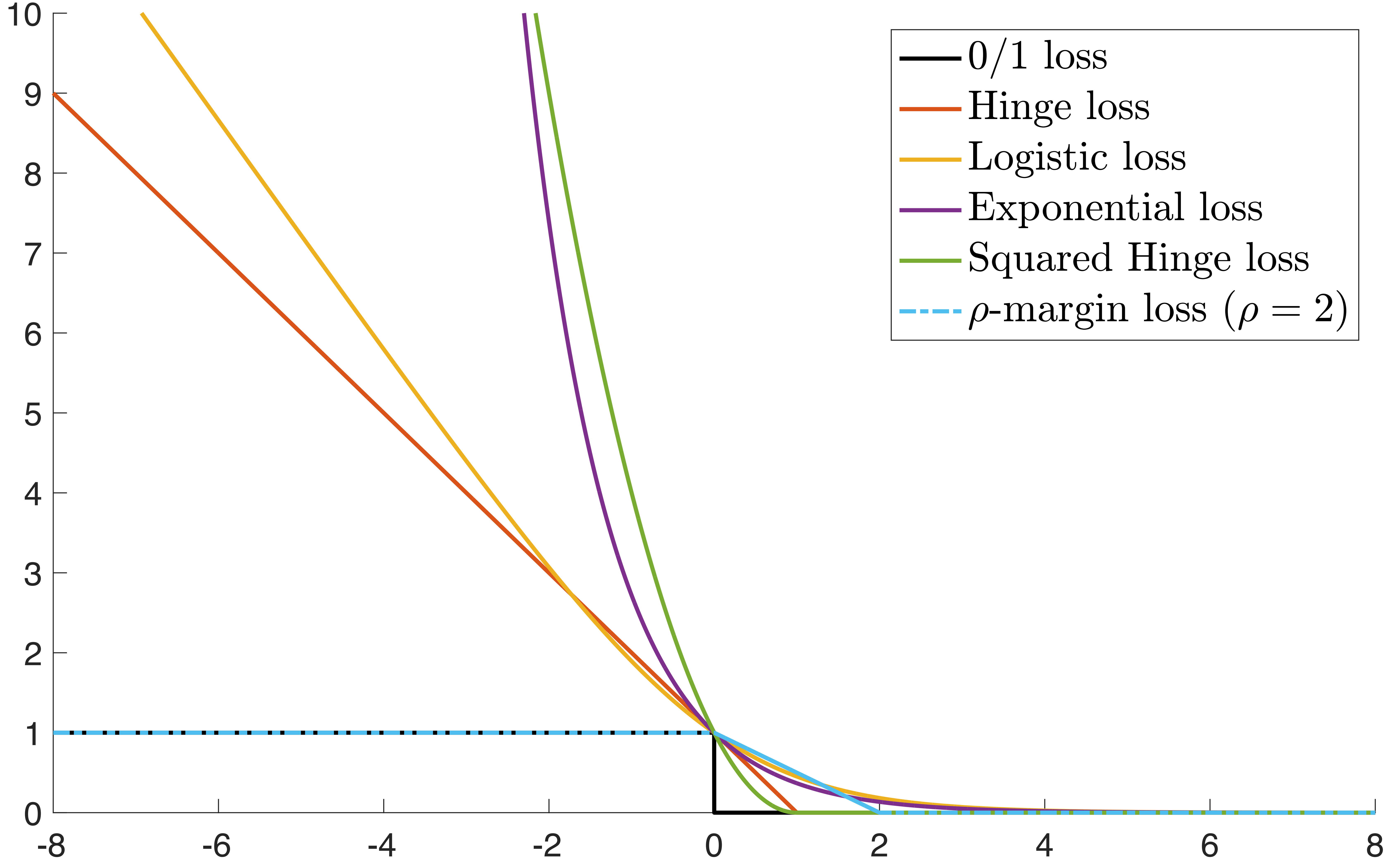
H-Consistency Bounds for Surrogate Loss Minimizers

