Constrained Classification and Policy Learning*

Toru Kitagawa†, Shosei Sakaguchi‡, and Aleksey Tetenov§

May 26, 2021

Abstract

Modern machine learning approaches to classification, including AdaBoost, support vector machines, and deep neural networks, utilize the surrogate-loss techniques to circumvent computational complexity in minimizing the empirical classification risk. These techniques are useful also for causal policy learning problems as estimation of individualized treatment rules can be cast as weighted (cost-sensitive) classification. Consistency of these surrogate-loss approaches studied in Zhang (2004) and Bartlett et al. (2006) crucially relies on the assumption of *correct specification*, meaning that the specified class of policies is rich enough to contain a first-best. This assumption is, however, less credible when the class of policies is constrained by interpretability and/or fairness, leaving applicability of the surrogate-loss based algorithms unknown in such second-best scenarios. This paper studies consistency of the surrogate-loss procedures under a constrained set of policies without assuming correct specification. We show that in the setting where the constraint restricts classifier's prediction set only, the hinge losses (i.e., ℓ_1-support vector machines) are the only surrogate losses that preserve consistency in the second-best scenarios. If the constraint additionally restricts a functional form of the classifiers, consistency of the surrogate loss approach is not guaranteed even with the hinge loss. We therefore characterize conditions for the constrained set of classifiers that can guarantee consistency of the hinge-risk minimizing classifiers. We illustrate implications and uses of our theoretical results in monotone classification by proposing computationally attractive hinge-loss based procedures that are robust to misspecification.

Keywords: Surrogate loss, support vector machine, monotone classification, fairness in machine learning, statistical treatment choice, personalized medicine

*We thank valuable comments from Max Tabord-Meehan. We also thank the seminar participants at Chicago, Penn State, SciencesPo, UC-Irvine, UC-Riverside, and Zürich for beneficial comments. The authors gratefully acknowledge financial support from ERC grant (number 715940) and the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001).

†Department of Economics, University College London. Email: t.kitagawa@ucl.ac.uk.

‡Department of Economics, University College London. Email: s.sakaguchi@ucl.ac.uk.

§Geneva School of Economics and Management, University of Geneva. Email: aleksey.tetenov@unige.ch

1 Introduction

Binary classification, a prediction problem for binary dependent variable $Y \in \{-1, +1\}$ based on covariate information $X \in \mathcal{X}$, is one of the fundamental problems in statistics and econometrics. Many modern machine learning algorithms build on statistically and computationally efficient classification algorithms, and their applications have been generating large impacts on various fields of study and our society in general, e.g., pattern recognition, credit approval systems, personalized recommendation systems, to list a few. Contributions to classification are also instrumental to causal problems of designing individualized treatment assignment policies, since estimation of an optimal treatment assignment policy can be cast to a weighted (cost-sensitive) classification problem (Zadrozny (2003)). As the allocations of the resources in both business and public policy settings become more evidence-based and dependent on algorithms, there have been increasingly active debates on how to make the allocation algorithms respect interpretability and fairness as desired by the society (Dwork et al. (2012)). Understanding theoretical performance guarantee and efficient implementation of classification algorithms under the interpretability or fairness constraints is therefore a problem of fundamental importance with tight connections to our real life.

In the supervised binary classification problem, a common objective is to learn a classification rule that minimizes the probability of false prediction. We denote the distribution of (Y, X) by P and a (non-randomized) classifier by $f : \mathcal{X} \to \mathbb{R}$ that predicts $Y \in \{-1, +1\}$ based on $\text{sign}(f(X))$ where $\text{sign}(\alpha) = 1\{\alpha \geq 0\} - 1\{\alpha < 0\}$. We denote the 0-level set of f by $G_f \equiv \{x \in \mathcal{X} : f(x) \geq 0\} \subset \mathcal{X}$, and refer to G_f as the *prediction set* of f. The goal is to learn a classifier that minimizes *classification risk*:

$$R(f) \equiv P(\text{sign}(f(X)) \neq Y) = E_P[1\{Y \cdot \text{sign}(f(X)) \leq 0\}]. \tag{1}$$

Given a training sample $\{(Y_i, X_i) \sim_{iid} P : i = 1, \ldots, n\}$, the empirical risk minimization principle of Vapnik (1998) recommends to estimate an optimal classifier by minimizing the empirical classification risk,

$$\hat{f} \in \arg\inf_{f \in \mathcal{F}} \widehat{R}(f), \tag{2}$$

$$\widehat{R}(f) \equiv \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \cdot \text{sign}(f(X_i)) \leq 0\}, \tag{3}$$

over a class of classifiers $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$. If complexity of $\mathcal{F}$ is properly constrained, the empirical risk minimizing (ERM) classifier $\hat{f}$ has statistically attractive properties including risk consistency and minimax rate optimality. See, e.g., Devroye et al. (1996) and Lugosi (2002).

Despite the desirable performance guarantee of the ERM classifer, computational complexity to solve optimization in (2) becomes a serious hurdle for practical implementation especially when the dimension of covariates is moderate to large. To get around this bottleneck, the literature has offered various alternatives to the ERM classifier, including support vector machines (Cortes and Vapnik (1995)), AdaBoost (Freund and Schapire (1997)), and neural networks. From the optimization point of view, each of these algorithms can be viewed as targeting to minimize a *surrogate risk*,

$$R_\phi(f) \equiv E_P[\phi(Yf(X))], \tag{4}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is called a *surrogate loss* function, a different specification of which corresponds to a different learning algorithm. A desirable choice of the surrogate loss is a convex function since combined with some functional form specification for f, the minimization problem for an empirical analogue of the surrogate risk (4) can becomes convex optimization. This is the major computational benefit that has been instrumental for these learning algorithms to handle large scale problems with high-dimensional features.

Can surrogate risk minimization lead to an optimal classifier in terms of the original classification risk? The seminal works of Zhang (2004) and Bartlett et al. (2006) provide theoretical justification for the uses of surrogate losses by clarifying the conditions that surrogate risk minimization also minimizes the original classification risk. A crucial assumption for this important result is *correct specification* of the classifiers, requiring that the class of classifiers $\mathcal{F}$ over which the surrogate risk is minimized contains a classifier that globally minimizes the original classification risk, i.e., a classifier that is identical to or performs as good as the Bayes classifier $f^*_{Bayes}(x) \equiv 2P(Y = 1|X = x) - 1$ in terms of the classification risk.

Credibility of the correct specification assumption is, however, limited if the set of implementable classifiers is constrained exogenously, independent of any belief concerning the underlying data generating process. Such situation is becoming more prevalent due to the increasing needs for interpretability and/or fairness of classification algorithms. Given that f determines the classification rule only through G_f, those constraints can be represented in terms of the shape restrictions on the prediction set of f, i.e., the class of feasible f's is represented by $\mathcal{F}_\mathcal{G} \equiv \{f \in \mathcal{F} : G_f \in \mathcal{G}\}$, where $\mathcal{G}$ is a restricted class of sets in $\mathcal{X}$ satisfying requirements for interpretability and/or fairness. To the best of our knowledge, it is not known how validity of the surrogate loss approaches can be affected if $\mathcal{F}_\mathcal{G}$ misses the first-best classifier.

The main contribution of this paper is to establish the conditions for the validity of surrogate loss approaches without assuming correct specification. We first characterize the condition for the surrogate loss such that minimization of the surrogate risk can lead

to a second-best rule (i.e., constrained optimum) in terms of the original classification risk. Specifically, we show that hinge losses $\phi_h(\alpha) = c\max\{0, 1-\alpha\}$, $c > 0$, are the only surrogate losses that guarantee consistency of the surrogate risk minimization to the second-best classifier. An important implication of this result is that ℓ_1-support vector machines are the only surrogate-loss based methods that are robust to misspecification.

Computational attractiveness of the surrogate loss approaches hinges not only on the convexity of the surrogate loss $\phi(\cdot)$ but also on functional form restrictions on the classifer f that lead to convex $\mathcal{F}$. We hence investigate how additional constraints on f on top of $G_f \in \mathcal{G}$ can affect the consistency of the hinge risk minimization. As the second contribution of this paper, we characterize a simple-to-check sufficient condition for consistency of the hinge risk minimization in terms of the additional functional form restrictions we can impose on $\mathcal{F}_\mathcal{G}$. We call a subclass of classifiers of $\mathcal{F}_\mathcal{G}$ satisfying the sufficient condition as a *classification-preserving reduction* of $\mathcal{F}_\mathcal{G}$.

Exploiting our main theoretical results, we develop novel procedures for monotone classification. In monotone classification, the prediction sets are constrained to

$$\mathcal{G}_M \equiv \{G \subset \mathcal{X} : x \in G \Rightarrow x' \in G \ \forall x' \leq x\},$$

where $x' \leq x$ is element-wise weak inequality. Since $\mathcal{G}_M$ agrees with the class of prediction sets spanned by the class of monotonically decreasing bounded functions $\mathcal{F}_M \equiv \{f : f \text{ decreasing in } x,\ -1 \leq f \leq 1\}$, hinge-loss based estimation for monotone classification can be performed by solving

$$\begin{aligned} &\hat{f}_M \in \arg\inf_{f\in\mathcal{F}_M} \widehat{R}_{\phi_h}(f), \qquad (5)\\ &\widehat{R}_{\phi_h}(f) \equiv \frac{1}{n}\sum_{i=1}^{n} \phi_h(y_i f(x_i)). \end{aligned}$$

We show that the class of monotone classifiers $\mathcal{F}_M$ is a constrained-classification-preserving reduction of $\mathcal{F}_{\mathcal{G}_M}$, guaranteeing consistency of the hinge-risk minimizing classifier $\hat{f}_M$. Furthermore, we show that convexity of $\mathcal{F}_M$ reduces optimization of (5) to finite dimensional linear programming and hence delivers significant computational gains relative to minimization of the original empirical classification risk. We also consider approximating $\mathcal{F}_M$ by a sieve of Bernstein polynomials and estimating a monotone classifier by solving (5) over the Bernstein polynomials. In either approach, an application of our main theorems guarantees

$$R(\hat{f}_M) - \inf_{f\in\mathcal{F}_M} R(f) \to_p 0,$$

as $n \to 0$, and this convergence is valid no matter whether $\mathcal{F}_M$ attains first-best, i.e., $\inf_{f\in\mathcal{F}_M} R(f) = \inf_{f\in\bar{\mathcal{F}}} R(f)$, or not, where $\bar{\mathcal{F}}$ is the class of measurable functions f :

$\mathcal{X} \to \mathbb{R}$. We also derive the uniform upper bound of the mean of $R(\hat{f}_M) - \inf_{f \in \mathcal{F}_M} R(f)$ to characterize the regret convergence rate attained by $\hat{f}_M$.

1.1 Connection and contributions to causal policy learning

For simplicity of exposition, this paper mainly focuses on the prototypical setting of binary classification. The main theoretical results can be extended straightforwardly to weighted (cost-sensitive) classification, where the canonical representation of the population risk criterion is given by

$$R^w(f) \equiv E_P[\omega \cdot 1\{Y \cdot \text{sign}(f(X)) \leq 0\}]. \tag{6}$$

Here, $\omega \in \mathbb{R}$ is a random variable defining the cost of misclassifying Y that typically depends on (Y, X). The cost of misclassification ω may represent the decision maker's economic cost (Lieli and White (2010)) or welfare weights over the individuals to be classified as considered in Rambachan et al. (2020) and Babii et al. (2020). Similarly to (4), the surrogate risk for weighted classification can be defined similarly,

$$R^w_\phi(f) = E_P[\omega \cdot \phi(Y f(X))]. \tag{7}$$

As discussed in Kitagawa and Tetenov (2018), the prediction problem of classification and the causal problem of treatment choice have fundamental conceptual differences. Nevertheless, if the training sample is obtained from a randomized control trial (RCT) study or an observational study satisfying unconfoundedness (selection on observables), we can view minimization of the weighted classification risk (6) as being equivalent to maximizing the additive welfare criterion commonly specified in the treatment choice problems. To see this equivalence, let $\{(Z_i, D_i, X_i) : i = 1, \ldots, n\}$ be an iid RCT sample of n experimental subjects, where $Z_i \in \mathbb{R}$ is subject i's observed outcome, $D_i \in \{-1, +1\}$ is an indicator for his assigned treatment, and $X_i \in \mathcal{X}$ is a vector of pretreatment covariates, and $(Z_i(d) : d \in \{-1, +1\})$ be i's potential outcomes satisfying $Z_i = Z_i(+1) \cdot 1\{D_i = +1\} + Z_i(-1) \cdot 1\{D_i = -1\}$. We denote the propensity score in the RCT sample by $e(x) \equiv P(D = +1|X = x)$ and assume that $e(x)$ is bounded away from 0 and 1 for all $x \in \mathcal{X}$. We denote the joint distribution of $(Z_i(+1), Z_i(-1), D_i, X_i)$ by P and assume P satisfies unconfoundedness, $(Z(+1), Z(-1)) \perp D|X$.

Similarly to classification, we represent (non-randomized) treatment assignment rule by the sign of $f : \mathcal{X} \to \mathbb{R}$, i.e., the 0-level set $G_f = \{x \in \mathcal{X} : f(x) \geq 0\} \subset \mathcal{X}$ specifies the subgroup of population assigned to treatment $+1$. Following Manski (2004), consider evaluating the welfare performance of assignment policy f by the average outcomes

attained by the assignment policy:

$$W(f) \equiv E_P\left[Z(+1) \cdot 1\{X \in G_f\} + Z(-1) \cdot 1\{X \notin G_f\}\right]$$

Relying on unconfoundedness of the experimental data and employing the inverse propensity score weighting technique, we can express this welfare in terms of observable variables:

$$\begin{aligned} W(f) &= E\left[\frac{Z}{De(X) + (1-D)/2} \cdot 1\{D = \text{sign}(f(X))\}\right] \\ &= E_P\left[\omega_p\right] - E_P\left[\omega_p \cdot 1\{D \cdot \text{sign}(f(X)) \leq 0\}\right], \quad \text{where} \\ \omega_p &= \frac{Z}{De(X) + (1-D)/2}. \end{aligned}$$

Provided that ω_p has the finite moment, maximizing $W(f)$ is therefore equivalent to minimizing the weighted classification risk $R^{\omega}(f)$ defined in (6) with $\omega = \omega_p$. As a result, optimal treatment assignment rules can be viewed as optimal classifiers for D in terms of the weighted classification risk. This equivalence implication holds also for other methods of policy learning, such as offset-tree learning of Beygelzimer and Langford (2009) and the doubly robust approaches of Swaminathan and Joachims (2015) and Athey and Wager (2021), as they correspond to different ways to construct or estimate the weighting term ω_p.

By such equivalence to weighted classification, the surrogate loss approach to policy learning proceeds by minimizing the empirical analogue of (7) with $\omega = \omega_p$. Section 7 of this paper shows that our main theoretical results established for constrained binary classification carry over to the setting of policy learning where the feasible treatment assignment policies are constrained exogenously due to fairness and legislative considerations. This paper therefore offers valuable and novel contributions to the current research and public debates on how to make use of machine learning algorithms for designing individualized policies. If the treatment assignment rules are constrained to monotone ones, our concrete proposals of monotone classification algorithms can be applied to policy learning and we gain significantly in terms of computational efficiency relative to the mixed integer programming approaches considered in Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2021).

1.2 Related literature

This paper is closely related to the literature of consistency and performance guarantees for the surrogate risk minimization. It includes Mannor et al. (2003), Jiang (2004), Lugosi and Vayatis (2004) , Zhang (2004), Steinwart (2005, 2007), Bartlett et al. (2006), Nguyen et al. (2009), Scott (2012). Assuming the correct specification, Zhang (2004) and Bartlett

et al. (2006) derive quantitative relationships between excess classification risk and excess surrogate risk, and then provide general conditions for surrogate risk minimization to achieve the risk consistency. Bartlett et al. (2006) show that the classification-calibration property of the surrogate loss, defined in Section 3 below, guarantees the risk consistency. Zhang (2004) and Bartlett et al. (2006) show that many commonly used surrogate loss functions, including the hinge loss, exponential loss, and truncated quadratic loss, satisfy his/their conditions. In a classification problem different from ours, where a pair of a quantizer and classifier is chosen, Nguyen et al. (2009) study sufficient and necessary conditions for surrogate risk minimization to yield risk consistency. They show that only the hinge loss functions satisfy the conditions for the consistency in their problem. The correct specification of the class of classifiers is an essential condition for consistency in all of the surrogate risk minimization approaches studied in the literature. The key contribution of our paper is to relax the correct specification assumption and clarifies conditions for the surrogate loss function to yield a consistent surrogate risk minimization procedure.

Relaxing the correct specification connects this paper to classification problems with exogenous constraints. Such problems are studied in several literatures of machine learning and statistics, such as interpretable classification (e.g., Zeng et al. (2017); Zhang et al. (2018)), fair classification (e.g., Dwork et al. (2012)), and monotone classification (e.g., Cano et al. (2019)). Some works in these literatures apply the surrogate loss approach. Donini et al. (2018) use ℓ_1-support vector machine in fair classification where the hinge risk minimization is subject to a statistical fairness constraint. Chen and Li (2014) apply ℓ_1-support vector machine with a monotonicity constraint, which constrains the class of feasible classifiers to a class of certain monotone functions. However, neither paper shows the consistency of their hinge risk minimization procedures.

From the optimization point of view, the empirical risk minimizing classification and the maximum score estimation (Manski (1975), Manski and Thompson (1989)) share the same objective function. Horowitz (1992) proposes smooth maximum score estimation where kernel smoothing is performed for the 1-0 loss to obtain a differentiable objective function. However, the smoothed objective function remains non-convex and does not offer computational gains that the surrogate risk minimization approach with convex surrogates could deliver.

This paper also contributes to a growing literature on statistical treatment rule in econometrics, including Manski (2004), Dehejia (2005), Hirano and Porter (2009), Stoye (2009, 2012), Chamberlain (2011), Bhattacharya and Dupas (2012), Tetenov (2012), Kasy (2018), Kitagawa and Tetenov (2018, 2021), Viviano (2019), Athey and Wager (2021), and Mbakop and Tabord-Meehan (2021), among others. As discussed above, the policy learning methods of Kitagawa and Tetenov (2018), Athey and Wager (2021), and Mbakop and

Tabord-Meehan (2021) build on similarity between empirical welfare maximizing treatment choice and empirical risk minimizing classification. Mbakop and Tabord-Meehan (2021) propose penalization methods to control complexity of treatment choice model, and derive relevant finite sample upper bounds for regret of the estimated treatment rules. Athey and Wager (2021) apply doubly-robust estimators to estimate the weight ω in (7), and show that $1/\sqrt{n}$-upper bound of the regret can be achieved also in the observational study setting. These works optimize the empirical welfare objective involving the indicator loss function. Hence, the computation of their methods are sometimes discouraging, especially when the sample size or number of the covariates is moderate to large.

Estimation of individualized treatment rules is a topic of active research in other fields including medical statistics, machine learning, and computer science; Zadrozny (2003), Beygelzimer and Langford (2009), Qian and Murphy (2011), Zhao et al. (2012), Swaminathan and Joachims (2015), Zhao et al. (2015), and Kallus (2020), among others. Zhao et al. (2012) propose to use ℓ_1-support vector machine to solve the weighted classification for individualized treatment choice problem, and show the risk consistency. They use a rich class of treatment choice models expressed by reproducing kernel Hilbert space, which certainly satisfies the correct specification. Zhao et al. (2015) extend this approach to estimate optimal dynamic treatment regimes.

2 Constrained classification with surrogate loss

Consider a binary classification problem for binary label $Y \in \{-1, +1\}$ based on covariate $X \in \mathcal{X}$, which have the joint distribution P. We let X be d_x-dimensional vector, $d_x < \infty$, and denote its marginal distribution by P_X. Let $\eta(x) \equiv P(Y = +1|X = x)$ denote the conditional probability of $Y = +1$ given $X = x$. We maintain the notations introduced in Introduction and set the ultimate objective to minimizing the classification risk of (1).

We study constrained classification problems where an optimal classifier is searched over a restricted class of functions. Section 2.1 first studies the consistency of surrogate risk minimization in a special case that the pre-specified class of classifiers contains a classifier whose prediction set agrees with the prediction set of the Bayes classifier. Section 2.2 introduces a classification problem with constraint on the prediction sets, which is a central problem throughout the paper.

2.1 Misspecification in constrained classification

Let $\mathcal{F}$ be a constrained class of classifiers $f : \mathcal{X} \to \mathbb{R}$. If the set of classifiers were unconstrained, it is well known that the Bayes classifier defined by

$$f^*_{Bayes} = 2\eta(x) - 1$$

minimizes the classification risk. Due to the constraints on the class of classifiers, however, the minimized classification risk on $\mathcal{F}$ can be strictly larger than the first-best minimal risk $R(f^*_{Bayes})$. We refer to this situation as R-misspecification of $\mathcal{F}$ as we state formally in the next definition.

Definition 2.1 (R-misspecification)**.** *$\mathcal{F}$ is R-misspecified if*

$$\inf_{f\in\mathcal{F}} R(f) > R(f^*_{Bayes}).$$

If the equality holds instead of the strict inequality, we say that $\mathcal{F}$ is R-correctly specified.

Because the 0-1 loss function is neither convex nor continuous, minimizing the empirical analog of $R(f)$ is computationally challenging and often infeasible in practical scale problems. Commonly used classification algorithms, such as boosting and support vector machines, alter the 0-1 loss with a surrogate loss function, $\phi : \mathbb{R} \to \mathbb{R}$, and aim to minimize the surrogate risk $R_\phi(f) \equiv E_P[\phi(Yf(X))]$. Table 1 below lists some commonly used surrogate loss functions including the *hinge loss* $\phi_h(\alpha) = \max\{0, 1-\alpha\}$, which corresponds to ℓ_1-support vector machines, and the *exponential loss* $\phi_e(\alpha) = \exp(-\alpha)$, which corresponds to AdaBoost.

We also introduce the concept of misspecification of $\mathcal{F}$ in terms of the surrogate risk as follows:

Definition 2.2 (R_ϕ-misspecification)**.** *Let $f^*_{\phi,FB}$ be a minimizer of R_ϕ over the unconstrained class of classifiers, i.e., the class of all measurable functions $f : \mathcal{X} \to \mathbb{R}$. A constrained class $\mathcal{F}$ is R_ϕ-misspecified if*

$$\inf_{f\in\mathcal{F}} R_\phi(f) > R_\phi(f^*_{\phi,FB}).$$

If the equality holds instead of the strict inequality, we say that $\mathcal{F}$ is R_ϕ-correctly specified.

The seminal theoretical results that guarantee consistency of surrogate-risk classification (Zhang (2004); Bartlett et al. (2006); Nguyen et al. (2009)) crucially rely on the

assumption that $\mathcal{F}$ is both R-correctly specified and R_ϕ-correctly specified in the sense of Definitions 2.1 and 2.2, respectively. The central question that this paper analyzes is what happens to the surrogate loss approaches if $\mathcal{F}$ is R-misspecified or R_ϕ-misspecified. This misspecification is a quite likely scenario especially when the origins of the constraints have nothing to do with assumptions on P, as is the case in the examples shown at the next subsection.

Throughout the paper, we limit our analysis to the class of classification-calibrated loss functions defined in Bartlett et al. (2006).

Definition 2.3 (Classification-calibrated loss functions)**.** *For $a \in \mathbb{R}$ and $0 \leq b \leq 1$, define $C_\phi(a,b) \equiv \phi(a)b + \phi(-a)(1-b)$. A loss function ϕ is classification-calibrated if for any $b \in [0,1]$,*

$$\inf_{\{a \in \mathbb{R} | a(2b-1)<0\}} C_\phi(a,b) > \inf_{\{a \in \mathbb{R} | a(2b-1) \geq 0\}} C_\phi(a,b)$$

Noting that the surrogate risk can be expressed as

$$E_P[\phi(Yf(X))] = E_{P_X}[C_\phi(f(X), \eta(X))], \tag{8}$$

the definition of classification-calibrated loss functions implies that at every $x \in \mathcal{X}$, a minimizer of $C_\phi(f(x), \eta(x))$ in $f(x)$ has the same sign as the Bayes classifier, $\text{sign}(2\eta(x) - 1)$). Bartlett et al. (2006) shows that many commonly used surrogate loss functions including those listed in Table 1 are classification-calibrated.[1]

Having introduced the concepts of misspecification, we first clarify the relationship between R-misspecification and R_ϕ-misspecification.

Proposition 2.1. *Let $\mathcal{F}$ be a constrained class of classifiers and $f_\phi^* \in \mathcal{F}$ be a minimizer of R_ϕ over $\mathcal{F}$. Suppose ϕ is a classification-calibrated loss function.*
(i) For any distribution P on $\{-1,1\} \times \mathcal{X}$, if $\mathcal{F}$ is R_ϕ-correctly specified, then $\mathcal{F}$ is R-correctly specified and $R(f_\phi^*) = R(f_{Bayes}^*)$ holds;
(ii) If ϕ is in addition convex, there exist a distribution P on $\{-1,1\} \times \mathcal{X}$ and a class of classifiers $\mathcal{F}$ under which $\mathcal{F}$ is R-correctly specified but R_ϕ-misspecified, and $R(f_\phi^*) > R(f_{Bayes}^*)$ holds.

Proof. (i) follows from claim 3 of Theorem 1 in Bartlett et al. (2006). To prove (ii), let $z_1, z_2 \in \mathbb{R}_+$ be such that $\phi(z_2) < \phi(z_1)$. Such a pair of (z_1, z_2) exists in a neighborhood

[1] Bartlett et al. (2006) also show that any convex loss function ϕ is classification-calibrated if and only if it is differentiable at 0 and $\phi'(0) < 0$.

of zero because from Theorem 2 of Bartlett et al. (2006), $\phi(\cdot)$ is differentiable at 0 and $\phi'(0) < 0$ if ϕ is classification-calibrated and convex. Then the pair (z_1, z_2) satisfies

$$\underset{(b_1,b_2)\in\{(-1,0),(0,1)\}}{\arg\min} \phi(-b_1 z_1) + \phi(b_2 z_2) = (0, 1).$$

Suppose that $\mathcal{F}$ is a constrained class such that $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ with $\mathcal{F}_1 = \{f(x) = x^T b : (b_1, b_2) = (-1, 0),\ b \in \mathbb{R}^{d_x}\}$ and $\mathcal{F}_2 = \{f(x) = x^T b : (b_1, b_2) = (0, 1),\ b \in \mathbb{R}^{d_x}\}$, where b_j denotes the j-th element of b. Let $x_1 = (z_1, 0, \ldots, 0)$ and $x_2 = (0, z_2, 0, \ldots, 0)$ be two points in $\mathcal{X}$, and let P be a distribution such that $\eta(x_1) = 0$, $\eta(x_2) = 1$, and $P_X(x_1) = P_X(x_2) = 1/2$. Under this pair of $(P, \mathcal{F})$, any classifier f_1 in $\mathcal{F}_1$ has the same sign as the Bayes classifier, P_X-almost everywhere, because $f_1(x_1) < 0$ and $f_1(x_2) = 0$ while $\eta(x_1) < 1/2$ and $\eta(x_2) \geq 1/2$. This means that $\mathcal{F}$, as well as $\mathcal{F}_1$, is R-correctly specified for such P. On the other hand, any classifier f_2 in $\mathcal{F}_2$ does not have the same sign as the Bayes classifier at x_1 because $f_2(x_1) = 0$ while $\eta(x_1) < 1/2$. f^*_ϕ must be in $\mathcal{F}_2$ because for any $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$, $R_\phi(f_2) = \phi(z_2)/2 < \phi(z_1)/2 = R_\phi(f_1)$. Hence $R(f^*_\phi) > R(f^*_{bayes})$ holds. $\mathcal{F}$ is also R_ϕ-misspecified because any classifier that minimizes R_ϕ over all measurable functions takes a negative value at x_1 whereas $f^*_\phi(x_1) > 0$. □

Proposition 2.1 (i), which rephrases claim 3 of Theorem 1 in Bartlett et al. (2006), implies that the surrogate risk minimization on R_ϕ-correctly specified class $\mathcal{F}$ leads to the (first-best) optimal classification in terms of the classification risk. An equivalent statement following Theorem 1 in Bartlett et al. (2006) is that for any P and every sequence of measurable functions $\{f_i : \mathcal{X} \to \mathbb{R}\}$,

$$R_\phi(f_i) \to \inf_{f\in\mathcal{F}} R_\phi(f) \text{ implies that } R(f_i) \to \inf_{f\in\mathcal{F}} R(f). \tag{9}$$

This result justifies the approach of surrogate risk minimization when we can introduce a sufficiently rich class of classifiers $\mathcal{F}$ (e.g., the reproducing kernel Hilbert space of functions with a large number of features as used in support vector machines), as R_ϕ-correct specification which is a credible assumption to make with the rich class of classifiers guarantees R-correct specification.

Proposition 2.1 (ii), in contrast, shows that R-correct specification of $\mathcal{F}$ does *not* guarantee R_ϕ-correct specification.[2] Since R_ϕ-misspecification of $\mathcal{F}$ can lead to a suboptimal

[2]Given a convex classification-calibrated loss function ϕ, our proof of Proposition 2.1 (ii) constructs a pair of a R-correctly specified class of classifiers $\mathcal{F}$ and distribution P that leads to R_ϕ-misspecification. In the construction, we assume $x_1 \neq x_2 \in \mathcal{X}$ supported by P_X on which $\phi(f(x_1)) < \phi(-f(x_2)))$ holds for all $f \in \mathcal{F}$ and $f(x_1) < 0 \leq f(x_2)$ holds for some $f \in \mathcal{F}$, and consider P that assigns a large positive value of $\eta(x_2)$ in $(0, 1]$ and slightly negative value of $\eta(x_1)$. Such construction of P is not pathological or limited to the specific class of classifiers considered in the proof.

classifier in $\mathcal{F}$ in terms of the classification risk, this result shows the cost of the surrogate loss approach with constrained classifiers. Even when we are confident that the constrained class $\mathcal{F}$ is R-correctly specified, we cannot justify a use of $\mathcal{F}$ in the surrogate risk minimization.

2.2 $\mathcal{G}$-constrained classification

In this section, we consider restricting the class of classifiers by requiring that their prediction sets belong to a prespecified class of sets, $\mathcal{G} \subset 2^{\mathcal{X}}$. See Examples 2.4–2.6 below for motivating examples.

We denote by

$$\mathcal{F}_{\mathcal{G}} \equiv \{f : G_f \in \mathcal{G},\ f(\cdot) \in [-1, 1]\}$$

the class of classifiers whose prediction sets are constrained to $\mathcal{G}$. In this definition, we restrict f to being bounded and, without loss of generality, normalize its range to $[-1, 1]$. Other than the shape of 0-level set and range, $\mathcal{F}_{\mathcal{G}}$ does not impose any constraints on the functional form of $f \in \mathcal{F}_{\mathcal{G}}$. The goal of the constrained classification problem is then to find a best classifier that minimizes the classification risk $R(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$. We refer to $\mathcal{F}_{\mathcal{G}}$ as *$\mathcal{G}$-constrained class of classifiers* and to the classification problem over $\mathcal{F}_{\mathcal{G}}$ as *$\mathcal{G}$-constrained classification*.

The specification of the class of prediction sets $\mathcal{G}$ represents interpretability, fairness, and other exogenous requirements desired for classification rules. Some examples follow.

Example 2.4 (Interpretbale classification)**.** *Decision makers may prefer simple decision or classification rules that are easily understandable or explainable even at the cost of harming prediction accuracy. This concept, often refereed to as interpretable machine learning, has been pursued, for instance, in the prediction analysis of recidivism (Zeng et al. (2017)) and the decision on medical intervention protocol (Zhang et al. (2018))). An example is a linear classification rule, in which $\mathcal{G}$ is a class of half-spaces with linear boundaries in $\mathcal{X}$,*

$$\mathcal{G} = \{x \in \mathbb{R}^{d_x} : x^T\beta \geq 0, \beta \in \mathbb{R}^{d_x}\}.$$

Note that $f \in \mathcal{F}_{\mathcal{G}}$ is not restricted to a liner function. Any function f including nonlinear ones is included in $\mathcal{F}_{\mathcal{G}}$ as long as its prediction set G_f is a hyperplane in $\mathcal{X}$. Another type of classifier that has merit in terms of interpretability is classification tree. See, e.g., Breiman et al. (1984).

Example 2.5 (Monotone classification)**.** *The framework we study can accommodate monotonicity constraints on classification. Formally, the monotonicity constraint corresponds*

to a partial order $\precsim$ on $\mathcal{X}$, and any prediction set G_f has to respect this partial order in the sense that if $x_1 \precsim x_2$ and $x_1 \in G_f$, then $x_2 \in G_f$. The monotonicity constraint have been utilized in classification of credit rating (Chen and Li (2014)) and assignment of job training program in the context of policy learning (Mbakop and Tabord-Meehan (2021)).

