution to the convexified problem and let $\mathcal{R}^*_\phi = \inf_{f:\mathcal{X}\to\mathbf{R}} \mathcal{R}_\phi(f)$ be the optimal convexified risk. Next, we can define the excess convexified risk of a prediction $f : \mathcal{X} \to \{-1, 1\}$ as

$$\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi = 0.5\mathbb{E}\left[(Ya(X) - b(X))\left(\phi(-Yf(X)) - \phi(-Yf^*_\phi(X))\right)\right]. \quad (1)$$

The excess convexified risk measures the deviation of the convexified risk of a given decision rule $f : \mathcal{X} \to \{-1, 1\}$ from the optimal convexified risk and can be controlled. Unfortunately, the convexified excess risk tells us little about the

[12]In particular, there does not exist a polynomial time algorithm, unless P=NP. Heuristically, since $Y_i \in \{-1, 1\}$, the exact solution may require the brute-force search over $\{-1, 1\}^n$ which involves 2^n operations growing exponentially with the sample size. Even for a very small sample size with $n = 1,000$ observations, the number operations exceeds the number of atoms in the observable universe. The computational difficulty is the main reason why the binary classification problem is typically not solved directly in modern machine learning applications.

performance of the binary decision in terms of the actual excess risk that the decision maker cares about

$$\mathcal{R}(f) - \mathcal{R}^* = 0.5\mathbb{E}\left[(Ya(X) - b(X))\left(\mathbb{1}\{-Yf(X) \geq 0\} - \mathbb{1}\{-Yf^*(X) \geq 0\}\right)\right];$$

see equation (A.1) in Appendix Section A.1. In the following subsection, we show that the excess risk is bounded from above by the convexified excess risk, a result that we refer to as the main convexification theorem.

3.2 Assumptions and main convexification theorem

Let $\eta(x) = \Pr(Y = 1|X = x)$ be the conditional choice probability. We impose the following assumption on the conditional probability η and the loss function.

Assumption 3.1. *(i) $\ell_{-1,1}(x) > \ell_{1,1}(x)$ and $\ell_{1,-1}(x) > \ell_{-1,-1}(x)$ a.s. over $x \in \mathcal{X}$; (ii) $\eta \in (\epsilon, 1-\epsilon)$* a.s. *for some $\epsilon > 0$; (iii) there exists $M < \infty$ such that $|\ell_{f,y}(x)| \leq M$ a.s. over $x \in \mathcal{X}$ for all $f, y \in \{-1, 1\}$, where $\ell_{f,y}$ is defined in Proposition 2.1.*

Assumption 3.1 (i) requires that the losses from getting a wrong binary prediction outweigh the benefits of getting it right. (ii) requires that $\eta(x)$ is non-degenerate for almost all states of the world $x \in \mathcal{X}$. (iii) requires that the loss function is bounded and is satisfied in our empirical application.

Next, we restrict the class of convexifying function ϕ in the following assumption.

Assumption 3.2. *(i) $\phi : \mathbf{R} \to [0, \infty)$ is a convex and non-decreasing function with $\phi(0) = 1$ and $\phi(1) < \infty$; (ii) there exists some $L < \infty$ such that $|\phi(z) - \phi(z')| \leq L|z - z'|$ for all z, z'; (iii) there exist $C > 0$ and $\gamma \in (0, 1]$ such that for all $x, c \in (0, 1)$*

$$|x - c| \leq C\left(x + c - 2xc - \inf_{y \in \mathbf{R}} Q_c(x, y)\right)^{\gamma},$$

where $Q_c(x, y) = x(1 - c)\phi(-y) + (1 - x)c\phi(y), x, y \in \mathbf{R}$.

Assumption 3.2 is satisfied by the logistic, exponential, and hinge convexifying functions; see Section 3.3.

We will also use the following function which depends on the asymmetry of the loss function ℓ :

$$c(x) = \frac{a(x) + b(x)}{2b(x)}, \tag{2}$$

where $a(x)$ and $b(x)$ are defined in Proposition 2.1.

Our first result establishes the link between the optimal prediction f^* and the solution to the convexified risk minimization problem f^*_ϕ.

Proposition 3.1. *Suppose that Assumption 3.1 (i) is satisfied. Then the optimal prediction is*

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) > c(x), \\ -1 & \text{if } \eta(x) < c(x), \end{cases}$$

where $c(x)$ is defined in equation (2). Suppose additionally that Assumption 3.1 (ii) is satisfied and that ϕ satisfies Assumption 3.2 (i)-(ii). Then

$$\operatorname{sign}(f^*_\phi(x)) = \begin{cases} 1 & \text{if } \eta(x) > c(x), \\ -1 & \text{if } \eta(x) < c(x). \end{cases}$$

The optimal decision rule $f^*(x) = \operatorname{sign}(\eta(x) - c(x))$ is well-known, see, e.g., Boyes, Hoffman, and Low (1989), Schervish (1989), Granger and Pesaran (2000), or Elliott and Lieli (2013).[13] More importantly, we show that the optimal decision rule for the convexified problem that can be easily solved in practice coincides with f^*. Assuming additionally that correct predictions have zero benefits, the threshold value $c(x)$ simplifies to

$$c(x) = \frac{\ell_{1,-1}(x)}{\ell_{1,-1}(x) + \ell_{-1,1}(x)},$$

where $\ell_{f,y}(x) = \ell(f, y, x)$. In the symmetric case losses from false positive and false negative predictions are the same, i.e., $\ell_{1,-1}(x) = \ell_{-1,1}(x)$, so that $c(x) = 1/2$, which corresponds to the standard binary classification.

Note that the boundary separating the two decisions is $\{x \in \mathcal{X} : \eta(x) - c(x) = 0\}$. Intuitively, the more the distribution of X, denoted P_X, is concentrated around this boundary, the harder it gets to predict accurately $Y \in \{-1, 1\}$. The following condition generalizes the so-called noise or margin condition of Tsybakov (2004) and quantifies how much X is concentrated near the decision boundary.

Assumption 3.3. *Suppose that there exist $\alpha, C > 0$ such that for all $u > 0$*

$$P_X\left(\{x : |\eta(x) - c(x)| \leq u\}\right) \leq C u^\alpha.$$

If $\alpha = 0$, then Assumption 3.3 does not impose any restrictions as we can always take $C = 1$. Our next result relates the convexified excess risk bound to the excess risk bound of the binary prediction problem with arbitrary loss function under the margin condition.

[13] For completeness of presentation, we provide a concise proof in the Appendix.

Theorem 3.1. *Suppose that Assumptions 3.1, 3.2, and 3.3 are satisfied. Then for every measurable function* $f : \mathcal{X} \to \mathbf{R}$

$$\mathcal{R}(\mathrm{sign}(f)) - \mathcal{R}^* \lesssim \left[\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}.$$

It is worth noting that this upper bound relates the excess risk of the sign of f to the convexified risk of f (note that the convexified risk is well-defined for arbitrary $f : \mathcal{X} \to \mathbf{R}$). In practice, we would solve the convexified empirical risk minimization problem

$$\inf_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))\phi(-Y_i f(X_i)),$$

where $\mathscr{F}$ is a class of measurable functions $f : \mathcal{X} \to [-1, 1]$, where we restrict the range to $[-1, 1]$ since the risk is determined by the sign of f only. We discuss several convexifying functions ϕ in the remaining of this section.

3.3 Examples

We consider three examples of convexifying functions widely used in empirical applications. They are the logistic, exponential and hinge functions used respectively in the logistic regression, adaptive boosting, and the support vector machines; see Figure 1. Applying Theorem 3.1 allows us to implement suitably reweighted versions of thereof.

Example 3.1 (Logistic convexification)**.** *Lemma A.1.3 in the Appendix shows that the function* $\phi(z) = \log(1 + e^z)$ *satisfies Assumption 3.2 with* $\gamma = 1/2$ *and* $C = 2\sqrt{2}$. *The empirical risk minimization objective function is*

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))\log\left(1 + e^{-Y_i f(X_i)}\right).$$

In the symmetric case where $a(x) = 0$ *and* $b(x) = b$*, and therefore the empirical risk minimization problem is equivalent to the logistic regression*

$$f \mapsto -\frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-Y_i f(X_i)}\right).$$

The logistic convexifying function is also a popular choice in several implementations of the gradient boosting, e.g., the XGBoost algorithm[14] as well as the deep learning.

[14] The gradient boosting is typically computed iteratively using a suitable version of the functional gradient descent algorithm, often with some additional regularization.

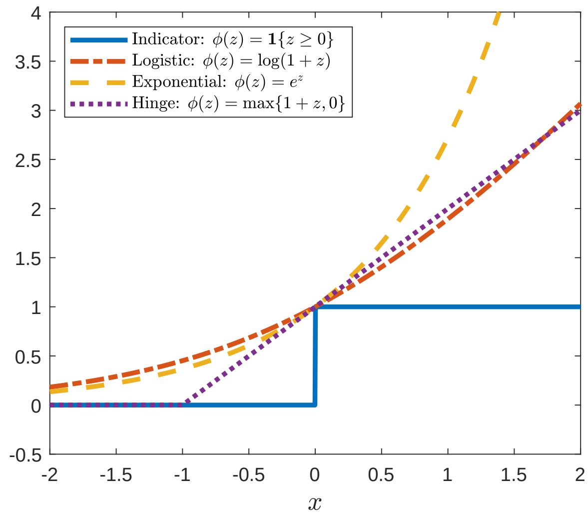


Figure 1: Convexifications corresponding to Logit, Boosting, and SVM.

This example, therefore, shows how a suitably reweighted for the asymmetries of the loss function logistic regression can be used for economic binary decisions.

Example 3.2 (Exponential convexification)**.** *Lemma A.1.5 shows that $\phi(z) = \exp(z)$, the exponential convexifying function, satisfies Assumption 3.2 with $\gamma = 1/2$ and $C = 2$. The empirical risk minimization objective function is*

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))e^{-Y_i f(X_i)}.$$

In the symmetric case it reduces to the objective function used in the adaptive boosting (AdaBoost)

$$f \mapsto -\frac{1}{n}\sum_{i=1}^{n} e^{-Y_i f(X_i)};$$

see Friedman, Hastie, and Tibshirani (2000).

This example, therefore, shows that one can use the adaptive boosting with a suitable reweighing for binary decision problems with a generic loss function.

Example 3.3 (Hinge convexification)**.** *Lemma A.1.4 shows that $\phi(z) = (1+z)_+$,*[15] *the hinge convexifying function, satisfies Assumption 3.2 with $\gamma = 1$ and $C = 1$. The*

[15] For $a \in \mathbb{R}$, $(a)_+ = \max\{a, 0\}$.

empirical risk minimization objective function is

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))(1 - Y_i f(X_i))_+.$$

In the symmetric case it reduces to the objective function used in the support vector machines

$$f \mapsto \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i f(X_i))_+,$$

see Vapnik (1995).[16] *The hinge convexification is also a popular choice for symmetric deep learning problems.*

The solution to the convexified empirical risk minimization problems, denoted $\hat{f}_n$, allows us to construct binary predictions taking $\operatorname{sign}(\hat{f}_n)$. Can we back out the estimates of conditional probabilities $\eta(x) = \Pr(Y = 1|X = x)$ from $\hat{f}_n$? Since $\hat{f}_n$ is an estimator of f_ϕ^*, we obtain from Lemma A.1.3 that

$$\eta(x) = \frac{1}{1 + \frac{1-c(x)}{c(x)} e^{-f_\phi^*(x)}}$$

for the Logistic convexifying function. Similarly, for the for the exponential convexifying function, Lemma A.1.5 gives

$$\eta(x) = \frac{1}{1 + \frac{1-c(x)}{c(x)} e^{-2f_\phi^*(x)}}.$$

Plugging $\hat{f}_n$ instead of f_ϕ^*, we obtain estimates of conditional probabilities $\hat{\eta}$ that can qualitatively supplement binary predictions if needed. In particular, in the symmetric case with $c(x) = 1/2$ (recall equation (2)), we recover the (nonparametric) estimate of the conditional probability that corresponds to the standard logistic regression.

4 Excess risk bounds

The convexified empirical risk minimization problem consists of minimizing the empirical risk

$$\widehat{\mathcal{R}}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))\phi(-Y_i f(X_i))$$

[16]The support vector machines has typically an additional Tikhonov regularization.

over some class of functions $f : \mathcal{X} \to [-1,1]$, denoted $\mathscr{F}_n$. Let $\hat{f}_n$ be a solution to $\inf_{f\in\mathscr{F}_n} \widehat{\mathcal{R}}_\phi(f)$, and let f_n^* be a solution to $\inf_{f\in\mathscr{F}_n} \mathcal{R}_\phi(f)$. Put also $\|f\|_q = (\mathbb{E}|f(X)|^q)^{1/q}, q \geq 1$ with a convention that $\|.\|_2 = \|.\|$. The following assumption requires that the risk function is sufficiently curved around the minimizer.

Assumption 4.1. *There exist some $c > 0$ and $\kappa \geq 1$ such that for every $f \in \mathscr{F}_n$*

$$\mathcal{R}_\phi(f) - \mathcal{R}_\phi^* \geq c\|f - f_\phi^*\|^{2\kappa}.$$

Since the convexified loss function depends on the convexifying function ϕ, verification of Assumption 4.1 is case-specific. We will see that it is satisfied with $\kappa = 1$ for the logistic and exponential functions and with $\kappa = 1 + 1/\alpha$ for the hinge function, where α is the margin parameter in Assumption 3.3.

We first state the oracle inequality for the excess risk in terms of the *local Rademacher complexity* of the class $\mathscr{F}_n$ defined as

$$\psi_n(\delta; \mathscr{F}_n) \triangleq \mathbb{E}\left[\sup_{f\in\mathscr{F}_n:\|f-f_n^*\|^2\leq\delta} |R_n(f - f_n^*)|\right],$$

where $R_n(f - f_n^*) = \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f_n^*(X_i))$ is a Rademacher process, i.e., $(\varepsilon_i)_{i=1}^n$ are i.i.d. in $\{-1, 1\}$ with probabilities 1/2. An attractive feature of this complexity measure is that it only depends on the local complexity of the parameter space in the neighborhood of the minimizer and provides a sharp description of the learning problem, cf. Koltchinskii (2006). At the same time, the local Rademacher complexities are general enough to provide a unified theoretical treatment for different methods and can be used to deduce oracle inequalities in many interesting examples. For a function $\psi : \mathbf{R}_+ \to \mathbf{R}_+$, put $\psi^\flat(\sigma) = \sup_{\delta\geq\sigma}[\psi(\delta)/\delta]$ and for a constant $\kappa \geq 1$, put $\psi_\kappa^\sharp(\epsilon) = \inf\left\{\sigma > 0 :\ \sigma^{1/\kappa-1}\psi^\flat(\sigma^{1/\kappa}) \leq \epsilon\right\}$. The transform $\psi_\kappa^\sharp$ is a suitable modification of the $\sharp$-transform introduced in Koltchinskii (2006) and describes the fixed point of the local Rademacher complexity in our setting. The following result holds.[17]

Theorem 4.1. *Suppose that Assumptions 3.1, 3.2, 3.3, and 4.1 are satisfied and $(Y_i, X_i)_{i=1}^n$ is an i.i.d. sample. Then there exist universal constants $c, \epsilon, K > 0$ such that for every $t > 0$ with probability at least $1 - ce^{-t}$*

$$\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^* \leq K\left[\psi_{n,\kappa}^\sharp(\epsilon) + \left(\frac{t}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \frac{t}{n} + \inf_{f\in\mathscr{F}_n} \mathcal{R}_\phi(f) - \mathcal{R}_\phi^*\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}.$$

[17] Proofs for all results in this section appear in Appendix Section A.2.

Theorem 4.1 tells us that the accuracy of the binary decision $\operatorname{sign}(\hat{f}_n)$ depends on the fixed point of the local Rademacher complexity of the class $\psi^\sharp_{n,\kappa}$, and the approximation error to the convexified risk of the optimal prediction. The accuracy also depends on the convexifying function through exponents γ and κ, as well as the margin parameter α. In the following subsections, we illustrate this result for parametric and nonparametric binary decision rules.

4.1 Parametric predictions

We start with illustrating our risk bounds for parametric binary decision rules. The parametric binary decision rule is defined $\operatorname{sign}(f_{\hat{\theta}})$ with $\hat{\theta}$ solving

$$\inf_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n (Y_i a(X_i) - b(X_i))\phi(-Y_i f_\theta(X_i)),$$

where $\Theta \subset \mathbf{R}^p$ such that $\{f_\theta : \theta \in \Theta\}$ is a convex subset of a linear subspace of $L_2(P_X)$. Linear predictions $f_\theta(x) = x^\top\theta$ with the logistic convexifying function $\phi(z) = \log(1+e^z)$ are the most popular choices. More generally, we have the following result for any convexifying function satisfying Assumption 3.2.

Theorem 4.2. *Under assumptions of Theorem 4.1*

$$\mathbb{E}\left[\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^*\right] \lesssim \left[\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \inf_{f\in\mathscr{F}_n}\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}.$$

In the parametric approach, we typically assume that the approximation error is zero. It follows from Lemmas A.1.3, A.1.4, A.1.5, A.2.1, and A.2.3 that for the logistic and the exponential functions $\gamma = 1/2$ and $\kappa = 1$ while for the hinge function $\gamma = 1$ and $\kappa = 1 + 1/\alpha$. Therefore, in all three cases, for parametric predictions we obtain

$$\sup_{P\in\mathcal{P}(\alpha)}\mathbb{E}[\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^*] \lesssim \left(\frac{p}{n}\right)^{\frac{1+\alpha}{2+\alpha}},$$

where $\mathcal{P}(\alpha)$ is a set of distributions restricted in Theorem 4.2. If the dimension of covariates p is fixed, then the convergence rate can be anywhere between $O(n^{-1/2})$ and $O(n^{-1})$ depending on margin exponent α. To achieve a better performance when p is large compared to n, one can also add the ℓ_1 and/or ℓ_2 penalties to the empirical risk minimization problem; see Belloni, Chernozhukov, Chetverikov, Hansen, and Kato (2018). Note also that there is virtually no difference between the three convexifying functions apart from constants, so one can safely use the logistic regression with the appropriate reweighing for the asymmetry of the loss function.

