On the Surrogate Gap Between Contrastive and Supervised Losses

Han Bao

Yoshihiro Nagano

Kento Nozawa

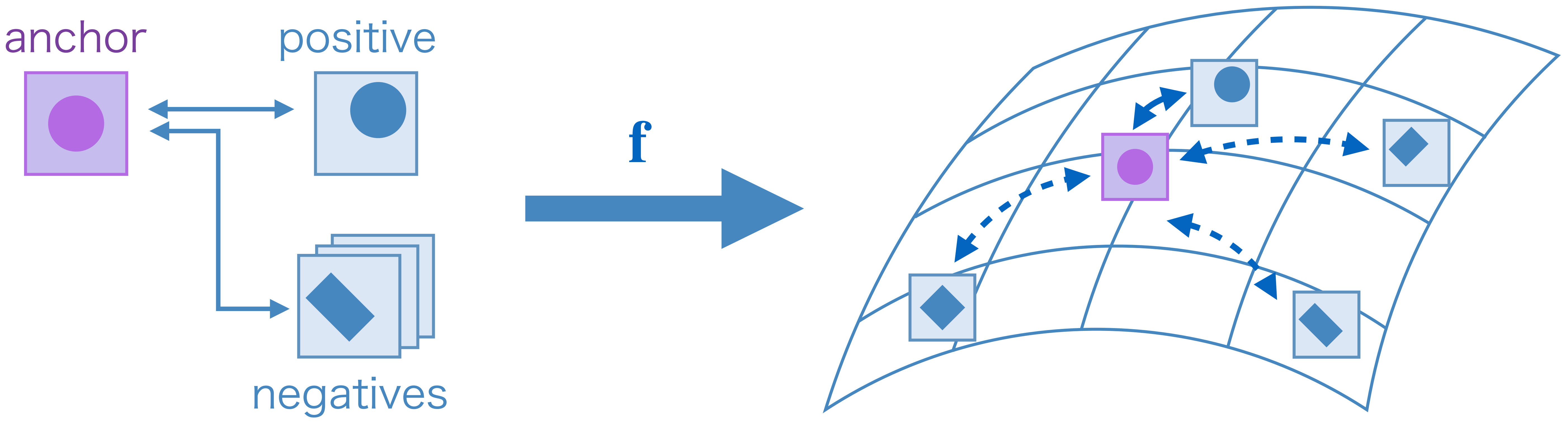
東京大学
THE UNIVERSITY OF TOKYO

RIKEN

(Authors are listed in alphabetical order)

Contrastive learning | Successful representation learning

- Learn a representation function $\mathbf{f}$ by making closer to positive/farther from negatives
    - ❖ Without labeled data