Pranjal Awasthi Anqi Mao Mehryar Mohri Yutao Zhong

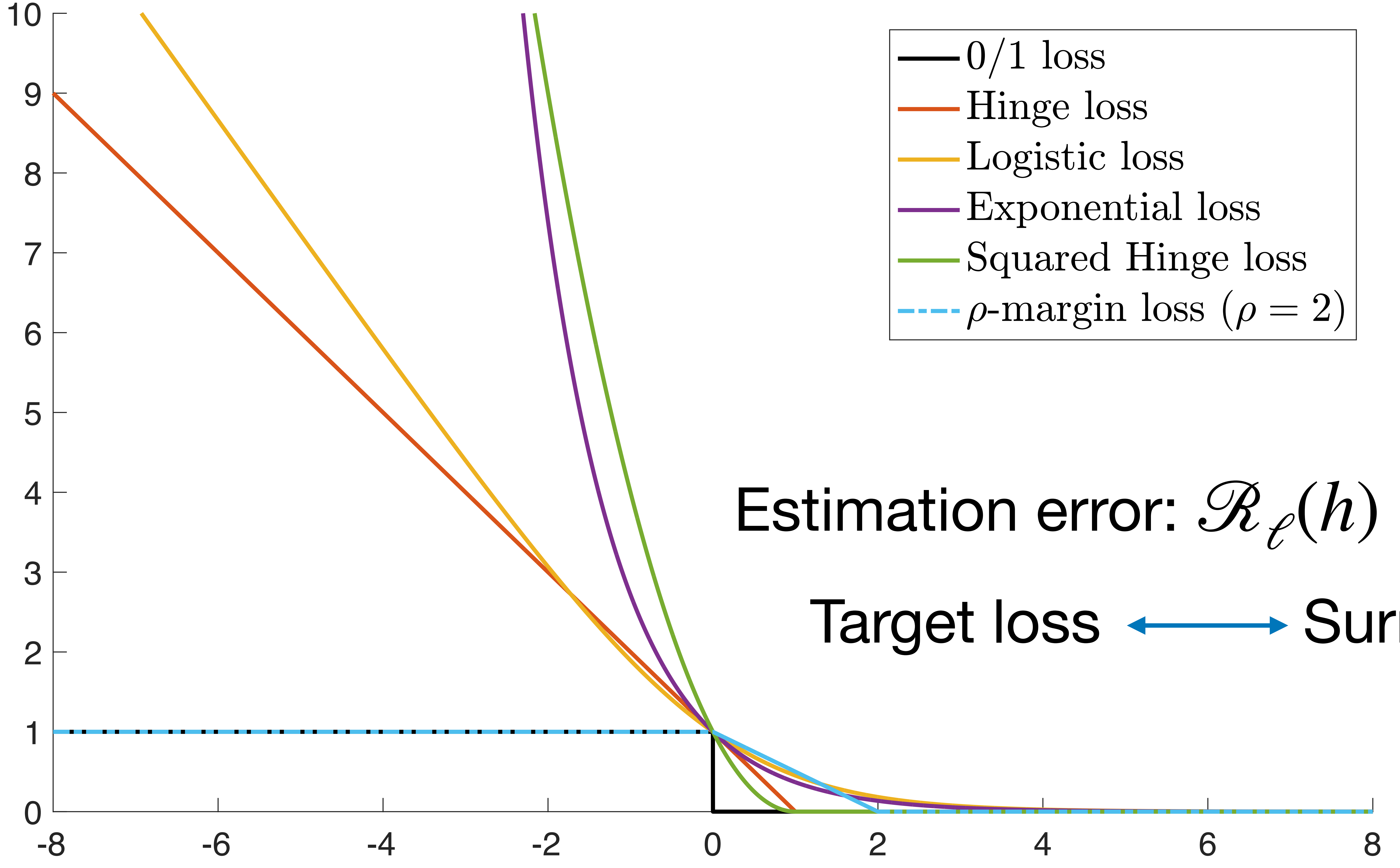
Google Research and Courant Institute

ICML
International Conference
On Machine Learning

Surrogate Loss versus Target Loss



Surrogate Loss versus Target Loss



Estimation error: $\mathscr{R}_{\ell}(h) - \mathscr{R}^{*}_{\ell,\mathscr{H}}$

Target loss ⟷ Surrogate loss

Bayes-consistency

$$\lim_{n\to\infty} \mathscr{R}_{\ell_1}(h_n) - \mathscr{R}^*_{\ell_1,\mathscr{H}_{\mathrm{all}}} = 0 \Rightarrow \lim_{n\to\infty} \mathscr{R}_{\ell_2}(h_n) - \mathscr{R}^*_{\ell_2,\mathscr{H}_{\mathrm{all}}} = 0$$

Bayes-consistency

$$\lim_{n\to\infty} \mathscr{R}_{\ell_1}(h_n) - \mathscr{R}^*_{\ell_1,\mathscr{H}_{\text{all}}} = 0 \Rightarrow \lim_{n\to\infty} \mathscr{R}_{\ell_2}(h_n) - \mathscr{R}^*_{\ell_2,\mathscr{H}_{\text{all}}} = 0$$

H-consistency

$$\lim_{n\to\infty} \mathscr{R}_{\ell_1}(h_n) - \mathscr{R}^*_{\ell_1,\mathscr{H}} = 0 \Rightarrow \lim_{n\to\infty} \mathscr{R}_{\ell_2}(h_n) - \mathscr{R}^*_{\ell_2,\mathscr{H}} = 0$$