4.2 Boosting

Boosting amounts to combining several "weak" binary decisions into more sophisticated and powerful decision rules; see Schapire, Freund, Bartlett, and Lee (1998). The weak binary decision are typically constructed with shallow decision trees. An interesting feature of boosting is that the weak decisions may be only slightly better than the random guessing while the ultimate combined binary decision may achieve the outstanding out-of-sample performance. For the asymmetric loss function, the boosting amounts to solving the following empirical risk minimization problem:

$$\inf_{f\in\mathscr{F}^{\mathrm{B}}} \frac{1}{n}\sum_{i=1}^{n}(Y_i a(X_i) - b(X_i))\phi(-Y_i f(X_i)),$$

where $\mathscr{F}^{\mathrm{B}} = \mathscr{F}^{\mathrm{B}}_{\lambda}$ is a class of a weighted sum of weak decisions

$$\mathscr{F}^{\mathrm{B}} = \left\{\sum_{j=1}^{J} w_j g_j(x) :\ |w|_1 \leq \lambda,\ g_j \in \mathscr{G},\ J \in \mathbf{N}\right\},$$

and $\mathscr{G}$ is a base class of weak predictions. Exponential convexifying function $\phi(z) = \exp(z)$ is a popular choice, but the logistic function is also often considered; see Friedman, Hastie, and Tibshirani (2000) and Blanchard, Lugosi, and Vayatis (2003). The problem is then solved using a functional version of the gradient descent algorithm, often with additional regularization and tuning.[18] Let $\hat{f}_n$ be a solution of the empirical risk minimization problem described above, then:

Theorem 4.3. *Suppose that $\mathscr{G}$ is a measurable class of functions from $\mathcal{X}$ to $[-1,1]$ with* VC*-dimension $V \geq 1$ and that ϕ is a logistic or exponential function. Then under assumptions of Theorem 4.1*

$$\mathbb{E}\left[\mathcal{R}(\mathrm{sign}(\hat{f})) - \mathcal{R}^*\right] \leq K\left[\left(\frac{C_V}{n}\right)^{\frac{2+V}{2(1+V)}} + \inf_{f\in\mathscr{F}^{\mathrm{B}}}\|f - f^*_{\phi}\|_{L_1(P_X)}\right]^{\frac{\alpha+1}{\alpha+2}}$$

for some constants $K, C_V > 0$, where C_V depends on V.

Note that the statistical accuracy of a binary decision is driven by the the complexity of the base class $\mathscr{G}$, which should have as low VC dimension as possible to minimize its impact on the first term and, at the same time, it should generate a sufficiently rich class $\mathscr{F}$ to make the approximation error as small as possible. We provide two examples of the baseline classes $\mathcal{G}$ below.

[18] Popular implementations include AdaBoost, XGBoost, or LightGBM.

Example 4.1 (Linear decision rules)**.** *Consider the class of signs of linear functions*

$$\mathscr{G} = \left\{x \mapsto 2\mathbb{1}_{x^\top a \leq b} - 1 : \ a \in \mathbf{R},\ b \in \mathbf{R}^d\right\}.$$

The VC*-dimension of* $\mathscr{G}$ *is* $V = d + 1$ *and the approximation error tends to zero for every* P_X *as* $\lambda \to \infty$*, cf., Cybenko (1989) and Hornik, Stinchcombe, and White (1989).*

Example 4.2. *Let* $(R_k)_{k=1}^K$ *be a tree-structured partition of* $\mathcal{X}$ *with cuts parallel to coordinate axes. Consider the class of decision trees with* K *terminal nodes*

$$\mathscr{G} = \left\{x \mapsto 2\sum_{k=1}^K \mathbb{1}_{R_k}(x) - 1\right\}.$$

The VC-dimension of $\mathscr{G}$ *is* $V \leq d\log(2d)$*, cf., Devroye, Györfi, and Lugosi (1996) and the approximation error tends to zero as* $\lambda \to \infty$*, cf., Breiman (2000).*

4.3 Shallow learning

The shallow learning amounts to fitting a neural network with one or two hidden layers. Neural networks are widely used in econometrics at least since Gallant and White (1988).[19] We focus on a very simple neural network class consisting of two hidden layers. Following the recent trends in big data applications, we refer to this approach shallow learning which in contrast to the deep learning does not allow for the number of layers to scale with the sample size. Consider a single layer neural network class

$$\Theta_n^{\mathrm{S}} = \left\{x \mapsto \sum_{j=1}^{p_n} b_j\sigma_0(a_j^\top x + a_0) + b_0, \quad a \in \mathbf{R}^{d+1},\ |b|_1 \leq \gamma_n\right\},$$

where $a = (a_0, a_1, \dots, a_d)$ and $b = (b_0, b_1, \dots, b_{p_n})$ are parameters to be estimated, σ_0 is some smooth function, and $|.|_1$ is the ℓ_1 norm. The shallow learning class is

$$\mathscr{F}_n^S = \left\{x \mapsto \sigma(\theta(x) + c(x)d + 1) - \sigma(\theta(x) + c(x)d - 1) - 1 : \ \theta \in \Theta_n^{\mathrm{S}},\ |d| \leq n\right\},$$

[19]Conceptually, neural networks can be traced to early mathematical models of the brain, see McCulloch and Pitts (1943) and Rosenblatt (1958). Among the early econometric applications, we may quote the nonlinear time series modeling, see Granger (1995), Lee, White, and Granger (1993), and Gallant and White (1992); prediction, see White and Racine (2001) and Chen, Racine, and Swanson (2001); asset pricing, see Hutchinson, Lo, and Poggio (1994) and Chen and Ludvigson (2009), among others. As noted in footnote 7 see also the comprehensive review of Chen (2007) covering various aspects of neural networks and other nonlinear sieves.

where $c(x)$ pertains to asymmetry (cf. equation (2)) and $\sigma(z) = (z)_+$ is the Rectified Linear Unit (ReLU) activation function. The shallow learning class can be visualized on a directed graph, see Figure 2. All covariates are fed first into the hidden layer 1 corresponding to Θ_n^{S}. This network class consists of p_n neurons with a smooth activation function σ_0. The output produced by the hidden layer 1 is fed subsequently into the two neurons with the ReLU activation function σ. Note that this last layer has a single free parameter b. The output $\hat{f}_n \in [-1, 1]$, also called a soft prediction, is obtained from summing up the two neurons from the ReLU layer. The final binary prediction is defined as a sign of the soft prediction $\hat{f}_n$.

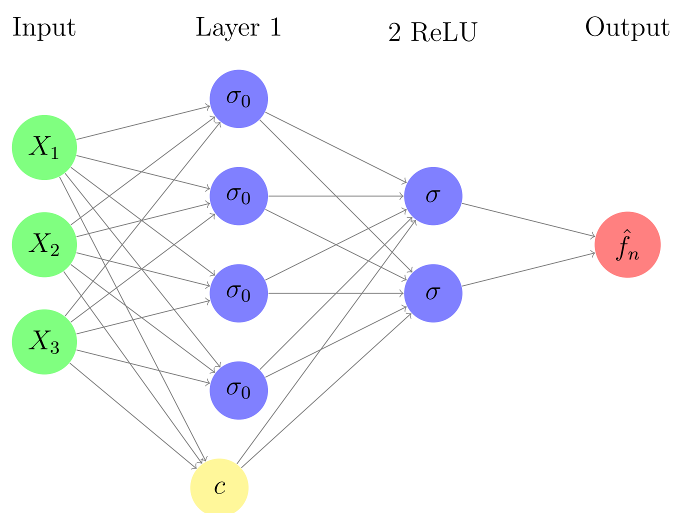


Figure 2: Directed graph of the shallow learning architechture with $d = 3$ covariates, single hidden layer with 4 neurons, and 2 outer ReLU neurons. The yellow neuron takes covariates $X \in \mathbf{R}^d$ as an input and produces $c(X) \in \mathbf{R}$, which is fed directly in 2 ReLU neurons.

The soft shallow learning prediction $\hat{f}_n : \mathcal{X} \to [-1, 1]$ is a solution to the convexified empirical risk minimization problem with hinge convexifying function

$$\inf_{f \in \mathscr{F}_n^{\mathrm{S}}} \frac{1}{n} \sum_{i=1}^{n} (Y_i a(X_i) - b(X_i))(1 - Y_i f(X_i))_+.$$

To describe the accuracy of the shallow learning binary decision $\operatorname{sign}(\hat{f}_n)$, consider the Sobolev ball of smoothness $\beta \in \mathbf{N}$ and radius $M \in (0, \infty)$

$$W_M^{\beta,\infty}(\mathcal{X}) = \left\{ f : \mathcal{X} \to \mathbf{R} : \max_{|k| \leq \beta} \operatorname*{ess\,sup}_{x \in \mathcal{X}} |D^k f(x)| \leq M \right\},$$

where we use the multi-index notation $k = (k_1, \ldots, k_d) \in \mathbf{N}^d$, $|k| = k_1 + \cdots + k_d$, and $D^k f = \frac{\partial^{|k|}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} f$.

The following assumption imposes some mild regularity conditions on the activation function σ_0 and the Sobolev smoothness of the conditional probability η.

Assumption 4.2. *(i) $\sigma_0 : \mathbf{R} \to [-b, b]$ is non-decreasing and Lipschitz continuous, infinitely differentiable on some open interval containing some x_0 with $D^k \sigma_0(x_0) \neq 0$ for all $k \in \mathbf{Z}_+$; (ii) $\eta \in W_M^{\beta,\infty}(\mathcal{X})$ for some $\beta \in \mathbf{N}$ and $0 < M < \infty$, where $\mathcal{X} \subset \mathbf{R}^d$ is a Cartesian product of compact intervals; (iii) p_n and γ_n are of polynomial order.*

Assumption 4.2 (i) rules out polynomial activation functions and allows for the sigmoid function $\sigma_0(x) = (1 + e^{-x})^{-1}$. It also rules out the ReLU activation function, which is a more natural choice for the deep learning problems and is considered in the subsequent section.

Theorem 4.4. *Suppose that $(Y_i, X_i)_{i=1}^n$ is an i.i.d. sample following a distributions satisfying Assumptions 3.1, 3.2, 3.3, and 4.2, and denoted $\mathcal{P}(\alpha, \beta)$. Then there exist universal constants $c, C > 0$ such that for every $t > 0$ with probability at least $1 - ce^{-t}$*

$$\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^* \leq C \left[\left(\frac{p_n \log^2 n}{n} \right)^{\frac{1+\alpha}{2+\alpha}} + p_n^{-(1+\alpha)\beta/d} + \left(\frac{t}{n} \right)^{\frac{1+\alpha}{2+\alpha}} + \frac{t}{n} \right]$$

uniformly over $\mathcal{P}(\alpha, \beta)$. In particular,

$$\sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_P \left[\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^* \right] \lesssim \left(\frac{\log^2 n}{n} \right)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta + d}}$$

provided that $p_n \sim (n / \log^2 n)^{\frac{d}{(2+\alpha)\beta+d}}$.

It is worth noting that the convergence rate of the excess risk of the shallow learning prediction can be anywhere between the slow nonparametric rate $O(n^{-\beta/(2\beta+d)})$ and the fast rate $O(n^{-1})$ depending on the margin exponent α. In particular, we can partially offset the curse of dimensionality if α is sufficiently large, e.g., for $\alpha = d/\beta$, the rate is $O(n^{-1/2})$. This is another manifestation of the fact that predicting a binary outcome is easier than predicting real-valued variables. Note also that the smoothness of the decision boundary itself $\{x \in [0,1]^d : \eta(x) - c(x) = 0\}$ does not directly play a role and only the smoothness of η is important. This is probably not surprising in light of the fact that c is known to the decision maker.

In the special case when the loss function is symmetric, under the mild assumption that the density of covariates is uniformly bounded, it follows from Audibert and Tsybakov (2007), Theorem 4.1 that there exists $C > 0$ such that for every $n \geq 1$

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}(\alpha,\beta)} \mathbb{E}_P \left[\mathcal{R}(\hat{f}_n) - \mathcal{R}^* \right] \geq C n^{-\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}}, \tag{3}$$

where the infimum is taken over all binary decisions $\hat{f}_n : \mathcal{X} \to \{-1, 1\}$ computed from an i.i.d. sample $(Y_i, X_i)_{i=1}^n$. Therefore, apart for a $\log n$ factor, our result shows that the shallow learning is the minimax optimal.

4.4 Deep learning

In this final subsection, we illustrate our excess risk bound for deep learning, cf., Goodfellow, Bengio, and Courville (2016). The deep learning amounts to fitting a neural network with several hidden layers, also known as a deep neural network.[20] Recent empirical research suggests that the such multi-layer neural networks may outperform shallow neural networks not only for traditional numerical data, see Chen, Pelger, and Zhu (2019) and Gu, Kelly, and Xiu (2020), but also for non-standard data such as images, text, and speech data that receive an increasing recent attention in econometrics.[21]

Fitting a deep neural network requires choosing the activation function $\sigma : \mathbf{R} \to \mathbf{R}$ and the network architecture. We focus on the ReLU activation function,

$$\sigma(z) = \max\{z, 0\},$$

which is the most popular choice for deep networks.[22]

The network architecture consists of d neurons corresponding to covariates $X = (X_1, \dots, X_d) \in \mathbf{R}^d$, one output neuron corresponding to the soft prediction $\hat{f}_n \in [-1, 1]$, and a number of hidden neurons. The final prediction is obtained with

[20] Multilayer neural network have been contemplated at least since Ivakhnenko and Lapa (1965), but their popularity increased tremendously only with more recent improvements in the computational power, optimization algorithms, and the empirical success.

[21] See Schmidt-Hieber (2017), Farrell, Liang, and Misra (2019), and Bauer and Kohler (2019) for a recent theoretical treatment of deep neural networks in the regression and semiparametric inference settings.

[22] Other activation functions used in the deep learning include: leaky ReLU, $\sigma(z) = \max\{\alpha z, 0\}, \alpha > 0$; exponential linear unit (ELU), $\sigma(z) = \alpha(e^z - 1)\mathbb{1}\{z < 0\} + z\mathbb{1}\{z \geq 0\}$; and scaled ELU.

$\text{sign}(\hat{f}_n) \in \{-1, 1\}$. Hidden neurons are grouped in L layers, known as the *depth* of the network. A hidden neuron $j \geq 1$ in a layer $l \geq 1$ operates as

$$z \mapsto \sigma(z^\top a_j^{(l)} + b_j^{(l)}),$$

where z is the output of neurons from the layer $l-1$ and $a_j^{(l)}, b_j^{(l)}$ are free parameters. The last layer and the output neuron produce together

$$z \mapsto \sigma(z + b^{(L)}c(x) + 1) - \sigma(z + b^{(L)}c(x) - 1) - 1 \in \mathbf{R},$$

where $b^{(L)}$ is the parameter to be estimated and $c(x)$ is a known decision cut-off function. The network architecture $(L, \mathbf{w})$ is described by the number of hidden layers L and a *width* vector $\mathbf{w} = (w_1, \ldots, w_L)$, where w_l denotes the number of hidden neurons at a layer $l = 1, 2, \ldots, L$. For completeness, put also $w_0 = d$ and $w_{L+1} = 2$. Graphically, the deep network can be arranged on a graph, see Figure 3. We focus on *feed-forward networks*, which means that we rule out backwards connections and loops. The network displayed in Figure 3 is called *fully connected* because it contains all possible connections between neurons on adjacent layers.

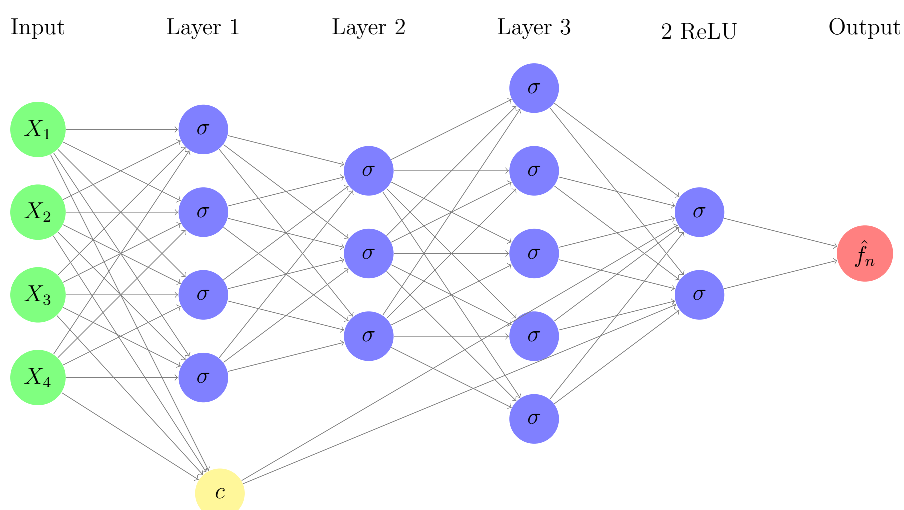


Figure 3: Directed graph of the deep learning architechture with $d = 4$ covariates, $L = 3$ hidden layers of width $\mathbf{w} = (4, 3, 5)$ neurons, and 2 outer ReLU neurons. The yellow neuron takes covariates $X \in \mathbf{R}^d$ as an input and produces $c(X) \in \mathbf{R}$, which is fed directly in 2 ReLU neurons.

Our final deep learning architecture $(L, \mathbf{w})$ is a natural extension of the shallow neural network

$$\mathscr{F}_n^{\text{DNN}} = \left\{x \mapsto \sigma(\theta(x) + c(x)d + 1) - \sigma(\theta(x) + c(x)d - 1) - 1 : \ |d| \leq n, \ \theta \in \Theta_n^{\text{DNN}}\right\},$$

with the inner layer replaced by the deep neural network

$$\Theta_n^{\mathrm{DNN}} = \left\{ x \mapsto A_{L-1}\sigma_{\mathbf{b}_{L-1}} \circ \cdots \circ A_1 \sigma_{\mathbf{b}_1} \circ A_0 x \right\},$$

where each A_l is $w_{l+1} \times w_l$ matrix of network weights and for two vectors $y = (y_1, \dots, y_r)$ and $\mathbf{b} = (b_1, \dots, b_r)$ (a bias vector), we put

$$\sigma_{\mathbf{b}} \circ y = \begin{pmatrix} \sigma(y_1 + b_1) \\ \vdots \\ \sigma(y_r + b_r) \end{pmatrix}.$$

The following assumption restricts the smoothness of the conditional probability and imposes some assumptions on how the network architecture should scale with the sample size. For simplicity, we define the width of the network as the maximum width across all layers and denote it as $W_n = \max_{1 \leq l \leq L} w_l$.

Assumption 4.3. *(i) $\eta \in W_M^{\beta,\infty}[0,1]^d$ for some $M > 0$ and $\beta \in \mathbf{N}$; (ii) the neural network architecture is such that the depth is $L_n = O(K_n \log K_n)$ and the width is $W_n = O(J_n \log J_n)$ for some $J_n, K_n \in \mathbf{N}$ of polynomial order satisfying $J_n K_n \sim n^{\frac{d}{2\beta(2+\alpha)+2d}}$.*

It is worth mentioning that we allow for neural networks with a fixed depth L_n and increasing width W_n as well as for deep networks with increasing depth L_n and a fixed width W_n. The key requirement is that the product $J_n K_n$ increases at the rate specified in Assumption 4.3 (ii). Let $\mathscr{F}_n^{\mathrm{DNN}}$ be a set of neural networks with the architecture satisfying Assumption 4.3, where weights and biases $\{A_0, A_l, b_l, l = 1, \dots, L\}$ are allowed to take arbitrary real values.

The deep learning soft prediction $\hat{f}_n$ is a solution to the empirical risk minimization problem with the hinge convexification[23]

$$\inf_{f \in \mathscr{F}_n^{\mathrm{DNN}}} \frac{1}{n} \sum_{i=1}^{n} (Y_i a(X_i) - b(X_i))(1 - Y_i f(X_i))_+.$$

The following result holds for the binary decision estimated with the deep learning.