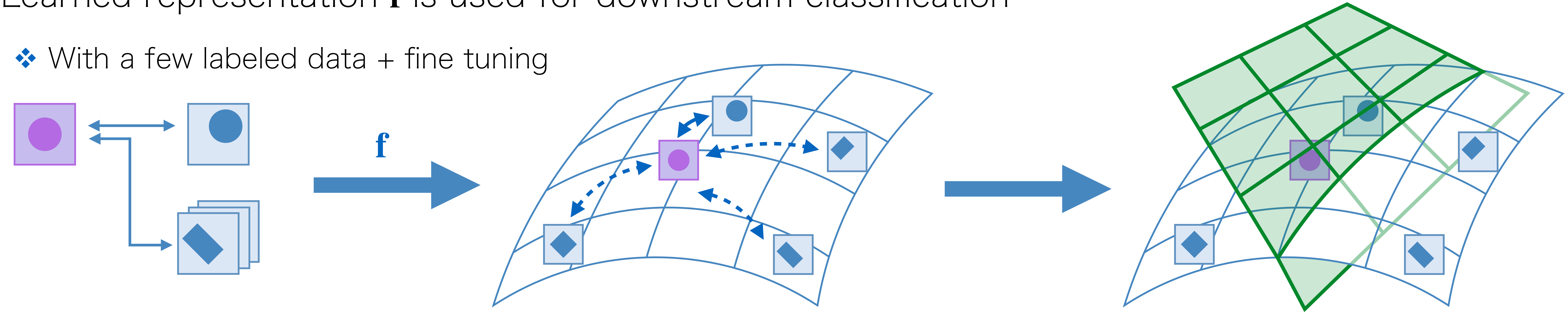


- Objective function: contrastive loss

$$R_{\text{cont}}(\mathbf{f}) = \mathbb{E}\left[-\ln \frac{\exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)\right)}{\exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}^+)\right) + \sum_{k \in [K]} \exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{f}(\mathbf{x}_k^-)\right)} \right]$$

Contrastive learning | Successful representation learning

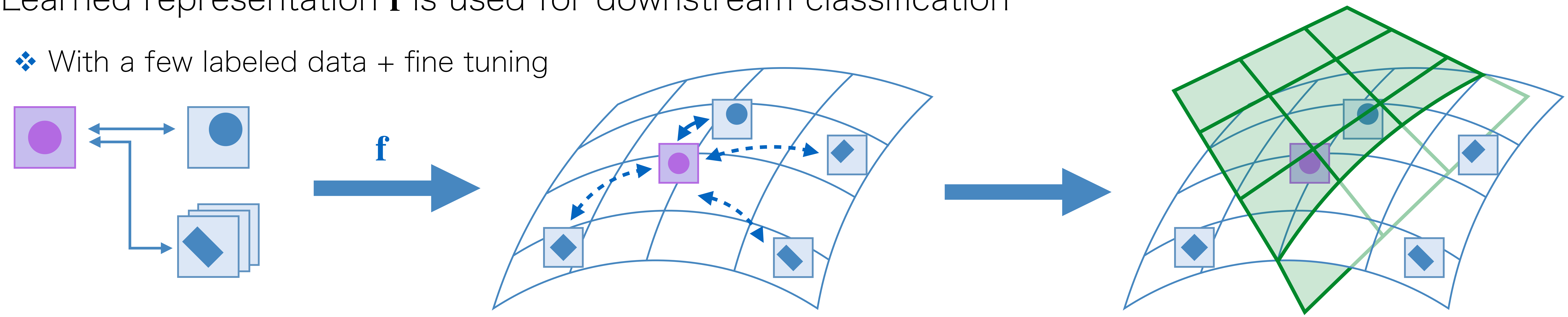
- Learned representation $\mathbf{f}$ is used for downstream classification
  - ❖ With a few labeled data + fine tuning



