Efficient PAC Learning for Realizable-Statistic Models via Convex Surrogates

Shivani Agarwal
University of Pennsylvania
ashivani@seas.upenn.edu

Abstract

A central question in the theory of machine learning concerns the identification of classes of data distributions for which one can provide computationally efficient learning algorithms with provable statistical learning guarantees. Indeed, in the context of probably approximately correct (PAC) learning, there has been much interest in exploring intermediate PAC learning models that, unlike the realizable PAC learning setting, allow for some stochasticity in the labels, and unlike the fully agnostic PAC learning setting, also admit computationally efficient learning algorithms with finite sample complexity bounds. Some examples of such models include random classification noise (RCN), probabilistic concepts, Massart noise, and generalized linear models (GLMs); in general, most of this work has focused on binary classification problems. In this paper, we study what we call *realizable-statistic models* (RSMs), wherein we allow stochastic labels but assume that some vector-valued statistic of the conditional label distribution comes from some known function class. RSMs are a flexible class of models that interpolate between the realizable and fully agnostic settings, and that also recover several previously studied models as special cases. We show that for a broad range of RSM learning problems, where the statistic of interest can be accurately estimated via a convex 'strongly proper composite' surrogate loss, minimizing this convex surrogate loss yields a computationally efficient learning algorithm with finite sample complexity bounds. We then apply this result to show that various commonly used (and in some cases, not so commonly used) convex surrogate risk minimization algorithms yield computationally efficient learning algorithms with finite sample complexity bounds for a variety of RSM learning problems including binary classification, multiclass classification, multi-label prediction, and subset ranking. For the special case of binary classification with sigmoid-of-linear class probabilities (also a special case of GLMs), our results show that minimizing the standard binary logistic loss has a similar sample complexity as the GLM-tron algorithm of Kakade et al. (2011), but is computationally more efficient. In terms of the distribution over the domain/instance space, our results are all distribution-independent. To our knowledge, these are the first such results for PAC learning with stochastic labels for such a broad range of learning problems.

1 Introduction

The probably approximately correct (PAC) learning model is a cornerstone in the theory of machine learning. The two most widely studied settings, namely the realizable and fully agnostic settings, both represent somewhat extreme tradeoffs between computational efficiency and statistical modeling power: The realizable setting, as originally proposed by Valiant [38], often admits computationally efficient learning algorithms, but makes the restrictive statistical assumption that examples are labeled by a deterministic target function (from some known function class); the (fully) agnostic setting [23, 29] allows for fully general joint probability distributions on the labeled examples, but often fails to admit computationally efficient learning algorithms. Consequently, there has been much interest

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

in exploring intermediate PAC learning models that both allow for some stochasticity in the labels, and admit computationally efficient learning algorithms with finite sample complexity bounds. Some examples of such models include random classification noise (RCN) [4, 13, 10, 17, 27, 21, 22, 30, 20], probabilistic concepts [28], Massart noise [36, 37, 35, 31, 6, 7, 8, 40, 45, 19, 15, 14], and (univariate) generalized linear models (GLMs) and single index models (SIMs) [26, 25]. In general, most of this work has focused on binary classification problems.

In this paper, we study what we call *realizable-statistic models* (RSMs), wherein we allow stochastic labels but assume that some vector-valued statistic of the conditional label distribution comes from some known function class. RSMs are a flexible class of models that interpolate between the realizable and fully agnostic settings, and that also recover several previously studied models as special cases. We show that for a broad range of RSM learning problems, where the statistic of interest can be accurately estimated via a convex 'strongly proper composite' surrogate loss, minimizing this convex surrogate loss yields a computationally efficient learning algorithm with finite sample complexity bounds. We then apply this result to show that various commonly used (and in some cases, not so commonly used) convex surrogate risk minimization algorithms yield computationally efficient learning algorithms with finite sample complexity bounds for a variety of RSM learning problems including binary classification, multiclass classification, multi-label prediction, and subset ranking. In terms of the distribution over the domain/instance space, our results are all distribution-independent.

Technically, our work involves the following components. First, after defining RSMs, we define the notion of 'strongly proper composite' surrogate losses for estimating a desired statistic $\boldsymbol{\tau}$ (generalizing previous definitions of strongly proper composite surrogate losses for binary and multiclass class probability estimation [3, 42]).[1] Second, we give a general surrogate regret transfer bound for any RSM learning problem for which the statistic of interest can be accurately estimated via a strongly proper composite surrogate loss; this allows us to upper bound the target loss based regret in terms of the surrogate regret. Third, we use uniform convergence techniques to upper bound the surrogate regret of an (approximate) surrogate risk minimization algorithm, thus also upper bounding the target loss based regret for such an algorithm. We give two such results: one using d_1 covering numbers, and the other using Rademacher complexities. For the result in terms of Rademacher complexities, we make use of a vector-contraction inequality due to [32] to upper bound the Rademacher complexities of the loss function class $\psi_{\mathcal{F}}$ associated with a vector-valued function class $\mathcal{F}$ and a surrogate loss ψ (that acts on vector-valued predictions and is Lipschitz w.r.t. the Euclidean metric) in terms of the Rademacher complexities of the real-valued projection classes $\mathcal{F}^j$. For the result in terms of d_1 covering numbers, we give a (to our knowledge, new) technical lemma that upper bounds the d_1 covering numbers of the loss function class $\psi_{\mathcal{F}}$ associated with a vector-valued function class $\mathcal{F}$ and a surrogate loss ψ (that acts on vector-valued predictions and is Lipschitz w.r.t. the L^1 metric) in terms of the d_1 covering numbers of the projection classes $\mathcal{F}^j$; this lemma may also be of independent interest. Finally, we show how these results can be applied to a variety of RSM learning problems.

While our results are broadly applicable to many RSM formulations, for each of the applications we consider, we include specific instantiations to RSM learning problems with sigmoid/softmax-of-(multi-)linear forms for the statistics of interest, which can also be viewed as (multivariate) GLMs (see Table 1 for a summary). For the applications to binary classification (with 0-1 loss), multi-label learning (with Hamming loss), and subset ranking (with discounted cummulative gain (DCG) based loss), the Rademacher complexity based result gives tighter sample complexity bounds than those based on d_1 covering numbers. For the application to multiclass classification (with 0-1 loss), the two results are complementary: for n classes and data dimension p, the d_1 covering number based result gives a dimension-dependent sample complexity bound of $\widetilde{O}(np/\epsilon^2)$ for achieving squared estimation error $\leq \epsilon$; the Rademacher complexity based result gives a dimension-independent bound of $\widetilde{O}(n^2/\epsilon^2)$. For the special case of binary classification with sigmoid-of-linear class probabilities, our results show that minimizing the standard binary logistic loss has a similar sample complexity as the GLM-tron algorithm of Kakade et al. (2011), but is computationally more efficient. In particular, the sample complexity for achieving squared estimation error $\leq \epsilon$ is $\widetilde{O}(1/\epsilon^2)$ for both algorithms; however, the computational complexity of GLM-tron is $\widetilde{O}(p/\epsilon^3)$, whereas that of the logistic regression algorithm is $\widetilde{O}(p/\epsilon^{5/2})$.

[1]We note that the usage of the term 'proper' here is related to that in 'proper scoring rules' in the probability forecasting literature (see for example [12, 33, 34, 39, 1, 2] and references therein), and is distinct from that in 'proper learner' as commonly used in the PAC learning literature.

Table 1: Summary of selected PAC learning results with stochastic labels (results selected for comparison with ours, which are shown in red). Note that in terms of the distribution on the domain $\mathcal{X}$, the results shown here are all distribution-independent. Here LTF stands for 'linear threshold function'. See Appendix A for details of the assumptions associated with RCN, Massart noise, GLM, and SIM. See Section 2 for details of notation used in the last row. The computational complexities listed for RSMs all assume implementations using Nesterov's accelerated gradient descent (AGD).

Assumption on conditional label distribution $\mathbf{P}(Y\|X=x)$	**Learning target**	**Sample complexity (for squared estimation error** $\leq \epsilon$**)**	**Sample complexity (for target loss based regret** $\leq \epsilon$**)**	**Computational complexity (**$m =$ **sample complexity from column 3 or 4)**
Binary classification with 0-1 loss $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}]$				
Noisy LTF: RCN [10, 17, 21]	Best LTF		$\mathrm{poly}(p, 1/\epsilon)$	$\mathrm{poly}(p, 1/\epsilon)$
Noisy LTF: Massart noise [15]	Upper bound η on Massart noise		$\widetilde{O}(\mathrm{poly}(p)/\epsilon^3)$	$\mathrm{poly}(p, 1/\epsilon)$
GLM [25]	Best LTF	$\widetilde{O}(1/\epsilon^2)$		$\widetilde{O}(m^{3/2}p)$
SIM [25]	Best LTF	*(i)* $\widetilde{O}(p/\epsilon^3)$ *(ii)* $\widetilde{O}(1/\epsilon^4)$		*(i)* $\widetilde{O}(m^{4/3}p)$ *(ii)* $\widetilde{O}(m^{5/4}p)$
Sigmoid-of-linear [as special case of RSMs]	Best LTF	$\widetilde{O}(1/\epsilon^2)$	$\widetilde{O}(1/\epsilon^4)$	$\widetilde{O}(m^{5/4}p)$
Multiclass classification with 0-1 loss (n classes) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = [n]]$				
Softmax-of-multilinear [as special case of RSMs]	Best multilinear multiclass classifier	*(i)* $\widetilde{O}(np/\epsilon^2)$ *(ii)* $\widetilde{O}(n^2/\epsilon^2)$	*(i)* $\widetilde{O}(np/\epsilon^4)$ *(ii)* $\widetilde{O}(n^2/\epsilon^4)$	*(i)* $\widetilde{O}(m^{5/4}np)$ *(ii)* $\widetilde{O}(m^{5/4}np)$
Multi-label prediction with Hamming loss (s tags) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}^s]$				
Sigmoid-of-linear marginals [as special case of RSMs]	Best multilinear multi-label prediction model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$
Subset ranking with DCG metric (s items, r rating levels) $[\mathcal{X} \subseteq \mathbb{R}^p, \mathcal{Y} = \{0,1,\ldots,r\}^s, \widehat{\mathcal{Y}} = \Pi_s]$				
Sigmoid-of-linear scaled marginal expectations [as special case of RSMs]	Best multilinear subset ranking model	$\widetilde{O}(s^3/\epsilon^2)$	$\widetilde{O}(r^4 s^5/\epsilon^4)$	$\widetilde{O}(m^{5/4}sp)$
General learning problem (general $\mathcal{X}, \mathcal{Y}, \widehat{\mathcal{Y}}$**) with general loss matrix** $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y} \times \widehat{\mathcal{Y}}}$				
RSM: $\boldsymbol{\tau} \circ \mathbf{p} \in \mathcal{Q}$, where $\mathbf{p} : \mathcal{X} \to \Delta_{\mathcal{Y}}$ with $p_y(x) = \mathbf{P}(Y = y\|X = x)$, $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathbb{R}^d$ and $\mathcal{Q} \subseteq (\mathbb{R}^d)^{\mathcal{X}}$	Best prediction model in $\mathcal{H} \subseteq \widehat{\mathcal{Y}}^{\mathcal{X}}$, where $\mathcal{H} = \mathrm{pred} \circ \mathcal{Q}$ for $\mathrm{pred} : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ s.t. $(\boldsymbol{\tau}, \mathrm{pred})$ is $\mathbf{L}$-calibrated	$\widetilde{O}\left(\frac{\rho_2^2 d^2 + B^2}{\gamma^2 \epsilon^2}\right)$ where $\mathcal{R}_m(\mathcal{F}^j) \leq \frac{C}{\sqrt{m}}$	$\widetilde{O}\left(\frac{\kappa^4(\rho_2^2 d^2 + B^2)}{\gamma^2 \epsilon^4}\right)$ where $\mathcal{R}_m(\mathcal{F}^j) \leq \frac{C}{\sqrt{m}}$	$\widetilde{O}(m^{5/4}t)$ where $t =$ number of parameters to be learned

Organization of the paper. Section 2 sets up the learning problem, defines RSMs, and gives our main results. Sections 3–6 then apply our results to binary classification, multiclass classification, multi-label learning, and subset ranking, respectively. All proofs can be found in the Appendix.

Notation. We denote by $\mathbb{Z}_+$ the positive integers, and denote $\mathbb{R}_+ = [0, \infty)$, $\mathbb{R}_{++} = (0, \infty)$, $\overline{\mathbb{R}} = [-\infty, \infty]$, $\overline{\mathbb{R}}_+ = [0, \infty]$. For a positive integer n, $[n] := \{1, \ldots, n\}$, and $\Pi_n = \{\pi : [n] \to [n] \,|\, \pi \text{ is a bijection}\}$. For a matrix $\mathbf{A}$, we denote by $\mathbf{a}_j$ the j-th column vector of $\mathbf{A}$. For a finite set $\mathcal{Y}$, $\Delta_{\mathcal{Y}} := \{\mathbf{p} \in \mathbb{R}_+^{\mathcal{Y}} \,|\, \sum_{y \in \mathcal{Y}} p_y = 1\}$; for $\mathcal{Y} = [n]$, abbreviate $\Delta_n := \Delta_{[n]}$. We denote by $\mathbf{1}(\cdot)$ the indicator function. For a vector $\mathbf{u} \in \mathbb{R}^n$, $\mathrm{argsort}(\mathbf{u}) := \{\pi \in \Pi_n \,|\, u_i > u_j \implies \pi(i) < \pi(j)\}$. For two vectors $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n$, the d_1 distance between them is $d_1(\mathbf{u}_1, \mathbf{u}_2) := \frac{1}{n}\|\mathbf{u}_1 - \mathbf{u}_2\|_1$. We use $\mathcal{N}_1$ to denote d_1 covering numbers. For a set $\mathcal{X}$, a class of real-valued functions $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}\}$, an integer $m \in \mathbb{Z}_+$, and an underlying probability distribution μ on $\mathcal{X}$, we denote the Rademacher complexity of $\mathcal{F}$ for sample size m as $\mathcal{R}_m(\mathcal{F}) := \mathbf{E}_{(X_1,\ldots,X_m)\sim\mu^m}[\mathbf{E}_{(\epsilon_1,\ldots,\epsilon_m)}[\sup_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^m \epsilon_i f(X_i)]]$, where ϵ_i are i.i.d. Rademacher random variables (each taking values $+1$ or -1 with probability $\frac{1}{2}$ each). For a set $\mathcal{C}$, an objective function $f : \mathcal{C} \to \mathbb{R}$, and a positive real number $\alpha > 0$, an α-approximate minimizer of f over $\mathcal{C}$ returns a solution $\widehat{c} \in \mathcal{C}$ satisfying $f(\widehat{c}) \leq \inf_{c \in \mathcal{C}} f(c) + \alpha$.

2 Realizable-Statistic Models (RSMs) and Main Results

Section 2.1 sets up the learning problem and formally defines RSMs. Section 2.2 starts by defining some useful tools and then gives our main results.

2.1 Realizable-Statistic Models (RSMs)

Problem setup. We will consider a fairly general supervised learning setup. Specifically, let $\mathcal{X}$ be an instance space, and $\mathcal{Y}, \widehat{\mathcal{Y}}$ be finite label and prediction spaces, respectively.[2] Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$ be a target loss function, where for each $y \in \mathcal{Y}$, $\widehat{y} \in \widehat{\mathcal{Y}}$, the loss $\ell(y, \widehat{y})$ is the cost of predicting $\widehat{y}$ when the true label is y; equivalently, we will represent the loss function via a loss matrix $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$, with $(y, \widehat{y})$-th element given by $L_{y,\widehat{y}} = \ell(y, \widehat{y})$. Let $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$ be a joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, which we will often write as $D = (\mu, \mathbf{p})$, where $\mu \in \Delta_{\mathcal{X}}$ is the marginal of D over $\mathcal{X}$ and $\mathbf{p} : \mathcal{X} \to \Delta_{\mathcal{Y}}$ denotes the conditional distribution over $\mathcal{Y}$ given an instance in $\mathcal{X}$. Given a training sample $S = ((X_1, Y_1), \ldots, (X_m, Y_m))$ containing labeled examples drawn i.i.d. from D, the goal is to learn a prediction model $h : \mathcal{X} \to \widehat{\mathcal{Y}}$ with small expected loss on a new example drawn from D, which we will refer to as the $\mathbf{L}$-*error* or $\mathbf{L}$-*risk* of h: $\mathrm{er}_D^{\mathbf{L}}[h] = \mathbf{E}_{(X,Y)\sim D}[L_{Y,h(X)}]$. In particular, for a class of models $\mathcal{H} \subseteq \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ and a class of probability distributions $\mathcal{D} \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}}$, a learning algorithm $\mathcal{A}$ that maps training samples $S \in \cup_{m=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^m$ to prediction models $\widehat{h}_S \in \mathcal{H}$ is a *probably approximately correct (PAC) learning algorithm for the learning problem* $(\mathbf{L}, \mathcal{H}, \mathcal{D})$ *with* ***target loss sample complexity function*** $m_{\mathcal{A}}^{\mathbf{L}} : \mathbb{R}_+ \times (0, 1] \to \mathbb{Z}_+$ if for every $\epsilon > 0, \delta \in (0, 1]$, every probability distribution $D \in \mathcal{D}$ and every $m \geq m_{\mathcal{A}}^{\mathbf{L}}(\epsilon, \delta)$, $\mathbf{P}_{S\sim D^m}\big(\mathrm{er}_D^{\mathbf{L}}[\widehat{h}_S] - \inf_{h\in\mathcal{H}} \mathrm{er}_D^{\mathbf{L}}[h] > \epsilon\big) < \delta$, and moreover, for every ϵ, δ, $m_{\mathcal{A}}^{\mathbf{L}}(\epsilon, \delta)$ is the smallest integer satisfying the above. We will sometimes denote by $\mathrm{er}_D^{\mathbf{L}}[\mathcal{H}] := \inf_{h\in\mathcal{H}} \mathrm{er}_D^{\mathbf{L}}[h]$ the *best* $\mathbf{L}$-*error for* D *within* $\mathcal{H}$.

Realizable-statistic models (RSMs). The essence of our realizable-statistic models (RSMs) is to allow the labels to be stochastic and assume that some (vector-valued) 'statistic' of the conditional label distribution $\mathbf{p}(x) = \big(\mathbf{P}(Y = y|X = x)\big)_{y\in\mathcal{Y}}$ (associated with the underlying data distribution D) belongs to some class of (vector-valued) functions $\mathcal{Q}$; in other words, we will assume that a statistic $\boldsymbol{\tau}$ of the conditional label distribution $\mathbf{p}(x)$ is '$\mathcal{Q}$-realizable'. Formally, for any $\mathcal{C} \subseteq \mathbb{R}^d$ and d-dimensional statistic $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$, and any class of functions $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X} \to \mathcal{C}\}$, define the class of $(\boldsymbol{\tau}, \mathcal{Q})$-RSM distributions over $\mathcal{X} \times \mathcal{Y}$ as follows:

$$\mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}} = \big\{D = (\mu, \mathbf{p}) \in \Delta_{\mathcal{X}\times\mathcal{Y}} \;\big|\; \exists\, \mathbf{q} \in \mathcal{Q} \text{ s.t. } \boldsymbol{\tau}(\mathbf{p}(x)) = \mathbf{q}(x)\ \forall x \in \mathcal{X}\big\}\,.$$

We will be interested in solving learning problems of the form $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$. We note that the realizable and (fully) agnostic PAC learning models can both be recovered as special cases of RSMs; all the previously studied intermediate PAC learning models listed in Table 1 can also be recovered as special cases of RSMs (see Appendix B). Our algorithms for solving (certain types of) RSM learning problems of the form $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$ will typically do the following: given a training sample S, they will first (sometimes implicitly) find an estimate $\widehat{\mathbf{q}}_S : \mathcal{X} \to \mathcal{C}$ for the true statistic function $\mathbf{q}^*(x) = \boldsymbol{\tau}(\mathbf{p}(x))$, and then will return a prediction model $\widehat{h}_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ effectively constructed from $\widehat{\mathbf{q}}_S$. Accordingly, for such an algorithm $\mathcal{A}$, in addition to its target loss sample complexity $m_{\mathcal{A}}^{\mathbf{L}}$ defined above, we will also be interested in its ***squared $\boldsymbol{\tau}$-estimation error sample complexity function*** $m_{\mathcal{A}}^{\boldsymbol{\tau}} : \mathbb{R}_+ \times (0, 1] \to \mathbb{Z}_+$, where for every $\epsilon > 0, \delta \in (0, 1]$, $m_{\mathcal{A}}^{\boldsymbol{\tau}}(\epsilon, \delta)$ is the smallest integer such that every probability distribution $D \in \mathcal{D}$ and every $m \geq m_{\mathcal{A}}^{\boldsymbol{\tau}}(\epsilon, \delta)$, $\mathbf{P}_{S\sim D^m}\big(\mathbf{E}_{X\sim\mu}[\|\widehat{\mathbf{q}}_S(X) - \mathbf{q}^*(X)\|_2^2] > \epsilon\big) < \delta$.

