

A survey on \mathcal{H} -consistency Bounds

Edwin Liu

EDWINLIU@NYU.EDU

Yizheng Li

YL8517@NYU.EDU

John Quigley

JWQ4271@NYU.EDU

Abstract

Many learning algorithms optimize surrogate losses instead of the target loss of interest, such as the 0–1 loss in classification. Thus, a central theoretical question is how minimizing surrogate risk translates into guarantees on target risk. This paper surveys recent advances in H-consistency bounds, which provide quantitative, non-asymptotic guarantees relating surrogate excess risk to target excess risk over restricted hypothesis classes. We review the framework introduced by Awasthi et al. (2022) and subsequent refinements by Mao et al. (2024b), emphasizing the role of the tight H-estimation error transformation that characterizes optimal transfers between losses. We highlight a universal result showing that for a broad class of smooth, convex margin-based and multi-class comp-sum losses, the transformation exhibits quadratic growth near zero, implying local square-root bounds on 0–1 excess risk. Beyond classical bounds, we present the Enhanced H-Consistency Bounds (EHCB) framework introduced by Mao et al. (2024a), which allows instance- and hypothesis-dependent reweighting and yields strictly sharper guarantees, including improved convergence exponents under Tsybakov noise conditions. In addition to synthesizing existing theory, we contribute several small theoretical supplements, including a proof of the converse of a previous result for calibrated margin losses and the identification of a gap in a previously published theorem. Finally, we present original numerical experiments that empirically validate the theoretical bounds derived in Mao et al. (2024a) on logistic regression and ReLU neural networks, and suggest even faster practical behavior in low-noise regimes.

1. Introduction

Many modern learning algorithms typically optimize a surrogate loss rather than directly optimizing the target loss of interest. This is especially relevant in classification, where the target loss (e.g., the 0–1 loss) is non-smooth and non-differentiable, making direct optimization difficult. Therefore the central question is: what guarantees can we give on the target loss estimation error when we minimize the surrogate loss estimation error? A series of recent papers (Awasthi et al. (2022); Mao et al. (2024b,a)) give explicit bounds for this objective which massively improve on previous asymptotic results. Below we summarize their relevant result, perform some numerical experiments, and include several extended results for theorems in Mao et al. (2024a).

1.1. Notations

We consider binary classification with labels $Y \in \{-1, +1\}$ and instances $X \in \mathcal{X}$ drawn from a distribution \mathcal{D} . Let \mathcal{H} be a hypothesis set (function class). For a loss function $\ell : \mathcal{H} \times \mathcal{X} \times Y \rightarrow \mathbb{R}_+$, define the *noise function* $\eta(x) := \mathbb{P}[Y = 1 \mid X = x]$ and the *conditional ℓ -risk* $\mathcal{C}_\ell(h, x) := \eta(x) \ell(h, x, +1) + (1 - \eta(x)) \ell(h, x, -1)$. The *generalization error* is $\mathcal{E}_\ell(h) := \mathbb{E}_X[\mathcal{C}_\ell(h, x)]$, and the best-in class error over \mathcal{H} is $\mathcal{E}_\ell^*(\mathcal{H}) := \inf_{h \in \mathcal{H}} \mathcal{E}_\ell(h)$. We also use the conditional regret (gap) notation $\Delta \mathcal{C}_{\ell, \mathcal{H}}(h, x) := \mathcal{C}_\ell(h, x) - \inf_{h \in \mathcal{H}} \mathcal{C}_\ell(h, x)$, and the *minimizability gap* $\mathcal{M}_\ell(\mathcal{H}) := \mathcal{E}_\ell^*(\mathcal{H}) - \mathbb{E}_X[\inf_{h \in \mathcal{H}} \mathcal{C}_\ell(h, x)]$. A useful identity is $\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H}) = \mathbb{E}_X[\Delta \mathcal{C}_{\ell, \mathcal{H}}(h, x)]$. For margin-based surrogates, we write $\ell(h, x, y) = \Phi(yh(x))$ for some function Φ .