Example 2.6 (Fair classification)**.** *Specification of $\mathcal{G}$ can accommodate some fairness constraints introduced in the literature of fair classification. Let $A = \{0, 1\}$ be an element of X indicating a binary protected group variable (e.g., race, gender). The decision maker wants to ensure fairness of classification, for instance, by equalizing raw positive classification rate (known as statistical parity): $P_X(f(X) \geq 0 \mid A = 1) = P_X(f(X) \geq 0 \mid A = 0)$. The classification problem under this constraint is equivalent to the $\mathcal{G}$-constrained classification with*

$$\mathcal{G} = \left\{G \in 2^{\mathcal{X}} : P_X(X \in G \mid A = 1) = P_X(X \in G \mid A = 0)\right\},$$

where $\mathcal{G}$ depends on P_X in this case. This fairness constraint is studied by Calders and Verwer (2010), Kamishima et al. (2011), Dwork et al. (2012), Feldman et al. (2015), among others. Some other forms of fairness constraint, such as equalized odds and equalized positive predictive value as reviewed by Chouldechova and Roth (2018), can be accommodated in our framework as well through a proper construction of $\mathcal{G}$.

In the $\mathcal{G}$-constrained classification problem, R-correct specification of $\mathcal{F}_{\mathcal{G}}$ is necessary and sufficient for the surrogate risk minimizer f_ϕ^* to achieve the first-best minimum risk.

Proposition 2.2. *Suppose ϕ is a classification-calibrated loss function. Let $\mathcal{G} \subseteq 2^{\mathcal{X}}$ be an arbitrary class of prediction sets and $f_\phi^* \in \mathcal{F}_{\mathcal{G}}$ be a minimizer of R_ϕ over $\mathcal{F}_{\mathcal{G}}$. Then, for any distribution P on $\{-1, 1\} \times \mathcal{X}$, $R(f_\phi^*) = R(f_{Bayes}^*)$ holds if and only if $\mathcal{F}_{\mathcal{G}}$ is R-correctly specified.*

Proof. Assume R-correct specification of $\mathcal{F}_{\mathcal{G}}$. Then, $\mathcal{F}_{\mathcal{G}}$ includes a classifier f^* that is identical to or shares the sign with $f_{Bayes}^*(x) = 2\eta(x) - 1$, P_X-almost everywhere. Since $f \in \mathcal{F}_{\mathcal{G}}$ is unconstrained except for $G_f \in \mathcal{G}$ and $-1 \leq f(\cdot) \leq 1$, the classification-calibrated property of ϕ and the representation of the surrogate risk $R_\phi(f) = E_{P_X}[C_\phi(f(X), \eta(X))]$ implies

$$f_\phi^*(x) \in \underset{a:(2\eta(x)-1)a \geq 0, |a| \leq 1}{\arg\min} C_\phi(a, \eta(x)),$$

P_X-almost everywhere, as otherwise f^* dominates f_ϕ^* in terms of the surrogate risk. This means that $f_\phi^*(x)$ has the same sign as $f_{Bayes}^*(x)$, P_X- almost everywhere, i.e., $R(f_\phi^*) = R(f_{Bayes}^*)$ holds.

Assume conversely that $\mathcal{F}_\mathcal{G}$ is R-misspecified. Then, $\text{sign}(f^*_\phi)$ has to differ from $\text{sign}(f^*_{Bayes}(x))$ for some x with a positive measure in terms of P_X, as otherwise it contradicts with R-misspecification of $\mathcal{F}_\mathcal{G}$. This then implies $R(f^*_\phi) > R(f^*_{Bayes})$. □

Proposition 2.2 shows that if ϕ is classification-calibrated, $f^*_\phi \in \mathcal{F}_\mathcal{G}$ that minimizes the surrogate risk over $\mathcal{F}_\mathcal{G}$ leads to a globally optimal classifier in terms of the classification risk if and only if $\mathcal{F}_\mathcal{G}$ is R-correctly specified. A comparison of Proposition 2.1 (ii) and Proposition 2.2 clarifies a special feature of the $\mathcal{G}$-constrained class of classifiers, i.e., Proposition 2.1 (ii) has shown that in general R-correct specification of a constrained class of classifiers $\mathcal{F}$ does not guarantee $R(f^*_\phi) = R(f^*_{Bayes})$. In contrast to the seminal results about surrogate risk consistency shown in Zhang (2004) and Bartlett et al. (2006), our claim does not require R_ϕ-correct specification of $\mathcal{F}_\mathcal{G}$.

If constraints defining $\mathcal{G}$ are motivated by some considerations that are independent of any belief on the underlying data generating process (e.g., Examples 2.4–2.6 above), the R-correct specification of $\mathcal{F}_\mathcal{G}$ is hard to justify. Therefore, an important question for our analysis to follow is whether or not the surrogate risk minimization procedures can yield a classifier achieving $\inf_{f\in\mathcal{F}_\mathcal{G}} R(f)$ *without* requiring R-correct specification of $\mathcal{F}_\mathcal{G}$.

3 Calibration of $\mathcal{G}$-constrained classification

This section investigates consistency of the surrogate risk minimization approach over $\mathcal{F}_\mathcal{G}$, where $\mathcal{F}_\mathcal{G}$ is now allowed to be R-misspecified. Let f^* be an optimal classifier that minimizes the classification risk over $\mathcal{F}_\mathcal{G}$:

$$f^* \in \arg\inf_{f\in\mathcal{F}_\mathcal{G}} R(f).$$

Similarly, we denote a best classifier among $\mathcal{F}_\mathcal{G}$ in terms of the surrogate risk by f^*_ϕ,

$$f^*_\phi \in \arg\inf_{f\in\mathcal{F}_\mathcal{G}} R_\phi(f),.$$

To begin our analysis, let us first perform a simple numerical example to assess influence of misspecification in constrained classification.

Example 3.1 (Numerical example 1)**.** *Let* $\mathcal{X} = \{0, 1, 2\}$ *and* $\mathcal{G} = \{\emptyset, \{2\}, \{2, 1\}, \{2, 1, 0\}\}$. *Here,* $\mathcal{G}$ *imposes monotonicity of the prediction sets in a way consistent to Example 2.5. We specify* P_X *to be uniform on* $\mathcal{X}$ *and* $P(Y = +1 \mid X = 0) = 0.9$, $P(Y = +1 \mid X = 1) = 0.3$, *and* $P(Y = +1 \mid X = 2) = 0.2$. *The Bayes classifier therefore predicts* $Y = +1$

at $x = 0$ and $Y = -1$ at $x = 1$ and 2, but such prediction set is excluded from $\mathcal{G}$. That is, $\mathcal{F}_{\mathcal{G}}$ is R-misspecified.

Under this specification, we compute the second-best (constrained optimum) classifier f^* and the attained classification risk $R(f^*)$. Also, for each of hinge loss ϕ_h, exponential loss ϕ_e, and truncated quadratic loss ϕ_q, we compute the classifier minimizing the surrogate risk f^*_ϕ and the classification risk at the surrogate optimal classifier $R(f^*_\phi)$. We obtain

$$R(f^*) = R(f^*_{\phi_h}) = 0.47, \quad R(f^*_{\phi_e}) = R(f^*_{\phi_q}) = 0.53,$$
$$G_{f^*} = G_{f^*_{\phi_h}} = \emptyset, \quad G_{f^*_{\phi_e}} = G_{f^*_{\phi_q}} = \{2, 1, 0\}.$$

In this specification, the hinge risk optimal classifier agrees with the second best optimal classifier, whereas that is not the case for the exponential or truncated quadratic loss.

This example illustrates that the hinge loss is robust to R-misspecification of $\mathcal{F}_{\mathcal{G}}$, but the exponential and truncated quadratic losses are not. To what extent, can we generalize this finding? What condition do we need for surrogate loss to guarantee consistency to the second best? We answer these questions below.

At any classifier f, we define the *$\mathcal{G}$-constrained excess risk* of f as

$$R(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f),$$

which is the *regret* of f relative to a constrained optimum f^* in terms of the classification risk. Similarly, we define the $\mathcal{G}$-constrained excess ϕ-risk of f as

$$R_\phi(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_\phi(f).$$

Fix $G \in \mathcal{G}$ and let

$$\mathcal{F}_G \equiv \{f : G_f = G, f(\cdot) \in [-1, 1]\}$$

be the class of classifiers that share the prediction set G. Then $\{\mathcal{F}_G : G \in \mathcal{G}\}$ form a partition of $\mathcal{F}_{\mathcal{G}}$ according to the prediction set, and they satisfy $\mathcal{F}_{\mathcal{G}} = \cup_{G \in \mathcal{G}} \mathcal{F}_G$ and $\mathcal{F}_G \cap \mathcal{F}_{G'} = \emptyset$ for $G, G' \in \mathcal{G}$ with $G \neq G'$. With this definition, choosing a classifier from $\mathcal{F}_{\mathcal{G}}$ can be decomposed into two steps: choosing a prediction set G from $\mathcal{G}$ and, then, choosing a classifier f from $\mathcal{F}_G$.

Denote the classification risk evaluated at a prediction set G by $\mathcal{R}(G) \equiv \inf_{f \in \mathcal{F}_G} R(f)$. Note that any $f \in \mathcal{F}_G$ attains the same level of classification risk, so $\mathcal{R}(G) = R(f)$ holds

for all $f \in \mathcal{F}_G$. $\mathcal{R}(G)$ can be written as

$$\begin{aligned}\mathcal{R}(G) &= \int_{\mathcal{X}} [\eta(x)1\{x \notin G\} + (1-\eta(x))1\{x \in G\}]\, dP_X(x), \\ &= \int_{\mathcal{X}} (1-2\eta(x)) \cdot 1\{x \in G\} dP_X(x) + P(Y=1). \end{aligned} \tag{10}$$

Similarly, we define the surrogate risk evaluated at G by $\mathcal{R}_\phi(G) \equiv \inf_{f \in \mathcal{F}_G} R_\phi(f)$, which can be written as

$$\begin{aligned}\mathcal{R}_\phi(G) &= \inf_{f \in \mathcal{F}_G} \int_{\mathcal{X}} [\eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x))]\, dP_X(x) \\ &= \int_G \inf_{0 \le f(x) \le 1} C_\phi(f(x), \eta(x)) dP_X(x) + \int_{G^c} \inf_{-1 \le f(x) < 0} C_\phi(f(x), \eta(x)) dP_X(x), \end{aligned} \tag{11}$$

where the second line follows from the fact that $f \in \mathcal{F}_G$ is unconstrained other than its prediction set and the minimization over $f \in \mathcal{F}_G$ can be done pointwise at each x. For $x \in G$ and $f \in \mathcal{F}_G$, $f(x)$ is constrained to $[0,1]$ and, for $x \in G^c$, $f(x)$ is constrained to $[-1,0)$. To simplify the notation, we define

$$\begin{aligned} C_\phi^+(\eta(x)) &\equiv \inf_{0 \le f(x) \le 1} C_\phi(f(x), \eta(x)), \\ C_\phi^-(\eta(x)) &\equiv \inf_{-1 \le f(x) < 0} C_\phi(f(x), \eta(x)), \\ \Delta C_\phi(\eta(x)) &\equiv C_\phi^+(\eta(x)) - C_\phi^-(\eta(x)), \end{aligned}$$

where $C_\phi^+(\eta(x))$ and $C_\phi^-(\eta(x))$ are the minimized surrogate risk conditional on $X = x$ under the constraints $f(x) \in [0,1]$ or $f(x) \in [-1,0)$, respectively. Using these, the surrogate risk at G can be written as

$$\begin{aligned}\mathcal{R}_\phi(G) &= \int_{\mathcal{X}} \left[C_\phi^+(\eta(x)) \cdot 1\{x \in G\} + C_\phi^-(\eta(x)) \cdot 1\{x \notin G\}\right] dP_X(x) \\ &= \int_{\mathcal{X}} \Delta C_\phi(\eta(x)) \cdot 1\{x \in G\} dP_X(x) + \int_{\mathcal{X}} C_\phi^-(\eta(x)) dP_X(x). \end{aligned} \tag{12}$$

By comparing the expressions of the risks between (10) and (12), we obtain the first main theorem that clarifies the condition for the surrogate risk $\mathcal{R}_\phi(G)$ to calibrate the global ordering of the classification risk $\mathcal{R}(G)$ over $G \in \mathcal{G}$.

Theorem 3.2 (Global calibration of the $\mathcal{G}$-constrained excess risk)**.** *Let P be an arbitrary distribution on $\{-1,1\} \times \mathcal{X}$ and $\mathcal{G} \subseteq 2^{\mathcal{X}}$ be an arbitrary class of prediction sets. For $G, G' \in \mathcal{G}$, their risk ordering in terms of the classification risk, $\mathcal{R}(G) \ge \mathcal{R}(G')$, is*

equivalent to

$$\int_{G\setminus G'}(1-2\eta(x))dP_X(x) \geq \int_{G'\setminus G}(1-2\eta(x))dP_X(x). \tag{13}$$

Their risk ordering in terms of the surrogate risk, $\mathcal{R}_\phi(G) \geq \mathcal{R}_\phi(G')$, is equivalent to

$$\int_{G\setminus G'}\Delta C_\phi(\eta(x))dP_X(x) \geq \int_{G'\setminus G}\Delta C_\phi(\eta(x))dP_X(x). \tag{14}$$

Hence, if $\Delta C_\phi(\eta(x))$ is proportional to $1-2\eta(x)$ up to a positive constant, i.e.,

$$\Delta C_\phi(\eta(x)) = c(1-2\eta(x)) \text{ for some } c>0, \tag{15}$$

the risk ordering over $\mathcal{G}$ in terms of the surrogate risk $\mathcal{R}_\phi(G)$ agrees with the risk ordering over $\mathcal{G}$ in terms of the classification risk $\mathcal{R}(G)$.

In particular, when ϕ is a hinge loss, $\phi_h(\alpha) = c\max\{0, 1-\alpha\}$, $c>0$,

$$\Delta C_\phi(\eta(x)) = c(1-2\eta(x))$$

holds, so the hinge risk preserves the risk ordering of the classification risk.

Proof. By equation (10),

$$\begin{aligned}
&\mathcal{R}(G) - \mathcal{R}(G') \\
&= \int_{\mathcal{X}}(1-2\eta(x))\cdot[1\{x\in G\} - 1\{x\in G'\}]dP_X(x) \\
&= \int_{\mathcal{X}}(1-2\eta(x))\cdot[1\{x\in G\setminus G'\} - 1\{x\in G'\setminus G\}]dP_X(x) \\
&= \int_{G\setminus G'}(1-2\eta(x))dP_X(x) - \int_{G'\setminus G}(1-2\eta(x))dP_X(x).
\end{aligned}$$

This proves (13), the first claim of the theorem.

Given the representation of the surrogate risk shown in (12), a similar argument yields (14), the second claim of the theorem.

For the hinge loss $\phi_h(\alpha) = c\max\{0, 1-\alpha\}$ and $f\in\mathcal{F}_G$, we have

$$C_{\phi_h}(f(x),\eta(x)) = c(1-2\eta(x))f(x) + c.$$

Hence, we obtain

$$C^+_{\phi_h}(\eta) = \begin{cases} c(1-2\eta)+c & \text{for } \eta > 1/2, \\ c & \text{for } \eta \leq 1/2, \end{cases}$$

$$C^-_{\phi_h}(\eta) = \begin{cases} c & \text{for } \eta > 1/2, \\ 2c\eta & \text{for } \eta \le 1/2. \end{cases}$$

Hence, $\Delta C_{\phi_h}(\eta) = c(1-2\eta)$ holds for all $\eta \in [0,1]$. □

Theorem 3.2 does not exploit the condition that ϕ is classification-calibrated, but if a surrogate loss function satisfies condition (15), it is automatically classification-calibrated. Another remark follows.

Remark 3.3. *Many commonly used surrogate loss functions do not satisfy the condition (15) in Theorem 3.2. Table 1 shows the forms of $\Delta C_\phi(\eta)$ for the hinge loss, exponential loss, logistic loss, quadratic loss, and truncated quadratic loss functions. None of them except for the hinge loss satisfies condition (15). That is, among the surrogate-loss based algorithms commonly used in practice, the ℓ_1-support vector machine corresponding to the hinge loss is the only algorithm whose surrogate risk preserves the classification risk.*

Table 1: Surrogate loss functions and their forms of ΔC_ϕ

Loss function	$\phi(\alpha)$	$\Delta C_\phi(\eta)$
0-1 loss	$1\{\alpha \le 0\}$	$1-2\eta$
Hinge loss	$c\max\{0, 1-\alpha\}$	$c(1-2\eta)$
Exponential loss	$e^{-\alpha}$	$\begin{cases} -2\sqrt{\eta(1-\eta)}+1 & \text{if } 0 \le \eta < 1/2 \\ 2\sqrt{\eta(1-\eta)}-1 & \text{if } 1/2 \le \eta \le 1 \end{cases}$
Logistic loss	$\log(1+e^{-\alpha})$	$\begin{cases} \log(2\eta^\eta(1-\eta)^{1-\eta}) & \text{if } 0 \le \eta < 1/2 \\ -\log(2\eta^\eta(1-\eta)^{1-\eta}) & \text{if } 1/2 \le \eta \le 1 \end{cases}$
Quadratic loss	$(1-\alpha)^2$	$\begin{cases} (1-2\eta)^2 & \text{if } 0 \le \eta < 1/2 \\ -(1-2\eta)^2 & \text{if } 1/2 \le \eta \le 1 \end{cases}$
Truncated quadratic loss	$(\max\{0, 1-\alpha\})^2$	$\begin{cases} (1-2\eta)^2 & \text{if } 0 \le \eta < 1/2 \\ -(1-2\eta)^2 & \text{if } 1/2 \le \eta \le 1 \end{cases}$

The well known inequality by Zhang (2004) relates the excess surrogate risk to the excess classification risk under R-correct specification. As a corollary of Theorem 3.2, if we set $\phi = \phi_h$, we can generalize Zhang's inequality by allowing R-misspecification of the classifiers. To formally state it, let $G^* \in \arg\inf_{G\in\mathcal{G}} \mathcal{R}(G)$, and set $G' = G^*$ in Theorem 3.2. Let $f \in \mathcal{F}_\mathcal{G}$ be arbitrary and $G_f = \{x \in \mathcal{X} : f(x) \ge 0\} \in \mathcal{G}$. The aligned risk ordering between the classification and hinge risks implies that the minimizers of $\mathcal{R}(\cdot)$ also minimize $\mathcal{R}_{\phi_h}(\cdot)$, i.e., $\inf_{f\in\mathcal{F}_\mathcal{G}} R_{\phi_h}(f) = \inf_{G\in\mathcal{G}} \mathcal{R}_{\phi_h}(G) = \mathcal{R}_{\phi_h}(G^*)$. Theorem 3.2 therefore implies that the $\mathcal{G}$-constrained excess classification risk of f satisfies the

following inequality:

$$
\begin{aligned}
& R(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f) = \mathcal{R}(G_f) - \mathcal{R}(G^*) \\
&= \int_{G_f \setminus G^*} (1 - 2\eta(x)) dP_X(x) - \int_{G^* \setminus G_f} (1 - 2\eta(x)) dP_X(x) \\
&= c^{-1} \left[\mathcal{R}_{\phi_h}(G_f) - \mathcal{R}_{\phi_h}(G^*) \right] = c^{-1} \left[\inf_{f' \in \mathcal{F}_{G_f}} R_{\phi_h}(f') - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_{\phi_h}(f) \right] \\
&\leq c^{-1} \left[R_{\phi_h}(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_{\phi_h}(f) \right], \qquad (16)
\end{aligned}
$$

where the second equality follows by equation (10); the third equality follows by equation (12) and $\Delta C_{\phi_h}(\eta) = c(1 - 2\eta)$. That is, when $\phi = \phi_h$, Zhang's inequality holds without requiring the R-correct specification of the classifiers.

Corollary 3.4. *For any distribution P on $\{-1, 1\} \times \mathcal{X}$ and any constraint $\mathcal{G} \subseteq 2^{\mathcal{X}}$, if $\Delta C_\phi(\eta(x))$ is proportional to $1 - 2\eta(x)$ with a proportionality constant $c > 0$, i.e., $\Delta C_\phi(\eta(x)) = c(1 - 2\eta(x))$, then the following inequality holds*

$$
c(R(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f)) \leq R_\phi(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_\phi(f)
$$

for any $f \in \mathcal{F}_{\mathcal{G}}$.

Proof. See equation (16). □

Corollary 3.4 shows that if the surrogate loss ϕ satisfies condition (15), then the classifier f^*_ϕ that minimizes the surrogate risk over $\mathcal{F}_{\mathcal{G}}$ also minimizes the classification risk over $\mathcal{F}_{\mathcal{G}}$. Importantly, this result holds without assuming the R-correct specification of $\mathcal{F}_{\mathcal{G}}$. It justifies the use of hinge loss in the constrained classification problem irrespective of whether or not $\mathcal{F}_{\mathcal{G}}$ is correctly R-specified. Note, however, that the result relies on the fact that at every $x \in \mathcal{X}$ we can choose any $f(x) \in [-1, 1]$ as long as the prediction set constraint is satisfied: $G_f \in \mathcal{G}$. We relax this requirement in the next section.

Further analysis can show that the condition (15) in Theorem 3.2 is not only sufficient but also necessary. To formally show that, we adopt the concept of universal equivalence of loss functions introduced by Nguyen et al. (2009) to the current setting.

Definition 3.5 (Universal equivalence)**.** *Loss functions ϕ_1 and ϕ_2 are universally equivalent, denoted by $\phi_1 \overset{u}{\sim} \phi_2$, if for any distribution P on $\{-1, 1\} \times \mathcal{X}$ and any constraint $\mathcal{G} \subseteq 2^{\mathcal{X}}$,*

$$
\mathcal{R}_{\phi_1}(G_1) \leq \mathcal{R}_{\phi_1}(G_2) \Leftrightarrow \mathcal{R}_{\phi_2}(G_1) \leq \mathcal{R}_{\phi_2}(G_2)
$$

holds for any $G_1, G_2 \in \mathcal{G}$.

Universally equivalent loss functions ϕ_1 and ϕ_2 lead to the same risk ordering over $\mathcal{G}$. Hence, if a loss function ϕ is universally equivalent to the 0-1 loss, the ϕ-risk shares the same risk ordering with the classification risk.

The following theorem establishes a necessary and sufficient condition for two classification-calibrated loss functions to be universally equivalent.

Theorem 3.6. *Let* ϕ_1 *and* ϕ_2 *be classification-calibrated loss functions. Then,* $\phi_1 \overset{u}{\sim} \phi_2$ *if and only if* $\Delta C_{\phi_2}(\eta) = c\Delta C_{\phi_1}(\eta)$ *for some* $c > 0$ *and any* $\eta \in [0,1]$*, i.e.,* ΔC_{ϕ_1} *is proportional to* ΔC_{ϕ_2} *up to a positive constant.*

The proof is given in Appendix A. The "if" part of the theorem is a generalization of Theorem 3.2 without assuming either of ϕ_1 or ϕ_2 to be the 0-1 loss function.

When we set ϕ_2 to the 0-1 loss function, Theorem 3.6 yields the class of loss functions that are universally equivalent to the 0-1 loss functions. It exactly coincides with the class of loss functions that satisfy the condition (15) in Theorem 3.2. Hence, the following corollary holds.

Corollary 3.7. *A classification-calibrated loss function* ϕ *is universally equivalent to the 0-1 loss function if and only if* ϕ *satisfies the condition (15) for any* $\eta(x) \in [0,1]$*. That is, the class of hinge loss functions* $\{\phi(\alpha) = a\max\{0, 1-\alpha\} + b : a > 0, b \geq 0\}$ *agrees with the class of loss functions that are universally equivalent to the 0-1 loss function.*

In the following sections, without loss of generality, we use the definition of the hinge loss as $\phi_h(\alpha) = \max\{0, 1-\alpha\}$, or equivalently suppose $c = 1$. We conclude this section with a remark to compare our constrained classification framework to that of Nguyen et al. (2009).

Remark 3.8. *Nguyen et al. (2009) show that, in the classification problem of choosing an optimal pair of a quantizer and classifier, the hinge loss function is also only a surrogate loss function that preserves the consistency of surrogate loss classification. In their framework, the quantizer is a stochastic mapping* $Q \in \mathcal{Q} : \mathcal{X} \mapsto \mathcal{Z}$*, where* $\mathcal{Z}$ *is a discrete space and* $\mathcal{Q}$ *is a possibly constrained class of conditional distributions of* Z *given* X*,* $Q(Z \mid X)$*. The classifier is a function* $\gamma \in \Gamma : \mathcal{Z} \mapsto \mathbb{R}$*, where* Γ *is a set of all measurable functions on* $\mathcal{Z}$*. One motivation of using* Z *as inputs, instead of* X*, is to reduce the dimension of* X*, which might be a high-dimensional vector. They consider to estimate the optimal pair of*

$(Q, \gamma) \in \mathcal{Q} \times \Gamma$ *that minimizes the risk* $R(\gamma, Q) := P(Y \neq sign(\gamma(Z)))$*, through solving a surrogate loss classification problem:* $\inf_{(Q,\gamma)\in\mathcal{Q}\times\Gamma} R_\phi(Q, \gamma)$ *where* $R_\phi(Q, \gamma) = E\phi(Y\gamma(Z))$*. They show that, among the commonly used surrogate loss functions, the hinge loss classification only leads to the optimal pair of* (Q, γ)*.*

The framework we study is different from that of Nguyen et al. (2009), and none nests the other. The framework Nguyen et al. (2009) study constrains the mapping $Q : \mathcal{X} \mapsto \mathcal{Z}$*, whereas the framework we study constrains prediction sets* G_f *for all classifiers* f*. Furthermore, the class of classifiers* Γ *considered in Nguyen et al. (2009) contains the Bayes classifier, whereas the class of classifiers* $\mathcal{F}_\mathcal{G}$ *we consider may not contain the Bayes classifier.*

4 Consistency of hinge-risk classification with functional form constraints

The previous section considers $\mathcal{F}_\mathcal{G}$, the class of all functions whose prediction sets are in $\mathcal{G}$. The generalized Zhang's inequality shown in Corollary 3.4 heavily relies on such richness of $\mathcal{F}_\mathcal{G}$. This richness, however, limits computational attractiveness of the surrogate-loss approach, since convexity of optimization for an empirical analogue of the surrogate risk does not directly follow from $\mathcal{F}_\mathcal{G}$ and typically requires additional functional form restrictions for f.

Unfortunately, once a functional form restriction on f is imposed on top of the prediction set constraint $G_f \in \mathcal{G}$, the global calibration property of the hinge risk as shown in Theorem 3.2 breaks down. The following example illustrates this phenomenon.

Example 4.1 (Numerical example 2)**.** *Maintain* $\mathcal{X} = \{0, 1, 2\}$ *and* $\mathcal{G} = \{\emptyset, \{2\}, \{2, 1\}, \{2, 1, 0\}\}$ *as in Example 3.1. We here consider choosing a classifier from the following class of non-decreasing linear functions:*

$$\mathcal{F}_L = \{f(x) = c_0 + c_1 x : c_0 \in \mathbb{R},\ c_1 \in \mathbb{R}_+,\ f(x) \in [-1, 1] \text{ for all } x \in \mathcal{X}\}.$$

Note that the class of prediction sets $\{G_f : f \in \mathcal{F}_L\}$ *agrees with* $\mathcal{G}$*; hence,* $\mathcal{F}_L$ *is a subclass of* $\mathcal{F}_\mathcal{G}$*. We set* X *to be uniformly distributed on* $\mathcal{X}$ *and* Y *to have conditional probabilities* $P(Y = 1 \mid X = 0) = 0.3$*,* $P(Y = 1 \mid X = 1) = 0.9$*, and* $P(Y = 1 \mid X = 2) = 0.2$*.*

The Bayes classifier predicts positive Y *only at* $x = 1$*. Hence, no classifier in* $\mathcal{F}_L$ *shares the prediction set with the Bayes classifier, and* $\mathcal{F}_L$ *is R-misspecified.*

The optimal classification risk $R(f^*)$ *over* $\mathcal{F}_L$ *(equivalently, over* $\mathcal{F}_\mathcal{G}$*) is* $R(f^*) = 0.4$ *with* $G_{f^*} = \{2, 1\}$*, while the classification risk at* $f^*_{\phi_h}$ *minimizing the hinge risk over* $\mathcal{F}_L$ *is* $R(f^*_{\phi_h}) = 0.47$ *with* $G_{f^*_{\phi_h}} = \emptyset$*. Thus, in contrast to Example 3.1 where* f *is unconstrained*

other than $G_f \in \mathcal{G}$, the linear functional form constraint on $\mathcal{F}_\mathcal{G}$ invalidates the calibration property of the hinge risk, and the hinge risk minimization is no longer consistent.

This example illustrates that even with the hinge loss, consistency to second best classifier becomes a fragile property once the functional form of f is constrained in addition to the prediction set constraint $G_f \in \mathcal{G}$. Consequently, it is natural to ask what additional functional form restriction we can safely introduce to $\mathcal{F}_\mathcal{G}$ without harming the consistency, i.e., for what subclass $\widetilde{\mathcal{F}}_\mathcal{G} \subset \mathcal{F}_\mathcal{G}$, minimizing the hinge risk $R_{\phi_h}(f)$ over $f \in \widetilde{\mathcal{F}}_\mathcal{G}$ leads to a classifier that minimizes the classification risk $R(f)$ over $f \in \mathcal{F}_\mathcal{G}$?

Formally, we introduce the following definition of *classification-preserving reduction of* $\mathcal{F}_\mathcal{G}$.

Definition 4.2 (Classification-preserving reduction)**.** *Let $\tilde{f}^* \in \arg\inf_{f \in \widetilde{\mathcal{F}}_\mathcal{G}} R_{\phi_h}(f)$. A subclass of classifiers $\widetilde{\mathcal{F}}_\mathcal{G}$ $(\subseteq \mathcal{F}_\mathcal{G})$ is a classification-preserving reduction of $\mathcal{F}_\mathcal{G}$ if*

$$R(\tilde{f}^*) = \inf_{f \in \mathcal{F}_\mathcal{G}} R(f)$$

holds for any P, distribution on $\{-1, 1\} \times \mathcal{X}$.

To start with the heuristic, consider a simple case where $\widetilde{\mathcal{F}}_\mathcal{G}$ consists of piecewise constant functions with at most $2J$ jumps, $J \geq 1$, in the following form:

$$\begin{aligned}\widetilde{\mathcal{F}}_{\mathcal{G},J} = \Bigg\{ & f(\cdot) = \sum_{j=1}^{J} c_j^+ 1\{\cdot \in G_j\} - \sum_{j=1}^{J} c_j^- 1\left\{\cdot \notin \tilde{G}_j\right\} : \\ & G_j, \tilde{G}_j \in \mathcal{G} \text{ and } c_j^+, c_j^- \geq 0 \text{ for } j = 1, \ldots, J; \\ & G_J \subseteq \cdots \subseteq G_1 \subseteq \tilde{G}_1 \subseteq \cdots \subseteq \tilde{G}_J; \sum_{j=1}^{J} c_j^+ = 1, \sum_{j=1}^{J} c_j^- = 1 \Bigg\}. \end{aligned} \tag{17}$$

By construction, any function in $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ is a step function bounded in $[-1, 1]$ and its sublevel sets $\{x \in \mathcal{X} : f(x) \leq t\}$ belong to $\mathcal{G}$ for any $t \in [-1, 1]$.

Let $\mathcal{G}^* \equiv \arg\inf_{G \in \mathcal{G}} \mathcal{R}(G)$ be the collection of the best prediction sets, and $\mathcal{R}^* \equiv \inf_{G \in \mathcal{G}} \mathcal{R}(G)$ be the optimal classification risk. The following lemma shows that $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ is a classification-preserving reduction of $\mathcal{F}_\mathcal{G}$.

Lemma 4.3. *Let $\mathcal{G} \subseteq 2^\mathcal{X}$ be an arbitrary class of prediction sets.*
(i) $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ is a classification-preserving reduction of $\mathcal{F}_\mathcal{G}$, and $\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}(f) = 2\mathcal{R}^*$ holds.
(ii) For $G^*, \tilde{G}^* \in \mathcal{G}^*$ such that $G^* \subseteq \tilde{G}^*$, $\tilde{f}^\dagger(\cdot) \equiv 1\{\cdot \in G^*\} - 1\{\cdot \notin \tilde{G}^*\}$ is a minimizer of $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$.

Proof. See Appendix A. □

Note that the function $\tilde{f}^\dagger(x) = 1\{x \in G^*\} - 1\{x \notin \tilde{G}^*\}$ takes a value 1 when $x \in G^*$, a value 0 when $x \in \tilde{G}^* \backslash G^*$, and a value -1 when $x \notin \tilde{G}^*$. When we set $G^* = \tilde{G}^*$, $\tilde{f}^\dagger$ is a step function indicating G^* and $(G^*)^c$ with values $+1$ and -1, respectively.