$\mathscr{H}$-Consistency Bound

$$\mathscr{R}_{\ell_2}(h) - \mathscr{R}^*_{\ell_2,\mathscr{H}} \leq f\left(\mathscr{R}_{\ell_1}(h) - \mathscr{R}^*_{\ell_1,\mathscr{H}}\right)$$

$\mathscr{H}$-Consistency Bound

$$\mathscr{R}_{\ell_2}(h) - \mathscr{R}^*_{\ell_2,\mathscr{H}} \leq f\left(\mathscr{R}_{\ell_1}(h) - \mathscr{R}^*_{\ell_1,\mathscr{H}}\right)$$

$$\forall\, h \in \mathscr{H}, \mathscr{D} \in \mathscr{P}$$

$\mathscr{P}_{\text{all}}$ — distribution independent

Standard Binary Classification

Family of all measurable functions: $\mathscr{H}_{\text{all}}$

All distributions: $\mathscr{P}_{\text{all}}$

(Zhang, 04a; Bartlett et al., 06; Mohri et al., 18)

ℓ_1: margin-based loss ϕ ⟷ ? ⟷ ℓ_2 : stanard $0/1$ loss ℓ_{0-1}

Excess error bound ($\mathscr{H}_{\text{all}}$-consistency bound)

Standard Binary Classification

General hypothesis sets: $\mathscr{H}$
Distribution-independent and distribution-dependent
(Our contribution)

ℓ_1: margin-based loss ϕ $\longleftarrow ? \longrightarrow$ ℓ_2 : stanard $0/1$ loss ℓ_{0-1}

$\mathscr{H}$-consistency bound

Adversarial Attacks (Szegedy et al., 13)

Correct Attack Ostrich Correct Attack Ostrich

Adversarially Robust Classification

General hypothesis sets: $\mathscr{H}$

Distribution-independent and distribution-dependent

(Our contribution)

ℓ_1: supremum-based margin-based loss $\tilde{\phi}$ $\longleftrightarrow$? ℓ_2 : adversarial $0/1$ loss ℓ_γ

$$\tilde{\phi} = \sup_{x' \colon \|x - x'\| \leq \gamma} \phi(yf(x'))$$

$$\ell_\gamma = \sup_{x' \colon \|x - x'\| \leq \gamma} \ell_{0-1}(yf(x'))$$

Adversarial $\mathscr{H}$-consistency bound

$\mathscr{H}$-Consistency Bounds Analysis

$$\mathscr{R}_{\ell_2}(h) - \mathscr{R}^*_{\ell_2,\mathscr{H}} \leq f\left(\mathscr{R}_{\ell_1}(h) - \mathscr{R}^*_{\ell_1,\mathscr{H}}\right)$$

Inverse of $\mathscr{H}$-estimation error transformation + Minimizability gap

Minimizability Gap

$$\mathscr{M}_{\ell,\mathscr{H}} = \mathscr{R}^*_{\ell,\mathscr{H}} - \mathbb{E}_X\left[\mathscr{C}^*_{\ell,\mathscr{H}}(x)\right]$$

difference of the best-in class error and
the expectation of the minimal conditional-risk

$$\mathscr{C}^*_{\ell,\mathscr{H}}(x) = \inf_{h\in\mathscr{H}}\left[\mathscr{D}(Y=1\,|\,X=x)\ell(h,x,+1) + (1-\mathscr{D}(Y=1\,|\,X=x))\ell(h,x,-1)\right]$$

Minimizability Gap

$$\mathscr{M}_{\ell,\mathscr{H}} = \mathscr{R}^*_{\ell,\mathscr{H}} - \mathbb{E}_X\left[\mathscr{C}^*_{\ell,\mathscr{H}}(x)\right]$$

difference of the best-in class error and
the expectation of the minimal conditional-risk

$$\mathscr{C}^*_{\ell,\mathscr{H}}(x) = \inf_{h\in\mathscr{H}}\left[\mathscr{D}(Y=1\,|\,X=x)\ell(h,x,+1) + (1-\mathscr{D}(Y=1\,|\,X=x))\ell(h,x,-1)\right]$$

we cannot hope to estimate or minimize.

