



# Tighter Information-Theoretic Generalization Bounds from Supersamples

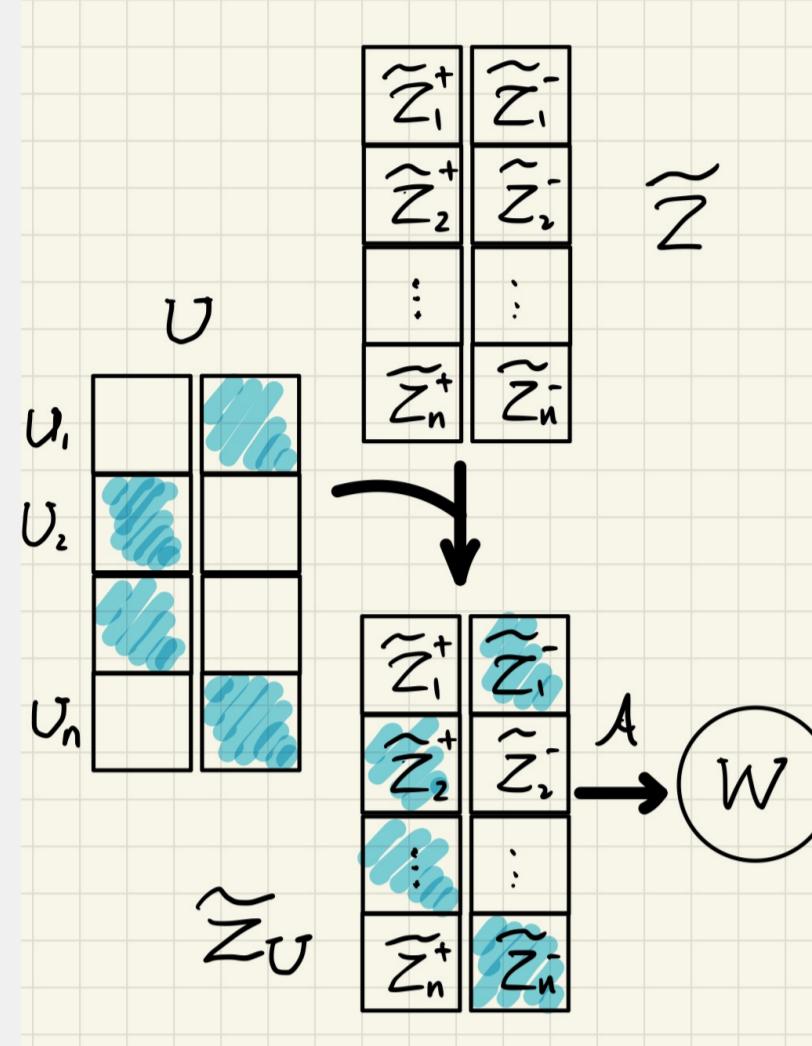
Ziqiao Wang and Yongyi Mao

University of Ottawa



uOttawa

## supersample and CMI bound (Steinke-Zakynthinou, 2020)

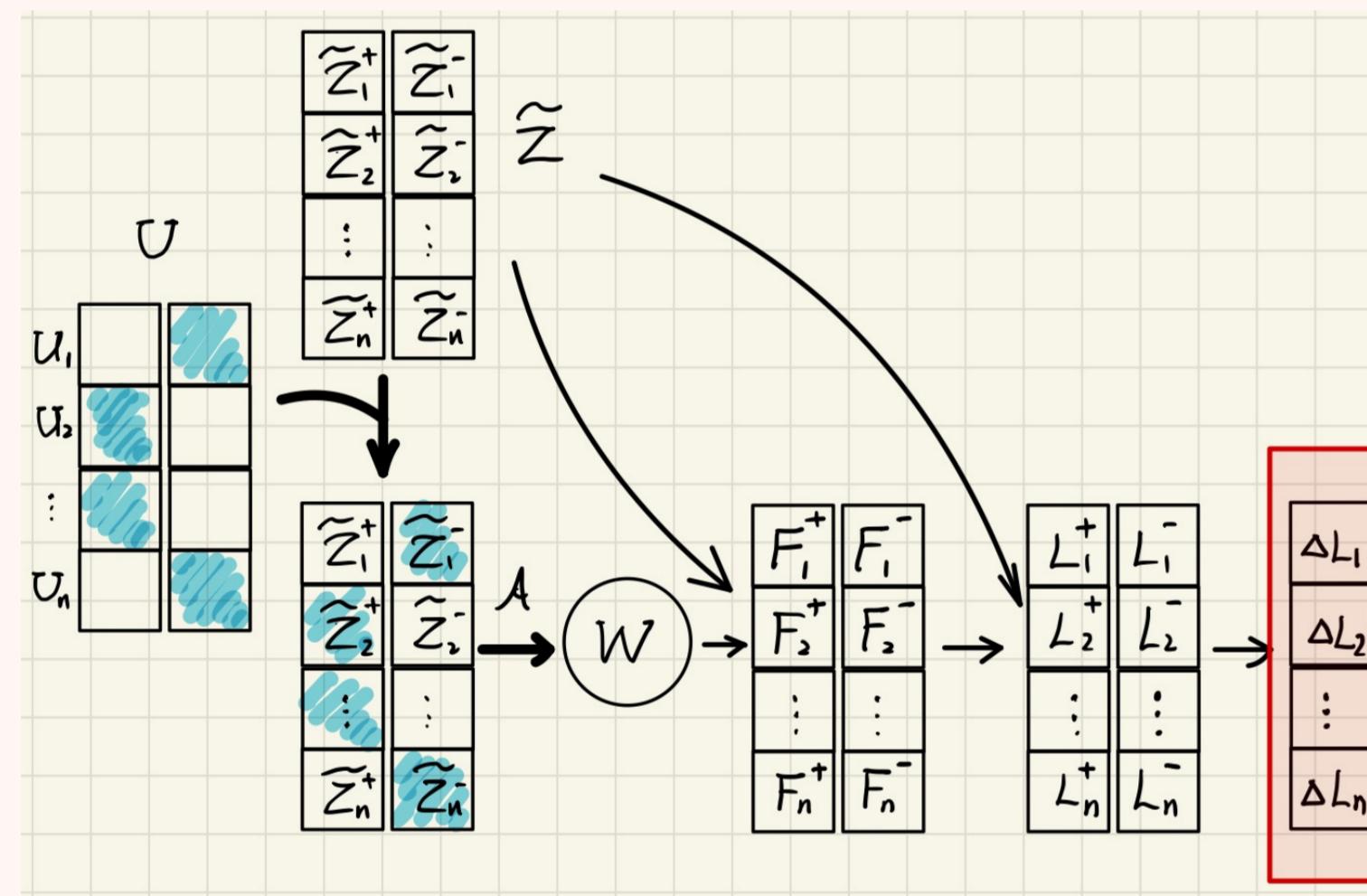


- Sample  $2n$  instances  $((X, Y)$  pairs) to fill matrix  $\tilde{Z}$
- Use a random binary mask  $U_i$  to select an instance from row  $i$  of  $\tilde{Z}$  to include in the training set  $\tilde{Z}_U$ :  $U_i = 0$  selects the left one.
- Apply learning algorithm  $A$  to  $\tilde{Z}_U$  to obtain model parameter  $W = \mathcal{A}(\tilde{Z}_U)$ .
- For bounded loss in  $[0, 1]$ , the generalization error (expected over  $W$  and  $\tilde{Z}_U$ ) is bounded by

$$|\text{Err}| \leq \frac{1}{n} \sum_{i=1}^n \sqrt{2I(W; U_i | \tilde{Z})} := B_n(I(W; U_i | \tilde{Z}))$$