Lemma 4.3 gives an example of a classification-preserving reduced class. Characteristic features of $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ are (i) sublevel sets of any $f \in \widetilde{\mathcal{F}}_{\mathcal{G},J}$ are in $\mathcal{G}$, and (ii) it contains $1\{x \in G^*\} - 1\{x \notin \tilde{G}^*\}$ for some $G^*, \tilde{G}^* \in \mathcal{G}^*$ with $G^* \subseteq \tilde{G}^*$.

It turns out that these two are the key features that need to be maintained for $\widetilde{\mathcal{F}}_{\mathcal{G}}$ to generalize Lemma 4.3. The next theorem is the second main theorem of the paper that extends Lemma 4.3 to a more general class of classifiers that can accommodate continuous ones.

Theorem 4.4 (Consistency under classification-preserving reduction)**.** *Given an arbitrary class of prediction sets $\mathcal{G} \subseteq 2^{\mathcal{X}}$ and $\mathcal{F}_{\mathcal{G}} = \{f : G_f \in \mathcal{G},\ f(\cdot) \in [-1, 1]\}$, suppose $\widetilde{\mathcal{F}}_{\mathcal{G}} \subset \mathcal{F}_{\mathcal{G}}$ satisfies the following two conditions:*

1. *For every $f \in \widetilde{\mathcal{F}}_{\mathcal{G}}$,*
$$\{x \in \mathcal{X} : f(x) \leq t\} \in \mathcal{G} \tag{18}$$
holds for all $t \in [-1, 1]$;

2. *For some $G^*, \tilde{G}^* \in \mathcal{G}^*$ with $G^* \subseteq \tilde{G}^*$,*
$$\tilde{f}^\dagger(\cdot) = 1\{\cdot \in G^*\} - 1\{\cdot \notin \tilde{G}^*\} \in \widetilde{\mathcal{F}}_{\mathcal{G}}. \tag{19}$$
holds.

Then, we the following holds:
(i) $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}}$, and $\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f) = 2\mathcal{R}^*$;
(ii) For $G^*, \tilde{G}^* \in \mathcal{G}^*$ such that $G^* \subseteq \tilde{G}^*$, $\tilde{f}^\dagger(\cdot) = 1\{\cdot \in G^*\} - 1\{\cdot \notin \tilde{G}^*\}$ is a minimizer of $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$.

Proof. See Appendix A. □

This theorem shows that the two conditions (18) and (19) aresufficient for $\widetilde{\mathcal{F}}_{\mathcal{G}}$ to be a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}}$. This result holds regardless of whether $\mathcal{F}_{\mathcal{G}}$ is correctly R-specified or not. Examples 4.6 and 4.7 in the end of this section show examples of classification-preserving reductions for linear classification and monotone classification.

The two conditions (18) and (19) are simple to interpret and guarantee the consistency of hinge risk minimization, while they do not imply that the empirical hinge risk

minimization over $\widetilde{\mathcal{F}}_{\mathcal{G}}$ can be reduced to a convex optimization. We do not know a general way to construct a classification-preserving reduction that makes the empirical hinge risk minimization a convex programming. For monotone classification analyzed in Section 6, we propose two constructions of $\widetilde{\mathcal{F}}_{\mathcal{G}_M}$, one of which is exactly and another is approximately classification-preserving reduction of $\mathcal{F}_{\mathcal{G}_M}$. Furthermore, we show that for both cases minimization of the empirical hinge risk becomes linear programming.

Although Theorem 4.4 shows the consistency of hinge risk minimization over $\widetilde{\mathcal{F}}_{\mathcal{G}}$, it does not lead to the generalized Zhang's (2004) inequality as in Corollary 3.4. Instead, the following corollary gives proportional equality between the $\mathcal{G}$-constrained excess classification risk and the $\mathcal{F}_{\mathcal{G}}$-constrained excess hinge risk with an extra term added.

Corollary 4.5. *Assume $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is a subclass of $\mathcal{F}_{\mathcal{G}}$ satisfying conditions (18) and (19) in Theorem 4.4. If $\Delta C_\phi(\eta) = c(1-2\eta)$ holds for some $c > 0$ and any $\eta \in [0,1]$, there holds*

$$c(R(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f)) = \frac{1}{2}\left(R_\phi(f) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_\phi(f)\right) + \frac{1}{2}\left(R_\phi(1\{\cdot \in G_f\} - 1\{\cdot \notin G_f\}) - R_\phi(f)\right) \tag{20}$$

for any classifier $f : \mathcal{X} \mapsto [-1,1]$. Moreover, the following holds:

$$c(R(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f)) \leq \frac{1}{2}\left\{\left(R_\phi(f) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_\phi(f)\right) + \left(R_\phi(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_\phi(f)\right)\right\} \tag{21}$$

for any $f \in \mathcal{F}_{\mathcal{G}}$.

Proof. See Appendix A. □

The extra term in (20) measures the difference of the hinge risks between a classifier f and a step function indicating the prediction set of f by the values $+1$ or -1. By the fact that some best classifiers have the form of $\tilde{f}^*(\cdot) = 1\{\cdot \in G^*\} - 1\{\cdot \notin G^*\}$ for $G^* \in \mathcal{G}^*$ (Theorem 4.4 (ii)), if f approximates well such classifier, the extra term gets close to zero. In the following section, we use equation (20) to derive statistical properties of the hinge risk minimization in terms of the $\mathcal{G}$-constrained excess classification risk. Equation (21) implies that the $\mathcal{G}$-constrained excess classification risk is bounded from above by the average of the two $\mathcal{F}_{\mathcal{G}}$-constrained excess hinge risks. One is over $\widetilde{\mathcal{F}}_{\mathcal{G}}$ and another is over $\mathcal{F}_{\mathcal{G}}$. We are not aware if we can bound from above the latter excess hinge risk by a term proportional to the former excess hinge risk, so we do not have Zhang's inequality in the form of Corollary 3.4 where constrained-classification-preserving reduction $\widetilde{\mathcal{F}}_{\mathcal{G}}$ replaces $\mathcal{F}_{\mathcal{G}}$.

We conclude this section by presenting examples of classes of classifiers that approximately or exactly satisfy the conditions for classification-preserving reduction.

Example 4.6 (Linear classification with a class of transformed logistic functions)**.** *Suppose that the prediction sets are subject to the linear index rules:*

$$\mathcal{G}_L = \{x \in \mathbb{R}^{d_x} : x^T\beta \geq 0 : \beta \in \mathbb{R}^{d_x}\},$$

where $\mathcal{X} = \mathbb{R}^{d_x}$*. Let* $\pi(\alpha, k) \equiv (1 - e^{-k\alpha})/(1 + e^{-k\alpha}) = 2/(1 + e^{-k\alpha}) - 1$ *be a transformed logistic function and define a class of classifiers*

$$\mathcal{F}_{Logit} = \{\pi(x^T\beta, k) : \beta \in \mathbb{R}^{d_x} \text{ and } k \in \mathbb{R}_+\}, \tag{22}$$

where k *is a tuning parameter that decides the steepness of the logistic curve.* $\mathcal{F}_{Logit}$ *satisfies condition (18).*[3] *Since* $\mathcal{F}_{Logit}$ *at fixed* $k < \infty$ *rules out any step function, condition (19) is not exactly met. Let* β^* *be such that* $\{x \in \mathcal{X} : x^T\beta^* \geq 0\} \in \mathcal{G}^*$*. As* $k \to \infty$*,* $\pi(x^T\beta^*, k)$ *approximates* $\text{sign}(x^T\beta^*)$*, so condition (19) is met approximately for large* k*. Every function in* $\mathcal{F}_{Logit}$ *is smooth and depends on a finite number of parameters. Hence, the empirical hinge risk becomes a smooth and continuous function with finite number of parameters, although it is not generally convex.*

Example 4.7 (Monotonic classification with a class of monotone functions)**.** *The hinge risk minimization with the monotonicity restriction retains the consistency when we use a class of monotone functions. Let* $\precsim$ *be a partial order on* $\mathcal{X}$*, and let* $\mathcal{G}_{\precsim}$ *be the collection of all* $G \in 2^{\mathcal{X}}$ *that respect the monotonicty (i.e., if* $x_1 \precsim x_2$ *and* $x_1 \in G$*, then* $x_2 \in G$*). Define* $\widetilde{\mathcal{F}}_{\mathcal{G}_{\precsim}}$ *as a class of functions* $f : \mathcal{X} \to [-1, 1]$ *that are weakly monotonic in* $\precsim$ *(i.e., satisfying* $f(x_1) \leq f(x_2)$ *if* $x_1 \precsim x_2$*). Then the prediction set of any* $f \in \widetilde{\mathcal{F}}_{\mathcal{G}_{\precsim}}$ *respects the partial order* $\precsim$ *(i.e., if* $x_1 \precsim x_2$ *and* $x_1 \in G_f$*, then* $x_2 \in G_f$*). For any* $t \in [-1, 1]$ *and* $f \in \widetilde{\mathcal{F}}_{\precsim}$*,* $\{x : f(x) \leq t\} = \{x : x \precsim \tilde{x} \text{ for any } \tilde{x} \text{ such that } f(\tilde{x}) = t\} \in \mathcal{G}_{\precsim}$ *holds, satisfying the condition (18). In addition, since* $\tilde{f}^\dagger(\cdot) = 1\{\cdot \in G^*\} - 1\{\cdot \notin \tilde{G}^*\}$*, for some* $G^*, \tilde{G}^* \in \mathcal{G}_{\precsim}$ *with* $G^* \subseteq \tilde{G}^*$*, is weakly monotonic in* $\precsim$*,* $\tilde{f}^\dagger \in \mathcal{G}_{\precsim}$ *holds, satisfying condition (19). Hence* $\widetilde{\mathcal{F}}_{\mathcal{G}_{\precsim}}$ *is a classification-preserving reduction of* $\mathcal{F}_{\mathcal{G}_{\precsim}}$*. Therefore, according to Theorem 4.4, the hinge risk minimization over* $\widetilde{\mathcal{F}}_{\precsim}$ *leads to the best classifier. Section 6 focuses on monotone classification and investigates statistical and computational properties.*

[3] Fix $\beta \in \mathbb{R}^{d_x}$ and $k \in \mathbb{R}_+$. The condition (18) is satisfied as, for any $t \in [-1, 1]$, $\{x : \pi(x^T\beta, k) \leq t\} = \{x : x^T\beta \leq \pi^{-1}(t, k)\} \in \mathcal{G}$, where $\pi^{-1}(\cdot, k)$ is an inverse function of $\pi(\cdot, k)$ with the fixed k.

5 Statistical property

The analyses presented so far concern consistency of the surrogate loss approach in terms of the population risk criterion. It is important to note that our results of Theorems 3.2 and 4.4 do not impose any restriction on the underlying distribution of (Y, X). Accordingly, equivalence of the risk orderings and risk-minimizing classifiers between the classification and hinge risks remains valid even if we consider empirical analogues of the risks constructed upon the empirical distribution of the sample. It hence guarantees that a classifier minimizing the empirical hinge risk over $\mathcal{F}_{\mathcal{G}}$ or over a classification-preserving reduction $\widetilde{\mathcal{F}}_{\mathcal{G}}$ also minimizes the empirical classification risk.

In this section, we assess the generalization performance of hinge-risk minimizing classifiers, allowing for general misspecification of the constrained class of classifiers. For that goal, let $\mathcal{G}$ be fixed and consider $\check{\mathcal{F}}$ a class of classifiers whose members satisfy $-1 \leq f \leq 1$. $\check{\mathcal{F}}$ may or may not be a subclass of $\mathcal{F}_{\mathcal{G}}$, while in our analysis of monotone classification below, $\check{\mathcal{F}}$ corresponds to an approximation of a classification-preserving reduction $\widetilde{\mathcal{F}}_{\mathcal{G}}$. Let $\{(Y_i, X_i) : i = 1, \ldots, n\}$ be a sample of observations that are independently and identically distributed (i.i.d) as (Y, X). We denote the joint distribution of a size n sample by P^n and the expectation with respect to P^n by $E_{P^n}(\cdot)$. Define the empirical classification risk and empirical hinge risk, respectively, by

$$\hat{R}(f) \equiv \frac{1}{n}\sum_{i=1}^{n} 1\{Y_i \cdot \operatorname{sign}(f(X_i)) \leq 0\},$$
$$\hat{R}_{\phi_h}(f) \equiv \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - Y_i f(X_i)\} = \frac{1}{n}\sum_{i=1}^{n} (1 - Y_i f(X_i)),$$

where the max operator in the hinge loss is redundant if we constrain $f(\cdot)$ to $[-1, 1]$. Let $\hat{f}$ be a classifier that minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\check{\mathcal{F}}$. We evaluate a statistical property of $\hat{f}$ in terms of the excess classification risk relative to the minimal risk over $\mathcal{F}_{\mathcal{G}}$. In particular, we will derive a distribution-free upper bound for the mean of the excess classification risk.

Let $\widetilde{\mathcal{F}}_{\mathcal{G}}$ be a subclass of $\mathcal{F}_{\mathcal{G}}$ and satisfy the conditions (18) and (19) in Theorem 4.4. $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is hence a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}}$ (Definition 4.2). Following Corollary 4.5, we have

$$R(\hat{f}) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f) = \frac{1}{2}\left(R_{\phi_h}(\hat{f}) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f)\right) + \frac{1}{2}\left(R_{\phi_h}\left(1\left\{\cdot \in G_{\hat{f}}\right\} - 1\left\{\cdot \notin G_{\hat{f}}\right\}\right) - R_{\phi_h}(\hat{f})\right). \tag{23}$$

When $\check{\mathcal{F}}$ coincides with $\widetilde{\mathcal{F}}_{\mathcal{G}}$, evaluating each term in the right hand side of (23) gives an

upper bound for the mean of the $\mathcal{G}$-constrained excess classification risk of $\hat{f}$.

Let $H_1^B(\epsilon, \mathcal{F}, P_X)$ be the $L_1(P_X)$-bracketing entropy of a class of functions $\mathcal{F}$ and $H_1^B(\epsilon, \mathcal{G}, P_X)$ be that of a class of prediction sets $\mathcal{G}$.[4] For these definitions, see Definition B.1 in Appendix B. When $\check{\mathcal{F}}$ coincides with $\widetilde{\mathcal{F}}_{\mathcal{G}}$, the following theorem gives a non-asymptotic distribution-free upper bound for the mean of the $\mathcal{G}$-constrained excess classification risk in terms of the bracketing entropy.

Theorem 5.1. *Let $\widetilde{\mathcal{F}}_{\mathcal{G}}$ be a subclass of $\mathcal{F}_{\mathcal{G}}$ and satisfy the conditions (18) and (19) in Theorem 4.4. Suppose that $\mathcal{P}$ is a class of distributions on $\{-1,1\} \times \mathcal{X}$ such that there exist positive constants C and r for which*

$$H_1^B(\epsilon, \mathcal{G}, P_X) \leq C\epsilon^{-r} \tag{24}$$

holds for any $P \in \mathcal{P}$ and $\epsilon > 0$, or

$$H_1^B\left(\epsilon, \widetilde{\mathcal{F}}_{\mathcal{G}}, P_X\right) \leq C\epsilon^{-r} \tag{25}$$

holds for any $P \in \mathcal{P}$ and $\epsilon > 0$. Define $\tau_n = n^{-1/2}$ if $r < 1$, $\tau_n = \log(n)/\sqrt{n}$ if $r = 1$, and $\tau_n = n^{-1/(r+1)}$ if $r \geq 2$. Let $q_n = \sqrt{n}\tau_n$. Then, for $\hat{f} \in \arg\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f)$, the following holds:

$$\sup_{P \in \mathcal{P}} E_{P^n}\left[R(\hat{f}) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f)\right] \leq L_C(r, n), \tag{26}$$

where

$$L_C(r,n) = \begin{cases} 2D_1\tau_n + 4D_2\exp\left(-D_1^2 q_n^2\right) & \text{if } r \geq 1 \\ 2D_3\tau_n + 2n^{-1}D_4 & \text{if } r < 1 \end{cases} \tag{27}$$

for some positive constants D_1, D_2, D_3, D_4, which depend only on C and r.

Proof. See Appendix B. □

The upper bound for the mean of the $\mathcal{G}$-constrained excess classification risk converges to zero at the rate of τ_n, which depends on r in the bracketing entropy condition (24) or (25). Dudley (1999) shows many examples that satisfy these bracketing entropy conditions. In particular, a class $\mathcal{G}_{\precsim} \subseteq 2^{\mathcal{X}}$ for the monotone classification, introduced in Example 4.7, satisfies the condition (24) with r being $d_x - 1$ (see Theorem 8.3.2 in Dudley (1999)).

[4] With a slight abuse of notation, we notate by $H_1^B(\epsilon, \mathcal{G}, P_X)$ the bracketing entropy number of the class of indicator functions, $H_1^B(\epsilon, \mathcal{H}_{\mathcal{G}}, P_X)$, where $\mathcal{H}_{\mathcal{G}} \equiv \{1\{\cdot \in G\} : G \in \mathcal{G}\}$.

We next consider the case when $\check{\mathcal{F}}$ does not coincide with $\widetilde{\mathcal{F}}_{\mathcal{G}}$. This case corresponds to a scenario that minimizing the empirical hinge risk over $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is difficult while minimizing over $\check{\mathcal{F}}$, a class approximating $\widetilde{\mathcal{F}}_{\mathcal{G}}$, is feasible.

A further decomposition of $R_{\phi_h}(\hat{f}) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f)$ in (23) leads to

$$R(\hat{f}) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f) = \frac{1}{2}\left(R_{\phi_h}\left(\hat{f}\right) - \inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f)\right) + \frac{1}{2}\left(\inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f)\right) + \frac{1}{2}\left(R_{\phi_h}\left(1\left\{\cdot \in G_{\hat{f}}\right\} - 1\left\{\cdot \notin G_{\hat{f}}\right\}\right) - R_{\phi_h}(\hat{f})\right). \tag{28}$$

Hence the $\mathcal{G}$-constrained excess classification risk is decomposed into the three terms. We call the first term estimation error, the second term approximation error to a best classifier, and the third term approximation error to a step classifier. Evaluating each error gives an upper bound for the $\mathcal{G}$-constrained excess classification risk.

The following theorem evaluates the estimation error in terms of the bracketing entropy.

Theorem 5.2. *Let $\widetilde{\mathcal{F}}_{\mathcal{G}}$ be a subclass of $\mathcal{F}_{\mathcal{G}}$ and satisfy the conditions (18) and (19) in Theorem 4.4. Suppose that $\mathcal{P}$ is a class of distributions on $\{-1,1\} \times \mathcal{X}$ such that there exist positive constants C' and r' for which*

$$H_1^B\left(\epsilon, \check{\mathcal{F}}, P_X\right) \leq C' \epsilon^{-r'} \tag{29}$$

holds for any $P \in \mathcal{P}$ and $\epsilon > 0$. Let $\hat{f} \in \arg\inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f)$. Then, there holds

$$\sup_{P \in \mathcal{P}} E_{P^n}\left[R(\hat{f}) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R(f)\right] \leq L_{C'}(r', n) + \frac{1}{2}\left(\inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R_{\phi_h}(f)\right) + \frac{1}{2}\left(R_{\phi_h}\left(1\left\{\cdot \in G_{\hat{f}}\right\} - 1\left\{\cdot \notin G_{\hat{f}}\right\}\right) - R_{\phi_h}(\hat{f})\right), \tag{30}$$

where $L_{C'}(r', n)$ is defined as in Theorem 5.1.

Proof. See Appendix B. □

Remark 5.3 (Approximation errors)**.** *Evaluating each approximation error in (30) depends on the functional form restriction placed for $f \in \check{\mathcal{F}}$. If $\check{\mathcal{F}}$ grows and approaches to $\widetilde{\mathcal{F}}_{\mathcal{G}}$ as $n \to \infty$, each approximation error converges to zero. In Section 6.2 below, we consider the monotone classification problem and set $\check{\mathcal{F}}$ being a sieve of Bernstein polynomials and characterize convergence of these two approximation errors. We then apply Theorem 5.2 to obtain the regret convergence rate of the estimated monotone classifier.*

6 Applications to monotone classification

This section applies the general theoretical results shown in Sections 3–5 to monotone classification problem (Example 2.5). By Theorem 3.2, we limit our analysis to the hinge loss. We assume that $\mathcal{X}$ is compact in $\mathbb{R}^{d_x}$, $d_x < \infty$, and without loss of generality, we represent it as the d_x-dimensional unit hypercube (i.e., $\mathcal{X} = [0,1]^{d_x}$). To be specific, we consider the class of monotone prediction sets $\mathcal{G}_M$ such that, for any $G \in \mathcal{G}_M$ and $x, \tilde{x} \in \mathcal{X}$, $x \in G$ and $x \leq \tilde{x}$ implies $\tilde{x} \in G$ holds[5] (i.e., $\mathcal{G}_M$ respects the partial order $\leq$ on $\mathcal{X}$). Accordingly, we have the class of monotonically increasing classifiers that can be represented as

$$\mathcal{F}_M \equiv \{f : f(x) \leq f(\tilde{x}) \text{ for any } x, \tilde{x} \in \mathcal{X} \text{ with } x \leq \tilde{x} \,;\, f(\cdot) \in [-1,1]\}.$$

In this section, we first study the monotone classification problem on $\mathcal{F}_M$. Note that $\mathcal{F}_M$ is a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}_M}$ (see Example 4.7). As an alternative to $\mathcal{F}_M$, we next consider using a sieve of Bernstein polynomials to approximate an hinge-risk minimizing classifier on $\mathcal{F}_M$. The Bernstein polynomial is known for its capability to accommodate bound constraints and various shape constraints on functions (e.g., monotonicity and/or convexity). The class of Bernstein polynomials becomes a classification-preserving reduction only at the limit with a growing order of polynomials.

6.1 Nonparametric monotone classification

We first consider the hinge risk minimization with the class of monotonically increasing classifiers $\mathcal{F}_M$. Let $\hat{f}_M$ be a minimizer of $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{F}_M$. Since the hinge risk for classifiers constrained to $-1 \leq f(x) \leq 1$ gives the linear loss $\phi_h(yf(x)) = 1 - yf(x)$, minimization of the empirical hinge risk can be formulated as the following linear programming:

$$\begin{aligned} &\max_{f_1,\ldots,f_n} \sum_{i=1}^{n} Y_i f_i \\ \text{s.t.} \quad &f_i \geq f_j \text{ for any } X_i \neq X_j \text{ with } X_i \geq X_j \text{ for } 1 \leq i \leq j \leq n; \\ &-1 \leq f_i \leq 1 \text{ for } 1 \leq i \leq n, \end{aligned} \tag{31}$$

where the first inequality constraints correspond to the monotononicity constraint on $\mathcal{F}_M$, and the second inequality constraints correspond to the range constraint for $f \in \mathcal{F}_M$. Solving this linear programming yields the values of $\hat{f}_M$ at the values of x observed in the training sample. Let $\left(\hat{f}_M(X_1), \ldots, \hat{f}_M(X_n)\right)$ be the solution of (31). Then any func-

[5] We define the partial order $\leq$ on $\mathcal{X}$ as follows. For any $x = (x_1, \ldots, x_d)^T$ and $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_d)^T$, we say $x \leq \tilde{x}$ if $x_j \leq \tilde{x}_j$ for every $j = 1, \ldots, d$. We further say $x < \tilde{x}$ if $x \leq \tilde{x}$ holds and for some $j \in \{1, \ldots, d\}$, $x_j < \tilde{x}_j$ holds.

tion in $\mathcal{F}_M$ that passes the points $\left(\left(X_1, \hat{f}_M(X_1)\right), \ldots, \left(X_1, \hat{f}_M(X_n)\right)\right)$ minimizes the empirical hinge risk over $\mathcal{F}_M$.[6] Since $\mathcal{F}_M$ is a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}_M}$, Theorem 4.4 with P replaced by P^n shows that any solution to (31) exactly minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{F}_{\mathcal{G}_M}$.

We investigate a statistical property of this procedure. Since $\mathcal{F}_M$ is a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}_M}$, we can apply Theorem 5.1. For this goal, we first characterize an upper bound of the bracketing entropy number of the class of monotone prediction sets. The next lemma, which we borrow from Theorem 8.3.2 in Dudley (1999), gives an upper bound for the the $L_1(P_X)$-bracketing entropy of $\mathcal{G}_M$. Here, we assume that X is continuously distributed with the bounded density.

Lemma 6.1. *Suppose that P_X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and has a density that is bounded from above by a finite constant $A > 0$. Then there exists a constant C, which depends only on A, such that*

$$H_1^B(\epsilon, \mathcal{G}_M, P_X) \leq C\epsilon^{1-d_x}.$$

holds for all $\epsilon > 0$.

Proof. See Appendix C. □

With this lemma, setting $r = 1 - d_x$ in Theorem 5.1 yields a finite sample uniform upper bound for the $\mathcal{G}$-constrained excess classification risk of $\hat{f}_M$. It shows that the excess risk convergence rate of $\hat{f}_M$ obtained by linear programming (31) attains the same convergence rate as the welfare regret of monotone treatment rules shown by by Mbakop and Tabord-Meehan (2021).

Theorem 6.2. *Let $\mathcal{P}$ be a class of distributions on $\{-1, 1\} \times \mathcal{X}$ such that the marginal distribution P_X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and has a density that is bounded from above by some finite constant $A > 0$. Define $\tau_n = n^{-1/2}$ if $d_x = 1$, $\tau_n = \log(n)/\sqrt{n}$ if $d_x = 2$, and $\tau_n = n^{-1/d_x}$ if $d_x \geq 3$. Let $q_n = \sqrt{n}\tau_n$. Then,*

[6] All classifiers obtained from this procedure predict a unique label at each point of x observed in the training sample, whereas they may not give a unique prediction at a point of x not observed in the training sample. One possible way to predict a label at an unobserved point of x without violating the monotonicity constraint is to predict its label by the largest label among those predicted by all classifiers in $\arg\inf_{f \in \mathcal{F}_M} \hat{R}_{\phi_h}(f)$. Let $\widetilde{\mathcal{X}}$ be a set of x observed in the training sample. Given any $\hat{f}_M \in \arg\inf_{f \in \mathcal{F}_M} \hat{R}_{\phi_h}(f)$, this way is equivalent to predict a label of $x \in \mathcal{X} \backslash \widetilde{\mathcal{X}}$ by the sign of $\min\{\hat{f}_M(\tilde{x}) : \tilde{x} \in \widetilde{\mathcal{X}}, \tilde{x} \geq x\}$ if there exists $\tilde{x} \in \widetilde{\mathcal{X}}$ such that $\tilde{x} \geq x$, and predict by 1 otherwise .

for $\hat{f}_M \in \arg\inf_{f\in\mathcal{F}_M} \hat{R}_{\phi_h}(f)$*, there holds*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[R(\hat{f}_M) - \inf_{f\in\mathcal{F}_{\mathcal{G}_M}} R(f)\right] \leq \begin{cases} 2D_1\tau_n + 4D_2\exp\left(-D_1^2 q_n^2\right) & \text{if } d_x \geq 2 \\ 2D_3\tau_n + 2n^{-1}D_4 & \text{if } d_x = 1 \end{cases}$$

for some positive constants D_1, D_2, D_3, D_4*, which depend only on* d_x *and* A*.*

Proof. Since $\mathcal{F}_M$ satisfies the conditions (18) and (19) in Theorem 4.4 with $\mathcal{G}$ being $\mathcal{G}_M$ (Example 4.7), the result follows from Theorem 5.1 and Lemma 6.1. □

This theorem guarantees the consistency of the monotone classification using the hinge loss and the class of monotone classifiers $\mathcal{F}_M$. The rate of convergence corresponds to τ_n.

6.2 Monotone classification with Bernstein polynomial

To illustrate our theoretical results in monotone classification, the second approach we consider is to use multivariate Bernstein polynomials to approximate a best classifier in $\mathcal{F}_M$.

Let $b_{kj}(x) = \binom{k}{j} x^j (1-x)^{k-j}$ be the Bernstein basis. The Bernstein polynomial for a d_x-dimensional function takes the following form:

$$B_{\mathbf{k}}(\theta, x) = \sum_{j_1=0}^{k_1} \cdots \sum_{j_{d_x}=0}^{k_{d_x}} \theta_{j_1\ldots j_d} \cdot \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x})\right),$$

where $\mathbf{k} = (k_1, \ldots, k_{d_x})^T$ is a vector collecting the orders of the Bernstein polynomial bases specified by the analyst, $\theta \equiv \left\{\theta_{j_1\ldots j_{d_x}}\right\}_{j_1=0,\ldots,k_1;\cdots;j_{d_x}=0,\ldots,k_{d_x}}$ is a $(k_1+1)\times\cdots\times(k_{d_x}+1)$-dimensional vector of the parameters to be estimated, and x_j denotes the j-th element of the d_x-dimensional vector x. If $-1 \leq \theta_{j_1\ldots j_{d_x}} \leq 1$ for all $(j_1, \ldots, j_{d_x})$, the range of the function $B_{\mathbf{k}}(\theta, \cdot)$ is bounded in $[-1, 1]$. Moreover, if $\theta_{j_1\ldots j_{d_x}} \geq \theta_{\tilde{j}_1\ldots\tilde{j}_{d_x}}$ for all $(j_1, \ldots, j_{d_x}) \geq \left(\tilde{j}_1, \ldots, \tilde{j}_{d_x}\right)$, $B_{\mathbf{k}}(\theta, \cdot)$ is non-decreasing in x.[7] Hence, to preserve the bound and non-decreasing constraints on $\mathcal{F}_M$, the class of Bernstein polynomials should be constrained on

$$\mathcal{B}_{\mathbf{k}} = \left\{B_{\mathbf{k}}(\theta, \cdot) : \theta \in \widetilde{\Theta}\right\},$$

[7] On the contrary, if $\theta_{j_1\ldots j_{d_x}} \leq \theta_{\tilde{j}_1\ldots\tilde{j}_{d_x}}$ for all $(j_1, \ldots, j_{d_x}) \geq \left(\tilde{j}_1, \ldots, \tilde{j}_{d_x}\right)$, $B_{\mathbf{k}}(\theta, \cdot)$ is non-increasing in x. See, e.g., Wang and Ghosh (2012) for the bound and shape preserving properties of the multivariate Bernstein polynomials.

where $\widetilde{\Theta}$ is a class of θ such that $\theta_{j_1 \ldots j_{d_x}} \in [-1,1]$ for all $(j_1, \ldots, j_{d_x})$ and $\theta_{j_1 \ldots j_{d_x}} \geq \theta_{\tilde{j}_1 \ldots \tilde{j}_{d_x}}$ for all $(j_1, \ldots, j_{d_x}) \geq (\tilde{j}_1, \ldots, \tilde{j}_{d_x})$. [8]

We propose to estimate the best classifier by a classifier $\hat{f}_B$ that minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{B}_{\mathbf{k}}$. A proper choice of $\mathbf{k}$ will be discussed later. This minimization problem is a convex optimization as the objective function $\hat{R}_{\phi}(\cdot)$ is linear in θ with the linear inequality constraints on $\widetilde{\Theta}$, and hence can be formulated as a linear programming (see Remark 6.6 below).

Note that $\mathcal{B}_{\mathbf{k}} \subseteq \mathcal{F}_M$. Setting $\check{F}$ to $\mathcal{B}_{\mathbf{k}}$ in the framework of Section 5, the excess classification risk of $\hat{f}_B$ is decomposed into three errors: estimation error ($R_{\phi_h}\left(\hat{f}_B\right) - \inf_{f \in \mathcal{B}_{\mathbf{k}}} R_{\phi_h}(f)$), approximation error to the best classifier ($\inf_{f \in \mathcal{B}_{\mathbf{k}}} R_{\phi_h}(f) - \inf_{f \in \mathcal{F}_M} R_{\phi}(f)$), and approximation error to the step function ($R_{\phi}(1\left\{\cdot \in G_{\hat{f}_B}\right\} - 1\left\{\cdot \notin G_{\hat{f}_B}\right\}) - R_{\phi}(\hat{f}_B)$). We evaluate each error below.

The following lemma gives finite upper bounds for the two approximation errors.