$\mathscr{H}$-Estimation Error Transformation

$$\forall t \in \left[0,1\right],\ \mathscr{T}_{\Phi}(t) = \mathscr{T}(t)\mathbb{I}_{t\in\left[\epsilon,1\right]} + (\mathscr{T}(\epsilon)/\epsilon)\, t\mathbb{I}_{t\in\left[0,\epsilon\right)}$$

Where $\mathscr{T}(t) := \inf_{x\in\mathscr{X},h\in\mathscr{H}:h(x)<0} \Delta\mathscr{C}_{\Phi,\mathscr{H}}\left(h, x, \frac{t+1}{2}\right)$

$\mathscr{H}$-Estimation Error Transformation

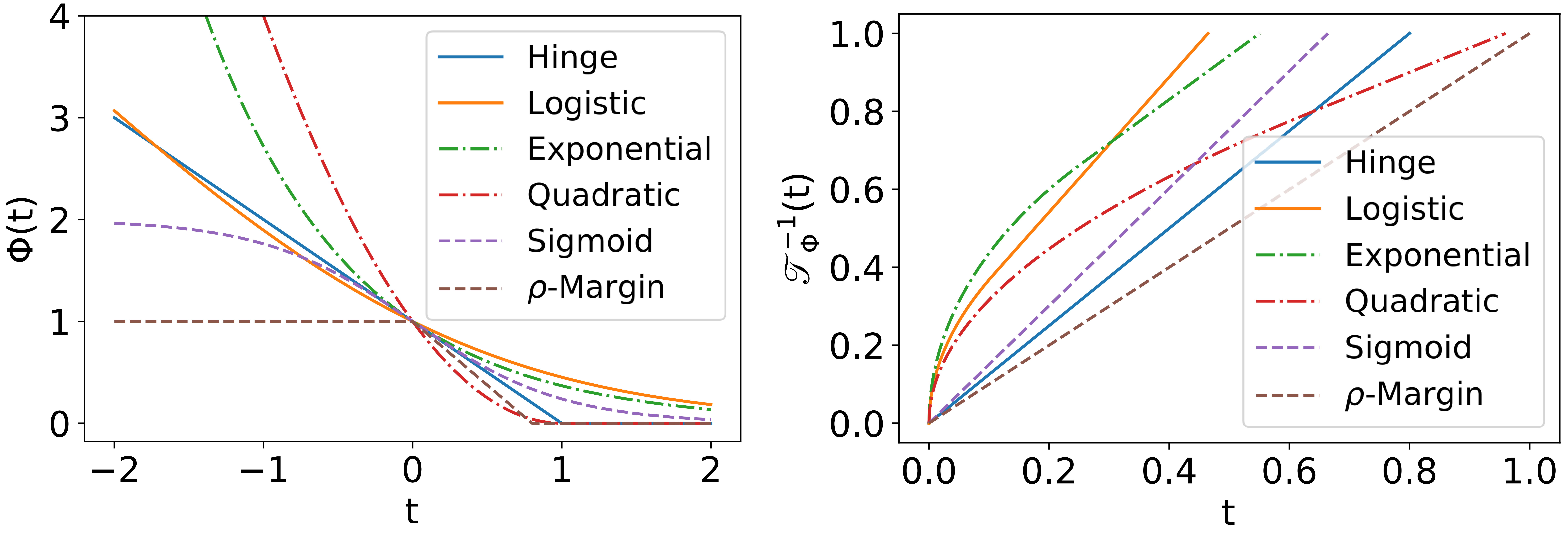
$$\forall t \in [0,1],\ \mathscr{T}_{\Phi}(t) = \mathscr{T}(t)\mathbb{1}_{t\in[\epsilon,1]} + (\mathscr{T}(\epsilon)/\epsilon)\, t\mathbb{1}_{t\in[0,\epsilon)}$$

Where $\mathscr{T}(t) := \inf_{x\in\mathscr{X},h\in\mathscr{H}:h(x)<0} \Delta\mathscr{C}_{\Phi,\mathscr{H}}\left(h, x, \frac{t+1}{2}\right)$

$\downarrow$

Tightness!

$\mathscr{H}$-Estimation Error Transformation



Conclusion

$\mathscr{H}$-consistency bounds for both standard and adversarial binary classification

- New estimation error guarantees for both the non-adversarial 0/1 loss function and the adversarial 0/1 loss function
- Compare different surrogate loss functions of the 0/1 loss or adversarial loss, given the specific hypothesis set used
- Theoretical and conceptual tools helpful for the analysis of other loss functions and other hypothesis sets

Conclusion

$\mathscr{H}$-consistency bounds for both standard and adversarial binary classification

- New estimation error guarantees for both the non-adversarial 0/1 loss function and the adversarial 0/1 loss function
- Compare different surrogate loss functions of the 0/1 loss or adversarial loss, given the specific hypothesis set used
- Theoretical and conceptual tools helpful for the analysis of other loss functions and other hypothesis sets

Future Directions

- $\mathscr{H}$-consistency bounds for other loss functions and other hypothesis sets
- Incorporating the trade-off of the optimization and $\mathscr{H}$-consistency bounds

Poster #1206 in Hall E