1.2. From Bayes-consistency to H-consistency to H-consistency bounds

The classical property that addresses our central question is *Bayes-consistency*, which compares losses in the idealized regime where the learner can range over all measurable predictors \mathcal{H}_{all} . A surrogate loss ℓ_1 is Bayes-consistent with respect to a target loss ℓ_2 if, for any distribution \mathcal{D} and any sequence $\{h_n\} \subset \mathcal{H}_{\text{all}}$, $\mathcal{E}_{\ell_1}(h_n) - \mathcal{E}_{\ell_1}^*(\mathcal{H}_{\text{all}}) \rightarrow 0 \Rightarrow \mathcal{E}_{\ell_2}(h_n) - \mathcal{E}_{\ell_2}^*(\mathcal{H}_{\text{all}}) \rightarrow 0$. In practice, however, learning is performed over a *restricted* hypothesis set \mathcal{H} (e.g., linear hypotheses). This motivates the notion of *H-consistency*, which requires the same implication but only for sequences $\{h_n\} \subset \mathcal{H}$ and relative to $\mathcal{E}_{\ell_1}^*(\mathcal{H})$. Going beyond these asymptotic notions, the key step in the H-consistency bounds framework is to derive *quantitative* guarantees that upper bound target estimation error in terms of surrogate estimation error (plus minimizability gaps) for approximate minimizers obtained from finite samples. In [Awasthi et al. \(2022\)](#), a class of such bounds is defined as follows.

Definition 1 (H-consistency bounds) *Given a hypothesis set \mathcal{H} , an H-consistency bound relating the loss function ℓ_1 to the loss function ℓ_2 is an inequality of the form*

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_{\ell_2}(h) - \mathcal{E}_{\ell_2}^*(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma\left(\mathcal{E}_{\ell_1}(h) - \mathcal{E}_{\ell_1}^*(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})\right),$$

that holds for any distribution \mathcal{D} , where $\Gamma : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is non-negative, non-decreasing, and concave. Note that Γ is invertible, so we can also consider an equivalent formulation where the function Γ^{-1} acts on the left hand side of the inequality.

1.3. The Error Transformation

We now specialize to the case where the target loss is the 0–1 loss. In this setting, the transfer can be captured by a single function \mathcal{T} , called the *H-estimation error transformation*. And in fact, \mathcal{T} is the optimal choice for Γ^{-1} in the sense that it is tight.

Definition 2 (H-estimation error transformation) *Let ℓ be a surrogate loss and let the target loss be ℓ_{0-1} . An H-estimation error transformation is a function $\mathcal{T} : [0, 1] \rightarrow \mathbb{R}_+$ such that the following holds tightly:*

$$\forall h \in \mathcal{H}, \quad \mathcal{T}\left(\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}_{\ell_{0-1}}^*(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H})\right) \leq \mathcal{E}_{\ell}(h) - \mathcal{E}_{\ell}^*(\mathcal{H}) + \mathcal{M}_{\ell}(\mathcal{H}).$$

Here, tight means that for any $t \in [0, 1]$, there exist a distribution and some $h \in \mathcal{H}$ such that $\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}_{\ell_{0-1}}^(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) = t$, and $\mathcal{E}_{\ell}(h) - \mathcal{E}_{\ell}^*(\mathcal{H}) + \mathcal{M}_{\ell}(\mathcal{H}) = \mathcal{T}(t)$.*

We focus on *complete* hypothesis sets, i.e., $\forall x \in \mathcal{X}, \{h(x) \mid h \in \mathcal{H}\} = \mathbb{R}$. For margin-based surrogates $\ell(h, x, y) = \Phi(yh(x))$, define for $t \in [0, 1]$, $f_t(u) := \frac{1-t}{2}\Phi(u) + \frac{1+t}{2}\Phi(-u)$. In this case, the transformation admits the variational characterization

$$\mathcal{T}(t) := \inf_{u \geq 0} f_t(u) - \inf_{u \in \mathbb{R}} f_t(u), \quad t \in [0, 1],$$

and in particular $\mathcal{T}(0) = 0$. If further Φ is differentiable at 0 and $\Phi'(0) < 0$, then for all $t \in [0, 1]$, we have $\inf_{u \geq 0} f_t(u) = f_t(0) \implies \mathcal{T}(t) = f_t(0) - \inf_{u \in \mathbb{R}} f_t(u)$.

The main idea of the proof of tightness is to consider, for any $t \in [0, 1]$ a degenerate distribution with support on a singleton where the noise on the singleton is $\frac{1}{2} + \frac{t}{2}$.

1.4. Examples: Transformation Table for Class of Linear Hypotheses

To build intuition, we can consider the \mathcal{T} transformations for the class of linear hypotheses with bounded parameters: $\mathcal{H}_{\text{lin}} = \{x \mapsto w \cdot x + b \mid \|w\|_q \leq W, |b| \leq B\}$. The paper [Mao et al. \(2024b\)](#) computed \mathcal{T} explicitly for a representative set of losses, shown in Table 1.