Lemma 6.3. *Let $k_j \geq 1$, for $j = 1, \ldots, d_x$, be fixed. Suppose that the density of P_X is bounded from above by some finite constant $A > 0$.*
(i) The following holds for the approximation error to the best classifier:

$$\inf_{f \in \mathcal{B}_{\mathbf{k}}} R_{\phi_h}(f) - \inf_{f \in \mathcal{F}_M} R_{\phi_h}(f) \leq 2A \sum_{j=1}^{d_x} \sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x} \frac{4}{\sqrt{k_j}}.$$

(ii) For $\hat{f}_B \in \arg\inf_{f \in \mathbf{B}_{\mathbf{k}}} \hat{R}_{\phi_h}(f)$ such that its coefficients of the Bernstein bases take values in $\{-1, 1\}$, the following holds for the approximation error to the step function:

$$R_{\phi_h}\left(1\left\{\cdot \in G_{\hat{f}_B}\right\} - 1\left\{\cdot \notin G_{\hat{f}_B}\right\}\right) - R_{\phi_h}(\hat{f}_B) \leq 2A \sum_{j=1}^{d_x} \sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x} \frac{4}{\sqrt{k_j}}.$$

Proof. See Appendix C. □

The two approximation errors have the same upper bound which converges to zero as k_j $(j = 1, \ldots, d_x)$ increase. The convergence rate is $\max_{j=1,\ldots,d_x} \sqrt{(\log k_j)/k_j}$. Note also that the upper bound for the approximation error to the step function does not depend on the sample size n. As for the coefficient restriction in (ii), Remark 6.5 introduces two-step procedure to compute $\hat{f}_B$ that minimizes $\hat{R}_{\phi_h}$ while satisfying the coefficient restriction.

The estimation error can be bounded by using Theorem 5.2 with Lemma C.1 in Appendix C. The following theorem shows a finite sample upper bound for the mean of the $\mathcal{G}$-constrained excess classification risk of $\hat{f}_B$.

[8] If $\mathcal{G}_M$ respects the opposite partial order $\geq$ on $\mathcal{X}$, $\widetilde{\Theta}$ should be replaced with a collection of θ such that $\theta_{j_1 \ldots j_{d_x}} \in [-1,1]$ for all $(j_1, \ldots, j_{d_x})$ and $\theta_{j_1 \ldots j_{d_x}} \leq \theta_{\tilde{j}_1 \ldots \tilde{j}_{d_x}}$ for all $(j_1, \ldots, j_{d_x}) \geq (\tilde{j}_1, \ldots, \tilde{j}_{d_x})$.

Theorem 6.4. *Let $\mathcal{P}$ be a class of distributions on $\{-1,1\} \times \mathcal{X}$ that satisfy the same conditions as in Theorem 6.2. Let $\tilde{\tau}_n = \log(n)/\sqrt{n}$ if $d_x = 1$ and $\tilde{\tau}_n = n^{-1/d_x}$ if $d_x \geq 2$. Define $\tilde{q}_n = \sqrt{n}\tilde{\tau}_n$. Then, for $\hat{f}_B \in \arg\inf_{f \in \mathcal{B}_{\mathbf{k}}} \hat{R}_{\phi_h}(f)$ such that its coefficients of the Bernstein bases take values in $\{-1,1\}$, the following holds:*

$$\sup_{P \in \mathcal{P}} E_{P^n}\left[R(\hat{f}_B) - \inf_{f \in \mathcal{F}_{\mathcal{G}_M}} R(f)\right] \leq 2D_1\tilde{\tau}_n + 4D_2 \exp\left(-D_1^2\tilde{q}_n^2\right) + 4A\sum_{j=1}^{d_x}\sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x}\frac{8}{\sqrt{k_j}}, \tag{32}$$

where D_1 and D_2 are some positive constants, which depend only on d_x and A.

Proof. From the fact that $\mathcal{B}_{\mathbf{k}} \subseteq \mathcal{F}_M$ and Lemma C.1 in Appendix C, we have $H_1^B(\epsilon, \mathcal{B}_{\mathbf{k}}, P_X) \leq C\epsilon^{-d_x}$ for some positive constant C, which depends only on A, and all $\epsilon > 0$. Then the result follows by combining Theorem 5.2 and Lemma 6.3. □

The upper bound in (32) converge to zero as the sample size n and the number of the Bernstein bases k_j $(j = 1, \ldots, d_x)$ increase. Note that the rate of convergence for the estimation error in this theorem, $\tilde{\tau}_n$, is slower than that in Theorem 6.2, τ_n. The difference comes from the different orders in the upper bounds of $H_1^B(\epsilon, \mathcal{G}_M, P_X)$ and $H_1^B(\epsilon, \mathcal{F}_M, P_X)$ in Lemmas 6.1 and C.1. To achieve the convergence rate of $\tilde{\tau}_n$ for the mean of the excess risk, the theorem suggests to set the tuning parameters k_j, $j = 1, \ldots, d_x$, sufficiently large so that $\sqrt{\log k_j / k_j} = O(\tilde{\tau}_n)$.

In practice, one may want to select the complexity of the Bernstein polynomials by minimizing penalized empirical surrogate risk. Classification and treatment choice literatures (Koltchinskii (2006), Mbakop and Tabord-Meehan (2021), and the references therein) analyze the regret properties and oracle inequalities for the penalized risk minimizing classifiers. We leave for future research theoretical investigation for applicability of penalization methods to the current hinge risk minimization with the Bernstein polynomials.

Several other remarks follow.

Remark 6.5 (Coefficient restriction)**.** *Theorem 6.4 requires that the estimated Bernstein polynomial classifier has binary coefficients taking values in $\{-1,1\}$. This restriction is needed to make the approximation error to the step function converge to zero. Actually, this restriction is not very strict.*

Fix $\hat{f}_B \in \arg\inf_{f \in \mathbf{B}_{\mathbf{k}}} \hat{R}_{\phi_h}(f)$, and let $\{\hat{\theta}_{j_1 \ldots j_{d_x}}\}_{j_1=0,\ldots,k_1;\cdots;j_{d_x}=0,\ldots,k_{d_x}}$ be the vector of the

coefficients in $\hat{f}_B$*. We make a modified classifier*

$$\hat{f}_B^{\dagger}(x) \equiv \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \operatorname{sign}\left(\hat{\theta}_{j_1 \ldots j_{d_x}}\right) \cdot \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x})\right),$$

which converts each estimated coefficient $\hat{\theta}_{j_1 \ldots j_{d_x}}$ *to either* -1 *or* 1 *depending on its sign. Then Lemma C.2 in Appendix C shows that* $\hat{f}_B^{\dagger}$ *minimizes* $R_{\phi_h}(\cdot)$ *over* $\mathcal{B}_{\mathbf{k}}$ *as well. We can hence apply Theorem 6.4 to the modified estimator* $\hat{f}_B^{\dagger}$ *and obtain the result (32). We therefore recommend using the modification* $\hat{f}_B^{\dagger}$*, instead of* $\hat{f}_B$*, if some coefficients in* $\hat{f}_B$ *are not equal to* -1 *or* 1*.*

Remark 6.6 (Linear programming)**.** *Denote* $\hat{f}_B \in \arg\inf_{f \in \mathbf{B_k}} \hat{R}_{\phi_h}(f)$ *by*

$$\hat{f}_B(x) = \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \hat{\theta}_{j_1 \ldots j_{d_x}} \cdot \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x})\right).$$

The vector of the coefficients $\hat{\theta} := \left\{\hat{\theta}_{j_1 \ldots j_{d_x}}\right\}_{j_1=0,\ldots,k_1;\cdots;j_{d_x}=0,\ldots,k_{d_x}}$ *can be obtained by solving the following linear programming:*

$$\max_{\theta} \sum_{i=1}^{n} Y_i \cdot \left(\sum_{j_1=0}^{k_1} \cdots \sum_{j_1=d_x}^{k_{d_x}} \theta_{j_1 \ldots j_{d_x}} \cdot \left(b_{k_1 j_1}(X_{i1}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{i d_x})\right) \right) \tag{33}$$

$$\begin{aligned} s.t. \quad & \theta_{j_1 \ldots j_{d_x}} \geq \theta_{\tilde{j}_1 \ldots \tilde{j}_{d_x}} \text{ for any } (j_1, \ldots, j_{d_x}) \geq \left(\tilde{j}_1, \ldots, \tilde{j}_{d_x}\right); \\ & -1 \leq \theta_{j_1 \ldots j_{d_x}} \leq 1 \text{ for all } (j_1, \ldots, j_{d_x}), \end{aligned}$$

where X_{ij} *denotes the* j*-th element of* X_i*. The first inequality constraints restrict feasible classifiers on a class of non-decreasing functions. The second inequality constraints bound feasible classifiers on* $[-1, 1]$*.*

The linear programming (31) for the nonparametric monotone classification has n*-decision variables, whereas that (33) has* $(k_1+1) \times \cdots \times (k_{d_x}+1)$*-decision variables. Thus when the dimension of* X *is small to moderate relative to the sample size* n*, the linear programming for the Bernstein polynomials would be easier to compute. The opposite is also true.*

7 Extension to individualized treatment rules

This section extends the primary results obtained in Sections 3–6 for binary classification to weighted classification introduced in Section 1.1, and importantly to causal policy

learning. We follow the same notations and definitions introduced in Section 1.1. We call R^w and R^w_ϕ, defined in (6) and (7), weighted classification risk and weighted ϕ-risk, respectively.

7.1 Consistency of weighted classification with hinge loss

We first show consistency of the weighted classification with hinge risk by following the line of analyses in Sections 3 and 4. Given prespecified $\mathcal{G}$, let $\mathcal{F}_\mathcal{G}$ be as in Section 2. Analogues to $\mathcal{R}(G)$ and $\mathcal{R}_\phi(G)$, we define $\mathcal{R}^w(G) \equiv \inf_{f\in\mathcal{F}_G} R^w(f)$ the weighted-classification risk evaluated at G, and $\mathcal{R}^w_\phi(G) \equiv \inf_{f\in\mathcal{F}_G} R^w_\phi(f)$ the weighted ϕ-risk evaluated at G. Note that $\mathcal{R}^w(G) = R^w(f)$ for all $f \in \mathcal{F}_G$. Let $\mathcal{R}^{w*} \equiv \inf_{G\in\mathcal{G}} \mathcal{R}^w(G) = \inf_{f\in\mathcal{F}_\mathcal{G}} R^w(f)$ be the optimal weighted risk, and $\mathcal{G}^* \equiv \arg\inf_{G\in\mathcal{G}} \mathcal{R}^w(G)$ be the collection of the best prediction sets.

For the weight variable ω, define

$$\begin{aligned}\omega_+(x) &\equiv E_P[\omega \mid X = x, Y = +1]\\ \omega_-(x) &\equiv E_P[\omega \mid X = x, Y = -1].\end{aligned}$$

In the setting of policy learning where $\omega = \omega_p$, $Y = D$, and P satisfies unconfoundedness, ω_+ and ω_- correspond to the regression equations of the potential outcomes divided by the propensity score $e(x) = \eta(x) = \Pr(Y = +1|X = x)$,

$$\begin{aligned}\omega_+(x) &= E_P[Z(+1) \mid X = x]/e(x)\\ \omega_-(x) &= E_P[Z(-1) \mid X = x]/(1 - e(x)).\end{aligned}$$

Let $C_\phi(a, b, c, d) \equiv a\phi(c)\, d + b\phi(-c)\,(1 - d)$, and

$$\begin{aligned}C^{w+}_\phi(\omega_+, \omega_-, \eta) &\equiv \inf_{0\le f\le 1} C_\phi(\omega_+, \omega_-, f, \eta),\\ C^{w-}_\phi(\omega_+, \omega_-, \eta) &\equiv \inf_{-1\le f<0} C_\phi(\omega_+, \omega_-, f, \eta),\\ \Delta C^w_\phi(\omega_+, \omega_-, \eta) &\equiv C^{w+}_\phi(\omega_+, \omega_-, \eta) - C^{w-}_\phi(\omega_+, \omega_-, \eta),\end{aligned}$$

which are analogues to C^+_ϕ, C^-_ϕ, and ΔC_ϕ defined in Section 3. Similarly to (10) and (12), we have

$$\begin{aligned}\mathcal{R}^w(G) = &\int_\mathcal{X} (-\omega_+(x)\,\eta(x) + \omega_-(x)\,(1 - \eta(x))) \cdot 1\{x \in G\}\, dP_X(x)\\ &+ \int_\mathcal{X} \omega_+(x)\,\eta(x)\, dP_X(x),\\ \mathcal{R}^w_\phi(G) = &\int_\mathcal{X} \Delta C^w_\phi(\omega_+(x), \omega_-(x), \eta(x)) \cdot 1\{x \in G\}\, dP_X(x)\end{aligned} \tag{34}$$

$$+\int_{\mathcal{X}} C_{\phi}^{w-}\left(\omega_{+}\left(x\right),\omega_{-}\left(x\right),\eta\left(x\right)\right)dP_{X}\left(x\right). \tag{35}$$

The next theorem generalizes Theorems 3.2, 3.6, and Corollary 3.7 to weighted classification, giving a necessary and sufficient condition for equivalence of the risk ordering among surrogate loss functions. In particular, we show that hinge loss functions share the risk ordering with the 0-1 loss function.

Theorem 7.1. *Let ϕ_1 and ϕ_2 be classification-calibrated loss functions in the sense of Definiton 2.3. Then the following holds for any distribution P of (ω, Y, X) and $G_1, G_2 \in \mathcal{G}$:*

$$\mathcal{R}_{\phi_1}^{w}\left(G_1\right) \leq \mathcal{R}_{\phi_1}^{w}\left(G_2\right) \Leftrightarrow \mathcal{R}_{\phi_2}^{w}\left(G_1\right) \leq \mathcal{R}_{\phi_2}^{w}\left(G_2\right) \tag{36}$$

holds if and only if $\Delta C_{\phi_2}^{w}\left(\omega_+, \omega_-, \eta\right) = c\Delta C_{\phi_1}^{w}\left(\omega_+, \omega_-, \eta\right)$ for some $c > 0$ and any $(\omega_+, \omega_-, \eta) \in \mathbb{R} \times \mathbb{R} \times [0,1]$. Furthermore, if ϕ_1 is the 0-1 loss function (i.e., $\phi_1(\alpha) = 1\{\alpha \leq 0\}$), (36) holds if and only if

$$\Delta C_{\phi_2}^{w}\left(\omega_+, \omega_-, \eta\right) = c\left(-\omega_+\eta + \omega_-\left(1-\eta\right)\right) \text{ for some } c > 0. \tag{37}$$

In particular, the hinge loss function $\phi_h(\alpha) = c\max\{0, 1-\alpha\}$, $c > 0$, satisfies the condition (37).

Proof. See Appendix D. □

In causal policy learning, since $\eta(x)$ corresponds to the propensity score $e(x)$, $-\omega_+(x)\eta(x) + \omega_-(x)\left(1-\eta(x)\right)$ in (37) coincides with $E_P[Z(-1) - Z(1) \mid X = x]$, the conditional average causal effect between $D = -1$ and $D = 1$.

Theorem 7.1 leads to a generalized Zhang's (2004) inequality for the weighted classification as follows.

Corollary 7.2. *For any distribution P of $(, Y, X)$ and any ϕ satisfying the condition (37),*

$$c(R^{w}(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R^{w}(f)) \leq R_{\phi}^{w}(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R_{\phi}^{w}(f) \tag{38}$$

holds for any $f \in \mathcal{F}_{\mathcal{G}}$.

Proof. See Appendix D. □

Remark 7.3. *Table 2 shows the forms of $\Delta C_\phi^w(\omega_+, \omega_-, \eta)$ for the hinge loss, exponential loss, logistic loss, quadratic loss, and truncated quadratic loss functions. We use $\mu_+ \equiv \omega_+\eta$ and $\mu_- \equiv \omega_-(1-\eta)$. None of them except for the hinge loss satisfies condition (37). That is, hinge losses have special status also in weighted classification, as they are the only surrogate losses that preserve the classification risk.*

Table 2: Surrogate loss functions and their forms of ΔC_ϕ^w

Loss function	$\phi(\alpha)$	$\Delta C_\phi^w(\omega_+, \omega_-, \eta)$
0-1 loss	$1\{\alpha \leq 0\}$	$-\mu_+ + \mu_-$
Hinge loss	$c\max\{0, 1-\alpha\}$	$c(-\mu_+ + \mu_-)$
Exponential loss	$e^{-\alpha}$	$\begin{cases} (\sqrt{\mu_+} - \sqrt{\mu_-})^2 & \text{if } \mu_+ \leq \mu_- \\ -(\sqrt{\mu_+} - \sqrt{\mu_-})^2 & \text{if } \mu_+ > \mu_- \end{cases}$
Logistic loss	$\log(1+e^{-\alpha})$	$\begin{cases} -\mu_+ \log\left(\frac{2\mu_+}{\mu_+ + \mu_-}\right) - \mu_- \log\left(\frac{2\mu_-}{\mu_+ + \mu_-}\right) & \text{if } \mu_+ \leq \mu_- \\ \mu_+ \log\left(\frac{2\mu_+}{\mu_+ + \mu_-}\right) + \mu_- \log\left(\frac{2\mu_-}{\mu_+ + \mu_-}\right) & \text{if } \mu_+ > \mu_- \end{cases}$
Quadratic loss	$(1-\alpha)^2$	$\begin{cases} \frac{(\mu_+ - \mu_-)^2}{\mu_+ + \mu_-} & \text{if } \mu_+ \leq \mu_- \\ -\frac{(\mu_+ - \mu_-)^2}{\mu_+ + \mu_-} & \text{if } \mu_+ > \mu_- \end{cases}$
Truncated quadratic loss	$(\max\{0, 1-\alpha\})^2$	$\begin{cases} \frac{(\mu_+ - \mu_-)^2}{\mu_+ + \mu_-} & \text{if } \mu_+ \leq \mu_- \\ -\frac{(\mu_+ - \mu_-)^2}{\mu_+ + \mu_-} & \text{if } \mu_+ > \mu_- \end{cases}$

Note: $\mu_+ = \omega_+\eta$ and $\mu_- = \omega_-(1-\eta)$.

Similarly to the analysis in Section 4, we consider adding functional form restrictions to the class of classifiers $\mathcal{F}_{MG}$. Let $\widetilde{\mathcal{F}}_{\mathcal{G}}$ be a subclass of $\mathcal{F}_{\mathcal{G}}$, functions in which may be constrained in form. In what follows, we suppose that the weight variable ω satisfies the following condition.

Condition 7.4 (Bounded weight variable)**.** *There exists $M < \infty$ such that the support of the weight variable ω is contained in $[-M, M]$.*

This condition requires a bounded support of ω. In the causal policy learning, Condition 7.4 holds if the outcome variable Z has a bounded support and the propensity score $e(x)$ satisfies the strict overlap condition. For example, if the support of Z is contained in $[-\tilde{M}, \tilde{M}]$, with some $\tilde{M} < \infty$, and the propensity score satisfies $\kappa < e(x) < 1-\kappa$ for some $\kappa \in (0, 1/2)$ and all $x \in \mathcal{X}$, then the support of the weight variable for the causal policy learning ω_p is contained in $[-\tilde{M}/\kappa, \tilde{M}/\kappa]$.

The following theorem, which is analogous to Theorem 4.4, shows that the two conditions (18) and (19) in Theorem 4.4 remain sufficient for $\widetilde{\mathcal{F}}_{\mathcal{G}}$ to guarantee the consistency of the hinge risk minimization approach to weighted classification.

Theorem 7.5. *Suppose that $\widetilde{\mathcal{F}}_{\mathcal{G}} \subset \mathcal{F}_{\mathcal{G}}$ satisfy the conditions (18) and (19) in Theorem 4.4 and that the weight variable ω satisfies condition 7.4.*
(i) $\tilde{f}^* \in \arg\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R^w_{\phi_h}(f)$ minimizes the weighted-classification risk $R^w(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$.
(ii) For $G^*, \tilde{G}^* \in \mathcal{G}^*$ such that $G^* \subseteq \tilde{G}^*$, $\tilde{f}^\dagger(\cdot) = 1\{\cdot \in G^*\} - 1\{\cdot \notin \tilde{G}^*\}$ is a minimizer of $R^w_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$.

Proof. See Appendix D. □

Then the similar relationship between the $\mathcal{G}$-constrained excess weighted-classification risk and $\mathcal{F}_{\mathcal{G}}$-constrained excess weighted-hinge risk as in Corollary 4.5 is obtained.

Corollary 7.6. *Assume $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is a subclass of $\mathcal{F}_{\mathcal{G}}$ and satisfies the conditions (18) and (19) in Theorem 4.4. If ϕ satisfies condition (37) with ϕ_2 replaced with ϕ, there holds*

$$c(R^w(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R^w(f)) = \frac{1}{2}\left(R^w_\phi(f) - \inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G}}} R^w_\phi(f)\right) + \frac{1}{2}\left(R^w_\phi(1\{\cdot \in G_f\} - 1\{\cdot \notin G_f\}) - R^w_\phi(f)\right)$$

for any $f \in \mathcal{F}_{\mathcal{G}}$.

Proof. See Appendix D. □

7.2 Statistical property for the weighted classification with hinge loss

This section extends the analysis of Section 5 to weighted classification with the hinge losses. Let $\{(\omega_i, Y_i, X_i) : i = 1, \ldots, n\}$ be a sample of observations that are independently and identically distributed (i.i.d) as (ω, Y, X). Given the sample, the empirical weighted classification risk and hinge risk for a classifier f are defined as

$$\hat{R}^w(f) \equiv n^{-1} \sum_{i=1}^{1} \omega_i 1\{Y_i \cdot \operatorname{sign}(f(X_i)) \leq 0\},$$
$$\hat{R}^w_{\phi_h}(f) \equiv n^{-1} \sum_{i=1}^{1} \omega_i \max\{0, 1 - Y_i f(X_i)\},$$

respectively. Let $\check{\mathcal{F}}$ be a subclass of $\mathcal{F}_{\mathcal{G}}$, on which we learn a best classifier, and $\widetilde{\mathcal{F}}_{\mathcal{G}}$ be a constrained-classification-preserving reduction of $\mathcal{F}_{\mathcal{G}}$.

As an analogue of Theorems 5.1 and 5.2, the following theorem gives general upper bounds for the mean of the $\mathcal{G}$-constrained excess weighted classification risk.

Theorem 7.7. *Suppose that $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is a subclass of $\mathcal{F}_{\mathcal{G}}$ and satisfy the conditions (18) and (19) in Theorem 4.4. Let $\hat{f} \in \arg\inf_{f\in\check{\mathcal{F}}} \hat{R}^w_{\phi_h}(f)$, and $(q_n, \tau_n, L_C(r,n))$ be as in Theorem 5.1.*

(i) Let $\mathcal{P}$ be a class of distributions of (ω, Y, X) such that, for any distribution $P \in \mathcal{P}$, the condition 7.4 holds and there exist positive constants C and r for which the condition (24) holds for all $\epsilon > 0$ or the condition (25) holds for all $\epsilon > 0$. Then if $\check{\mathcal{F}}$ coincides with $\widetilde{\mathcal{F}}_{\mathcal{G}}$, there holds

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[R^w(\hat{f}) - \inf_{f\in\mathcal{F}_{\mathcal{G}}} R^w(f)\right] \leq ML_C(r,n). \tag{39}$$

(ii) Suppose that $\mathcal{P}$ is a class of distributions of (ω, Y, X) such that, for any distribution $P \in \mathcal{P}$, the condition 7.4 holds and there exist positive constants C' and r' for which the condition (29) holds for all $\epsilon > 0$. Then the following holds:

$$\begin{aligned}\sup_{P\in\mathcal{P}} E_{P^n}\left[R^w(\hat{f}) - \inf_{f\in\mathcal{F}_{\mathcal{G}}} R^w(f)\right] \leq{}& ML_{C'}(r',n) + \frac{1}{2}\left(\inf_{f\in\check{\mathcal{F}}} R^w_{\phi_h}(f) - \inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}} R^w_{\phi_h}(f)\right) \\ &+ \frac{1}{2}\left(R^w_{\phi_h}(1\left\{\cdot \in G_{\hat{f}}\right\} - 1\left\{\cdot \notin G_{\hat{f}}\right\}) - R^w_{\phi_h}(\hat{f})\right).\end{aligned} \tag{40}$$

Proof. See Appendix D. □

Similar comments as in Remark 5.3 apply to Theorem 7.7. The two approximation errors in (40) are small as $\check{\mathcal{F}}$ approximates $\widetilde{\mathcal{F}}_{\mathcal{G}}$ well. When $\check{\mathcal{F}}$ coincides with $\widetilde{\mathcal{F}}_{\mathcal{G}}$, they disappear.

7.3 Monotone weighted classification

Finally, we extend the results for the monotone classification in Section 6 to the weighted classification. Let $\mathcal{F}_{\mathcal{G}_M}$, $\mathcal{F}_M$, and $\mathcal{B}_{\mathbf{k}}$ be as in Section 6, and suppose $\mathcal{X} = [0,1]^{d_x}$. Our aim is to find a best classifier that minimizes $R^w(\cdot)$ over $\mathcal{F}_{\mathcal{G}_M}$. We again consider to use the whole class of monotone classifiers $\mathcal{F}_M$ and sieve of Bernstein polynomials $\mathcal{B}_{\mathbf{k}}$ in the empirical hinge risk minimization for the weighted classification. The following theorems show statistical properties of these methods.

Theorem 7.8. *Let $\mathcal{P}$ be a class of distributions of (ω, Y, X) such that the condition 7.4 holds for any $P \in \mathcal{P}$ and that for any $P \in \mathcal{P}$ the marginal distribution P_X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and has a density that is bounded from above by some finite constant $A > 0$. Let q_n and τ_n be as in Theorem 6.2, and let*

$\hat{f}_M \in \arg\inf_{f\in\mathcal{F}_M} \hat{R}^w_{\phi_h}(f)$. *Then the following holds:*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[R^w(\hat{f}_M) - \inf_{f\in\mathcal{F}_{\mathcal{G}_M}} R^w(f)\right] \leq \begin{cases} 2MD_1\tau_n + 4MD_2\exp\left(-D_1^2 q_n^2\right) & \text{if } d_x \geq 2 \\ 2MD_3\tau_n + 2Mn^{-1}D_4 & \text{if } d_x = 1 \end{cases} \tag{41}$$

for some positive constants D_1 and D_2, which depend only on d_x and A.

Proof. Since $\mathcal{F}_M$ satisfies the conditions (18) and (19) in Theorem 4.4 with $\mathcal{G}$ being $\mathcal{G}_M$ (Example 4.7), the result follows from Theorem 7.7 (i) and Lemma 6.1. □

Theorem 7.9. *Let $\mathcal{P}$ be a class of distributions of (ω, Y, X) that satisfies the same conditions as in Theorem 7.8. Let $\tilde{q}_n$ and $\tilde{\tau}_n$ be as in Theorem 6.4 . Then, for $\hat{f}_B \in \arg\inf_{f\in\mathbf{B_k}} \hat{R}^w_{\phi_h}(f)$ such that its coefficients take values in $\{-1, 1\}$, the following holds:*

$$\sup_{P\in\mathcal{P}} E_{P^n}\left[R^w(\hat{f}_B) - \inf_{f\in\mathcal{F}_{\mathcal{G}_M}} R^w(f)\right] \leq 2MD_1\tilde{\tau}_n + 4MD_2\exp\left(-D_1^2\tilde{q}_n^2\right) + 4MA\sum_{j=1}^{d_x}\sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x}\frac{8M}{\sqrt{k_j}},$$

where D_1 and D_2 are the same constants as in Theorem 7.8, which depend only on d_x and A.

Proof. The result follows by combining Theorem 7.7 (ii), Lemma C.1 in Appendix C, and Lemma D.4 in Appendix D. □

Similar comments apply to Theorems 7.8 and 7.9 as those in Section 6. Using $\mathcal{F}_M$ leads to the faster convergence rate than using the Bernstein polynomials $\mathcal{B}_\mathbf{k}$. When using the Bernstein polynomials $\mathcal{B}_\mathbf{k}$, to achieve the convergence rate of $\tilde{\tau}_n$ for the excess risk, Theorem 7.9 suggests to set the tuning parameters k_j, $j = 1, \ldots, d_x$, sufficiently large so that $\sqrt{\log k_j / k_j} = O(\tilde{\tau}_n)$. Furthermore, for any $\hat{f}_B \in \arg\inf_{f\in\mathbf{B_k}} \hat{R}^w_{\phi_h}(f)$, the modification $\hat{f}_B^\dagger$ introduced in Remark 6.5 minimizes the empirical hinge risk $\hat{R}^w_{\phi_h}(\cdot)$, while satisfying the coefficient restriction in Theorem 7.9 (see Lemma D.2 (iii) in Appendx D). The weighted hinge risk minimization problems using $\mathcal{F}_M$ and $\mathcal{B}_\mathbf{k}$ can be formulated as linear programmings similar to those in Section 6.2 and Remark 6.6, respectively.

8 Conclusion

This paper studies consistency of surrogate risk minimization approaches to classification and weighted classification under a constrained set of classifiers, where the latter includes

policy learning for individualized treatment assignment rules. Our focus is on how surrogate risk minimizing classifiers behave if the constrained class of classifiers fails the correct specification assumption. Our first main result shows that when the constraint restricts classifiers' prediction sets only, the hinge losses are the only loss functions that secure consistency of the surrogate-risk minimizing classifier without the correct specification assumption. When the constraint additionally restricts the functional form of the classifiers, the surrogate risk minimizing classifier cannot be generally consistent even with the hinge loss. Our second main result is to show that in this case the condition of constrained-classification-preserving reduction becomes a sufficient condition for the consistency of the hinge-risk minimizing classifier.

The paper also investigates statistical properties of the hinge risk minimizing classifiers in terms of the uniform upper bounds of the excess regret. We illustrate usefulness and implications of our theoretical results in monotone classification. Exploiting the hinge loss and the class of monotone classifiers, we show that the empirical surrogate-risk minimizing classifier can be computed by linear programming. All of the results obtained in the standard classification setting are naturally extended to the weighted classification problem, so our contributions carry over to its important application to causal policy learning.

Appendix

A Proofs of the results in Sections 3 and 4

In this appendix, we provide the proofs of our main results in Sections 3 and 4 with some auxiliary lemmas. We here let ϕ be any surrogate loss function. Before proceeding to the proofs, we note that if ϕ is classification-calibrated, ΔC_ϕ has the same sign with the Bayes classifier:

$$\Delta C_\phi\left(\eta\left(x\right)\right)\begin{cases} >0 & \text{if } \eta\left(x\right)>1/2 \\ <0 & \text{if } \eta\left(x\right)<1/2 \end{cases}, \tag{42}$$

which will be used in the following proofs.

Proof of Theorem 3.6.
("if" part)

For any $G_1, G_2 \in \mathcal{G}$, we have shown in Theorem 3.2 that $\mathcal{R}_{\phi_1}\left(G_2\right) \geq \mathcal{R}_{\phi_1}\left(G_1\right)$ is

equivalent to

$$\int_{G_2\backslash G_1} \Delta C_{\phi_1}(\eta(x))\, dP_X(x) \geq \int_{G_1\backslash G_2} \Delta C_{\phi_1}(\eta(x))\, dP_X(x).$$

The inequality does not change if we replace $\Delta C_{\phi_1}(\eta(x))$ by $\Delta C_{\phi_2}(\eta(x)) = c\Delta C_{\phi_1}(\eta(x))$ with $c > 0$. Furthermore, the above inequality with $\Delta C_{\phi_1}(\eta(x))$ replaced by $\Delta C_{\phi_2}(\eta(x))$ is equivalent to $\mathcal{R}_{\phi_2}(G_2) \geq \mathcal{R}_{\phi_2}(G_1)$ from Theorem 3.2. Therefore if $\Delta C_{\phi_2}(\cdot) = c\Delta C_{\phi_1}(\cdot)$ with $c > 0$, $\phi_1 \overset{u}{\sim} \phi_2$ holds.

("only if" part)
We prove the "only if" part of the theorem by exploiting a specific class of data generating processes (DGPs). Suppose $\mathcal{X} = \{1, 2\}$ and $\mathcal{G} = \{\emptyset, G_1, G_2, \mathcal{X}\}$ with $G_1 = \{1\}$ and $G_2 = \{2\}$. Let $\alpha = P(X = 1)\,(= 1 - P(X = 2))$ and $(\eta_1, \eta_2) = (\eta(1), \eta(2))$. The DGP varies depending on the values of $(\alpha, \eta_1, \eta_2) \in [0, 1]^3$.

In what follows, we will show that

$$\frac{\Delta C_{\phi_1}(\eta_1)}{\Delta C_{\phi_1}(\eta_2)} = \frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_2}(\eta_2)}$$

holds for any $(\eta_1, \eta_2) \in ([0, 1]\backslash\{1/2\})^2$. Then, applying Lemma A.1 below proves the "only if" part of the theorem.