[23] Our focus on the hinge convexification function is motivated by the objective of achieving the minimax optimal convergence rates. It is not obvious whether the logistic convexification can achieve the minimax optimal convergence rate and we leave more detailed investigation of this for future research.

Theorem 4.5. *Suppose that $(Y_i, X_i)_{i=1}^n$ is an i.i.d. sample from a distribution satisfying Assumptions 3.1, 3.2, 3.3, and 4.3, and denoted $\mathcal{P}(\alpha,\beta)$. Then there exist universal constants $c, C > 0$ such that for every $t > 0$ with probability at least $1 - ce^{-t}$*

$$\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^* \leq C\left[\left(\frac{\log^6 n}{n}\right)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}} + \left(\frac{t}{n}\right)^{\frac{1+\alpha}{2+\alpha}} + \frac{t}{n}\right]$$

uniformly over $\mathcal{P}(\alpha,\beta)$. In particular,

$$\sup_{P\in\mathcal{P}(\alpha,\beta)} \mathbb{E}_P\left[\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^*\right] \lesssim \left(\frac{\log^6 n}{n}\right)^{\frac{(1+\alpha)\beta}{(2+\alpha)\beta+d}}.$$

Deep learning vs. shallow learning. Theorem 4.5 shows that the deep learning binary decision achieves the minimax optimal convergence rate apart from the $\log n$ factor. The same convergence rate is achieved by the shallow learning with a single sigmoid layer. Since it is impossible to improve upon the minimax optimal convergence rate, this raises the question on the relative merits of the deep and the shallow learning and their advantage compared to other nonparametric methods that may also be minimax-optimal, such as the kernel smoothers. For the nonparametric regression problems, this question has been addressed in several recent papers demonstrating that the deep learning may achieve better convergence rates for smaller classes of functions and we also expect that similar results could be obtained for the nonparametric binary decision problems studied in the present paper.

For example, if the conditional probability η depends only on a smaller number of covariates $d^* < d$, or more generally satisfies the generalized hierarchical model of order $d^* < d$, we expect that apart of the $\log n$ factors, the rate might be improved to $O(n^{-(1+\alpha)\beta/(2+\alpha)\beta+d^*})$, offsetting the curse of the dimensionality; see Kohler and Langer (2020). This assumption is similar to the sparsity in the linear regression case and might be plausible in various real-world applications.

Second, when the number of covariates d is large, assuming that the choice probability η has the same smoothness β in all directions may be overly restrictive in the real-world applications. A more plausible assumption is that the smoothness of η is heterogeneous over $[0,1]^d$, e.g., we may have the anisotropic smoothness in each coordinate described by a vector $\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_d)$; see Suzuki and Nitanda (2019). We expect that for the anisotropic smoothness classes, the rate may be improved to $O(n^{-(1+\alpha)\beta^*/((2+\alpha)\beta^*+1)})$ with $\beta^{*-1} = \sum_{j=1}^d \beta_j^{-1}$, offsetting the curse of the dimensionality at another level.[24]

[24]Note that for kernel smoothers, capturing the anisotropic smoothness requires selecting a dif-

5 A simulation study

Anticipating the empirical application to the economic prediction of recidivism, we report on a simulation study pertaining to pretrial detention decisions. As explained in the next section, we present a design pertaining to judges who needs to decide whether to release an offender, facing the possibility that the defendant might commit other crimes versus keeping in jail a defendant who would obey the law and the terms of the release. There are two groups, $G = 0$ and 1, with one group assumed to be a protected segment of the population. We set $Y =$ -1 if the person does not commit a crime upon pretrial release, and $Y = 1$ otherwise. This is a much studied topic and we approach it from a social planner point of view using a simplified stylized example for the purpose of a Monte Carlo simulation study with a loss function from Example 2.2:

	G = 0		G = 1	
	$f(0,z) = 1$	$f(0,z) = -1$	$f(1,z) = 1$	$f(1,z) = -1$
Y = 1	0	ψ_0	0	ψ_1
Y = -1	φ_0	0	φ_1	0

where z is a vector of observable characteristics and $\psi_g, \varphi_g > 0$ for $g \in \{0, 1\}$. From the above, we do not suffer any losses (or gains) when a defendant being released becomes a productive member of society or when a defendant who would commit another crime is kept in jail. This is of course an simplification for the purpose of keeping the simulation design simple.

Keeping someone in jail not intending to commit another crime comes with costs $\varphi_g, g \in \{0, 1\}$. In particular, if $\varphi_1 > \varphi_0$, then the cost of keeping an individual in jail not indenting to commit another crime is higher if that individual is in the group $G = 1$. Similarly, releasing a recidivist comes with costs $\psi_g, g \in \{0, 1\}$, so that if $\psi_1 \neq \psi_0$, the costs of releasing a recidivist is different for the protected group and everyone else.[25]

ferent bandwidth parameter for each coordinate, which may be statistically too costly if d is large. Similarly, a linear tensor product sieve requires selecting a different number of series terms for each coordinate.

[25]It is worth mentioning that when predicting defaults the bank might care more about the false negative mistakes (failing to predict defaults), while the social planner might be more concerned

According to equation (2), the threshold between the two binary decisions is:

$$c(g,z) = \frac{\varphi_g}{\varphi_g + \psi_g}, \qquad g \in \{0,1\}$$

and according to Proposition 3.1 the optimal decision rule is $f^*(g,z) = \operatorname{sign}(\eta(g,z) - c(g,z))$ with $\eta(g,z) = \Pr(Y=1|G=g, Z=z)$. Note also that $a(g,z) = \psi_g - \varphi_g$ and $b(g,z) = -(\psi_g + \varphi_g)$. The design of the data generating process is

$$Y = 2\mathbb{1}\left\{2G + Z^\top\gamma + \tau\left(\frac{1}{d}\sum_{j=1}^d Z_j^2 + 2Z_1\sum_{j=2}^d Z_j\right) \geq \varepsilon\right\} - 1,$$

where $\varepsilon \sim N(0,1)$, $G \sim \text{Bernoulli}(\rho)$, and $Z_1, \dots, Z_d \sim_{i.i.d.} N(0,1)$. Therefore, the protected segment is a fraction ρ of the population (determined by the Bernoulli distribution parameter) and the conditional probability is

$$\begin{aligned} \eta(g,z) &= \Pr(Y=1|G=g, Z=z) \\ &= \Phi\left(2G + Z^\top\gamma + \tau\left(\frac{1}{d}\sum_{j=1}^d Z_j^2 + 2Z_1\sum_{j=2}^d Z_j\right)\right), \end{aligned} \tag{4}$$

where Φ is the CDF of $N(0,1)$. Note that the DGP may feature a very simple example of nonlinearities with quadratic terms and interactions with Z_1 and that setting $\tau = 0$, we obtain the linear model. Lastly, we also set $\gamma = (1, 0.9, 0.8, 0, 0, \dots, 0)^\top$.

Let $(Y_i, G_i, Z_i)_{i=1}^n$ be i.i.d. draws of (Y, G, Z). To evaluate the performance of our approach, we split the sample into the training or estimation sample $(Y_i, G_i, Z_i)_{i=1}^{n_e}$ and the test sample $(Y_i, G_i, Z_i)_{i=n_e+1}^n$. For parametric predictions, we estimate the decision rule solving the weighted logistic regression over the class of linear functions $\{(g,z) \mapsto \theta_0 + \theta_1 g + z^\top\gamma : \ \theta_0, \theta_1 \in \mathbf{R}, \gamma \in \mathbf{R}^{d-1}\}$. Note that according to our theory if $\tau = 0$, then the weighted linear logistic regression provides valid binary predictions even if the choice probabilities are not Logistic; see Eq. 4. However, since in practice we typically do not know the parameteric class that can capture all the relevant nonlinearities (i.e., that $\tau \neq 0$), we would often estimate the linear prediction rule

$$\min_{(\theta_0,\theta_1,\gamma)\in\mathbf{R}^{d+2}} \frac{1}{n_e}\sum_{i=1}^{n_e}(Y_i(\psi_{G_i} - \varphi_{G_i}) + (\psi_{G_i} + \varphi_{G_i}))\log\left(1 + e^{-Y_i(\theta_0 + \theta_1 G_i + Z_i^\top\gamma)}\right).$$

with the false positive mistakes (failing to predict that the loan will be repaid) and equal credit opportunities regardless of the group membership. Since our general framework can be applied to different economic binary prediction problems, in this simulation study, we will look at how both mistakes change with φ_g and ψ_g for $g \in \{0,1\}$.

Then the estimated prediction rule is $(g,z) \mapsto \operatorname{sign}(\hat{\theta}_0 + \hat{\theta}_1 g + z^\top \hat{\gamma})$, where $(\hat{\theta}_0, \hat{\theta}_1, \hat{\gamma})$ are estimated above, and the binary predictions evaluated on the test sample are

$$\operatorname{sign}(\hat{\theta}_0 + \hat{\theta}_1 G_i + Z_i^\top \hat{\gamma}), \qquad i = n_e + 1, \dots, n.$$

To obtain binary predictions with neural networks, we solve

$$\min_{f \in \mathscr{F}_n^{\mathrm{NN}}} \frac{1}{n_e} \sum_{i=1}^{n_e} (Y_i(\psi_{G_i} - \varphi_{G_i}) + (\psi_{G_i} + \varphi_{G_i}))(1 - Y_i f(G_i, Z_i))_+,$$

where $\mathscr{F}_n^{\mathrm{NN}}$ is a relevant neural network class. Then the estimated prediction rule is $(g,z) \mapsto \operatorname{sign}(\hat{f}(g,z))$, and the binary predictions evaluated on the test data are

$$\operatorname{sign}(\hat{f}(G_i, Z_i)), \qquad i = n_e + 1, \dots, n.$$

We set $n = 100$ and 1,000 with 30 % set aside as test sample in the simulations - corresponding to a relatively small sample compared to what is often found in applications. Hence, the design emphasizes how good our asymptotic results are in small samples.

To benchmark our asymmetric binary choice approach, we focus first on the unweighted approach with the logistic regression, gradient boosted trees, shallow and deep learning, and support vector machines. For each method, we compute the group-specific false positive (FP) and false negative (FN) mistakes estimating

$$\begin{aligned} \mathrm{FP}_g &= \Pr(\operatorname{sign}(\hat{f}(X)) = 1, Y = -1 | G = g) \\ \mathrm{FN}_g &= \Pr(\operatorname{sign}(\hat{f}(X)) = -1, Y = 1 | G = g) \end{aligned}$$

for $g \in \{0, 1\}$ on the test sample. We also compute the total misclassification error estimating

$$\mathrm{Error} = \Pr(\operatorname{sign}(\hat{f}(X)) \neq Y)$$

on the test sample. We use TensorFlow, scikit-learn, and XGBoost packages in Python to compute machine learning methods. The gradient boosted trees are computed with the number of trees selected using 10-fold cross-validated. All other parameters are kept to their default values in the XGBoost package. The regularization parameter of the support vector machines is computed using the 10-fold cross-validation. We also use the radial basis function and the default value of the scaling parameter in the scikit-learn package. The neural networks have width of 15 neurons. It is worth stressing that we use the architectures described in Figures 2 and 3 with two outer ReLU units. For the shallow learning, we use the sigmoid activation function. The deep neural network has 5 hidden layers.

Results appear in Table 2. We find that in terms of the total misclassification error, the ML methods outperform the logistic regression for the nonlinear DGP. Interestingly, the neural networks outperform the logistic regression when we only have $n = 100$ observations. In this case, we observe almost 4-fold reduction in the total misclassification error for the nonlinear DGP and, srikingly, the neural networks outperform the Logistic regression even for the linear DGP. The deep learning and the shallow learning perform similarly in general, except for the case of the linear DGP with $\rho = 0.5$.[26] Importantly, we observe the disproportionate number of false positive and false negative mistakes across two groups in many cases.

Next, we investigate whether group-specific misclassification rates can be equalized across two groups with weighted ML methods. For simplicity, we focus on the the setting with $\tau = 0$, $\rho = 0.2$, and $n = 1,000$, as in this case we observe a disproportionate number of FP and FN across the two groups. Figure 4 shows that the asymmetric weighted logistic regression can equalize the FP probabilities across the two groups for $\varphi_0 \approx 1.65$ and FN probabilities across the two groups for $\psi_0 \approx 3$. Note that equalizing the FP probabilities comes with the increase in FN probabilities in the group $G = 0$ and equalizing the FN probabilites comes with the increase in the FP probabilities in the group $G = 0$. Therefore, the decision maker or the social planner has to think carefully about all these trade-offs when calibrating the asymmetric loss function.[27] The results for the gradient boosting are similar, except for the fact that larger weight factors are required to equalize FP/FN probabilities across groups.

Lastly, we compare the performance of the standard logistic regression approach, which ignores the asymmetric loss function, with the asymmetric logistic regression in terms of the average loss of a social planner. The social planner loss for the former will be denoted by ℓ_{logit} while our new estimator yields $\ell_{w-logit}$.

Simulation results are reported in Table 3. We report several measures to appraise the findings. First, we report $\mathrm{P}(\ell_{logit} > \ell_{w-logit})$ which is the percent that the standard logistic regressions generate larger social planner costs compared to

[26] In our experience, the deep ReLU network is computationally more stable with less variability across MC experiments as opposed to the shallow a single layer sigmoid network. Note also that the single layer ReLU network might have important approximation-theoretic limitations as piecewise constant approximations typically can not take advantage of the higher-order smoothness.

[27] Note that we estimate the FP and FN probabilities splitting the sample into two parts, also known as the validation set approach. In practice, the K-fold cross-validation may provide better estimates of these probabilities.

our weighted regression. Hence, this measure reports how often our estimator outperforms the standard procedure, where the probabilities are computed from the Monte Carlo simulated samples. Next, we report summary statistics for the ratio $\ell_{logit}/\ell_{w-logit}$. When the ratio is above 1.00 then the weighted logistic approach is better. We report the minimum, maximum, mean, and three quartiles of the simulation distribution. All simulations involve 5000 replications.

We start from a baseline case for the parameter setting, namely: $\psi_0 = 3$, $\psi_1 = 1$, $\varphi_0 = 1.7$ and finally $\varphi_1 = 1$. Moreover, the dimension d of covariates Z is set to 15. For the baseline case and $n = 1{,}000$, $\mathrm{P}(\ell_{logit} > \ell_{w-logit})$ is 0.73, meaning that in a large majority of cases our procedure is superior (lower social costs) to the standard logistic regression. According to the summary statistics for the ratio $\ell_{logit}/\ell_{w-logit}$ the mean and median is roughly 1.07/1.06, meaning a 6-7 % reduction in costs, with a max of 1.61 (60% reduction) and a min of 0.74. We also examine various deviations from the baseline case. For the larger sample size $n = 5{,}000$ the gains are similar, although with larger values for $\mathrm{P}(\ell_{logit} > \ell_{w-logit})$, which is now 0.94, meaning that the probability that the symmetric logistic regression leads to larger losses increases as we get more data. Note that the min and max of the distribution move in opposite direction, with the latter now 1.29.

Next we report two columns in Table 3 where we change the fraction of protected population from 0.2 to respectively 0.5. These changes do not seem to have a significant impact on any of the results. In contrast, however, if we change the cost structure we see, as might be expected, more variation. More precisely, introducing more asymmetries in the loss function implies better performance of the asymmetric logistic regression.

6 Pretrial Detention Decisions - Racial Bias and Recidivism Revisted

The U.S. criminal justice system costs have skyrocketed over the past decades. In California, thirty years ago, 10% of the state general fund went to higher education and 3% went to prisons; today, 11% goes to prisons and 7.5% to higher education according to figures quoted by Baughman (2017). The main purpose of this section is to apply our novel econometric methods to the problem of recidivism and bias in pretrial detention. At the outset we should note that we only provide a succinct discussion here of a vast number of papers written by scholars across different fields and do not try to cite all relevant works.

Among economists, the idea to apply machine learning to pretrial detention decisions has recently been explored by Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018), among others.[28] They note that pretrial detention decisions provide an attractive template for when and how machine learning might be used to improve on decisions made by judges and does not pertain to uncovering causal relationships. The main contribution of our paper is to show how machine learning taking into account social planner preferences can be incorporated directly into the digital decision process. To that end, we use a comprehensive cost-benefit analysis by Baughman (2017) to build a preference-based approach for this particular application. It is beyond the scope of the current paper to provide an exhaustive and comprehensive empirical analysis, an endeavor we leave for future research. The focus of this section is therefore limited: to assess how asymmetric risks affect machine learning decisions in a simple ideal setting of great practical relevance. The purpose here is to apply the new econometric tools developed in this paper and show how they can potentially improve machine learning application applied at a larger scale.

Judges have to assess the risk whether a defendant, if released, would fail to appear in court or be rearrested for a new crime.[29] The decision to detain or release a defendant has economic and social benefits and costs. A decision to detain an individual has costs directly affecting the detainee and indirect/social costs to the detainee's family, employer, government, and the detention center. On the flip side, releasing the individual has direct and societal benefits, provided no criminal acts will ensue. Recidivism, one of the most fundamental concepts in criminal justice, refers to a person's relapse into criminal behavior. While this is already a complex problem, things become even more complicated when the economic and social costs of racial discrimination are factored into the discussions. Black - low, moderate or high risk felony arrestees - are treated differently, which brings us to the fairness issues. Even after accounting for demographic and charge characteristics of defendants, there are significant differences across counties. For example Harris County in Texas is 34 percent more likely to detain black defendants compared to white defendants with the same observable characteristics, while Baltimore County in Maryland is 1 percent less likely to detain black defendants compared to white defendants; see Reaves (2013) and Dobbie and Yang (2019).

We use data from Broward County, Florida originally compiled by ProPublica; see Larson, Mattu, Kirchner, and Angwin (2016). Following their analysis, we only

[28]A recent economic policy initiative at the Brookings Institution (Hamilton Project 2019-05, see Dobbie and Yang (2019)) is an excellent source of academic as well as policy discussions.

[29]In some states judges are asked to only consider flight risk, not public safety risk.

consider defendants who were assigned COMPAS risk scores within 30 days of their arrest, and were not arrested for an ordinary traffic offense. In addition, we restrict our analysis to only those defendants who spent at least two years (after their COMPAS evaluation) outside a correctional facility without being arrested for a violent crime, or were arrested for a violent crime within this two-year period. Following standard practice, we use this two-year violent recidivism metric to approximate the benefit of detention. In particular we set $Y_i = 1$ for those who re-offended, and $Y_i =$ -1 for those who did not. In Table 4 we report some summary statistics for our data. The total number of records is 11181, with 8972 male defendants. We have 3695 cases of recidivism and a racial mix dominated by African-Americans and Caucasian, respectively 5751 and 3822 in numbers. The binary outcome *is_recid* indicates that 3695 out of the total of 11181 resulted in recidivism. The largest crime category is the aggravated assault, with 2771 cases and the smallest is murder with 9 observations. The minimum age is 18 with a median of 31.

It is not the purpose to compare machine learning outcomes with human decisions (as in Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018)) or to compare machine learning outcomes with COMPAS.[30] Instead we examine how preference-based machine learning, explicitly taking into account asymmetries, compares to standard machine learning methods (such as those applied by Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) who use boosted trees or the various fairness enhanced algorithms as in Corbett-Davies, Pierson, Feller, Goel, and Huq (2017), among others).