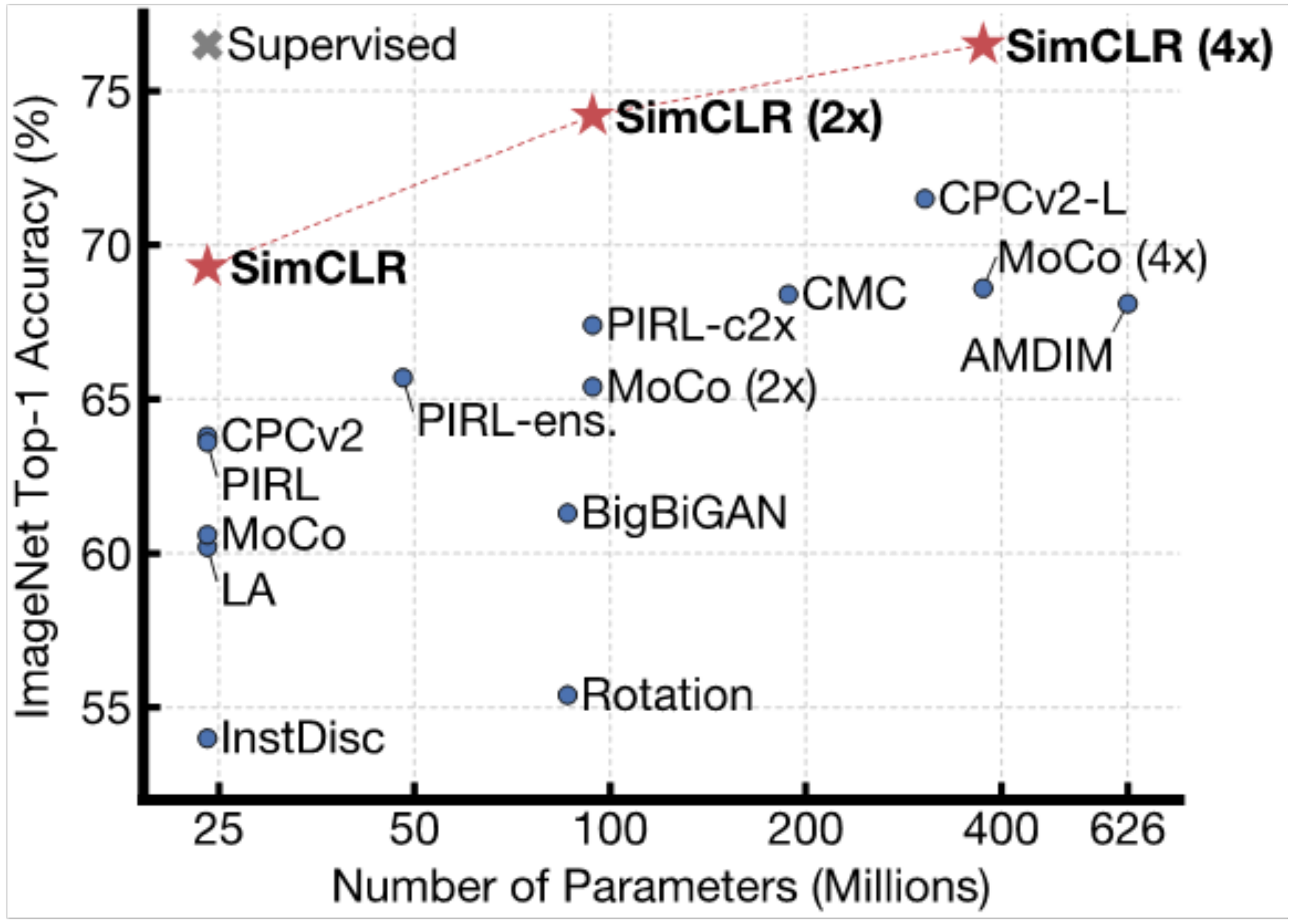
Chen et al. "A simple framework for contrastive learning of visual representations" (ICML2020)

Contrastive learning | Successful representation learning

- Learned representation $\mathbf{f}$ is used for downstream classification
    - ❖ With a few labeled data + fine tuning



- Good empirical classification performance [Chen *et al.*, 2020]
    - ❖ Linear classifier built upon the learned representation achieves accuracy close to complex supervised models