Surrogate	$\mathcal{T}_\Phi(t)$ for $t \in [0, 1]$
Hinge	$\min\{B, 1\} t$
Logistic	$\begin{cases} \frac{t+1}{2} \log_2(t+1) + \frac{1-t}{2} \log_2(1-t), & t \leq \frac{e^B - 1}{e^B + 1}, \\ 1 - \frac{t+1}{2} \log_2(1 + e^{-B}) - \frac{1-t}{2} \log_2(1 + e^B), & t > \frac{e^B - 1}{e^B + 1}, \end{cases}$
Exponential	$\begin{cases} 1 - \sqrt{1 - t^2}, & t \leq \frac{e^{2B} - 1}{e^{2B} + 1}, \\ 1 - \frac{t+1}{2} e^{-B} - \frac{1-t}{2} e^B, & t > \frac{e^{2B} - 1}{e^{2B} + 1}, \end{cases}$
Quadratic	$\begin{cases} t^2, & t \leq B, \\ 2Bt - B^2, & t > B, \end{cases}$
Sigmoid	$\tanh(kB) t$
ρ -Margin	$\frac{\min\{B, \rho\}}{\rho} t$

Table 1: Example \mathcal{H}_{lin} -estimation error transformations.

2. Universal Growth Rate of H-Consistency Bounds

2.1. Motivation

The \mathcal{H} -consistency bounds introduced in the previous section are driven by a concave transfer function Γ that converts (gap-corrected) surrogate estimation error into (gap-corrected) target estimation error. To better understand the sharpness and limitations of such bounds, we ask a finer question: *what is the asymptotic shape of these transfers as the error approaches 0?* In particular, when the surrogate excess risk becomes small, we are interested in whether the induced control on the 0-1 excess risk is linear, square-root, or follows some other local rate.

2.2. Main result: universal quadratic growth of \mathcal{T}

The universal growth-rate result in [Mao et al. \(2024b\)](#) states that for a broad class of smooth, convex margin-based losses, the transformation behaves quadratically near 0.

Theorem 3 (Universal growth for binary margin-based losses) *Let \mathcal{H} be a complete hypothesis set. Assume that Φ is convex, twice continuously differentiable, and satisfies $\Phi'(0) > 0$ and $\Phi''(0) > 0$. Then $\mathcal{T}(t) = \Theta(t^2)$; i.e., there exist constants $C > 0$, $c > 0$, and $T > 0$ such that*

$$ct^2 \leq \mathcal{T}(t) \leq Ct^2, \quad \forall 0 < t \leq T.$$

Proof sketch: Use the implicit function theorem on the first-order condition $f'_t(a_t^*) = 0$ to show a unique minimizer a_t^* exists, with $a_0^* = 0$ and $\left.\frac{da_t^*}{dt}\right|_{t=0} = \frac{\Phi'(0)}{\Phi''(0)} > 0$, hence $a_t^* = \Theta(t)$. Then represent $\mathcal{T}(t) = f_t(0) - \inf_u f_t(u)$ as $\mathcal{T}(t) = \int_0^{a_t^*} u f_t''(u) du$. By continuity and $\Phi''(0) > 0$, bound the second derivative on a small interval: $c \leq f_t''(u) \leq C$ for all $u \in [0, a_t^*]$. Then $\frac{c}{2}(a_t^*)^2 \leq \mathcal{T}(t) \leq \frac{C}{2}(a_t^*)^2$, and since $a_t^* = \Theta(t)$, this gives $\mathcal{T}(t) = \Theta(t^2)$.

This implies that the \mathcal{H} -consistency bounds are locally square-root: define the quantity of estimation loss plus minimizability gap as $V_\ell(h) := \mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$. Since $\mathcal{T}(t)$ lower-bounds the surrogate excess required to achieve target excess t , the relation $\mathcal{T}(V_{\ell_{0-1}}(h)) \leq V_\ell(h)$ combined with $\mathcal{T}(t) = \Theta(t^2)$ yields a local *square-root* control of $V_{\ell_{0-1}}(h)$ in terms of $V_\ell(h)$ (up to constants), for smooth losses.

