Beyond Tsybakov: Model Margin Noise and $\mathcal{H}$-Consistency Bounds

Mehryar Mohri[1,2], **Yutao Zhong**[1]

[1] Google Research, New York
[2] Courant Institute of Mathematical Sciences, New York University
mohri@google.com, yutaozhong@google.com

Abstract

We introduce a new low-noise condition for classification, the *Model Margin Noise (MM noise)* assumption, and derive enhanced $\mathcal{H}$-consistency bounds under this condition. MM noise is *weaker* than Tsybakov noise condition: it is implied by Tsybakov noise condition but can hold even when Tsybakov fails, because it depends on the discrepancy between a given hypothesis and the Bayes-classifier rather than on the intrinsic distributional minimal margin (see Figure 1 for an illustration of an explicit example). This hypothesis-dependent assumption yields enhanced $\mathcal{H}$-consistency bounds for both binary and multi-class classification. Our results extend the enhanced $\mathcal{H}$-consistency bounds of Mao, Mohri, and Zhong (2025a) with the same favorable exponents but under a weaker assumption than the Tsybakov noise condition; they interpolate smoothly between linear and square-root regimes for intermediate noise levels. We also instantiate these bounds for common surrogate loss families and provide illustrative tables.

1 Introduction

The design and analysis of surrogate losses are fundamental to classification. Classical analyses established *Bayes-consistency* for large families of convex surrogates and derived bounds on the surrogate-to-target excess error in the binary setting (Zhang 2004b; Bartlett, Jordan, and McAuliffe 2006; Steinwart 2007), with sharper constants for specific losses such as q-norm SVMs and proper losses (Chen et al. 2004; Reid and Williamson 2009). In multi-class classification, subsequent work identified which surrogates are Bayes-consistent and which fail (e.g., certain hinge variants), while establishing consistency for sum-exponential/logistic and constrained families (Zhang 2004a; Tewari and Bartlett 2007; Crammer and Singer 2001; Weston and Watkins 1999; Lee, Lin, and Wahba 2004). A complementary thread investigated how *growth rates* of these bounds behave near zero: smoothing often degrades rates, polyhedral losses can achieve linear behavior, and broad smooth/proper losses admit square-root lower bounds (Mahdavi, Zhang, and Jin 2014; Frongillo and Waggoner 2021; Bao 2023). Yet, these guarantees apply only to the family of all measurable functions $\mathcal{H}_{\text{all}}$ and are thus not relevant to the hypothesis set that is actually used in practice.

To close that gap, $\mathcal{H}$-*consistency bounds* provide non-asymptotic guarantees specific to a fixed hypothesis set $\mathcal{H}$ (Awasthi et al. 2022a; Zhong 2025), and have since been developed broadly, from multi-class (including max, sum, constrained, and comp-sum families) to ranking, structured prediction, abstention/defer, multi-label learning, adversarial settings, and beyond (Awasthi et al. 2022b; Zheng et al. 2023; Mao, Mohri, and Zhong 2023a,c,e, 2024d; Mao et al. 2023; Awasthi et al. 2023b; Mao, Mohri, and Zhong 2024b; Cortes et al. 2024; Mao, Mohri, and Zhong 2024h). More recently, Mao, Mohri, and Zhong (2025a) showed how *enhanced $\mathcal{H}$-consistency bounds* can be derived by relating *conditional regrets* via more general inequalities with non-constant factors depending on the input or predictor instance. They showed, in particular, that when a lower bound of the surrogate loss *conditional regret* is given as a power function of the target *conditional regret* with exponent s, this yields enhanced bounds with low-noise exponents of the form $(1/(s-\alpha(s-1)))$ under Tsybakov noise condition.

The Tsybakov noise condition constrains the minimal margin $\gamma(x) = \mathbb{P}(y_{\max} \mid x) - \sup_{y \neq y_{\max}} \mathbb{P}(y \mid x)$ with $y_{\max} = \operatorname{argmax}_{y \in \mathcal{Y}} \mathbb{P}(y \mid x)$ and is therefore purely distributional, describing a worst-case property of the data, regardless of the hypothesis set $\mathcal{H}$ being used. We introduce a hypothesis-dependent low-noise assumption, the *Model Margin (MM) noise*, that depends on the *model margin* $\mu(h,x) = \mathbb{P}(\mathsf{h}^*(x) \mid x) - \mathbb{P}(\mathsf{h}(x) \mid x) \geq 0$, where $h^* \in \mathcal{H}_{\text{all}}$ is a Bayes classifier and $\mathsf{h}(x) = \operatorname{argmax}_{y \in \mathcal{Y}} h(x,y)$ is the prediction made by hypothesis $h: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. This is the gap between the Bayes label's conditional probability and the hypothesis's predicted label's conditional probability at x This hypothesis-dependent condition is *weaker* than the distributional Tsybakov noise condition: since $\mu(h,x) \geq \gamma(x)$ holds whenever $\mathsf{h}^*(x) \neq \mathsf{h}(x)$, any bound on the minimal margin tail immediately bounds the model margin tail (hence, Tsybakov implies MM), but the converse need not hold, and it enables $\mathcal{H}$-consistency bounds with the same favorable noise exponents, but under a weaker assumption.

Our contributions. We summarize our main results below.

- **Model Margin (MM) noise condition.** We introduce the MM noise condition (Section 4.1), a hypothesis-dependent alternative to Tsybakov noise condition. We formally establish that it is a *weaker* condition ("Tsybakov $\Rightarrow$ MM") and can hold even when Tsybakov noise condition fails (Theorem 4).
- **Key property.** In Section 4.2, we establish a key property

of the MM noise condition (Lemma 5), which provides an inequality bounding the disagreement mass $\mathbb{E}\big[1_{\mu(h,X)>0}\big]$ by a power of the 0-1 excess error. This property is central to the derivation of our $\mathcal{H}$-consistency bounds.

- **Enhanced $\mathcal{H}$-consistency bounds under MM noise.** We derive enhanced $\mathcal{H}$-consistency bounds under MM noise for both binary (Section 5.2) and multi-class classification (Section 5.1). These bounds preserve the favorable exponents from (Mao, Mohri, and Zhong 2025a) while requiring only weaker assumptions.
- **Applications.** In Section 6, we provide structural properties (monotonicity and invariance) of MM noise and illustrate our bounds for common surrogate losses, including binary margin-based losses and multi-class comp-sum losses.