Let $G \in \mathcal{G}$. In the current setting, $\mathcal{R}_\phi(G)$ can be written as

$$\begin{aligned}
\mathcal{R}_\phi(G) &= P(X = 1)\,\Delta C_\phi(\eta_1)\,1\{1 \in G\} + P(X = 2)\,\Delta C_\phi(\eta_2)\,1\{2 \in G\} \\
&\quad + \sum_{x=1}^{2} P(X = x)\, C_\phi^-(\eta(x)) \\
&= \alpha\Delta C_\phi(\eta_1)\,1\{1 \in G\} + (1 - \alpha)\,\Delta C_\phi(\eta_2)\,1\{2 \in G\} + C_{\alpha,\eta_1,\eta_2},
\end{aligned}$$

where $C_{\alpha,\eta_1,\eta_2} \equiv \alpha\Delta C_\phi(\eta_1) + (1 - \alpha)\,\Delta C_\phi(\eta_2)$ which does not depend on G. Thus, we have

$$\begin{aligned}
\mathcal{R}_\phi(\emptyset) &= C_{\alpha,\eta_1,\eta_2} \\
\mathcal{R}_\phi(G_1) &= \alpha\Delta C_\phi(\eta_1) + C_{\alpha,\eta_1,\eta_2}, \\
\mathcal{R}_\phi(G_2) &= (1 - \alpha)\,\Delta C_\phi(\eta_2) + C_{\alpha,\eta_1,\eta_2}, \\
\mathcal{R}_\phi(\mathcal{X}) &= \alpha\Delta C_\phi(\eta_1) + (1 - \alpha)\,\Delta C_\phi(\eta_2) + C_{\alpha,\eta_1,\eta_2}.
\end{aligned}$$

In what follows, we separately consider four cases: (i) $\eta_1 > 1/2$ and $\eta_2 > 1/2$; (ii) $\eta_1 < 1/2$ and $\eta_2 < 1/2$; (iii) $\eta_1 < 1/2$ and $\eta_2 > 1/2$; (iv) $\eta_1 > 1/2$ and $\eta_2 < 1/2$.

First, we consider the case (i): $\eta_1 > 1/2$ and $\eta_2 > 1/2$. Because we assume $\phi_1 \overset{u}{\sim} \phi_2$,

$$\mathcal{R}_{\phi_1}(G_1) \leq \mathcal{R}_{\phi_1}(G_2) \Leftrightarrow \mathcal{R}_{\phi_2}(G_1) \leq \mathcal{R}_{\phi_2}(G_2),$$

holds for any $(\alpha, \eta_1, \eta_2) \in (0,1) \times (1/2,1]^2$. This is equivalent to

$$\alpha \Delta C_{\phi_1}(\eta_1) \leq (1-\alpha) \Delta C_{\phi_1}(\eta_2) \Leftrightarrow \alpha \Delta C_{\phi_2}(\eta_1) \leq (1-\alpha) \Delta C_{\phi_2}(\eta_2)$$

for any $(\alpha, \eta_1, \eta_2) \in (0,1) \times (1/2,1]^2$. Let $\gamma^+ \equiv (1-\alpha)/\alpha$, which may take any value in $(0,+\infty)$ by varying α on $(0,1)$. From the classification-calibrated property (42), both $\Delta C_{\phi_1}(\eta)$ and $\Delta C_{\phi_2}(\eta)$ are positive for $\eta \in (1/2,1]$. Thus, it follows for any $(\gamma^+, \eta_1, \eta_2) \in (0,+\infty) \times (1/2,1]^2$ that

$$\frac{\Delta C_{\phi_1}(\eta_1)}{\Delta C_{\phi_1}(\eta_2)} \leq \gamma^+ \Leftrightarrow \frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_2}(\eta_2)} \leq \gamma^+, \tag{43}$$

where both $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2)$ and $\Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ are positive. Since (43) holds for any value of $\eta^+ \in (0,+\infty)$, $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ holds for any $(\eta_1, \eta_2) \in (1/2,1]^2$.

Similarly, in the case (ii): $\eta_1 < 1/2$ and $\eta < 1/2$, the following equivalence holds for any $(\gamma^+, \eta_1, \eta_2) \in (0,+\infty) \times [0,1/2)^2$:

$$\frac{\Delta C_{\phi_1}(\eta_1)}{\Delta C_{\phi_1}(\eta_2)} \geq \gamma^+ \Leftrightarrow \frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_2}(\eta_2)} \geq \gamma^+, \tag{44}$$

where both $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2)$ and $\Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ are positive. Thus, varying the value of γ^+ on $(0,+\infty)$ in (44) shows that $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ holds for any $(\eta_1, \eta_2) \in [0,1/2)^2$.

Next, we consider the case (iii): $\eta_1 < 1/2$ and $\eta_2 > 1/2$. Because we assume $\phi_1 \overset{u}{\sim} \phi_2$, it follows for any $(\alpha, \eta_1, \eta_2) \in (0,1) \times [0,1/2) \times (1/2,1]$ that

$$\mathcal{R}_{\phi_1}(\emptyset) \leq \mathcal{R}_{\phi_1}(\mathcal{X}) \Leftrightarrow \mathcal{R}_{\phi_2}(\emptyset) \leq \mathcal{R}_{\phi_2}(\mathcal{X}),$$

which is equivalent to

$$0 \leq \alpha \Delta C_{\phi_1}(\eta_1) + (1-\alpha) \Delta C_{\phi_1}(\eta_2) \Leftrightarrow 0 \leq \alpha \Delta C_{\phi_2}(\eta_1) + (1-\alpha) \Delta C_{\phi_2}(\eta_2).$$

Let $\gamma^- \equiv (\alpha-1)/\alpha$, which takes any value in $(-\infty, 0)$ by varying the value of α on $(0,1)$. Because $\Delta C_{\phi_1}(\eta_1) < 0$ and $\Delta C_{\phi_2}(\eta_2) > 0$ hold for $(\eta_1, \eta_2) \in [0,1/2) \times (1/2,1]$ due to the

classification-calibrated property (42), it follows that

$$\frac{\Delta C_{\phi_1}(\eta_1)}{\Delta C_{\phi_1}(\eta_2)} \geq \gamma^- \Leftrightarrow \frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_2}(\eta_2)} \geq \gamma^- \tag{45}$$

for any $(\gamma^-, \eta_1, \eta_2) \in (-\infty, 0) \times [0, 1/2) \times (1/2, 1]$, where both $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2)$ and $\Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ are also negative. Thus, varying the value of γ^- on $(-\infty, 0)$ in (45) shows that $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ holds for any $(\eta_1, \eta_2) \in [0, 1/2) \times (1/2, 1]$.

Similarly, in the case (iv): $\eta_1 > 1/2$ and $\eta_2 < 1/2$, the following equivalence holds for any $(\gamma^-, \eta_1, \eta_2) \in (-\infty, 0) \times (1/2, 1] \times [0, 1/2)$:

$$\frac{\Delta C_{\phi_1}(\eta_1)}{\Delta C_{\phi_1}(\eta_2)} \leq \gamma^- \Leftrightarrow \frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_2}(\eta_2)} \leq \gamma^-, \tag{46}$$

where both $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2)$ and $\Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ are negative. Therefore, varying the value of γ^- in (46) shows that $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ holds for any $(\eta_1, \eta_2) \in (1/2, 1] \times [0, 1/2)$.

Combining these four results, we have $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ for any $(\eta_1, \eta_2) \in ([0, 1] \setminus \{1/2\})^2$. Then the proof follows from Lemma A.1 below. □

Lemma A.1. *Let ϕ_1 and ϕ_2 be classification-calibrated loss functions. If $\Delta C_{\phi_1}(\eta_1)/\Delta C_{\phi_1}(\eta_2) = \Delta C_{\phi_2}(\eta_1)/\Delta C_{\phi_2}(\eta_2)$ holds for $(\eta_1, \eta_2) \in ([0, 1] \setminus \{1/2\})^2$, then there exists some constant $c > 0$ such that $\Delta C_{\phi_2}(\eta) = c\Delta C_{\phi_1}(\eta)$ for $\eta \in [0, 1]$.*

Proof. For $\eta \in [0, 1] \setminus \{1/2\}$, let $c(\eta)$ be a value such that

$$\Delta C_{\phi_2}(\eta) = c(\eta)\, \Delta C_{\phi_1}(\eta). \tag{47}$$

Because ϕ_1 and ϕ_2 are classification-calibrated, $c(\eta)$ must be positive from (42). We will show that $c(\eta)$ is constant over $\eta \in [0, 1] \setminus \{1/2\}$ by contradiction.

Suppose there exists $(\eta_1, \eta_2) \in ([0, 1] \setminus \{1/2\})^2$ such that $c(\eta_1) \neq c(\eta_2)$. From the assumption, the following equations hold

$$\Delta C_{\phi_2}(\eta_1) = \left(\frac{\Delta C_{\phi_2}(\eta_2)}{\Delta C_{\phi_1}(\eta_2)}\right) \Delta C_{\phi_1}(\eta_1),$$
$$\Delta C_{\phi_2}(\eta_2) = \left(\frac{\Delta C_{\phi_2}(\eta_1)}{\Delta C_{\phi_1}(\eta_1)}\right) \Delta C_{\phi_2}(\eta_2).$$

Combining these equations with equation (47), we have $\Delta C_{\phi_2}(\eta_2) = c(\eta_1)\, \Delta C_{\phi_1}(\eta_2)$ and $\Delta C_{\phi_2}(\eta_2) = c(\eta_2)\, \Delta C_{\phi_1}(\eta_2)$. However, this contradicts that $c(\eta_1) \neq c(\eta_2)$. There-

fore, $c(\eta)$ must be constant over $\eta \in [0,1] \setminus \{1/2\}$.

When $\eta = 1/2$, $\Delta C_{\phi_1}(\eta) = \Delta C_{\phi_2}(\eta) = 0$ holds by the definition. In this case, $\Delta C_{\phi_2}(\eta) = c\Delta C_{\phi_1}(\eta)$ holds for any c. □

For the proofs of Lemma 4.3 and Theorem 4.4, we introduce some algebraic results. Firstly, the hinge risk $R_{\phi_h}(f)$ has the following expression:

$$\begin{aligned} R_{\phi_h}(f) &= \int_{\mathcal{X}} \left[\eta(x)(1-f(x)) + (1-\eta(x))(1+f(x))\right] dP_X(x) \\ &= \int_{\mathcal{X}} (1-2\eta(x))\, f(x)\, dP_X(x) + 1. \end{aligned} \tag{48}$$

Secondly, for $G \in \mathcal{G}$, $\mathcal{R}(G)$ can be written as

$$\begin{aligned} \mathcal{R}(G) &= \int_{\mathcal{X}} \left[\eta(x) 1\{x \notin G\} + (1-\eta(x)) 1\{x \in G\}\right] dP_X(x) \\ &= \int_{\mathcal{X}} \left[\eta(x) 1\{x \in G^c\} + (1-\eta(x))\left(1 - 1\{x \in G^c\}\right)\right] dP_X(x) \\ &= -\int_{G^c} (1-2\eta(x))\, dP_X(x) + P(Y=-1). \end{aligned} \tag{49}$$

Proof of Lemma 4.3. Fix $\tilde{f} \in \widetilde{\mathcal{F}}_{\mathcal{G},J}$. It has the form of

$$\tilde{f}(x) = \sum_{j=1}^{J} c_j^+ 1\{x \in G_j\} - \sum_{j=1}^{J} c_j^- 1\left\{x \notin \tilde{G}_j\right\} \tag{50}$$

for some $G_J \subseteq \cdots \subseteq G_1 \subseteq \tilde{G}_1 \subseteq \cdots \subseteq \tilde{G}_J$ and $c_j^+, c_j^- \geq 0$ for $j = 1, \ldots, J$ with $\sum_{j=1}^J c_j^+ = \sum_{j=1}^J c_j^- = 1$. Accordingly, from equation (48), the hinge risk of $\tilde{f}$ can be written as

$$\begin{aligned} R_{\phi_h}\left(\tilde{f}\right) = &\sum_{j=1}^{J} \left[\left(c_j^+\right) \int_{G_j} (1-2\eta(x))\, dP_X(x)\right] \\ &+ \sum_{j=1}^{J} \left[\left(-c_j^-\right) \int_{\left(\tilde{G}_j\right)^c} (1-2\eta(x))\, dP_X(x)\right] + 1. \end{aligned} \tag{51}$$

Denote the first and second terms in (51) by

$$R^I_{\phi_h}(\tilde{f}) \equiv \sum_{j=1}^{J}\left[\left(c_j^+\right)\int_{G_j}\left(1-2\eta\left(x\right)\right)dP_X\left(x\right)\right],$$
$$R^{II}_{\phi_h}(\tilde{f}) \equiv \sum_{j=1}^{J}\left[\left(-c_j^-\right)\int_{\left(\tilde{G}_j\right)^c}\left(1-2\eta\left(x\right)\right)dP_X\left(x\right)\right].$$

$R^I_{\phi_h}(\tilde{f})$ can be rewritten as

$$\begin{aligned}R^I_{\phi_h}(\tilde{f}) &= \sum_{j=1}^{J}\left[\left(c_j^+\right)\left(\mathcal{R}\left(G_j\right)-P\left(Y=1\right)\right)\right]\\ &= \sum_{j=1}^{J}\left(c_j^+\right)\mathcal{R}\left(G_j\right)-P\left(Y=1\right),\end{aligned}$$

where the first equality follows from (10) and the second equality follows from $\sum_{j=1}^{J} c_j^+ = 1$. Similarly, $R^{II}_{\phi_h}(\tilde{f})$ can be written as

$$\begin{aligned}R^{II}_{\phi_h}(\tilde{f}) &= \sum_{j=1}^{J}\left[\left(c_j^-\right)\left(\mathcal{R}\left(\tilde{G}_j\right)-P\left(Y=-1\right)\right)\right]\\ &= \sum_{j=1}^{J}\left(c_j^-\right)\mathcal{R}\left(\tilde{G}_j\right)-P\left(Y=-1\right),\end{aligned}$$

where the first equality follows from (49) and the second equality follows from $\sum_{j=1}^{n} c_j^- = 1$.

Combining these expressions, $R_{\phi_h}(\tilde{f})$ can be written as

$$\begin{aligned}R_{\phi_h}(\tilde{f}) &= R^I_{\phi_h}(\tilde{f}) + R^{II}_{\phi_h}(\tilde{f}) + 1\\ &= \sum_{j=1}^{J}\left(c_j^+\right)\mathcal{R}\left(G_j\right)+\sum_{j=1}^{J}\left(c_j^-\right)\mathcal{R}\left(\tilde{G}_j\right).\end{aligned} \tag{52}$$

From this expressions, we can see that $R_{\phi_h}(\tilde{f})$ is bounded from below by $2\mathcal{R}^*$.

Let $G^*, \tilde{G}^* \in \mathcal{G}^*$ such that $G^* \subseteq \tilde{G}^*$, and define $\tilde{f}^\dagger(x) = 1\{x \in G^*\} - 1\{x \notin \tilde{G}^*\}$. $\tilde{f}^\dagger$ can be taken from $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ by setting $G_1 = G^*$ with $c_1^+ = 1$ and $\tilde{G}_1 = \tilde{G}^*$ with $c_1^- = 1$. Then, from (52), $R_{\phi_h}(\tilde{f}^\dagger)$ takes its lower bound $2\mathcal{R}^*$. Thus, $\tilde{f}^\dagger$ minimizes $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. This proves $R_{\phi_h}(\tilde{f}^*) = 2\mathcal{R}^*$ and the statement (ii) of the lemma.

Next, we prove that a minimizer of the hinge risk $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ also minimizes the classification risk $R(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$. To obtain contradiction, suppose $\tilde{f}$ minimizes $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ but does not minimize $R(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$. As $\tilde{f}$ does not minimize the classification

risk $R(\cdot)$, $G_{\tilde{f}} \notin \mathcal{G}^*$ holds. Let m be the smallest number in $\{1, \ldots, J\}$ such that $c_m^- > 0$. Because the corresponding set $\tilde{G}_m$ coincides with $G_{\tilde{f}}$, $\tilde{G}_m \notin \mathcal{G}^*$ holds. Then it follows that

$$\begin{aligned} R_{\phi_h}\left(\tilde{f}\right) &= \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}\left(G_j\right) + \sum_{j=1}^{J} \left(c_j^-\right) \mathcal{R}\left(\tilde{G}_j\right) \\ &= c_m^- \mathcal{R}\left(\tilde{G}_m\right) + \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}\left(G_j\right) + \sum_{j \in \{1,\ldots,m-1,m+1,\ldots,J\}} \left(c_j^-\right) \mathcal{R}\left(\tilde{G}_j\right) \\ &\geq c_m^- \mathcal{R}\left(\tilde{G}_m\right) + \left(2 - c_m^-\right) \mathcal{R}^* \\ &> 2\mathcal{R}^*, \end{aligned}$$

where the last line follows from $c_m^- > 0$ and $\tilde{G}_m \notin \mathcal{G}^*$. $R_{\phi_h}(\tilde{f})$ does not take the minimum value of $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ that is $2\mathcal{R}^*$. This contradicts that $\tilde{f}$ minimizes the hinge risk over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. Thus, $\tilde{f}$ minimizes the classification risk over $\mathcal{F}_{\mathcal{G}}$. □

Proof of Theorem 4.4. Define a class of step functions

$$\begin{aligned} \bar{\mathcal{F}}_J^* \equiv \Bigg\{ & f = \sum_{j=0}^{J} c_j^+ 1\left\{x \in G_j\right\} - \sum_{j=1}^{J} c_j^- 1\left\{x \notin \tilde{G}_j\right\} : \\ & G_j, \tilde{G}_j \in \mathcal{G} \text{ and } c_j^+, c_j^- \geq 0 \text{ for } j = 1, \ldots, J;\ c_1^- > 0; \\ & G_J \subseteq \cdots \subseteq G_1 \subseteq G_{\tilde{f}^*} = \tilde{G}_1 \subseteq \cdots \subseteq \tilde{G}_J;\ \sum_{j=1}^{J} c_j^+ = 1, \sum_{j=1}^{J} c_j^- = 1 \Bigg\}. \end{aligned}$$

Any function in $\bar{\mathcal{F}}_J^*$ has the prediction set corresponding to $G_{\tilde{f}^*}$, i.e., $G_f = G_{\tilde{f}^*}$ for any $f \in \bar{\mathcal{F}}_J^*$. We can find a sequence of functions $\left\{\bar{f}_J^*\right\}_{J=1}^{\infty}$ such that $\bar{f}_J^* \in \bar{\mathcal{F}}_J^*$ for any J and $\bar{f}_J^*(X) \to \tilde{f}^*(X)$ as $J \to \infty$ with probability one. Such a sequence of functions can be made in $\bar{\mathcal{F}}_J^*$ by choosing $G_j = \left\{x : \tilde{f}^*\left(x\right) \geq j/J\right\}$ and $\tilde{G}_j = \left\{x : \tilde{f}^*\left(x\right) \geq -j/J\right\}$ and setting $c_j^+ = c_j^- = 1/J$ for $j = 1, \ldots, J$, i.e.,

$$\bar{f}_J^*(\cdot) \equiv J^{-1} \sum_{j=1}^{J} \frac{1}{J} \left(1\left\{\tilde{f}^*\left(\cdot\right) \geq j/J\right\} - 1\left\{\tilde{f}^*\left(\cdot\right) < -j/J\right\}\right).$$

Then it follows for all $x \in \mathcal{X}$ that

$$\left|\bar{f}_J^*\left(x\right) - \tilde{f}^*\left(x\right)\right| = \left|\sum_{j=1}^{J} \frac{1}{J} \left(1\left\{\tilde{f}^*\left(x\right) \geq j/J\right\} - 1\left\{\tilde{f}^*\left(x\right) < -j/J\right\}\right) - \tilde{f}^*\left(x\right)\right|$$

$$< \frac{1}{J} \to 0 \text{ as } J \to \infty.$$

Thus, $\bar{f}_J^*(X) \to \tilde{f}^*(X)$ holds with probability one.

Then it flows that

$$\begin{aligned} R_{\phi_h}(\tilde{f}^*) &= \int_{\mathcal{X}} (1 - 2\eta(x))\, \tilde{f}^*(x)\, dP_X(x) + 1 \\ &= \lim_{J\to\infty} \int_{\mathcal{X}} (1 - 2\eta(x))\, \bar{f}_J^*(x)\, dP_X(x) + 1 \\ &= \lim_{J\to\infty} R_{\phi_h}\left(\bar{f}_J^*\right) \geq \lim_{J\to\infty} \inf_{\bar{f}\in\bar{\mathcal{F}}_J^*} R_{\phi_h}\left(\bar{f}_J^*\right) \qquad (53) \\ &\geq \lim_{J\to\infty} \inf_{\tilde{f}\in\tilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}(\tilde{f}), \qquad (54) \end{aligned}$$

where the first and third equalities follow from equation (48); the second equality follows from the dominated convergence theorem, which holds because both $(1-2\eta)\,\bar{f}_J^*(X) \to (1-2\eta(X))\,\tilde{f}^*(X)$ and $\left|(1-2\eta(X))\,\bar{f}_J^*(X)\right| < 1$ hold with probability one; the first inequality follows from $\bar{f}_J^* \in \bar{\mathcal{F}}_J^*$; the last inequality follows from $\bar{\mathcal{F}}_J^* \subseteq \tilde{\mathcal{F}}_{\mathcal{G},J}$.

Lemma 4.3 has shown that $\inf_{\tilde{f}\in\tilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}\left(\tilde{f}\right) = 2\mathcal{R}^*$ for any J. Thus, from equation (54), we have

$$R_{\phi_h}\left(\tilde{f}^*\right) \geq \lim_{J\to\infty} \inf_{\tilde{f}\in\tilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}\left(\tilde{f}\right) \geq 2\mathcal{R}^*,$$

which means that the minimal value of $R_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is at least $2\mathcal{R}^*$. Lemma 4.3 has also shown that $\tilde{f}^\dagger$ leads to $R_{\phi_h}(\tilde{f}^\dagger) = 2\mathcal{R}^*$. Therefore, $\tilde{f}^\dagger$ minimizes $R_{\phi_h}(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$, which proves the second statement of the theorem.

Next we will prove the first statement of the theorem by contradiction. Suppose that $\tilde{f}^*$ does not minimize the classification risk $R(\cdot)$ over $\mathcal{F}_{\mathcal{G}}$, or equivalently $G_{\tilde{f}^*} \notin \mathcal{G}^*$. Then, it follows for any $\bar{f} \in \bar{\mathcal{F}}_J^*$ that

$$\begin{aligned} R_{\phi_h}\left(\bar{f}\right) &= \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}\left(G_j\right) + \sum_{j=1}^{J} \left(c_j^-\right) \mathcal{R}\left(\tilde{G}_j\right) \\ &\geq \left(c_1^-\right) \mathcal{R}\left(\tilde{G}_1\right) + \left(2 - c_1^-\right) \mathcal{R}^* \\ &> 2\mathcal{R}^*, \end{aligned}$$

where the last inequality follows from $\tilde{G}_1 = G_{\tilde{f}^*} \notin \mathcal{G}^*$ and $c_1^- > 0$. Therefore, we have from equation (53) that

$$R_{\phi_h}\left(\tilde{f}^*\right) \geq \lim_{J\to\infty} \inf_{\bar{f}\in\bar{\mathcal{F}}_J^*} R_{\phi_h}\left(\bar{f}\right) > 2\mathcal{R}^*.$$

This contradicts that $\tilde{f}^*$ minimizes $R_\phi(\cdot)$ over $\tilde{\mathcal{F}}_\mathcal{G}$ because, as we have seen, $\tilde{f}^\dagger$ achieves $R_{\phi_h}(\tilde{f}^\dagger) = 2\mathcal{R}^*$. Therefore, $G_{\tilde{f}^*} \in \mathcal{G}^*$ must hold. □

Proof of Corollar 4.5. By equations (10) and (49), $cR(f)$ can be written as

$$cR(f) = \frac{1}{2}\left\{\int_\mathcal{X} c\left(1-2\eta\left(x\right)\right)\left(1\left\{x \in G_f\right\} - 1\left\{x \notin G_f\right\}\right) dP_X\left(x\right) + c\right\}.$$

By equation (48), the terms in the curly brackets equal

$$R_\phi\left(1\left\{\cdot \in G_f\right\} - 1\left\{\cdot \notin G_f\right\}\right).$$

Combining this result with Theorem 4.4 (i) leads to equation (20).

When $\widetilde{\mathcal{F}}_\mathcal{G}$ coincides with $\mathcal{F}_\mathcal{G}$, by equation (20) and Corollary 3.4,

$$R_\phi(1\left\{\cdot \in G_f\right\} - 1\left\{\cdot \notin G_f\right\}) - R_\phi(f) \leq R_\phi(f) - \inf_{f\in\mathcal{F}_\mathcal{G}} R_\phi(f)$$

holds. Combining this with equation (20) leads to the second result. □

B Proof of Theorems 5.1 and 5.2

This appendix provides the proof of the results in Section 5 with some auxiliary results. The results below are related to the theory of empirical processes. We refer to Alexander (1984), Mammen and Tsybakov (1999), Tsybakov (2004), and Mbakop and Tabord-Meehan (2021) for the general strategy of the proof.

Given $G \in \mathcal{G}$, let $\mathbf{1}_G$ be an indicator function from $\mathcal{X}$ such that $\mathbf{1}_G(x) = 1\{x \in G\}$. We first give the definition of the bracketing entropy for a class of functions and a class of sets.

Definition B.1 (Bracketing entropy)**.** *(i) Let $\mathcal{F}$ be a class of functions on $\mathcal{X}$. For $f \in \mathcal{F}$, let $\|f\|_{p,Q} := \left(\int_\mathcal{X} |f\left(x\right)|^p \, dQ\left(x\right)\right)^{1/p}$. $\|\cdot\|_{p,Q}$ is the $L_p\left(Q\right)$-metric on $\mathcal{X}$, where Q is a measure on $\mathcal{X}$. Given a pair of functions (f_1, f_2) with $f_1 \leq f_2$, let $[f_1, f_2] := \{f \in \mathcal{F} : f_1 \leq f \leq f_2\}$ be the bracket. Given $\epsilon > 0$, let $N_p^B\left(\epsilon, \mathcal{F}, Q\right)$ be the smallest k such that there exist pairs of functions $\left(f_j^L, f_j^U\right)$, $j = 1, \ldots, k$, with $f_j^L \leq f_j^U$ that satisfy $\left\|f_j^U - f_j^L\right\|_{p,Q} < \epsilon$ and*

$$\mathcal{F} \subseteq \cup_{j=1}^k \left[f_j^L, f_j^U\right].$$

We call $N_p^B(\epsilon, \mathcal{F}, Q)$ *the* $L_p(Q)$*-bracketing number of* $\mathcal{F}$*, and* $H_p^B(\epsilon, \mathcal{F}, Q) \equiv \log N_p^B(\epsilon, \mathcal{F}, Q)$ *the* $L_p(Q)$*-bracketing entropy of* $\mathcal{F}$*. We also refer* $[f_j^L, f_j^U]$ *to the* ϵ*-bracket with respect to* $L_p(Q)$ *if and only if* $\left\|f_j^U - f_j^L\right\|_{p,Q} < \epsilon$ *holds.*
(ii) Given a class of sets $\mathcal{G} \subseteq \mathcal{X}$*, let* $\mathcal{H}_\mathcal{G} \equiv \{\mathbf{1}_G : G \in \mathcal{G}\}$*. We define* $H_p^B(\epsilon, \mathcal{G}, Q) \equiv H_p^B(\epsilon, \mathcal{H}_\mathcal{G}, Q)$ *and call it the* $L_p(Q)$*-bracketing entropy of* $\mathcal{G}$*.*

Note that in the definition of $N_p^B(\epsilon, \mathcal{F}, Q)$, the functions f_j^L and f_j^U do not have to belong to $\mathcal{F}$. Note also that if $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$, $H_p^B(\epsilon, \mathcal{F}, Q) \leq H_p^B\left(\epsilon, \widetilde{\mathcal{F}}, Q\right)$ holds. When $\mathbf{1}_G \in \mathcal{F}$ for all $G \in \mathcal{G}$, $H_p^B(\epsilon, \mathcal{G}, Q) \leq H_p^B(\epsilon, \mathcal{F}, Q)$ holds.

The following theorem gives a finite-sample upper bound for the mean of the estimation error in Section 5, auxiliary results of which are provided below.

Theorem B.2. *Let* $\check{\mathcal{F}}$ *a class of classifiers whose members satisfy* $-1 \leq f \leq 1$*. Suppose that* $\mathcal{P}$ *is a class of distributions on* $\{-1, 1\} \times \mathcal{X}$ *such that there exist positive constants* C *and* r *for which*

$$H_1^B\left(\epsilon, \check{\mathcal{F}}, P_X\right) \leq C\epsilon^{-r}$$

holds for any $P \in \mathcal{P}$ *and* $\epsilon > 0$*. Let* q_n *and* τ_n *be as in Theorem 5.1. Let* $\hat{f}$ *minimizes* $\hat{R}_{\phi_h}(\cdot)$ *over* $\check{\mathcal{F}}$*. Then the following holds:*

$$\sup_{P \in \mathcal{P}} E_{P^n}\left[R_{\phi_h}\left(\hat{f}\right) - \inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f)\right] \leq \begin{cases} 4D_1\tau_n + 8D_2\exp\left(-D_1^2 q_n^2\right) & \text{for } r \geq 1 \\ 4D_3\tau_n + 4n^{-1}D_4 & \text{for } 0 < r < 1 \end{cases},$$

for some positive constants D_1, D_2, D_3, D_4*, which depend only on* C *and* r*.*

Proof. Fix $P \in \mathcal{P}$. Let $\check{f}^*$ minimizes $R_{\phi_h}(\cdot)$ over $\check{\mathcal{F}}$. Define a class of functions $\dot{\mathcal{F}} = \left\{(f+1)/2 : f \in \check{\mathcal{F}}\right\}$, which normalizes $\check{\mathcal{F}}$ so that $0 \leq f \leq 1$ for all $f \in \dot{\mathcal{F}}$.

A standard argument gives

$$\begin{aligned} E_{P^n}\left[R_{\phi_h}\left(\hat{f}\right) - \inf_{f \in \check{\mathcal{F}}} R_{\phi_h}(f)\right] \leq & E_{P^n}\left[R_{\phi_h}\left(\hat{f}\right) - \hat{R}_{\phi_h}\left(\hat{f}\right) + \hat{R}_{\phi_h}\left(\check{f}^*\right) - R_{\phi_h}\left(\check{f}^*\right)\right] \\ & \left(\because \hat{R}_{\phi_h}\left(\hat{f}\right) \leq \hat{R}_{\phi_h}\left(\check{f}^*\right)\right) \\ = & 2E_{P^n}\left[R_{\phi_h}\left(\frac{\hat{f}+1}{2}\right) - \hat{R}_{\phi_h}\left(\frac{\hat{f}+1}{2}\right)\right] \\ & + 2E_{P^n}\left[\hat{R}_{\phi_h}\left(\frac{\check{f}^*+1}{2}\right) - R_{\phi_h}\left(\frac{\check{f}^*+1}{2}\right)\right] \\ \leq & 4\sup_{f \in \dot{\mathcal{F}}} E_{P^n}\left[\left|R_{\phi_h}(f) - \hat{R}_{\phi_h}(f)\right|\right]. \end{aligned} \tag{55}$$

Since $R_{\phi_h}(f) - \hat{R}_{\phi_h}(f)$ can be seen as the centered empirical process indexed by $f \in \dot{\mathcal{F}}$, we can apply results in empirical process theory to (55) to obtain a finite-sample upper bound for the mean of the excess hinge risk.

We follow the general strategy of Theorem 1 in Mammen and Tsybakov (1999) and Proposition B.1 in Mbakop and Tabord-Meehan (2021). Note that

$$\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|R_{\phi_h}(f) - \hat{R}_{\phi_h}(f)\right|\right] = \sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|E_P\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right|\right] \tag{56}$$

and that

$$\sup_{f\in\dot{\mathcal{F}}}\left|E_P\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right| \leq 2$$

with probability one.

We first prove the result for the case of $r \geq 1$. For any $f \in \dot{\mathcal{F}}$ and $D > 0$,

$$\begin{aligned}
&\frac{\sqrt{n}}{q_n}\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|E\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right|\right] \\
&\leq D + \frac{2\sqrt{n}}{q_n} P^n\left(\sup_{f\in\dot{\mathcal{F}}}\frac{\sqrt{n}}{q_n}\left|E\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right| > D\right).
\end{aligned}$$

We consider to apply Corollary B.3. Set $Z = (Y, X)$, $g(z_1) = z_1$, and $\mathcal{H} = \dot{\mathcal{F}}$ in Corollary B.3. Note that, by the transformation, $H_1^B(\epsilon, \mathcal{H}, P_2) \leq H_1^B(2\epsilon, \mathcal{H}, P_2)$ holds. Then, by Corollary B.3 with $K = 2^{-r}C$, there exist $D_1, D_2, D_3 > 0$, depending only on C and r, such that

$$P^n\left(\sup_{f\in\dot{\mathcal{F}}}\frac{\sqrt{n}}{q_n}\left|E\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right| > D\right) \leq D_2\exp\left(-D^2 q_n^2\right),$$

for $D_1 \leq D \leq D_3\sqrt{n}/q_n$. Thus when $r \geq 1$, we have

$$\begin{aligned}
\tau_n^{-1} E_{P^n}\left[R_{\phi_h}(\hat{f}) - \inf_{f\in\dot{\mathcal{F}}} R_{\phi_h}(f)\right] &\leq 4\tau_n^{-1}\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|E\left(Yf\left(X\right)\right) - \frac{1}{n}\sum_{i=1}^{n} Y_i f\left(X_i\right)\right|\right] \\
&\leq 4D_1 + 8\tau_n^{-1} D_2\exp\left(-D_1^2 q_n^2\right),
\end{aligned}$$

which leads to the result for the case of $r \geq 1$.