$\Phi(u)$	surrogate loss $\ell(h, x, y)$	induced bound (illustrative)
e^{-u}	$e^{-y h(x)}$	$V_{\ell_{0-1}} \lesssim \sqrt{V_\ell}$
$\log(1 + e^{-u})$	$\log(1 + e^{-y h(x)})$	$V_{\ell_{0-1}} \lesssim \sqrt{V_\ell}$
$\max\{0, 1 - u\}^2$	$\max\{0, 1 - y h(x)\}^2$	$V_{\ell_{0-1}} \lesssim \sqrt{V_\ell}$
$\max\{0, 1 - u\}$	$\max\{0, 1 - y h(x)\}$	$V_{\ell_{0-1}} \lesssim V_\ell$

Table 2: Illustrative local forms of the induced \mathcal{H} -consistency bounds via $\mathcal{T}(t)$ (constants omitted).

2.3. Extension to multi-class comp-sum losses

The same quadratic growth phenomenon extends beyond binary margin losses to multi-class comp-sum losses such as cross-entropy loss.

Theorem 4 (Universal growth for multi-class comp-sum losses) *Assume that Φ is convex, twice continuously differentiable, and satisfies $\Phi'(u) < 0$ and $\Phi''(u) > 0$ for all $u \in (0, \frac{1}{2}]$. Then the associated \mathcal{H} -estimation error transformation satisfies $\mathcal{T}(t) = \Theta(t^2)$.*

3. Enhanced H-Consistency Bounds (EHCB)

3.1. Motivation: beyond a fixed function Γ

Previous \mathcal{H} -consistency bounds control the target estimation error by applying a *single* concave transfer function Γ to the surrogate estimation error. Enhanced H-consistency bounds ask whether this transfer can be sharpened by allowing *instance- and hypothesis-dependent* reweightings of the conditional regrets. Concretely, the EHCB framework introduces two positive functions $\alpha(h, x)$ and $\beta(h, x)$ that modulate the comparison between conditional surrogate and target regrets, yielding a more general bound with a *hypothesis-dependent* factor $\gamma(h)$.

3.2. A general enhanced Γ -bound with α, β and $\gamma(h)$

The core technical tool is the following enhanced bound: if a (reweighted) conditional target regret is controlled by a concave transform of a (reweighted) conditional surrogate regret, then one obtains a global \mathcal{H} -consistency bound with an explicit multiplicative factor $\gamma(h)$.

Theorem 5 (Enhanced Γ -bound) *Assume that there exist a concave function $\Gamma : \mathbb{R}_+ \rightarrow \mathbb{R}$ and two positive functions $\alpha : \mathcal{H} \times \mathcal{X} \rightarrow \mathbb{R}_+^*$ and $\beta : \mathcal{H} \times \mathcal{X} \rightarrow \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h, x) < +\infty$ and $\mathbb{E}_X[\beta(h, X)] < +\infty$ for all $h \in \mathcal{H}$, such that for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$,*

$$\frac{\Delta \mathcal{C}_{\ell_2, \mathcal{H}}(h, x) \mathbb{E}_X[\beta(h, X)]}{\beta(h, x)} \leq \Gamma(\alpha(h, x) \Delta \mathcal{C}_{\ell_1, \mathcal{H}}(h, x)).$$

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

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}_{\ell_2}^*(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma(\gamma(h)) (\mathcal{E}_{\ell_1}(h) - \mathcal{E}_{\ell_1}^*(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})), \quad (\text{EHCB-1})$$

with

$$\gamma(h) = \left[\frac{\sup_{x \in \mathcal{X}} \alpha(h, x) \beta(h, x)}{\mathbb{E}_X[\beta(h, X)]} \right].$$

If additionally $\mathcal{X} \subset \mathbb{R}^n$ and, for any $h \in \mathcal{H}$, the maps $x \mapsto \Delta\mathcal{C}_{\ell_1, \mathcal{H}}(h, x)$ and $x \mapsto \alpha(h, x)\beta(h, x)$ are monotone in opposite directions (or vice-versa), then (EHCB-1) holds with

$$\gamma(h) = \mathbb{E}_X \left[\frac{\alpha(h, X) \beta(h, X)}{\mathbb{E}_X[\beta(h, X)]} \right].$$

Interpretation. The functions $\alpha(h, x)$ and $\beta(h, x)$ can be viewed as *reweighting/conditioning* terms that allow the conditional comparison to adapt to where the distribution puts mass and where the hypothesis incurs regret. The resulting parameter $\gamma(h)$ acts like a hypothesis-dependent “condition number”: when $\gamma(h)$ is small, the global transfer from surrogate regret to target regret becomes strictly tighter than what one would obtain from a uniform (constant) comparison.