Relation to prior work. Our results improve enhanced $\mathcal{H}$-consistency bounds from (Mao, Mohri, and Zhong 2025a) in two ways. First, shifting from distributional minimal margin γ to hypothesis-dependent model margin μ yields a *weaker* low-noise assumption: the Tsybakov condition implies MM (for any fixed $\mathcal{H}$), but MM may still hold when the distribution violates Tsybakov, especially for restricted hypothesis sets, as $\mathcal{H}$ may not contain the 'bad' classifiers that Tsybakov's worst-case margin $\gamma(x)$ is designed to guard against. Second, this shift produces *predictor-dependent* constants, quantified by $\mathbb{E}\big[1_{\mu(h,X)>0}\big]^{1/t}$, which are absent from Tsybakov-only analyses and which align better with model selection in practice. At the same time, we retain the desirable exponents proven under Tsybakov noise — so practitioners gain bounds with the same exponents under a weaker assumption.

2 Related work.

Bayes-consistency and surrogate losses. The study of Bayes-consistency for convex surrogate losses has a long history. For binary classification, early foundational analyses by Zhang (2004b), Bartlett, Jordan, and McAuliffe (2006), and Steinwart (2007) established Bayes-consistency of several convex and margin-based losses, while also deriving excess-error or surrogate-regret bounds. Specific examples include q-norm SVM losses with optimal square-root rates (Chen et al. 2004) and tight regret bounds for proper losses (Reid and Williamson 2009).

In multi-class classification, analogous results were developed by Zhang (2004a) and Tewari and Bartlett (2007), who analyzed *max*-, *sum*-, and *constrained*-type surrogates (Crammer and Singer 2001; Weston and Watkins 1999; Lee, Lin, and Wahba 2004). They demonstrated that max and sum multi-class hinge variants fail to be Bayes-consistent, whereas sum-exponential and sum-logistic losses, and constrained families achieve Bayes-consistency. Later work unified binary and multi-class analyses under a general supervised learning framework (Steinwart 2007).

Growth rates and smoothness effects. The growth rates of excess-error bounds, that is, the behavior of the bound's functional form Γ near zero, have been studied extensively. Smoothing a hinge-type loss can worsen this growth (Mahdavi, Zhang, and Jin 2014), while local strong convexity and Lipschitz gradients imply at best square-root rates (Frongillo and Waggoner 2021; Bao 2023). Polyhedral losses attain linear rates (Finocchiaro, Frongillo, and Waggoner 2019), clarifying why piecewise-linear surrogates such as the hinge are statistically optimal in that sense. These results, however, apply only to the family of all measurable functions and thus ignore the hypothesis-set choices that dominate practical performance.

Hypothesis-dependent analysis. Bayes-consistency guarantees are asymptotic and model-agnostic. As emphasized by Long and Servedio (2013) and Zhang and Agarwal (2020), a Bayes-consistent surrogate may yield constant test error on restricted model families, while an inconsistent one may succeed. This observation motivated the introduction of $\mathcal{H}$-*consistency bounds* by Awasthi et al. (2022a), which relate the target estimation error within a restricted hypothesis set $\mathcal{H}$ to the surrogate estimation error in a non-asymptotic manner.

$\mathcal{H}$-consistency bounds. Following the binary framework of Awasthi et al. (2022a), Awasthi et al. (2022b) extended $\mathcal{H}$-consistency to multi-class settings, covering *max*, *sum*, and *constrained loss* families (Crammer and Singer 2001; Weston and Watkins 1998; Lee, Lin, and Wahba 2004). Mao, Mohri, and Zhong (2023a) further extended these analyses to the *comp-sum losses*, encompassing cross-entropy, generalized cross-entropy, mean absolute error, and other hybrid surrogate losses. A general characterization for comp-sum and constrained losses was later provided in Mao, Mohri, and Zhong (2023b). Recent refinements revealed that smooth surrogates across both binary and multi-class settings exhibit a universal square-root growth rate (Mao, Mohri, and Zhong 2024h). The $\mathcal{H}$-consistency framework has also been adapted to ranking (Mao, Mohri, and Zhong 2023d,c), abstention and rejection learning (Mao, Mohri, and Zhong 2024c,g; Mohri et al. 2024), learning to defer (Mao, Mohri, and Zhong 2024d,e, 2025b; Mao et al. 2023; Mao 2025; DeSalvo et al. 2025), top-k classification (Cortes et al. 2024), multi-label learning (Mao, Mohri, and Zhong 2024b), adversarial robustness (Awasthi et al. 2021a,b, 2023a,b), bounded regression (Mao, Mohri, and Zhong 2024f,a), optimization of generalized metrics (Mao, Mohri, and Zhong 2025c), imbalanced learning (Cortes et al. 2025; Cortes, Mohri, and Zhong 2025), and structured prediction (Mao, Mohri, and Zhong 2023e).

Enhanced $\mathcal{H}$-consistency bounds. Mao, Mohri, and Zhong (2025a) generalized the previous setting by introducing *enhanced $\mathcal{H}$-consistency bounds* based on refined inequalities between surrogate and target conditional regrets. This produced distribution-dependent exponents of the form $1/(s-\alpha(s-1))$ under the Tsybakov noise assumption across binary and multi-class classification. Nevertheless, the Tsybakov noise condition constrains only the minimal margin $\gamma(x)$ and is thus purely distributional.

This work: model-dependent low-noise conditions. Our results strengthen the above framework by replacing the minimal margin γ with the *model margin* μ, which depends on both the data distribution and the chosen hypothesis. This yields the *Model Margin (MM) noise* assumption, a *weaker*, hypothesis-dependent condition that is implied by the classical Tsybakov noise assumption but remains valid for a broader range of models and distributions, as it measures

noise relative to each hypothesis rather than the minimal margin alone. Under MM noise, $\mathcal{H}$-consistency bounds preserve the same exponent as in the Tsybakov case. We further establish the theoretical robustness of the MM noise condition by demonstrating its monotonicity with respect to hypothesis class inclusion and its invariance under monotone score transformations.

In summary, MM noise extends enhanced $\mathcal{H}$-consistency bounds to a weaker yet more flexible, hypothesis-dependent setting, allowing more adaptive generalization guarantees with the same favorable exponents to be established for both binary and multi-class surrogate losses.

3 Preliminaries

Learning setup and notation. We consider a supervised learning problem with an unknown distribution $\mathcal{D}$ over pairs $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the label space. A hypothesis h is selected from a hypothesis set $\mathcal{H} \subseteq \mathcal{H}_{\text{all}} \coloneqq \{h \colon \mathcal{X} \to \mathcal{Y}_p \mid h \text{ measurable}\}$, where $\mathcal{Y}_p$ denotes the prediction space that specifies the form of model outputs. For instance, $\mathcal{Y}_p = \mathbb{R}$ for scalar scores in binary classification, and $\mathcal{Y}_p = \mathbb{R}^n$ for vector-valued scores in multi-class classification, where $n \in \mathbb{Z}_+$ is the number of labels.