We rely on a two-group setup appearing in Table 1 also used in the previous section where the protected group ($G = 1$) are African-American offenders. Our analysis involves asymmetric costs that are covariate-driven and based on a comprehensive cost-benefit analysis for the U.S. pretrial detention decision provided Baughman (2017) which we summarize in Table A.2 in the Appendix. The cost-benefit covariate-

[30]COMPAS, assigns defendants risk scores between 1 and 10 that indicate how likely they are to commit a violent crime based on more than 100 factors, including age, sex and criminal history. Defendants classified as high risk are much more likely to be detained while awaiting trial than those classified as low risk. COMPAS does not explicitly use race as an input. Nevertheless, the aforementioned ProPublica article revealed that black defendants are substantially more likely to be classified as high risk. In addition, among defendants who ultimately did not re-offend, blacks were more than twice as likely as whites to be labeled as risky.

Table 1: Asymmetric costs with protected group

Asymmetric costs for two-group population, where x_i, c_i, and d_i are individual i characteristics, with d_i the pretrial duration, c_i the crime being arrested for, and x_i a vector of other characteristics. The protected group is represented by $G = 1$.

	G = 0		G = 1	
	$f(0,z)=1$	$f(0,z)=-1$	$f(1,z)=1$	$f(1,z)=-1$
Y = 1	$\ell^0_{1,1}(x_i,c_i,d_i)$	$\ell^0_{-1,1}(x_i,c_i,d_i)$	$\ell^1_{1,1}(x_i,c_i,d_i)$	$\ell^1_{-1,1}(x_i,c_i,d_i)$
Y = -1	$\ell^0_{1,-1}(x_i,c_i,d_i)$	$\ell^0_{-1,-1}(x_i,c_i,d_i)$	$\ell^1_{1,-1}(x_i,c_i,d_i)$	$\ell^1_{-1,-1}(x_i,c_i,d_i)$

driven costs functions to characterize the risk $\mathcal{R}(f)$ are as follows:

$$\begin{array}{lcl} \ell^G_{1,1}(x_i,c_i,d_i) & = & \gamma^G_{eb}(x_i)EBD(c_i) + ECD(d_i) \\ \ell^G_{1,-1}(x_i,c_i,d_i) & = & \gamma^G_{eb}(x_i)C(x_i,c_i) \\ \ell^G_{-1,1}(x_i,c_i,d_i) & = & \lambda^G_{ec}(x_i)ECD(d_i) \\ \ell^G_{-1,-1}(x_i,c_i,d_i) & = & \rho^G(x_i) \end{array} \tag{5}$$

with $EBD(c_i)$ the economic benefit of detention which depends on the type of crime committed by individual i, $ECD(d_i)$ the economic cost of detention which depends on detention duration d_i, and $C(x_i, c_i)$ the expected cost of recidivism in the event of a false negative verdict. Details regarding the functions $EBD(c_i)$, $ECD(d_i)$ and $C(x_i, c_i)$ appear in Appendix Section A.4. Finally $\gamma^G_{eb}(x_i)$, $\lambda^G_{ec}(x_i)$ and $\rho^G(x_i)$ are scaling functions which reflect preference attitudes towards members of the protect population where we put $\gamma^G_{eb}(x_i) = \lambda^G_{ec}(x_i)$ and equal to one for $G = 0$ and equal to two for $G = 1$. Finally we set $\rho^G(x_i) = 0$.

We consider the following empirical model specifications: (1) logistic regression covered in Section 4.1, (2) shallow and deep learning in Sections 4.3 and 4.4, and (3) boosting covered in Section 4.2. In each case we compare symmetric versus asymmetric costs, with the latter involving two costs schemes. In all specifications we use as dependent variable a dummy of Recidivism occurrence. The explanatory variables are: (a) race as a categorical variable, (b) gender using female indicator, (c) crime history: prior count of crimes, (d) COMPASS score, (e) crime factor: whether crime is felony or not and (e) interaction between race factor and compass score.

The results are reported in Table 5. We report respectively: (a) True & False Positive/Negative Costs, (b) overall cost, (c) True & Positives/Negatives, (d) True/False

Positive Rates, (e) area under the ROC curve (AUC) during the training and testing sample. For each estimation procedure we compare side-by-side the unweighted, i.e. traditional symmetric, and weighted procedure. Let us start with the logistic regression model. The overall costs are smaller for the weighted, down roughly 10 % compared with the unweighted estimation. This is driven by smaller True Positive Costs (or more precisely larger gains) and False Negative Costs. In contrast, False Positive Costs - meaning keeping the wrong people in jail - are higher with the weighted estimator. In terms of AUC both in- and out-of-sample we do not see much improvement, however. We keep more criminals in jail, release slightly less people (true negatives), and release less the wrong people and thereby reduce recidivism with the weighted estimator of the model. The Boosting model yields results that are worse, both in terms of weighted versus unweighted and vis-à-vis the logistic regression, although the AUC results are overall better than for the logistic regression.

The No Hidden Layer model appears in Table 5 since it allows us to bridge the logistic regression with the shallow and deep learning models. We look at hinge and Logistic Loss functions, and again the standard unweighted versus the novel weighted estimator approach proposed in our paper. Let us start with Logistic Loss estimates, which should match those of the logistic regression model reported in the first panel of the table, as is indeed the case. More interesting is to compare the hinge Loss with the Logistic one. Here, we clearly see that for the unweighted estimator we observe a lower cost with the logistic, but the reverse is true for the weighted estimators. This being said, the difference between hinge and Logistic are typically small.

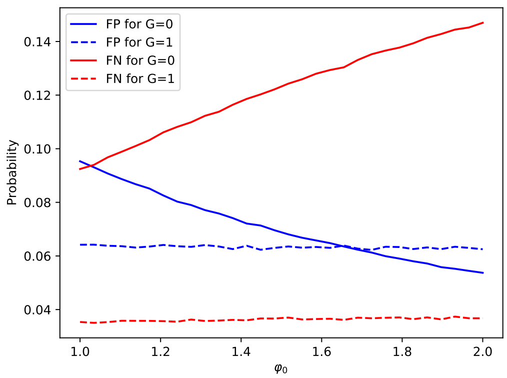
Turning to shallow and deep learning models, we obtain the best results with a two-layer deep learning model using hinge Loss when we compare overall costs (at 5442) which is roughly a 10 % reduction compared to the weighted logistic regression model we started out with. In terms of AUC, however, one would favor the Logistic Loss with two hidden layers, or even the three-layer deep learning model with Logistic Loss. The patterns in terms of True & False Positive/Negative Costs, True & Positives/Negatives, or True/False Positive Rates are mostly similar to the findings reported for the unweighted versus weighted logistic regression model in the first panel.

7 Conclusions

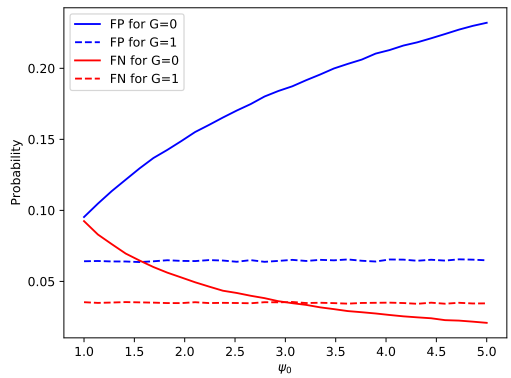
This paper provides a new perspective on the problem of the data-driven binary decision problems with arbitrary loss functions and contributes more broadly to the growing literature at the intersection of the econometrics and machine learning. We suggest an extremely simple reweighting of the logistic regression, or other state-of-the art machine learning techniques, such as boosting or (deep) neural networks and establish several supporting theoretical results. This constitutes a significant computational advantage relative to other approaches previously considered in the literature and allows obtaining valid binary forecasts for outsized and high-dimensional datasets frequently encountered in the modern empirical practice.

Our theory shows that the valid parametric and nonparametric binary predictions can be obtained for arbitrary loss functions based on the appropriately reweighed convexified empirical risk minimization. In particular, the reweighed logistic regression delivers valid binary forecasts even when the choice probabilities are not Logistic. We also show that the carefully crafted shallow and deep neural networks deliver nonparametric predictions that are nearly optimal from the minimax point of view. It is also worth recalling that the nonparametric binary prediction problem is easier than the nonparametric regression and the convergence rate of the risk depends on the amount of the probability mass near the decision boundary which is reflected in our theoretical results.

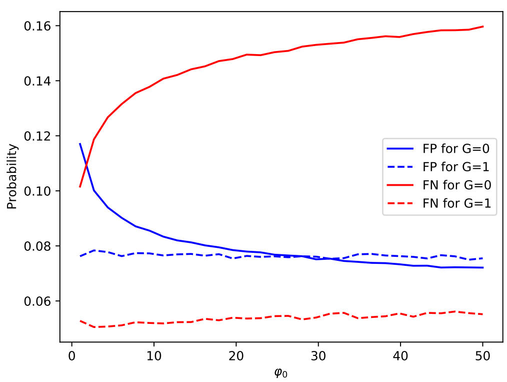
Lastly, we apply our methodology to the problem of predicting the recidivism and find that the binary decisions produced by the ML methods taking into account the economic consequences can reduce the total costs of such decisions.



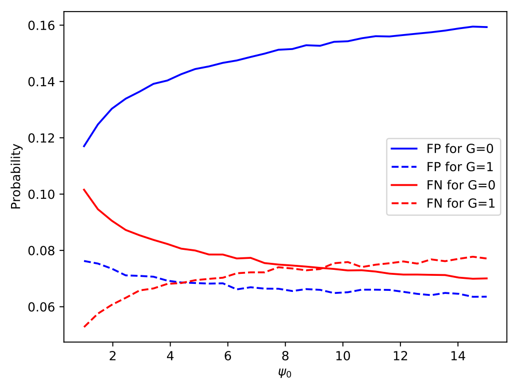
(a) Logit: false positives



(b) Logit: false negatives



(c) Boosting: false positives



(d) Boosting: false negatives

Figure 4: Asymmetric binary choice. The figure shows that introducing asymmetries in the loss function can equalize the False Positive and the False Negative mistakes across groups. Setting: $\rho = 0.2$, $\tau = 0$, $n = 1,000$. Results based on $5,000$ Monte Carlo experiments.

References

Aigner, D. J., T. Amemiya, and D. J. Poirier (1976): "On the estimation of production frontiers: maximum likelihood estimation of the parameters of a discontinuous density function," *International Economic Review*, pp. 377–396.

Anthony, M., and P. L. Bartlett (2009): *Neural network learning: Theoretical foundations*. Cambridge University Press.

Athey, S., and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Athey, S., and S. Wager (2020): "Policy learning with observational data," *Econometrica*.

Audibert, J.-Y., and A. B. Tsybakov (2007): "Fast learning rates for plug-in classifiers," *Annals of Statistics*, 35(2), 608–633.

Babii, A., E. Ghysels, and J. Striaukas (2020): "Inference for high-dimensional regressions with heteroskedasticity and autocorrelation," arXiv preprint arXiv:1912.06307.

Barocas, S., M. Hardt, and A. Narayanan (2020): "Fairness and machine learning: Limitations and Opportunities," Incomplete working draft available at http://Fairmlbook.org.

Bartlett, P. L., N. Harvey, C. Liaw, and A. Mehrabian (2019): "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks.," *Journal of Machine Learning Research*, 20(63), 1–17.

Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006): "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, 101(473), 138–156.

Bauer, B., and M. Kohler (2019): "On deep learning as a remedy for the curse of dimensionality in nonparametric regression," *Annals of Statistics*, 47(4), 2261–2285.

Baughman, S. B. (2017): "Costs of Pretrial Detention," *Boston University Law Review*, 97(1).

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018): "High-dimensional econometrics and regularized GMM," *arXiv preprint arXiv:1806.01888*.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "High-dimensional methods and inference on structural and treatment effects," *Journal of Economic Perspectives*, 28(2), 29–50.

Blanchard, G., G. Lugosi, and N. Vayatis (2003): "On the rate of convergence of regularized boosting classifiers," *Journal of Machine Learning Research*, 4, 861–894.

Boucheron, S., O. Bousquet, and G. Lugosi (2005): "Theory of classification: A survey of some recent advances," *ESAIM: Probability and Statistics*, 9, 323–375.

Boyes, W. J., D. L. Hoffman, and S. A. Low (1989): "An econometric analysis of the bank credit scoring problem," *Journal of Econometrics*, 40(1), 3–14.

Breiman, L. (2000): "Some infinity theory for predictor ensembles," Discussion paper, Technical Report 579, Statistics Dept. UCB.

Chen, J., X. Chen, and E. Tamer (2020): "Optimal Estimation of Expectation Functionals of Nonparametric Conditional Moments via Neural Networks," Discussion paper Harvard and Yale.

Chen, L., M. Pelger, and J. Zhu (2019): "Deep learning in asset pricing," *Available at SSRN 3350138*.

Chen, X. (2007): "Large sample sieve estimation of semi-nonparametric models," in *Handbook of Econometrics - Volume 6b*, ed. by J. J. Heckman, and E. E. Leamer, pp. 5549–5632. Elsevier.

Chen, X., and S. C. Ludvigson (2009): "Land of addicts? an empirical investigation of habit-based asset pricing models," *Journal of Applied Econometrics*, 24(7), 1057–1093.

Chen, X., J. Racine, and N. R. Swanson (2001): "Semiparametric ARX neural-network models with an application to forecasting inflation," *IEEE Transactions on Neural Networks*, 12(4), 674–683.

Chen, X., and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, pp. 289–314.

CHEN, X., AND H. WHITE (1999): "Improved rates and asymptotic normality for nonparametric neural network estimators," *IEEE Transactions on Information Theory*, 45(2), 682–691.

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

CORBETT-DAVIES, S., E. PIERSON, A. FELLER, S. GOEL, AND A. HUQ (2017): "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806.

COWGILL, B., AND C. E. TUCKER (2019): "Economics, fairness and algorithmic bias," *Journal of Economic Perspectives*, forthcoming.

CYBENKO, G. (1989): "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, 2(4), 303–314.

DATTA, A., M. C. TSCHANTZ, AND A. DATTA (2015): "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92–112.

DEVROYE, L., L. GYÖRFI, AND G. LUGOSI (1996): *A probabilistic theory of pattern recognition*, vol. 31. Springer.

DOBBIE, W., AND C. YANG (2019): "Proposals for Improving the US Pretrial System," Hamilton Project Policy Proposal 2019-05, Brookings Institution.

DWORK, C., M. HARDT, T. PITASSI, O. REINGOLD, AND R. ZEMEL (2012): "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226.

ELLIOTT, G., AND R. P. LIELI (2013): "Predicting binary outcomes," *Journal of Econometrics*, 174(1), 15–26.

ELLIOTT, G., AND A. TIMMERMANN (2016): *Economic Forecasting*. Princeton University Press.

FARRELL, M. H., T. LIANG, AND S. MISRA (2019): "Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands," *arXiv preprint arXiv:1809.09953*.

Friedman, J., T. Hastie, and R. Tibshirani (2000): "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *Annals of Statistics*, 28(2), 337–407.

Gallant, A. R., and H. White (1988): "There exists a neural network that does not make avoidable mistakes," in *IEEE 1988 International Conference on Neural Networks*, pp. 657–664.

——— (1992): "On learning the derivatives of an unknown mapping with multilayer feedforward networks," *Neural Networks*, 5(1), 129–138.

Goodfellow, I., Y. Bengio, and A. Courville (2016): *Deep learning*. MIT Press.

Granger, C. W. (1969): "Prediction with a generalized cost of error function," *Journal of the Operational Research Society*, 20(2), 199–207.

——— (1995): "Modelling nonlinear relationships between extended-memory variables," *Econometrica*, 63, 265–279.

Granger, C. W., and M. H. Pesaran (2000): "Economic and statistical measures of forecast accuracy," *Journal of Forecasting*, 19(7), 537–560.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical asset pricing via machine learning," *Review of Financial Studies*, 33(5), 2223–2273.

Hornik, K., M. Stinchcombe, and H. White (1989): "Multilayer feedforward networks are universal approximators.," *Neural networks*, 2(5), 359–366.

Hutchinson, J. M., A. W. Lo, and T. Poggio (1994): "A nonparametric approach to pricing and hedging derivative securities via learning networks," *Journal of Finance*, 49(3), 851–889.

Ivakhnenko, A., and V. Lapa (1965): "Cybernetic Predictive Devices (in Russian)," .

Kamishima, T., S. Akaho, H. Asoh, and J. Sakuma (2012): "Fairness-aware classifier with prejudice remover regularizer," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human decisions and machine predictions," *Quarterly Journal of Economics*, 133(1), 237–293.

Kleinberg, J., J. Ludwig, S. Mullainathan, and A. Rambachan (2018): "Algorithmic fairness," in *AEA Papers and Proceedings*, vol. 108, pp. 22–27.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2016): "Inherent trade-offs in the fair determination of risk scores," arXiv preprint arXiv:1609.05807.

Koenker, R., and G. Bassett Jr (1978): "Regression quantiles," *Econometrica*, pp. 33–50.

Kohler, M., and S. Langer (2020): "On the rate of convergence of fully connected deep neural network regression estimates," *arXiv preprint arXiv:1908.11133*.

Koltchinskii, V. (2006): "Local Rademacher complexities and oracle inequalities in risk minimization," *Annals of Statistics*, 34(6), 2593–2656.

——— (2011): *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, vol. 2033. Springer.

Larson, J., S. Mattu, L. Kirchner, and J. Angwin (2016): "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica (5 2016)*.

Lee, T.-H., H. White, and C. W. Granger (1993): "Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests," *Journal of Econometrics*, 56(3), 269–290.

Lu, J., Z. Shen, H. Yang, and S. Zhang (2020): "Deep network approximation for smooth functions," *arXiv preprint arXiv:2001.03040*.

Manski, C. F. (1975): "Maximum score estimation of the stochastic utility model of choice," *Journal of econometrics*, 3(3), 205–228.

McCulloch, W. S., and W. Pitts (1943): "A logical calculus of the ideas immanent in nervous activity," *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.

Mhaskar, H. N. (1996): "Neural networks for optimal approximation of smooth and analytic functions," *Neural computation*, 8(1), 164–177.

Pesaran, M. H., and S. Skouras (2002): "Decision-based methods for forecast evaluation," *A companion to economic forecasting*, pp. 241–267.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020): "An economic approach to regulating algorithms," Discussion paper, National Bureau of Economic Research.

Reaves, B. (2013): *Felony defendants in large urban counties, 1990 - 2009*. State Court Processing Statistics, Bureau of Justice Statistics, Office of Justice Programs, U.S. Department of Justice, Washington, DC.

Rosenblatt, F. (1958): "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, 65(6), 386.

Schapire, R. E., Y. Freund, P. Bartlett, and W. S. Lee (1998): "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, 26(5), 1651–1686.

Schervish, M. J. (1989): "A general method for comparing probability assessors," *Annals of Statistics*, 17(4), 1856–1879.

Schmidt-Hieber, J. (2017): "Nonparametric regression using deep neural networks with ReLU activation function," *arXiv preprint arXiv:1708.06633*.