2.2 Main Results

We start by defining some tools that will be needed for our main results – specifically, the tools of $\mathbf{L}$-calibrated statistics and strongly proper composite surrogate losses. Before doing so, we recall:

Definition 1 (Bayes L-error and Bayes L-optimal model). *Let* $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$ *be any loss matrix. The* Bayes $\mathbf{L}$-error for D, *denoted* $\mathrm{er}_D^{\mathbf{L},*}$, *is the smallest* $\mathbf{L}$-*error under* D *over all possible prediction models:* $\mathrm{er}_D^{\mathbf{L},*} = \inf_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} \mathrm{er}_D^{\mathbf{L}}[h]$. *A* Bayes $\mathbf{L}$-optimal model for D, *denoted* $h_D^{\mathbf{L},*} : \mathcal{X} \to \widehat{\mathcal{Y}}$, *is any prediction model that achieves the Bayes* $\mathbf{L}$-*error for* D*:* $\mathrm{er}_D^{\mathbf{L}}\big[h_D^{\mathbf{L},*}\big] = \mathrm{er}_D^{\mathbf{L},*}$.

[2] Our model and results easily extend to more general $\mathcal{Y}, \widehat{\mathcal{Y}}$; we take these to be finite for simplicity.

Definition 2 (**L-calibrated statistics [2]**). *Let* $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$ *be any loss matrix. Let* $d \in \mathbb{Z}_+$ *and* $\mathcal{C} \subseteq \mathbb{R}^d$. *A statistic* $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$ *is* $\mathbf{L}$-calibrated *if* $\exists$ *a mapping* $\text{pred} : \mathcal{C} \to \widehat{\mathcal{Y}}$ *such that for all distributions* $D = (\mu, \mathbf{p}) \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, *a Bayes* $\mathbf{L}$*-optimal model for* D *can be obtained from* $\boldsymbol{\tau}(\mathbf{p}(x))$ *as* $h_D^{\mathbf{L},*}(x) = \text{pred}(\boldsymbol{\tau}(\mathbf{p}(x)))$. *We will also say the statistic-mapping pair* $(\boldsymbol{\tau}, \text{pred})$ *is* $\mathbf{L}$*-calibrated.*

The convex surrogate risk minimization algorithms we will consider will minimize the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$, for some suitably defined convex surrogate loss $\psi : \mathcal{Y} \times \mathcal{C}' \to \mathbb{R}_+$ that acts on vector predictions in some convex set $\mathcal{C}' \subseteq \mathbb{R}^{d'}$ (for a suitable integer d'), over some class of vector-valued functions $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathcal{C}'\}$ to learn a vector-valued function $\widehat{\mathbf{f}}_S \in \mathcal{F}$, and then will return a prediction model $\widehat{h}_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ of the form $\widehat{h}_S(x) = \text{decode}(\widehat{\mathbf{f}}_S(x))$ for a suitable decoding function $\text{decode} : \mathcal{C}' \to \widehat{\mathcal{Y}}$. We will be especially interested in surrogate losses whose minimization yields accurate estimates of a desired statistic $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$. To this end, we define below the notion of strongly proper (composite) surrogate losses ψ for a statistic $\boldsymbol{\tau}$, for which the expected surrogate loss $\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, \mathbf{u})]$ is 'strongly' minimized at (possibly an invertible transformation of) the correct statistic value $\boldsymbol{\tau}(\mathbf{p})$; this generalizes the definition of strongly proper (composite) surrogate losses for binary and multiclass class probability estimation [3, 42] to estimation of general statistics:[3]

Definition 3 (**Strongly proper composite surrogate losses for a statistic** $\boldsymbol{\tau}$). *Let* $d \in \mathbb{Z}_+$ *and* $\mathcal{C} \subseteq \mathbb{R}^d$, *and let* $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$ *be any statistic of interest. Let* $d' \in \mathbb{Z}_+$, *and let* $\mathcal{C}' \subseteq \mathbb{R}^{d'}$ *be such that* $\mathcal{C}$ *is in one-to-one correspondence with a subset of* $\mathcal{C}'$. *If* $\mathcal{C}$ *is in one-to-one correspondence with* $\mathcal{C}'$ *itself, then let* $\boldsymbol{\lambda} : \mathcal{C} \to \mathcal{C}'$ *be an invertible mapping with inverse* $\boldsymbol{\lambda}^{-1} : \mathcal{C}' \to \mathcal{C}$; *otherwise, let* $\boldsymbol{\lambda} : \mathcal{C} \to \mathcal{C}'$ *be a one-to-one mapping and let* $\mathcal{S} = \{\mathcal{S}_{\mathbf{q}} : \mathbf{q} \in \mathcal{C}\}$ *be a partition of* $\mathcal{C}'$ *such that* $\boldsymbol{\lambda}(\mathbf{q}) \in \mathcal{S}_{\mathbf{q}}\ \forall \mathbf{q} \in \mathcal{C}$, *and let* $\boldsymbol{\lambda}^{-1} : \mathcal{C}' \to \mathcal{C}$ *denote an 'extended' inverse that assigns* $\boldsymbol{\lambda}^{-1}(\mathbf{u}) = \mathbf{q}\ \forall \mathbf{u} \in \mathcal{S}_{\mathbf{q}}$. *Let* $\gamma > 0$. *A surrogate loss* $\psi : \mathcal{Y} \times \mathcal{C}' \to \mathbb{R}_+$ *acting on* $\mathcal{C}'$ *is* γ-strongly proper composite for statistic $\boldsymbol{\tau}$ with link function $\boldsymbol{\lambda}$ *if* $\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, \mathbf{u}) - \psi(Y, \boldsymbol{\lambda}(\boldsymbol{\tau}(\mathbf{p})))] \geq \frac{\gamma}{2}\|\boldsymbol{\lambda}^{-1}(\mathbf{u}) - \boldsymbol{\tau}(\mathbf{p})\|_2^2\ \forall \mathbf{p} \in \Delta_{\mathcal{Y}},\ \mathbf{u} \in \mathcal{C}'$.

We are now ready to state our main results. We start by giving a general surrogate regret transfer bound for RSM learning problems for which the statistic of interest admits a strongly proper composite surrogate loss; this allows us to upper bound the target loss based regret in terms of the surrogate regret. Specifically, the theorem below effectively shows that given a target loss $\mathbf{L}$, an $\mathbf{L}$-calibrated statistic-mapping pair $(\boldsymbol{\tau}, \text{pred})$ satisfying a certain condition (which allows the $\mathbf{L}$-regret to be upper-bounded by the squared $\boldsymbol{\tau}$-estimation error), a class of 'statistic' functions $\mathcal{Q}$, and a strongly proper composite surrogate loss ψ for $\boldsymbol{\tau}$ with link function $\boldsymbol{\lambda}$, for any data distribution $D \in \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}}$, both the squared $\boldsymbol{\tau}$-estimation error of any $\mathbf{q} \in \mathcal{Q}$ and the target $\mathbf{L}$-regret (excess $\mathbf{L}$-risk) of a model $h = \text{pred} \circ \mathbf{q}$ in the class of models $\mathcal{H} = \text{pred} \circ \mathcal{Q}$ can be upper bounded in terms of the surrogate ψ-regret (excess ψ-risk) of the vector-valued function $\mathbf{f} = \boldsymbol{\lambda} \circ \mathbf{q}$ in the class of vector-valued functions $\mathcal{F} = \boldsymbol{\lambda} \circ \mathcal{Q}$. The proof of this theorem is inspired by the proof of a surrogate regret transfer bound given in a different context (Bayes consistent multi-label learning with the F-measure) by [43].

Theorem 1 (**Surrogate regret transfer bound for RSMs that admit strongly proper composite surrogate losses**). *Let* $\mathcal{X}$ *be any instance space and* $\mathcal{Y}, \widehat{\mathcal{Y}}$ *be any label and prediction spaces, respectively. Let* $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$ *be a loss matrix. Let* $d \in \mathbb{Z}_+$ *and* $\mathcal{C} \subseteq \mathbb{R}^d$. *Let* $\boldsymbol{\tau} : \Delta_{\mathcal{Y}} \to \mathcal{C}$ *and* $\text{pred} : \mathcal{C} \to \widehat{\mathcal{Y}}$ *be such that* $(\boldsymbol{\tau}, \text{pred})$ *is an* $\mathbf{L}$*-calibrated statistic-mapping pair, and suppose* $\exists \kappa > 0$ *s.t.*

$$\mathbf{E}_{Y\sim\mathbf{p}}[L_{Y,\text{pred}(\mathbf{q})}] - \min_{\widehat{y}\in\mathcal{Y}} \mathbf{E}_{Y\sim\mathbf{p}}[L_{Y,\widehat{y}}] \ \leq\ \kappa\|\mathbf{q} - \boldsymbol{\tau}(\mathbf{p})\|_2 \quad \forall \mathbf{p} \in \Delta_{\mathcal{Y}}, \mathbf{q} \in \mathcal{C}\,.$$

Let $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X} \to \mathcal{C}\}$ *be a class of 'statistic' functions, and let* $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ *be a* γ*-strongly proper composite surrogate loss for* $\boldsymbol{\tau}$ *with link function* $\boldsymbol{\lambda} : \mathcal{C} \to \mathbb{R}^d$.[4] *Let* $\mathcal{H} \subseteq \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ *be defined as* $\mathcal{H} := \text{pred} \circ \mathcal{Q} = \{h : \mathcal{X} \to \widehat{\mathcal{Y}} \mid \exists \mathbf{q} \in \mathcal{Q} \text{ s.t. } h(x) = \text{pred}(\mathbf{q}(x))\ \forall x \in \mathcal{X}\}$, *let* $\mathcal{F} \subseteq \{f : \mathcal{X} \to \mathbb{R}^d\}$ *be defined as* $\mathcal{F} := \boldsymbol{\lambda} \circ \mathcal{Q} = \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d \mid \exists \mathbf{q} \in \mathcal{Q} \text{ s.t. } \mathbf{f}(x) = \boldsymbol{\lambda}(\mathbf{q}(x))\ \forall x \in \mathcal{X}\}$, *and*

[3]The reason for introducing a new space $\mathcal{C}' \subseteq \mathbb{R}^{d'}$ is that often it is easier to minimize a surrogate loss acting on a space $\mathcal{C}'$ different from $\mathcal{C}$ (in many of our examples, we will have $\mathcal{C} \subsetneq \mathbb{R}^d$, $d' = d$ and $\mathcal{C}' = \mathbb{R}^d$).

[4]As in Definition 3, if $\mathcal{C}$ is in one-to-one correspondence with $\mathbb{R}^d$ itself, then we will assume that $\boldsymbol{\lambda} : \mathcal{C} \to \mathbb{R}^d$ is an invertible mapping with inverse $\boldsymbol{\lambda}^{-1} : \mathbb{R}^d \to \mathcal{C}$; otherwise, we will assume that $\boldsymbol{\lambda} : \mathcal{C} \to \mathbb{R}^d$ is a one-to-one mapping and $\mathcal{S} = \{\mathcal{S}_{\mathbf{q}} : \mathbf{q} \in \mathcal{C}\}$ is a partition of $\mathbb{R}^d$ such that $\boldsymbol{\lambda}(\mathbf{q}) \in \mathcal{S}_{\mathbf{q}}\ \forall \mathbf{q} \in \mathcal{C}$, and $\boldsymbol{\lambda}^{-1} : \mathbb{R}^d \to \mathcal{C}$ denotes an 'extended' inverse that assigns $\boldsymbol{\lambda}^{-1}(\mathbf{u}) = \mathbf{q}\ \forall \mathbf{u} \in \mathcal{S}_{\mathbf{q}}$. Note that in the notation of Definition 3, here we have set $d' = d$ and $\mathcal{C}' = \mathbb{R}^d$ (this is both for simplicity and because this suffices for our examples); however, the theorem easily extends to any suitable d' and $\mathcal{C}'$.

define $\text{decode} : \mathbb{R}^d \to \widehat{\mathcal{Y}}$ *as* $\text{decode} := \text{pred} \circ \boldsymbol{\lambda}^{-1}$. *Suppose that* $\psi(y, \mathbf{f}(x)) \in [0, B]\ \forall x \in \mathcal{X}, y \in \mathcal{Y}, \mathbf{f} \in \mathcal{F}$ *for some* $B > 0$. *Then for any* $\mathbf{f} \in \mathcal{F}$ *and any* $D \in \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}}$,

$$\mathrm{er}_D^{\mathbf{L}}[\underbrace{\text{decode} \circ \mathbf{f}}_{h}] - \mathrm{er}_D^{\mathbf{L}}[\mathcal{H}] \;\leq\; \kappa \cdot \sqrt{\mathbf{E}_X[\|\boldsymbol{\lambda}^{-1}(\mathbf{f}(X)) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2^2]} \;\leq\; \kappa \cdot \sqrt{\tfrac{2}{\gamma}(\mathrm{er}_D^{\psi}[\mathbf{f}] - \mathrm{er}_D^{\psi}[\mathcal{F}])}\,.$$

In practice, when applying the above theorem, it will often be the case that the class of 'statistic' functions $\mathcal{Q}$ is of the form $\mathcal{Q} = \boldsymbol{\sigma} \circ \mathcal{F}$ for some pre-specified class of vector-valued functions $\mathcal{F}$ (such as bounded multi-linear functions) and some 'transfer' function $\boldsymbol{\sigma}$; in such settings, it can be helpful to choose a strongly proper composite surrogate loss whose inverse link function $\boldsymbol{\lambda}^{-1}$ is matched to $\boldsymbol{\sigma}$ (we will see several examples of this in the next few sections).

The above result can be combined with any upper bound on the surrogate ψ-regret in $\mathcal{F}$ to yield upper bounds on both the squared $\boldsymbol{\tau}$-estimation error and the target $\mathbf{L}$-regret in $\mathcal{H}$, which in turn can then be converted to sample complexity bounds. The following two results make this concrete for standard unregularized surrogate risk minimization; the first result makes use of d_1 covering numbers, while the second makes use of Rademacher complexities. For the result in terms of d_1 covering numbers, we make use of standard uniform convergence techniques, together with a (to our knowledge, new) technical lemma (given in Appendix B) that upper bounds the d_1 covering numbers of the loss function class $\psi_{\mathcal{F}} = \{\psi_{\mathbf{f}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \,|\, \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } \psi_{\mathbf{f}}(x, y) = \psi(y, \mathbf{f}(x))\}$ associated with a vector-valued function class $\mathcal{F}$ and a surrogate loss ψ (that acts on vector-valued predictions and is Lipschitz w.r.t. the L^1 metric) in terms of the d_1 covering numbers of the real-valued projection function classes $\{\mathcal{F}^j\}_j$ (defined below); this lemma may also be of independent interest. For the result in terms of Rademacher complexities, we make use of uniform convergence techniques, together with a vector-contraction inequality due to [32] that upper bounds the Rademacher complexities of the loss function class $\psi_{\mathcal{F}}$ associated with a vector-valued function class $\mathcal{F}$ and a surrogate loss ψ (that acts on vector-valued predictions and is Lipschitz w.r.t. the Euclidean metric) in terms of the Rademacher complexities of the real-valued projection classes $\{\mathcal{F}^j\}_j$.

Theorem 2 (**RSM learning bounds for surrogate risk minimizers via d_1 covering numbers**). *Under the conditions of Theorem 1, suppose the surrogate loss ψ is ρ_1-Lipschitz in the second argument with respect to the L^1 metric, so that $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \leq \rho_1 \|\mathbf{u}_1 - \mathbf{u}_2\|_1\ \forall y, \mathbf{u}_1, \mathbf{u}_2$, and suppose that the function classes $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \,|\, \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j\ \forall x\}$, $j \in [d]$ each have bounded d_1 covering numbers $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m)$ (polynomial in m and $1/\epsilon$). Then a surrogate risk minimization algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(16B/\sqrt{m})$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over $\mathcal{F}$, and produces a $\boldsymbol{\tau}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\lambda}^{-1}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S \in \mathcal{H}$ given by $\widehat{h}_S(x) = \text{decode}(\widehat{\mathbf{f}}_S(x))$ (or equivalently, $\widehat{h}_S(x) = \text{pred}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$ with squared $\boldsymbol{\tau}$-estimation error sample complexity $m_{\mathcal{A}}^{\boldsymbol{\tau}}(\epsilon, \delta) \leq \min\big\{m_0 \in \mathbb{Z}_+ : m \geq m_0 \implies m \geq \frac{1152B^2}{\gamma^2\epsilon^2}\big(\sum_{j=1}^d \ln\big(\mathcal{N}_1\big(\frac{\gamma\epsilon}{48\rho_1 d}, \mathcal{F}^j, 2m\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)\big\}$, and with target loss sample complexity $m_{\mathcal{A}}^{\mathbf{L}}(\epsilon, \delta) \leq \min\big\{m \in \mathbb{Z}_+ : m \geq m_0 \implies m \geq \frac{1152\kappa^4B^2}{\gamma^2\epsilon^4}\big(\sum_{j=1}^d \ln\big(\mathcal{N}_1\big(\frac{\gamma\epsilon^2}{48\kappa^2\rho_1 d}, \mathcal{F}^j, 2m\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)\big\}$. In particular, if the d_1 covering numbers of the function classes $\mathcal{F}^j$ have upper bounds of the form $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m) \leq \phi(\epsilon, \mathcal{F}^j)$ (i.e., bounds independent of sample size m), then $m_{\mathcal{A}}^{\boldsymbol{\tau}}(\epsilon, \delta) \leq \frac{1152B^2}{\gamma^2\epsilon^2}\big(\sum_{j=1}^d \ln\big(\phi\big(\frac{\gamma\epsilon}{48\rho_1 d}, \mathcal{F}^j\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)$, and $m_{\mathcal{A}}^{\mathbf{L}}(\epsilon, \delta) \leq \frac{1152\kappa^4B^2}{\gamma^2\epsilon^4}\big(\sum_{j=1}^d \ln\big(\phi\big(\frac{\gamma\epsilon^2}{48\kappa^2\rho_1 d}, \mathcal{F}^j\big)\big) + \ln\big(\frac{4}{\delta}\big)\big)$.*

Theorem 3 (**RSM learning bounds for surrogate risk minimizers via Rademacher complexities**). *Under the conditions of Theorem 1, suppose the surrogate loss ψ is ρ_2-Lipschitz in the second argument with respect to the Euclidean metric, so that $\psi(y, \mathbf{u}_1) - \psi(y, \mathbf{u}_2) \leq \rho_2 \|\mathbf{u}_1 - \mathbf{u}_2\|_2\ \forall y, \mathbf{u}_1, \mathbf{u}_2$, and suppose that the function classes $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \,|\, \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j\ \forall x\}$, $j \in [d]$ each have non-negative, decreasing Rademacher complexities $\mathcal{R}_m(\mathcal{F}^j)$ (decreasing in m). Then a surrogate risk minimization algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(B/(2\sqrt{m}))$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over $\mathcal{F}$, and produces a $\boldsymbol{\tau}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\lambda}^{-1}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S \in \mathcal{H}$ given by $\widehat{h}_S(x) = \text{decode}(\widehat{\mathbf{f}}_S(x))$ (or equivalently, $\widehat{h}_S(x) = \text{pred}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}})$ with squared $\boldsymbol{\tau}$-estimation error sample complexity $m_{\mathcal{A}}^{\boldsymbol{\tau}}(\epsilon, \delta) \leq \min\big\{m_0 \in \mathbb{Z}_+ : m \geq$*

$m_0 \implies 3\Big(2\sqrt{2}\rho_2 \cdot \sum_{j=1}^d \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \le \frac{\gamma\epsilon}{2}\big\}$*, and with target loss sample complexity* $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon,\delta) \le \min\big\{m \in \mathbb{Z}_+ : m \ge m_0 \implies 3\Big(2\sqrt{2}\rho_2 \cdot \sum_{j=1}^d \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \le \frac{\gamma\epsilon^2}{2\kappa^2}\big\}$*. In particular, if* $\exists C > 0$ *such that the Rademacher complexities of the function classes* $\mathcal{F}^j$ *have upper bounds of the form* $\mathcal{R}_m(\mathcal{F}^j) \le C/\sqrt{m}\ \forall j \in [d]$*, then* $m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon,\delta) \le \frac{36}{\gamma^2\epsilon^2}\big(2\sqrt{2}\rho_2 Cd + B\sqrt{\ln(2/\delta)}\,\big)^2$*, and* $m^{\mathbf{L}}_{\mathcal{A}}(\epsilon,\delta) \le \frac{36\kappa^4}{\gamma^2\epsilon^4}\big(2\sqrt{2}\rho_2 Cd + B\sqrt{\ln(2/\delta)}\,\big)^2$*.*

In Sections 3–6 below, we apply the above results to a variety of RSM learning problems, including binary classification, multiclass classification, multi-label prediction, and subset ranking. While our results are broadly applicable to many RSM formulations, for each of the applications below, we will include specific instantiations to RSM learning problems with sigmoid/softmax-of-(multi-)linear forms for the statistics of interest. To this end, we will make use of the following upper bounds on the d_1 covering numbers and the Rademacher complexity of (bounded) linear functions:

Proposition 4. *Let* $R, W > 0$*. Let* $\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 \le R\}$*. Let* $\mathcal{F}_{\text{linear}} = \{\mathbf{f} : \mathcal{X}\to\mathbb{R} \mid \exists \mathbf{w} \in \mathbb{R}^p, \|\mathbf{w}\|_2 \le W \text{ s.t. } \mathbf{f}(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}\ \forall \mathbf{x}\}$*. Then for any* $m \in \mathbb{Z}_+$ *and any* $\epsilon > 0$*:*
(i) $\mathcal{N}_1(\epsilon, \mathcal{F}_{\text{linear}}, m) \le (1/\epsilon)^p$*; (ii)* $\mathcal{N}_1(\epsilon, \mathcal{F}_{\text{linear}}, m) \le (4R^2W^2/\epsilon^2 + 1)^{\lceil 2R^2W^2/\epsilon^2\rceil}$*; and (iii)* $0 \le \mathcal{R}_m(\mathcal{F}_{\text{linear}}) \le RW/\sqrt{m}$*.*

3 Binary Classification

Consider a binary classification problem with instance space $\mathcal{X}$, label and prediction spaces $\mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm 1\}$, and the standard 0-1 loss $\mathbf{L}^{\text{0-1}} \in \mathbb{R}_+^{\{\pm1\}\times\{\pm1\}}$ with $\ell^{\text{0-1}}(y,\widehat{y}) = \mathbf{1}(\widehat{y} \ne y)$. Let $\mathcal{C} = [0,1]$, and define the 'projection-onto-$(+1)$th-component' statistic $\tau^{+1} : \Delta_{\{\pm1\}}\to[0,1]$ and mapping $\text{pred}^{\text{0-1}} : [0,1]\to\{\pm1\}$ as

$$\tau^{+1}\big(\mathbf{p} \equiv (p_{+1}, p_{-1})^\top\big) = p_{+1}\,; \qquad \text{pred}^{\text{0-1}}(q) = \text{sign}(q - 1/2)\,.$$

Then $(\tau^{+1}, \text{pred}^{\text{0-1}})$ is an $\mathbf{L}^{\text{0-1}}$-calibrated pair. Moreover, as is well known (also see Appendix C),

$$\mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}}_{Y,\text{pred}^{\text{0-1}}(q)}] - \min_{\widehat{y}\in\{\pm1\}} \mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}}_{Y,\widehat{y}}] \ \le\ 2\,|q - p_{+1}| \quad \forall \mathbf{p} \in \Delta_{\{\pm1\}}, q \in [0,1]\,.$$

Therefore, for any class of 'statistic' functions $\mathcal{Q} \subseteq \{q : \mathcal{X}\to[0,1]\}$ and corresponding hypothesis class $\mathcal{H} = \text{pred}^{\text{0-1}} \circ \mathcal{Q}$, Theorems 2 and 3 establish that any convex surrogate risk minimization algorithm minimizing a strongly proper composite surrogate loss for τ^{+1} over a suitable class of functions $\mathcal{F} \subseteq \{f : \mathcal{X}\to\mathbb{R}\}$ yields an efficient PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{0-1}}, \mathcal{H}, \mathcal{D}_{(\tau^{+1},\mathcal{Q})\text{-RSM}})$. While this result can be applied to any class $\mathcal{Q}$ and suitable surrogate loss ψ, the following theorem makes this concrete for the class of sigmoid-of-linear models $\mathcal{Q}_{\text{sigmoid-of-linear}}$ and the binary logistic loss $\psi^{\log}$ (defined below).

Theorem 5 (PAC learning algorithm for binary classification with sigmoid-of-linear class probabilities). *Consider a binary classification problem, with* $\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 \le R\}$ *for some* $R > 0$*,* $\mathcal{Y} = \widehat{\mathcal{Y}} = \{\pm1\}$*, and with the standard 0-1 loss* $\mathbf{L}^{\text{0-1}}$ *as above. Let* τ^{+1} *and* $\text{pred}^{\text{0-1}}$ *be as defined above. Let* $\sigma : \mathbb{R}\to[0,1]$ *be the sigmoid function* $\sigma(u) = 1/(1+e^{-u})$*, and let*

$$\mathcal{Q}_{\text{sigmoid-of-linear}} = \{q : \mathcal{X}\to[0,1] \mid \exists \mathbf{w} \in \mathbb{R}^p, \|\mathbf{w}\|_2 \le W \text{ s.t. } q(\mathbf{x}) = \sigma(\mathbf{w}^\top\mathbf{x})\ \forall \mathbf{x}\}$$

for some $W > 0$*. Let* $\mathcal{H}_{\text{linear}} := \text{pred}^{\text{0-1}} \circ \mathcal{Q}_{\text{sigmoid-of-linear}}$*, i.e.* $\mathcal{H}_{\text{linear}} = \{h : \mathcal{X}\to\{\pm1\} \mid \exists \mathbf{w} \in \mathbb{R}^p, \|\mathbf{w}\|_2 \le W \text{ s.t. } h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top\mathbf{x})\ \forall \mathbf{x}\}$*. Let* $\psi^{\log} : \{\pm1\}\times\mathbb{R}\to\mathbb{R}_+$ *be the binary logistic loss:*

$$\psi^{\log}(y,u) \ =\ \ln(1 + e^{-yu})\,.$$

Let $\mathcal{F}_{\text{linear}} = \{f : \mathcal{X}\to\mathbb{R} \mid \exists \mathbf{w} \in \mathbb{R}^p, \|\mathbf{w}\|_2 \le W \text{ s.t. } f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}\ \forall \mathbf{x}\}$*. Then an algorithm* $\mathcal{A}$ *which, given a training sample* S *of size* m*, finds an* $(\ln(1+e^{RW})/(2\sqrt{m}))$*-approximate minimizer* $\widehat{f}_S \in \mathcal{F}_{\text{linear}}$ *of the empirical surrogate risk* $\frac{1}{m}\sum_{i=1}^m \psi^{\log}(y_i, f(\mathbf{x}_i))$ *over* $\mathcal{F}_{\text{linear}}$*, and produces a* τ^{+1}*-statistic estimate* $\widehat{q}_S(x) = \sigma(\widehat{f}_S(x))$ *and prediction model* $\widehat{h}_S \in \mathcal{H}_{\text{linear}}$ *given by* $\widehat{h}_S = \text{sign} \circ \widehat{f}_S$ *(equivalently,* $\widehat{h}_S = \text{pred}^{\text{0-1}} \circ \widehat{q}_S$*), is a PAC learning algorithm for the RSM learning problem* $(\mathbf{L}^{\text{0-1}}, \mathcal{H}_{\text{linear}}, \mathcal{D}_{(\tau^{+1},\mathcal{Q}_{\text{sigmoid-of-linear}})\text{-RSM}})$ *with squared* τ^{+1}*-estimation error sample complexity* $m^{\tau^{+1}}_{\mathcal{A}}(\epsilon,\delta) = O\big(\frac{1}{\epsilon^2}\ln\big(\frac{1}{\delta}\big)\big)$*, and with target loss sample complexity* $m^{\mathbf{L}^{\text{0-1}}}_{\mathcal{A}}(\epsilon,\delta) = O\big(\frac{1}{\epsilon^4}\ln\big(\frac{1}{\delta}\big)\big)$*.*

4 Multiclass Classification

Consider now a multiclass classification problem with instance space $\mathcal{X}$, label and prediction spaces $\mathcal{Y} = \widehat{\mathcal{Y}} = [n]$ for $n > 2$, and the multiclass 0-1 loss $\mathbf{L}^{\text{0-1}(n)} \in \mathbb{R}_+^{n\times n}$ with $\ell^{\text{0-1}}(y,\widehat{y}) = \mathbf{1}(\widehat{y} \neq y)$. Let $\mathcal{C} = \Delta_n$, and define the 'identity' statistic $\boldsymbol{\tau}^{\text{id}} : \Delta_n \to \Delta_n$ and mapping $\text{pred}^{\text{0-1}(n)} : \Delta_n \to [n]$ as

$$\boldsymbol{\tau}^{\text{id}}(\mathbf{p}) = \mathbf{p}\,; \qquad \text{pred}^{\text{0-1}(n)}(\mathbf{q}) = \text{argmax}_{\widehat{y}\in[n]}\, q_{\widehat{y}}\,.$$

Then $(\boldsymbol{\tau}^{\text{id}}, \text{pred}^{\text{0-1}(n)})$ is an $\mathbf{L}^{\text{0-1}(n)}$-calibrated pair. Moreover, as shown in Appendix D,

$$\mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\text{pred}^{\text{0-1}(n)}(\mathbf{q})}] - \min_{\widehat{y}\in[n]}\mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\widehat{y}}] \;\leq\; \sqrt{2}\cdot\|\mathbf{q}-\mathbf{p}\|_2 \quad \forall \mathbf{p},\mathbf{q}\in\Delta_n\,.$$

Therefore, for any class of 'statistic' functions $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X}\to\Delta_n\}$ and corresponding hypothesis class $\mathcal{H} = \text{pred}^{\text{0-1}(n)} \circ \mathcal{Q}$, Theorems 2 and 3 establish that any convex surrogate risk minimization algorithm minimizing a strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{id}}$ over a suitable class of functions $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X}\to\mathbb{R}^n\}$ yields an efficient PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{0-1}(n)}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{id}},\mathcal{Q})\text{-RSM}})$. While this result can be applied to any class $\mathcal{Q}$ and suitable surrogate loss ψ, the following theorem makes this concrete for the class of softmax-of-multilinear models $\mathcal{Q}_{\text{softmax-of-multilinear}}$ and the multiclass logistic loss ψ^{mlog} (defined below).

Theorem 6 (PAC learning algorithm for multiclass classification with softmax-of-multilinear class probabilities). *Consider a multiclass classification problem, with $\mathcal{X} \subseteq \{\mathbf{x}\in\mathbb{R}^p \mid \|\mathbf{x}\|_2 \leq R\}$ for some $R > 0$, $\mathcal{Y} = \widehat{\mathcal{Y}} = [n]$, and with the multiclass 0-1 loss $\mathbf{L}^{\text{0-1}(n)}$ as above. Let $\boldsymbol{\tau}^{\text{id}}$ and $\text{pred}^{\text{0-1}(n)}$ be as defined above. Let $\boldsymbol{\sigma} : \mathbb{R}^n\to\Delta_n$ be the softmax function $(\boldsymbol{\sigma}(\mathbf{u}))_y = e^{u_y}/(\sum_{y'=1}^n e^{u_{y'}})\ \forall y\in[n]$, and let*

$$\mathcal{Q}_{\text{softmax-of-mlinear}} \;=\; \{\mathbf{q} : \mathcal{X}\to\Delta_n \mid \exists \mathbf{W}\in\mathbb{R}^{p\times n}, \|\mathbf{w}_y\|_2 \leq W\ \forall y \text{ s.t. } \mathbf{q}(\mathbf{x}) = \boldsymbol{\sigma}(\mathbf{W}^\top\mathbf{x})\ \forall \mathbf{x}\}$$

for some $W > 0$. Let $\mathcal{H}_{\text{multiclass-linear}} := \text{pred}^{\text{0-1}(n)} \circ \mathcal{Q}_{\text{softmax-of-mlinear}}$, i.e. $\mathcal{H}_{\text{multiclass-linear}} = \{h : \mathcal{X}\to[n] \mid \exists \mathbf{W}\in\mathbb{R}^{p\times n}, \|\mathbf{w}_y\|_2 \leq W\ \forall y \text{ s.t. } h(\mathbf{x}) \in \text{argmax}_{y\in[n]}(\mathbf{w}_y^\top\mathbf{x})\ \forall\mathbf{x}\}$. Let $\psi^{\text{mlog}} : [n]\times\mathbb{R}^n\to\mathbb{R}_+$ be the multiclass logistic loss

$$\psi^{\text{mlog}}(y,\mathbf{u}) \;=\; -u_y + \ln(\textstyle\sum_{y'=1}^n e^{u_{y'}})\,.$$

Define $\text{decode}^{\text{0-1}(n)} : \mathbb{R}^n\to[n]$ as $\text{decode}^{\text{0-1}(n)}(\mathbf{u}) \in \text{argmax}_{\widehat{y}\in[n]}\, u_{\widehat{y}}$, and let $\mathcal{F}_{\text{multiclass-linear}} = \{\mathbf{f} : \mathcal{X}\to\mathbb{R}^n \mid \exists\mathbf{W}\in\mathbb{R}^{p\times n}, \|\mathbf{w}_y\|_2 \leq W\ \forall y \text{ s.t. } \mathbf{f}(\mathbf{x}) = \mathbf{W}^\top\mathbf{x}\ \forall\mathbf{x}\}$. Then an algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $((\ln(n) + 2RW)/(2\sqrt{m}))$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}_{\text{multiclass-linear}}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi^{\text{mlog}}(y_i,\mathbf{f}(\mathbf{x}_i))$ over $\mathcal{F}_{\text{multiclass-linear}}$, and produces a $\boldsymbol{\tau}^{\text{id}}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\sigma}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S \in \mathcal{H}_{\text{multiclass-linear}}$ given by $\widehat{h}_S(\mathbf{x}) = \text{decode}^{\text{0-1}(n)}(\widehat{\mathbf{f}}_S(\mathbf{x}))$ (or equivalently, $\widehat{h}_S(x) = \text{pred}^{\text{0-1}(n)}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{0-1}(n)}, \mathcal{H}_{\text{multiclass-linear}}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{id}},\mathcal{Q}_{\text{softmax-of-mlinear}})\text{-RSM}})$ with squared $\boldsymbol{\tau}^{\text{id}}$-estimation error sample complexity $m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{id}}}(\epsilon,\delta)$ and target loss sample complexity $m_{\mathcal{A}}^{\mathbf{L}^{\text{0-1}(n)}}(\epsilon,\delta)$ upper bounded as follows:
(i) (Dimension-dependent)
$m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{id}}}(\epsilon,\delta) = O\big(\frac{(\ln n)^2}{\epsilon^2}\left(np\ln\left(\frac{n}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\big)$; $m_{\mathcal{A}}^{\mathbf{L}^{\text{0-1}(n)}}(\epsilon,\delta) = O\big(\frac{(\ln n)^2}{\epsilon^4}\left(np\ln\left(\frac{n}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\big)$.
(ii) (Dimension-independent)
$m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{id}}}(\epsilon,\delta) = O\left(\frac{1}{\epsilon^2}\left(n^2 + (\ln(n))^2\cdot\ln\left(\frac{1}{\delta}\right)\right)\right)$; $m_{\mathcal{A}}^{\mathbf{L}^{\text{0-1}(n)}}(\epsilon,\delta) = O\left(\frac{1}{\epsilon^4}\left(n^2 + (\ln(n))^2\cdot\ln\left(\frac{1}{\delta}\right)\right)\right)$.

5 Multi-Label Learning

Next, consider a multi-label prediction problem such as in image tagging, with s tags $[s] = \{1,\ldots,s\}$, several of which can be active in an instance simultaneously, and the goal is to predict for a new instance which of the s tags are active. Specifically, let $\mathcal{X}$ be any instance space, with label and prediction spaces $\mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}^s$ (labels are represented as vectors $\mathbf{y}\in\{0,1\}^s$, with $y_j = 1$ indicating that the j-th tag is active), and consider the Hamming loss $\mathbf{L}^{\text{Ham}} \in \mathbb{R}_+^{\{0,1\}^s\times\{0,1\}^s}$ with $\ell^{\text{Ham}}(\mathbf{y},\widehat{\mathbf{y}}) = \sum_{j=1}^s \mathbf{1}(\widehat{y}_j \neq y_j)$. Let $\mathcal{C} = [0,1]^s$, and define the s-dimensional 'marginals' statistic

$\boldsymbol{\tau}^{\text{marginals}} : \Delta_{\{0,1\}^s} \to [0,1]^s$ and mapping $\text{pred}^{\text{Ham}} : [0,1]^s \to \{0,1\}^s$ as

$$(\boldsymbol{\tau}^{\text{marginals}}(\mathbf{p}))_j = \sum_{\mathbf{y}\in\{0,1\}^s : y_j = 1} p_{\mathbf{y}} \,; \qquad (\text{pred}^{\text{Ham}}(\mathbf{q}))_j \;=\; \text{sign}(q_j - 1/2) \qquad \forall j \in [s]\,.$$

Then $(\boldsymbol{\tau}^{\text{marginals}}, \text{pred}^{\text{Ham}})$ is an $\mathbf{L}^{\text{Ham}}$-calibrated pair. Moreover, as shown in Appendix E,

$$\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\text{pred}^{\text{Ham}}(\mathbf{q})}] - \min_{\widehat{\mathbf{y}}\in\{0,1\}^s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\widehat{\mathbf{y}}}] \;\leq\; 2\sqrt{s}\cdot\|\mathbf{q} - \boldsymbol{\tau}^{\text{marginals}}(\mathbf{p})\|_2 \quad \forall \mathbf{p} \in \Delta_{\{0,1\}^s}, \mathbf{q} \in [0,1]^s\,.$$

Therefore, for any class of 'statistic' functions $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X} \to [0,1]^s\}$ and corresponding hypothesis class $\mathcal{H} = \text{pred}^{\text{Ham}} \circ \mathcal{Q}$, Theorems 2 and 3 establish that any convex surrogate risk minimization algorithm minimizing a strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{marginals}}$ over a suitable class of functions $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^s\}$ yields an efficient PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{Ham}}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{marginals}}, \mathcal{Q})\text{-RSM}})$. While this result can be applied to any class $\mathcal{Q}$ and suitable surrogate loss ψ, below we make this concrete for the class of sigmoid-of-multilinear models $\mathcal{Q}_{\text{sigmoid-of-multilinear}}$ and the 'binary relevance' logistic-based multi-label surrogate loss ψ^{BRlog} (defined below).[5]

Theorem 7 (PAC learning algorithm for multi-label prediction with sigmoid-of-multilinear marginals). *Consider a multi-label prediction problem, with $\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 \leq R\}$ for some $R > 0$, $\mathcal{Y} = \widehat{\mathcal{Y}} = \{0,1\}^s$, and with the Hamming loss $\mathbf{L}^{\text{Ham}}$ as above. Let $\boldsymbol{\tau}^{\text{marginals}}$ and* pred^{Ham} *be as defined above. Let $\sigma : \mathbb{R} \to [0,1]$ be the sigmoid function $\sigma(u) = 1/(1+e^{-u})$, and let*

$$\mathcal{Q}_{\text{sigmoid-of-multilinear}} = \{\mathbf{q} : \mathcal{X} \to [0,1]^s \mid \exists \mathbf{W} \in \mathbb{R}^{p\times s}, \|\mathbf{w}_j\|_2 \leq W \;\forall j \text{ s.t. } q_j(\mathbf{x}) = \sigma(\mathbf{w}_j^\top \mathbf{x}) \;\forall \mathbf{x}, j\}$$

for some $W > 0$. Let $\mathcal{H}^{\text{sign}}_{\text{multilinear}} := \text{pred}^{\text{Ham}} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$, i.e. $\mathcal{H}^{\text{sign}}_{\text{multilinear}} = \{\mathbf{h} : \mathcal{X} \to \{0,1\}^s \mid \exists \mathbf{W} \in \mathbb{R}^{p\times s}, \|\mathbf{w}_j\|_2 \leq W \;\forall j \text{ s.t. } h_j(\mathbf{x}) = \text{sign}(\mathbf{w}_j^\top \mathbf{x}) \;\forall \mathbf{x}, j\}$. Let $\psi^{\text{BRlog}} : \{0,1\}^s \times \mathbb{R}^s \to \mathbb{R}_+$ be the 'binary relevance' logistic-based multi-label surrogate loss defined by

$$\psi^{\text{BRlog}}(\mathbf{y}, \mathbf{u}) \;=\; \textstyle\sum_{j=1}^s \ln(1 + e^{-(2y_j - 1)u_j})\,.$$

Define $\text{decode}^{\text{Ham}} : \mathbb{R}^s \to \{0,1\}^s$ *as* $(\text{decode}^{\text{Ham}}(\mathbf{u}))_j := \text{sign}(u_j) \;\forall j \in [s]$, *and let* $\mathcal{F}_{\text{multilinear}} = \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^s \mid \exists \mathbf{W} \in \mathbb{R}^{p\times s}, \|\mathbf{w}_j\|_2 \leq W \;\forall j \text{ s.t. } \mathbf{f}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x} \;\forall \mathbf{x}\}$. *Then an algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(s\ln(1+e^{RW})/(2\sqrt{m}))$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}_{\text{multilinear}}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^m \psi^{\text{BRlog}}(y_i, \mathbf{f}(\mathbf{x}_i))$ over $\mathcal{F}_{\text{multilinear}}$, and produces a $\boldsymbol{\tau}^{\text{marginals}}$-statistic estimate $(\widehat{\mathbf{q}}_S(\mathbf{x}))_j = \sigma((\widehat{\mathbf{f}}_S(\mathbf{x}))_j)$ and a prediction model $\widehat{h}_S \in \mathcal{H}^{\text{sign}}_{\text{multilinear}}$ given by $\widehat{h}_S(\mathbf{x}) =$* $\text{decode}^{\text{Ham}}(\widehat{\mathbf{f}}_S(\mathbf{x}))$ *(or equivalently, $\widehat{h}_S(x) \;=\;$* $\text{pred}^{\text{Ham}}(\widehat{\mathbf{q}}_S(x))$*), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{Ham}}, \mathcal{H}^{\text{sign}}_{\text{multilinear}}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{marginals}}, \mathcal{Q}_{\text{sigmoid-of-multilinear}})\text{-RSM}})$ with squared $\boldsymbol{\tau}^{\text{marginals}}$-estimation error sample complexity $m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{marginals}}}(\epsilon, \delta) \;=\; O\big(\frac{s^2}{\epsilon^2}\big(s + \ln\big(\frac{1}{\delta}\big)\big)\big)$, and with target loss sample complexity $m_{\mathcal{A}}^{\mathbf{L}^{\text{Ham}}}(\epsilon, \delta) = O\big(\frac{s^4}{\epsilon^4}\big(s + \ln\big(\frac{1}{\delta}\big)\big)\big)$.*

6 Subset Ranking

As a final example, consider a subset ranking problem such as those that arise in information retrieval, wherein each instance contains a query and a subset of s documents, together with some relevance judgments for each of the s documents as labels, and given a new instance containing a new query and a new subset of s documents, the goal is to find a good ranking of the s documents for that query. Specifically, let $\mathcal{X}$ be any instance space, and let the label space be $\mathcal{Y} = \{0,1,\ldots,r\}^s$, where each document is graded on a scale of 0 to r; the prediction space is $\widehat{\mathcal{Y}} = \Pi_s$. A widely used performance measure for such problems is the discounted cumulative gain (DCG); in loss form, one version of the DCG loss $\mathbf{L}^{\text{DCG}}$ is given by $\ell^{\text{DCG}}(\mathbf{y}, \widehat{\pi}) \;=\; Z - \sum_{j=1}^s y_j \cdot \text{disc}(\widehat{\pi}(j))$, where $\text{disc} : [s] \to [0,1]$ is a non-increasing 'discount' function that discounts documents placed lower in the ranking, often taken to be $\text{disc}(a) = 1/(\log_2(a+1))$, and Z is a constant that ensures non-negativity of the loss [18, 24]. Let $\mathcal{C} = [0,1]^s$, and define the s-dimensional 'scaled marginal expectations' property

[5] The 'binary relevance' approach effectively solves s binary problems, one for each tag [41, 11]. One could also apply Theorem 5 s times (drawing a fresh sample of size $O(\frac{s^2}{\epsilon^2}(\ln(\frac{s}{\delta})))$ for each tag), yielding a sample complexity of $O(\frac{s^3}{\epsilon^2}(\ln(\frac{s}{\delta})))$. The result of Theorem 7 improves over this by removing a multiplicative $\ln(s)$ factor. We also note that contrary to popular belief, Theorem 7 indicates that the binary relevance approach *does not* require the s tags to be conditionally independent given x in order to be an effective learning algorithm.