The result for the case of $0 < r < 1$ follows immediately by applying Lemma B.4 to equation (56), where we set $Z = (Y, X)$, $g(z_1) = z_1$, and $\mathcal{H} = \dot{\mathcal{F}}$. □

We now give the proofs of Theorems 5.1 and 5.2.

Proof of Theorem 5.2. The result in Theorem 5.2 follows by combining equation (28) and Theorem B.2. □

Proof of Theorem 5.1. Define a new classifier

$$\hat{f}^\dagger(x) \equiv 1\left\{x \in G_{\hat{f}}\right\} - 1\left\{x \notin G_{\hat{f}}\right\}, \tag{57}$$

which is a step function indicating $x \in G_{\hat{f}}$ and $x \notin G_{\hat{f}}$ by 1 and -1, respectively. Note that $R(\hat{f}^\dagger) = R(\hat{f})$ holds. Then equation (23) becomes

$$R(\hat{f}) - \inf_{f \in \mathcal{F}_\mathcal{G}} R(f) = R(\hat{f}^\dagger) - \inf_{f \in \mathcal{F}_\mathcal{G}} R(f) = \frac{1}{2}\left(R_{\phi_h}(\hat{f}^\dagger) - \inf_{f \in \widetilde{\mathcal{F}}_\mathcal{G}} R_{\phi_h}(f)\right). \tag{58}$$

When (24) holds for all $\epsilon > 0$, the result follows by applying Theorem B.2 to (58).

We consider the case when (25) holds for all $\epsilon > 0$. Define a class of step functions $\mathcal{I}_\mathcal{G} \equiv \{2 \cdot \mathbf{1}_G - 1 : G \in \mathcal{G}\}$. We now show that (A) $\hat{f}^\dagger$ minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{I}_\mathcal{G}$ and that (B) $\inf_{f \in \widetilde{\mathcal{F}}_\mathcal{G}} R_{\phi_h}(f) = \inf_{f \in \mathcal{I}_\mathcal{G}} R_{\phi_h}(f)$. If they hold, we can apply Theorem B.2 to the excess hinge rink in (58) with $\breve{\mathcal{F}}$ replaced by $\mathcal{I}_\mathcal{G}$.

We first prove (B). Since

$$\inf_{f \in \mathcal{I}_\mathcal{G}} R_{\phi_h}(f) = 2 \inf_{G \in \mathcal{G}} \mathcal{R}(G) = 2\mathcal{R}^*,$$

Theorem 4.4 (i) shows that $\inf_{f \in \widetilde{\mathcal{F}}_\mathcal{G}} R_{\phi_h}(f) = \inf_{f \in \mathcal{I}_\mathcal{G}} R_{\phi_h}(f)$.

We next prove (A). Note first that $\hat{f}^\dagger \in \mathcal{I}_\mathcal{G}$ holds. Let P_n be the empirical distribution on the sample $\{(Y_i, X_i) : i = 1, \ldots, n\}$. Replacing P with P_n in Theorem 4.4 shows that $\hat{f}$ minimizes $\hat{R}(\cdot)$ over $\widetilde{\mathcal{F}}_\mathcal{G}$. Hence $\hat{\mathcal{R}}(G) \equiv \inf_{f \in \mathcal{F}_G} \hat{R}(f)$ is minimized by $G_{\hat{f}}$ over $\mathcal{G}$. Then Theorem 4.4 (ii) with P replaced by P_n shows that the new classifier $\hat{f}^\dagger$ also minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_\mathcal{G}$. Then by the statement (B) with R_{ϕ_h} replaced with $\hat{R}_{\phi_h}$, $\hat{f}^\dagger$ minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{I}_\mathcal{G}$.

From the definitions of $H_1^B(\epsilon, \mathcal{G}, P_X)$ and $\mathcal{I}_\mathcal{G}$, we have $H_1^B(\epsilon, \mathcal{I}_\mathcal{G}, P_X) = H_1^B(2\epsilon, \mathcal{G}, P_X)$. Therefore, we can apply Theorem B.2, with $\breve{\mathcal{F}}$ replaced by $\mathcal{I}_\mathcal{G}$, to (58) and then obtain the inequality in (26). □

The following corollary is similar to Corollary D.1 in Mbakop and Tabord-Meehan (2021). The difference is that a class of functions $\mathcal{H}$ in the following corollary does not need to be a class of binary functions.

Corollary B.3. *Let $Z = (Z_1, Z_2) \sim P$, and $\{Z_i\}_{i=1}^n$ be a sequence of random variables that are i.i.d distributed as Z. Denote by P_2 the marginal distribution of Z_2. Suppose*

P_2 is absolutely continuous with respect to Lebesugue measure and its density is bounded form above by a finite constant $A > 0$. Let $\mathcal{F}$ be a class of real-valued functions of the form $f(z) = f(z_1, z_2) = g(z_1) \cdot h(z_2)$, where $h \in \mathcal{H}$, $\mathcal{H}$ is a class of functions with values in $[0,1]$, and g takes values in $[-1,1]$. Suppose $\mathcal{H}$ satisfies

$$H_1^B(\epsilon, \mathcal{H}, P_2) \leq K\epsilon^{-r}$$

for some constants $K > 0$ and $r \geq 1$ and for all $\epsilon > 0$. Then there exist positive constants D_1, D_2, D_3, depending only on K and r, such that for $n \geq 3$:

$$P^n\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(f(Z_i) - E_P[f(Z_i)]\right)\right| > xq_n\right) \leq D_2 \exp\left(-x^2 q_n^2\right),$$

for $D_1 \leq x \leq D_3\sqrt{n}/q_n$, where

$$q_n = \begin{cases} \log n & r = 1 \\ n^{(r-1)/2(r+2)} & r > 1 \end{cases}.$$

Proof. Let $\left[h_j^L, h_j^U\right]$, $j = 1, \ldots, N_1^B(\epsilon, \mathcal{H}, P_2)$, be a set of ϵ-brackets of $\mathcal{H}$ with respect to $L_1(P_2)$ such that $\left\|h_j^U - h_j^L\right\|_{1,P_2} \leq \epsilon$ and that $\mathcal{H} \subseteq \cup_{j=1}^{N_1^B(\epsilon,\mathcal{H},P_2)}\left[h_j^L, h_j^U\right]$. Since $\left|h_j^U - h_j^L\right| < 1$, $\left\|h_j^U - h_j^L\right\|_{2,P_2}^2 \leq \left\|h_j^U - h_j^L\right\|_{1,P_2} \leq \epsilon$ holds. We hence have

$$N_2^B(\epsilon, \mathcal{H}, P_2) \leq N_1^B\left(\epsilon^2, \mathcal{H}, P_2\right) \leq K\epsilon^{-2r}.$$

The result immediately follows by applying Proposition B.1. □

Lemma B.4. *Maintain the same definitions and assumptions as in Corollary B.3 with $r \geq 1$ replaced by $0 < r < 1$. Then, there exist positive constants D_3 and D_4, depending only on K and r, such that:*

$$\sup_{f\in\mathcal{F}} E_{P^n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E_P\left[f(Z)\right]\right|\right] \leq \frac{D_3}{\sqrt{n}} + \frac{D_4}{n}.$$

Proof. We consider to apply Proposition 3.5.15 in Giné and Nickl (2016). Note first that $|f| \leq 1$ and $\|f\|_{2,P} \leq 1$ for all $f \in \mathcal{F}$. Then we can apply Proposition 3.5.15 in Giné and

Nickl (2016), with $F = 1$ and $\delta = 1$, and obtain

$$\begin{aligned}\sup_{f\in\mathcal{F}} E_{P^n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E_P\left[f(Z)\right]\right|\right] &\leq \left(\frac{58}{\sqrt{n}} + \frac{1}{3n}\int_0^2 \sqrt{\log\left(2N_2^B\left(\epsilon,\mathcal{F},P\right)\right)}d\epsilon\right) \\ &\quad \times \int_0^2 \sqrt{\log\left(2N_2^B\left(\epsilon,\mathcal{F},P\right)\right)}d\epsilon. \\ &\leq \left(\frac{58}{\sqrt{n}} + \frac{2}{3n} + \frac{1}{3n}\int_0^2 \sqrt{H_2^B\left(\epsilon,\mathcal{F},P\right)}d\epsilon\right) \\ &\quad \times \left(\frac{2}{3} + \frac{1}{3}\int_0^2 \sqrt{H_2^B\left(\epsilon,\mathcal{F},P\right)}d\epsilon\right). \end{aligned} \tag{59}$$

By combining the arguments form the proofs of Corollary B.3 and Proposition B.1 below, we have

$$H_2^B\left(\epsilon,\mathcal{F},P\right) \leq K\epsilon^{-2r}.$$

Therefore, substituting this upper bound into (59) yields

$$\begin{aligned}\sup_{f\in\mathcal{F}} E_{P^n}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - E_P\left[f(Z)\right]\right|\right] &\leq \left(\frac{58}{\sqrt{n}} + \frac{2}{3n} + \frac{1}{3n}\int_0^2 K\epsilon^{-r}d\epsilon\right) \\ &\quad \times \left(\frac{2}{3} + \frac{1}{3}\int_0^2 K\epsilon^{-r}d\epsilon\right) \\ &= \left(\frac{58}{\sqrt{n}} + \frac{2}{3n} + \frac{2^{1-r}K}{3n(1-r)}\right)\left(\frac{2}{3} + \frac{2^{1-r}K}{3(1-r)}\right).\end{aligned}$$

Therefore, setting

$$\begin{aligned}D_3 &:= \left(\frac{116}{3} + \frac{29\cdot 2^{2-r}K}{3(1-r)}\right), \\ D_4 &:= \left(\frac{2}{3} + \frac{2^{1-r}K}{3(1-r)}\right)^2,\end{aligned}$$

leads to the result. □

Proposition B.1. *Let $Z = (Z_1, Z_2) \sim P$, and $\{Z_i\}_{i=1}^n$ be a sequence of random variables that are i.i.d distributed as Z. Denote by P_2 the marginal distribution of Z_2. Let $\mathcal{F}$ be a class of real-valued functions of the form $f(z) = f(z_1, z_2) = g(z_1)\cdot h(z_2)$, where $h \in \mathcal{H}$, $\mathcal{H}$ is a class of functions with values in $[0,1]$, and g takes values in $[-1,1]$. Suppose $\mathcal{H}$ satisfies*

$$H_2^B\left(\epsilon,\mathcal{H},P_2\right) \leq K\epsilon^{-r} \tag{60}$$

for some constants $K > 0$ and $r \geq 2$ and for all $\epsilon > 0$. Then there exist positive constants C_1, C_2, C_3, depending only on K and r, such that if

$$\xi \leq \frac{\sqrt{n}}{128} \tag{61}$$

and

$$\xi \geq \begin{cases} C_1 n^{(r-2)/2(r+2)} & r \geq 2 \\ C_2 \log \max(n, e) & r = 2 \end{cases}, \tag{62}$$

then

$$P^n \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(Z_i) - E_P[f(Z_i)] \right) \right| > \xi \right) \leq C_3 \exp\left(-\xi^2\right).$$

Proof. We follow the general strategy of Theorem 2.3 and Corollary 2.4 in Alexander (1984) and Proposition D.1 in Mbakop and Tabord-Meehan (2021). Define

$$v_n(f) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[f(Z_i) - E(f(Z_i)) \right].$$

We start by giving some definitions. Let $\delta_0 > \delta_1 > \cdots > \delta_N > 0$ be a sequence of real numbers where $\{\delta_j\}_{j=0}^N$ and N will be specified later. For each δ_j, there exists a set of δ_j-brackets $\mathcal{H}_j^B$ of $\mathcal{H}$ with respect to $L_2(P_2)$ such that $\left|\mathcal{H}_j^B\right| = N_2^B(\delta_j, \mathcal{H}, P_2)$. Define the function $H(\cdot) : (0, \infty) \to [0, \infty)$ as follows:

$$H(u) = \begin{cases} Ku^{-r} & \text{if } u < 1 \\ 0 & \text{if } u \geq 1 \end{cases}.$$

Note that by Assumption (60) and the fact that $\mathcal{H}$ has the diameter 1 by definition, $N_2^B(\delta_j, \mathcal{H}, P_2) \leq \exp(H(\delta_j))$ for all $\delta_j > 0$. For each $0 \leq j \leq N$ and any $f = g \cdot h \in \mathcal{F}$, define $f_j^L := g \cdot h_j^L 1\{g \geq 0\} + g \cdot h_j^U 1\{g < 0\}$ and $f_j^U := g \cdot h_j^U 1\{g \geq 0\} + g \cdot h_j^L 1\{g < 0\}$ for some $\left(h_j^L, h_j^U\right)$ that forms a δ_j-bracket for h with respect to $L_2(P_2)$ such that $h \in \left[h_j^L, h_j^U\right]$ and $\left[h_j^L, h_j^U\right] \in \mathcal{H}_j^B$. From the construction, $\left[f_j^U, f_j^L\right]$ is a δ_j-bracket for f with respect to $L_2(P)$. Let $f_j = f_j^L$, and let $\mathcal{F}_j = \{f_j : f \in \mathcal{F}\}$. We have $|\mathcal{F}_j| \leq \exp(H(\delta_j))$ and $\|f - f_j\|_{2,P} < \delta_j$ for every $f \in \mathcal{F}$.

By a standard chaining argument,

$$P\left(\sup_{f \in \mathcal{F}} |v_n(f)| > \xi \right) \leq |\mathcal{F}_0| \sup_{f \in \mathcal{F}} P\left(|v_n(f)| > \frac{7}{8}\xi \right)$$

$$+\sum_{j=0}^{N-1}|\mathcal{F}_j|\,|\mathcal{F}_{j+1}|\sup_{f\in\mathcal{F}}P\left(|v_n\left(f_j-f_{j+1}\right)|>\eta_j\right)$$
$$+P\left(\sup_{f\in\mathcal{F}}|v_n\left(f_N-f\right)|>\frac{\xi}{16}+\eta_N\right),$$

where $\{\eta_j\}_{j=0}^N$ are chosen such that $\sum_{j=0}^N\eta_j\leq\xi/16$ and will be specified later. Define

$$R_1=|\mathcal{F}_0|\sup_{f\in\mathcal{F}}P\left(|v_n\left(f\right)|>\frac{7}{8}\xi\right),$$
$$R_2=\sum_{j=0}^{N-1}|\mathcal{F}_j|\,|\mathcal{F}_{j+1}|\sup_{f\in\mathcal{F}}P\left(|v_n\left(f_j-f_{j+1}\right)|>\eta_j\right),$$
$$R_3=P\left(\sup_{f\in\mathcal{F}}|v_n\left(f_N-f\right)|>\frac{\xi}{16}+\eta_N\right).$$

We now choose $\{\delta_j\}_{j=0}^N$, $\{\eta_j\}_{j=0}^N$, and N to make the three terms sufficiently small.

First we study R_1. Set δ_0 such that $H\left(\delta_0\right)=\xi^2/4$. Then, applying Hoeffdings's inequality, we have

$$R_1\leq 2\,|\mathcal{F}_0|\exp\left(-2\left(\frac{7}{8}\xi\right)^2\right)\leq 2\exp\left(-\xi^2\right),$$

where we use that $|\mathcal{F}_0|\leq\exp\left(H\left(\delta_0\right)\right)=\exp\left(\xi^2/4\right)$ in the second inequality.

Next, we study R_2. Since $\|f_j-f_{j+1}\|_{2,P}\leq 2\delta_j$ by construction, applying Bennet's inequality (Lemma B.5) to each $\sup_{f\in\mathcal{F}}P\left(|v_n\left(f_j-f_{j+1}\right)|>\eta_j\right)$ in R_2 leads to

$$R_2\leq\sum_{j=0}^{N-1}2\exp\left(2H\left(\delta_{j+1}\right)\right)\exp\left(-\psi_1\left(\eta_j,n,4\delta_j^2\right)\right),$$

where ψ_1 satisfies the properties described in Lemma B.5.

Next, we study R_3. Given the construction of $\mathcal{F}_N$,

$$\begin{aligned}|v_n\left(f_N-f\right)|&\leq\left|v_n\left(f_N^U-f_N^L\right)\right|+2\sqrt{n}\left\|f_N^U-f_L^L\right\|_{1,P}\\&\leq\left|v_n\left(f_N^U-f_N^L\right)\right|+2\sqrt{n}\delta_N.\end{aligned}$$

The last inequality holds because $\left\|f_N^U-f_N^L\right\|_{1,P}\leq\left\|h_N^U-h_N^L\right\|_{1,P_2}$ and

$$\left\|h_N^U-h_N^L\right\|_{1,P_2}\leq\left\|h_N^U-h_N^L\right\|_{2,P_2}\leq\delta_N,$$

which holds from Hölder's inequality. Set $\delta_N\leq s:=\xi/\left(32\sqrt{n}\right)$. Then, by the above

derivation and Bennet's inequality,

$$\begin{aligned} R_3 &\leq P\left(\sup_{f\in\mathcal{F}}\left|v_n\left(f_N^U - f_N^L\right)\right| > \eta_N\right) \\ &\leq 2\left|\mathcal{F}_N\right|\exp\left(-\psi_1\left(\eta_N, n, \delta_N^2\right)\right). \end{aligned}$$

To develop upper bounds of R_2 and R_3, we consider two cases separately. First we consider the case $\delta_0 \leq s$. Set $N = 0$ and $\eta_0 = \xi/16$. Then we have that $R_2 = 0$ and

$$R_3 \leq 2\left|\mathcal{F}_0\right|\exp\left(-\psi_1\left(\eta_0, n, \delta_0^2\right)\right).$$

Since Assumption (61) and $\delta_0 \leq s$, we have that

$$2\eta_0 = \frac{\xi}{8} \geq 4\sqrt{n}\left(\frac{\xi}{32\sqrt{n}}\right)^2 \geq 4\sqrt{n}\delta_0^2.$$

Hence by the property of ψ_1 in Lemma B.5,

$$\psi_1\left(\eta_0, n, \delta_0^2\right) \geq \frac{1}{4}\psi_1\left(2\eta_0, n, \delta_0^2\right) \geq \frac{1}{4}\eta_0\sqrt{n}.$$

Using $\eta_0 = \xi/16$ and Assumption (61), we obtain

$$\psi_1\left(\eta_0, n, \delta_0^2\right) \geq \frac{1}{4}\eta_0\sqrt{n} = \frac{\xi}{64}\sqrt{n} \geq 2\xi^2.$$

By the definition of δ_0, we also have $\left|\mathcal{F}_0\right| \leq \exp\left(\xi^2/4\right)$. Therefore, combining these results gives

$$R_2 + R_3 \leq 2\exp\left(-\xi^2\right).$$

Next we consider the case $\delta_0 > s$. We here consider to apply Lemma B.6 where we let N and $\{\delta_j\}_{j=0}^N$ be as in Lemma B.6 and $t = \delta_0$ and s be as defined above. Let $\eta_j = 8\sqrt{2}\delta_j H\left(\delta_{j+1}\right)^{1/2}$ for $0 \leq j < N$ and $\eta_N = 8\sqrt{2}\delta_N H\left(\delta_N\right)^{1/2}$. Then Lemma B.6 leads to

$$\sum_{j=0}^N \eta_j = 8\sqrt{2}\sum_{j=0}^N H\left(\delta_{j+1}\right)^{1/2} \leq 64\sqrt{2}\int_{s/4}^{\delta_0} H\left(u\right)^{1/2} du.$$

We have that for $0 < s < t$,

$$\int_s^t H\left(u\right)^{1/2} du \leq \begin{cases} K^{1/2}\log\left(1/s\right) & r = 2 \\ 2K^{1/2}\left(r-2\right)^{-1} s^{(2-r)/2} & r > 2. \end{cases}$$

Combining this with Assumption (62), where C_1 and C_1 are set to be sufficiently large, we have

$$\sum_{j=0}^{N} \eta_j \leq \frac{\xi}{16},$$

which is consistent with our choice of $\{\eta_j\}_j$. Setting C_1 and C_2 sufficiently large, it follows from Assumption (62) that

$$H(s) \leq \frac{\xi\sqrt{n}}{16}.$$

Hence we have

$$\left(\frac{\eta_j}{4\delta_j^2\sqrt{n}}\right)^2 < \frac{8H(s)}{ns^2} \leq 16.$$

Then from the property of ψ_1,

$$\psi_1\left(\eta_j, n, 4\delta_j^2\right) \geq \frac{\eta_j^2}{16\delta_j^2}.$$

Using our bound on R_2, we obtain that

$$R_2 \leq \sum_{j=0}^{N-1} 2\exp\left(2H(\delta_{j+1}) - \frac{\eta_j^2}{16\delta_j^2}\right) \leq \sum_{j=0}^{N-1} 2\exp\left(-4^{j+1}H(\delta_0)\right).$$

Similarly, we obtain that

$$R_3 \leq 2\exp\left(-4^{N+1}H(\delta_0)\right).$$

Putting these together and using Assumption (62), we have

$$R_2 + R_3 \leq \sum_{j=0}^{\infty} 2\exp\left(-4^{j+1}H(\delta_0)\right) \leq C\exp\left(-\xi^2\right),$$

where C is a constant that depends only on K and r. □

Lemma B.5 (Bennet's inequality: see Theorem 2.9 in Boucheron et al. (2013))**.** *Let $\{Z_i\}_{i=1}^n$ be a sequence of independent random vectors with distribution P. Let f be some*

function taking values in $[0,1]$ *and define*

$$v_n(f) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[f(Z_i) - E_P(f(Z_i)) \right].$$

Then, for any $\xi \geq 0$*, the following holds:*

$$P^n(|v_n(f)| > \xi) \leq 2\exp(-\psi_1(\xi, n, a)),$$

where $a = var(v_n(f))$ *and*

$$\psi_1(\xi, n, a) = \xi\sqrt{n}h\left(\frac{\xi}{\sqrt{n}\alpha}\right),$$

with $h(x) = (1 + x^{-1})\log(1 + x) - 1$.

Furthermore, ψ_1 *has the following two properties:*

$$\psi_1(\xi, n, \alpha) \geq \psi_1(C\xi, n, \rho\alpha) \geq C^2\rho^{-1}\psi_1(\xi, n, \alpha)$$

for $C \leq 1$ *and* $\rho \geq 1$*, and*

$$\psi_1(\xi, n, \alpha) \geq \begin{cases} \frac{\xi^2}{4\alpha} & \text{if } \xi < 4\sqrt{n}\alpha \\ \frac{\xi}{2}\sqrt{n} & \text{if } \xi \geq 4\sqrt{n}\alpha \end{cases}.$$

Lemma B.6 (Lemma 3.1 in Alexander (1984))**.** *Let* $H : (0, t] \to \mathbb{R}^+$ *be a decreasing function, and let* $0 < s < t$*. Set* $\delta_0 := t$*,* $\delta_{j+1} := s \vee \sup\{x \leq \delta_j/2 : H(x) \geq 4H(\delta_j)\}$ *for* $j \geq 0$*, and* $N := \min\{j : \delta_j = s\}$*. Then*

$$\sum_{j=0}^{N} \delta_j H(\delta_{j+1})^{1/2} \leq 8\int_{s/4}^{t} H(x)^{1/2}\, dx.$$

C Proofs of the results in Section 6

This appendix provides the proofs of the results in Section 6. Throughout this appendix, we suppose $\mathcal{X} = [0,1]^{d_x}$ as in Section 6.

We first provide the proof of Lemma 6.1.

Proof of Lemma 6.1. Let μ_X be the Lebesgue measure on $\mathcal{X}$. From Theorem 8.3.2 in Dudley (1999), $H_1^B(\epsilon, \mathcal{G}_M, \mu_X) \leq K\epsilon^{d_x - 1}$ holds for some positive constant K and for

all $\epsilon > 0$. Since P_X is absolutely continuous with respect to μ_X and has a density that is bounded from above by A, we have $H_1^B(A^{-1}\epsilon, \mathcal{G}_M, P_X) \leq H_1^B(\epsilon, \mathcal{G}_M, \mu_X)$. Thus the result (i) follows by setting $C = A^{-d_x}K$.□

The following lemma will be used in the proof of Theorem 6.4.

Lemma C.1. *Suppose that P_X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and has a density that is bounded from above by a finite constant $A > 0$. Then there exists a constant $\widetilde{C}$, which depends only on A, such that*

$$H_1^B\left(\epsilon, \mathcal{F}_M, P_X\right) \leq \widetilde{C}\epsilon^{-d_x}.$$

holds for all $\epsilon > 0$.

Proof. Transform $\mathcal{F}_M$ into $\widetilde{\mathcal{F}}_M = \{(f+1)/2 : f \in \mathcal{F}_M\}$, which is a class of monotonically increasing functions taking values in $[0,1]$. By this transformation, $N_1^B\left(\epsilon, \mathcal{F}_M, P_X\right) = N_1^B\left(\epsilon/2, \widetilde{\mathcal{F}}_M, P_X\right)$ holds. Then the result follows by applying Corollary 1.3 in Gao and Wellner (2007) to $\widetilde{\mathcal{F}}_M$, in which we set $\tilde{C} = 2^{-d_x}C_2$, where C_2 is the same constant that appears in Corollary 1.3 in Gao and Wellner (2007). Note that this corollary requires that P_X is absolutely continuous with respect to the Lebesgue measure on $\mathcal{X}$ and has a bounded density. □

The following two lemmas will be used in the proof of Lemma 6.3.

Lemma C.2. *Let $\hat{f}_B \in \arg\inf_{f \in \mathcal{B}_\mathbf{k}} \hat{R}_{\phi_h}(f)$, and $\hat{\theta} := \left\{\hat{\theta}_{j_1 \cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ be the vector of the coefficients of the Bernstein bases in $\hat{f}_B$. Let r_1^+ and r_1^- be the smallest non-negative value and the largest negative value in $\hat{\theta}$, respectively.*
(i) If all non-negative elements in $\hat{\theta}$ take the same value r_1^+, let r_2^+ be 1; otherwise, let r_2^+ be the second smallest non-negative value in $\hat{\theta}$. Make a $(k_1+1) \times \cdots \times (k_{d_x}+1)$-th dimensional vector $\tilde{\theta} := \left\{\widetilde{\Theta}_{j_1 \cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ such that, for all $j_1, \ldots, j_{d_x}$, if $\hat{\theta}_{j_1 \cdots j_{d_x}} = r_1^+$, $\widetilde{\Theta}_{j_1 \cdots j_{d_x}} = r_2^+$; otherwise, $\widetilde{\Theta}_{j_1 \cdots j_{d_x}} = \hat{\theta}_{j_1 \cdots j_{d_x}}$. Then a new classifier

$$\tilde{f}_B(x) := \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \widetilde{\Theta}_{j_1 \cdots j_{d_x}} \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_1 j_1}(x_{d_x})\right)$$

minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{B}_\mathbf{k}$.
(ii) Similarly, if all negative elements in $\hat{\theta}$ take the same value r_1^-, let r_2^- be -1; otherwise, let r_2^- be the second largest negative value in $\hat{\theta}$. Make a $(k_1+1) \times \cdots \times (k_{d_x}+1)$-th dimensional vector $\check{\theta} := \left\{\check{\theta}_{j_1 \cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ such that, for all $j_1, \ldots, j_{d_x}$, if

$\hat{\theta}_{j_1\ldots j_{d_x}} = r_1^-$, $\check{\theta}_{j_1\ldots j_{d_x}} = r_2^-$; *otherwise,* $\check{\theta}_{j_1\ldots j_{d_x}} = \hat{\theta}_{j_1\ldots j_{d_x}}$. *Then a new classifier*

$$\check{f}_B(x) := \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \check{\theta}_{j_1\ldots j_{d_x}} \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_1 j_1}(x_{d_x})\right)$$

minimizes $\hat{R}_{\phi_h}(\cdot)$ *over* $\mathcal{B}_{\mathbf{k}}$.
(iii) A classifier

$$\hat{f}_B^{\dagger}(x) := \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x=1}}^{k_{d_x}} \operatorname{sign}\left(\hat{\theta}_{j_1\ldots j_{d_x}}\right) \cdot \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x})\right)$$

minimizes $\hat{R}_{\phi_h}(\cdot)$ *over* $\mathcal{B}_{\mathbf{k}}$.

Proof. First, note that $\tilde{\theta}, \check{\theta} \in \widetilde{\Theta}$ holds by their constructions. We now prove (i). The proof of (ii) follows by the similar argument. Define

$$L_n(\theta) \equiv \sum_{i=1}^{n} \left\{ Y_i \cdot \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \theta_{j_1\ldots j_{d_x}} \sum_{i=1}^{n} \left(b_{k_1 j_1}(X_{1i}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{d_x i})\right) \right\}.$$

Minimization problem of $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{B}_{\mathbf{k}}$ is equivalent to the maximization problem of $L_n(\cdot)$ over $\widetilde{\Theta}$. Thus, $\hat{\theta}$ maximizes $L_n(\cdot)$ over $\widetilde{\Theta}$.

We prove the result by contradiction. Suppose $\tilde{\theta} \notin \arg\max_{\theta \in \widetilde{\Theta}} L_n(\theta)$. Let

$$J_1 \equiv \left\{ (j_1, \ldots, j_{d_x}) : \hat{\theta}_{j_1\ldots j_{d_x}} = r_1^+ \right\}.$$

Then,

$$L_n\left(\tilde{\theta}\right) - L_n\left(\hat{\theta}\right) = \sum_{(j_1,\ldots,j_{d_x}) \in J_1} \left\{ \left(r_2^+ - r_1^+\right) \sum_{i=1}^{n} Y_i \left(b_{k_1 j_1}(X_{1i}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{d_x i})\right) \right\} < 0.$$

Since $r_2^+ - r_1^+ \geq 0$, the above inequality implies that there exists some $(j_1, \ldots, j_{d_x}) \in J_1$ such that $\sum_{i=1}^{n} Y_i \left(b_{k_1 j_1}(X_{1i}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{d_x i})\right) < 0$. For such $(j_1, \ldots, j_{d_x})$, setting $\hat{\theta}_{j_1\ldots j_{d_x}}$ to r_1^- can increase the value of $L_n\left(\hat{\theta}\right)$ without violating the constraints in $\widetilde{\Theta}$. But this contradicts that $\hat{\theta}_{j_1\ldots j_{d_x}}$ is non-negative. Therefore, $\tilde{\theta}$ maximizes $L_n(\cdot)$ over $\widetilde{\Theta}$, or equivalently $\tilde{f}_B$ minimizes $\hat{R}_{\phi_h}(\cdot)$ over $\mathcal{B}_{\mathbf{k}}$.