Suzuki, T., and A. Nitanda (2019): "Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space," *arXiv preprint arXiv:1910.12799*.

Tsybakov, A. B. (2004): "Optimal aggregation of classifiers in statistical learning," *Annals of Statistics*, 32(1), 135–166.

Vapnik, V. (1995): *The nature of statistical learning theory*. Springer.

White, H., and J. Racine (2001): "Statistical inference, the bootstrap, and neural-network modeling with application to foreign exchange rates," *IEEE Transactions on Neural Networks*, 12(4), 657–673.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork (2013): "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333.

Zhang, T. (2004): "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, pp. 56–85.

Table 2: Monte Carlo Simulation Results: ML Prediction

The Monte Carlo simulation design is presented in Section 5, which represents a stylized social planner facing disproportionate number of false positive and false negative mistakes across two groups with the standard ML classification approach. The population consists of two groups, $G = 0$ and 1. Constituents of group $G = 1$ are a fraction ρ. Moreover, the dimension d of covariates Z is set to 15. FP and FN are false positive and false negative mistakes computed as a share in the corresponding group, Total = misclassification rate. Logit = logistic regression, GB = Gradient Boosted trees, SL = shallow learning, DL = deep learning, SVM = support vector machines. All results are based on $5,000$ MC experiments.

		Nonlinear DGP: $\tau = 1$						Linear DGP: $\tau = 0$					
		$\rho = 0.2$			$\rho = 0.5$			$\rho = 0.2$			$\rho = 0.5$		
	G	FP	FN	Error	FP	FN	Error	FP	FN	Error	FP	FN	Error
						Sample size n = 1,000							
Logit	0	0.28	0.12	0.37	0.28	0.12	0.33	0.10	0.09	0.17	0.10	0.09	0.14
	1	0.25	0.02		0.25	0.01		0.06	0.04		0.07	0.03	
GB	0	0.15	0.09	0.22	0.17	0.08	0.21	0.12	0.10	0.2	0.12	0.10	0.17
	1	0.12	0.06		0.12	0.05		0.08	0.05		0.08	0.04	
SVM	0	0.13	0.08	0.2	0.14	0.08	0.18	0.14	0.08	0.2	0.16	0.07	0.17
	1	0.10	0.05		0.10	0.04		0.04	0.11		0.07	0.05	
SL	0	0.07	0.03	0.09	0.04	0.06	0.08	0.10	0.08	0.17	0.10	0.08	0.14
	1	0.04	0.02		0.03	0.03		0.07	0.03		0.07	0.03	
DL	0	0.06	0.05	0.1	0.06	0.04	0.08	0.10	0.09	0.18	0.28	0.05	0.23
	1	0.04	0.03		0.04	0.03		0.06	0.05		0.10	0.02	
						Sample size n = 100							
Logit	0	0.25	0.20	0.43	0.25	0.19	0.39	0.14	0.10	0.23	0.17	0.09	0.20
	1	0.19	0.16		0.20	0.13		0.04	0.14		0.06	0.08	
GB	0	0.25	0.17	0.41	0.27	0.15	0.37	0.18	0.11	0.29	0.21	0.11	0.24
	1	0.16	0.20		0.18	0.14		0.05	0.22		0.09	0.07	
SVM	0	0.30	0.10	0.38	0.33	0.07	0.33	0.21	0.09	0.29	0.29	0.06	0.25
	1	0.19	0.10		0.21	0.06		0.05	0.18		0.08	0.08	
SL	0	0.08	0.03	0.10	0.04	0.07	0.09	0.10	0.09	0.17	0.11	0.08	0.14
	1	0.05	0.02		0.03	0.04		0.06	0.03		0.07	0.03	
DL	0	0.08	0.04	0.11	0.09	0.03	0.10	0.15	0.06	0.19	0.09	0.11	0.15
	1	0.06	0.02		0.06	0.02		0.08	0.02		0.06	0.04	

Table 3: Monte Carlo Simulation Results

The Monte Carlo simulation design is presented in Section 5, which represents a stylized social planner with a loss function from Example 2.2 featuring asymmetries for false positives and false negatives. The population consists of two groups, $G = 0$ and 1, with the former assumed to be a protected segment of the population. Constituents of group $G = 1$ are a fraction ρ. The baseline case has the loss function with the following setting: $\psi_1 = \varphi_1 = 1$, $\varphi_0 = 1.7$, and $\psi_0 = 3$. We also set $\rho = 0.2$ and $\tau = 0$. We compare the performance of a standard logistic regression approach, which ignores the asymmetric loss function, with our convexified weighted logistic model appearing in equation (4). The social planner loss for the former will be denoted by ℓ_{logit} while our new estimator yields $\ell_{w-logit}$. We report $\mathrm{P}(\ell_{logit} > \ell_{w-logit})$ which is the percent that standard logistic regressions generate larger social planner costs compared to our weighted regression. Hence, this measure reports how often our estimator outperforms the standard procedure, where the probabilities are computed from the Monte Carlo simulated samples. Next, we report summary statistics for the ratio $\ell_{logit}/\ell_{w-logit}$. When the ratio is above 1.00 then the weighted logistic approach is better. We report the minimum, maximum, mean, and three quartiles of the simulation distribution. Columns with $\varphi_0, \varphi_1, \psi_0$, or ψ_1 as headers represent deviations from the baseline case.

	Baseline case	ρ 0.5	τ 1	φ_0 2	φ_1 2	φ_1 3	ψ_0 4	ψ_1 2	ψ_1 3
				Sample size $n = 1{,}000$					
$\mathrm{P}(\ell_{logit} > \ell_{w-logit})$	0.73	0.65	0.94	0.60	0.78	0.8	0.87	0.77	0.80
Summary statistics for $\ell_{logit}/\ell_{w-logit}$									
Mean	1.07	1.05	1.16	1.03	1.08	1.10	1.16	1.08	1.10
Min	0.74	0.68	0.85	0.76	0.77	0.74	0.71	0.77	0.73
1st Quantile	0.99	0.97	1.08	0.97	1.01	1.02	1.06	1.01	1.02
Median	1.06	1.04	1.15	1.02	1.08	1.09	1.14	1.08	1.09
3rd Quantile	1.13	1.11	1.23	1.08	1.15	1.17	1.25	1.15	1.17
Max	1.61	1.60	1.70	1.46	1.60	1.60	1.83	1.52	1.67
				Sample size $n = 5{,}000$					
$\mathrm{P}(\ell_{logit} > \ell_{w-logit})$	0.94	0.87	1.00	0.8	0.95	0.97	0.99	0.95	0.97
Summary statistics for $\ell_{logit}/\ell_{w-logit}$									
Mean	1.08	1.06	1.12	1.03	1.08	1.09	1.17	1.08	1.09
Min	0.91	0.87	0.98	0.90	0.88	0.86	0.93	0.92	0.92
1st Quantile	1.04	1.02	1.08	1.01	1.05	1.06	1.12	1.05	1.06
Median	1.07	1.06	1.11	1.03	1.08	1.09	1.17	1.08	1.09
3rd Quantile	1.11	1.10	1.15	1.06	1.11	1.13	1.21	1.11	1.13
Max	1.28	1.29	1.34	1.21	1.28	1.32	1.47	1.28	1.32

Table 4: Summary Statistics Data

Summary statistics of recidivism data set from Broward County, Florida originally compiled by ProPublica (see Larson, Mattu, Kirchner, and Angwin (2016)). We use an excerpt of the original data, considering only black and white defendants who were assigned COMPAS risk scores within 30 days of their arrest, were not arrested for an ordinary traffic offense, and defendants who spent at least two years (after their COMPAS evaluation) outside a correctional facility without being arrested for a violent crime, or were arrested for a violent crime within this two-year period. The entry *is_recid* pertains to the binary outcome of recidivism, Decile Score refers to the COMPAS score.

Gender	Male	Female				
	8972	2209				
is_recid	0	1				
	7486	3695				
Race	African American	Asian	Caucasian	Hispanic	Native American	Other
	5751	52	3822	910	30	616
Crime	Aggravated Assault	Arson	Fraud	Household Burglary	Larceny Theft	
	2771	6275	224	577	1028	
	Rape Sexual Assault	Robbery	Murder	Motor Vehicle Theft		
	38	82	9	177		
	Min.	25% Quantile	Median	Mean	75% Quantile	Max
Decile Score	-1	2	4	4.577	7	10
Prior crime count	0	0	1	3.263	4	38
Age	18	25	31	34.33	42	96
Detention Days	0	0	1	21.69	8	2152

Table 5: Empirical Results

Empirical results with (1) logistic regression covered in Section 4.1, (2) shallow and deep learning in Sections 4.3 and 4.4, and (3) boosting covered in Section 4.2. The setting involves two groups with cost structure appearing in Table 1.

	Logistics		Boosting	
	Unweighted	Weighted	Unweighted	Weighted
True Positive Cost	-1310	-2150	-1674	-2332
False Negative Cost	5575	4771	5141	4085
True Negative Cost	0	0	0	0
False Positive Cost	2415	3289	3036	3680
Overall cost	6680	5909	6503	5433
TP	178	206	228	228
FN	577	549	527	527
TN	1438	1400	1411	1383
FP	105	143	132	160
TP Rate	0.24	0.27	0.30	0.30
FP Rate	0.07	0.09	0.09	0.10
AUC_train	0.69	0.67	0.73	0.71
AUC_test	0.69	0.67	0.7	0.68

	No Hidden Layer			
	Hinge Loss		Logistic Loss	
	Unweighted	Weighted	Unweighted	Weighted
True Positive Cost	-1930	-2360	-1310	-2150
False Negative Cost	4395	3971	5575	4771
True Negative Cost	0	0	0	0
False Positive Cost	4646	4186	2415	3289
Overall cost	7111	5797	6680	5909
TP	247	251	178	206
FN	508	504	577	549
TN	1341	1361	1438	1400
FP	202	182	105	143
TP Rate	0.33	0.33	0.24	0.27
FP Rate	0.13	0.12	0.07	0.09
AUC_train	0.7	0.69	0.69	0.67
AUC_test	0.69	0.68	0.69	0.67

Table 5 continued

	Shallow Learning - One Hidden Layer			
	Hinge Loss		Logistic Loss	
	Unweighted	Weighted	Unweighted	Weighted
True Positive Cost	-1793	-2429	-2100	-2648
False Negative Cost	5402	4482	5748	4586
True Negative Cost	0	0	0	0
False Positive Cost	2806	3634	3059	3933
Overall cost	6415	5687	6707	5871
TP	199	227	211	234
FN	556	528	544	521
TN	1421	1385	1410	1372
FP	122	158	133	171
TP Rate	0.26	0.3	0.28	0.31
FP Rate	0.08	0.1	0.09	0.11
AUC_train	0.69	0.66	0.7	0.69
AUC_test	0.69	0.66	0.7	0.68

	Deep Learning: Two Hidden Layers			
	Hinge Loss		Logistic Loss	
	Unweighted	Weighted	Unweighted	Weighted
True Positive Cost	-1677	-2266	-2045	-2149
False Negative Cost	5288	4488	5232	4763
True Negative Cost	0	0	0	0
False Positive Cost	2645	3220	3335	3381
Overall cost	6256	5442	6521	5994
TP	202	236	229	203
FN	553	519	526	552
TN	1428	1403	1398	1396
FP	115	140	145	147
TP Rate	0.27	0.31	0.3	0.27
FP Rate	0.07	0.09	0.09	0.1
AUC_train	0.64	0.62	0.7	0.69
AUC_test	0.65	0.64	0.7	0.68

	Deep Learning: Three Hidden Layers			
	Hinge Loss		Logistic Loss	
	Unweighted	Weighted	Unweighted	Weighted
True Positive Cost	-1774	-2379	-2272	-2500
False Negative Cost	5495	4652	5642	4967
True Negative Cost	0	0	0	0
False Positive Cost	2530	3220	3427	3864
Overall cost	6251	5493	6796	6331
TP	192	235	234	217
FN	563	520	521	538
TN	1433	1403	1394	1375
FP	110	140	149	168
TP Rate	0.25	0.31	0.31	0.29
FP Rate	0.07	0.09	0.1	0.11
AUC_train	0.64	0.62	0.71	0.69
AUC_test	0.65	0.62	0.7	0.68

APPENDIX

A.1 Convexification

We will use $a \lesssim b$ if $a \leq Cb$ for some universal constant C that does not depend on the distribution of (Y, X).

Proof of Proposition 2.1. Note that for every $f, y \in \{-1, 1\}$ and $x \in \mathcal{X}$

$$\begin{aligned}
\ell(f, y, x) &= \ell_{1,1}(x)\frac{(1+f)(1+y)}{4} + \ell_{1,-1}(x)\frac{(1+f)(1-y)}{4} \\
&\quad + \ell_{-1,1}(x)\frac{(1-f)(1+y)}{4} + \ell_{-1,-1}(x)\frac{(1-f)(1-y)}{4} \\
&= 0.25(yb(x) - a(x))f + 0.25\bar{c}(y, x)
\end{aligned}$$

with $\bar{c}(y, x) = (\ell_{1,1}(x) + \ell_{-1,1}(x))(1 + y) + (\ell_{1,-1}(x) + \ell_{-1,-1}(x))(1 - y)$. Next, for every $(y, x) \in \{-1, 1\} \times \mathcal{X}$ and every $f : \mathcal{X} \to \{-1, 1\}$

$$(yb(x) - a(x))f(x) = (b(x) - ya(x))(1 - 2\mathbb{1}\{-yf(x) \geq 0\}).$$

The result follows since

$$\begin{aligned}
\mathcal{R}(f) &= \mathbb{E}[\ell(f(X), Y, X)] \\
&= 0.5\mathbb{E}[(Ya(X) - b(X))\mathbb{1}\{-Yf(X) \geq 0\}] \\
&\quad + 0.25\mathbb{E}[b(X) - Ya(X) + \bar{c}(Y, X)].
\end{aligned} \tag{A.1}$$

□

Proof of Proposition 3.1. Note that $f^*(x)$ solves

$$\inf_{f \in \{-1,1\}} \mathbb{E}[(Ya(X) + b(X))\mathbb{1}\{-Yf \geq 0\}|X = x].$$

By the law of iterated expectations, for every $f \in \{-1, 1\}$, the objective function equals to

$$\begin{aligned}
\mathbb{E}[(Ya(X) - b(X))\mathbb{1}\{-Yf \geq 0\}|X = x] &= \eta(x)(a(x) - b(x))\mathbb{1}\{f \leq 0\} \\
&\quad - (1 - \eta(x))(a(x) + b(x))\mathbb{1}\{f \geq 0\} \\
&= [a(x) + b(x) - 2\eta(x)b(x)]\, \mathbb{1}\{f \leq 0\} \\
&\quad - (1 - \eta(x))(a(x) + b(x)).
\end{aligned}$$

Under Assumption 3.1 (i), $b < 0$ a.s., and whence we obtain the first statement

$$f^*(x) = 1 \iff \eta(x) > \frac{a(x) + b(x)}{2b(x)}.$$

For the second statement, note that $f_\phi^*(x)$ solves

$$\inf_{f \in \mathbf{R}} \mathbb{E}[(Ya(X) - b(X))\phi(-Yf)|X = x].$$

By the law of iterated expectations, for every $f \in \mathbf{R}$, the objective function is

$$H_\eta(f) \triangleq \eta(x)(a(x) - b(x))\phi(-f(x)) - (1 - \eta(x))(a(x) + b(x))\phi(f(x)).$$

Since ϕ is differentiable under Assumption 3.2 (ii), the optimum solves

$$H'_\eta(f_\phi^*) = \eta(x)(b(x) - a(x))\phi'(-f_\phi^*(x)) - (1 - \eta(x))(a(x) + b(x))\phi'(f_\phi^*(x)) = 0.$$

Under Assumption 3.1 (i), $b < a < -b$ a.s. and under Assumption 3.1 (ii), $\eta \notin \{0, 1\}$. Therefore,

$$\frac{\eta(x)(b(x) - a(x))}{(1 - \eta(x))(a(x) + b(x))} = \frac{\phi'(f_\phi^*(x))}{\phi'(-f_\phi^*(x))}.$$

Since ϕ is a convex function, its derivative ϕ' is non-decreasing. Then from the equation above, we have the following equivalence relation

$$\begin{aligned} f_\phi^*(x) > 0 &\iff \frac{\eta(x)(b(x) - a(x))}{(1 - \eta(x))(a(x) + b(x))} > 1 \\ &\iff \eta(x) > \frac{a(x) + b(x)}{2b(x)}, \end{aligned}$$

where the second line follows since $a + b < 0$ a.s. under Assumption 3.1 (i). $\square$

Lemma A.1.1. *Suppose that Assumption 3.1 (i) is satisfied. Then for every decision $f : \mathcal{X} \to \{-1, 1\}$, the excess risk is*

$$\mathcal{R}(f) - \mathcal{R}^* = -\mathbb{E}_{ff^* < 0} b(X) \left| \eta(X) - c(X) \right|,$$

where for event A, we put $\mathbb{E}_A \xi = \mathbb{E} \mathbb{1}_A \xi$.