This paper: tightest bounds and novel perspectives for this setting

## from CMI, f-CMI, e-CMI, to loss-difference CMI



- $F_i^+ := f_W(\tilde{Z}_i^+)$ ,  $F_i^- := f_W(\tilde{Z}_i^-)$ ,  $F_i := (F_i^+, F_i^-)$   
⇒ f-CMI Bound:  $|\text{Err}| \leq B_n(I(F_i; U_i | \tilde{Z}))$  (Harutyunyan et al., 2021)
- $L_i^+ := \ell(W, \tilde{Z}_i^+)$ ,  $L_i^- := \ell(W, \tilde{Z}_i^-)$ ,  $L_i := (L_i^+, L_i^-)$   
⇒ e-CMI Bound:  $|\text{Err}| \leq B_n(I(L_i; U_i | \tilde{Z}))$  (Hellström-Durisi, 2022)
- This paper:  $\Delta L_i := L_i^- - L_i^+$

## Theorem A: Id-CMI Bound

$$|\text{Err}| \leq B_n(I(\Delta L_i; U_i | \tilde{Z}))$$

## Id-CMI bound is tighter than e-CMI, f-CMI and CMI bounds

Noting the Markov Chain (conditioned on  $\tilde{Z}$ ):  $U_i - W - F_i - L_i - \Delta L_i$  we see

$$I(W; U_i | \tilde{Z}) \geq I(F_i; U_i | \tilde{Z}) \geq I(L_i; U_i | \tilde{Z}) \geq I(\Delta L_i; U_i | \tilde{Z})$$

Id-CMI Benefits from reducing loss-pairs to their equivalence classes.

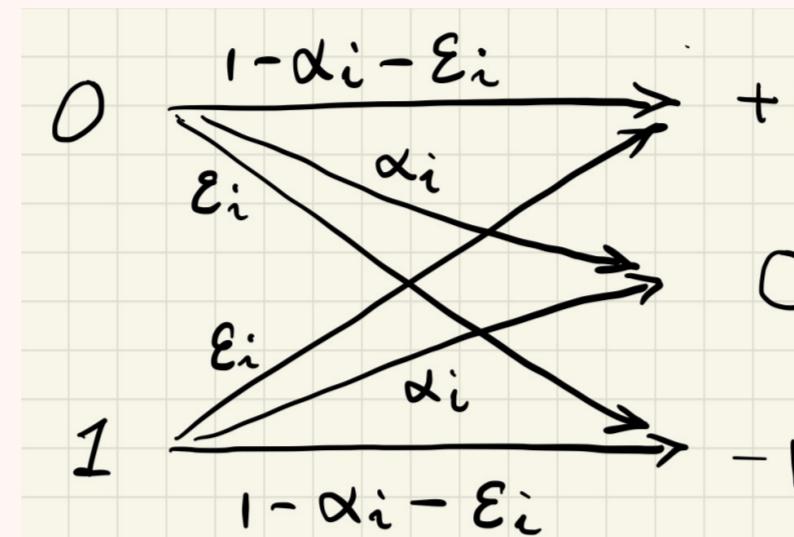
## an even tighter bound: loss-difference MI bound

## Theorem B: Id-MI bound

$$|\text{Err}| \leq B_n(I(\Delta L_i; U_i))$$

Since  $U_i$  and  $\tilde{Z}$  are independent,  $I(\Delta_i; U) \leq I(\Delta_i; U) + I(U_i; \tilde{Z} | \Delta_i) = I(\Delta_i; U | \tilde{Z})$ 

## loss-difference MI bound characterizes generalization error for interpolation algorithms under 0/1-loss



- $I(\Delta L_i; U_i)$  measures the rate of communication for channel  $U_i \rightarrow \Delta L_i$ .
- Consider 0/1-loss. Then  $\Delta L_i \in \{-1, 0, +1\}$ .
- $\alpha_i$ :  $\mathbb{P}\{\text{trainLoss} = \text{testLoss} \text{ in row } i \text{ of } \tilde{Z}\}$
- $\epsilon_i$ :  $\mathbb{P}\{\text{trainLoss} > \text{testLoss} \text{ in row } i \text{ of } \tilde{Z}\}$

- Consider an interpolation algorithm:  $\epsilon_i = 0$ . ⇒ Binary Erasure Channel.