A loss function $\ell \colon \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ measures the prediction error. Its generalization error and the best-in-class generalization error within $\mathcal{H}$ are defined as

$$\mathcal{E}_\ell(h) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h,x,y)], \qquad \mathcal{E}^*_\ell(\mathcal{H}) \coloneqq \inf_{h\in\mathcal{H}} \mathcal{E}_\ell(h).$$

Conditional errors. For every input $x \in \mathcal{X}$, we define the *conditional error* and *best-in-class conditional error* as

$$\mathcal{C}_\ell(h,x) \coloneqq \mathop{\mathbb{E}}_{y|x}[\ell(h,x,y)], \quad \mathcal{C}^*_\ell(\mathcal{H},x) = \inf_{h\in\mathcal{H}} \mathcal{C}_\ell(h,x).$$

The generalization error can be rewritten as $\mathcal{E}_\ell(h) = \mathbb{E}_X[\mathcal{C}_\ell(h,x)]$. We also define the *conditional regret*, (how suboptimal h is at a single point x) and *estimation error* as:

$$\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x) \coloneqq \mathcal{C}_\ell(h,x) - \mathcal{C}^*_\ell(\mathcal{H},x), \quad \mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}).$$

The *minimizability gap* (a technical term capturing how well $\mathcal{H}$ can optimize the loss pointwise) is defined as

$$\mathcal{M}_\ell(\mathcal{H}) \coloneqq \mathcal{E}^*_\ell(\mathcal{H}) - \mathop{\mathbb{E}}_X[\mathcal{C}^*_\ell(\mathcal{H},x)] \geq 0.$$

When $\mathcal{H}$ is sufficiently rich (for example, $\mathcal{H} = \mathcal{H}_{\text{all}}$ or $\mathcal{E}^*_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}_{\text{all}})$), this gap vanishes. In general, it is non-zero and can be upper bounded by the approximation error $\mathcal{E}^*_\ell(\mathcal{H}) - \mathcal{E}^*_\ell(\mathcal{H}_{\text{all}})$ (Mao, Mohri, and Zhong 2024h).

$\mathcal{H}$-consistency bounds. Let ℓ_1 be a surrogate loss and ℓ_2 the target loss. An *$\mathcal{H}$-consistency bound* relates their estimation errors via a non-asymptotic bound:

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma\Big(\mathcal{E}_{\ell_1}(h) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})\Big),$$

for a non-decreasing concave function Γ with $\Gamma(0) = 0$ (Mao, Mohri, and Zhong 2023a). This guarantee shows that reducing the surrogate estimation error implies a proportional reduction in the target error within the same hypothesis set.

Enhanced $\mathcal{H}$-consistency bound. An enhanced form introduces a multiplicative hypothesis-dependent factor $\gamma(h)$:

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma\Big(\gamma(h)\Big(\mathcal{E}_{\ell_1}(h) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})\Big)\Big).$$

In our analysis, Γ will be expressed as the power function $\Gamma(u) = u^{1/s}$ with $s \geq 1$, and $\gamma(h)$ will be expressed in terms of the model margin via the term $\mathbb{E}_X\big[1_{\mu(h,X)>0}\big]^{\frac{1}{t}}$, where t is the conjugate of s (i.e., $\frac{1}{s} + \frac{1}{t} = 1$).

Pointwise bounds. An $\mathcal{H}$-consistency bound typically follows from a pointwise bound relating the conditional regrets:

$$\Delta\mathcal{C}_{\ell_2,\mathcal{H}}(h,x) \leq \Gamma(\Delta\mathcal{C}_{\ell_1,\mathcal{H}}(h,x)).$$

Taking expectation over X and applying Jensen's inequality yields a $\mathcal{H}$-consistency bound.

Binary classification. For binary classification, the label space is $\mathcal{Y} = \{-1,+1\}$ and the prediction space is $\mathcal{Y}_p = \mathbb{R}$. The target loss is the binary zero–one loss:

$$\ell^{\text{bi}}_{0-1}(h,x,y) = 1_{\text{sign}(h(x))\neq y}, \qquad \text{sign}(t) = \begin{cases} +1, & t \geq 0, \\ -1, & t < 0. \end{cases}$$

Let $\mathsf{h}(x) = \text{sign}(h(x))$ be the hypothesis's prediction and $\eta(x) = (Y = +1 \mid X = x)$ be the conditional probability of $Y = +1$ given $X = x$. For any loss function ℓ, the conditional error can be written as

$$\mathcal{C}_\ell(h,x) = \eta(x)\ell(h,x,+1) + (1-\eta(x))\ell(h,x,-1).$$

Typical surrogates include margin-based losses $\ell_\Phi(h,x,y) = \Phi(yh(x))$, with Φ convex, non-increasing, and nonnegative.

Multi-class classification. In the multi-class setting, $\mathcal{Y} = [n] = \{1,\ldots,n\}$ and $\mathcal{Y}_p = \mathbb{R}^n$. The scaler $h(x,y)$ denotes the score assigned to label y, and the predicted label is $\mathsf{h}(x) = \text{argmax}_{y\in\mathcal{Y}}\, h(x,y)$ (with a fixed deterministic tie-breaking rule). The target loss is the multi-class zero–one loss

$$\ell_{0-1}(h,x,y) = 1_{\mathsf{h}(x)\neq y}.$$

Let $\mathbb{P}(y \mid x)$ be conditional probability of y given x. The conditional error is

$$\mathcal{C}_\ell(h,x) = \sum_{y\in\mathcal{Y}} \mathbb{P}(y \mid x)\ell(h,x,y).$$

Common surrogate families include max losses (Crammer and Singer 2001), constrained losses (Lee, Lin, and Wahba 2004), and comp–sum losses (Mao, Mohri, and Zhong 2023a). The following result from Awasthi et al. (2022b) characterizes the conditional regret of the multi-class zero-one loss.

Lemma 1. *For every input $x \in \mathcal{X}$,*

$$\mathcal{C}^*_{\ell_{0-1}}(\mathcal{H},x) = 1 - \max_{y\in\{\mathsf{h}(x):h\in\mathcal{H}\}} \mathbb{P}(y \mid x),$$

$$\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) = \max_{y\in\{\mathsf{h}(x):h\in\mathcal{H}\}} \mathbb{P}(y \mid x) - \mathbb{P}(\mathsf{h}(x) \mid x).$$

4 Model Margin (MM) Noise Assumption

This section introduces the Model Margin (MM) noise condition, a hypothesis-dependent low-noise assumption. In contrast to classical Tsybakov noise assumption, the proposed condition depends on a predictor h via its *model margin*.