Proof. By the law of iterated expectations for every decision $f : \mathcal{X} \to \{-1, 1\}$

$$\begin{aligned} \mathbb{E}[(Ya(X) - b(X))\mathbb{1}\{-Yf(X) \geq 0\}] &= \mathbb{E}[\eta(X)(a(X) - b(X))\mathbb{1}\{f(X) \leq 0\}] \\ &\quad - \mathbb{E}[(1 - \eta(X))(a(X) + b(X))\mathbb{1}\{f(X) \geq 0\}] \\ &= \mathbb{E}[(a(X) + b(X) - 2\eta(X)b(X))\mathbb{1}\{f(X) \leq 0\}] \\ &\quad + \mathbb{E}[(\eta(X) - 1)(a(X) + b(X))]. \end{aligned}$$

Combining this observation with equation (A.1)

$$\begin{aligned}
2(\mathcal{R}(f)-\mathcal{R}^*) &= \mathbb{E}[(Ya(X)-b(X))(\mathbb{1}\{-Yf(X)\geq 0\}-\mathbb{1}\{-Yf^*(X)\geq 0\})] \\
&= \mathbb{E}[(a(X)+b(X)-2\eta(X)b(X))(\mathbb{1}\{f(X)\leq 0\}-\mathbb{1}\{f^*(X)\leq 0\})] \\
&= \mathbb{E}[-2b(X)(\eta(X)-c(X))(\mathbb{1}\{f(X)\leq 0\}-\mathbb{1}\{f^*(X)\leq 0\})] \\
&= \mathbb{E}_{ff^*<0}\left[-2b(X)\left|\eta(X)-c(X)\right|\right],
\end{aligned}$$

where the last line follows since under Assumption 3.1 (i), by Proposition 3.1

$$f^*(x)>0 \iff \eta(x)>\frac{a(x)+b(x)}{2b(x)}.$$

□

Lemma A.1.2. *Suppose that Assumptions 3.1 and 3.2 are satisfied. Then for every measurable function* $f:\mathcal{X}\to\mathbf{R}$

$$\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^*\lesssim\left[\mathcal{R}_\phi(f)-\mathcal{R}_\phi^*\right]^\gamma.$$

Proof. Under Assumption 3.1 (i), by Lemma A.1.1

$$\begin{aligned}
\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^* &= -\mathbb{E}_{ff^*<0}b(X)\left|\eta(X)-c(X)\right| \\
&\lesssim \mathbb{E}_{ff^*<0}\left[-b(X)\left((\eta(X)+c(X)-2\eta(X)c(X))-\inf_{y\in\mathbf{R}}Q_c(\eta,y)\right)^\gamma\right] \\
&\lesssim \left(\mathbb{E}_{ff^*<0}\left[(-b(X))^{1/\gamma}\left(\eta(X)+c(X)-2\eta(X)c(X)-\inf_{y\in\mathbf{R}}Q_c(\eta,y)\right)\right]\right)^\gamma \\
&\lesssim \left(\mathbb{E}_{ff^*<0}\left[-b(X)\left(\eta(X)+c(X)-2\eta(X)c(X)-\inf_{y\in\mathbf{R}}Q_c(\eta,y)\right)\right]\right)^\gamma,
\end{aligned}$$

where the second line follows under Assumption 3.2 (iii) since $\eta,c\in(0,1)$ under Assumption 3.1 (i)-(ii); the third by Jensen's inequality since $\gamma\in(0,1]$ under Assumption 3.2 (iii); and the last since $-b\lesssim 1$ under Assumption 3.1 (ii). Next, since

$$\mathcal{R}_\phi(f)-\mathcal{R}_\phi^*=0.5\mathbb{E}\left[(Ya(X)-b(X))\left(\phi(-Yf(X))-\phi(-Yf_\phi^*(X))\right)\right],$$

see equation 1, if we show that

$$\begin{aligned}
&-2b(x)\mathbb{1}\{(\eta(x)-c(x))f(x)<0\}\left(\eta(x)+c(x)-2\eta(x)c(x)-\inf_{y\in\mathbf{R}}Q_{c(x)}(\eta(x),y)\right) \\
&\quad\leq\mathbb{E}[(Ya(X)-b(X))(\phi(-Yf(X))-\phi(-Yf_\phi^*(X)))|X=x]
\end{aligned}\tag{A.2}$$

the result will follow from integrating over x by the law of iterated expectations since $f^*(x)=\mathrm{sign}(\eta(x)-c(x))$ by Proposition 3.1. To that end, if $(\eta(x)-c(x))f(x)\geq 0$, then

Eq. A.2 follows trivially since f_ϕ^* minimizes $f \mapsto \mathbb{E}[Y(a(X)-b(X))\phi(-Yf(X))]$. Suppose that $(\eta(x)-c(x))f(x) < 0$. Then by the law of iterated expectations and the definition of Q in the Assumption 3.2 (iii)

$$\mathbb{E}[(Ya(X)-b(X))\phi(-Yf_\phi^*(X))|X=x] = -2b(x)\inf_{y\in\mathbf{R}} Q_{c(x)}(\eta(x),y).$$

Therefore, the inequality in Eq. A.2 follows if we can show that

$$-2b(x)(\eta(x)+c(x)-2\eta(x)c(x)) \leq \mathbb{E}[(Ya(X)-b(X))\phi(-Yf(X))|X=x].$$

But this follows from

$$\begin{aligned}
&\mathbb{E}[(Ya(X)-b(X))\phi(-Yf(X))|X=x] \\
&= \eta(x)(a(x)-b(x))\phi(-f(x)) - (1-\eta(x))(a(x)+b(x))\phi(f(x)) \\
&= -2b(x)[\eta(x)(1-c(x))\phi(-f(x)) + (1-\eta(x))c(x)\phi(f(x))] \\
&\geq -2b(x)(\eta(x)+c(x)-2\eta(x)c(x))\phi\left(\frac{(c(x)-\eta(x))f(x)}{\eta(x)+c(x)-2\eta(x)c(x)}\right) \\
&\geq -2b(x)(\eta(x)+c(x)-2\eta(x)c(x))\phi(0) \\
&= -2b(x)(\eta(x)+c(x)-2\eta(x)c(x)),
\end{aligned}$$

where the first inequality follows by the convexity of ϕ under Assumption 3.2 (i) and since $\eta+c-2\eta c > 0$ under Assumption 3.1 (i)-(ii); the second inequality since ϕ is non-decreasing under Assumption 3.2 (i) and $(c(x)-\eta(x))f(x) \geq 0$ by assumption; and the last line since $\phi(0) = 1$ under Assumption 3.2 (i). □

Lemma A.1.3. *For the logistic convexifying function $\phi(z) = \log(1+e^z)$,*

$$f_\phi^*(x) = \log\left(\frac{\eta(x)(1-c(x))}{(1-\eta(x))c(x)}\right).$$

Assumption 3.2 is satisfied with $\gamma = 1/2$ and $C = 2\sqrt{2}$.

Proof. Note that the minimum of $y \mapsto Q_c(x,y)$ is achieved at $y^* = \log\left(\frac{x(1-c)}{(1-x)c}\right)$. Then

$$\begin{aligned}
\inf_{y\in\mathbf{R}} Q_c(x,y) &= x(1-c)\log(1+e^{-y^*}) + (1-x)c\log(1+e^{y^*}) \\
&= -x(1-c)\log\frac{x(1-c)}{x(1-c)+(1-x)c} - (1-x)c\log\frac{(1-x)c}{x(1-c)+(1-x)c}.
\end{aligned}$$

Then, since $\log(x) \geq \frac{1}{2}\left(x - \frac{1}{x}\right), \forall x \in (0,1]$,

$$
\begin{aligned}
&2(x+c-2xc) - \inf_{y\in\mathbf{R}} Q_c(x,y) \\
&\geq 2(x+c-2xc) + x(1-c)\left[\frac{x(1-c)}{x(1-c)+c(1-x)} - \frac{x(1-c)+c(1-x)}{x(1-c)}\right] \\
&\quad + c(1-x)\left[\frac{c(1-x)}{x(1-c)+c(1-x)} - \frac{x(1-c)+c(1-x)}{c(1-x)}\right] \\
&= \frac{[x(1-c)]^2}{x(1-c)+c(1-x)} + \frac{[c(1-x)]^2}{x(1-c)+c(1-x)} \\
&\geq \frac{1}{2}\frac{[x(1-c)-c(1-x)]^2}{x(1-c)+c(1-x)} \geq \frac{(x-c)^2}{4},
\end{aligned}
$$

where the last line follows since $x, c \in (0,1)$. Therefore, Assumption 3.2 is verified with $\gamma = 1/2$ and $C = 2\sqrt{2}$. □

Lemma A.1.4. *For the Hinge convexifying function $\phi(z) = (1+z)_+$,*

$$
f^*_\phi(x) = \begin{cases} 1 & \textit{if } \ \eta(x) > c(x), \\ -1 & \textit{if } \ \eta(x) < c(x). \end{cases}
$$

Assumption 3.2 is satisfied with $\gamma = 1$ and $C = 1$.

Proof. Note that the minimum of $y \mapsto Q_c(x,y)$ is achieved at

$$
\begin{cases} 1 & \text{if } \ x > c, \\ -1 & \text{if } \ x < c. \end{cases}
$$

Then

$$
\begin{aligned}
\inf_{y\in\mathbf{R}} Q_c(x,y) &= \inf_{y\in\mathbf{R}} x(1-c)(1-y)_+ + c(1-x)(1+y)_+ \\
&= \min\{2x(1-c), 2c(1-x)\}.
\end{aligned}
$$

If $x \leq c$, then

$$
\begin{aligned}
(x+c-2xc) - \inf_{y\in\mathbf{R}} Q_c(x,y) &= (x+c-2xc) - 2x(1-c) \\
&= (c-x).
\end{aligned}
$$

If $x > c$, then

$$
\begin{aligned}
(x+c-2xc) - \inf_{y\in\mathbf{R}} Q_c(x,y) &= (x+c-2xc) - 2c(1-x) \\
&= (x-c).
\end{aligned}
$$

Therefore,

$$
(x+c-2xc) - \inf_{y\in\mathbf{R}} Q_c(x,y) = |x-c|
$$

and Assumption 3.2 is satisfied with $\gamma = 1$ and $C = 1$. □

Lemma A.1.5. *For the exponential convexifying function $\phi(z) = e^z$,*

$$f^*_\phi(x) = \frac{1}{2}\log\left(\frac{\eta(x)(1-c(x))}{(1-\eta(x))c(x)}\right).$$

Assumption 3.2 is satisfied with $\gamma = 1/2$ and $C = 2$.

Proof. Note that the minimum of $y \mapsto Q_c(x,y)$ is achieved at

$$y^* = \frac{1}{2}\log\left(\frac{x(1-c)}{(1-x)c}\right).$$

Then

$$\begin{aligned}\inf_{y\in\mathbf{R}} Q_c(x,y) &= \inf_{y\in\mathbf{R}} x(1-c)e^{-y} + c(1-x)e^y \\ &= 2\sqrt{xc(1-x)(1-c)}.\end{aligned}$$

Then

$$\begin{aligned}(x+c-2xc) - \inf_{y\in\mathbf{R}} Q_c(x,y) &= (x+c-2xc) - 2\sqrt{xc(1-x)(1-c)} \\ &= \left(\sqrt{x(1-c)} - \sqrt{c(1-x)}\right)^2 \\ &= \frac{(x-c)^2}{(\sqrt{x(1-c)} + \sqrt{c(1-x)})^2} \geq \frac{(x-c)^2}{4}.\end{aligned}$$

where last line is due to the fact $x, c \in (0,1)$. Therefore, Assumption 3.2 is satisfied with $\gamma = 1/2$ and $C = 2$. □

Lemma A.1.6. *Suppose that Assumptions 3.1 (i) and 3.3 are satisfied. Then for every measurable $f : \mathcal{X} \to \mathbf{R}$*

$$\mathcal{R}(f) - \mathcal{R}^* \gtrsim P_X^{\frac{1+\alpha}{\alpha}}\left(\{x :\ f(x)f^*(x) \leq 0\}\right).$$

Proof. By Lemma A.1.1

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}^* &= -\mathbb{E}_{ff^*\leq 0} b(X)|\eta(X) - c(X)| \\ &\gtrsim \int_{\mathcal{X}} \mathbb{1}\{f(x)f^*(x) \leq 0\}|\eta(x) - c(x)|\mathrm{d}P_X(x) \\ &\geq uP_X\left(\{x :\ f(x)f^*(x) \leq 0\} \cap \{x :\ |\eta(x) - c(x)| > u\}\right) \\ &\geq uP_X\left(\{x :\ f(x)f^*(x) \leq 0\}\right) - uP_X\left(\{x :\ |\eta(x) - c(x)| \leq u\}\right) \\ &\geq uP_X\left(\{x :\ f(x)f^*(x) \leq 0\}\right) - Cu^{1+\alpha},\end{aligned}$$

where the first inequality follows under Assumption 3.1 (i); the second by Markov's inequality for every $u > 0$; the third by $\Pr(A \cap B) \geq \Pr(A) - \Pr(B^c)$; and the fourth under Assumption 3.3. The result follows from substituting u solving

$$P_X\left(\{x :\ f(x)f^*(x) \leq 0\}\right) = 2Cu^\alpha$$

in the last equation. □

Proof of Theorem 3.1. Under Assumption 3.1 (i), by Lemma A.1.1

$$\begin{aligned}
\mathcal{R}(\mathrm{sign}(f)) - \mathcal{R}^* &= -\mathbb{E}\left[b(X)|\eta(X)-c(X)|\mathbb{1}\{ff^*<0\}\right] \\
&= -\mathbb{E}_{ff^*<0}\left[b(X)|\eta(X)-c(X)|\mathbb{1}\{|\eta(X)-c(X)|<\epsilon\}\right] \\
&\quad - \mathbb{E}_{ff^*<0}\left[b(X)|\eta(X)-c(X)|\mathbb{1}\{|\eta(X)-c(X)|\geq\epsilon\}\right] \\
&\lesssim \epsilon P_X(\{x:\ f(x)f^*(x)<0\}) \\
&\quad - \epsilon^{1-1/\gamma}\mathbb{E}_{ff^*<0} b(X)|\eta(X)-c(X)|^{1/\gamma}\mathbb{1}\{|\eta(X)-c(X)|\geq\epsilon\} \\
&\lesssim \epsilon\left[\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^*\right]^{\frac{\alpha}{1+\alpha}} + \epsilon^{1-1/\gamma}\left(-\mathbb{E}_{ff^*<0} b(X)|\eta(X)-c(X)|\right)^{1/\gamma},
\end{aligned}$$

where the second line follows for every $\epsilon > 0$; the third since $-b \lesssim 1$ and $\gamma \in (0,1]$ under Assumptions 3.1 (iii) and Assumption 3.2 (iii); and the last by Lemma A.1.6 under Assumptions 3.1 (i) and 3.3, and Jensen's inequality. Next, under Assumption 3.1 and 3.2, by Lemmas A.1.1 and A.1.2

$$\begin{aligned}
-\mathbb{E}_{ff^*<0} b(X)|\eta(X)-c(X)| &= \mathcal{R}(\mathrm{sign}(f)) - \mathcal{R}^* \\
&\lesssim \left[\mathcal{R}_\phi(f)-\mathcal{R}_\phi^*\right]^\gamma.
\end{aligned}$$

Therefore,

$$\mathcal{R}(\mathrm{sign}(f)) - \mathcal{R}^* \lesssim \epsilon\left[\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^*\right]^{\frac{\alpha}{1+\alpha}} + \epsilon^{1-1/\gamma}\left[\mathcal{R}_\phi(f)-\mathcal{R}_\phi^*\right].$$

Setting $\epsilon = 0.5\left[\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^*\right]^{\frac{1}{1+\alpha}}$ and rearranging, we obtain

$$\left[\mathcal{R}(\mathrm{sign}(f))-\mathcal{R}^*\right]^{\frac{\gamma\alpha+1}{\gamma(1+\alpha)}} \lesssim \left[\mathcal{R}_\phi(f)-\mathcal{R}_\phi^*\right].$$

□

A.2 Excess risk bounds

Lemma A.2.1. *Suppose that Assumptions 3.1 (i)-(ii) are satisfied and that there exists a constant $F<\infty$ such that $|f|\leq F$ for all $f\in\mathscr{F}_n$. Then the exponential and the logistic convexifying functions satisfy Assumption 4.1 with $\kappa=1$.*

Proof. First note that for every $f\in\mathscr{F}_n$

$$\begin{aligned}
\mathcal{R}_\phi(f) &= \mathbb{E}\left[(Ya(X)-b(X))\phi(-Yf(X))\right] \\
&= \mathbb{E}\left[\eta(X)(a(X)-b(X))\phi(-f(X)) + (\eta(X)-1)(a(X)+b(X))\phi(f(X))\right].
\end{aligned}$$

Then, since f_ϕ^* minimizes the convexified risk, by Taylor's theorem there exists $\xi\in[0,1]$ such that

$$\begin{aligned}
\mathcal{R}_\phi(f) - \mathcal{R}_\phi^* &= \frac{1}{2}\mathbb{E}\left[\eta(X)(a(X)-b(X))\phi''(-\xi f(X)-(1-\xi)f_\phi^*(X))(f(X)-f_\phi^*(X))^2\right] \\
&\quad + \frac{1}{2}\mathbb{E}\left[(\eta(X)-1)(a(X)+b(X))\phi''(\xi f(X)+(1-\xi)f_\phi^*(X))(f(X)-f_\phi^*(X))^2\right].
\end{aligned}$$

For the exponential convexifying function $\phi''(z) = e^z$ while for the logistic convexifying function $\phi''(z) = \frac{e^z}{(1+e^z)^2}$. In both cases ϕ'' is strictly positive and uniformly bounded away from zero on $[-F, F]$. Under Assumption 3.1 this shows that

$$\mathcal{R}_\phi(f) - \mathcal{R}_\phi^* \gtrsim \mathbb{E}|f(X) - f_\phi^*(X)|^2.$$

□

Lemma A.2.2. *Suppose that $|f| \leq 1$. Then the Hinge convexifying function satisfies*

$$\mathcal{R}_\phi(f) - \mathcal{R}_\phi^* \quad = -2 \int_{\mathcal{X}} b \left| f - f_\phi^* \right| |\eta - c| \mathrm{d}P_X.$$

Proof. Since, $|f| \leq 1$ and $f_\phi^*(x) = \operatorname{sign}(\eta(x) - c(x))$, cf., Lemma A.1.4, we have

$$\begin{aligned}
\mathcal{R}_\phi(f) - \mathcal{R}_\phi^* &= \mathbb{E}[(a(X)Y - b(X))Y(f_\phi^*(X) - f(X))] \\
&= \mathbb{E}[(a(X) + b(X) - 2\eta(X)b(X))(f_\phi^*(X) - f(X))] \\
&= -2\mathbb{E}b(X)(\eta(X) - c(X))(f_\phi^*(X) - f(X)) \\
&= -2 \int_{\mathcal{X}} b \left| f - f_\phi^* \right| |\eta - c| \mathrm{d}P_X.
\end{aligned}$$

□

Lemma A.2.3. *Suppose that Assumptions 3.1 and 3.3 are satisfied and that $|f| \leq 1$ for all $f \in \mathscr{F}_n$. Then the Hinge convexifying function satisfies Assumption 4.1 with $\kappa = 1 + 1/\alpha$.*

Proof. Under Assumption 3.1, for all $u > 0$

$$\begin{aligned}
\|f - f_\phi^*\|^2 &\leq 2 \int_{\mathcal{X}} |f - f_\phi^*| \mathrm{d}P_X \\
&= 2 \int_{|\eta - c| > u} \left| f - f_\phi^* \right| \mathrm{d}P_X + 2 \int_{|\eta - c| \leq u} \left| f - f_\phi^* \right| \mathrm{d}P_X \\
&\lesssim u^{-1} \int_{\mathcal{X}} [-2b] \left| f - f_\phi^* \right| |\eta - c| \mathrm{d}P_X + P_X(|\eta - c| \leq u) \\
&\lesssim u^{-1} [\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*] + u^\alpha,
\end{aligned}$$

where last line follows by Lemma A.2.2. Setting $u = [\mathcal{R}_\phi(f) - \mathcal{R}_\phi^*]^{\frac{1}{1+\alpha}}$ yields the result. □