$\boldsymbol{\tau}^{\text{sc-marg-exp}} : \Delta_{\{0,1,\ldots,r\}^s} \to [0,1]^s$ and mapping $\text{pred}^{\text{DCG}} : [0,1]^s \to \Pi_s$ as

$$(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j = \frac{\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[Y_j]}{r} = \frac{1}{r}\sum_{k=0}^{r} k \cdot \Bigg(\sum_{\mathbf{y}\in\{0,1,\ldots,r\}^s : y_j = k} p_{\mathbf{y}}\Bigg)\,; \qquad \text{pred}^{\text{DCG}}(\mathbf{q}) \in \text{argsort}(\mathbf{q})\,.$$

Then $(\boldsymbol{\tau}^{\text{sc-marg-exp}}, \text{pred}^{\text{DCG}})$ is an $\mathbf{L}^{\text{DCG}}$-calibrated pair. Moreover, as shown in Appendix F,

$$\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\text{pred}^{\text{DCG}}(\mathbf{q})}] - \min_{\widehat{\pi}\in\Pi_s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\widehat{\mathbf{y}}}] \;\le\; 2r \cdot \|\mathbf{disc}\|_2 \cdot \|\mathbf{q} - \boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})\|_2 \quad \forall \mathbf{p}, \mathbf{q}\,,$$

where $\mathbf{disc} = (\text{disc}(1), \ldots, \text{disc}(s))^\top \in [0,1]^s$. Therefore, for any class of 'statistic' functions $\mathcal{Q} \subseteq \{\mathbf{q} : \mathcal{X} \to [0,1]^s\}$ and corresponding hypothesis class $\mathcal{H} = \text{pred}^{\text{DCG}} \circ \mathcal{Q}$, Theorems 2 and 3 establish that any convex surrogate risk minimization algorithm minimizing a strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ over a suitable class of functions $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^s\}$ yields an efficient PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{DCG}}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{sc-marg-exp}}, \mathcal{Q})\text{-RSM}})$. While this result can be applied to any class $\mathcal{Q}$ and suitable surrogate loss ψ, the following theorem makes this concrete for the class of sigmoid-of-multilinear models $\mathcal{Q}_{\text{sigmoid-of-multilinear}}$ and a suitably weighted multivariate logistic-based surrogate loss ψ^{wlog} that we introduce here (defined below).

Theorem 8 (**PAC learning algorithm for subset ranking with sigmoid-of-multilinear scaled marginal expectations**). *Consider a subset ranking problem, with $\mathcal{X} \subseteq \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 \le R\}$ for some $R > 0$, $\mathcal{Y} = \{0,1,\ldots,r\}^s$ and $\widehat{\mathcal{Y}} = \Pi_s$, and with the DCG loss $\mathbf{L}^{\text{DCG}}$ as above. Let $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ and* pred^{DCG} *be as defined above. Let σ and $\mathcal{Q}_{\text{sigmoid-of-multilinear}}$ be as defined in Theorem 7, and let $\mathcal{H}^{\text{sort}}_{\text{multilinear}} := \text{pred}^{\text{DCG}} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$, i.e. $\mathcal{H}^{\text{sort}}_{\text{multilinear}} = \{h : \mathcal{X} \to \Pi_s \mid \exists \mathbf{W} \in \mathbb{R}^{p\times s}, \|\mathbf{w}_j\|_2 \le W\ \forall j \text{ s.t. } h(\mathbf{x}) \in \text{argsort}(\mathbf{W}^\top\mathbf{x})\ \forall \mathbf{x}\}$. Let $\psi^{\text{wlog}} : \{0,1,\ldots,r\}^s \times \mathbb{R}^s \to \mathbb{R}_+$ be a multivariate weighted logistic-based surrogate loss defined by*

$$\psi^{\text{wlog}}(\mathbf{y}, \mathbf{u}) \;=\; \textstyle\sum_{j=1}^{s} \left(\frac{y_j}{r}\right) \cdot \ln(1 + e^{-u_j}) + \left(1 - \frac{y_j}{r}\right) \cdot \ln(1 + e^{u_j})\,.$$

Define $\text{decode}^{\text{DCG}} : \mathbb{R}^s \to \Pi_s$ *as* $\text{decode}^{\text{DCG}}(\mathbf{u}) \in \text{argsort}(\mathbf{u})$, *and let $\mathcal{F}_{\text{multilinear}}$ be as defined in Theorem 7. Then an algorithm $\mathcal{A}$ which, given a training sample S of size m, finds an $(s\ln(1+e^{RW})/(2\sqrt{m}))$-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}_{\text{multilinear}}$ of the empirical surrogate risk $\frac{1}{m}\sum_{i=1}^{m} \psi^{\text{wlog}}(y_i, \mathbf{f}(\mathbf{x}_i))$ over $\mathcal{F}_{\text{multilinear}}$, and produces a $\boldsymbol{\tau}^{\text{sc-marg-exp}}$-statistic estimate $(\widehat{\mathbf{q}}_S(\mathbf{x}))_j = \sigma((\widehat{\mathbf{f}}_S(\mathbf{x}))_j)$ and a prediction model $\widehat{h}_S \in \mathcal{H}^{\text{sort}}_{\text{multilinear}}$ given by $\widehat{h}_S(\mathbf{x}) = \text{decode}^{\text{DCG}}(\widehat{\mathbf{f}}_S(\mathbf{x}))$ (or equivalently, $\widehat{h}_S(x) = \text{pred}^{\text{DCG}}(\widehat{\mathbf{q}}_S(x))$), is a PAC learning algorithm for the RSM learning problem $(\mathbf{L}^{\text{DCG}}, \mathcal{H}^{\text{sort}}_{\text{multilinear}}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{sc-marg-exp}}, \mathcal{Q}_{\text{sigmoid-of-multilinear}})\text{-RSM}})$ with squared $\boldsymbol{\tau}^{\text{sc-marg-exp}}$-estimation error sample complexity $m^{\boldsymbol{\tau}^{\text{sc-marg-exp}}}_{\mathcal{A}}(\epsilon,\delta) = O\big(\frac{s^2}{\epsilon^2}\big(s + \ln\big(\frac{1}{\delta}\big)\big)\big)$, and with target loss sample complexity $m^{\mathbf{L}^{\text{DCG}}}_{\mathcal{A}}(\epsilon,\delta) = O\big(\frac{r^4 s^2 \cdot \|\mathbf{disc}\|_2^4}{\epsilon^4}\big(s + \ln\big(\frac{1}{\delta}\big)\big)\big) = O\big(\frac{r^4 s^4}{\epsilon^4}\big(s + \ln\big(\frac{1}{\delta}\big)\big)\big)$.*

7 Conclusion

We have studied a flexible class of intermediate PAC leaning models that we call *realizable-statistic models* (RSMs), wherein we allow labels to be stochastic but assume that some vector-valued statistic of the conditional label distribution comes from a known function class. RSMs interpolate between the realizable and fully agnostic settings, and also recover several previously studied intermediate PAC learning models as special cases. We have shown that for RSMs where the statistic of interest can be estimated via a convex 'strongly proper composite' surrogate loss, minimizing this convex surrogate loss yields a computationally efficient learning algorithm with finite sample complexity bounds, and have demonstrated applications of these results to a broad range of RSM learning problems including binary and multiclass classification, multi-label learning, and subset ranking.

RSMs are also connected to the structured prediction framework studied in [16], where the target loss function can be written as $\ell(y,\widehat{y}) = \boldsymbol{\phi}_1(y)^\top \mathbf{A} \boldsymbol{\phi}_2(\widehat{y})$ for some embedding functions $\boldsymbol{\phi}_1 : \mathcal{Y} \to \mathbb{R}^k$, $\boldsymbol{\phi}_2 : \widehat{\mathcal{Y}} \to \mathbb{R}^k$ and matrix $\mathbf{A} \in \mathbb{R}^{k\times k}$.[6] In particular, [16] effectively considers the 'conditional mean embedding' statistic $\mathbf{q}^*(x) = \mathbf{E}[\boldsymbol{\phi}_1(Y)|X = x]$, and assumes that this statistic belongs to some class of functions (such as multilinear functions or a vector-valued RKHS); this statistic is then estimated to produce $\widehat{\mathbf{q}}(x)$. Thus this setting can also be viewed as a special case of our RSM framework (indeed, the quadratic surrogate used in [16] is also a strongly proper composite surrogate for the above statistic; the target loss based sample complexity bounds of [16] are of the form $\widetilde{O}(\beta/\epsilon^4)$, where β captures problem-dependent parameters, and are therefore comparable to our bounds).

[6]More generally, [16] allows embedding into a Hilbert space $\mathcal{F}$.

Acknowledgments and Disclosure of Funding

Warm thanks to Rob Schapire for valuable discussions and suggestions related to this work; to Peter Bartlett for valuable pointers; and to Nishant Agarwal and Ananya Mukherjee for help with typesetting parts of this paper. Thanks also to the anonymous reviewers of this work for helpful comments. This material is based upon work supported in part by the US National Science Foundation (NSF) under Grant No. 1934876. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Jacob D. Abernethy and Rafael M. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

[2] Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, 2015.

[3] Shivani Agarwal. Surrogate regret bounds for the area under the ROC curve via strongly proper losses. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013.

[4] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Mach. Learn.*, 2(4):343–370, 1988.

[5] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[6] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. Efficient learning of linear separators under bounded noise. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 167–190. JMLR.org, 2015.

[7] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 152–192. JMLR.org, 2016.

[8] Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *J. ACM*, 63(6):50:1–50:27, 2017.

[9] Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, 2002.

[10] Avrim Blum, Alan M. Frieze, Ravi Kannan, and Santosh S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[11] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognit.*, 37(9):1757–1771, 2004.

[12] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation: Structure and applications. Technical report, University of Pennsylvania, November 2005.

[13] Tom Bylander. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory, COLT 1994*, pages 340–347. ACM, 1994.

[14] Gautam Chandrasekaran, Vasilis Kontonis, Konstantinos Stavropoulos, and Kevin Tian. Learning noisy halfspaces with a margin: Massart is no harder than random. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024.

[15] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Classification under misspecification: Halfspaces, generalized linear models, and evolvability. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.

[16] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A general framework for consistent structured prediction with implicit loss embeddings. *Journal of Machine Learning Research*, 21(98):1–67, 2020.

[17] Edith Cohen. Learning noisy perceptrons by a perceptron in polynomial time. In *38th Annual Symposium on Foundations of Computer Science, FOCS '97, 1997*, pages 514–523. IEEE Computer Society, 1997.

[18] David Cossock and Tong Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

[19] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with Massart noise. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 4751–4762, 2019.

[20] Ilias Diakonikolas, Mingchen Ma, Lisheng Ren, and Christos Tzamos. Statistical query hardness of multiclass linear classification with random classification noise. *arXiv* 2502.11413, 2025.

[21] John Dunagan and Santosh S. Vempala. A simple polynomial-time rescaling algorithm for solving linear programs. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 315–320. ACM, 2004.

[22] Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. New results for learning noisy parities and halfspaces. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006)*, pages 563–574. IEEE Computer Society, 2006.

[23] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput.*, 100(1):78–150, 1992.

[24] Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2000.

[25] Sham M. Kakade, Adam Kalai, Varun Kanade, and Ohad Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*, pages 927–935, 2011.

[26] Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

[27] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

[28] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994.

[29] Michael J. Kearns, Robert E. Schapire, and Linda Sellie. Toward efficient agnostic learning. *Mach. Learn.*, 17(2-3):115–141, 1994.

[30] Philip M. Long and Rocco A. Servedio. Random classification noise defeats all convex potential boosters. *Mach. Learn.*, 78(3):287–304, 2010.

[31] Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[32] Andreas Maurer. A vector-contraction inequality for rademacher complexities. *arXiv* 1605.00251, 2016.

[33] Mark D. Reid and Robert C. Williamson. Surrogate regret bounds for proper losses. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009*, 2009.

[34] Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

[35] Ronald L. Rivest and Robert H. Sloan. A formal model of hierarchical concept learning. *Inf. Comput.*, 114(1):88–114, 1994.

[36] Robert H. Sloan. Types of noise in data for concept learning. In *Proceedings of the First Annual Workshop on Computational Learning Theory, COLT '88*, pages 91–96. ACM/MIT, 1988.

[37] Robert H. Sloan. Corrigendum to types of noise in data for concept learning. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992*, page 450. ACM, 1992.

[38] Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

[39] Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In *Neural Information Processing Systems*, 2011.

[40] Songbai Yan and Chicheng Zhang. Revisiting perceptron: Efficient and label-optimal learning of halfspaces. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 1056–1066, 2017.

[41] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 26:1819–1837, 08 2014.

[42] Mingyuan Zhang, Jane H. Lee, and Shivani Agarwal. Learning from noisy labels with no change to the training process. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[43] Mingyuan Zhang, Harish Guruprasad Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label f-measure. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[44] Tong Zhang. Some theoretical results concerning the convergence of compositions of regularized linear functions. In *Advances in Neural Information Processing Systems 12*, pages 370–378, 1999.

[45] Yuchen Zhang, Percy Liang, and Moses Charikar. A hitting time analysis of stochastic gradient langevin dynamics. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1980–2022. PMLR, 2017.

Appendix

Organization of the Appendix. Appendix A is a supplement to Section 1 (introduction). Appendix B is a supplement to Section 2 (RSMs and main results). Appendix C is a supplement to Section 3 (binary classification). Appendix D is a supplement to Section 4 (multiclass classification). Appendix E is a supplement to Section 5 (multi-label learning). Appendix F is a supplement to Section 6 (subset ranking). Proofs of all the theorems in the main paper can be found in the relevant sections of this Appendix.

A Supplement to Section 1 (Introduction)

Here we give details of the assumptions associated with the previously studied intermediate PAC learning models listed in Table 1.

Noisy LTF with random classification noise (RCN): This model assumes that there is a weight vector $\mathbf{w} \in \mathbb{R}^p$ and a noise parameter $\eta \in (0, 1/2)$ such that for any instance $\mathbf{x}$, a deterministic binary label is first generated according to the sign of $\mathbf{w}^\top \mathbf{x}$, and then with probability η the label is flipped to the opposite sign. Equivalently, the model can be viewed as assuming that the conditional label distribution is of the form $\mathbf{P}(Y = 1|X = \mathbf{x}) = (1 - \eta) \cdot \mathbf{1}(\mathbf{w}^\top \mathbf{x} \geq 0) + \eta \cdot \mathbf{1}(\mathbf{w}^\top \mathbf{x} < 0)$.

Noisy LTF with Massart noise: This model assumes that there is a weight vector $\mathbf{w} \in \mathbb{R}^p$ and a "noise upper bound" parameter $\eta \in (0, 1/2)$ such that for any instance $\mathbf{x}$, a deterministic binary label is first generated according to the sign of $\mathbf{w}^\top \mathbf{x}$, and then with some (unknown) probability $\eta(\mathbf{x}) \leq \eta$ the label is flipped to the opposite sign. Equivalently, the model can be viewed as assuming that the conditional label distribution satisfies $\mathbf{P}(Y = 1|X = \mathbf{x}) \geq (1 - \eta)$ if $\mathbf{w}^\top \mathbf{x} \geq 0$ and $\mathbf{P}(Y = 1|X = \mathbf{x}) \leq \eta$ if $\mathbf{w}^\top \mathbf{x} < 0$.

Generalized linear model (GLM): The (univariate) GLMs considered in [26, 25] are for real-valued regression problems with bounded label spaces $\mathcal{Y} = \widehat{\mathcal{Y}} \subseteq [0, 1]$, and assume that there is a weight vector $\mathbf{w} \in \mathbb{R}^p$ such that $\mathbf{E}[Y \mid X = \mathbf{x}] = \theta(\mathbf{w}^\top \mathbf{x})$ for some known transfer function $\theta : \mathbb{R} \to [0, 1]$ (it is common to assume that θ is monotonically increasing and Lipschitz continuous). These include as a special case binary classification by setting $\mathcal{Y} = \{0, 1\}$.

Single index model (SIM): The assumption here is of a similar form as that for GLMs above, namely that $\mathbf{E}[Y|X = \mathbf{x}] = \theta(\mathbf{w}^\top \mathbf{x})$ for some weight vector $\mathbf{w}$ and transfer function $\theta : \mathbb{R} \to [0, 1]$; however unlike GLMs, where θ is assumed to be known, in SIMs, both the weight vector $\mathbf{w}$ and the transfer function θ are unknown (it is common to assume that θ is monotonically increasing and Lipschitz continuous).

B Supplement to Section 2 (RSMs and Main Results)

B.1 Realizable and Agnostic PAC Learning as Special Cases of RSMs

We note here that both realizable and (fully) agnostic PAC learning can be recovered as extreme cases of RSMs. In the following, for a finite set $\mathcal{Y}$ and $y \in \mathcal{Y}$, $\mathbf{e}_y \in \{0, 1\}^{\mathcal{Y}}$ denotes the unit vector with y-th element equal to 1 and all other elements equal to 0.

Example 1 (**Realizable PAC learning as RSM**). *Let $\widehat{\mathcal{Y}} = \mathcal{Y}$, and let $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis class/class of prediction models. Let $\mathcal{C} = \Delta_{\mathcal{Y}}$, and consider the identity property $\boldsymbol{\tau}^{\text{id}} : \Delta_{\mathcal{Y}} \to \Delta_{\mathcal{Y}}$ defined as $\boldsymbol{\tau}^{\text{id}}(\mathbf{p}) = \mathbf{p}$. Also define the class of functions $\mathcal{Q}_{\text{one-hot-}\mathcal{H}}$ as $\mathcal{Q}_{\text{one-hot-}\mathcal{H}} = \big\{\mathbf{q} : \mathcal{X} \to \Delta_{\mathcal{Y}} \bigm| \exists h \in \mathcal{H} \text{ s.t. } \mathbf{q}(x) = \mathbf{e}_{h(x)}\ \forall x \in \mathcal{X}\big\}$. Then it can be seen that*

$$\begin{aligned}
\mathcal{D}_{(\boldsymbol{\tau}^{\text{id}}, \mathcal{Q}_{\text{one-hot-}\mathcal{H}})\text{-}RSM} &= \big\{D = (\mu, \mathbf{p}) \in \Delta_{\mathcal{X} \times \mathcal{Y}} \bigm| \exists h \in \mathcal{H} \text{ s.t. } \mathbf{p}(x) = \mathbf{e}_{h(x)}\ \forall x \in \mathcal{X}\big\} \\
&\equiv \mathcal{D}_{\mathcal{H}\text{-realizable}} ,
\end{aligned}$$

where $\mathcal{D}_{\mathcal{H}\text{-realizable}}$ denotes the class of probability distributions $D = (\mu, \mathbf{p}) \in \Delta_{\mathcal{X} \times \mathcal{Y}}$ wherein the label Y is (with probability 1) given by a deterministic function of the instance X, with the function belonging to $\mathcal{H}$. Therefore, realizable PAC learning w.r.t. $\mathcal{H}$ for any loss $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y} \times \mathcal{Y}}$ is equivalent to the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{id}}, \mathcal{Q}_{\text{one-hot-}\mathcal{H}})\text{-}RSM})$.