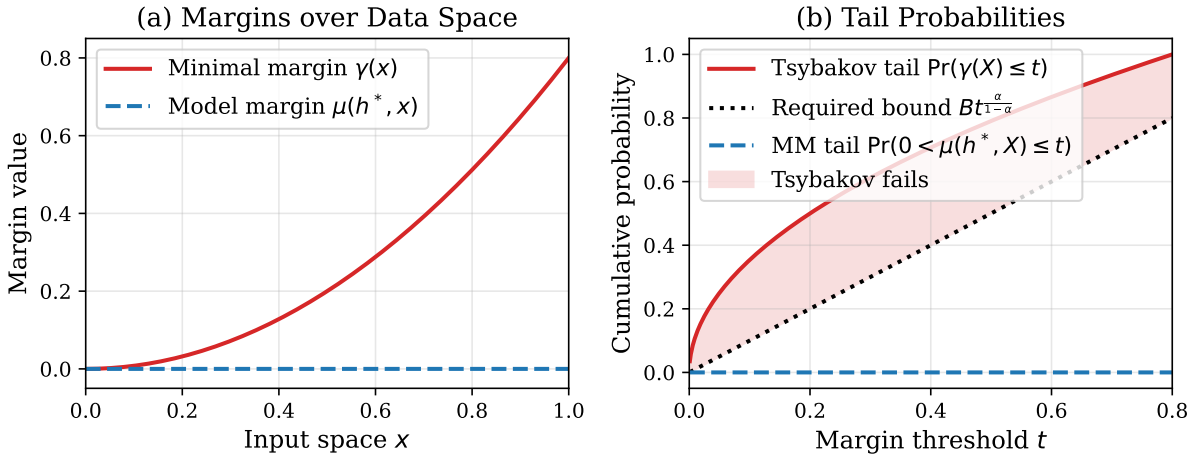

Figure 1: **MM noise holds while Tsybakov noise fails.** We visually illustrate the explicit example from Theorem 4, where $X \sim \mathrm{Unif}[0,1]$ and $\eta(x) = \frac{1}{2} + cx^\beta$ (plotted with parameters $c = 0.4, \beta = 2$). **(a)** The distributional minimal margin $\gamma(x) = 2cx^\beta$ measures intrinsic label ambiguity and is very small near the boundary ($x = 0$). In contrast, for the restricted class $\mathcal{H} = \{h^*\}$ where $h^*(x) \equiv +1$, the model margin $\mu(h^*, x)$ perfectly matches the Bayes optimal prediction and is identically zero everywhere. **(b)** The noise conditions require the probability of observing small margins to fall below a polynomial bound $Bt^{\frac{\alpha}{1-\alpha}}$ (black dotted line, plotted for $\alpha = 0.5, B = 1$). Because $\gamma(x)$ is small for a large portion of the input space, its cumulative probability tail $\mathbb{P}[\gamma(X) \le t]$ (red solid curve) decays too slowly and severely violates the required upper bound (shaded red area). Conversely, since $\mu(h^*, x) = 0$, the probability of the model making an incorrect prediction with a small margin is zero. Thus, the MM tail $\mathbb{P}[0 < \mu(h^*, X) \le t] = 0$ (blue dashed curve) trivially satisfies the bound. This visually demonstrates that MM noise provides a more flexible assumption by safely disregarding purely distributional minimal margins when models already predict optimally.

4.1 Definition and comparison with Tsybakov noise

Let $h^* \in \mathcal{H}_{\mathrm{all}}$ be a Bayes classifier. By (Mao, Mohri, and Zhong 2024h, Lemma 2.1), there exists indeed a measurable function h^* and for all $x \in \mathcal{X}$, it satisfies $\mathbb{P}(\mathsf{h}^*(x) \mid x) = \max_{y \in \mathcal{Y}} \mathbb{P}(y \mid x)$. For each input x, define the *model margin*

$$\mu(h, x) \coloneqq \mathbb{P}(\mathsf{h}^*(x) \mid x) - \mathbb{P}(\mathsf{h}(x) \mid x) \ge 0.$$

Note that $\mu(h, x) = \Delta\mathcal{C}_{\ell_{0-1}, \mathcal{H}_{\mathrm{all}}}(h, x)$ by Lemma 1.

Definition 2 (*Model Margin (MM) noise*)**.** *There exist $B > 0$ and $\alpha \in [0, 1]$ such that for all $h \in \mathcal{H}$ and all $t > 0$,*

$$\mathbb{P}[0 < \mu(h, X) \le t] \le Bt^{\frac{\alpha}{1-\alpha}}.$$

Intuitively, this condition requires that for each hypothesis h, the probability mass of points where h disagrees with the Bayes classifier but with a small model-dependent margin is controlled. It is therefore a *weaker* assumption than Tsybakov noise (Mammen and Tsybakov 1999).

Definition 3 (Tsybakov noise (Mao, Mohri, and Zhong 2025a))**.** *There exist $B > 0$ and $\alpha \in [0, 1]$ such that*

$$\forall t > 0, \quad \mathbb{P}[\gamma(X) \le t] \le Bt^{\frac{\alpha}{1-\alpha}}.$$

where $\gamma(x) = \mathbb{P}(y_{\max} \mid x) - \sup_{y \ne y_{\max}} \mathbb{P}(y \mid x)$ with $y_{\max} = \operatorname{argmax}_{y \in \mathcal{Y}} \mathbb{P}(y \mid x)$ is the minimal margin for a point $x \in \mathcal{X}$.

The Tsybakov condition is strong because it is independent of any specific hypothesis. However, this is also its main limitation. It can be overly pessimistic if the regions where the minimal margin $\gamma(x)$ is small are regions where the classifiers in $\mathcal{H}$ already perform well (i.e., they predict the Bayes label $\mathsf{h}^*(x)$). Our MM noise condition, $\mu(h, x)$, is designed to handle exactly this scenario, as $\mu(h, x)$ becomes zero for such correct predictions, effectively ignoring the small minimal margin.