Proof of Theorem 4.1. By Theorem 3.1

$$\begin{aligned}
\mathcal{R}(\operatorname{sign}(\hat{f}_n)) - \mathcal{R}^* &\lesssim \left[\mathcal{R}_\phi(\hat{f}_n) - \mathcal{R}_\phi^* \right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}} \\
&\triangleq \left[\mathcal{R}_\phi(\hat{f}_n) - \mathcal{R}_\phi(f_n^*) + \triangle_n \right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}
\end{aligned}$$

with $\triangle_n \triangleq \inf_{f\in\mathscr{F}_n} \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi$. We will bound the stochastic term by Koltchinskii (2011), Theorem 4.3. To that end, put $\mathscr{F}(\delta) = \{\ell\circ f:\ \mathcal{R}_\phi(f) - \mathcal{R}_\phi(f^*_n) \leq \delta,\ f\in\mathscr{F}_n\}$ for some $\delta > 0$ and $(\ell\circ f)(y,x) = [ya(x) - b(x)]\phi(-yf(x))$. Then for every $f\in\mathscr{F}_n$

$$\begin{aligned}
\|f - f^*_n\| &\leq \|f - f^*_\phi\| + \|f^*_n - f^*_\phi\| \\
&\lesssim \left[\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right]^{\frac{1}{2\kappa}} + \left[\mathcal{R}_\phi(f^*_n) - \mathcal{R}^*_\phi\right]^{\frac{1}{2\kappa}} \\
&\leq \left[\mathcal{R}_\phi(f) - \mathcal{R}^*_\phi + \mathcal{R}_\phi(f^*_n) - \mathcal{R}^*_\phi\right]^{\frac{1}{2\kappa}} \leq [\delta + 2\triangle_n]^{\frac{1}{2\kappa}},
\end{aligned} \tag{A.3}$$

where the second inequality follows under Assumption 4.1 and the third by Jensen's inequality since $\kappa \geq 1$. Therefore,

$$\mathscr{F}(\delta) \subset \left\{\ell\circ f:\ \|f - f^*_n\| \leq K[\delta + 2\triangle_n]^{\frac{1}{2\kappa}},\ f\in\mathscr{F}_n\right\}. \tag{A.4}$$

Under Assumptions 3.1 and 3.2 for all $f_1, f_2 \in \mathscr{F}_n$ and all $(y,x)\in\{-1,1\}\times\mathcal{X}$

$$\begin{aligned}
|(\ell\circ f_1)(y,x) - (\ell\circ f_2)(y,x)| &= |(a(x)y - b(x))\,(\phi(-yf_1(x)) - \phi(-yf_2(x)))| \\
&\lesssim |f_1(x) - f_2(x)|.
\end{aligned}$$

In conjunction with inequalities in equations (A.3) and (A.4) this shows that the L_2-diameter of $\mathscr{F}(\delta)$ satisfies

$$D(\delta) \triangleq \sup_{g_1,g_2\in\mathscr{F}(\delta)} \|g_1 - g_2\| \lesssim [\delta + 2\triangle_n]^{\frac{1}{2\kappa}},$$

and whence

$$(D^2)^\flat(\sigma) \triangleq \sup_{\delta\geq\sigma} \frac{D^2(\delta)}{\delta} \lesssim \sup_{\delta\geq\sigma} \delta^{\frac{1}{\kappa}-1}[1 + 2\triangle_n/\delta]^{\frac{1}{\kappa}} \lesssim \sigma^{\frac{1}{\kappa}-1}[1+2\tau]^{\frac{1}{\kappa}},$$

where $\tau = \triangle_n/\sigma$. Likewise, it follows from equation (A.4) that

$$\begin{aligned}
\phi_n(\delta) &\triangleq \mathbb{E}\left[\sup_{g_1,g_2\in\mathscr{F}(\delta)} |(P_n - P)(g_1 - g_2)|\right] \\
&\leq 2\mathbb{E}\left[\sup_{g\in\mathscr{F}(\delta)} |(P_n - P)(g - \ell\circ f_n^*)|\right] \\
&\leq 2\mathbb{E}\left[\sup_{f\in\mathscr{F}_n:\|f-f_n^*\|\leq K[\delta+2\triangle_n]^{\frac{1}{2\kappa}}} |(P_n - P)(\ell\circ f - \ell\circ f_n^*)|\right] \\
&\leq 4\mathbb{E}\left[\sup_{f\in\mathscr{F}_n:\|f-f_n^*\|\leq K[\delta+2\triangle_n]^{\frac{1}{2\kappa}}} |R_n(\ell\circ f - \ell\circ f_n^*)|\right] \\
&\lesssim \mathbb{E}\left[\sup_{f\in\mathscr{F}_n:\|f-f_n^*\|\leq K[\delta+2\triangle_n]^{\frac{1}{2\kappa}}} |R_n(f - f_n^*)|\right] \\
&= \psi_n\left(K^2[\delta + 2\triangle_n]^{\frac{1}{\kappa}}; \mathscr{F}_n\right),
\end{aligned}$$

where the last two by the symmetrization and contraction inequalities, cf., Koltchinskii (2011), Theorems 2.1 and 2.3. This gives

$$\begin{aligned}
\phi_n^\flat(\sigma) &= \sup_{\delta\geq\sigma}\frac{\phi_n(\delta)}{\delta} \\
&\lesssim \sup_{\delta\geq\sigma}\frac{\psi_n\left(\delta^{\frac{1}{\kappa}}K^2\left[1+2\tau\right]^{\frac{1}{\kappa}}; \mathscr{F}_n\right)}{\delta} \\
&\leq \sigma^{\frac{1}{\kappa}-1}[1+2\tau]^{\frac{1}{\kappa}}\psi_n^\flat\left(\sigma^{\frac{1}{\kappa}}K^2\left[1+2\tau\right]^{\frac{1}{\kappa}}\right).
\end{aligned}$$

Therefore, by Koltchinskii (2011), Theorem 4.3, there exists a universal constant $c > 0$ such that for every $t > 0$

$$\Pr\left(\mathcal{R}_\phi(\hat{f}_n) - \mathcal{R}_\phi(f_n^*) \leq \inf\left\{\sigma :\ V_n^t(\sigma) \leq 1\right\}\right) \geq 1 - ce^{-t}$$

with

$$\begin{aligned}
V_n^t(\sigma) &\triangleq 2q\left[\phi_n^\flat(\sigma) + \sqrt{\frac{(D^2)^\flat(\sigma)t}{n\sigma}} + \frac{t}{n\sigma}\right] \\
&\lesssim \sigma^{\frac{1-\kappa}{\kappa}}[1+2\tau]^{\frac{1}{\kappa}}\psi_n^\flat\left(\sigma^{\frac{1}{\kappa}}K^2\left[1+2\tau\right]^{\frac{1}{\kappa}}\right) + \sqrt{\frac{[1+2\tau]^{\frac{1}{\kappa}}t}{n\sigma^{2-\frac{1}{\kappa}}}} + \frac{t}{n\sigma}.
\end{aligned}$$

If $\sigma \leq \triangle_n$, then $\inf\{\sigma :\ V_n^t(\sigma)\} \leq \triangle_n$. If $\sigma \geq \triangle_n$, then $\tau = \triangle_n/\sigma \leq 1$, and so

$$V_n^t(\sigma) \lesssim \sigma^{\frac{1-\kappa}{\kappa}} \psi_n^\flat \left(\sigma^{\frac{1}{\kappa}} C\right) + \sqrt{\frac{t}{n\sigma^{2-\frac{1}{\kappa}}}} + \frac{t}{n\sigma}$$

with $C = K^2 3^{\frac{1}{\kappa}}$. Since all functions in this upper bound are decreasing in σ, we have $V_n^t(\sigma) \leq 1$ as soon as

$$\sigma \geq K \left[\psi_{n,\kappa}^\sharp(\epsilon) \vee \left(\frac{t}{n}\right)^{\frac{\kappa}{2\kappa-1}} \vee \frac{t}{n}\right]$$

with a sufficiently large constants $\epsilon, K > 0$. Therefore,

$$\inf\{\sigma :\ V_n^t(\sigma) \leq 1\} \leq K \left[\triangle_n \vee \psi_{n,\kappa}^\sharp(\epsilon) \vee \left(\frac{t}{n}\right)^{\frac{\kappa}{2\kappa-1}} \vee \frac{t}{n}\right],$$

which implies the bound of the theorem. □

Proof of Theorem 4.2. By Koltchinskii (2011), Proposition 3.2, $\psi_n(\delta) \lesssim \sqrt{\frac{\delta p}{n}}$. Consequently, $\psi_n^\flat(\sigma) \lesssim \sqrt{\frac{p}{\sigma n}}$, and whence

$$\psi_{n,\kappa}^\sharp(\epsilon) \lesssim \left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Put $\triangle_n = \inf_{f\in\mathscr{F}_n} \mathcal{R}_\phi(f) - \mathcal{R}_\phi^*$. Under maintained assumption, by Theorem 4.1, there exists a constant C such that for every $t > 0$

$$\Pr\left(\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^* > C\left[\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} \bigvee tn^{-\frac{\kappa}{2\kappa-1}} + \triangle_n\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}\right) \leq ce^{-t},$$

where we use $t^{\kappa/(2\kappa-1)} \leq p \vee t$ since $\kappa \geq 1$. Therefore, there exists $C > 0$ such that for every $t > 0$

$$\Pr\left(\left(\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^*\right)^{\frac{\gamma\alpha+1}{\gamma(\alpha+1)}} > C\left(\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \triangle_n + tn^{-\frac{\kappa}{2\kappa-1}}\right)\right) \leq ce^{-t}.$$

Integrating the tail bound

$$\begin{aligned}
\mathbb{E}\left[\left(\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^*\right)^{\frac{\gamma\alpha+1}{\gamma(\alpha+1)}}\right] &= \int_0^\infty \Pr\left(\left(\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^*\right)^{\frac{\gamma\alpha+1}{\gamma(\alpha+1)}} > u\right)\mathrm{d}u \\
&\leq Cn^{-\frac{\kappa}{2\kappa-1}}\left(Cn^{\frac{\kappa}{2\kappa-1}}\left(\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \triangle_n\right) + \int_0^\infty ce^{-t}\mathrm{d}t\right) \\
&\lesssim \left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \triangle_n,
\end{aligned}$$

where the second line follows by the change of variables $u = C\left(\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \triangle_n + tn^{-\frac{\kappa}{2\kappa-1}}\right)$ and bounding the probability by 1 for $t < 0$. Since $\gamma \in (0, 1]$, by Jensen's inequality, this gives

$$\mathbb{E}\left[\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^*\right] \lesssim \left[\left(\frac{p}{n}\right)^{\frac{\kappa}{2\kappa-1}} + \triangle_n\right]^{\frac{\gamma(\alpha+1)}{\gamma\alpha+1}}.$$

$\square$

Proof of Theorem 4.3. By Lemmas A.1.3, A.1.5, and A.2.1, $\gamma = 1/2$ and $\kappa = 1$. Therefore, by Theorem 4.1, for every $t > 0$ with probability at least $1 - ce^{-t}$

$$\mathcal{R}(\mathrm{sign}(\hat{f})) - \mathcal{R}^* \leq K\left[\psi^\sharp_{n,1}(\epsilon) + \frac{t}{n} + \inf_{f\in\mathscr{F}^{\mathrm{B}}} \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi\right]^{\frac{\alpha+1}{\alpha+2}}.$$

By Koltchinskii (2011), Example 5 on p. 87,

$$\psi^\sharp_{n,1}(\epsilon) \leq \left(\frac{C_V}{n\epsilon^2}\right)^{\frac{2+V}{2(1+V)}}$$

for some universal constant $C_V > 0$ that depends only on V.

For the approximation error, under Assumptions 3.1 and 3.2

$$\begin{aligned}\inf_{f\in\mathscr{F}^{\mathrm{B}}} \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi &= \inf_{f\in\mathscr{F}^{\mathrm{B}}} \mathbb{E}[(a(X)Y - b(X))(\phi(-Yf(X)) - \phi(-Yf^*_\phi(X)))] \\ &\lesssim \inf_{f\in\mathscr{F}^{\mathrm{B}}} \|f - f^*_\phi\|_{L_1(P_X)}.\end{aligned}$$

The result follows from integrating the tail bound. $\square$

Proof of Theorem 4.4. By Theorem 4.1, for every $t > 0$ with probability at least $1 - ce^{-t}$

$$\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^* \lesssim \psi^\sharp_{n,1+1/\alpha}(\epsilon) + \left(\frac{t}{n}\right)^{\frac{1+\alpha}{2+\alpha}} + \frac{t}{n} + \inf_{f\in\mathscr{F}^{\mathrm{S}}_n} \mathcal{R}_\phi(f) - \mathcal{R}^*_\phi,$$

where we use the fact that $\gamma = 1$ and $\kappa = 1 + 1/\alpha$ by Lemmas A.1.4 and A.2.3. By Lemmas A.3.1 and A.3.3

$$\psi^\sharp_{n,1+1/\alpha}(\epsilon) \leq C\left(\frac{p_n \log p_n}{n} \log\left(\frac{n}{p_n \log p_n}\right)\right)^{\frac{1+\alpha}{2+\alpha}}.$$

Next, under Assumption 4.2 (ii), by Mhaskar (1996), Theorem 2.1, there exists $\eta_n \in \Theta^{\mathrm{S}}_n$ such that

$$\|\eta_n - \eta\|_\infty \leq Cp_n^{-\beta/d} \triangleq \varepsilon_n.$$

Define

$$f_n(x) = \left(\frac{\eta_n(x) - c(x)}{\varepsilon_n} + 1\right)_+ - \left(\frac{\eta_n(x) - c(x)}{\varepsilon_n} - 1\right)_+ - 1$$

and note that $f_n \in \mathscr{F}_n^{\mathrm{S}}$. Note also that

$$f_n(x) = \begin{cases} 1, & \text{if } \eta_n(x) - c(x) > \varepsilon_n \\ \frac{\eta_n(x) - c(x)}{\varepsilon_n}, & \text{if } |\eta_n(x) - c(x)| \leq \varepsilon_n \\ -1, & \text{if } \eta_n(x) - c(x) < -\varepsilon_n \end{cases}$$

and that on the event $\{x \in [0,1]^d : \ |\eta(x) - c(x)| > 2\varepsilon_n\}$ we have $f_n = f_\phi^*$. To see this, recall that by Lemma A.1.4, $f_\phi^* = \mathrm{sign}(\eta - c)$. Then if $\eta - c > 0$, we have $\eta_n - c = (\eta - c) - (\eta - \eta_n) > \varepsilon_n$ while if $\eta - c < 0$, we have $\eta_n - c = (\eta - c) + (\eta_n - \eta) < -\varepsilon_n$. Therefore, by Lemma A.2.2

$$\begin{aligned}
\inf_{f \in \mathscr{F}_n^{\mathrm{S}}} \mathcal{R}_\phi(f) - \mathcal{R}_\phi^* &= \inf_{f \in \mathscr{F}_n^{\mathrm{S}}} -2 \int_{[0,1]^d} b|f - f_\phi^*||\eta - c| \mathrm{d}P_X \\
&\leq -2 \int_{[0,1]^d} b|f_n - f_\phi^*||\eta - c| \mathrm{d}P_X \\
&= -2 \int_{|\eta - c| \leq 2\varepsilon_n} b|f_n - f_\phi^*||\eta - c| \mathrm{d}P_X \\
&\leq -4\varepsilon_n \int_{|\eta - c| \leq 2\varepsilon_n} b|f_n - f_\phi^*| \mathrm{d}P_X \\
&\lesssim \varepsilon_n P_X(|\eta - c| \leq 2\varepsilon_n) \\
&\lesssim \varepsilon_n^{1+\alpha}
\end{aligned}$$

where the last two lines follow under Assumptions 3.1 (iii) and 3.3. Therefore, for every $t > 0$ with probability at least $1 - ce^{-t}$

$$\mathcal{R}(\mathrm{sign}(\hat{f}_n)) - \mathcal{R}^* \lesssim \left(\frac{p_n \log^2 n}{n}\right)^{\frac{1+\alpha}{2+\alpha}} + p_n^{-(1+\alpha)\beta/d} + \left(\frac{t}{n}\right)^{\frac{1+\alpha}{2+\alpha}} + \frac{t}{n}.$$

The second statement follows from integrating this tail bound in the same way as in the proof of Theorem 4.2. The uniformity follows from the fact that all constants do no depend on the specific distribution of (X, Y) in $\mathcal{P}(\alpha, \beta)$. □

Proof of Theorem 4.5. Under Assumption 4.3, by Lu, Shen, Yang, and Zhang (2020), Theorem 1.1, there exists $\eta_n \in \mathcal{F}_n^{\mathrm{DNN}}$ with width $W_n = O(J_n \log J_n)$ and depth $L_n = O(K_n \log K_n)$ such that

$$\|\eta_n - \eta\|_\infty = O\left((J_n K_n)^{-2\beta/d}\right) \triangleq \varepsilon_n.$$

Put

$$f_n(x) \triangleq \sigma\left(\frac{\eta_n(x)-c(x)}{\varepsilon_n}+1\right)-\sigma\left(\frac{\eta_n(x)-c(x)}{\varepsilon_n}-1\right)-1.$$

Then

$$f_n(x)=\begin{cases}1 & \text{if } \eta_n(x)-c(x)>\varepsilon_n,\\ \frac{\eta_n(x)-c(x)}{\varepsilon_n} & \text{if } |\eta_n(x)-c(x)|\leq\varepsilon_n,\\ -1 & \text{if } \eta_n(x)-c(x)<-\varepsilon_n.\end{cases}$$

Then on the event $\{x\in\mathcal{X}:|\eta(x)-c(x)|>2\varepsilon_n\}$, we have $f_n(x)=f^*_\phi(x)$. To see this note that $f^*_\phi(x)=\operatorname{sign}(\eta(x)-c(x))$ and that if $\eta(x)>c(x)$, we have $\eta_n(x)-c(x)=(\eta(x)-c(x))-(\eta(x)-\eta_n(x))\geq\varepsilon_n$ while if $\eta(x)<c(x)$, we have $\eta_n(x)-c(x)<-\varepsilon_n$. Therefore, by Lemma A.2.2

$$\begin{aligned}\inf_{f\in\mathscr{F}_n^{\mathrm{DNN}}}\mathcal{R}_\phi(f)-\mathcal{R}^*_\phi &\leq \mathcal{R}_\phi(f_n)-\mathcal{R}^*_\phi\\ &=\int_{\mathcal{X}}[-2b(x)]|\eta(x)-c(x)||f_n(x)-f^*_\phi(x)|\mathrm{d}P_X(x)\\ &\lesssim\int_{|\eta-c|\leq 2\varepsilon_n}|\eta(x)-c(x)||f_n(x)-f^*_\phi(x)|\mathrm{d}P_X(x)\\ &\lesssim\varepsilon_n P_X(|\eta-c|\leq 2\varepsilon_n)\\ &\leq\varepsilon_n^{1+\alpha},\end{aligned}$$

where the third line follows since $-b,f,f^*_\phi,\eta,c\lesssim 1$ under Assumption 3.1 and the last under Assumption 3.3.