The result (iii) is shown by applying Lemma C.2 (i) and (ii) repeatedly to $\hat{f}_B$. □

Lemma C.3. *Fix $G \in \mathcal{G}$ and $k_j \geq 1$ for $j = 1, \ldots, d_x$. Define a classifier*

$$f_G(x) := \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \theta_{j_1 \ldots j_{d_x}} \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x}) \right),$$

such that, for all $j_1, \ldots, j_{d_x}$,

$$\theta_{j_1 \ldots j_{d_x}} = \begin{cases} 1 & \text{if } (j_1/k_1, \ldots, j_{d_x}/k_{d_x}) \in G \\ -1 & \text{if } (j_1/k_1, \ldots, j_{d_x}/k_{d_x}) \notin G \end{cases}.$$

Then the following holds:

$$\left| R_{\phi_h}(f_G) - R_{\phi_h}(1\{\cdot \in G\} - 1\{\cdot \notin G\}) \right| \leq 2A \sum_{j=1}^{d_x} \sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x} \frac{4}{\sqrt{k_j}}.$$

Proof. Define

$$J_{\mathbf{k}} := \{(j_1, \ldots, j_{d_x}) : (j_1/k_1, \ldots, j_{d_x}/k_{d_x}) \in G\},$$

which is a set of grid points on G. It follows that

$$\begin{aligned}
& R_{\phi_h}(f_G) - R_{\phi_h}(1\{\cdot \in G\} - 1\{\cdot \notin G\}) \\
= & \int_{[0,1]^{d_x}} (2\eta(x) - 1)(1\{x \in G\} - 1\{x \notin G\} - B_{\mathbf{k}}(\theta, x)) \, dP_X(x) \\
= & \int_{[0,1]^{d_x}} (2\eta(x) - 1) 1\{x \in G\} \, dP_X(x) - \int_{[0,1]^{d_x}} (2\eta(x) - 1) 1\{x \notin G\} \, dP_X(x) \\
& - \underbrace{\int_{[0,1]^{d_x}} (2\eta(x) - 1) B_{\mathbf{k}}(\theta, x) \, dP_X(x)}_{(I)}. \qquad (63)
\end{aligned}$$

(I) can be written as

$$\begin{aligned}
(I) = & \int_{[0,1]^{d_x}} (2\eta(x) - 1) \sum_{(j_1, \ldots, j_{d_x}) \in J_{\mathbf{k}}} \left(\prod_{v=1}^{d_x} b_{k_v j_v}(x_v) \right) dP_X(x) \\
& - \int_{[0,1]^{d_x}} (2\eta(x) - 1) \sum_{(j_1, \ldots, j_{d_x}) \notin J_{\mathbf{k}}} \left(\prod_{v=1}^{d_x} b_{k_v j_v}(x_v) \right) dP_X(x).
\end{aligned}$$

Thus,

$$(63) = \int_{[0,1]^{d_x}} (2\eta(x) - 1) \underbrace{\left(1\{x \in G\} - \sum_{(j_1,\ldots,j_{d_x}) \in J_{\mathbf{k}}} \left[\prod_{v=1}^{d_x} b_{k_v j_v}(x_v) \right] \right)}_{(II)} dP_X(x)$$
$$+ \int_{[0,1]^{d_x}} (2\eta(x) - 1) \underbrace{\left(\sum_{(j_1,\ldots,j_{d_x}) \notin J_k} \left[\prod_{v=1}^{d_x} b_{k_v j_v}(x_v) \right] - 1\{x \in (G)^c\} \right)}_{(III)} dP_X(x).$$

Let $Bin(k_j, x_j)$, $j = 1, \ldots, d_x$, be independent binomial variables with parameters (k_j, x_j). Then, both (II) and (III) are equivalent to

$$\Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in (J_{\mathbf{k}})^c) 1\{x \in G\}$$
$$- \Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}}) 1\{x \in (G)^c\}.$$

Hence,

$$(63) = 2 \int_G (2\eta(x) - 1) \Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in (J_{\mathbf{k}})^c) dP_X(x).$$
$$- 2 \int_{(G)^c} \Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}}) dP_X(x),$$

and therefore

$$|R_{\phi_h}(f_G) - R_{\phi_h}(1\{\cdot \in G\} - 1\{\cdot \notin G\})|$$
$$\leq 2 \underbrace{\int_G \Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in (J_{\mathbf{k}})^c) dP_X(x)}_{(IV)}$$
$$+ 2 \underbrace{\int_{(G)^c} \Pr((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}}) dP_X(x)}_{(V)}.$$

We first evaluate (V). Let ϵ be a small positive value which will will converge to zero as $k_v \to \infty$. For small $\Delta_v \leq \epsilon/\sqrt{d_x}$, $v = 1, \ldots, d_x$, which will converge to zero as $k_v \to \infty$, define

$$\tilde{G}^c := \{x \in (G)^c : (x_1 + \Delta_1, \ldots, x_{d_x} + \Delta_{d_x}) \in (G)^c\}.$$

This set might be nonempty or empty. We consider these cases separately. First, suppose

that $\tilde{G}^c$ is nonempty. For each $x \in \tilde{G}^c$, let

$$(j_1(x), \ldots, j_{d_x}(x)) \in \underset{(j_1,\ldots,j_{d_x}) \in J_{\mathbf{k}}: j_1/k_1 \geq x_1+\Delta_1, \ldots, j_{d_x}/k_{d_x} \geq x_{d_x}+\Delta_{d_x}}{\arg\min} \|x - (j_1/k_1, \ldots, j_{d_x}/k_{d_x})\|.$$

Then,

$$\begin{aligned}
(V) \leq & \int_{(G)^c \setminus \tilde{G}^c} \Pr\left((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}} \right) dP_X(x) \\
& + \int_{\tilde{G}^c} \left(\sum_{v=1}^{d_x} \Pr\left(Bin(k_v, x_v) \geq j_v(x) \right) \right) dP_X(x) \\
\leq & A \cdot (\Delta_1 + \cdots + \Delta_{d_x}) \\
& + \int_{\tilde{G}^c} \sum_{v=1}^{d_x} \exp\left\{ -2k_v \left(\frac{j_v(x)}{k_v} - x_v \right)^2 \right\} dP_X(x) \\
\leq & A \cdot (\Delta_1 + \cdots + \Delta_{d_x}) + \sum_{v=1}^{d_x} \int_{\tilde{G}^c} \exp\left(-2k_v \Delta_v^2 \right) dP_X(x).
\end{aligned}$$

To obtain the second inequality, we apply the Hoeffding's inequality to $\Pr\left(Bin(k_v, x_v) \geq j_v(x) \right)$, which is applicable since $k_v x_v \leq j_v(x)$ for each $x \in \tilde{G}^c$, and use the following:

$$\begin{aligned}
& \int_{(G)^c \setminus \tilde{G}^c} \Pr\left((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}} \right) dP_X(x) \\
\leq & A \int_{(G)^c \setminus \tilde{G}^c} dx \leq A \cdot (\Delta_1 + \cdots + \Delta_{d_x}),
\end{aligned}$$

where the second inequality holds because $\int_{(G)^c \setminus \tilde{G}^c} dx$ is bounded from above by $\sum_{v=1}^{d_x} \Delta_v - (d_x - 1) \prod_{v=1}^{d_x} \Delta_v$, which is taken when $(G)^c = \mathcal{X}$. The last inequality follows from that $j_v(x)/k_v - x_v \geq \Delta_v$ for all $v = 1, \ldots, d_x$ and $x \in \tilde{G}$.

Next, we consider the case that $\tilde{G}^c$ is empty. In this case,

$$\begin{aligned}
(V) = & \int_{(G)^c} \Pr\left((Bin(k_1, x_1), \ldots, Bin(k_{d_x}, x_{d_x})) \in J_{\mathbf{k}} \right) dP_X(x) \\
\leq & \int_{(G)^c} dP_X(x) \leq A \cdot \max_{v=1,\ldots,d_x} \Delta_v.
\end{aligned}$$

The inequality follows because $\int_{(G)^c} dx$ is bounded from above by $\max_{v=1,\ldots,d_x} \Delta_v$ when $\tilde{G}^c$ is empty. Therefore, regardless of whether $\tilde{G}^c$ is empty or not, we have

$$(V) \leq A \cdot (\Delta_1 + \cdots + \Delta_{d_x}) + \sum_{v=1}^{d_x} \int_{\tilde{G}^c} \exp\left(-2k_v \Delta_v^2 \right) dP_X(x).$$

Set $\Delta_v = \sqrt{\log k_v}/\left(2\sqrt{k_v}\right)$ for each $v = 1, \ldots, d_x$. Then we have

$$\begin{aligned}(V) &\leq \frac{A}{2}\left(\sum_{v=1}^{d_x}\sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x}\exp\left(-\frac{1}{2}\log k_v\right) \\ &= \frac{A}{2}\left(\sum_{v=1}^{d_x}\sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x}\frac{1}{\sqrt{k_v}}.\end{aligned}$$

Next, we evaluate (IV). For small $\Delta_v \leq \epsilon/\sqrt{d_x}$, $v = 1, \ldots, d_x$, which will converge to zero as $k_v \to \infty$, define

$$\tilde{G} := \{x \in G : (x_1 - \Delta_1, \ldots, x_{d_x} - \Delta_{d_x}) \in G\}.$$

We again separately consider two cases: $\tilde{G}$ is empty or not. First, suppose that $\tilde{G}$ is nonempty. For each $x \in \tilde{G}$, let

$$\left(\tilde{j}_1(x), \ldots, \tilde{j}_{d_x}(x)\right) \in \underset{(j_1,\ldots,j_{d_x})\in(J_k)^c : j_1/k_1 \leq x_1-\Delta_1, \ldots, j_{d_x}/k_{d_x} \leq x_{d_x}-\Delta_{d_x}}{\arg\min} \|x - (j_1/k_1, \ldots, j_{d_x}/k_{d_x})\|.$$

Then,

$$\begin{aligned}(IV) \leq & \int_{G\setminus\tilde{G}} \Pr\left(\left(Bin\left(k_1, x_1\right), \ldots, Bin\left(k_{d_x}, x_{d_x}\right)\right) \in (J_{\mathbf{k}})^c\right) dP_X(x) \\ & + \int_{\tilde{G}} \left(\sum_{v=1}^{d_x} \Pr\left(Bin\left(k_v, x_v\right) \leq \tilde{j}_v(x)\right)\right) dP_X(x) \\ \leq & A \cdot (\Delta_1 + \cdots + \Delta_{d_x}) + \int_{\tilde{G}} \left(\sum_{v=1}^{d_x} \exp\left\{-2k_v\left(x_v - \frac{\tilde{j}_v(x)}{k_v}\right)^2\right\}\right) dP_X(x) \\ \leq & A \cdot (\Delta_1 + \cdots + \Delta_{d_x}) + \sum_{v=1}^{d_x} \int_{\tilde{G}} \exp\left(-2k_v\Delta_v^2\right) dP_X(x).\end{aligned}$$

The second inequality follows from Hoeffding's inequality and that

$$\begin{aligned}& \int_{G\setminus\tilde{G}} \Pr\left(\left(Bin\left(k_1, x_1\right), \ldots, Bin\left(k_{d_x}, x_{d_x}\right)\right) \in (J_{\mathbf{k}})^c\right) dP_X(x) \\ \leq & A \int_{G\setminus\tilde{G}} dx \leq A \cdot (\Delta_1 + \cdots + \Delta_{d_x}),\end{aligned}$$

where the inequality holds because $\int_{G\setminus\tilde{G}} dx$ takes the largest value, $\sum_{v=1}^{d_x}\Delta_v - \prod_{v=1}^{d_x}\Delta_v$, when $G = \mathcal{X}$. The last inequality follows from that $j_v(x)/k_v \leq x_v - \Delta_v$ for all $v = 1, \ldots d_x$ and $x \in \tilde{G}$.

Next, we consider the case that $\tilde{G}$ is empty. In this case,

$$\begin{aligned}(IV) &= \int_G \Pr\left(\left(Bin\left(k_1, x_1\right), \ldots, Bin\left(k_{d_x}, x_{d_x}\right)\right) \in \left(J_{\mathbf{k}}\right)^c\right) dP_X\left(x\right) \\ &\leq \int_G dP_X\left(x\right) \leq A \cdot \max_{v=1,\ldots,d_x} \Delta_v,\end{aligned}$$

where the inequality follows because $\int_G dx$ is bounded from above by $\max_{v=1,\ldots,d_x} \Delta_v$ when $\tilde{G}$ is empty. Therefore, regardless of whether $\tilde{G}$ is empty or not, we have

$$(IV) \leq A \cdot \left(\Delta_1 + \cdots + \Delta_{d_x}\right) + \sum_{v=1}^{d_x} \int_{\tilde{G}} \exp\left(-2k_v\Delta_v^2\right) dP_X\left(x\right).$$

Set $\Delta_v = \sqrt{\log k_v}/\left(2\sqrt{k_v}\right)$ for each $v = 1, \ldots, d_x$. Then, we have

$$\begin{aligned}(IV) &\leq \frac{A}{2}\left(\sum_{v=1}^{d_x} \sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x} \exp\left(-\frac{1}{2}\log k_v\right) \\ &= \frac{A}{2}\left(\sum_{v=1}^{d_x} \sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x} \frac{1}{\sqrt{k_v}}.\end{aligned}$$

Consequently, combining above the results, we obtain

$$\left|R_{\phi_h}\left(f_G\right) - R_{\phi_h}\left(1\left\{\cdot \in G\right\} - 1\left\{\cdot \notin G\right\}\right)\right| \leq 2A\left(\sum_{v=1}^{d_x} \sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x} \frac{4}{\sqrt{k_v}}.$$

□

Finally, the following is the proof of Lemma 6.3.

Proof of Lemma 6.3. We first prove (i). Let G^* minimizes $\mathcal{R}(\cdot)$ over $\mathcal{G}_M$. From Theorem 4.4, a classifier $\tilde{f}^*\left(x\right) := 1\left\{x \in G^*\right\} - 1\left\{x \in \left(G^*\right)^c\right\}$ minimizes the hinge risk $R_{\phi_h}\left(\cdot\right)$ over $\mathcal{F}_M$. Define a vector $\theta^* = \left\{\theta^*_{j_1\ldots j_d}\right\}_{j_1=0,\ldots,k_1;\ldots;j_d=0,\ldots,k_d}$ such that for each $j_1, \ldots, j_d$,

$$\theta^*_{j_1\ldots j_d} = \begin{cases} 1 & \text{if } \left(j_1/k_1, \ldots, j_d/k_d\right) \in G^* \\ -1 & \text{otherwise.}\end{cases}$$

Note that θ^* is contained by $\widetilde{\Theta}$. Thus, it follows that

$$\inf_{f \in \mathcal{B}_{\mathbf{k}}} R_{\phi_h}\left(f\right) - R_{\phi_h}\left(\tilde{f}^*\right) \leq R_{\phi_h}\left(B_{\mathbf{k}}\left(\theta^*, \cdot\right)\right) - R_{\phi_h}\left(\tilde{f}^*\right).$$

Then, applying Lemma C.3 to $R_{\phi_h}\left(B_{\mathbf{k}}\left(\theta^*,\cdot\right)\right) - R_{\phi_h}\left(\tilde{f}^*\right)$ shows the result (i).

The result (ii) follows immediately from Lemmas C.2 (iii) and C.3. □

D Proofs of the results in Section 7

This section gives the proofs of the results in Section 7 for the weighted classification. Most of the proofs are natural extensions of the proofs of the results in Sections 3–6. For the simplicity of the notation, define a function

$$L(\omega_+, \omega_-, \eta) \equiv -\omega_+\eta + \omega_-(1-\eta),$$

the right hand side of which appears in the condition (37).

We first give the proofs of Theorem 7.1 and Corollary 7.2.

Proof of Theorem 7.1. We first prove the "if part" of the first statement. By equation (35),

$$\begin{aligned}\mathcal{R}_\phi^w\left(G_1\right) - \mathcal{R}_\phi^w\left(G_2\right) = &\int_{G_1\setminus G_2} \Delta C_\phi^w\left(\omega_+\left(x\right), \omega_-\left(x\right), \eta\left(x\right)\right) dP_X\left(x\right) \\ &- \int_{G_2\setminus G_1} \Delta C_\phi^w\left(\omega_+\left(x\right), \omega_-\left(x\right), \eta\left(x\right)\right) dP_X\left(x\right).\end{aligned}$$

Thus, $\mathcal{R}_{\phi_1}^w\left(G_1\right) \geq \mathcal{R}_{\phi_1}^w\left(G_2\right)$ is equivalent to

$$\int_{G_1\setminus G_2} \Delta C_{\phi_1}^w\left(\omega_+\left(x\right), \omega_-\left(x\right), \eta\left(x\right)\right) dP_X\left(x\right) \geq \int_{G_2\setminus G_1} \Delta C_{\phi_1}^w\left(\omega_+\left(x\right), \omega_-\left(x\right), \eta\left(x\right)\right) dP_X\left(x\right).$$

Replacing $\Delta C_{\phi_1}^w$ by $\Delta C_{\phi_2}^w = c\Delta C_{\phi_1}^w$ with $c > 0$ does not change the above inequality. Moreover, the above inequality with $\Delta C_{\phi_1}^w$ replaced by $\Delta C_{\phi_2}^w$ is equivalent to $\mathcal{R}_{\phi_2}^w\left(G_1\right) \geq \mathcal{R}_{\phi_2}^w\left(G_2\right)$. Therefore, the "if part" of the first statement holds.

The "only if" part follows directly from Theorem 3.6 by setting $\Delta C_\phi^w\left(\omega_+\left(x\right), \omega_-\left(x\right), \eta\left(x\right)\right) = \Delta C_\phi\left(\eta\left(x\right)\right)$, or equivalently $\omega_+(x) = \omega_-(x) = 1$, for all $x \in \mathcal{X}$.

Next, we prove the second statement. For the 0-1 loss function $\phi_{0-1}\left(\alpha\right) = 1\left\{\alpha \leq 0\right\}$,

$$\Delta C_{\phi_{0-1}}^w\left(\omega_+, \omega_-, \eta\right) = L(\omega_+, \omega_-, \eta)$$

holds. Thus, according to the first statement, condition (37) is a necessary and sufficient condition for ϕ_2 to share the same risk ordering with ϕ_{0-1}.

Finally, we prove the last statement. For the hinge loss function $\phi_h(\alpha) = a\max\left\{0, 1-\alpha\right\}$

and $f \in \mathcal{F}_{\mathcal{G}}$, we have

$$C_{\phi_h}(\omega_+, \omega_-, f, \eta) = a\left(\omega_+ (1-f)\eta + \omega_- (1+f)(1-\eta)\right).$$

Hence, we obtain

$$\Delta C^{w+}_{\phi_h}(\omega_+, \omega_-, \eta) = \begin{cases} 2a\omega_-(1-\eta) & \text{for } L(\omega_+, \omega_-, \eta) < 0 \\ a\left(\omega_+\eta + \omega_-(1-\eta)\right) & \text{for } L(\omega_+, \omega_-, \eta) \geq 0 \end{cases},$$

$$\Delta C^{w}_{\phi_h}(\omega_+, \omega_-, \eta) = \begin{cases} a\left(\omega_+\eta + \omega_-(1-\eta)\right) & \text{for } L(\omega_+, \omega_-, \eta) < 0 \\ 2a\omega_+\eta & \text{for } L(\omega_+, \omega_-, \eta) \geq 0 \end{cases}.$$

Hence, $\Delta C^{w}_{\phi_h}(\omega_+, \omega_-, \eta) = aL(\omega_+, \omega_-, \eta)$ holds for all $(\omega_+, \omega_-, \eta) \in \mathbb{R} \times \mathbb{R} \times [0,1]$.□

Proof of Corollary 7.2. Equation (38) follows by

$$\begin{aligned}
&R^w(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R^w(f) = \mathcal{R}^w(G_f) - \mathcal{R}^w(G^*) \\
&= \int_{\mathcal{X}} L(\omega_+(x), \omega_-(x), \eta(x)) \left(1\{x \in G_f\} - 1\{x \in G^*\}\right) dP_X(x) \\
&= c^{-1} \int_{\mathcal{X}} \Delta C^w_{\phi}(\omega_+(x), \omega_-(x), \eta(x)) \left(1\{x \in G_f\} - 1\{x \in G^*\}\right) dP_X(x) \\
&= c^{-1}\left(\mathcal{R}^w_{\phi}(G_f) - \mathcal{R}^w_{\phi}(G^*)\right) = c^{-1}\left(\inf_{\tilde{f} \in \mathcal{F}_{G_f}} R^w_{\phi}\left(\tilde{f}\right) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R^w_{\phi}(f)\right) \\
&\leq c^{-1}\left(R^w_{\phi}(f) - \inf_{f \in \mathcal{F}_{\mathcal{G}}} R^w_{\phi}(f)\right),
\end{aligned}$$

where the first equality follows from (34); the second equality follows from the assumption; the third equality follows from (35).□

We next provide the proof of Theorem 7.5. Beforehand, note that the weighted hinge risk can be expressed as

$$\begin{aligned}
R^w_{\phi_h}(f) &= \int_{\mathcal{X}} \left(\omega_+(x)(1 - f(x))\eta(x) + \omega_-(x)(1 + f(x))(1 - \eta(x))\right) dP_X(x) \\
&= \int_{\mathcal{X}} L(\omega_+(x), \omega_-(x), \eta(x)) f(x)\, dP_X(x) + E_P[\omega]. \qquad (64)
\end{aligned}$$

Moreover, for $G \in \mathcal{G}$, $\mathcal{R}(G)$ can be written as

$$\begin{aligned}\mathcal{R}^{w}(G) &= \int_{\mathcal{X}} \left(\omega_{+}(x)\,\eta(x)\,1\{x \in G^{c}\} + \omega_{-}(x)\,(1-\eta(x))\,1\{x \in G\}\right) dP_{X}(x) \\ &= -\int_{G^{c}} L(\omega_{+}(x), \omega_{-}(x), \eta(x))\,(1-\eta(x))\,dP_{X}(x) \\ &\quad + \int_{\mathcal{X}} \omega_{-}(x)\,(1-\eta(x))\,dP_{X}(x). \end{aligned} \tag{65}$$

The following lemma, which is an analogue of Lemma 4.3, will be used in the proof of Theorem 7.5.

Lemma D.1. *(i) Let $\tilde{f}^{*}$ be a minimizer of the weighted hinge risk $R^{w}_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. Then $\tilde{f}^{*}$ minimizes the weighted classification risk $R^{w}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$, and leads to $R^{w}_{\phi_h}\left(\tilde{f}^{*}\right) = 2\mathcal{R}^{w*}$.*
(ii) For $G^{*}, \tilde{G}^{*} \in \mathcal{G}^{*}$ such that $G^{*} \subseteq \tilde{G}^{*}$, $\tilde{f}^{\dagger}(\cdot) = 1\{\cdot \in G^{*}\} - 1\left\{\cdot \notin \tilde{G}^{*}\right\}$ minimizes $R^{w}_{\phi_h}(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$.

Proof. Fix $\tilde{f} \in \widetilde{\mathcal{F}}_{\mathcal{G},J}$. The classifier $\tilde{f}$ has the form of (50) for some $G_J \subseteq \cdots \subseteq G_1 \subseteq \tilde{G}_1 \subseteq \cdots \subseteq \tilde{G}_J$ and $c_j^{+}, c_j^{-} \geq 0$ for $j = 1, \ldots, J$ with $\sum_{j=1}^{J} c_j^{+} = \sum_{j=1}^{J} c_j^{-} = 1$. From (64), the weighted hinge risk of $\tilde{f}$ can be written as

$$\begin{aligned} R^{w}_{\phi_h}\left(\tilde{f}\right) &= \sum_{j=1}^{J}\left[\left(c_j^{+}\right)\int_{G_j} L(\omega_{+}(x), \omega_{-}(x), \eta(x))\,(1-\eta(x))\,dP_{X}(x)\right] \\ &\quad + \sum_{j=1}^{J}\left[\left(-c_j^{-}\right)\int_{(\tilde{G}_j)^{c}} L(\omega_{+}(x), \omega_{-}(x), \eta(x))\,(1-\eta(x))\,dP_{X}(x)\right] \\ &\quad + E_{P}[\omega]. \end{aligned} \tag{66}$$

Denote the first and second terms in (66) by

$$\begin{aligned} R^{wI}_{\phi_h}\left(\tilde{f}\right) &= \sum_{j=1}^{J}\left[\left(c_j^{+}\right)\int_{G_j} L(\omega_{+}(x), \omega_{-}(x), \eta(x))\,(1-\eta(x))\,dP_{X}(x)\right], \\ R^{wII}_{\phi_h}\left(\tilde{f}\right) &= \sum_{j=1}^{J}\left[\left(-c_j^{-}\right)\int_{(\tilde{G}_j)^{c}} L(\omega_{+}(x), \omega_{-}(x), \eta(x))\,(1-\eta(x))\,dP_{X}(x)\right]. \end{aligned}$$

By equation (35), $R^{wI}_{\phi_h}\left(\tilde{f}\right)$ can be written as

$$R^{wI}_{\phi_h}\left(\tilde{f}\right) = \sum_{j=1}^{J}\left[\left(c_j^{+}\right)\left(\mathcal{R}^{w}(G_j) - \int_{\mathcal{X}} \omega_{+}(x)\,\eta(x)\,dP_{X}(x)\right)\right]$$

$$= \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}^w \left(G_j\right) - \int_{\mathcal{X}} \omega_+ \left(x\right) \eta \left(x\right) dP_X \left(x\right),$$

where the second equality follows from $\sum_{j=1}^{J} c_j^+ = 1$. Similarly, by equation (65), $R_{\phi_h}^{wII}\left(\tilde{f}\right)$ can be written as

$$\begin{aligned} R_{\phi_h}^{wII}\left(\tilde{f}\right) &= \sum_{j=1}^{J} \left[\left(c_j^-\right)\left(\mathcal{R}^w\left(\tilde{G}_j\right) - \int_{\mathcal{X}} \omega_- \left(x\right)\left(1-\eta\left(x\right)\right) dP_X\left(x\right)\right)\right] \\ &= \sum_{j=1}^{J} \left(c_j^-\right) \mathcal{R}^w\left(\tilde{G}_j\right) - \int_{\mathcal{X}} \omega_- \left(x\right)\left(1-\eta\left(x\right)\right) dP_X\left(x\right), \end{aligned}$$

where the second equality follows from $\sum_{j=1}^{J} c_j^- = 1$.

Combining these expressions gives

$$\begin{aligned} R_{\phi_h}^{w}\left(\tilde{f}\right) &= R_{\phi_h}^{wI}\left(\tilde{f}\right) + R_{\phi_h}^{wII}\left(\tilde{f}\right) + E_P\left[\omega\right] \\ &= \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}^w\left(G_j\right) + \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}^w\left(\tilde{G}_j\right). \end{aligned} \tag{67}$$

Because $\mathcal{R}^w\left(G_j\right), \mathcal{R}^w\left(\tilde{G}_j\right) \geq \mathcal{R}^{w*}$ and $\sum_{j=1}^{J} c_j^+ = \sum_{j=1}^{J} c_j^- = 1$, the above expression implies that $R_{\phi_h}^{w}\left(\tilde{f}\right) \geq 2\mathcal{R}^{w*}$ for all $\tilde{f} \in \widetilde{\mathcal{F}}_{\mathcal{G},J}$.

Let $G^*, \tilde{G}^* \in \mathcal{G}^*$ such that $G^* \subseteq \tilde{G}^*$, and let $\tilde{f}^\dagger\left(x\right) = 1\left\{x \in G^*\right\} - 1\left\{x \notin \tilde{G}^*\right\}$. $\tilde{f}^\dagger$ can be taken from $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ by setting $G_1 = G^*$ with $c_1^+ = 1$ and $\tilde{G}_1 = \tilde{G}^*$ with $c_1^- = 1$. Then, from (67), $R_{\phi_h}^{w}\left(\tilde{f}^\dagger\right)$ takes its lower bound $2\mathcal{R}^{w*}$. Therefore, $\tilde{f}^\dagger$ minimizes $R_{\phi_h}^{w}\left(\cdot\right)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. This proves $\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}^{w}\left(f\right) = 2\mathcal{R}^{w*}$ and the statement (ii).

We next prove that a minimizer of $R_{\phi_h}^{w}\left(\cdot\right)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ also minimizes $R^w\left(\cdot\right)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. To obtain contradiction, suppose $\tilde{f}$ minimizes $R_{\phi_h}^{w}\left(\cdot\right)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$ but do does not minimize $R^w\left(\cdot\right)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. As $\tilde{f}$ does not minimize the weighted classification risk, $G_{\tilde{f}} \notin \mathcal{G}^*$ holds. Let m be the smallest number in $\{1, \ldots, J\}$ such that $c_m^- > 0$. Because $\tilde{G}_m = G_{\tilde{f}}$, $\tilde{G}_m \notin \mathcal{G}^*$ holds. Then,

$$\begin{aligned} R_{\phi_h}^{w}\left(\tilde{f}\right) &= \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}^w\left(G_j\right) + \sum_{j=1}^{J} \left(c_j^-\right) \mathcal{R}^w\left(\tilde{G}_j\right) \\ &= c_m^- \mathcal{R}^w\left(\tilde{G}_m\right) + \sum_{j=1}^{J} \left(c_j^+\right) \mathcal{R}^w\left(G_j\right) + \sum_{j \in \{1,\ldots,m-1,m+1,\ldots,J\}} \left(c_j^+\right) \mathcal{R}^w\left(\tilde{G}_j\right) \\ &\geq c_m^- \mathcal{R}^w\left(\tilde{G}_m\right) + \left(2 - c_m^-\right) \mathcal{R}^{w*} \\ &> 2\mathcal{R}^{w*}, \end{aligned}$$

where the last line comes from $c_m^- > 0$ and $\tilde{G}_m \notin \mathcal{G}^*$. As $\inf_{f \in \widetilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}^w(f) = 2\mathcal{R}^{w*}$, this contradicts that $\tilde{f}$ minimizes $R_{\phi_h}^w(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G},J}$. □

We are now prepared to give the proof of Theorem 7.5.

Proof of Theorem 7.5. Let $\bar{\mathcal{F}}_J^*$ be as in the proof of Theorem 4.4. Note that $G_f = G_{\tilde{f}^*}$ holds for any $f \in \bar{\mathcal{F}}_J^*$. Similarly to the the proof of Theorem 4.4, define a sequence of functions $\left\{\bar{f}_J^*\right\}_{J=1}^{\infty}$ with

$$\bar{f}_J^*(x) \equiv \sum_{j=1}^{J} \frac{1}{J}\left(1\left\{\tilde{f}^*(x) \geq j/J\right\} - 1\left\{\tilde{f}^*(x) < j/J\right\}\right).$$

It is shown in the the proof of Theorem 4.4 that $\bar{f}_J^*(X) \to \tilde{f}^*(X)$ as $J \to \infty$ with probability one.

Then the following holds:

$$\begin{aligned} R_{\phi_h}^w\left(\tilde{f}^*\right) &= \int_{\mathcal{X}} L(\omega_+(x), \omega_-(x), \eta(x))\tilde{f}^*(x)\, dP_X(x) + E_P[\omega] \\ &= \lim_{J\to\infty} \int_{\mathcal{X}} v\bar{f}_J^*(x)\, dP_X(x) + E_P[\omega] \\ &= \lim_{J\to\infty} R_{\phi_h}^w\left(\bar{f}_J^*(x)\right) \geq \lim_{J\to\infty} \inf_{\bar{f}\in\bar{\mathcal{F}}_J^*} R_{\phi_h}^w\left(\bar{f}\right) \\ &\geq \lim_{J\to\infty} \inf_{\tilde{f}\in\widetilde{\mathcal{F}}_{\mathcal{G}.J}} R_{\phi_h}^w\left(\tilde{f}\right), \end{aligned} \tag{68}$$

where the first and third equalities follow from (66). The second equality follows from the dominated convergence theorem, which holds as both

$$L(\omega_+(X), \omega_-(X), \eta(X))\bar{f}_J^*(X) \to L(\omega_+(X), \omega_-(X), \eta(X))\tilde{f}^*(X)$$

and

$$\left|L(\omega_+(X), \omega_-(X), \eta(X))\bar{f}_J^*(X)\right| < \infty$$

hold with probability, where the second condition holds from condition (7.4). The first inequality follows from $\bar{f}_J^* \in \bar{\mathcal{F}}_J^*$, and the last inequality follows from $\bar{\mathcal{F}}_J^* \subseteq \widetilde{\mathcal{F}}_{\mathcal{G}.J}$.

Lemma D.1 shows that $\inf_{\tilde{f}\in\widetilde{\mathcal{F}}_{\mathcal{G},J}} R_{\phi_h}^w\left(\tilde{f}\right) = 2\mathcal{R}^{w*}$ for any J. Thus, we have

$$R_{\phi_h}^w\left(\tilde{f}^*\right) \geq \lim_{J\to\infty} \inf_{\tilde{f}\in\widetilde{\mathcal{F}}_{\mathcal{G}.J}} R_{\phi_h}^w\left(\tilde{f}\right) = 2\mathcal{R}^{w*}.$$

This means that the minimal value of $R_{\phi_h}^w$ on $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is at least $2\mathcal{R}^{w*}$. Lemma D.1 also shows

that $\tilde{f}^{\dagger}$ defined in Theorem 7.5 leads to $R^w_{\phi_h}\left(\tilde{f}^{\dagger}\right) = 2\mathcal{R}^{w*}$. Therefore, $\tilde{f}^{\dagger}$ minimizes $R^w_{\phi_h}$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$. This proves the second statement of the theorem.

We next prove the first statement of the theorem by contradiction. Suppose that $\tilde{f}^*$ does not minimize $R^w(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$, or equivalently $G_{\tilde{f}^*} \notin \mathcal{G}^*$. Then, for any $\bar{f} \in \bar{\mathcal{F}}^*_J$,

$$\begin{aligned} R^w_{\phi_h}\left(\bar{f}\right) &= \sum_{j=1}^{J}\left(c_j^+\right)\mathcal{R}^w\left(G_j\right) + \sum_{j=1}^{J}\left(c_j^-\right)\mathcal{R}^w\left(\tilde{G}_j\right) \\ &\geq c_1^-\mathcal{R}^w\left(\tilde{G}_1\right) + \left(2 - c_1^-\right)\mathcal{R}^{w*} \\ &> 2\mathcal{R}^{w*}, \end{aligned}$$

where the last line follows from $\tilde{G}_1 = G_{\tilde{f}^*} \notin \mathcal{G}^*$ and $c_1^- > 0$. Therefore, we have from equation (68) that

$$R^w_{\phi_h}\left(\tilde{f}^*\right) \geq \lim_{J\to\infty}\inf_{\bar{f}\in\bar{\mathcal{F}}^*_J} R^w_{\phi_h}\left(\bar{f}\right) > 2\mathcal{R}^{w*}.$$

This contradicts that $\tilde{f}^*$ minimizes $R^w_{\phi_h}$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$ because $R^w_{\phi_h}\left(\tilde{f}^{\dagger}\right) = 2\mathcal{R}^{w*}$. Therefore, $\tilde{f}^*$ minimizes $R^w(\cdot)$ over $\widetilde{\mathcal{F}}_{\mathcal{G}}$.□

The following is the proof of Corollary 7.6.