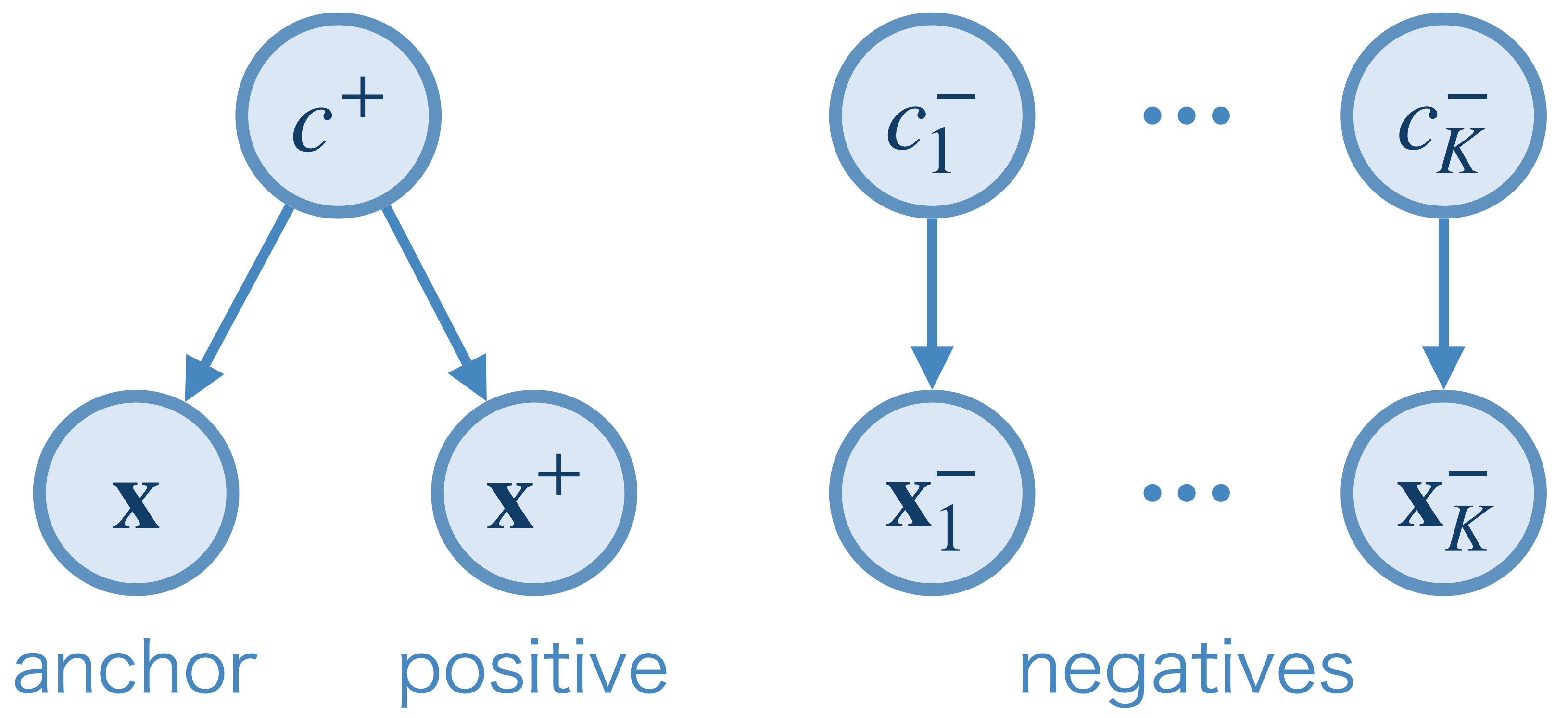


Q. What is the underlying mechanism of the success?

Existing theoretical analysis of contrastive learning

[Arora *et al.*, 2019]