Proof sketch: Use the identity, $\mathcal{E}_{\ell}(h) - \mathcal{E}_{\ell}^*(\mathcal{H}) + \mathcal{M}_{\ell}(\mathcal{H}) = \mathbb{E}_X[\Delta\mathcal{C}_{\ell, \mathcal{H}}(h, x)]$, then the assumption, and then Jensen’s and Holder’s inequalities.

3.3. Low-noise refinement: Tsybakov noise condition and improved exponents

EHCB also yields stronger exponents under low-noise assumptions. We use the Tsybakov noise condition: there exist $B > 0$ and $\alpha \in [0, 1)$ such that

$$\forall t > 0, \quad \mathbb{P}[|\eta(X) - 1/2| \leq t] \leq B t^{\frac{\alpha}{1-\alpha}}.$$

As $\alpha \rightarrow 1$, this approaches the Massart (bounded noise) regime; when $\alpha = 0$, the condition is vacuous. Additionally, the above condition is equivalent to the existence of a constant $c > 0$ and $\alpha \in [0, 1)$ such that for all $h \in \mathcal{H}$, the following inequality holds:

$$\mathbb{E} [\mathbf{1}_{h(X) \neq h^*(X)}] \leq c \left[\mathcal{E}_{\ell_{0-1}^{\text{bi}}}(h) - \mathcal{E}_{\ell_{0-1}^{\text{bi}}}(h^*) \right]^{\alpha}.$$

Under Tsybakov noise, one can strengthen \mathcal{H} -consistency bounds from the generic “square-root” behavior to faster exponents.

Theorem 6 (Bounds under Tsybakov noise condition) Consider binary classification where the Tsybakov noise assumption holds. Assume there is no approximation error and $\mathcal{M}_{\ell_{0-1}^{\text{bi}}}(\mathcal{H}) = 0$. If for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$,

$$\Delta\mathcal{C}_{\ell_{0-1}^{\text{bi}}, \mathcal{H}}(h, x) \leq \Gamma(\Delta\mathcal{C}_{\ell, \mathcal{H}}(h, x)), \quad \text{with } \Gamma(x) = x^{\frac{1}{s}} \text{ for some } s \geq 1,$$

then for any $h \in \mathcal{H}$,

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

Loss functions	Φ	Γ	\mathcal{H} -consistency bounds
Hinge	$\Phi_{\text{hinge}}(u) = \max\{0, 1 - u\}$	x^1	$\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$
Logistic	$\Phi_{\log}(u) = \log(1 + e^{-u})$	x^2	$c^{\frac{1}{2-\alpha}} [\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Exponential	$\Phi_{\exp}(u) = e^{-u}$	x^2	$c^{\frac{1}{2-\alpha}} [\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Squared-hinge	$\Phi_{\text{sq-hinge}}(u) = (1 - u)^2 \mathbf{1}_{u \leq 1}$	x^2	$c^{\frac{1}{2-\alpha}} [\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})]^{\frac{1}{2-\alpha}}$
Sigmoid	$\Phi_{\text{sig}}(u) = 1 - \tanh(ku)$, $k > 0$	x^1	$\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$
ρ -Margin	$\Phi_\rho(u) = \min \left\{ 1, \max \left\{ 0, 1 - \frac{u}{\rho} \right\} \right\}$, $\rho > 0$	x^1	$\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$

Table 3: Enhanced bounds under Tsybakov noise

3.4. Examples: common surrogate losses under Tsybakov noise

Using Theorem 6 on standard margin losses yields the following bounds:

We note that

- **Hinge / sigmoid / ρ -margin:** behave like $s = 1$ (linear), giving essentially linear bounds. This does not contradict the quadratic universal growth rate because these losses are either non-convex or non-smooth.
- **Logistic / exponential / squared-hinge:** behave like $s = 2$ (quadratic), giving

$$\mathcal{E}_{0-1}(h) - \mathcal{E}_{0-1}^*(\mathcal{H}) \lesssim \left(\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H}) \right)^{\frac{1}{2-\alpha}}.$$

4. Significance of H-Consistency Bounds

4.1. Generalization bounds

One can use Rademacher complexity bounds on the surrogate loss functions to transform H-Consistency Bounds into generalization bounds.

For instance, H-Consistency bounds are of the form $\mathcal{E}_l - \mathcal{E}_l^* + \mathcal{M}_l \leq \Gamma(\mathcal{E}_{l_{\text{surr}}} - \mathcal{E}_{l_{\text{surr}}}^* + \mathcal{M}_{l_{\text{surr}}})$, and you can use standard Rademacher complexity bounds given in the textbook as follows.