Tsybakov ⇒ MM (MM is weaker). Let $\eta_1(x) = \mathbb{P}(y_{\max} \mid x) = \mathbb{P}(\mathsf{h}^*(x) \mid x)$ be the probability of the Bayes-optimal label and $\eta_2(x) = \max_{y \ne \mathsf{h}^*(x)} \mathbb{P}(y \mid x)$ be the probability of the most likely incorrect label. The minimal margin is $\gamma(x) = \eta_1(x) - \eta_2(x)$. On any x where $\mathsf{h}(x) \ne \mathsf{h}^*(x)$, we have $\mathbb{P}(\mathsf{h}(x) \mid x) \le \eta_2(x)$. Thus,

$$\begin{aligned} \mu(h, x) &= \mathbb{P}(\mathsf{h}^*(x) \mid x) - \mathbb{P}(\mathsf{h}(x) \mid x) \\ &\ge \eta_1(x) - \eta_2(x) \\ &= \gamma(x). \end{aligned}$$

Since $\mu(h, x) = 0$ when $\mathsf{h}(x) = \mathsf{h}^*(x)$, the region of disagreement is $\{x : \mu(h, x) > 0\}$. This implies that $\mu(h, x) > 0$ only if $\mathsf{h}(x) \ne \mathsf{h}^*(x)$, in which case $\mu(h, x) \ge \gamma(x)$. Therefore, for all $t > 0$,

$$\begin{aligned} \{0 < \mu(h, X) \le t\} &\subseteq \{\gamma(X) \le t\} \\ \mathbb{P}[0 < \mu(h, X) \le t] &\le \mathbb{P}[\gamma(X) \le t]. \end{aligned}$$

Any Tsybakov noise tail bound on γ immediately yields the MM noise tail bound for all $h \in \mathcal{H}$.

Theorem 4 (MM $\not\Rightarrow$ Tsybakov)**.** *There exist a distribution $\mathcal{D}$ and a hypothesis set $\mathcal{H}$ such that the MM noise condition holds while the Tsybakov noise condition fails.*

Proof. **Explicit example (see Figure 1 for illustration).** Let $\mathcal{X} = [0,1]$ with $X \sim \mathrm{Unif}[0,1]$, $\mathcal{Y} = \{-1,+1\}$, and define

$$\eta(x) \coloneqq \mathbb{P}(Y = +1 \mid X = x) = \tfrac{1}{2} + cx^{\beta}, \quad c \in \left(0, \tfrac{1}{2}\right), \beta > 0.$$

Then one Bayes classifier is $h^*(x) \equiv +1$ and the minimal margin is $\gamma(x) = |2\eta(x) - 1| = 2cx^{\beta}$. For any $t > 0$,

$$\mathbb{P}[\gamma(X) \le t] = \mathbb{P}\left(X \le \left(\tfrac{t}{2c}\right)^{1/\beta}\right) = \left(\tfrac{t}{2c}\right)^{1/\beta}.$$

Fix a target Tsybakov exponent $\alpha \in [0,1)$ and choose $\beta > \frac{1-\alpha}{\alpha}$ (e.g., for $\alpha = \frac{1}{2}$ take $\beta > 1$). Then $t^{1/\beta}$ decays more slowly than $t^{\alpha/(1-\alpha)}$, so the Tsybakov noise condition fails:

$$\mathbb{P}[\gamma(X) \le t] \not\le B\, t^{\alpha/(1-\alpha)}$$
$$\text{for any } B > 0 \text{ and sufficiently small } t.$$

Now let $\mathcal{H} \coloneqq \{h^*\}$ be the singleton hypothesis set. For $h = h^*$, $\mu(h,x) \equiv 0$, hence

$$\mathbb{P}[0 < \mu(h^*, X) \le t] = 0 \quad \text{for all } t > 0,$$

which trivially satisfies the MM noise tail bound for any α and B. Therefore, MM noise condition holds while Tsybakov noise condition fails for the same α and B. □

4.2 Key property

Next, we establish a key property of the MM noise condition, which provides an inequality bounding the disagreement mass $\mathbb{E}\big[1_{\mu(h,X)>0}\big]$ by a power of the 0–1 excess error.

Lemma 5 (Disagreement mass vs. 0–1 excess error)**.** *Under MM noise, there exists $c = \frac{B^{1-\alpha}}{\alpha^{\alpha}} > 0$ such that for all $h \in \mathcal{H}$,*

$$\begin{aligned}\mathbb{E}\big[1_{\mu(h,X)>0}\big] &\le c\,\mathbb{E}[\mu(h,X)1_{\mu(h,X)>0}]^{\alpha} \\ &= c(\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}_{\ell_{0-1}}(h^*))^{\alpha}.\end{aligned}$$

Proof. By definition of the expectation and the Lebesgue integral, for any $u > 0$,

$$\begin{aligned}&\mathbb{E}\big[\mu(h,X)1_{\mu(h,X)>0}\big] \\ &= \int_0^{+\infty} \mathbb{P}\big[\mu(h,X)1_{\mu(h,X)>0} > t\big]\,dt \\ &\ge \int_0^{u} \mathbb{P}[\mu(h,X) > t]\,dt.\end{aligned}$$

Since the equality $\mathbb{P}[\mu(h,X) > t] = \mathbb{P}[\mu(h,X) > 0] - \mathbb{P}[0 < \mu(h,X) \le t]$ holds, and by the MM noise assumption $\mathbb{P}[0 < \mu(h,X) \le t] \le Bt^{\frac{\alpha}{1-\alpha}}$, we have

$$\begin{aligned}&\mathbb{E}\big[\mu(h,X)1_{\mu(h,X)>0}\big] \\ &\ge \int_0^{u} \left(\mathbb{P}[\mu(h,X) > 0] - Bt^{\frac{\alpha}{1-\alpha}}\right)dt \\ &= u\,\mathbb{P}[\mu(h,X) > 0] - B(1-\alpha)u^{\frac{1}{1-\alpha}}.\end{aligned}$$

Maximizing the right-hand side over $u > 0$ gives

$$u^* = \left(\frac{\mathbb{P}[\mu(h,X) > 0]}{B}\right)^{\frac{1-\alpha}{\alpha}},$$

$$\mathbb{E}\big[\mu(h,X)1_{\mu(h,X)>0}\big] \ge \alpha B^{-\frac{1-\alpha}{\alpha}}\,\mathbb{P}[\mu(h,X) > 0]^{\frac{1}{\alpha}}.$$

Rearranging,

$$\mathbb{P}[\mu(h,X) > 0] \le \frac{B^{1-\alpha}}{\alpha^{\alpha}}\left(\mathbb{E}\big[\mu(h,X)1_{\mu(h,X)>0}\big]\right)^{\alpha}.$$

By Lemma 1,

$$\begin{aligned}\mathbb{E}\big[\mu(h,X)1_{\mu(h,X)>0}\big] &= \mathbb{E}[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}_{\mathrm{all}}}(h,x)] \\ &= \mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}_{\ell_{0-1}}(h^*).\end{aligned}$$

Thus the claim holds with $c = \frac{B^{1-\alpha}}{\alpha^{\alpha}}$. □

The left-hand side of Lemma 5 quantifies the disagreement mass between the hypothesis h and a Bayes classifier. The right-hand side establishes that this probability is bounded by a power of the 0–1 excess error. This property is central to the derivation of our $\mathcal{H}$-consistency bounds.

