H-Consistency Bounds

Yizheng Li, Edwin Liu, John Quigley

December 16, 2025

Motivation

- Algorithms optimize using a surrogate loss function different from the target loss function
  - Target loss (eg. 0-1 loss) is hard to optimize (non-smooth, non-differentiable)
- What guarantees can we give on the target loss estimation error when we minimize the surrogate loss estimation error?

Historical result: H-consistency bound setup

Definitions and notations:

- Noise: $\eta(x) = \Pr[Y = 1|X = x]$.
- Conditional ℓ-risk: $\mathcal{C}_\ell(h, x) = \eta(x)\ell(h, x, +1) + (1 - \eta(x))\ell(h, x, -1)$.
- notation for gap: $\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x) = \mathcal{C}_\ell(h, x) - \inf_{h \in \mathcal{H}} \mathcal{C}_\ell(h, x)$
- generalization error: $\mathcal{E}_\ell(h) := \mathbb{E}_X[\mathcal{C}_\ell(h, x)]$. $\mathcal{E}^*_\ell(\mathcal{H}) := \inf_{h \in \mathcal{H}} \mathcal{E}_\ell(h)$
- Minimizability gap: $\mathcal{M}_\ell(\mathcal{H}) = \mathcal{E}^*_\ell(\mathcal{H}) - \mathbb{E}_x[\inf_{h \in \mathcal{H}} \mathcal{C}_\ell(h, x)]$
- Note that $\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(h) + \mathcal{M}_\ell(\mathcal{H}) = \mathbb{E}_X[\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x)]$
- Margin based losses: $\ell(h, x, y) = \Phi(yh(x))$

The History of H-consistency

Definition 1 (Bayes-consistency) *A loss function ℓ_1 is Bayes-consistent with respect to a loss function ℓ_2, if for any distribution $\mathcal{D}$ and any sequence $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_{all}$,*

$$\lim_{n\to+\infty} \mathcal{E}_{\ell_1}(h_n) - \mathcal{E}^*_{\ell_1}(\mathcal{H}_{\text{all}}) = 0 \quad \text{implies} \quad \lim_{n\to+\infty} \mathcal{E}_{\ell_2}(h_n) - \mathcal{E}^*_{\ell_2}(\mathcal{H}_{\text{all}}) = 0.$$

Definition 2 ($\mathcal{H}$-consistency). *We say that ℓ_1 is $\mathcal{H}$-consistent with respect to ℓ_2, if, for all distributions $\mathcal{D}$ and sequences $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{H}$, we have*

$$\lim_{n\to+\infty} \mathcal{E}_{\ell_1}(h_n) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) = 0 \implies \lim_{n\to+\infty} \mathcal{E}_{\ell_2}(h_n) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) = 0.$$

Definition 3 ($\mathcal{H}$-consistency bounds) *Given a hypothesis set $\mathcal{H}$, an $\mathcal{H}$-consistency bound relating the loss function ℓ_1 to the loss function ℓ_2 for a hypothesis set $\mathcal{H}$ is an inequality of the form*

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

that holds for any distribution $\mathcal{D}$, where $\Gamma : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing concave function with $\Gamma \ge 0$ (Awasthi et al., 2022b,a). Here, $\mathcal{M}_{\ell_1}(\mathcal{H})$ and $\mathcal{M}_{\ell_2}(\mathcal{H})$ are minimizability gaps for the respective loss functions.

Universal growth rate bounds based on HCB

Consider the case when the target loss is just the **0-1 loss**, then we have a function $\mathcal{T}$, and we call it the $\mathcal{H}$**-estimation error transformation** for the surrogate loss ℓ and the following holds **tightly**

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

Tight means that for any $t \in [0,1]$, there exists a hypothesis $h \in \mathcal{H}$ and a distribution such that $\mathcal{E}_{\ell_{0-1}}(h) - \mathcal{E}^*_{\ell_{0-1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0-1}}(\mathcal{H}) = t$ and $\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H}) = \mathcal{T}(t)$. And in the case where $\mathcal{H}$ is complete $(\forall x, \{h(x) | h \in \mathcal{H}\} = \mathbb{R})$, we have that $\mathcal{T}$ takes the form:

$$f_t(u) = \frac{1-t}{2}\Phi(u) + \frac{1+t}{2}\Phi(-u), t \in [0,1]$$

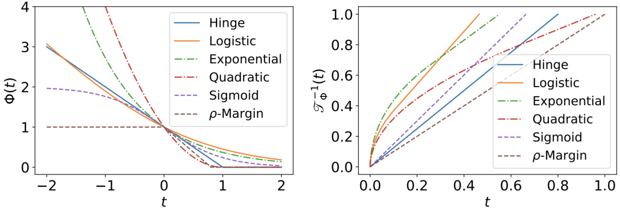
$$\mathcal{T}(t) := \inf_{u \geq 0} f_t(u) - \inf_{u \in \mathbb{R}} f_t(u)$$