Example 2 (**Agnostic PAC learning as RSM).** *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ be a class of prediction models. Let $\mathcal{C} = \Delta_{\mathcal{Y}}$, and consider the identity property $\boldsymbol{\tau}^{\text{id}} : \Delta_{\mathcal{Y}} \to \Delta_{\mathcal{Y}}$ defined as $\boldsymbol{\tau}^{\text{id}}(\mathbf{p}) = \mathbf{p}$. Define $\mathcal{D}_{\text{all}} = \Delta_{\mathcal{X}\times\mathcal{Y}}$ and $\mathcal{Q}_{\text{all}} = \big\{\mathbf{q} : \mathcal{X} \to \Delta_{\mathcal{Y}}\big\}$. Then it can be seen that*

$$
\begin{aligned}
\mathcal{D}_{(\boldsymbol{\tau}^{\text{id}},\mathcal{Q}_{\text{all}})\text{-}RSM} &= \big\{D = (\mu,\mathbf{p}) \in \Delta_{\mathcal{X}\times\mathcal{Y}} \,\big|\, \exists \mathbf{q} \in \mathcal{Q}_{\text{all}} \text{ s.t. } \mathbf{p}(x) = \mathbf{q}(x)\ \forall x \in \mathcal{X}\big\} \\
&\equiv \mathcal{D}_{\text{all}}\,,
\end{aligned}
$$

and therefore (fully) agnostic PAC learning w.r.t. $\mathcal{H}$ and any loss $\mathbf{L} \in \mathbb{R}_+^{\mathcal{Y}\times\widehat{\mathcal{Y}}}$ is equivalent to the RSM learning problem $(\mathbf{L}, \mathcal{H}, \mathcal{D}_{(\boldsymbol{\tau}^{\text{id}},\mathcal{Q}_{\text{all}})\text{-}RSM})$.

B.2 Previously Studied Intermediate PAC Learning Models as Special Cases of RSMs

The RSM framework also recovers as special cases all the previously studied intermediate PAC learning models listed in Table 1.

Noisy LTF with RCN: Consider the statistic τ^{+1} defined in Section 3 and the class of statistic functions $Q_{\text{RCN-linear}} := \{q : \mathcal{X} \to [0,1] \mid \exists \mathbf{w} \in \mathbb{R}^p, \eta \in (0,1/2) \text{ s.t. } q(x) = (1-\eta)\cdot \mathbf{1}(\mathbf{w}^\top\mathbf{x} \geq 0) + \eta \cdot \mathbf{1}(\mathbf{w}^\top\mathbf{x} < 0)\}$. Then the RSM learning problem $(\mathbf{L}^{\text{0-1}}, \mathcal{H}_{\text{linear}}, \mathcal{D}_{(\tau^{+1},Q_{\text{RCN-linear}})\text{-RSM}})$ captures exactly the problem of learning linear threshold functions with RCN.

Noisy LTF with Massart noise: Consider again the statistic τ^{+1} defined in Section 3, and now the class of statistic functions $Q_{\text{Massart-linear}} := \{q : \mathcal{X} \to [0,1] \mid \exists \mathbf{w} \in \mathbb{R}^p, \eta \in (0,1/2) \text{ s.t. } q(x) \geq (1-\eta) \text{ if } \mathbf{w}^\top\mathbf{x} \geq 0 \text{ and } q(x) \leq \eta \text{ if } \mathbf{w}^\top\mathbf{x} < 0\}$. Then the RSM learning problem $(\mathbf{L}^{\text{0-1}}, H_{\text{linear}}, \mathcal{D}_{(\tau^{+1},Q_{\text{Massart-linear}})\text{-RSM}})$ captures exactly the problem of learning linear threshold functions with Massart noise.

GLM: Let $\theta : \mathbb{R} \to [0,1]$ be a fixed (known) transfer function (it is common to assume that θ is monotonically increasing and Lipschitz continuous). Consider again the statistic τ^{+1} defined in Section 3, and now the class of statistic functions $Q^{\theta}_{\text{GLM}} := \{q : \mathcal{X} \to [0,1] \mid \exists \mathbf{w} \in \mathbb{R}^p \text{ s.t. } q(x) = \theta(\mathbf{w}^\top\mathbf{x})\}$. Then the RSM learning problem $(\mathbf{L}^{\text{0-1}}, \mathcal{H}_{\text{linear}}, \mathcal{D}_{(\tau^{+1},Q^{\theta}_{\text{GLM}})\text{-RSM}})$ captures exactly the problem of learning GLMs with transfer function θ.

SIM: Let $\mathcal{T} \subseteq \{\theta : \mathbb{R} \to [0,1]\}$ be a class of transfer functions of interest (it is common to let $\mathcal{T}$ be a class of transfer functions that are monotonically increasing and Lipschitz continuous). Consider again the statistic τ^{+1} defined in Section 3, and now the class of statistic functions $Q^{\mathcal{T}}_{\text{SIM}} := \{q : \mathcal{X} \to [0,1] \mid \exists \mathbf{w} \in \mathbb{R}^p, \theta \in \mathcal{T} \text{ s.t. } q(x) = \theta(\mathbf{w}^\top\mathbf{x})\}$. Then the RSM learning problem $(\mathbf{L}^{\text{0-1}}, \mathcal{H}_{\text{linear}}, \mathcal{D}_{(\tau^{+1},Q^{\mathcal{T}}_{\text{SIM}})\text{-RSM}})$ captures exactly the problem of learning SIMs with class of transfer functions $\mathcal{T}$.

B.3 Proof of Theorem 1

Recall that we denote

$$
\begin{aligned}
\text{er}^{\mathbf{L}}_D[h] &= \mathbf{E}_{(X,Y)\sim D}[L_{Y,h(X)}]\,; \\
\text{er}^{\mathbf{L}}_D[\mathcal{H}] &= \inf_{h\in\mathcal{H}} \text{er}^{\mathbf{L}}_D[h]\,; \\
\text{er}^{\mathbf{L},*}_D &= \inf_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} \text{er}^{\mathbf{L}}_D[h]\,; \\
\text{er}^{\psi}_D[\mathbf{f}] &= \mathbf{E}_{(X,Y)\sim D}[\psi(Y,\mathbf{f}(X))]\,; \\
\text{er}^{\psi}_D[\mathcal{F}] &= \inf_{\mathbf{f}\in\mathcal{F}} \text{er}^{\psi}_D[\mathbf{f}]\,; \\
\text{er}^{\psi,*}_D &= \inf_{\mathbf{f}:\mathcal{X}\to\mathbb{R}^d} \text{er}^{\psi}_D[\mathbf{f}]\,.
\end{aligned}
$$

Proof. **(of Theorem 1)** Let $\mathbf{f} \in \mathcal{F}$ and $D = (\mu,\mathbf{p}) \in \mathcal{D}_{(\boldsymbol{\tau},\mathcal{Q})\text{-RSM}}$. We start by setting up some notation. Define $\mathbf{q}^* \in \mathcal{Q}$ as $\mathbf{q}^*(x) = \boldsymbol{\tau}(\mathbf{p}(x))$; define $\mathbf{f}^* \in \mathcal{F}$ as $\mathbf{f}^*(x) = \boldsymbol{\lambda}(\mathbf{q}^*(x)) = \boldsymbol{\lambda}(\boldsymbol{\tau}(\mathbf{p}(x)))$; and define $\mathbf{q} \in \mathcal{Q}$ as $\mathbf{q}(x) = \boldsymbol{\lambda}^{-1}(\mathbf{f}(x))$.

Now, we have $\mathrm{er}_D^{\mathbf{L}}[\mathcal{H}] = \mathrm{er}_D^{\mathbf{L},*}$ – to see this, note that since $(\boldsymbol{\tau}, \mathrm{pred})$ is an $\mathbf{L}$-calibrated statistic-mapping pair, the Bayes optimal classifier $h_D^{\mathbf{L},*}$ satisfies

$$h_D^{\mathbf{L},*}(x) \;=\; \mathrm{pred}(\boldsymbol{\tau}(\mathbf{p}(x))) \;=\; \mathrm{pred}(\mathbf{q}^*(x))\,,$$

and so $h_D^{\mathbf{L},*} \in \mathrm{pred} \circ \mathcal{Q} = \mathcal{H}$, which gives $\mathrm{er}_D^{\mathbf{L},*} = \mathrm{er}_D^{\mathbf{L}}[\mathcal{H}]$.

Also note that

$$h(x) \;=\; \mathrm{decode}(\mathbf{f}(x)) \;=\; \mathrm{pred}(\boldsymbol{\lambda}^{-1}(\boldsymbol{\lambda}(\mathbf{q}(x)))) \;=\; \mathrm{pred}(\mathbf{q}(x))\,.$$

Thus, we have,

$$\begin{aligned}
\mathrm{er}_D^{\mathbf{L}}[h] - \mathrm{er}_D^{\mathbf{L}}[\mathcal{H}] \;&=\; \mathrm{er}_D^{\mathbf{L}}[h] - \mathrm{er}_D^{\mathbf{L},*} \quad (\text{since } \mathrm{er}_D^{\mathbf{L}}[\mathcal{H}] = \mathrm{er}_D^{\mathbf{L},*})\\
&=\; \mathbf{E}_X\Big[\mathbf{E}_{Y\sim\mathbf{p}(X)}[L_{Y,h(X)}] - \min_{\widehat{y}\in\mathcal{Y}} \mathbf{E}_{Y\sim\mathbf{p}(X)}[L_{Y,\widehat{y}}]\Big]\\
&=\; \mathbf{E}_X\Big[\mathbf{E}_{Y\sim\mathbf{p}(X)}[L_{Y,\mathrm{pred}(\mathbf{q}(X))}] - \min_{\widehat{y}\in\mathcal{Y}} \mathbf{E}_{Y\sim\mathbf{p}(X)}[L_{Y,\widehat{y}}]\Big]\\
&\le\; \kappa \cdot \mathbf{E}_X[\|\mathbf{q}(X) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2] \quad (\text{by given condition})\\
&=\; \kappa \cdot \mathbf{E}_X[\|\boldsymbol{\lambda}^{-1}(\mathbf{f}(X)) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2]\\
&\le\; \kappa \cdot \sqrt{\mathbf{E}_X[\|\boldsymbol{\lambda}^{-1}(\mathbf{f}(X)) - \boldsymbol{\tau}(\mathbf{p}(X))\|_2^2]}\\
&\qquad (\text{by Jensen's inequality applied to the function } \phi(u) = u^2)\\
&\le\; \kappa \cdot \sqrt{\frac{2}{\gamma}\mathbf{E}_X\Big[\Big(\mathbf{E}_{Y\sim\mathbf{p}(X)}[\psi(Y,\mathbf{f}(X))] - \mathbf{E}_{Y\sim\mathbf{p}(X)}[\psi(Y,\boldsymbol{\lambda}(\boldsymbol{\tau}(\mathbf{p}(X))))]\Big)\Big]}\\
&\qquad (\text{by } \gamma\text{-strong proper compositeness of } \psi)\\
&=\; \kappa \cdot \sqrt{\frac{2}{\gamma}\mathbf{E}_X\Big[\Big(\mathbf{E}_{Y\sim\mathbf{p}(X)}[\psi(Y,\mathbf{f}(X))] - \mathbf{E}_{Y\sim\mathbf{p}(X)}[\psi(Y,\mathbf{f}^*(X))]\Big)\Big]}\\
&=\; \kappa \cdot \sqrt{\frac{2}{\gamma}\Big(\mathrm{er}_D^{\psi}[\mathbf{f}] - \mathrm{er}_D^{\psi}[\mathbf{f}^*]\Big)}\\
&=\; \kappa \cdot \sqrt{\frac{2}{\gamma}\Big(\mathrm{er}_D^{\psi}[\mathbf{f}] - \mathrm{er}_D^{\psi}[\mathcal{F}]\Big)}\\
&\qquad (\text{since strong proper compositeness of } \psi \text{ and the definition of } \mathbf{f}^*\\
&\qquad\ \ \text{also imply – for } D \text{ of the given form – that}\\
&\qquad\ \ \mathbf{E}_{(X,Y)\sim D}[\psi(Y,\mathbf{f}^*(X))] \le \mathbf{E}_{(X,Y)\sim D}[\psi(Y,\widetilde{\mathbf{f}}(X))]\ \forall\widetilde{\mathbf{f}}:\mathcal{X}\to\mathbb{R}^d,\\
&\qquad\ \ \text{and therefore } \mathrm{er}_D^{\psi}[\mathbf{f}^*] = \mathrm{er}_D^{\psi,*} = \mathrm{er}_D^{\psi}[\mathcal{F}])\,.
\end{aligned}$$

□

B.4 Proof of Theorem 2

Let us start by stating the following uniform convergence result, which relates the empirical and expected surrogate risks for a bounded surrogate loss ψ acting on vector-valued predictions, uniformly for all functions in a vector-valued function class $\mathcal{F}$, in terms of the d_1 covering numbers of the loss class $\psi_{\mathcal{F}}$. The result follows from a straightforward generalization of standard uniform convergence results for real-valued function classes (such as given in Chapter 17 of [5]) to vector-valued function classes.

Theorem 9 (**Uniform convergence for bounded (surrogate) loss classes in terms of d_1 covering numbers**). *Let $\mathcal{X}$ be any instance space and $\mathcal{Y}$ be any label space. Let $d \in \mathbb{Z}_+$ and let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be a (surrogate) loss function. Let $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d\}$ and suppose $\psi(y, \mathbf{f}(x)) \in [0, B]\ \forall x \in \mathcal{X}, y \in \mathcal{Y}, \mathbf{f} \in \mathcal{F}$ for some $B > 0$. Then for any $m \in \mathbb{Z}_+$, any $\epsilon > 0$, and any $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,*

$$\mathbf{P}_{S\sim D^m}\left(\sup_{\mathbf{f}\in\mathcal{F}}\big|\mathrm{er}_D^{\psi}[\mathbf{f}] - \widehat{\mathrm{er}}_S^{\psi}[\mathbf{f}]\big| \ge \epsilon\right) \;\le\; 4\mathcal{N}_1(\epsilon/8, \psi_{\mathcal{F}}, 2m)e^{-m\epsilon^2/(32B^2)}\,.$$

We will now prove the following technical lemma, which upper bounds the d_1 covering numbers of the surrogate loss class $\psi_{\mathcal{F}}$ – for surrogate losses ψ that act on vector-valued predictions and that are Lipschitz with respect to the L^1 metric – in terms of the d_1 covering numbers of the real-valued 'projection' function classes $\mathcal{F}^j$. This lemma may also be of independent interest.

Lemma 1 (**Bounding d_1 covering numbers of loss function classes $\psi_{\mathcal{F}}$ for Lipschitz losses ψ acting on vector-valued predictions).** *Let $\mathcal{X}$ and $\mathcal{Y}$ be any sets. Let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be any (surrogate) loss function that is ρ_1-Lipschitz in the second argument with respect to the L^1 metric, and $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d\}$ be any class of vector-valued functions on $\mathcal{X}$. Let*

$$\psi_{\mathcal{F}} \;:=\; \{\psi_{\mathbf{f}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } \psi_{\mathbf{f}}(x,y) = \psi(y, \mathbf{f}(x))\ \forall x \in \mathcal{X}, y \in \mathcal{Y}\}\,.$$

For each $j \in [d]$, let $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j\ \forall x\}$. Then for any $\epsilon > 0$ and $m \in \mathbb{Z}_+$,

$$\mathcal{N}_1(\epsilon, \psi_{\mathcal{F}}, m) \;\leq\; \prod_{j=1}^{d} \mathcal{N}_1(\epsilon/(\rho_1 d), \mathcal{F}^j, m)\,.$$

Proof. **(of Lemma 1)** Let $\epsilon > 0$ and $m \in \mathbb{Z}_+$. Fix any $\mathbf{z} = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, and denote $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}^m$. For each $j \in [d]$, let $C_j \subset \mathbb{R}^m$ be an (ϵ/ρ_1)-cover for $(\mathcal{F}^j)_{|\mathbf{x}}$ with respect to the d_1 distance. We will construct an ϵ-cover $C \subset \mathbb{R}^m$ for $(\psi_{\mathcal{F}})_{|\mathbf{z}}$ with respect to the d_1 distance of size $|C| \leq \prod_{j=1}^d |C_j|$.

Let $\mathbf{f} \in \mathcal{F}$, and denote $(\psi_{\mathbf{f}})_{|\mathbf{z}} = (\psi_{\mathbf{f}}(z_1), \ldots, \psi_{\mathbf{f}}(z_m)) \in \mathbb{R}^m$; moreover, for each $j \in [d]$, let $f_j : \mathcal{X} \to \mathbb{R}$ be defined as $f_j(x) = (\mathbf{f}(x))_j$, and denote $(f_j)_{|\mathbf{x}} = (f_j(x_1), \ldots, f_j(x_m))$. For each $j \in [d]$, let $\mathbf{u}^j = (u_1^j, \ldots, u_m^j) \in C_j$ be such that $d_1((f_j)_{|\mathbf{x}}, \mathbf{u}^j) \leq \epsilon/(\rho_1 d)$. For each $i \in [m]$, define the d-dimensional vector $\mathbf{u}_i = (u_i^1, \ldots, u_i^d) \in \mathbb{R}^d$. Now consider the m-dimensional point $\mathbf{v} := \psi_{|((y_i, \mathbf{u}_i))_{i=1}^m} = (\psi(y_1, \mathbf{u}_1), \ldots, \psi(y_m, \mathbf{u}_m)) \in \mathbb{R}^m$. Then we have

$$\begin{aligned}
d_1((\psi_{\mathbf{f}})_{|\mathbf{z}}, \mathbf{v}) &= \frac{1}{m} \sum_{i=1}^m |\psi_{\mathbf{f}}(z_i) - v_i| \\
&= \frac{1}{m} \sum_{i=1}^m |\psi(y_i, \mathbf{f}(x_i)) - \psi(y_i, \mathbf{u}_i)| \\
&\leq \frac{1}{m} \sum_{i=1}^m \left(\rho_1 \cdot \sum_{j=1}^d |f_j(x_i) - u_i^j| \right) \quad \text{(by } \rho_1\text{-Lipschitzness of } \psi \text{ w.r.t. } L^1) \\
&= \rho_1 \cdot \sum_{j=1}^d d_1((f_j)_{|\mathbf{x}}, \mathbf{u}^j) \\
&\leq \rho_1 \cdot \sum_{j=1}^d \left(\frac{\epsilon}{\rho_1 d} \right) \\
&= \epsilon\,.
\end{aligned}$$

Therefore the set

$$C \;=\; \left\{ \mathbf{v} := \psi_{|((y_i, \mathbf{u}_i))_{i=1}^m} \;\middle|\; \mathbf{u}^j \in C_j\ \forall j \right\} \subset \mathbb{R}^m$$

is an ϵ-cover for $(\psi_{\mathcal{F}})_{|\mathbf{z}}$ with respect to the d_1 distance. Since $|C| \leq \prod_{j=1}^d |C_j|$, the claim follows. □

Next, the following result shows that uniform convergence of surrogate risks also implies (surrogate) learning results for approximate empirical risk minimizers. The proof technique is standard (such as given in Chapter 19 of [5]); we include a self-contained proof here for completeness.