$|\mathcal{E}_l(h) - \widehat{\mathcal{E}}_{l,S}(h)| \leq 2\mathcal{R}_m^l(\mathcal{H}) + \mathcal{B}_l \sqrt{\frac{\log(2/\delta)}{2m}}$ with probability at least δ , where the sample size is m .

For instance,

$$\mathcal{E}_l(\hat{h}) - \mathcal{E}_l^*(\mathcal{H}) \leq \mathcal{E}_l(\hat{h}) - \widehat{\mathcal{E}}_{l,S}(\hat{h}) + \widehat{\mathcal{E}}_{l,S}(\hat{h}^*) - \mathcal{E}_l^*(h^*) + \epsilon \leq 2(2\mathcal{R}_m^l(\mathcal{H}) + \mathcal{B}_l \sqrt{\frac{\log(2/\delta)}{2m}}) + \epsilon$$

Then you would place this bound into the H-Consistency Bound and receive a bound of the form

$$\mathcal{E}_l - \mathcal{E}_l^* \leq \Gamma(4\mathcal{R}_m^l(\mathcal{H}) + 2\mathcal{B}_l \sqrt{\frac{\log(2/\delta)}{2m}} + \mathcal{M}_{l_{\text{surr}}}) - \mathcal{M}_l$$

4.2. Useful framework for future analysis

The proof techniques and tools can be further used to analyze \mathcal{H} -consistency bounds in other situations, hypothesis sets, and loss functions.

5. Numerical Experiments

As an extension to the project, we conducted a few extra experiments. We primarily tested the enhanced excess error bound under the Tsybakov noise condition for small toy 1D binary classification logistic regression and ReLU neural network models.

Setup: We use a sample of 1000 data points generated from inverse transform sampling that follow a "maximal" Tsybakov noise condition, $\mathbb{P}[|\eta(X) - 1/2| \leq t] = B t^{\frac{\alpha}{1-\alpha}}$, where the original \leq is now equality and $B = 2^{\frac{\alpha}{1-\alpha}}$. The noise follows a sigmoid function. We assume that the minimizability gaps and best-in class errors are 0, which is reasonable given the simplicity of the problem. Both models are trained using SGD with batch size of 5 and learning rates of 0.1 and 0.005 respectively. Both models use logistic loss, so we can set $s = 2$ (quadratic). We plot surrogate errors as x axis and target errors as y axis. We approximate the generalization error with the mean error over the dataset. Note that we cannot give an explicit value for the c constant since we only assume its existence, thus the graphs may be inaccurate up to a translation of the theoretical bounds.