Furthermore, a theorem proved in *Universal Growth Rate* says that when Φ is differentiable at 0, and $\Phi'(0) < 0$, we have that $\mathcal{T}(t) = f_t(0) - \inf_{u \in \mathbb{R}} f_t(u)$
(one small note is that $\mathcal{T}(0) = 0$

Universal Growth Rate: Transformation Table

To demonstrate, consider $\mathcal{H}_{\mathrm{lin}} = \{x \mapsto w \cdot x + b \mid \|w\|_q \leq W, \ |b| \leq B\}$

Surrogates	$\mathcal{T}_\Phi(t), t \in [0, 1]$
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	$\frac{\min\{B, \rho\}}{\rho}t$



Universal Growth Rate: Main theoretical results

Theorem 5 (Upper and lower bound for binary margin-based losses) Let $\mathcal{H}$ be a complete hypothesis set. Assume that Φ is convex, twice continuously differentiable, and satisfies the inequalities $\Phi'(0) > 0$ and $\Phi''(0) > 0$. Then, the following property holds: $\mathcal{T}(t) = \Theta(t^2)$; that is, there exist positive constants $C > 0$, $c > 0$, and $T > 0$ such that $Ct^2 \geq \mathcal{T}(t) \geq ct^2$, for all $0 < t \leq T$.

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 $\frac{da_t^*}{dt}\Big|_{t=0} = \frac{\Phi'(0)}{\Phi''(0)} > 0$, hence $a_t^* = \Theta(t)$.
- 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 $\mathcal{T}(t) = \Theta(t^2)$.

Universal Grow Rate: Results and extension

Define $V_\ell := \mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$.

$\Phi(u)$	margin-based losses ℓ	$\mathcal{H}$-Consistency bounds
e^{-u}	$e^{-yh(x)}$	$V_{\ell_{0-1}} \leq \sqrt{2(V_\ell)}$
$\log(1+e^{-u})$	$\log(1+e^{-yh(x)})$	$V_{\ell_{0-1}} \leq \sqrt{2(V_\ell)}$
$\max\{0, 1-u\}^2$	$\max\{0, 1-yh(x)\}^2$	$V_{\ell_{0-1}} \leq \sqrt{V_\ell}$
$\max\{0, 1-u\}$	$\max\{0, 1-yh(x)\}$	$V_{\ell_{0-1}} \leq V_\ell$

The result can also be extended to multi-class comp-sum losses:

Theorem 8 (Upper and lower bound for comp-sum losses) *Assume that Φ is convex, twice continuously differentiable, and satisfies the properties $\Phi'(u) < 0$ and $\Phi''(u) > 0$ for any $u \in (0, \frac{1}{2}]$. Then, the following property holds: $\mathcal{T}(t) = \Theta(t^2)$.*

Enhanced HCB

What if we allow possibly non-constant α and β to modify the Γ?
Result: a more general bound with a hypothesis-dependent parameter γ
Theorem 2 Assume that there exist a concave function $\Gamma : \mathbb{R}_+ \to \mathbb{R}$ and two positive functions $\alpha : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ and $\beta : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_+^*$ with $\sup_{x \in \mathcal{X}} \alpha(h, x) < +\infty$ and $\mathbb{E}_{x \in \mathcal{X}}[\beta(h, x)] < +\infty$ for all $h \in \mathcal{H}$ such that the following holds for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$:

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

Then, the following inequality holds for any hypothesis $h \in \mathcal{H}$:

$$\mathcal{E}_{\ell_2}(h) - \mathcal{E}^*_{\ell_2}(\mathcal{H}) + \mathcal{M}_{\ell_2}(\mathcal{H}) \leq \Gamma\left(\gamma(h)\left(\mathcal{E}_{\ell_1}(h) - \mathcal{E}^*_{\ell_1}(\mathcal{H}) + \mathcal{M}_{\ell_1}(\mathcal{H})\right)\right), \tag{3}$$

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, $\mathcal{X}$ is a subset of $\mathbb{R}^n$ and, for any $h \in \mathcal{H}$, $x \mapsto \Delta \mathcal{C}_{\ell_1, \mathcal{H}}(h, x)$ is non-decreasing and $x \mapsto \alpha(h, x)\beta(h, x)$ is non-increasing, or vice-versa, then, the inequality holds with $\gamma(h) = \mathbb{E}_X\left[\frac{\alpha(h,x)\beta(h,x)}{\mathbb{E}_X[\beta(h,x)]}\right]$.

Enhanced HCB, cont'd

Consider the Tsybakov noise condition (Mammen and Tsybakov, 1999), that is there exist $B > 0$ and $\alpha \in [0, 1)$ such that

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

Note that as $\alpha \to 1$, $t^{\frac{\alpha}{1-\alpha}} \to 0$, corresponding to Massart's noise condition. When $\alpha = 0$, the condition is void. This condition is equivalent to assuming the existence of a universal constant $c > 0$ and $\alpha \in [0, 1)$ such that for all $h \in \mathcal{H}$, the following inequality holds (Bartlett et al., 2006):

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

where h^* is the Bayes-classifier. We also assume that there is no approximation error and that $\mathcal{M}_{\ell_{0-1}^{\mathrm{bi}}}(\mathcal{H}) = 0$.