5 $\mathcal{H}$-Consistency Bounds under MM Noise

We now derive the main $\mathcal{H}$-consistency bounds under the MM noise assumption, first for the multi-class setting and then for the binary case.

5.1 Multi-class classification

The following result provides the $\mathcal{H}$-consistency bound based on the notion of $\mu(h,x)$.

Theorem 6. *Suppose there exists $s \ge 1$ with conjugate number $t \ge 1$, that is $\frac{1}{s} + \frac{1}{t} = 1$, such that*

$$\forall (h,x) \in \mathcal{H} \times \mathcal{X}, \quad \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) \le (\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x))^{\frac{1}{s}}.$$

Then, for every $h \in \mathcal{H}$,

$$\begin{aligned}&\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \\ &\quad \le \mathop{\mathbb{E}}_{X}\big[1_{\mu(h,X)>0}\big]^{\frac{1}{t}} \left(\mathcal{E}_{\ell}(h) - \mathcal{E}^*_{\ell}(\mathcal{H}) + \mathcal{M}_{\ell}(\mathcal{H})\right)^{\frac{1}{s}}.\end{aligned}$$

Proof. The total 0-1 estimation error is the expectation of the conditional regret:

$$\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) = \mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,X)].$$

From Lemma 1 and the definition of $\mu(h,x)$, we have $\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) \le \mu(h,x)$. This implies that if $\mu(h,x) = 0$, then $\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) = 0$. We can therefore write

$$\begin{aligned}\mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,X)] &= \mathop{\mathbb{E}}_{X}\big[\Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,X)\cdot 1_{\mu(h,X)>0}\big] \\ &\le \mathop{\mathbb{E}}_{X}\left[(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,X))^{\frac{1}{s}}\cdot 1_{\mu(h,X)>0}\right],\end{aligned}$$

where the inequality uses the theorem's pointwise assumption. We now apply Hölder's inequality with conjugate exponents $s \ge 1$ and $t \ge 1$ such that $\frac{1}{s} + \frac{1}{t} = 1$:

$$\begin{aligned}&\mathop{\mathbb{E}}_{X}\left[(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,X))^{\frac{1}{s}}\cdot 1_{\mu(h,X)>0}\right] \\ &\le \left(\mathop{\mathbb{E}}_{X}\left[\left((\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,X))^{\frac{1}{s}}\right)^{s}\right]\right)^{\frac{1}{s}}\left(\mathop{\mathbb{E}}_{X}\left[\left(1_{\mu(h,X)>0}\right)^{t}\right]\right)^{\frac{1}{t}} \\ &= \left(\mathop{\mathbb{E}}_{X}[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,X)]\right)^{\frac{1}{s}}\cdot\left(\mathop{\mathbb{E}}_{X}\big[1_{\mu(h,X)>0}\big]\right)^{\frac{1}{t}}\end{aligned}$$

since $1^t = 1$. Substituting the definition of the surrogate estimation error, $\mathbb{E}_X[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,X)] = \mathcal{E}_{\ell}(h) - \mathcal{E}^*_{\ell}(\mathcal{H}) + \mathcal{M}_{\ell}(\mathcal{H})$, yields the claimed bound. □

Theorem 6 provides a strong theoretical guarantee. Specifically, it shows that if the functional form of $\Gamma(x) = x^{\frac{1}{s}}$ is applied for standard $\mathcal{H}$-consistency bounds, then, the constant can be improved from 1 to a more refined, hypothesis-dependent quantity: $\mathbb{E}_X\big[1_{\mu(h,X)>0}\big]^{\frac{1}{t}} \leq 1$. This refinement applies to all existing $\mathcal{H}$-consistency bounds in (Awasthi et al. 2022b; Mao, Mohri, and Zhong 2023a, 2024h). In particular, for many cases where the surrogate loss ℓ is smooth, we have $t = s = \frac{1}{2}$, as shown by (Mao, Mohri, and Zhong 2024h). Next, we assume that the MM noise assumption holds and that the approximation error $\mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) = \mathcal{E}_{\ell_{0-1}}(h^*) = \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}_{\text{all}}) = 0$. This also implies that the minimizability gap $\mathcal{M}_{\ell_{0-1}}(\mathcal{H}) = 0$. The following result provides the corresponding $\mathcal{H}$-consistency bound in this case.

Theorem 7. *Suppose $\mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) = \mathcal{E}_{\ell_{0-1}}(h^*)$ and there exists $s \geq 1$ with conjugate number $t \geq 1$, that is $\frac{1}{s} + \frac{1}{t} = 1$, such that*

$$\forall (h,x) \in \mathcal{H} \times \mathcal{X}, \quad \Delta\mathcal{C}_{\ell_{0-1},\mathcal{H}}(h,x) \leq \big(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)\big)^{\frac{1}{s}}.$$

Then, under MM noise, there exists $c > 0$ such that for all $h \in \mathcal{H}$,

$$\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) \leq c^{\frac{s-1}{s-\alpha(s-1)}} \big[\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\big]^{\frac{1}{s-\alpha(s-1)}}.$$

Proof. We start from Theorem 6. Let $A = \mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H})$ and $B = \mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$. By Lemma 5 and the assumption $\mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) = \mathcal{E}_{\ell_{0-1}}(h^*)$, there exists $c > 0$ with $\mathbb{E}_X\big[1_{\mu(h,X)>0}\big] \leq cA^\alpha$. Using $\mathcal{M}_{\ell_{0-1}}(\mathcal{H}) = 0$, Theorem 6 gives $A \leq (cA^\alpha)^{\frac{1}{t}} B^{\frac{1}{s}} = c^{\frac{1}{t}} A^{\frac{\alpha}{t}} B^{\frac{1}{s}}$. Rearranging yields $A^{1-\alpha/t} \leq c^{\frac{1}{t}} B^{\frac{1}{s}}$. Using $\frac{1}{t} = \frac{s-1}{s}$, the exponent is $1 - \alpha/t = \frac{s-\alpha(s-1)}{s}$. Raising both sides to the power of $\frac{1}{1-\alpha/t} = \frac{s}{s-\alpha(s-1)}$ gives:

$$A \leq \left(c^{\frac{s-1}{s}} B^{\frac{1}{s}}\right)^{\frac{s}{s-\alpha(s-1)}} = c^{\frac{s-1}{s-\alpha(s-1)}} B^{\frac{1}{s-\alpha(s-1)}},$$

which is the claimed bound. □

The exponent $\frac{1}{s-\alpha(s-1)}$ matches that derived from Tsybakov noise condition (Mao, Mohri, and Zhong 2025a, Theorem 9), but our bounds are established under the weaker MM noise assumption. For smooth binary surrogates ($s = 2$), this result demonstrates the interpolation between a square-root rate (as $\alpha \to 0$) and a linear rate (as $\alpha \to 1$).