We see  $I(\Delta L_i; U_i) = (1 - \alpha_i) \log 2$  nats (= channel capacity), and

## Theorem C: MI Characterization of Err for Interpolation Algorithm

$$|\text{Err}| = \sum_{i=1}^n \frac{I(\Delta L_i; U_i)}{n \log 2} = 1 - \sum_{i=1}^n \alpha_i/n$$

## weighted generalization error from correlating with shifted Rademacher sequence

- $R$ : population risk expected over  $W$ .
- $\hat{R}_n$ : empirical risk expected over  $W$  and training set of size  $n$
- For any  $\rho > 0$ , define  $\rho$ -weighted generalization error (Catoni, 2007) as

$$\text{Err}_\rho := R - (1 + \rho)\hat{R}_n$$

and define  $\rho$ -shifted Rademacher random variable

$$\tilde{\rho}_i := (-1)^{1-U_i} - \frac{\rho}{\rho + 2}$$

## Lemma

$$\text{Err}_\rho = \frac{\rho + 2}{n} \sum_{i=1}^n \mathbb{E}_{L_i^+, \tilde{\rho}_i} \tilde{\rho}_i L_i^+$$

Such a perspective allows the development of fast-rate bounds.

## fast-rate MI bounds

### Theorem D: Fast-Rate MI Bounds

1. There exist  $\rho, \delta > 0$  such that  $\text{Err}_\rho \leq \sum_{i=1}^n \frac{I(L_i^+; U_i)}{\delta n}$
2.  $R \leq \hat{R}_n + \sum_{i=1}^n \frac{4I(L_i^+; U_i)}{n} + 4\sqrt{\sum_{i=1}^n \frac{\hat{R}_n I(L_i^+; U_i)}{n}}$ .
3. For interpolation algorithms,  $\text{Err} \leq \sum_{i=1}^n \frac{2I(L_i^+; U_i)}{n \ln 2}$ .

## variance-based fast-rate MI bound

- $L_S(W)$ : empirical risk of  $W$  on training sample  $S$
- For  $\gamma \in (0, 1)$ ,  $\gamma$ -variance is defined as

$$V(\gamma) := \mathbb{E}_{W, S} \frac{1}{n} \sum_{i=1}^n (\ell(W, Z_i) - (1 + \gamma)L_S(W))^2.$$

### Theorem E: Variance-Based MI Bound for 0/1-Loss

$$\text{There exist } \rho, \delta > 0 \text{ such that } \text{Err} \leq \rho V(\gamma) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{n \delta}$$