Theorem 10 (**Uniform convergence implies bounded (surrogate) regret of approximate (surrogate) risk minimizers).** *Let $\mathcal{X}$ be any instance space and $\mathcal{Y}$ be any label space. Let $d \in Z_+$ and let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be a (surrogate) loss function. Let $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d\}$. Let $m_{\text{uc}} : \mathbb{R}_+ \times (0, 1] \to \mathbb{Z}_+$*

be such that for every $\epsilon > 0$, every $\delta \in (0,1]$, every $m \geq m_{\text{uc}}(\epsilon, \delta)$, and every $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{\mathbf{f}\in\mathcal{F}}\left|\text{er}_D^\psi[\mathbf{f}] - \widehat{\text{er}}_S^\psi[\mathbf{f}]\right| \geq \epsilon\right) \quad \leq \quad \delta\,.$$

Let $(\alpha_m)_{m\in\mathbb{Z}_+}$ be a sequence of positive real numbers such that for every $\epsilon > 0$, every $\delta \in (0,1]$, and every $m \geq m_{\text{uc}}(\epsilon/3, \delta)$, we have $\alpha_m \leq \epsilon/3$. Let $\mathcal{A}$ be an approximate surrogate risk minimization algorithm which, given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ of size m, returns an α_m-approximate minimizer $\widehat{\mathbf{f}}_S \in \mathcal{F}$ of the empirical ψ-risk $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over $\mathcal{F}$, so that $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \widehat{\mathbf{f}}_S(x_i)) \leq \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i)) + \alpha_m$. Then for every $\epsilon > 0$, every $\delta \in (0,1]$, every $m \geq m_{\text{uc}}(\epsilon/3, \delta)$, and every $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\text{er}_D^\psi[\widehat{\mathbf{f}}_S] - \inf_{\mathbf{f}\in\mathcal{F}}\text{er}_D^\psi[\mathbf{f}] \geq \epsilon\right) \quad \leq \quad \delta\,.$$

Proof. **(of Theorem 10)** Let $\epsilon > 0$, $\delta \in (0,1]$, and $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$. Let $\beta > 0$, and let $\mathbf{f}^* \in \mathcal{F}$ be such that

$$\text{er}_D^\psi[\mathbf{f}^*] \leq \inf_{\mathbf{f}\in\mathcal{F}}\text{er}_D^\psi[\mathbf{f}] + \beta\,.$$

Let $m \geq m_{\text{uc}}(\epsilon/3, \delta)$. Then we have the following with probability at least $1 - \delta$ over the draw of $S \sim D^m$:

$$\sup_{\mathbf{f}\in\mathcal{F}}|\text{er}_D^\psi[\mathbf{f}] - \widehat{\text{er}}_S^\psi[\mathbf{f}]| \leq \epsilon\,,$$

and therefore,

$$\begin{aligned}
\text{er}_D^\psi[\widehat{\mathbf{f}}_S] &\leq \widehat{\text{er}}_S^\psi[\widehat{\mathbf{f}}_S] + \frac{\epsilon}{3}\\
&\leq \left(\inf_{\mathbf{f}\in\mathcal{F}}\widehat{\text{er}}_S^\psi[\mathbf{f}] + \alpha_m\right) + \frac{\epsilon}{3}\\
&\leq \inf_{\mathbf{f}\in\mathcal{F}}\widehat{\text{er}}_S^\psi[\mathbf{f}] + \frac{2\epsilon}{3}\\
&\leq \widehat{\text{er}}_S^\psi[\mathbf{f}^*] + \frac{2\epsilon}{3}\\
&\leq \left(\text{er}_D^\psi[\mathbf{f}^*] + \frac{\epsilon}{3}\right) + \frac{2\epsilon}{3}\\
&\leq \inf_{\mathbf{f}\in\mathcal{F}}\text{er}_D^\psi[\mathbf{f}] + \beta + \epsilon\,.
\end{aligned}$$

Since the above holds for all $\beta > 0$, we have that with probability at least $1 - \delta$ over $S \sim D^m$,

$$\text{er}_D^\psi[\widehat{\mathbf{f}}_S] \leq \inf_{\mathbf{f}\in\mathcal{F}}\text{er}_D^\psi[\mathcal{F}] + \epsilon\,.$$

This proves the claim. □

Next, we define the surrogate sample complexity below:

Definition 4 (Surrogate sample complexity). *Let $\mathcal{C}' \subseteq \mathbb{R}^{d'}$. Let $\psi : \mathcal{Y} \times \mathcal{C}' \to \mathbb{R}_+$ be any surrogate loss, $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathcal{C}'\}$ be a class of surrogate prediction models, and $\mathcal{D} \subseteq \Delta_{\mathcal{X}\times\mathcal{Y}}$ be a class of probability distributions. We will say an algorithm $\mathcal{A}$ that given a training sample $S \in \cup_{m=1}^\infty (\mathcal{X} \times \mathcal{Y})^m$ returns a surrogate prediction model $\widehat{\mathbf{f}}_S \in \mathcal{F}$ is a learning algorithm for the surrogate loss learning problem $(\psi, \mathcal{F}, \mathcal{D})$ with surrogate sample complexity function $m_{\mathcal{A}}^\psi : \mathbb{R}_+ \times (0,1] \to \mathbb{Z}_+$ if for every $\epsilon > 0, \delta \in (0,1]$, every distribution $D \in \mathcal{D}$, and every $m \geq m_{\mathcal{A}}^\psi(\epsilon, \delta)$,*

$$\mathbf{P}_{S\sim D^m}\left(\text{er}_D^\psi[\widehat{\mathbf{f}}_S] - \inf_{\mathbf{f}\in\mathcal{F}}\text{er}_D^\psi[\mathbf{f}] \geq \epsilon\right) \;\leq\; \delta\,,$$

and moreover, for every ϵ, δ, $m_{\mathcal{A}}^\psi(\epsilon, \delta)$ is the smallest integer satisfying the above.

Bringing all the above together, under the conditions of Theorem 2, the following result upper bounds the surrogate sample complexity of an approximate surrogate risk minimization algorithm in terms of the d_1 covering numbers of the real-valued projection classes $\mathcal{F}^j$.

Theorem 11 (Upper bounding surrogate sample complexity of an approximate surrogate risk minimizer via d_1 covering numbers). *Under the conditions of Theorem 2, the $(16B/\sqrt{m})$-approximate surrogate risk minimization algorithm $\mathcal{A}$ is a learning algorithm for the surrogate learning problem $(\psi, \mathcal{F}, \Delta_{\mathcal{X}\times\mathcal{Y}})$ with surrogate sample complexity upper bounded as*

$$m^{\psi}_{\mathcal{A}}(\epsilon,\delta) \le \min\Big\{m_0 \in \mathbb{Z}_+ : m \ge m_0 \implies m \ge \tfrac{288B^2}{\epsilon^2}\Big(\textstyle\sum_{j=1}^d \ln\Big(\mathcal{N}_1\Big(\tfrac{\epsilon}{24\rho_1 d}, \mathcal{F}^j, 2m\Big)\Big) + \ln\big(\tfrac{4}{\delta}\big)\Big)\Big\} .$$

In particular, if $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m) \le \phi(\epsilon, \mathcal{F}^j)\ \forall j \in [d]$, then we have

$$m^{\psi}_{\mathcal{A}}(\epsilon,\delta) \le \tfrac{288B^2}{\epsilon^2}\Big(\textstyle\sum_{j=1}^d \ln\Big(\phi\Big(\tfrac{\epsilon}{24\rho_1 d}, \mathcal{F}^j\Big)\Big) + \ln\big(\tfrac{4}{\delta}\big)\Big) .$$

Proof. **(of Theorem 11)** Define $m_{\text{uc}} : \mathbb{R}_+ \times (0,1] \to \mathbb{Z}_+$ as

$$m_{\text{uc}}(\epsilon,\delta) := \min\Big\{m_0 \in \mathbb{Z}_+ : m \ge m_0 \implies m \ge \tfrac{32B^2}{\epsilon^2}\Big(\textstyle\sum_{j=1}^d \ln\Big(\mathcal{N}_1\Big(\tfrac{\epsilon}{8\rho_1 d}, \mathcal{F}^j, 2m\Big)\Big) + \ln\big(\tfrac{4}{\delta}\big)\Big)\Big\} .$$

Then by Theorem 9 and Lemma 1, we have that for every $\epsilon > 0$, $\delta \in (0,1]$, $m \ge m_{\text{uc}}(\epsilon,\delta)$, and $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{\mathbf{f}\in\mathcal{F}} \big|\text{er}^{\psi}_D[\mathbf{f}] - \widehat{\text{er}}^{\psi}_S[\mathbf{f}]\big| \ge \epsilon\right) \le \delta .$$

Next, define a sequence of positive real numbers $(\alpha_m)_{m\in\mathbb{Z}_+}$ as

$$\alpha_m := \frac{16B}{\sqrt{m}}.$$

Then it can be verified that for every $\epsilon > 0$, $\delta \in (0,1]$, and $m \ge m_{\text{uc}}(\epsilon/3,\delta)$, we have $\alpha_m \le \epsilon/3$. Therefore, by Theorem 10, an α_m-approximate surrogate risk minimization algorithm as described satisfies for every $\epsilon > 0$, $\delta \in (0,1]$, $m \ge m_{\text{uc}}(\epsilon/3,\delta)$, and $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\text{er}^{\psi}_D[\widehat{\mathbf{f}}_S] - \inf_{\mathbf{f}\in\mathcal{F}} \text{er}^{\psi}_D[\mathbf{f}] \ge \epsilon\right) \le \delta .$$

Thus we have

$$\begin{aligned} m^{\psi}_{\mathcal{A}}(\epsilon,\delta) &\le m_{\text{uc}}(\epsilon/3,\delta) \\ &\le \min\Big\{m_0 \in \mathbb{Z}_+ : m \ge m_0 \implies m \ge \tfrac{288B^2}{\epsilon^2}\Big(\textstyle\sum_{j=1}^d \ln\Big(\mathcal{N}_1\Big(\tfrac{\epsilon}{24\rho_1 d}, \mathcal{F}^j, 2m\Big)\Big) + \ln\big(\tfrac{4}{\delta}\big)\Big)\Big\} . \end{aligned}$$

Moreover, if $\mathcal{N}_1(\epsilon, \mathcal{F}^j, m) \le \phi(\epsilon, \mathcal{F}^j)\ \forall j \in [d]$, this yields the stated bound. □

Finally, we will also make use of the following proposition, whose proof follows directly from Theorem 1.

Proposition 12 (Upper bounding squared $\boldsymbol{\tau}$-estimation error sample complexity and target loss sample complexity in terms of surrogate sample complexity). *Under the conditions of Theorem 1, any learning algorithm $\mathcal{A}$ which given a training sample S, finds a surrogate prediction model $\widehat{\mathbf{f}}_S \in \mathcal{F}$ and produces a $\boldsymbol{\tau}$-statistic estimate $\widehat{\mathbf{q}}_S(x) = \boldsymbol{\lambda}^{-1}(\widehat{\mathbf{f}}_S(x))$ and a prediction model $\widehat{h}_S(x) = \text{decode}(\widehat{\mathbf{f}}_S(x))$, satisfies*

$$\begin{aligned} m^{\boldsymbol{\tau}}_{\mathcal{A}}(\epsilon,\delta) &\le m^{\psi}_{\mathcal{A}}\left(\frac{\gamma\epsilon}{2},\delta\right); \\ m^{\mathbf{L}}_{\mathcal{A}}(\epsilon,\delta) &\le m^{\psi}_{\mathcal{A}}\left(\frac{\gamma\epsilon^2}{2\kappa^2},\delta\right). \end{aligned}$$

Proof. **(of Proposition 12)** Follows directly from Theorem 1. □

The proof of Theorem 2 is now immediate:

Proof. **(of Theorem 2)** Follows directly from Theorem 11 and Proposition 12. □

B.5 Proof of Theorem 3

Let us start by stating the following uniform convergence result, which relates the empirical and expected surrogate risks for a bounded surrogate loss ψ acting on vector-valued predictions, uniformly for all functions in a vector-valued function class $\mathcal{F}$, in terms of the Rademacher complexity of the loss class $\psi_{\mathcal{F}}$. The proof is standard (via an application of McDiarmid's inequality; see e.g., [9]).

Theorem 13 (Uniform convergence for bounded (surrogate) loss classes in terms of Rademacher complexity). *Let $\mathcal{X}$ be any instance space and $\mathcal{Y}$ be any label space. Let $d \in \mathbb{Z}_+$ and let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be a (surrogate) loss function. Let $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d\}$ and suppose $\psi(y, \mathbf{f}(x)) \in [0, B]$ $\forall x \in \mathcal{X}, y \in \mathcal{Y}, \mathbf{f} \in \mathcal{F}$ for some $B > 0$. Then for any $m \in \mathbb{Z}_+$, any $\delta \in (0, 1]$, and any $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$, we have with probability at least $1 - \delta$ over the draw of $S \sim D^m$:*

$$\sup_{\mathbf{f}\in\mathcal{F}} \left|\mathrm{er}_D^{\psi}[\mathbf{f}] - \widehat{\mathrm{er}}_S^{\psi}[\mathbf{f}]\right| \quad \leq \quad 2\,\mathcal{R}_m(\psi_{\mathcal{F}}) + B\sqrt{\frac{\ln(2/\delta)}{m}}\,.$$

We will make use of the vector-contraction inequality for Rademacher complexities, due to Maurer [32], which upper bounds the Rademacher complexity of the surrogate loss class $\psi_{\mathcal{F}}$ – for surrogate losses ψ that act on vector-valued predictions and that are Lipschitz with respect to the Euclidean metric – in terms of the Rademacher complexities of the real-valued 'projection' function classes $\mathcal{F}^j$.

Lemma 2 (Bounding Rademacher complexities of loss function classes $\psi_{\mathcal{F}}$ for Lipschitz losses ψ acting on vector-valued predictions [32]). *Let $\mathcal{X}$ and $\mathcal{Y}$ be any sets. Let $\psi : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ be any (surrogate) loss function that is ρ_2-Lipschitz in the second argument with respect to the Euclidean metric, and $\mathcal{F} \subseteq \{\mathbf{f} : \mathcal{X} \to \mathbb{R}^d\}$ be any class of vector-valued functions on $\mathcal{X}$. Let*

$$\psi_{\mathcal{F}} \;:=\; \{\psi_{\mathbf{f}} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } \psi_{\mathbf{f}}(x, y) = \psi(y, \mathbf{f}(x))\; \forall x \in \mathcal{X}, y \in \mathcal{Y}\}\,.$$

For each $j \in [d]$, let $\mathcal{F}^j = \{f_j : \mathcal{X} \to \mathbb{R} \mid \exists \mathbf{f} \in \mathcal{F} \text{ s.t. } f_j(x) = (\mathbf{f}(x))_j\; \forall x\}$. Then for any $m \in \mathbb{Z}_+$,

$$\mathcal{R}_m(\psi_{\mathcal{F}}) \;\leq\; \sqrt{2}\rho_2 \cdot \sum_{j=1}^{d} \mathcal{R}_m(\mathcal{F}^j)\,.$$

Bringing the above together, under the conditions of Theorem 3, the following result upper bounds the surrogate sample complexity of an approximate surrogate risk minimization algorithm in terms of the Rademacher complexities of the real-valued projection classes $\mathcal{F}^j$.

Theorem 14 (Upper bounding surrogate sample complexity of an approximate surrogate risk minimizer via Rademacher complexities). *Under the conditions of Theorem 3, the $(B/(2\sqrt{m}))$-approximate surrogate risk minimization algorithm $\mathcal{A}$ is a learning algorithm for the surrogate learning problem $(\psi, \mathcal{F}, \Delta_{\mathcal{X}\times\mathcal{Y}})$ with surrogate sample complexity upper bounded as*

$$m_{\mathcal{A}}^{\psi}(\epsilon, \delta) \quad \leq \quad \min\Big\{ m_0 \in \mathbb{Z}_+ : m \geq m_0 \implies 3\Big(2\sqrt{2}\rho_2 \cdot \textstyle\sum_{j=1}^{d} \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \leq \epsilon \Big\}$$

In particular, if $\exists C > 0$ such that the Rademacher complexities of the function classes $\mathcal{F}^j$ have upper bounds of the form $\mathcal{R}_m(\mathcal{F}^j) \leq C/\sqrt{m}\; \forall j \in [d]$, then we have

$$m_{\mathcal{A}}^{\psi}(\epsilon, \delta) \;\leq\; \frac{9}{\epsilon^2}\left(2\sqrt{2}\rho_2 C d + B\sqrt{\ln(2/\delta)}\right)^2,$$

Proof. **(of Theorem 14)** Define $m_{\mathrm{uc}} : \mathbb{R}_+ \times (0, 1] \to \mathbb{Z}_+$ as

$$m_{\mathrm{uc}}(\epsilon, \delta) \quad := \quad \min\Big\{ m_0 \in \mathbb{Z}_+ : m \geq m_0 \implies \Big(2\sqrt{2}\rho_2 \cdot \textstyle\sum_{j=1}^{d} \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \leq \epsilon \Big\}$$

Then by Theorem 13 and Lemma 2, we have that for every $\epsilon > 0$, $\delta \in (0,1]$, $m \geq m_{\text{uc}}(\epsilon, \delta)$, and $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\sup_{\mathbf{f}\in\mathcal{F}} \left|\text{er}_D^{\psi}[\mathbf{f}] - \widehat{\text{er}}_S^{\psi}[\mathbf{f}]\right| \geq \epsilon\right) \quad \leq \quad \delta\,.$$

Next, define a sequence of positive real numbers $(\alpha_m)_{m\in\mathbb{Z}_+}$ as

$$\alpha_m \ := \ \frac{B}{2\sqrt{m}}.$$

Then it can be verified that for every $\epsilon > 0$, $\delta \in (0,1]$, and $m \geq m_{\text{uc}}(\epsilon/3, \delta)$, we have $\alpha_m \leq \epsilon/3$. Therefore, by Theorem 10, an α_m-approximate surrogate risk minimization algorithm as described satisfies for every $\epsilon > 0$, $\delta \in (0,1]$, $m \geq m_{\text{uc}}(\epsilon/3, \delta)$, and $D \in \Delta_{\mathcal{X}\times\mathcal{Y}}$,

$$\mathbf{P}_{S\sim D^m}\left(\text{er}_D^{\psi}[\widehat{\mathbf{f}}_S] - \inf_{\mathbf{f}\in\mathcal{F}} \text{er}_D^{\psi}[\mathbf{f}] \geq \epsilon\right) \quad \leq \quad \delta\,.$$

Thus we have

$$\begin{aligned}
m_{\mathcal{A}}^{\psi}(\epsilon, \delta) \quad &\leq \quad m_{\text{uc}}(\epsilon/3, \delta) \\
&\leq \quad \min\Big\{m_0 \in \mathbb{Z}_+ : m \geq m_0 \implies \\
&\qquad\qquad\qquad 3\Big(2\sqrt{2}\rho_2 \cdot \textstyle\sum_{j=1}^d \mathcal{R}_m(\mathcal{F}^j) + B\sqrt{\frac{\ln(2/\delta)}{m}}\Big) \leq \epsilon\Big\}\,.
\end{aligned}$$

Moreover, if $\mathcal{R}_m(\mathcal{F}^j) \leq C/\sqrt{m}\ \forall j \in [d]$, this yields the stated bound. □

The proof of Theorem 3 is now immediate:

Proof. **(of Theorem 3)** Follows directly from Theorem 14 and Proposition 12. □

B.6 Proof of Proposition 4

Proof. **(of Proposition 4)**

(i) This is a well-known result (e.g., see [5]).

(ii) This is a well-known result (e.g., see [44]).