5.2 Binary classification

The binary analogue, which can be viewed as a special case of the multi-class setting, satisfies the same bound via an identical proof.

Theorem 8. *Suppose there exists $s \geq 1$ with conjugate number $t \geq 1$, that is $\frac{1}{s} + \frac{1}{t} = 1$, such that*

$$\forall (h,x) \in \mathcal{H} \times \mathcal{X}, \quad \Delta\mathcal{C}_{\ell^{\text{bi}}_{0-1},\mathcal{H}}(h,x) \leq \big(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h,x)\big)^{\frac{1}{s}}.$$

Then, for every $h \in \mathcal{H}$,

$$\mathcal{E}_{\ell^{\text{bi}}_{0-1}}(h) - \mathcal{E}^*_{\ell^{\text{bi}}_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell^{\text{bi}}_{0-1}}(\mathcal{H}) \leq \mathbb{E}_X\big[1_{\mu(h,X)>0}\big]^{\frac{1}{t}} \big(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\big)^{\frac{1}{s}}.$$

Under MM noise and $\mathcal{M}_{\ell^{\text{bi}}_{0-1}}(\mathcal{H}) = 0$, there exists $c > 0$,

$$\forall h \in \mathcal{H}, \ \mathcal{E}_{\ell^{\text{bi}}_{0-1}}(h) - \mathcal{E}^*_{\ell^{\text{bi}}_{0-1}}(\mathcal{H}) \leq c^{\frac{s-1}{s-\alpha(s-1)}} \big(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\big)^{\frac{1}{s-\alpha(s-1)}}.$$

Proof. The proofs are identical to that of Theorem 6 and Theorem 7, replacing the multi-class zero-one loss ℓ_{0-1} with the binary zero-one loss ℓ^{bi}_{0-1}. □

These results are analogous to the enhanced $\mathcal{H}$-consistency bounds established under Tsybakov noise by Mao, Mohri, and Zhong (2025a), but with the key distinction that the MM condition is *weaker* and hypothesis-dependent. This shift also yields *predictor-dependent* constants, quantified by $\mathbb{E}\big[1_{\mu(h,X)>0}\big]^{1/t}$, which are absent from purely distributional Tsybakov analyses and align better with model selection. Crucially, we retain the desirable noise exponent $1/(s - \alpha(s-1))$ derived under Tsybakov noise. Consequently, practitioners gain bounds with the same favorable rates but under a less restrictive, model-aware assumption.

6 Applications

Structural properties. The MM noise condition exhibits two key structural properties: *(i) Monotonicity in $\mathcal{H}$.* If $\mathcal{H}_1 \subseteq \mathcal{H}_2$ and the MM noise condition holds uniformly over $\mathcal{H}_2$, then it also holds over $\mathcal{H}_1$ with the same parameters. *(ii) Invariance.* If the scores $h(x,\cdot)$ are replaced by $\psi \circ h(x,\cdot)$ for any strictly increasing function ψ, the resulting predictions h and the Bayes prediction h^* remain unchanged. Consequently, the model margin $\mu(h,x)$ and the MM noise condition are also invariant.

These properties demonstrate the flexibility of the MM noise condition, highlighting its robustness to the specific choice of hypothesis set and the scaling of the prediction scores.

Examples. We now instantiate the derived bounds for common surrogate loss families. Table 1 applies Theorem 8 to common binary margin-based losses in (Awasthi et al. 2022a). The results confirm the linear rate ($s = 1$) for the hinge loss and the $1/(2-\alpha)$ exponent for smooth surrogates ($s = 2$), which interpolates between the square-root and linear regimes based on the noise parameter α.

Loss	$\Phi(z)$	s	Bound under MM noise
Hinge	$[1-z]_+$	1	$\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H})$
Logistic	$\log(1+e^{-z})$	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Exponential	e^{-z}	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Sq-hinge	$[1-z]^2_+$	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$

Table 1: Binary margin-based losses of the form $\ell_\Phi(h,x,y) = \Phi(y\,h(x))$ with corresponding $\mathcal{H}$-consistency bounds.

Table 2 illustrates the bounds for comp-sum losses in (Mao, Mohri, and Zhong 2023a), a broad family that generalizes sum-type and cross-entropy-like surrogates. Under MM noise, the same exponents as in the enhanced $\mathcal{H}$-consistency bounds of Mao, Mohri, and Zhong (2025a) are recovered, but they hold under the weaker MM noise assumptions.

Loss	$\ell(h, x, y)$	s	Bound under MM noise
Mean absolute error (MAE)	$1 - p_\theta(y \mid x)$	1	$\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H})$
Cross-entropy (Logistic)	$-\log(p_\theta(y \mid x))$	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Exponential comp-sum	$\frac{1}{p_\theta(y\mid x)} - 1$	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Generalized cross-entropy	$\frac{1 - p_\theta(y \mid x)^{q-1}}{q-1}$	2	$c^{\frac{1}{2-\alpha}}[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H})]^{\frac{1}{2-\alpha}}$

Table 2: Multi-class comp-sum losses. Here $p_\theta(y \mid x) = \exp(h(x, y))\big/\sum_{y'} \exp(h(x, y'))$ is the softmax model.

7 Conclusion

We introduced the MM noise condition, a *weaker yet more flexible* hypothesis-dependent noise model than Tsybakov noise, and derived enhanced $\mathcal{H}$-consistency bounds for both binary and multi-class classification. These results broaden the applicability of existing bounds and provide adaptive guarantees under hypothesis-dependent low-noise conditions. Future work includes extending MM noise analysis to regression settings and developing adaptive algorithms that leverage its flexibility in practice.

References

Awasthi, P.; Frank, N.; Mao, A.; Mohri, M.; and Zhong, Y. 2021a. Calibration and consistency of adversarial surrogate losses. In *Advances in Neural Information Processing Systems*, 9804–9815.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2021b. A finer calibration analysis for adversarial robustness. *arXiv preprint arXiv:2105.01550*.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2022a. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, 1117–1174.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2022b. Multi-Class $\mathcal{H}$-Consistency Bounds. In *Advances in neural information processing systems*.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2023a. DC-Programming for Neural Network Optimizations. *Journal of Global Optimization*.