Proof of Corollary 7.6. By equations (34) and (65), $R^w(f)$ can be written as

$$cR^w(f) = \frac{c}{2}\left\{\int_{\mathcal{X}} L(\omega_+(x), \omega_-(x), \eta(x))\left(1\{x \in G_f\} - 1\{x \notin G_f\}\right) dP_X(x) + E_P[\omega]\right\}.$$

By equation (64), the right-hand side is equal to $2^{-1}R^w_{\phi}\left(1\{\cdot \in G_f\} - 1\{\cdot \notin G_f\}\right)$.□

The following gives the proof Theorem 7.7, which is an extension of the proof of Theorem 5.1.

Proof of Theorem 7.7 (ii). For the convenience of notation, we prove the result with C' and r' replaced by C and r, respectively. Fix $P \in \mathcal{P}$. First of all, Corollary 7.6 and decomposing $R^w_{\phi_h}(\hat{f}) - \inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}} R^w_{\phi_h}(f)$ gives

$$\begin{aligned} R^w(\hat{f}) - \inf_{f\in\mathcal{F}_{\mathcal{G}}} R^w(f) = \frac{1}{2}\left(R^w_{\phi_h}\left(\hat{f}\right) - \inf_{f\in\tilde{\mathcal{F}}} R^w_{\phi_h}(f)\right) + \frac{1}{2}\left(\inf_{f\in\tilde{\mathcal{F}}} R^w_{\phi_h}(f) - \inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}} R^w_{\phi_h}(f)\right) \\ + \frac{1}{2}\left(R^w_{\phi_h}(1\left\{\cdot \in G_{\hat{f}}\right\} - 1\left\{\cdot \notin G_{\hat{f}}\right\}) - R^w_{\phi_h}(\hat{f})\right). \end{aligned} \tag{69}$$

Hence, to obtain the inequality in (40), we need to prove

$$R^w_{\phi_h}\left(\hat{f}\right) - \inf_{f\in\check{\mathcal{F}}} R^w_{\phi_h}(f) \le L_C(r,n).$$

We follow the same strategy as the proof of Theorem B.2. Let $\check{f}^*$ minimizes $R^w_{\phi_h}(\cdot)$ over $\check{\mathcal{F}}$. A standard argument gives

$$\begin{aligned}
E_{P^n}\left[R^w_{\phi_h}\left(\hat{f}\right) - \inf_{f\in\check{\mathcal{F}}} R^w_{\phi_h}(f)\right] &\le E_{P^n}\left[R^w_{\phi_h}\left(\hat{f}\right) - \hat{R}^w_{\phi_h}\left(\hat{f}\right) + \hat{R}^w_{\phi_h}\left(\check{f}^*\right) - R^w_{\phi_h}\left(\check{f}^*\right)\right] \\
&\quad \left(\because \hat{R}^w_{\phi_h}\left(\hat{f}\right) \le \hat{R}^w_{\phi_h}\left(\check{f}^*\right)\right) \\
&= 2E_{P^n}\left[R^w_{\phi_h}\left(\frac{\hat{f}+1}{2}\right) - \hat{R}^w_{\phi_h}\left(\frac{\hat{f}+1}{2}\right)\right] \\
&\quad + 2E_{P^n}\left[\hat{R}^w_{\phi_h}\left(\frac{\check{f}^*+1}{2}\right) - R^w_{\phi_h}\left(\frac{\check{f}^*+1}{2}\right)\right] \\
&\le 4\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|R^w_{\phi_h}(f) - \hat{R}^w_{\phi_h}(f)\right|\right],
\end{aligned}$$

where $\dot{\mathcal{F}} = \left\{(f+1)/2 : f \in \check{\mathcal{F}}\right\}$ be as in the proof of Theorem B.2.

Note that

$$\begin{aligned}
&\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|R^w_{\phi_h}(f) - \hat{R}^w_{\phi_h}(f)\right|\right] \\
&= M\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|E_P\left(\left(\frac{\omega}{M}\right)Yf(X)\right) - \frac{1}{n}\sum_{i=1}^n\left(\frac{\omega_i}{M}\right)Y_i f(X_i)\right|\right] \qquad (70)
\end{aligned}$$

and that

$$\sup_{f\in\dot{\mathcal{F}}}\left|E_P\left(\left(\frac{\omega}{M}\right)Yf(X)\right) - \frac{1}{n}\sum_{i=1}^n\left(\frac{\omega_i}{M}\right)Y_i f(X_i)\right| \le 2.$$

We first prove the result for the case of $r \ge 1$. For any $f \in \dot{\mathcal{F}}$ and $D > 0$, it holds that

$$\begin{aligned}
&\frac{\sqrt{n}}{q_n}\sup_{f\in\dot{\mathcal{F}}} E_{P^n}\left[\left|E\left(\left(\frac{\omega}{M}\right)Yf(X)\right) - \frac{1}{n}\sum_{i=1}^n\left(\frac{\omega_i}{M}\right)Y_i f(X_i)\right|\right] \\
&\le D + \frac{2\sqrt{n}}{q_n}P^n\left(\sup_{f\in\dot{\mathcal{F}}}\frac{\sqrt{n}}{q_n}\left|E\left(\left(\frac{\omega}{M}\right)Yf(X)\right) - \frac{1}{n}\sum_{i=1}^n\left(\frac{\omega_i}{M}\right)Y_i f(X_i)\right| > D\right).
\end{aligned}$$

Then, applying Corollary B.3, where we set $Z_1 = (\omega, Y)$, $Z_2 = X$, $g(Z_1) = (\omega/M)\cdot Y$,

and $\mathcal{H} = \dot{\mathcal{F}}$, shows that there exist $D_1, D_2, D_3 > 0$, depending only on r and C, such that

$$P^n\left(\sup_{f\in\dot{\mathcal{F}}}\frac{\sqrt{n}}{q_n}\left|E\left(\left(\frac{\omega}{M}\right)Yf\left(X\right)\right)-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\omega_i}{M}\right)Y_if\left(X_i\right)\right|>D\right)\leq D_2\exp\left(-D^2q_n^2\right),$$

for $D_1 \leq D \leq D_3\sqrt{n}/q_n$. Therefore, when $r \geq 1$, we have

$$\tau_n^{-1}E_{P^n}\left[R^w_{\phi_h}(\hat{f})-\inf_{f\in\breve{\mathcal{F}}}R^w_{\phi_h}(f)\right]\leq 4MD_1+8M\tau_n^{-1}D_2\exp\left(-D_1^2q_n^2\right).$$

Combining this result with (69) leads to the inequality (40) for the case of $r \geq 1$.

The inequality in (40) for the case of $r < 1$ follows immediately by applying Lemma B.4 to equation (70).

It remains to prove the inequality (39) for the case of $\breve{\mathcal{F}} = \widetilde{\mathcal{F}}_{\mathcal{G}}$. By the similar argument as in the proof of Theorem 5.1, when $\breve{\mathcal{F}} = \widetilde{\mathcal{F}}_{\mathcal{G}}$, the second and third terms in equation (69) are ignorable. Thus, the inequality (39) follows from the above argument. □

Proof of Theorem 7.7 (i). We follow the same strategy as in the proof of Theorem 5.1. Define $\hat{f}^\dagger(x) = 1\{x \in G_{\hat{f}}\} - 1\{x \notin G_{\hat{f}}\}$. Then equation (69) becomes

$$\begin{aligned}R^\omega(\hat{f})-\inf_{f\in\mathcal{F}_{\mathcal{G}}}R^\omega(f)&=R^\omega(\hat{f}^\dagger)-\inf_{f\in\mathcal{F}_{\mathcal{G}}}R^\omega(f)\\&=\frac{1}{2}\left(R^\omega_{\phi_h}(\hat{f}^\dagger)-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}R^\omega_{\phi_h}(f)\right).\end{aligned}\tag{71}$$

It follows that

$$\begin{aligned}R^\omega_{\phi_h}(\hat{f}^\dagger)-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}R^\omega_{\phi_h}(f)&=E_P(\omega Y\hat{f}^\dagger(X))-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}E_P(\omega Yf(X))\\&\leq M\left|E_P(Y\hat{f}^\dagger(X))-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}E_P(Yf(X))\right|\\&=M\left(R_{\phi_h}(\hat{f}^\dagger)-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}R_{\phi_h}(f)\right),\end{aligned}$$

where $R_{\phi_h}(\cdot)$ is defined here with respect to the marginal distribution of (Y, X) generated by P. The third line follows because $\widetilde{\mathcal{F}}_{\mathcal{G}}$ is a classification-preserving reduction of $\mathcal{F}_{\mathcal{G}}$ and, accordingly, $E_P(Y\hat{f}^\dagger(X)) \geq \inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}} E_P(Yf(X))$ holds. Thus we have

$$R^\omega(\hat{f})-\inf_{f\in\mathcal{F}_{\mathcal{G}}}R^\omega(f)\leq M\left(R_{\phi_h}(\hat{f}^\dagger)-\inf_{f\in\widetilde{\mathcal{F}}_{\mathcal{G}}}R_{\phi_h}(f)\right).$$

Then the result follows by applying the same argument in the proof of Theorem 5.1 to the above equation. □

The following are extensions of Lemmas C.2, C.3, and 6.3.

Lemma D.2. *Let* $\hat{f}_B \in \arg\inf_{f\in\mathcal{B}_{\mathbf{k}}} \hat{R}^w_{\phi_h}(f)$, *and* $\hat{\theta} := \left\{\hat{\theta}_{j_1\cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ *be the vector of the coefficients in* $\hat{f}_B$. *Let* r_1^+ *and* r_1^- *be the smallest non-negative value and the largest negative value in* $\hat{\theta}$, *respectively.*
(i) If all non-negative elements in $\hat{\theta}$ *take the same value* r_1^+, *let* r_2^+ *be* 1; *otherwise, let* r_2^+ *be the second smallest non-negative value in* $\hat{\theta}$. *Make a* $(k_1+1)\times\cdots\times(k_{d_x}+1)$*-th dimensional vector* $\tilde{\theta} := \left\{\widetilde{\Theta}_{j_1\cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ *such that for all* $j_1,\ldots,j_{d_x}$ *if* $\hat{\theta}_{j_1\cdots j_{d_x}} = r_1^+$, $\widetilde{\Theta}_{j_1\cdots j_{d_x}} = r_2^+$; *otherwise,* $\widetilde{\Theta}_{j_1\cdots j_{d_x}} = \hat{\theta}_{j_1\cdots j_{d_x}}$. *Then, a classifier*

$$\tilde{f}_B(x) := \sum_{j_1=1}^{k_1}\cdots\sum_{j_{d_x}=1}^{k_{d_x}} \widetilde{\Theta}_{j_1\cdots j_{d_x}} \left(b_{k_1 j_1}(x_1)\times\cdots\times b_{k_1 j_1}(x_{d_x})\right)$$

minimizes $\hat{R}_{\phi_h}$ *over* $\mathcal{B}_{\mathbf{k}}$.
(ii) Similarly, if all negative elements in $\hat{\theta}$ *take the same value* r_1^-, *let* r_2^- *be* -1; *otherwise, let* r_2^- *be the second largest negative value in* $\hat{\theta}$. *Make a* $(k_1+1)\times\cdots\times(k_{d_x}+1)$*-vector* $\check{\theta} := \left\{\check{\theta}_{j_1\cdots j_{d_x}}\right\}_{j_1=1,\ldots,k_1;\ldots;j_{d_x}=1,\ldots,k_{d_x}}$ *such that for all* $j_1,\ldots,j_{d_x}$ *if* $\check{\theta}_{j_1\cdots j_{d_x}} = r_1^-$, $\check{\theta}_{j_1\cdots j_{d_x}} = r_2^-$; *otherwise,* $\check{\theta}_{j_1\cdots j_{d_x}} = \hat{\theta}_{j_1\cdots j_{d_x}}$. *Then, a classifier*

$$\check{f}_B(x) := \sum_{j_1=1}^{k_1}\cdots\sum_{j_{d_x}=1}^{k_{d_x}} \check{\theta}_{j_1\cdots j_{d_x}} \left(b_{k_1 j_1}(x_1)\times\cdots\times b_{k_1 j_1}(x_{d_x})\right)$$

minimizes $\hat{R}^w_{\phi_h}$ *over* $\mathcal{B}_{\mathbf{k}}$.
(iii) A classifier

$$\hat{f}^{\dagger}_B(x) := \sum_{j_1=1}^{k_1}\cdots\sum_{j_{d_x=1}}^{k_{d_x}} \operatorname{sign}\left(\hat{\theta}_{j_1\cdots j_{d_x}}\right)\cdot\left(b_{k_1 j_1}(x_1)\times\cdots\times b_{k_{d_x} j_{d_x}}(x_{d_x})\right)$$

minimizes $\hat{R}_{\phi_h}(\cdot)$ *over* $\mathcal{B}_{\mathbf{k}}$.

Proof. First, note that $\tilde{\theta}, \check{\theta} \in \widetilde{\Theta}$ holds by their constructions. We here prove (i). The proof of (ii) follows by the similar argument. Define

$$L_n(\theta) = \sum_{i=1}^{n}\left\{\omega_i Y_i \sum_{j_1=1}^{k_1}\cdots\sum_{j_{d_x}=1}^{k_{d_x}} \theta_{j_1\cdots j_{d_x}} \sum_{i=1}^{n}\left(b_{k_1 j_1}(X_{1i})\times\cdots\times b_{k_{d_x} j_{d_x}}(X_{d_x i})\right)\right\}.$$

Minimization problem of $\hat{R}_{\phi_h}$ over $\mathcal{B}_{\mathbf{k}}$ is equivalent to the following maximization problem of $L_n(\theta)$ over $\widetilde{\Theta}$. Thus, $\hat{\theta}$ maximizes $L_n(\theta)$ over $\widetilde{\Theta}$.

We prove the result by contradiction. Suppose $\tilde{\theta} \notin \arg\max_{\theta \in \widetilde{\Theta}} L_n(\theta)$. Let

$$J_1 \equiv \left\{ (j_1, \ldots, j_{d_x}) : \hat{\theta}_{j_1 \ldots j_{d_x}} = r_1^+ \right\}.$$

Then,

$$\begin{aligned} L_n\left(\tilde{\theta}\right) - L_n\left(\hat{\theta}\right) = & \sum_{(j_1,\ldots,j_{d_x}) \in J_1} \left\{ \left(r_2^+ - r_1^+\right) \sum_{i=1}^n \omega_i Y_i \left(b_{k_1 j_1}(X_{1i}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{d_x i}) \right) \right\} \\ & < 0 \end{aligned}$$

holds. Since $r_2^+ - r_1^+ \geq 0$, the above equation implies that there exists some $(j_1, \ldots, j_{d_x})$ in J_1 such that $\sum_{i=1}^n \omega_i Y_i \left(b_{k_1 j_1}(X_{1i}) \times \cdots \times b_{k_{d_x} j_{d_x}}(X_{d_x i}) \right) < 0$. For such $(j_1, \ldots, j_{d_x})$, setting $\hat{\theta}_{j_1 \ldots j_{d_x}}$ to r_1^- increases the value of $L_n\left(\hat{\theta}\right)$ without violating the constraints in $\widetilde{\Theta}$. But this contradicts that $\hat{\theta}_{j_1 \ldots j_{d_x}}$ is non-negative. Therefore, $\tilde{\theta}$ maximizes $L_n(\theta)$ over $\widetilde{\Theta}$, or equivalently $\tilde{f}_B$ minimizes $\hat{R}_{\phi_h}$ over $\mathcal{B}_{\mathbf{k}}$.

The result (iii) follows by applying the results (i) and (ii) repeatedly to $\hat{f}_B$. □

Lemma D.3. *Fix $G \in \mathcal{G}$ and $k_1, \ldots, k_{d_x}$. Define a classifier*

$$f_G(x) = \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \theta_{j_1 \ldots j_{d_x}} \left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x}) \right),$$

such that, for all $j_1, \ldots, j_{d_x}$, $\theta_{j_1 \ldots j_{d_x}} = 1$ if $(j_1/k_1, \ldots, j_{d_x}/k_{d_x}) \in G$, and $\theta_{j_1 \ldots j_{d_x}} = -1$ otherwise. Then, the following holds:

$$\left| R_{\phi_h}^w(f_G) - R_{\phi_h}^w\left(1\{\cdot \in G\} - 1\{\cdot \notin G\}\right) \right| \leq 2MA \sum_{j=1}^{d_x} \sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x} \frac{4M}{\sqrt{k_j}}.$$

Proof. Define

$$J_{\mathbf{k}} \equiv \{ (j_1, \ldots, j_{d_x}) : (j_1/k_1, \ldots, j_{d_x}/k_{d_x}) \in G \},$$

which is a set of grid points on G, and

$$L(x) \equiv -\omega_+(x)\eta(x) + \omega_-(x)(1 - \eta(x)).$$

It follows that

$$\begin{aligned}
&R_{\phi_h}^w\left(f_G\right)-R_{\phi_h}^w\left(1\left\{\cdot\in G\right\}-1\left\{\cdot\notin G\right\}\right)\\
=&\int_{[0,1]^{d_x}}L(x)\left(1\left\{x\in G\right\}-1\left\{x\notin G\right\}-B_k\left(\theta,x\right)\right)dP_X\left(x\right)\\
=&\int_{[0,1]^{d_x}}L(x)1\left\{x\in G\right\}dP_X\left(x\right)-\int_{[0,1]^{d_x}}L(x)1\left\{x\notin G\right\}dP_X\left(x\right)\\
&-\underbrace{\int_{[0,1]^{d_x}}L(x)B_{\mathbf{k}}\left(\theta,x\right)dP_X\left(x\right)}_{(I)}.
\end{aligned}\tag{72}$$

(I) can be written as

$$\begin{aligned}
(I)=&\int_{[0,1]^{d_x}}L(x)\sum_{(j_1,...,j_{d_x})\in J_{\mathbf{k}}}\left(\prod_{v=1}^{d_x}b_{k_vj_v}(x_v)\right)dP_X\left(x\right)\\
&-\int_{[0,1]^{d_x}}L(x)\sum_{(j_1,...,j_{d_x})\notin J_{\mathbf{k}}}\left(\prod_{v=1}^{d_x}b_{k_vj_v}(x_v)\right)dP_X\left(x\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
(72)=&\int_{[0,1]^{d_x}}L(x)\underbrace{\left(1\left\{x\in G\right\}-\sum_{(j_1,...,j_{d_x})\in J_{\mathbf{k}}}b_{k_vj_v}(x_v)\right)}_{(II)}dP_X\left(x\right)\\
&+\int_{[0,1]^{d_x}}L(x)\underbrace{\left(\sum_{(j_1,...,j_{d_x})\notin J_k}b_{k_vj_v}(x_v)-1\left\{x\in(G)^c\right\}\right)}_{(III)}dP_X\left(x\right).
\end{aligned}$$

Let $Bin\left(k_j,x_j\right)$, $j=1,\ldots,d_x$, be independent binomial variables with parameters (k_j,x_j). Then, (II) and (III) are equivalent to

$$\begin{aligned}
&\Pr\left(\left(Bin\left(k_1,x_1\right),\ldots,Bin\left(k_{d_x},x_{d_x}\right)\right)\in\left(J_{\mathbf{k}}\right)^c\right)1\left\{x\in G\right\}\\
-&\Pr\left(\left(Bin\left(k_1,x_1\right),\ldots,Bin\left(k_{d_x},x_{d_x}\right)\right)\in J_{\mathbf{k}}\right)1\left\{x\in(G)^c\right\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
(72)=&\,2\int_G L(x)\Pr\left(\left(Bin\left(k_1,x_1\right),\ldots,Bin\left(k_{d_x},x_{d_x}\right)\right)\in\left(J_{\mathbf{k}}\right)^c\right)dP_x(x)\\
&-2\int_{(G)^c}L(x)\Pr\left(\left(Bin\left(k_1,x_1\right),\ldots,Bin\left(k_{d_x},x_{d_x}\right)\right)\in J_{\mathbf{k}}\right)dP_X\left(x\right),
\end{aligned}$$

and therefore

$$\begin{aligned}
&\left|R^w_{\phi_h}\left(f_G\right) - R^w_{\phi_h}\left(1\left\{\cdot \in G\right\} - 1\left\{\cdot \notin G\right\}\right)\right| \\
&\leq 2M \underbrace{\int_G \Pr\left(\left(Bin\left(k_1, x_1\right), \ldots, Bin\left(k_{d_x}, x_{d_x}\right)\right) \in \left(J_{\mathbf{k}}\right)^c\right) dP_X\left(x\right)}_{(IV)} \\
&+ 2M \underbrace{\int_{(G)^c} \Pr\left(\left(Bin\left(k_1, x_1\right), \ldots, Bin\left(k_{d_x}, x_{d_x}\right)\right) \in J_{\mathbf{k}}\right) dP_X\left(x\right)}_{(V)},
\end{aligned}$$

because $\|L(x)\| < M$ for all $x \in \mathcal{X}$. The proof of Lemma 6.3 shows that

Then, applying the same argument in the proof of Corollary C.2 shows

$$\left|R^w_{\phi_h}\left(f_G\right) - R^w_{\phi_h}\left(1\left\{\cdot \in G\right\} - 1\left\{\cdot \notin G\right\}\right)\right| \leq 2MA\left(\sum_{v=1}^{d_x}\sqrt{\frac{\log k_v}{k_v}}\right) + \sum_{v=1}^{d_x}\frac{4M}{\sqrt{k_v}}.$$

□

Lemma D.4. *Let $k_j \geq 1$, for $j = 1, \ldots, d_x$, be fixed. Suppose that the density of P_X is bounded from above by $A > 0$. Suppose further that $\mathbf{k} = (k_1, \ldots, k_{d_x})$ satisfies $\sqrt{d_x \log k_j}/\left(2\sqrt{k_j}\right) \leq \epsilon$ for all $j = 1, \ldots, d_x$ and some $\epsilon > 0$.*
(i) The following holds for the approximation error to the best classifier:

$$\inf_{f \in \mathcal{B}_{\mathbf{k}}} R^w_{\phi_h}\left(f\right) - \inf_{f \in \mathcal{F}_M} R^w_{\phi_h}\left(f\right) \leq 2AM\sum_{j=1}^{d_x}\sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x}\frac{4M}{\sqrt{k_j}}.$$

(ii) For $\hat{f}_B \in \arg\inf_{f \in \mathbf{B}_{\mathbf{k}}} \hat{R}^w_{\phi_h}\left(f\right)$ such that its coefficients take values in $\{-1, 1\}$, the following holds for the approximation error to the step function:

$$R^w_{\phi_h}\left(1\left\{\cdot \in G_{\hat{f}_B}\right\} - 1\left\{\cdot \notin G_{\hat{f}_B}\right\}\right) - R^w_{\phi_h}(\hat{f}_B) \leq 2AM\sum_{j=1}^{d_x}\sqrt{\frac{\log k_j}{k_j}} + \sum_{j=1}^{d_x}\frac{4M}{\sqrt{k_j}}.$$

Proof. We first prove (i). Let G^* minimizes $\mathcal{R}^w(\cdot)$ over $\mathcal{G}_M$. From Theorem 7.5, a classifier $\tilde{f}^*\left(x\right) := 1\left\{x \in G^*\right\} - 1\left\{x \in \left(G^*\right)^c\right\}$ minimizes the hinge risk $R^w_{\phi_h}\left(\cdot\right)$ over $\mathcal{F}_M$. Define a vector $\theta^* = \left\{\theta^*_{j_1 \ldots j_d}\right\}_{j_1=0,\ldots,k_1;\ldots;j_d=0,\ldots,k_d}$ such that for each $j_1, \ldots, j_d$,

$$\theta^*_{j_1 \ldots j_d} = \begin{cases} 1 & \text{if } \left(j_1/k_1, \ldots, j_d/k_d\right) \in G^* \\ -1 & \text{otherwise.} \end{cases}$$

Note that θ^* is contained by $\widetilde{\Theta}$. Thus, it follows that

$$\inf_{f\in\mathcal{B}_{\mathbf{k}}} R^w_{\phi_h}(f) - \inf_{f\in\mathcal{F}_M} R^w_{\phi_h}(f) \leq R^w_{\phi_h}(B_{\mathbf{k}}(\theta^*,\cdot)) - R^w_{\phi_h}\left(\tilde{f}^*\right).$$

Then, applying Lemma D.3 to $R^w_{\phi_h}(B_{\mathbf{k}}(\theta^*,\cdot)) - R^w_{\phi_h}(\tilde{f}^*)$ shows the result (i).

The inequality in Lemma D.4 (ii) follows immediately from Lemma D.3. Applying Lemma D.2 (iii) to any $\hat{f}_B \in \arg\inf_{f\in\mathbf{B}_{\mathbf{k}}} \hat{R}_{\phi_h}(f)$ shows that a classifier

$$\hat{f}^{\dagger}_B(x) = \sum_{j_1=1}^{k_1} \cdots \sum_{j_{d_x}=1}^{k_{d_x}} \operatorname{sign}\left(\hat{\theta}_{j_1\cdots j_{d_x}}\right)\left(b_{k_1 j_1}(x_1) \times \cdots \times b_{k_{d_x} j_{d_x}}(x_{d_x})\right)$$

minimizes $\hat{R}^w_{\phi_h}(\cdot)$ over $\mathcal{B}_{\mathbf{k}}$, which proves the existence of $\hat{f}_B \in \arg\inf_{f\in\mathbf{B}_{\mathbf{k}}} \hat{R}^w_{\phi_h}(f)$ such that its coefficients take values in $\{-1, 1\}$. □

References

Alexander, K. S. (1984): "Probability inequalities for empirical processes and a law of the iterated logarithm," *Annals of Probability*, 12, 1041–1067.

Athey, S. and S. Wager (2021): "Policy learning with observational data," *Econometrica*, 89, 133–161.

Babii, A., X. Chen, E. Ghysels, and R. Kumar (2020): "Binary Choice with Asymmetric Loss in a Data-Rich Environment: Theory and an Application to Racial Justice," *arXiv*.

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

Beygelzimer, A. and J. Langford (2009): "The offset tree for learning with partial labels," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 129–137.

Bhattacharya, D. and P. Dupas (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167, 168–196.

Boucheron, S., G. Lugosi, and P. Massart (2013): *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford, UK: Oxford University Press.

Breiman, L., J. Friedman, C. Stone, and R. Olshen (1984): *Classification and Regression Trees*, The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis.

Calders, T. and S. Verwer (2010): "Three naive Bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, 21, 277–292.

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

Chamberlain, G. (2011): "Bayesian aspects of treatment choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, Oxford University Press, 11–39.

Chen, C. C. and S. T. Li (2014): "Credit rating with a monotonicity-constrained support vector machine model," *Expert Systems with Applications*, 41, 7235–7247.

Chouldechova, A. and A. Roth (2018): "The frontiers of fairness in machine learning," ArXiv:1810.08810.

Cortes, C. and V. Vapnik (1995): "Support-vector networks," *Machine Learning*, 20, 273–297.

Dehejia (2005): "Program evaluation as a decision problem," *Journal of Econometrics*, 125, 141–173.

Devroye, L., L. Györfi, and G. Lugosi (1996): *A Probabilistic Theory of Pattern Recognition*, Springer.

Donini, M., L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil (2018): "Empirical risk minimization under fairness constraints," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2796–2806.

Dudley, R. M. (1999): *Uniform Central Limit Theorems*, Cambridge University Press.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012): "Fairness through awareness," in *Proceedings of the 2012 Innovations in Theoretical Computer Science Conference (ITCS 2012)*, 214–226.

Feldman, M., S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015): "Certifying and removing disparate impact," in *Proceedings of the 2015 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Freund, Y. and R. Schapire (1997): "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer System Sciences*, 55, 119–139.

Gao, F. and J. A. Wellner (2007): "Entropy estimate for high-dimensional monotonic functions," *Journal of Multivariate Analysis*, 1751–1764.

Giné, E. and R. Nickl (2016): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, New York: Cambridge University Press.

Hirano, K. and J. R. Porter (2009): "Asymptotics for statistical treatment rules," *Econometrica*, 77, 1683–1701.

Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica: journal of the Econometric Society*, 505–531.

Jiang, W. (2004): "Process consistency for adaboost," *Annals of Statistics*, 32, 13–29.

Kallus, N. (2020): "More efficient policy learning via optimal retargeting," *Journal of the American Statistical Association*, 1–13.

Kamishima, T., S. Akaho, and J. Sakuma (2011): "Fairness-aware learning through regularization approach," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM 2011)*, 643–650.

Kasy, M. (2018): "Optimal taxation and insurance using machine learning — Sufficient statistics and beyond," *Journal of Public Economics*, 167, 205–219.

Kitagawa, T. and A. Tetenov (2018): "Who should be treated? Empirical welfare maximization methods for treatment choice," *Econometrica*, 86, 591–616.

——— (2021): "Equality-Minded Treatment Choice," *Journal of Business Economics and Statistics*, 39, 561–574.

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

Lieli, R. P. and H. White (2010): "The Construction of Empirical Credit Scoring Rules Based on Maximization Principles," *Journal of Econometrics*, 157, 110–119.

Lugosi, G. (2002): "Pattern classification and learning theory," in *Principles of Nonparametric Learning*, ed. by L. Györfi, Vienna: Springer, 1–56.

Lugosi, G. and N. Vayatis (2004): "On the Bayes-risk consistency of regularized boosting methods," *Annals of Statistics*, 32, 30–55.

Mammen, E. and A. B. Tsybakov (1999): "Smooth discrimination analysis," *Annals of Statistics*, 27, 1808–1829.

Mannor, S., R. Meir, and T. Zhang (2003): "Greedy algorithms for classification – consistency, convergence rates, and adaptivity," *Journal of Machine Learning Research*, 4, 713–742.

Manski, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

——— (2004): "Statistical treatment rules for heterogeneous populations," *Econometrica*, 72, 1221–1246.

Manski, C. F. and T. Thompson (1989): "Estimation of Best Predictors of Binary Response," *Journal of Econometrics*, 40, 97–123.

Mbakop, E. and M. Tabord-Meehan (2021): "Model selection for treatment choice: Penalized welfare maximization," *Econometrica*, 89, 825–848.

Nguyen, X., M. J. Wainwright, and M. I. Jordan (2009): "On surrogate loss functions and f-divergences," *Annals of Statistics*, 37, 876–904.

Qian, M. and S. A. Murphy (2011): "Performance guarantees for individualized treatment rules," *Annals of statistics*, 39, 1180–1210.

Rambachan, A., J. Kleinberg, J. Ludwig, and S. Mullainathan (2020): "An Economic Approach to Regulating Algorithms," *NBER Working paper*.

Scott, C. (2012): "Calibrated asymmetric surrogate losses," *Electronic Journal of Statistics*, 6, 958–992.

Steinwart, I. (2005): "Consistency of support vector machines and other regularized kernel classifiers," *IEEE Transactions on Information Theory*, 51, 713–742.

——— (2007): "How to compare different loss functions and their risks," *Constructive Approximation*, 26, 225–287.

Stoye, J. (2009): "Minimax regret treatment choice with finite samples," *Journal of Econometrics*, 151, 70–81.

——— (2012): "Minimax regret treatment choice with covariates or with limited validity of experiments," *Journal of Econometrics*, 166, 138–156.

Swaminathan, A. and T. Joachims (2015): "Counterfactual risk minimization: Learning from logged bandit feedback," *Journal of Machine Learning Research*, 16, 1731–1755.

Tetenov, A. (2012): "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 166, 157–165.

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

Vapnik, V. N. (1998): *Statistical Learning Theory*, John Wiley & Sons.

Viviano, D. (2019): "Policy Targeting under Network Interference," *arXiv*.

Wang, J. and S. K. Ghosh (2012): "Shape restricted nonparametric regression based on multivariate Bernstein polynomials," North Carolina State University Department of Statistics Technical report.

Zadrozny, B. (2003): "Policy mining: Learning decision policies from fixed sets of data," *Ph.D Thesis, University of California, San Diego*.

Zeng, J., B. Ustun, and C. Rudin (2017): "Interpretable classification models for recidivism prediction," *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 180, 689–722.

Zhang, T. (2004): "Statistical behavior and consistency of classification methods based on convex risk minimization," *Annals of Statistics*, 32, 56–85.

Zhang, Y., E. B. Laber, M. Davidian, and A. A. Tsiatis (2018): "Interpretable dynamic treatment regimes," *Journal of the American Statistical Association*, 113, 1541–1549.

Zhao, Y., D. Zeng, E. B. Laber, and M. R. Kosorok (2015): "New statistical learning methods for estimating optimal dynamic treatment regimes," *Journal of the American Statistical Association*, 110, 583–598.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012): "Estimating individualized treatment rules using outcome weighted learning," *Journal of the American Statistical Association*, 107, 1106–1118.