By Bartlett, Harvey, Liaw, and Mehrabian (2019), Theorem 7

$$V\lesssim p_nL_n\log U_n\lesssim(W_nL_n)^2\log(W_nL_n).$$

Therefore, by Lemmas A.3.2 and A.3.3 under Assumption 4.3

$$\psi^\sharp_{n,\kappa}(\epsilon)\leq C\left(\frac{(W_nL_n)^2\log(W_nL_n)}{n}\log\left(\frac{n}{(W_nL_n)^2\log(W_nL_n)}\right)\right)^{\frac{1+\alpha}{2+\alpha}}.$$

If W_n,L_n are of polynomial order, then by Theorem 4.1 there exists $c>0$ such that for every $t>0$ with probability at least $1-ce^{-t}$

$$\mathcal{R}(\operatorname{sign}(\hat f_n))-\mathcal{R}^*\lesssim\left(\frac{(J_nK_n)^2\log^2(J_n)\log^2(K_n)\log^2 n}{n}\right)^{\frac{1+\alpha}{2+\alpha}}+(J_nK_n)^{-2\beta(1+\alpha)/d}+\left(\frac{t}{n}\right)^{\frac{1+\alpha}{2+\alpha}}+\frac{t}{n}.$$

Lastly, since $J_nK_n\sim n^{\frac{d}{2\beta(2+\alpha)+2d}}$ under Assumption 4.3 (ii), we obtain

$$\mathcal{R}(\operatorname{sign}(\hat f_n))-\mathcal{R}^*\lesssim\left(\frac{\log^6 n}{n}\right)^{\frac{\beta(1+\alpha)}{\beta(2+\alpha)+d}}+\left(\frac{t}{n}\right)^{\frac{1+\alpha}{2+\alpha}}+\frac{t}{n}.$$

The bound in expectations follows from integrating the tail bound. $\square$

A.3 Local Rademacher complexities

In this section, we obtain useful bounds on the local Rademacher complexities for shallow and deep neural network classes. First, we consider the shallow learning class

$$\mathscr{F}_n^{\mathrm{S}} = \left\{\sigma(\theta + cd + 1) - \sigma(\theta + cd - 1) - 1 : \ \theta \in \Theta_n^{\mathrm{S}}, \ |d| \leq n\right\},$$

where $\sigma(z) = \max\{z, 0\}$,

$$\Theta_n^{\mathrm{S}} = \left\{ x \mapsto \sum_{j=1}^{p_n} b_j \sigma_0(a_j^\top x + a_0) + b_0, \quad a \in \mathbf{R}^{d+1}, \ |b|_1 \leq \gamma_n \right\},$$

and $a = (a_0, a_1, \dots, a_{p_n})$, $b = (b_0, b_1, \dots b_{p_n})$.

The following result bounds the local Rademacher complexity of the shallow learning class $\mathscr{F}_n^{\mathrm{S}}$ in terms of the number of neurons p_n in the inner neural network class Θ_n^{S}.

Lemma A.3.1. *Suppose that (i) $\sigma_0 : \mathbf{R} \to [-b, b]$ is non-decreasing with bounded Lipschitz constant; (ii) $p_n \sim n^{c_1}$ and $\gamma_n \sim n^{c_2}$ for some constants $c_1, c_2 > 0$; (iii) $\|c\|_\infty < \infty$. Then there exist universal constants $A, C > 0$ such that*

$$\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) \leq C \left[\sqrt{\frac{p_n \delta (1 \vee \log(An/\delta^{1/2}))}{n}} \bigvee \frac{p_n(1 \vee \log(An/\delta^{1/2}))}{n} \right].$$

Proof. Put

$$\mathscr{F}_n(\delta) \triangleq \{f - f_n^* : \ f \in \mathscr{F}_n^{\mathrm{S}}, \ \|f - f_n^*\| \leq \sqrt{\delta}\}$$

and $\sigma_n^2 \triangleq \sup_{g \in \mathscr{F}_n(\delta)} P_n g^2$. By Dudley's entropy bound, cf., Koltchinskii (2011), Theorem 3.11

$$\begin{aligned} \psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) &= \mathbb{E}\left[\sup_{g \in \mathscr{F}_n(\delta)} |R_n g|\right] \\ &\lesssim \frac{1}{\sqrt{n}} \mathbb{E} \int_0^{2\sigma_n} \sqrt{\log N(\mathscr{F}_n(\delta), L_2(P_n), \varepsilon)} \mathrm{d}\varepsilon \end{aligned} \tag{A.5}$$

and by symmetrization and contraction inequalities, cf., Koltchinskii (2011), Theorems 2.1 and 2.3

$$\begin{aligned} \mathbb{E}\sigma_n^2 &\leq \mathbb{E}\left[\sup_{f \in \mathscr{F}_n(\delta)} |(P_n - P)g^2|\right] + \sup_{f \in \mathscr{F}_n(\delta)} Pg^2 \\ &\leq 2\mathbb{E}\left[\sup_{g \in \mathscr{F}_n(\delta)} |R_n g^2|\right] + \delta \\ &\leq 16\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) + \delta \\ &\triangleq B. \end{aligned}$$

Given a sequence $x = (x_1, \dots, x_n) \in \mathcal{X}^n$ and a class of functions from $\mathcal{X}$ to $\mathbf{R}$, denoted $\mathscr{F}$, put $\mathscr{F}|_x \triangleq \{(f(x_1), f(x_2), \dots, f(x_n)) : f \in \mathscr{F}\} \subset \mathbf{R}^n$. For $\varepsilon > 0$, let $N(\mathscr{F}|_x, \ell_\infty, \varepsilon)$ be the ε-covering number of $\mathscr{F}|_x$ with respect to the ℓ_∞ distance. The uniform covering number is defined as

$$N_\infty(\mathscr{F}_n(\delta), n, \varepsilon) \triangleq \max\{N(\mathscr{F}_n(\delta)|_x, \ell_\infty, \varepsilon) : x \in \mathcal{X}^n\}.$$

By Anthony and Bartlett (2009), Lemma 10.5

$$\begin{aligned} N(\mathscr{F}_n(\delta), L_2(P_n), \varepsilon) &= N(\mathscr{F}_n(\delta)|_{X_1,\dots,X_n}, \ell_2, \varepsilon) \\ &\leq \max\{N(\mathscr{F}_n^{\mathrm{S}}|_x, \ell_2, \varepsilon) : x \in \mathcal{X}^n\} \\ &\leq N_\infty(\mathscr{F}_n^{\mathrm{S}}, n, \varepsilon) \end{aligned}$$

Note that for $f_1, f_2 \in \mathscr{F}_n^{\mathrm{S}}$, we have $f_j(x) = \sigma(\theta_j(x) + c(x)d_j + 1) - \sigma(\theta_j(x) + c(x)d_j - 1) - 1$ for some $\theta_j \in \Theta_n^{\mathrm{S}}$ and $|d_j| \leq n$ with $j = 1, 2$. Since $x \mapsto \sigma(x+1) - \sigma(x-1) - 1$ is Lipschitz continuous with Lipschitz constant 1

$$\max_{1\leq i\leq n} |f_1(X_i) - f_2(X_i)| \leq \max_{1\leq i\leq n} |\theta_1(X_i) - \theta_2(X_i)| + \|c\|_\infty |d_1 - d_2|$$

Note that the ε-covering number of $[-n, n]$ is n/ε. Therefore,

$$\begin{aligned} N_\infty(\mathscr{F}_n^{\mathrm{S}}, n, \varepsilon) &\leq \frac{2\|c\|_\infty n}{\varepsilon} N_\infty(\Theta_n^{\mathrm{S}}, n, \varepsilon/2) \\ &\leq \frac{2\|c\|_\infty n}{\varepsilon}\left(\frac{Anp_n\gamma_n}{\varepsilon}\right)^{(d+1)(p_n+1)} \\ &\lesssim \left(\frac{Anp_n\gamma_n}{\varepsilon}\right)^{2(d+1)(p_n+1)}, \end{aligned}$$

where the second inequality follows by Anthony and Bartlett (2009), Theorem 14.5 with some universal constant $A > 0$. In conjunction with equation (A.5), this shows that

$$\begin{aligned} \psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) &\lesssim \sqrt{\frac{p_n}{n}} \mathbb{E} \int_0^{2\sigma_n} \sqrt{\log(Anp_n\gamma_n/\varepsilon)} \mathrm{d}\varepsilon \\ &\leq \sqrt{\frac{p_n}{n}} \int_0^{2\sqrt{\mathbb{E}\sigma_n^2}} \sqrt{\log(Anp_n\gamma_n/\varepsilon)} \mathrm{d}\varepsilon \\ &\lesssim \sqrt{\frac{p_n}{n}} \int_0^{2\sqrt{\mathbb{E}\sigma_n^2}} \sqrt{\log(An/\varepsilon)} \mathrm{d}\varepsilon \\ &= A\sqrt{p_n n} \int_0^{2\sqrt{\mathbb{E}\sigma_n^2}/An} \sqrt{\log(1/u)} \mathrm{d}u \\ &= A\sqrt{p_n n} \int_0^{2\sqrt{B}/An} \sqrt{\log(1/u)} \mathrm{d}u \\ &\leq 4\sqrt{\frac{p_n B}{n}} \left\{1 \vee \sqrt{\log\left(\frac{An}{\delta^{1/2}}\right)}\right\}, \end{aligned}$$

where the second line follows by Jensen's inequality since $t \mapsto \int_0^t h(u)\mathrm{d}u$ is concave whenever h is non-decreasing, and the last from

$$\int_0^a \sqrt{\log(1/u)}\mathrm{d}u \leq 2a\left\{1 \vee \sqrt{\log(1/a)}\right\},$$

and $\delta \leq B$. Therefore,

$$\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) \lesssim \sqrt{\frac{p_n}{n}}\left\{1 \vee \sqrt{\log(An/\delta^{1/2})}\right\}\left(\sqrt{\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}})} + \sqrt{\delta}\right).$$

Solving this inequality for $\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}})$ gives

$$\psi_n(\delta; \mathscr{F}_n^{\mathrm{S}}) \lesssim \sqrt{\frac{p_n\delta\{1 \vee \log(An/\delta^{1/2})\}}{n}} \bigvee \frac{p_n\{1 \vee \log(An/\delta^{1/2})\}}{n}.$$

□

Next, we consider the deep learning class

$$\mathscr{F}_n^{\mathrm{DNN}} = \left\{\sigma(\theta + cd + 1) - \sigma(\theta + cd - 1) - 1 :\ |d| \leq n,\ \theta \in \Theta_n^{\mathrm{DNN}}\right\},$$

where

$$\Theta_n^{\mathrm{DNN}} = \left\{x \mapsto A_L\sigma_{\mathbf{b}_L} \circ A_{L-1}\sigma_{\mathbf{b}_{L-1}} \circ \cdots \circ A_1\sigma_{\mathbf{b}_1} \circ A_0 x\right\}$$

is a deep neural network class such that $\{A_l, \mathbf{b}_l :\ 1 \leq l \leq L\}$ and A_0 are unrestricted free parameters.

The following result bounds the local Rademacher complexity of the deep learning class $\mathscr{F}_n^{\mathrm{DNN}}$ in terms of the pseudo-dimension of the class Θ_n^{DNN}. To define the pseudo-dimension, let Θ be arbitrary class of functions from $\mathcal{X}$ to $\mathbf{R}$. The pseudo-dimension of Θ is the largest integer m for which there exists $(x_1, \ldots, x_m, y_1, \ldots, y_m) \in \mathcal{X}^m \times \mathbf{R}^m$ such that for every $(b_1, \ldots, b_m) \in \{0,1\}^m$, there exists $\theta \in \Theta$ such that

$$\theta(x_i) > y_i \iff b_i = 1, \qquad \forall 1 \leq i \leq m.$$

Lemma A.3.2. *Suppose that $\|c\|_\infty < \infty$ and that Θ_n^{DNN} has pseudo-dimension $V \leq n$. Then there exist universal constants $A, C > 0$ such that*

$$\psi_n(\delta; \mathscr{F}_n^{\mathrm{DNN}}) \leq C\left[\sqrt{\frac{V\delta(1 \vee \log(An/\delta^{1/2}))}{n}} \bigvee \frac{V(1 \vee \log(An/\delta^{1/2}))}{n}\right].$$

Proof. The proof follows the same steps as the the proof of Lemma A.3.1, but now we bound the uniform covering numbers relying on Anthony and Bartlett (2009), Theorem 12.2

$$N_\infty(\Theta_n^{\mathrm{DNN}}, n, \varepsilon/2) \leq \left(\frac{8en}{\varepsilon V}\right)^V.$$

□

Lemma A.3.3. *Suppose that*

$$\psi_n(\delta;\mathscr{F}_n) \leq C\left[\sqrt{\frac{V_n\delta(1\vee\log(An/\delta^{1/2}))}{n}}\bigvee\frac{V_n(1\vee\log(An/\delta^{1/2}))}{n}\right]$$

for some $V_n = o(n)$ and $\delta, A > 0$. Then there exists a universal constant $C > 0$ such that

$$\psi^{\sharp}_{n,\kappa}(\epsilon) \leq C\left(\frac{V_n\log(n/V_n)}{n}\right)^{\frac{\kappa}{2\kappa-1}}.$$

Proof. By Lemma A.3.1

$$\psi^{\flat}_n(\sigma) = \sup_{\delta\geq\sigma}\frac{\psi_n(\delta;\mathscr{F}_n)}{\delta} = \sqrt{\frac{V_n(1\vee\log(An/\delta^{1/2}))}{\sigma n}}\bigvee\frac{V_n(1\vee\log(An/\delta^{1/2}))}{\sigma n}$$

and whence

$$\begin{aligned}\psi^{\sharp}_{n,\kappa}(\epsilon) &= \inf\left\{\sigma>0:\ \sigma^{1/\kappa-1}\psi^{\flat}_n(\sigma^{1/\kappa})\leq K\right\}\\ &\leq \inf\left\{\sigma>0:\ \sqrt{\frac{V_n(1\vee\log(An/\sigma^{1/2\kappa}))}{\sigma^{2-1/\kappa}n}}\bigvee\frac{V_n(1\vee\log(An/\sigma^{1/2\kappa}))}{\sigma n}\leq K\right\}.\end{aligned}$$

Since the two functions inside the infimum are decreasing in σ, we have $\psi^{\sharp}_{n,\kappa}(\epsilon)\leq\sigma_1\vee\sigma_2$ with σ_1 and σ_2 solving

$$\frac{V_n\log(An\sigma_1^{-1/2\kappa})}{n\sigma_1^{2-1/\kappa}} = K^2 \qquad\text{and}\qquad \frac{V_n\log(An\sigma_2^{-1/2\kappa})}{n\sigma_2} = K.$$

To bound σ_1 and σ_2, note that

$$\frac{v\log(c/x)}{x^a} = b \iff x = \left(\frac{v}{ab}W_0\left(\frac{abc^a}{v}\right)\right)^{1/a},$$

where W_0 is the Lambert W-function. Since $W_0(z)\leq\log z, \forall z\geq e$ and $W_0(z)\leq W(e)$ for all $z\in(0,e]$, this yields

$$x\leq\left(\frac{v}{ab}W_0(e)\right)^{1/a}\bigvee\left(\frac{v}{ab}\log\left(\frac{abc^a}{v}\right)\right)^{1/a}$$

Therefore, there exists a constant $A>0$ such that

$$\psi^{\sharp}_{n,\kappa}(\epsilon)\lesssim\left(\frac{V_n}{n}\right)^{\frac{\kappa}{2\kappa-1}}\bigvee\left(\frac{V_n\log(An/V_n)}{n}\right)^{\frac{\kappa}{2\kappa-1}}\bigvee\frac{V_n}{n}\bigvee\frac{V_n\log(An/V_n)}{n}.$$

for some $A > 0$, and whence since $V_n/n = o(1)$ and $\kappa \geq 1$

$$\begin{aligned}\psi^{\sharp}_{n,\kappa}(\epsilon) &= \inf\left\{\sigma > 0 :\ \sigma^{1/\kappa - 1}\psi^{\flat}_{n}(\sigma^{1/\kappa}) \leq K\right\} \\ &\lesssim \left(\frac{V_n \log(n/V_n)}{n}\right)^{\frac{\kappa}{2\kappa-1}}.\end{aligned}$$

$\square$

Table A.1: Economic costs and benefits - Summary entries for EBD and ECD functions

crimetype	$C_{crimetype}$	$EBD_{crimetype}$
Murder	10,754	11,732
Rape/Sexual Assault	266	353
Aggravated Assault	126	127
Robbery	48	230
Arson/Other	23	292
Motor Vehicle Theft	11	53
Household Burglary	7	64
Forgery/Counterfeiting	5	46
Fraud	5	49
Larceny/Theft	3	43

A.4 Pretrial detention costs and benefits

In this section we provide a summary of the cost benefit analysis reported by Baughman (2017) to build a preference-based approach for the empirical application appearing in Section 6. In Table A.2 we provide the key elements of the aforementioned study, which will serve as inputs to the $EBD(c_i)$, $ECD(d_i)$ and $C(x_i, c_i)$ appearing in equation (5). In particular, based on Table A.2 we define:

$$EBD(c_i) = \sum_{crimetype} \mathbb{1}_{crimetype=c_i} EBD_{crimetype}$$

where $EBD_{crimetype}$ is obtained from averaging the entries to Panel B in Table A.2 (and dividing by 1000) and adding the entries of Panel C. These entries are summarized in Table A.1. The function is scaled by 0.05 to balance the costs.

The function $C(x_i, C_i)$ is computed as the median of future recidivism costs, which is 23. Finally, from Panel D in Table A.2 we calculate $ECD(d_i)$, using both individual and public costs, again scaling by 1000. Namely (assuming $m_i = 30d_i$), which yields $0.347d_i$ namely:

$$\begin{aligned}
\mathrm{ECD}(d_i) &= \underbrace{\frac{1}{1000} \times \left[\frac{1036+590}{90} + \frac{1565}{30}\right]}_{\text{Individual Costs I}} d_i + \\
&\quad \underbrace{\frac{1}{1000} \times \left[\frac{31028+1938+103670*(.17)+136191*(.032)}{365}\right]}_{\text{Individual Costs II}} d_i + \\
&\quad \underbrace{\frac{1}{1000} \times \left[\frac{31406+5142+1249+8293}{365}\right]}_{\text{Public Costs}} d_i \\
&= 0.347d_i
\end{aligned}$$

Table A.2: Economic costs and benefits

This table uses inputs from Baughman (2017) Tables 1 through 3 to construct the costs and benefits used in our empirical applications.

	Panel A: Total Per-Offense Cost for Different Crimes ($)	**Panel B: Economic Benefits of Detention ($)**	
Type of Offense		Low Estimate	High Estimate
Murder	10,754,332	4,602,326	18,780,120
Rape/Sexual Assault	266,332	136,191	488,243
Aggravated Assault	126,585	14,715	158,250
Robbery	48,589	12,523	364,898
Arson/Other	23,839	75,453	426,571
Motor Vehicle Theft	11,936	5949	19,299
Household Burglary	7175	2192	44,875
Forgery/Counterfeiting	5821	5731	10,439
Fraud	5563	3950	5478
Larceny/Theft	3906	580	3839
Panel C: Benefit Categories			
Avoidance of Felony for Which No Arrest Is Made			40,338
Avoidance of Failure to Appear			518

Panel D: Economic Costs of Detention

Individual Costs per individual	
Loss of Freedom	$(\$1036/90)d_i$
Loss of Income	$(\$31,028/365)d_i$
Loss of Housing	$\$1565m_i$
Childcare Costs	$(\$1938/365)d_i$
Stolen or Lost Property	$(\$590/3)m_i$
Strain on Intimate Relationships	$(\$103,670(.17)/365)d_i$
Possibility of Violent or Sexual Assault	$(\$136,191(.032)/365)d_i$
Public Costs	
Prison Operation Costs	$(\$31,406/365)d_i$
Loss of Federal Tax	$(5142/365)d_i$
Loss of State Tax	$(\$1249/365)d_i$
Welfare for Detainee's Family	$(\$8293/365)d_i$