## sharpness-based fast-rate MI bounds

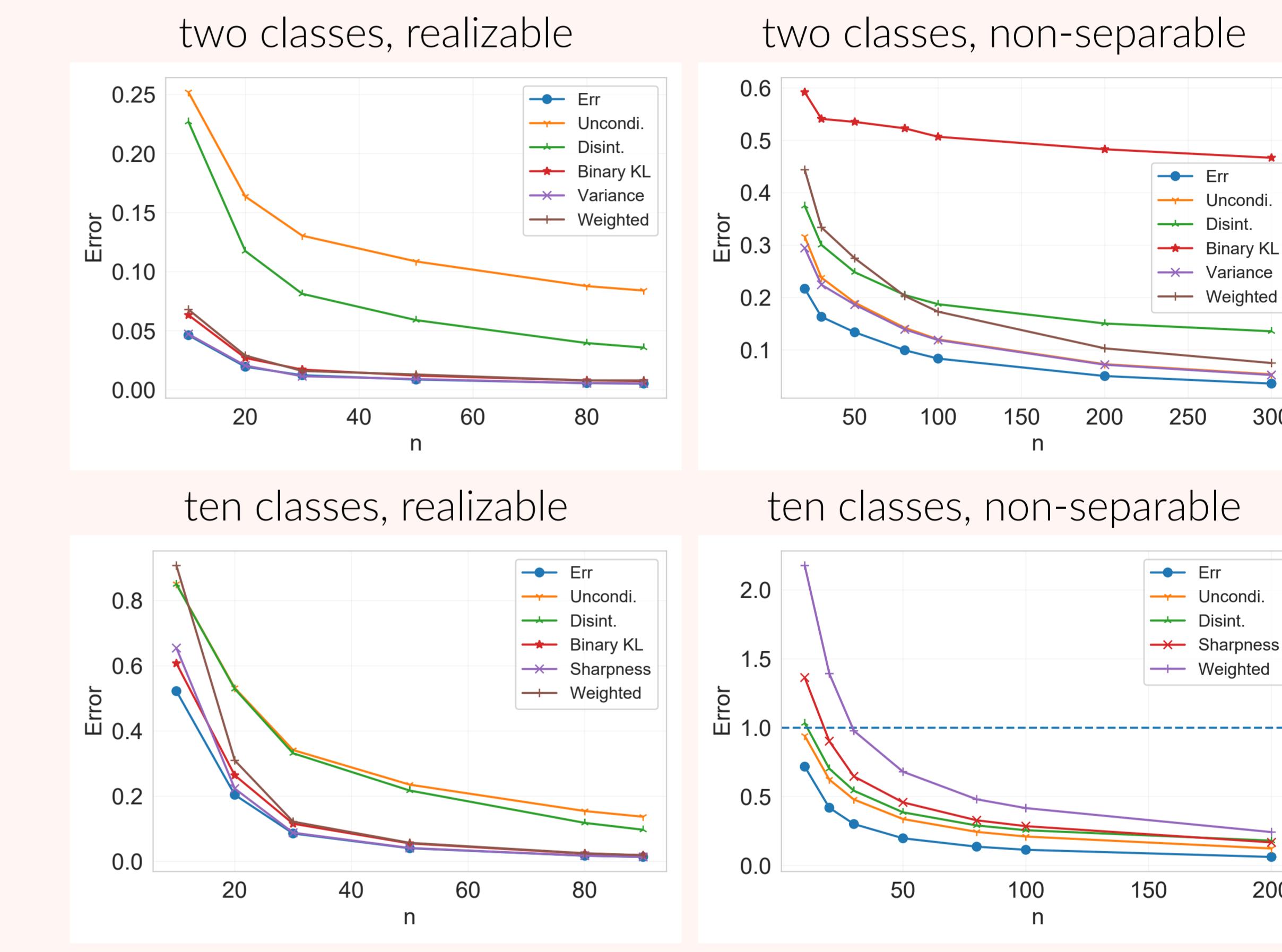
For any  $\lambda \in (0, 1)$ ,

- $F_i(\lambda) \triangleq \mathbb{E}_{W, Z_i} \ell(W, Z_i) - (1 + \lambda) \mathbb{E}_{W|Z_i} \ell(W, Z_i)^2$ .
- $\lambda$ -sharpness is defined as  $F(\lambda) = \frac{1}{n} \sum_{i=1}^n F_i(\lambda)$

### Theorem F: Sharpness-Based MI Bound for 0/1-Loss

$$\text{There exist } \rho, \delta > 0 \text{ such that } \text{Err} \leq \rho F(\lambda) + \sum_{i=1}^n \frac{I(L_i^+; U_i)}{\delta n}$$

## experimental comparison: linear classifier on synthetic data



## experimental comparison: neural networks on real data