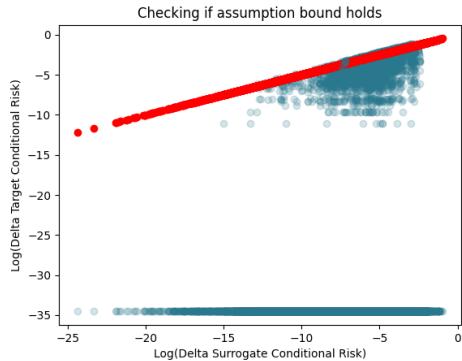


Figure 1: Logistic Regression: Assumptions, $\alpha = 0.4$

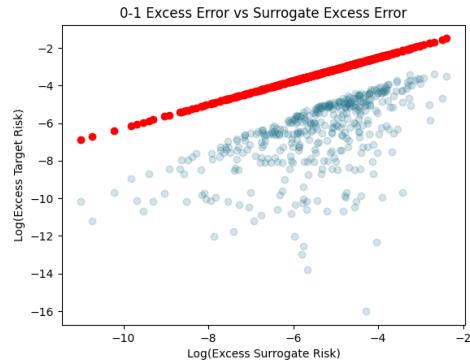
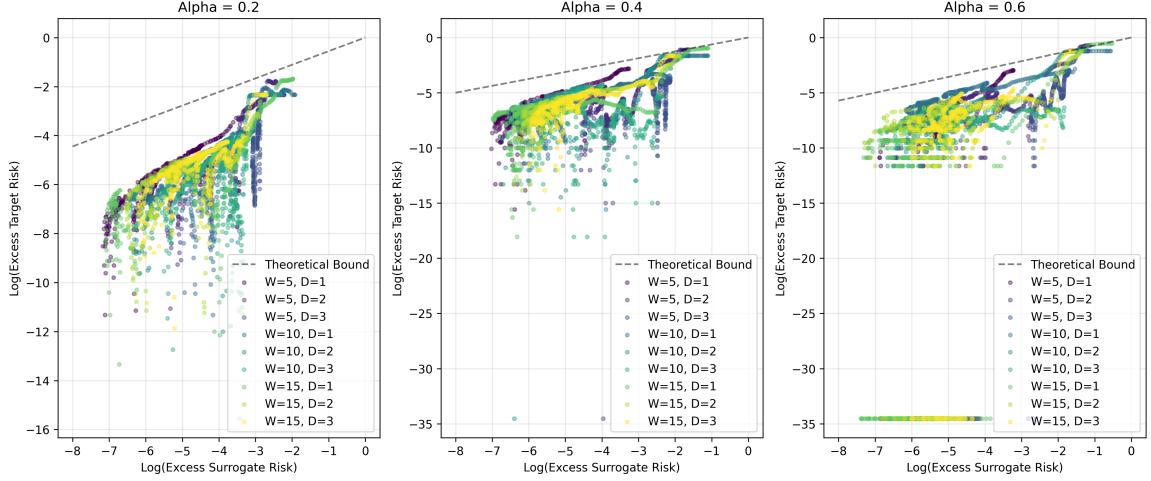


Figure 2: Logistic Regression: Excess Error, $\alpha = 0.4$

Interpretation: For the logistic regression graphs, the red points are the theoretical upper bounds and the other points correspond to the generalization errors computed by running a forward pass for each hypothesis obtained after each iteration of training over a batch. As we can see, the assumption holds and is very tight. What is remarkable is that the excess error has a nice uniform linear bound (in log plot). Additionally, the theoretical bound actually seems to be less tight because as surrogate error decreases, the gap between the red line and the best linear upper bound for the blue points seems to widen which suggests that the relationship is even closer to linear than what the theorem proposes. For the ReLU NN graphs, we omit the assumption graphs for sake of clarity and space. Again, we see that the excess error bounds are satisfied, but the rate seems faster than what the theoretical bound suggests, especially for smaller α . Additionally, the rate seems to not be uniform as we see very rapid decrease in target error to begin with and then it levels off to uniform.

Figure 3: ReLU Neural Networks: Excess Error, many α , widths, and depths

References

- Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1117–1174. PMLR, 17–23 Jul 2022. URL <https://proceedings.mlr.press/v162/awasthi22c.html>.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. Enhanced h -consistency bounds, 2024a. URL <https://arxiv.org/abs/2407.13722>.
- Anqi Mao, Mehryar Mohri, and Yutao Zhong. A universal growth rate for learning with smooth surrogate losses, 2024b. URL <https://arxiv.org/abs/2405.05968>.

Appendix A. Extended results

We note some more mathematical aspects related to [Mao et al. \(2024a\)](#). In Enhanced H-Consistency Bounds, Theorem 11 is as follows:

Theorem 11 Assume that Φ is convex and differentiable, and satisfies $\Phi'(t) < 0$ for all $t \in \mathbb{R}$, and $\frac{\Phi'(t)}{\Phi'(-t)} = e^{-\nu t}$ for some $\nu > 0$. Then ℓ_Φ is calibrated with respect to L_Φ .

Where calibrated means that for any $x, x' \in X$ $\Delta C_{\ell, \mathcal{H}_{all}}(h, x) = 0$ and $\Delta C_{\ell, \mathcal{H}_{all}}(h, x') = 0$, then $\Delta \bar{C}_{L, \mathcal{H}_{all}}(h, x, x') = 0$.

Theorem 11 gives examples of margin-based loss function that could potentially have useful H-Consistency bounds with respect to misranking. Our improvement proves the converse (with an additional distribution assumption) – if a margin-based loss function Φ has useful H-Consistency bounds for misranking, then it must satisfy $\frac{\Phi'(t)}{\Phi'(-t)} = e^{-\nu t}$. We outline this more precisely below.

Theorem 11 Converse Assume that Φ is strictly convex and twice differentiable, satisfies $\Phi'(t) < 0$, $\{\eta(x) : x \in X\} = [0, 1]$, and ℓ_Φ is calibrated with respect to L_Φ . Then $\frac{\Phi'(t)}{\Phi'(-t)} = e^{-\nu t}$ for some $\nu > 0$.