Awasthi, P.; Mao, A.; Mohri, M.; and Zhong, Y. 2023b. Theoretically Grounded Loss Functions and Algorithms for Adversarial Robustness. In *International Conference on Artificial Intelligence and Statistics*, 10077–10094.

Bao, H. 2023. Proper Losses, Moduli of Convexity, and Surrogate Regret Bounds. In *Conference on Learning Theory*, 525–547.

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

Chen, D.-R.; Wu, Q.; Ying, Y.; and Zhou, D.-X. 2004. Support vector machine soft margin classifiers: error analysis. *The Journal of Machine Learning Research*, 5: 1143–1175.

Cortes, C.; Mao, A.; Mohri, C.; Mohri, M.; and Zhong, Y. 2024. Cardinality-Aware Set Prediction and Top-k Classification. In *Advances in Neural Information Processing Systems*.

Cortes, C.; Mao, A.; Mohri, M.; and Zhong, Y. 2025. Balancing the Scales: A Theoretical and Algorithmic Framework for Learning from Imbalanced Data. In *International Conference on Machine Learning*.

Cortes, C.; Mohri, M.; and Zhong, Y. 2025. Improved Balanced Classification with Theoretically Grounded Loss Functions. In *Advances in neural information processing systems*.

Crammer, K.; and Singer, Y. 2001. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of machine learning research*, 2(Dec): 265–292.

DeSalvo, G.; Mohri, C.; Mohri, M.; and Zhong, Y. 2025. Budgeted Multiple-Expert Deferral. *arXiv preprint arXiv:2510.26706*.

Finocchiaro, J.; Frongillo, R.; and Waggoner, B. 2019. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*.

Frongillo, R.; and Waggoner, B. 2021. Surrogate Regret Bounds for Polyhedral Losses. In *Advances in Neural Information Processing Systems*, 21569–21580.

Lee, Y.; Lin, Y.; and Wahba, G. 2004. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465): 67–81.

Long, P.; and Servedio, R. 2013. Consistency versus realizable H-consistency for multiclass classification. In *International Conference on Machine Learning*, 801–809.

Mahdavi, M.; Zhang, L.; and Jin, R. 2014. Binary excess risk for smooth convex surrogates. *arXiv preprint arXiv:1402.1792*.

Mammen, E.; and Tsybakov, A. B. 1999. Smooth Discrimination Analysis. *The Annals of Statistics*, 27(6): 1808–1829.

Mao, A. 2025. *Theory and Algorithms for Learning with Multi-Class Abstention and Multi-Expert Deferral*. Ph.D. thesis, New York University.

Mao, A.; Mohri, C.; Mohri, M.; and Zhong, Y. 2023. Two-Stage Learning to Defer with Multiple Experts. In *Advances in neural information processing systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023a. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. In *International Conference on Machine Learning*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023b. H-Consistency Bounds: Characterization and Extensions. In *Advances in Neural Information Processing Systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023c. H-Consistency Bounds for Pairwise Misranking Loss Surrogates. In *International conference on Machine learning*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023d. Ranking with Abstention. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*.

Mao, A.; Mohri, M.; and Zhong, Y. 2023e. Structured Prediction with Stronger Consistency Guarantees. In *Advances in Neural Information Processing Systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024a. H-Consistency Guarantees for Regression. In *International Conference on Machine Learning*, 34712–34737.

Mao, A.; Mohri, M.; and Zhong, Y. 2024b. Multi-Label Learning with Stronger Consistency Guarantees. In *Advances in neural information processing systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024c. Predictor-Rejector Multi-Class Abstention: Theoretical Analysis and Algorithms. In *International Conference on Algorithmic Learning Theory*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024d. Principled Approaches for Learning to Defer with Multiple Experts. In *International Symposium on Artificial Intelligence and Mathematics*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024e. Realizable H-Consistent and Bayes-Consistent Loss Functions for Learning to Defer. In *Advances in neural information processing systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2024f. Regression with Multi-Expert Deferral. In *International Conference on Machine Learning*, 34738–34759.

Mao, A.; Mohri, M.; and Zhong, Y. 2024g. Theoretically grounded loss functions and algorithms for score-based multi-class abstention. In *International Conference on Artificial Intelligence and Statistics*, 4753–4761.

Mao, A.; Mohri, M.; and Zhong, Y. 2024h. A Universal Growth Rate for Learning with Smooth Surrogate Losses. In *Advances in Neural Information Processing Systems*.

Mao, A.; Mohri, M.; and Zhong, Y. 2025a. Enhanced H-Consistency Bounds. In *International Conference on Algorithmic Learning Theory*.

Mao, A.; Mohri, M.; and Zhong, Y. 2025b. Mastering Multiple-Expert Routing: Realizable H-Consistency and Strong Guarantees for Learning to Defer. In *International Conference on Machine Learning*.

Mao, A.; Mohri, M.; and Zhong, Y. 2025c. Principled Algorithms for Optimizing Generalized Metrics in Binary Classification. In *International Conference on Machine Learning*.

Mohri, C.; Andor, D.; Choi, E.; Collins, M.; Mao, A.; and Zhong, Y. 2024. Learning to Reject with a Fixed Predictor: Application to Decontextualization. In *International Conference on Learning Representations*.

Reid, M. D.; and Williamson, R. C. 2009. Surrogate regret bounds for proper losses. In *International Conference on Machine Learning*, 897–904.

Steinwart, I. 2007. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2): 225–287.

Tewari, A.; and Bartlett, P. L. 2007. On the Consistency of Multiclass Classification Methods. *Journal of Machine Learning Research*, 8(36): 1007–1025.

Weston, J.; and Watkins, C. 1998. Multi-class support vector machines. Technical report, Citeseer.

Weston, J.; and Watkins, C. 1999. Support vector machines for multi-class pattern recognition. *European Symposium on Artificial Neural Networks*, 4(6).

Zhang, M.; and Agarwal, S. 2020. Bayes Consistency vs. H-Consistency: The Interplay between Surrogate Loss Functions and the Scoring Function Class. In *Advances in Neural Information Processing Systems*, 16927–16936.

Zhang, T. 2004a. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct): 1225–1251.

Zhang, T. 2004b. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1): 56–85.

Zheng, C.; Wu, G.; Bao, F.; Cao, Y.; Li, C.; and Zhu, J. 2023. Revisiting Discriminative vs. Generative Classifiers: Theory and Implications. In *International Conference on Machine Learning*.

Zhong, Y. 2025. *Fundamental Novel Consistency Theory: H-Consistency Bounds*. Ph.D. thesis, New York University.