(iii) This is also a well-known result; we provide a self-contained proof here for completeness. The fact that $\mathcal{R}_m(\mathcal{F}_{\text{linear}}) \geq 0$ follows directly from the fact $\mathcal{F}_{\text{linear}}$ is closed under negation. For the upper bound, we have

$$
\begin{aligned}
\mathcal{R}_m(\mathcal{F}_{\text{linear}}) &= \mathbf{E}\left[\sup_{\mathbf{w},\|\mathbf{w}\|_2 \le W}\left(\frac{1}{m}\sum_{i=1}^{n}\epsilon_i \mathbf{w}^\top \mathbf{x}_i\right)\right] \\
&= \frac{1}{m}\mathbf{E}\left[\sup_{\mathbf{w},\|\mathbf{w}\|_2 \le W}\left(\mathbf{w}^\top \sum_{i=1}^{n}\epsilon_i \mathbf{x}_i\right)\right] \\
&\le \frac{1}{m}\mathbf{E}\left[\sup_{\mathbf{w},\|\mathbf{w}\|_2 \le W}\|\mathbf{w}\|_2 \left\|\sum_{i=1}^{n}\epsilon_i \mathbf{x}_i\right\|_2\right] \quad \text{(by Cauchy-Schwarz)} \\
&= \frac{1}{m}\mathbf{E}\left[W\left\|\sum_{i=1}^{n}\epsilon_i \mathbf{x}_i\right\|_2\right] \\
&= \frac{W}{m}\mathbf{E}\left[\sqrt{\left(\sum_{i=1}^{n}\epsilon_i \mathbf{x}_i\right)^\top\left(\sum_{i=1}^{n}\epsilon_i \mathbf{x}_i\right)}\right] \\
&= \frac{W}{m}\mathbf{E}\left[\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\left(\mathbf{x}_i^\top \mathbf{x}_j\right)}\right] \\
&\le \frac{W}{m}\sqrt{\mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\left(\mathbf{x}_i^\top \mathbf{x}_j\right)\right]} \quad \text{(by Jensen's inquality)} \\
&= \frac{W}{m}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left[\epsilon_i\epsilon_j\right]\left(\mathbf{x}_i^\top \mathbf{x}_j\right)} \\
&= \frac{W}{m}\sqrt{\sum_{i=1}^{n}\|\mathbf{x}_i\|_2^2} \\
&\le \frac{RW}{\sqrt{m}}
\end{aligned}
$$

□

C Supplement to Section 3 (Binary Classification)

Lemma 3. $(\tau^{+1}, \text{pred}^{0\text{-}1})$ *is an* $\mathbf{L}^{0\text{-}1}$*-calibrated statistic-mapping pair, with*

$$\mathbf{E}_{Y\sim\mathbf{p}}[L^{0\text{-}1}_{Y,\text{pred}^{0\text{-}1}(q)}] - \min_{\widehat{y}\in\{\pm 1\}}\mathbf{E}_{Y\sim\mathbf{p}}[L^{0\text{-}1}_{Y,\widehat{y}}] \;\le\; 2\,|q - p_{+1}| \quad \forall \mathbf{p}\in\Delta_{\{\pm 1\}}, q\in[0,1]\,.$$

Proof. **(of Lemma 3)** Calibration of $(\tau^{+1}, \text{pred}^{0\text{-}1})$ for $\mathbf{L}^{0\text{-}1}$ is immediate, since the Bayes optimal classifier for $\mathbf{L}^{0\text{-}1}$ is given by $h_D^{\mathbf{L}^{0\text{-}1},*}(\mathbf{x}) = \text{sign}(p_{+1}(\mathbf{x}) - \frac{1}{2}) = \text{pred}^{0\text{-}1}(\tau^{+1}(\mathbf{x}))$. Moreover, for any

$\mathbf{p} \in \Delta_{\{\pm 1\}}, q \in [0,1]$, we have

$$\begin{aligned}
&\mathbf{E}_{Y\sim\mathbf{p}}[L^{0\text{-}1}_{Y,\text{pred}^{0\text{-}1}(q)}] - \min_{\widehat{y}\in\{\pm 1\}} \mathbf{E}_{Y\sim\mathbf{p}}[L^{0\text{-}1}_{Y,\widehat{y}}] \\
&= \mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{1}(\text{pred}^{0\text{-}1}(q) \neq Y)] - \min_{\widehat{y}\in\{\pm 1\}} \mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{1}(\widehat{y} \neq Y)] \\
&= p_{+1} \cdot \mathbf{1}(\text{pred}^{0\text{-}1}(q) \neq +1) + (1-p_{+1}) \cdot \mathbf{1}(\text{pred}^{0\text{-}1}(q) \neq -1) - \min(p_{+1}, 1-p_{+1}) \\
&= p_{+1} \cdot \mathbf{1}(q < \frac{1}{2}) + (1-p_{+1}) \cdot \mathbf{1}(q \geq \frac{1}{2}) - \min(p_{+1}, 1-p_{+1}) \\
&= (2p_{+1}-1) \cdot \mathbf{1}(q < \frac{1}{2}, p_{+1} \geq \frac{1}{2}) + (1-2p_{+1}) \cdot \mathbf{1}(q \geq \frac{1}{2}, p_{+1} < \frac{1}{2}) \\
&= 2\,|p_{+1} - \frac{1}{2}| \cdot \left(\mathbf{1}(q < \frac{1}{2}, p_{+1} \geq \frac{1}{2}) + \mathbf{1}(q \geq \frac{1}{2}, p_{+1} < \frac{1}{2})\right) \\
&\leq 2\,|q - p_{+1}|\,.
\end{aligned}$$

□

Proof. **(of Theorem 5)** Consider the (invertible) logit link function $\lambda : [0,1] \to \overline{\mathbb{R}}$ and its inverse $\lambda^{-1} : \overline{\mathbb{R}} \to [0,1]$ given by[7]

$$\begin{aligned}
\lambda(p) &= \ln\left(\frac{p}{1-p}\right), \\
\lambda^{-1}(u) &= \frac{1}{1+e^{-u}}\,.
\end{aligned}$$

Note that λ^{-1} here is equivalent to the sigmoid function σ (defined in the theorem statement). We will observe/prove the following:

(1) $\mathcal{H}_{\text{linear}} = \text{pred}^{0\text{-}1} \circ \mathcal{Q}_{\text{sigmoid-of-linear}}$;

(2) $\psi^{\log}$ is a 4-strongly proper composite surrogate loss for τ^{+1} with link function λ;

(3) $\text{sign} = \text{pred}^{0\text{-}1} \circ \lambda^{-1}$;

(4) $\mathcal{F}_{\text{linear}} = \lambda \circ \mathcal{Q}_{\text{sigmoid-of-linear}}$;

(5) $\psi^{\log}(y, f(\mathbf{x})) \leq \ln(1 + e^{RW}) \quad \forall(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, f \in \mathcal{F}_{\text{linear}}$;

(6) $\psi^{\log}$ is 1-Lipschitz with respect to the L^1 metric (equivalently Euclidean metric) on $\mathbb{R}$;

(7) $\mathcal{N}_1(\epsilon, \mathcal{F}_{\text{linear}}, m) \leq \left(\frac{1}{\epsilon}\right)^p$;

(8) $0 \leq \mathcal{R}_m(\mathcal{F}_{\text{linear}}) \leq RW/\sqrt{m}$.

The result will then follow from Lemma 3 and Theorem 3.

Parts (1), (3), (4), (5) are immediate from the definitions.

Part (7) is a well-known result (e.g. see [5]).

Part (8) is a well-known result (see Proposition 4).

Part (2): $\psi^{\log}$ is known to be a 4-strongly proper composite loss for the property τ^{+1} (i.e., for binary class probability estimation) with link function λ as above [3].

Part (6): It is well known (and easy to verify) that the binary logistic loss $\psi^{\log}$ is 1-Lipschitz with respect to the L^1 (equivalently Euclidean) metric on $\mathbb{R}$ (to verify this, note that the absolute value of the derivative of $\psi^{\log}$ with respect to the second argument is upper bounded by 1).

[7] Note that in the notation of Definition 3, we use $\mathcal{C}' = \overline{\mathbb{R}}$ here. Technically, we would also need to extend the definitions of the surrogate loss $\psi^{\log}$ (and the mapping decode = sign) to act on $\overline{\mathbb{R}}$ instead of $\mathbb{R}$: we ignore this issue here for simplicity.

Combining all the above together with Lemma 3 and applying Theorem 3 (with $\kappa = 2$, $\gamma = 4$, $\rho_2 = 1$, $d = 1$, $0 \le \mathcal{R}_m(\mathcal{F}_{\text{linear}}) \le RW/\sqrt{m}$, and $B \le \ln(1 + e^{RW})$) gives the desired result with squared τ^{+1} estimation error sample complexity

$$\begin{aligned} m_{\mathcal{A}}^{\tau^{+1}}(\epsilon, \delta) &\le \frac{9}{4\epsilon^2}\Big(2\sqrt{2}RW + (\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\Big)^2 \\ &= O\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right) \end{aligned}$$

and with target loss sample complexity

$$\begin{aligned} m_{\mathcal{A}}^{\mathbf{L}^{\text{0-1}}}(\epsilon, \delta) &\le \frac{36}{\epsilon^4}\Big(2\sqrt{2}RW + (\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\Big)^2 \\ &= O\left(\frac{1}{\epsilon^4}\ln\left(\frac{1}{\delta}\right)\right). \end{aligned}$$

□

D Supplement to Section 4 (Multiclass Classification)

Lemma 4. $(\boldsymbol{\tau}^{\text{id}}, \text{pred}^{\text{0-1}(n)})$ *is an* $\mathbf{L}^{\text{0-1}(n)}$*-calibrated statistic-mapping pair, with*

$$\mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\text{pred}^{\text{0-1}(n)}(\mathbf{q})}] - \min_{\widehat{y}\in[n]} \mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\widehat{y}}] \;\le\; \sqrt{2}\cdot\|\mathbf{q}-\mathbf{p}\|_2 \quad \forall \mathbf{p},\mathbf{q}\in\Delta_n\,.$$

Proof. **(of Lemma 4)** Calibration of $(\boldsymbol{\tau}^{\text{id}}, \text{pred}^{\text{0-1}(n)})$ for $\mathbf{L}^{\text{0-1}(n)}$ is immediate, since the Bayes optimal classifier for $\mathbf{L}^{\text{0-1}(n)}$ is given by $h_D^{\mathbf{L}^{\text{0-1}(n)},*}(\mathbf{x}) = \text{pred}^{\text{0-1}(n)}(\mathbf{p}(\mathbf{x})) = \text{pred}^{\text{0-1}(n)}(\boldsymbol{\tau}^{\text{id}}(\mathbf{p}(\mathbf{x})))$. Moreover, for any $\mathbf{p},\mathbf{q}\in\Delta_n$, we have

$$\begin{aligned} &\mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\text{pred}^{\text{0-1}(n)}(\mathbf{q})}] - \min_{\widehat{y}\in[n]} \mathbf{E}_{Y\sim\mathbf{p}}[L^{\text{0-1}(n)}_{Y,\widehat{y}}] \\ &= \sum_{y=1}^n p_y\cdot L^{\text{0-1}(n)}_{y,\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \min_{\widehat{y}\in[n]}\sum_{y=1}^n p_y\cdot L^{\text{0-1}(n)}_{y,\widehat{y}} \\ &= \mathbf{p}^\top\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \min_{\widehat{y}\in[n]}\mathbf{p}^\top\boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}} \\ &= \max_{\widehat{y}\in[n]}\Big(\mathbf{p}^\top(\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}})\Big) \\ &= \max_{\widehat{y}\in[n]}\Big((\mathbf{p}-\mathbf{q})^\top(\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}}) + \mathbf{q}^\top(\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}})\Big) \\ &\le \max_{\widehat{y}\in[n]}\Big((\mathbf{p}-\mathbf{q})^\top(\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}})\Big) \quad \text{(by definition of pred}^{\text{0-1}(n)}) \\ &\le \|\mathbf{p}-\mathbf{q}\|_2\cdot\max_{\widehat{y}\in[n]}\|\boldsymbol{\ell}^{\text{0-1}(n)}_{\text{pred}^{\text{0-1}(n)}(\mathbf{q})} - \boldsymbol{\ell}^{\text{0-1}(n)}_{\widehat{y}}\|_2 \quad \text{(by the Cauchy-Schwarz inequality)} \\ &\le \sqrt{2}\cdot\|\mathbf{q}-\mathbf{p}\|_2 \end{aligned}$$

(since the difference between any two columns of $\mathbf{L}^{\text{0-1}(n)}$ has at most two non-zero entries, each with magnitude at most 1)

□

Proof. **(of Theorem 6)** Consider the link function $\boldsymbol{\lambda} : \Delta_n \to \overline{\mathbb{R}}^n$ with extended inverse $\boldsymbol{\lambda}^{-1} : \overline{\mathbb{R}}^n \to \Delta_n$ given by[8]

$$\begin{aligned}(\boldsymbol{\lambda}(\mathbf{p}))_y &= \ln(p_y)\,, \\ (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_y &= \frac{e^{u_y}}{\sum_{y'=1}^n e^{u_{y'}}}\,.\end{aligned}$$

Note that $\boldsymbol{\lambda}^{-1}$ here is equivalent to the softmax function $\boldsymbol{\sigma}$ (defined in the theorem statement). We will observe/prove the following:

(1) $\mathcal{H}_{\text{multiclass-linear}} = \text{pred}^{\text{0-1}(n)} \circ \mathcal{Q}_{\text{softmax-of-mlinear}}$;

(2) ψ^{mlog} is a 1-strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{id}}$ with link function $\boldsymbol{\lambda}$;

(3) $\text{decode}^{\text{0-1}(n)} = \text{pred}^{\text{0-1}(n)} \circ \boldsymbol{\lambda}^{-1}$;

(4) $\mathcal{F}_{\text{multiclass-linear}} = \boldsymbol{\lambda} \circ \mathcal{Q}_{\text{softmax-of-mlinear}}$;

(5) $\psi^{\text{mlog}}(y, \mathbf{f}(\mathbf{x})) \leq \ln(n) + 2RW \quad \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \mathbf{f} \in \mathcal{F}_{\text{multiclass-linear}}$;

(6) ψ^{mlog} is 1-Lipschitz with respect to the L^1 metric and 2-Lipschitz with respect to the Euclidean metric on $\mathbb{R}^n$;

(7) $\mathcal{N}_1(\epsilon, \mathcal{F}^y_{\text{multiclass-linear}}, m) \leq \left(\frac{1}{\epsilon}\right)^p \ \forall y \in [n]$;

(8) $0 \leq \mathcal{R}_m(\mathcal{F}^y_{\text{multiclass-linear}}) \leq RW/\sqrt{m}\ \forall y \in [n]$.

The result will then follow from Lemma 4, Theorem 2, and Theorem 3.

Parts (1), (3), (4), (5) are immediate from the definitions.

Part (7) is a well-known result (e.g. see [5]).

Part (8) is a well-known result (see Proposition 4).

Part (2): ψ^{mlog} has been shown to be a 1-strongly proper composite loss for the property $\boldsymbol{\tau}^{\text{id}}$ (i.e., for multiclass class probability estimation) with link function $\boldsymbol{\lambda}$ as above [42].[9]

Part (6): To see that ψ^{mlog} is 1-Lipschitz with respect to the L^1 metric, note that

$$\begin{aligned}\frac{\partial \psi^{\text{mlog}}(y, \mathbf{u})}{\partial u_y} &= -1 + \frac{e^{u_y}}{\sum_{y'=1}^n e^{u_{y'}}}\,, \\ \frac{\partial \psi^{\text{mlog}}(y, \mathbf{u})}{\partial u_{y''}} &= \frac{e^{u_{y''}}}{\sum_{y'=1}^n e^{u_{y'}}} \quad \forall y'' \neq y\,.\end{aligned}$$

Thus we have,

$$\begin{aligned}\psi^{\text{mlog}}(y, \mathbf{u}_1) - \psi^{\text{mlog}}(y, \mathbf{u}_2) &\leq (\nabla_{\mathbf{u}} \psi^{\text{mlog}}(y, \mathbf{u}_1))^\top (\mathbf{u}_1 - \mathbf{u}_2) \quad \text{(by convexity of } \psi^{\text{mlog}}(y, \cdot)\text{)} \\ &\leq \|\nabla_{\mathbf{u}} \psi^{\text{mlog}}(y, \mathbf{u}_1)\|_\infty \cdot \|\mathbf{u}_1 - \mathbf{u}_2\|_1 \quad \text{(by Hölder's inequality)} \\ &\leq \|\mathbf{u}_1 - \mathbf{u}_2\|_1 \quad \text{(since } |\partial \psi^{\text{mlog}}(y, \mathbf{u})/\partial u_{y''}| \leq 1\ \forall y'' \in [n]\text{)}\,.\end{aligned}$$

[8] Note that in the notation of Definition 3, we use $\mathcal{C}' = \overline{\mathbb{R}}^n$ here. Technically, we would also need to extend the definitions of the surrogate loss ψ^{mlog} and the mapping $\text{decode}^{\text{0-1}(n)}$ to act on $\overline{\mathbb{R}}^n$ instead of $\mathbb{R}^n$: we ignore this issue here for simplicity. Also note that here, $\mathcal{C} = \Delta_n$ is in one-to-one correspondence with only a strict subset of $\mathcal{C}' = \overline{\mathbb{R}}^n$, and so we use an extended inverse; in particular, we use the partition $\mathcal{S} = \{\mathcal{S}_{\mathbf{p}} : \mathbf{p} \in \Delta_n\}$ of $\mathcal{C}' = \overline{\mathbb{R}}^n$ given by $\mathcal{S}_{\mathbf{p}} = \{\mathbf{u} \in \overline{\mathbb{R}}^n \mid \exists c \in \mathbb{R} \text{ s.t. } u_y = \ln(p_y) + c\ \forall y\}$.

[9] Note that [42] show this result for a slight variant of ψ^{mlog} that acts on $\overline{\mathbb{R}}^{n-1}$ rather than $\overline{\mathbb{R}}^n$; however, essentially the same proof works for the variant we use here as well.

Next, to see that ψ^{mlog} is 2-Lipschitz with respect to the Euclidean metric, note that

$$\begin{aligned}
\psi^{\text{mlog}}(y,\mathbf{u}_1) - \psi^{\text{mlog}}(y,\mathbf{u}_2) &\leq (\nabla_{\mathbf{u}}\psi^{\text{mlog}}(y,\mathbf{u}_1))^\top(\mathbf{u}_1-\mathbf{u}_2) \quad \text{(by convexity of } \psi^{\text{mlog}}(y,\cdot)\text{)} \\
&\leq \|\nabla_{\mathbf{u}}\psi^{\text{mlog}}(y,\mathbf{u}_1)\|_2 \cdot \|\mathbf{u}_1-\mathbf{u}_2\|_2 \quad \text{(Cauchy-Schwarz inequality)} \\
&\leq \|\nabla_{\mathbf{u}}\psi^{\text{mlog}}(y,\mathbf{u}_1)\|_1 \cdot \|\mathbf{u}_1-\mathbf{u}_2\|_2 \\
&\leq 2\cdot\|\mathbf{u}_1-\mathbf{u}_2\|_2 \\
&\qquad \text{(since } |\partial\psi^{\text{mlog}}(y,\mathbf{u})/\partial u_y| \leq 1 \text{ and} \\
&\qquad \sum_{y''\neq y} |\partial\psi^{\text{mlog}}(y,\mathbf{u})/\partial u_{y''}| = 1 - \frac{e^{u_y}}{\sum_{y'=1}^n e^{u_{y'}}} \leq 1 \; \forall y'' \neq y)\,.
\end{aligned}$$

(i) Combining all the above together with Lemma 4 and applying Theorem 2 (with $\kappa = \sqrt{2}$, $\gamma = 1$, $\rho_1 = 1$, $d = n$, $\mathcal{N}_1(\epsilon, \mathcal{F}^y_{\text{multiclass-linear}}, m) \leq \left(\frac{1}{\epsilon}\right)^p \;\; \forall y \in [n]$, and $B \leq \ln(n) + 2RW$) gives the desired result with squared $\boldsymbol{\tau}^{\text{id}}$ estimation error sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{id}}}(\epsilon,\delta) &\leq \frac{1152\,(\ln(n)+2RW)^2}{\epsilon^2}\left(np\ln\left(\frac{48n}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right) \\
&= O\left(\frac{(\ln(n))^2}{\epsilon^2}\left(np\ln\left(\frac{n}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)\,.
\end{aligned}$$

and with target loss sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\mathbf{L}^{0\text{-}1(n)}}(\epsilon,\delta) &\leq \frac{4608\,(\ln(n)+2RW)^2}{\epsilon^4}\left(np\ln\left(\frac{96n}{\epsilon^2}\right) + \ln\left(\frac{4}{\delta}\right)\right) \\
&= O\left(\frac{(\ln(n))^2}{\epsilon^4}\left(np\ln\left(\frac{n}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)\,.
\end{aligned}$$

(ii) Next, combining all the above together with Lemma 4 and applying Theorem 3 (with $\kappa = \sqrt{2}$, $\gamma = 1$, $\rho_2 = 2$, $d = n$, $0 \leq \mathcal{R}_m(\mathcal{F}^y_{\text{multiclass-linear}}) \leq RW/\sqrt{m} \;\forall y \in [n]$, and $B \leq \ln(n) + 2RW$) gives the desired result with squared $\boldsymbol{\tau}^{\text{id}}$ estimation error sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{id}}}(\epsilon,\delta) &\leq \frac{36}{\epsilon^2}\left(4\sqrt{2}RWn + (\ln(n)+2RW)\sqrt{\ln(2/\delta)}\right)^2 \\
&= O\left(\frac{1}{\epsilon^2}\left(n^2 + (\ln(n))^2\cdot\ln\left(\frac{1}{\delta}\right)\right)\right)\,.
\end{aligned}$$

and with target loss sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\mathbf{L}^{0\text{-}1(n)}}(\epsilon,\delta) &\leq \frac{144}{\epsilon^4}\left(4\sqrt{2}RWn + (\ln(n)+2RW)\sqrt{\ln(2/\delta)}\right)^2 \\
&= O\left(\frac{1}{\epsilon^4}\left(n^2 + (\ln(n))^2\cdot\ln\left(\frac{1}{\delta}\right)\right)\right)\,.
\end{aligned}$$

Combining the above bounds yields the desired results. □

E Supplement to Section 5 (Multi-Label Learning)

Lemma 5. $(\boldsymbol{\tau}^{\text{marginals}}, \text{pred}^{\text{Ham}})$ *is an* $\mathbf{L}^{\text{Ham}}$*-calibrated statistic-mapping pair, with*

$$\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\text{pred}^{\text{Ham}}(\mathbf{q})}] - \min_{\widehat{\mathbf{y}}\in\{0,1\}^s}\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\widehat{\mathbf{y}}}] \;\leq\; 2\sqrt{s}\,\|\mathbf{q}-\boldsymbol{\tau}^{\text{marginals}}(\mathbf{p})\|_2 \quad \forall \mathbf{p}\in\Delta_{\{0,1\}^s}, \mathbf{q}\in[0,1]^s\,.$$

Proof. **(of Lemma 5)** Calibration of $(\boldsymbol{\tau}^{\text{marginals}}, \text{pred}^{\text{Ham}})$ for $\mathbf{L}^{\text{Ham}}$ is immediate, since the Bayes optimal classifier for $\mathbf{L}^{\text{Ham}}$ is given by $h_D^{\mathbf{L}^{\text{Ham}},*}(\mathbf{x}) = \text{pred}^{\text{Ham}}(\boldsymbol{\tau}^{\text{marginals}}(\mathbf{p}(\mathbf{x})))$. Moreover, for any