Proof of Converse of Theorem 11 Use the assumption that Φ is calibrated, then take an instance where $\Delta C_{\ell, \mathcal{H}_{all}}(h, x) = 0$ and $\Delta C_{\ell, \mathcal{H}_{all}}(h, x') = 0$, then $\Delta \bar{C}_{L_\Phi} = 0$. This is equivalent to saying that $\frac{dC_{\ell, \Phi}(h, x)}{dh(x)} = 0$ (by strict convexity and differentiability), i.e. $(\eta(x)\Phi(h(x)) + (1 - \eta(x))\Phi(-h(x)))' = 0$, which implies $\Phi'(h^*(x))/\Phi'(-h^*(x)) = \frac{1 - \eta(x)}{\eta(x)}$.

Applying a similar reasoning to the other loss functions gives the equation

$$\frac{\Phi'(h^*(x))}{\Phi'(-h^*(x))} \frac{\Phi'(-h^*(x'))}{\Phi'(h^*(x'))} = \frac{\Phi'(h^*(x) - h^*(x'))}{\Phi'(-h^*(x) + h^*(x'))}$$

Define $f(x) = \frac{\Phi'(x)}{\Phi'(-x)}$, then we have the equation that $f(h^*(x))f(-h^*(x')) = f(h^*(x) - h^*(x'))$ where $(\eta(x), \eta(x'))$ achieves all values in $[0, 1]^2$, so $f(h^*(x)), f(h^*(x'))$ achieves all values in $[0, \infty)^2$ (from the assumption on $\eta(x)$ and the fact that $f(h^*(x)) = \frac{1 - \eta(x)}{\eta(x)}$).

Also, $f'(x) < 0$ (strict convexity, and twice differentiability), and since $\{(f(h^*(x)), f(h^*(x'))): x, x' \in X^2\} = [0, \infty)^2$, $(h^*(x), h^*(x'))$ can take any value in \mathbb{R}^2 (using the fact that f 's range is $[0, \infty)$, and f is continuously monotone (Φ' is continuous), and the exercise that $f(g(x))$ being bijective, and f being bijective, implies g is bijective). Thus $f(x)f(y) = f(x+y)$ for all $x, y \in \mathbb{R}$, and the result follows by Cauchy's theorem. i.e. $\frac{\Phi'(x)}{\Phi'(-x)} = e^{-\nu t}$ ■

Additionally, regarding theorem 13 of Enhanced H-Consistency Bounds, we believe there is a small mistake in the authors' proof, and we give an explanation below.

Theorem 13 attempts to prove the inequality

$$\Delta \bar{C}_{L_{\Phi_{\log}}}(h, x, x') \leq \max(\eta', 1 - \eta')\Delta C_{\ell_{\Phi_{\log}}}(h, x) + \max(\eta, 1 - \eta)\Delta C_{\ell_{\Phi_{\log}}}(h, x')$$

. The proof can be found in [Mao et al. \(2024a\)](#) Appendix F.4. Within the proof, there is a jump in logic that implicitly assumes that

$$\begin{aligned} & \eta(1 - \eta')(\Phi_{\log}(h) + \Phi_{\log}(-h')) + \eta'(1 - \eta)(\Phi_{\log}(-h) + \Phi_{\log}(h')) \\ & + \eta(1 - \eta')(\log \eta + \log(1 - \eta')) + \eta'(1 - \eta)(\log \eta' + \log(1 - \eta)). \end{aligned}$$

is bounded above by

$$\begin{aligned} & \max(\eta', 1 - \eta') \left[\eta(\Phi_{\log}(h) + \log(\eta)) + (1 - \eta)(\Phi(h) + \log(1 - \eta)) \right] \\ & + \max(\eta, 1 - \eta) \left[\eta'(\Phi_{\log}(h) + \log(\eta')) + (1 - \eta')(\Phi(h) + \log(1 - \eta')) \right] \end{aligned}$$

However, one cannot bound $\eta(1 - \eta') \log(\eta)$ above by $\max(\eta', 1 - \eta')\eta \log(\eta)$ because both terms are negative. Figuring out a precise counterexample from this step is probably doable, but nuanced (lots of writing), and our preliminary attempts to create a working bound of the form seen in Theorem 13 haven't worked. If we had more time to try and prove something, we would begin with a Pinsker style inequality, since $\Delta C_\ell = KL(\eta || \sigma(h))$, where $\sigma(h) = \frac{1}{1+e^{-h}}$, and then we would try to do some sort of Taylor series bound for the LHS.