Theorem 6 *Consider a binary classification setting where the Tsybakov noise assumption holds. Assume that the following holds for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$: $\Delta\mathcal{C}_{\ell_{0-1}^{bi},\mathcal{H}}(h, x) < \Gamma(\Delta\mathcal{C}_{\ell,\mathcal{H}}(h, x))$, with $\Gamma(x) = x^{\frac{1}{s}}$, for some $s \geq 1$. Then, for any $h \in \mathcal{H}$,*

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

Enhanced HCB, cont'd

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	$c^{\frac{1}{2-\alpha}} \left[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\right]^{\frac{1}{2-\alpha}}$
Exponential	$\Phi_{\exp}(u) = e^{-u}$	x^2	$c^{\frac{1}{2-\alpha}} \left[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\right]^{\frac{1}{2-\alpha}}$
Squared-hinge	$\Phi_{\text{sq-hinge}}(u) = (1-u)^2 1_{u \leq 1}$	x^2	$c^{\frac{1}{2-\alpha}} \left[\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\right]^{\frac{1}{2-\alpha}}$
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	$\mathcal{E}_\ell(h) - \mathcal{E}_\ell^*(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})$

Generalization Bounds

$$\hat{h}_S = \arg\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell(h, x_i, y_i).$$

$\mathcal{H} = \{(x, y) \mapsto \ell(h, x, y) : h \in \mathcal{H}\}$ and $\mathfrak{R}_m^{\ell}(\mathcal{H})$ its Rademacher complexity. We also write B_ℓ to denote an upper bound for ℓ. Then, given the following $\mathcal{H}$-consistency bound:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_{\ell_{0\text{-}1}}(h) - \mathcal{E}^*_{\ell_{0\text{-}1}}(\mathcal{H}) + \mathcal{M}_{\ell_{0\text{-}1}}(\mathcal{H}) \leq \Gamma\left(\mathcal{E}_\ell(h) - \mathcal{E}^*_\ell(\mathcal{H}) + \mathcal{M}_\ell(\mathcal{H})\right), \tag{21}$$

for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, the following estimation bound holds for $\hat{h}_S$:

$$\forall h \in \mathcal{H}, \quad \mathcal{E}_{\ell_{0\text{-}1}}(h) - \mathcal{E}^*_{\ell_{0\text{-}1}}(\mathcal{H}) \leq \Gamma\left(4\mathfrak{R}_m^{L}(\mathcal{H}) + 2B_L\sqrt{\frac{\log\frac{2}{\delta}}{2m}} + \mathcal{M}_\ell(\mathcal{H})\right) - \mathcal{M}_{\ell_{0\text{-}1}}(\mathcal{H}).$$

Proof Sketch:

$$\mathcal{E}_\ell(\hat{h}_S) - \mathcal{E}^*_\ell(\mathcal{H}) \leq 4\,\mathfrak{R}_m^{\ell}(\mathcal{H}) + 2B_\ell\sqrt{\frac{\log(2/\delta)}{2m}}\,.$$

Numerical Experiments

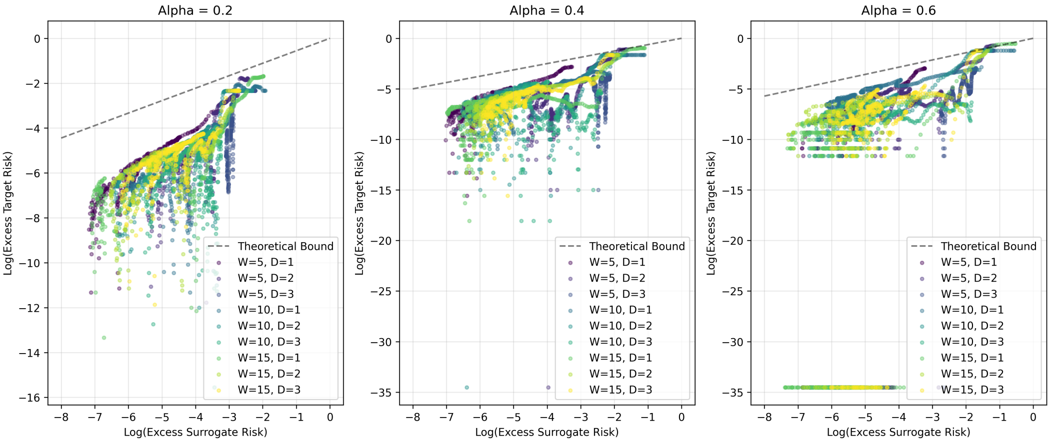


Figure: Validating theorem 6 of EHCB on ReLU neural networks of varying width and depth against varying alphas