$\mathbf{p} \in \Delta_{\{0,1\}^s}, \mathbf{q} \in [0,1]^s$, we have

$$\begin{aligned}
&\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\text{pred}^{\text{Ham}}(\mathbf{q})}] - \min_{\widehat{\mathbf{y}}\in\{0,1\}^s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{Ham}}_{\mathbf{Y},\widehat{\mathbf{y}}}] \\
&= \sum_{j=1}^s \mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,(\text{pred}^{\text{Ham}}(\mathbf{q}))_j}] - \min_{\widehat{\mathbf{y}}\in\{0,1\}^s} \sum_{j=1}^s \mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,\widehat{y}_j}] \\
&\qquad \text{(by linearity of expectation; here } \mathbf{L}^{\text{0-1}} \in \mathbb{R}_+^{\{0,1\}\times\{0,1\}} \text{ denotes the} \\
&\qquad \text{binary loss } L^{\text{0-1}}_{y,\widehat{y}} = \mathbf{1}(\widehat{y} \neq y)) \\
&= \sum_{j=1}^s \mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,(\text{pred}^{\text{Ham}}(\mathbf{q}))_j}] - \sum_{j=1}^s \min_{\widehat{y}_j\in\{0,1\}} \mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,\widehat{y}_j}] \\
&= \sum_{j=1}^s \left(\mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,(\text{pred}^{\text{Ham}}(\mathbf{q}))_j}] - \min_{\widehat{y}_j\in\{0,1\}} \mathbf{E}_{Y_j}[L^{\text{0-1}}_{Y_j,\widehat{y}_j}] \right) \\
&\le \sum_{j=1}^s 2\,|q_j - (\boldsymbol{\tau}^{\text{marginals}}(\mathbf{p}))_j| \\
&\qquad \text{(by well-known result for binary 0-1 loss, as also shown in the proof of Theorem 5)} \\
&= 2\,\|\mathbf{q} - \boldsymbol{\tau}^{\text{marginals}}(\mathbf{p})\|_1 \\
&\le 2\sqrt{s}\,\|\mathbf{q} - \boldsymbol{\tau}^{\text{marginals}}(\mathbf{p})\|_2\,.
\end{aligned}$$

□

Proof. **(of Theorem 6)** Consider the (invertible) link function $\boldsymbol{\lambda} : [0,1]^s \to \overline{\mathbb{R}}^s$ and its inverse $\boldsymbol{\lambda}^{-1} : \overline{\mathbb{R}}^s \to [0,1]^s$ given by[10]

$$\begin{aligned}
(\boldsymbol{\lambda}(\mathbf{q}))_j &= \ln\left(\frac{q_j}{1-q_j}\right), \\
(\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j &= \frac{1}{1+e^{-u_j}}\,.
\end{aligned}$$

Note that each component of $\boldsymbol{\lambda}^{-1}$ here is equivalent to the sigmoid function σ (defined in the theorem statement). We will observe/prove the following:

(1) $\mathcal{H}^{\text{sign}}_{\text{multilinear}} = \text{pred}^{\text{Ham}} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$;

(2) ψ^{BRlog} is a 4-strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{marginals}}$ with link function $\boldsymbol{\lambda}$;

(3) $\text{decode}^{\text{Ham}} = \text{pred}^{\text{Ham}} \circ \boldsymbol{\lambda}^{-1}$;

(4) $\mathcal{F}_{\text{multilinear}} = \boldsymbol{\lambda} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$;

(5) $\psi^{\text{BRlog}}(y, \mathbf{f}(\mathbf{x})) \le s\ln(1+e^{RW}) \;\; \forall(\mathbf{x}, y) \in \mathcal{X}\times\mathcal{Y}, \mathbf{f} \in \mathcal{F}_{\text{multilinear}}$;

(6) ψ^{BRlog} is 1-Lipschitz with respect to the L^1 metric and $\sqrt{s}$-Lipschitz with respect to the Euclidean metric on $\mathbb{R}^s$;

(7) $\mathcal{N}_1(\epsilon, \mathcal{F}^j_{\text{multilinear}}, m) \le \left(\frac{1}{\epsilon}\right)^p \;\; \forall j \in [s]$;

(8) $0 \le \mathcal{R}_m(\mathcal{F}^j_{\text{multilinear}}) \le RW/\sqrt{m} \; \forall j \in [s]$.

The result will then follow from Lemma 5, and Theorem 3.

Parts (1), (3), (4), (5) are immediate from the definitions.

[10]Note that in the notation of Definition 3, we use $\mathcal{C}' = \overline{\mathbb{R}}^s$ here. Technically, we would also need to extend the definitions of the surrogate loss ψ and the mapping decode to act on $\overline{\mathbb{R}}^s$ instead of $\mathbb{R}^s$: we ignore this issue here for simplicity.

Part (7) is a well-known result (e.g. see [5]).

Part (8) is a well-known result (see Proposition 4).

Part (2): The fact that ψ^{BRlog} is a 4-strongly proper composite loss for the property $\boldsymbol{\tau}^{\text{marginals}}$ with link function $\boldsymbol{\lambda}$ as above follows from 4-strong proper compositeness of the binary logistic loss ($\psi^{\log}$ as defined in Theorem 5) for binary class probability estimation, applied separately to each component of the loss [3].

Part (6): The fact that ψ^{BRlog} is 1-Lipschitz with respect to the L^1 metric follows directly from the fact that the binary logistic loss ($\psi^{\log}$ as defined in Theorem 5) is 1-Lipschitz with respect to the L^1 metric, applied separately to each component of the loss. This also implies it is $\sqrt{s}$-Lipschitz with respect to the Euclidean metric.

Combining all the above together with Lemma 5 and applying Theorem 3 (with $\kappa = 2\sqrt{s}$, $\gamma = 4$, $\rho_2 = \sqrt{s}$, $d = s$, $0 \leq \mathcal{R}_m(\mathcal{F}^j) \leq RW/\sqrt{m}\ \forall j$, and $B \leq s\ln(1+e^{RW})$) gives the desired result with squared $\boldsymbol{\tau}^{\text{marginals}}$ estimation error sample complexity

$$\begin{aligned} m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{marginals}}}(\epsilon,\delta) &\leq \frac{9}{4\epsilon^2}\left(2\sqrt{2}RWs^{3/2} + s(\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\right)^2 \\ &= O\left(\frac{s^2}{\epsilon^2}\left(s + \ln\left(\frac{1}{\delta}\right)\right)\right) . \end{aligned}$$

and with target loss sample complexity

$$\begin{aligned} m_{\mathcal{A}}^{\mathbf{L}^{\text{Ham}}}(\epsilon,\delta) &\leq \frac{36s^2}{\epsilon^4}\left(2\sqrt{2}RWs^{3/2} + s(\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\right)^2 \\ &= O\left(\frac{s^4}{\epsilon^4}\left(s + \ln\left(\frac{1}{\delta}\right)\right)\right) . \end{aligned}$$

□

F Supplement to Section 6 (Subset Ranking)

Lemma 6. $(\boldsymbol{\tau}^{\text{sc-marg-exp}}, \text{pred}^{\text{DCG}})$ *is an* $\mathbf{L}^{\text{DCG}}$*-calibrated statistic-mapping pair, with*

$$\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\text{pred}^{\text{DCG}}(\mathbf{q})}] - \min_{\widehat{\pi}\in\Pi_s}\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\widehat{\pi}}] \quad \leq \quad 2r\cdot\|\mathbf{disc}\|_2\cdot\|\mathbf{q}-\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})\|_2 \qquad \forall \mathbf{p}\in\Delta_{\{0,1,\ldots,r\}^s}, \mathbf{q}\in[0,1]^s ,$$

where $\mathbf{disc} = (\text{disc}(1),\ldots,\text{disc}(s))^\top \in [0,1]^s$.

Proof. **(of Lemma 6)** Calibration of $(\boldsymbol{\tau}^{\text{sc-marg-exp}}, \text{pred}^{\text{DCG}})$ for $\mathbf{L}^{\text{DCG}}$ is immediate, since the Bayes optimal classifier for $\mathbf{L}^{\text{DCG}}$ is given by $h_D^{\mathbf{L}^{\text{DCG}},*}(\mathbf{x}) \in \text{argsort}(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}(\mathbf{x}))) = \text{pred}^{\text{DCG}}(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}(\mathbf{x})))$. In the following, for any $\mathbf{q}\in[0,1]^s$, we will denote

$$\widehat{\pi}^{\mathbf{q}} \ := \ \text{pred}^{\text{DCG}}(\mathbf{q}) \ \in \ \Pi_s ,$$

and for any $\widehat{\pi}\in\Pi_s$, we will denote

$$\mathbf{disc}_{\widehat{\pi}} := (\text{disc}(\widehat{\pi}(1)),\ldots,\text{disc}(\widehat{\pi}(s)))^\top \in [0,1]^s .$$

[CONTINUED ON NEXT PAGE]

Then for any $\mathbf{p} \in \Delta_{\{0,1,\ldots,r\}^s}, \mathbf{q} \in [0,1]^s$, we have

$$
\begin{aligned}
&\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\text{pred}^{\text{DCG}}(\mathbf{q})}] - \min_{\widehat{\pi}\in\Pi_s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[L^{\text{DCG}}_{\mathbf{Y},\widehat{\pi}}] \\
&= \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}\left[Z - \sum_{j=1}^{s} Y_j \cdot \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\right] - \min_{\widehat{\pi}\in\Pi_s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}\left[Z - \sum_{j=1}^{s} Y_j \cdot \text{disc}(\widehat{\pi}(j))\right] \\
&= \max_{\widehat{\pi}\in\Pi_s} \sum_{j=1}^{s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}\left[Y_j \cdot \text{disc}(\widehat{\pi}(j))\right] - \sum_{j=1}^{s} \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}\left[Y_j \cdot \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\right] \quad \text{(by linearity of expectation)} \\
&= \max_{\widehat{\pi}\in\Pi_s} \sum_{j=1}^{s} \mathbf{E}_{Y_j}[Y_j] \cdot \text{disc}(\widehat{\pi}(j)) - \sum_{j=1}^{s} \mathbf{E}_{Y_j}[Y_j] \cdot \text{disc}(\widehat{\pi}^{\mathbf{q}}(j)) \\
&= r \max_{\widehat{\pi}\in\Pi_s} \sum_{j=1}^{s} (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \cdot \text{disc}(\widehat{\pi}(j)) - \sum_{j=1}^{s} (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \cdot \text{disc}(\widehat{\pi}^{\mathbf{q}}(j)) \\
&= r \max_{\widehat{\pi}\in\Pi_s} \sum_{j=1}^{s} (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \cdot \Big(\text{disc}(\widehat{\pi}(j)) - \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\Big) \\
&= r \max_{\widehat{\pi}\in\Pi_s} \left(\sum_{j=1}^{s} \Big((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j - q_j\Big) \cdot \Big(\text{disc}(\widehat{\pi}(j)) - \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\Big) \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{s} q_j \cdot \Big(\text{disc}(\widehat{\pi}(j)) - \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\Big) \right) \\
&\leq r \max_{\widehat{\pi}\in\Pi_s} \sum_{j=1}^{s} \Big((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j - q_j\Big) \cdot \Big(\text{disc}(\widehat{\pi}(j)) - \text{disc}(\widehat{\pi}^{\mathbf{q}}(j))\Big) \quad \text{(by definition of } \widehat{\pi}^{\mathbf{q}}\text{)} \\
&= r \max_{\widehat{\pi}\in\Pi_s} \left(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}) - \mathbf{q}\right)^{\top} \left(\mathbf{disc}_{\widehat{\pi}} - \mathbf{disc}_{\widehat{\pi}^{\mathbf{q}}}\right) \\
&\leq r \max_{\widehat{\pi}\in\Pi_s} \|\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}) - \mathbf{q}\|_2 \cdot \|\mathbf{disc}_{\widehat{\pi}} - \mathbf{disc}_{\widehat{\pi}^{\mathbf{q}}}\|_2 \quad \text{(by the Cauchy-Schwarz inequality)} \\
&\leq 2r \Big(\max_{\widehat{\pi}\in\Pi_s} \|\mathbf{disc}_{\widehat{\pi}}\|_2\Big) \cdot \|\mathbf{q} - \boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})\|_2 \\
&= 2r \cdot \|\mathbf{disc}\|_2 \cdot \|\mathbf{q} - \boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})\|_2 \quad (\text{since } \|\mathbf{disc}_{\widehat{\pi}}\|_2 = \|\mathbf{disc}\|_2 \; \forall \widehat{\pi} \in \Pi_s)\,.
\end{aligned}
$$

□

Proof. **(of Theorem 8)** Consider the (invertible) link function $\boldsymbol{\lambda} : [0,1]^s \to \overline{\mathbb{R}}^s$ and its inverse $\boldsymbol{\lambda}^{-1} : \overline{\mathbb{R}}^s \to [0,1]^s$ given by[11]

$$
\begin{aligned}
(\boldsymbol{\lambda}(\mathbf{q}))_j &= \ln\left(\frac{q_j}{1-q_j}\right), \\
(\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j &= \frac{1}{1+e^{-u_j}}\,.
\end{aligned}
$$

Note that each component of $\boldsymbol{\lambda}^{-1}$ here is equivalent to the sigmoid function σ (defined in the theorem statement). We will observe/prove the following:

(1) $\mathcal{H}^{\text{sort}}_{\text{multilinear}} = \text{pred}^{\text{DCG}} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$;

(2) ψ^{wlog} is a 4-strongly proper composite surrogate loss for $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ with link function $\boldsymbol{\lambda}$;

(3) $\text{decode}^{\text{DCG}} = \text{pred}^{\text{DCG}} \circ \boldsymbol{\lambda}^{-1}$;

(4) $\mathcal{F}_{\text{multilinear}} = \boldsymbol{\lambda} \circ \mathcal{Q}_{\text{sigmoid-of-multilinear}}$;

[11]Note that in the notation of Definition 3, we use $\mathcal{C}' = \overline{\mathbb{R}}^s$ here. Technically, we would also need to extend the definitions of the surrogate loss ψ and the mapping decode to act on $\overline{\mathbb{R}}^s$ instead of $\mathbb{R}^s$: we ignore this issue here for simplicity.

(5) $\psi^{\text{wlog}}(y, \mathbf{f}(\mathbf{x})) \leq s \ln(1 + e^{RW}) \quad \forall (\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}, \mathbf{f} \in \mathcal{F}_{\text{multilinear}}$;

(6) ψ^{wlog} is 1-Lipschitz with respect to the L^1 metric and $\sqrt{s}$-Lipschitz with respect to the Euclidean metric on $\mathbb{R}^s$;

(7) $\mathcal{N}_1(\epsilon, \mathcal{F}^j_{\text{multilinear}}, m) \leq \left(\frac{1}{\epsilon}\right)^p \ \forall j \in [s]$;

(8) $0 \leq \mathcal{R}_m(\mathcal{F}^j_{\text{multilinear}}) \leq RW/\sqrt{m} \ \forall j \in [s]$.

The result will then follow from Lemma 6 and Theorem 3.

Parts (1), (3), (4), (5) are immediate from the definitions.

Part (7) is a well-known result (e.g. see [5]).

Part (8) is a well-known result (see Proposition 4).

Part (2): We show here that ψ^{wlog} is a 4-strongly proper composite loss for the property $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ with link function $\boldsymbol{\lambda}$ as above. In particular, we have:

$$\begin{aligned}
&\mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[\psi^{\text{wlog}}(\mathbf{Y}, \mathbf{u}) - \psi^{\text{wlog}}(\mathbf{Y}, \boldsymbol{\lambda}(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})))] \\
&= \sum_{j=1}^{s} \Bigg(\left(\frac{\mathbf{E}[Y_j]}{r} \right) \cdot \Big(\ln(1 + e^{-u_j}) - \ln(1 + e^{-(\boldsymbol{\lambda}(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})))_j}) \Big) \\
&\qquad\qquad + \left(1 - \frac{\mathbf{E}[Y_j]}{r} \right) \cdot \Big(\ln(1 + e^{u_j}) - \ln(1 + e^{(\boldsymbol{\lambda}(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p})))_j}) \Big) \Bigg) \\
&= \sum_{j=1}^{s} \Bigg((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \cdot \Big(- \ln((\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j) + \ln((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j) \Big) \\
&\qquad\qquad + \Big(1 - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \Big) \cdot \Big(- \ln(1 - (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j) + \ln(1 - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j) \Big) \Bigg) \\
&\qquad\qquad\qquad\qquad \text{(by definition of } \boldsymbol{\lambda} \text{ and } \boldsymbol{\lambda}^{-1}) \\
&= \sum_{j=1}^{s} \Bigg((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \cdot \ln \left(\frac{(\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j}{(\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j} \right) \\
&\qquad\qquad + \Big(1 - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \Big) \cdot \ln \left(\frac{1 - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j}{1 - (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j} \right) \Bigg) \\
&= \sum_{j=1}^{s} D_{KL}\Big((\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \ \Big|\Big| \ (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j \Big) \\
&\qquad\qquad \text{(by definition of Kullback-Leibler divergence for binary-valued random variables)} \\
&\geq \frac{1}{2} \sum_{j=1}^{s} \left(\left| (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \right| + \left| (1 - (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j) - (1 - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j) \right| \right)^2 \\
&\qquad\qquad \text{(by Pinsker's inequality and properties of the total variation distance)} \\
&= \frac{1}{2} \sum_{j=1}^{s} \left(2 \left| (\boldsymbol{\lambda}^{-1}(\mathbf{u}))_j - (\boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}))_j \right| \right)^2 \\
&= 2 \, \| \boldsymbol{\lambda}^{-1}(\mathbf{u}) - \boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}) \|_1^2 \\
&\geq 2 \, \| \boldsymbol{\lambda}^{-1}(\mathbf{u}) - \boldsymbol{\tau}^{\text{sc-marg-exp}}(\mathbf{p}) \|_2^2 \, .
\end{aligned}$$

Thus ψ^{wlog} is a 4-strongly proper composite loss for the property $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ with link function $\boldsymbol{\lambda}$.

Part (6): It is easy to see that the weighted binary logistic loss $\psi^{\text{wlog,bin}} : [0,1] \times \mathbb{R} \to \mathbb{R}_+$ defined as

$$\psi^{\text{wlog,bin}}(\alpha, u) \;=\; \alpha \cdot \ln(1 + e^{-u}) + (1 - \alpha) \cdot \ln(1 + e^{u})$$

is 1-Lipschitz (in particular, the absolute value of the derivative with respect to u is upper bounded by 1). The fact that the surrogate loss ψ^{wlog} is 1-Lipschitz with respect to the L^1 metric then follows directly from this observation, applied separately to each component of the loss (with weight $\alpha = y_j/r$ for component j). This also implies it is $\sqrt{s}$-Lipschitz with respect to the Euclidean metric.

Combining all the above together with Lemma 5 and applying Theorem 3 (with $\kappa = 2r \cdot \|\textbf{disc}\|_2$, $\gamma = 4$, $\rho_2 = \sqrt{s}$, $d = s$, $0 \leq \mathcal{R}_m(\mathcal{F}^j) \leq RW/\sqrt{m}\ \forall j$, and $B \leq s\ln(1+e^{RW})$) gives the desired result with squared $\boldsymbol{\tau}^{\text{sc-marg-exp}}$ estimation error sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\boldsymbol{\tau}^{\text{sc-marg-exp}}}(\epsilon,\delta) &\leq \frac{9}{4\epsilon^2}\left(2\sqrt{2}RWs^{3/2} + s(\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\right)^2 \\
&= O\left(\frac{s^2}{\epsilon^2}\left(s + \ln\left(\frac{1}{\delta}\right)\right)\right) .
\end{aligned}$$

and with target loss sample complexity

$$\begin{aligned}
m_{\mathcal{A}}^{\mathbf{L}^{\text{DCG}}}(\epsilon,\delta) &\leq \frac{36r^4 \cdot \|\textbf{disc}\|_2^4}{\epsilon^4}\left(2\sqrt{2}RWs^{3/2} + s(\ln(1+e^{RW}))\sqrt{\ln(2/\delta)}\right)^2 \\
&= O\left(\frac{r^4 s^2 \cdot \|\textbf{disc}\|_2^4}{\epsilon^4}\left(s + \ln\left(\frac{1}{\delta}\right)\right)\right) \\
&= O\left(\frac{r^4 s^4}{\epsilon^4}\left(s + \ln\left(\frac{1}{\delta}\right)\right)\right) \qquad (\text{since } \|\textbf{disc}\|_2 \leq \sqrt{s})\,.
\end{aligned}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the results are clearly stated and explained in the paper, and proofs are provided in the Appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: The paper studies learning problems that in many ways are more general than those studied in previous work. Both sample and computational complexity bounds are provided for the algorithms discussed. We do not foresee any significant limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For all results in the paper, all assumptions are clearly stated and complete proofs are provided in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The paper does not include experiments requiring code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: The paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have made every effort to conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: While this work can lead to a better understanding of various types of learning problems and their possible solutions, the work is largely theoretical/foundational in nature and does not have immediate societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks (there are no data or models to be released).

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: The paper does not use existing assets of the form described.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components (nor does the writing of the paper use LLMs in any form).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.