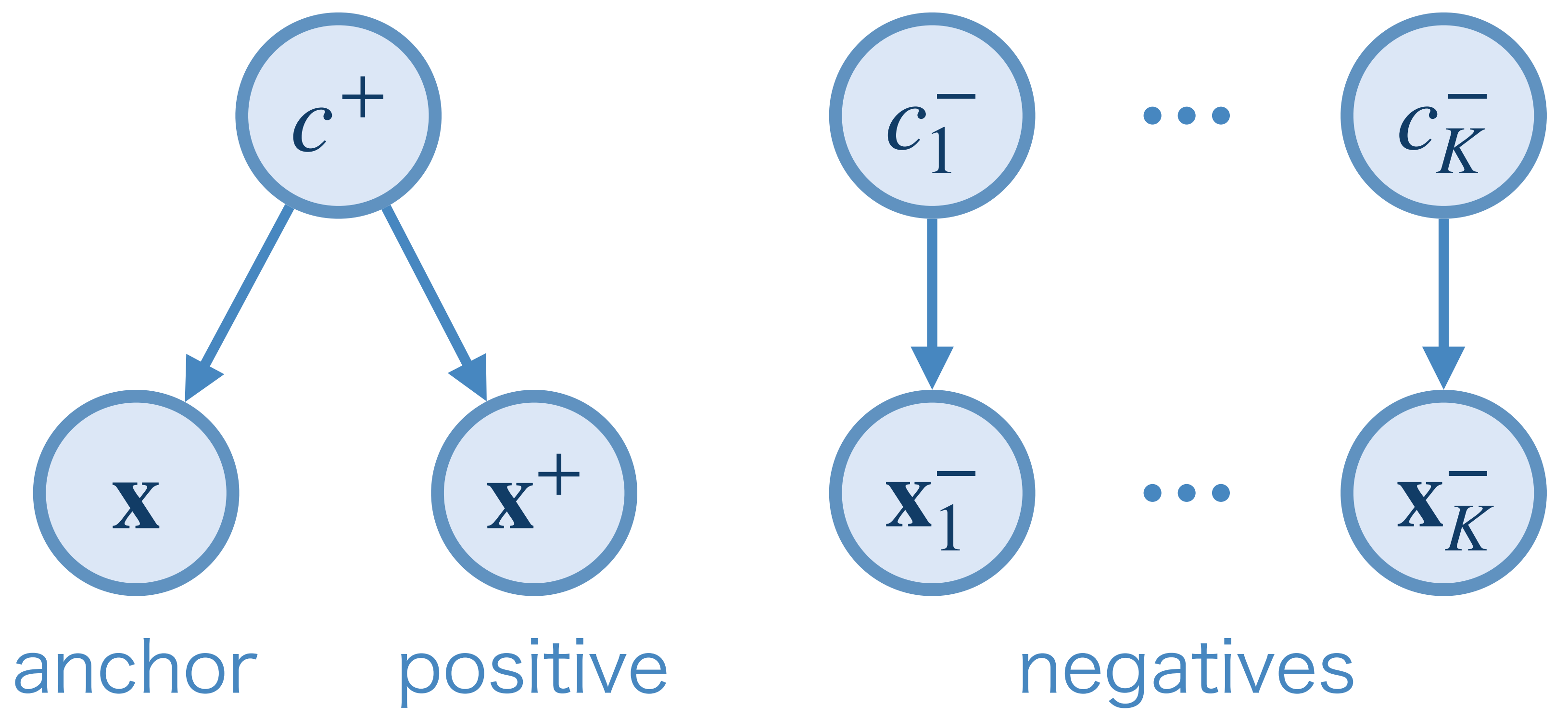
- Class set $\mathcal{Y} = \{1,2,\ldots,C\}$, the number of negative samples K
- Data generating process
  - Draw positive/negative classes $c^+, \{c_k^-\}_{k\in[K]} \sim \mathbb{P}(Y)$
  - Draw an anchor/positive sample $\mathbf{x}, \mathbf{x}^+ \sim \mathbb{P}(X \mid Y = c^+)$
  - Draw K negative samples $\mathbf{x}_k^- \sim \mathbb{P}(X \mid Y = c_k^-)$



Existing theoretical analysis of contrastive learning

[Arora *et al.*, 2019]

- Class set $\mathcal{Y} = \{1,2,\ldots,C\}$, the number of negative samples K
- Data generating process
  - Draw positive/negative classes $c^+, \{c_k^-\}_{k\in[K]} \sim \mathbb{P}(Y)$
  - Draw an anchor/positive sample $\mathbf{x}, \mathbf{x}^+ \sim \mathbb{P}(X \mid Y = c^+)$
  - Draw K negative samples $\mathbf{x}_k^- \sim \mathbb{P}(X \mid Y = c_k^-)$



- Result: contrastive loss $R_{\mathrm{cont}}(\mathbf{f})$ upper bounds downstream linear classification loss $R_{\mu-\mathrm{supv}}(\mathbf{f})$

$$R_{\mu-\mathrm{supv}}(\mathbf{f}) \leq \frac{1}{(1-\tau_K)v_{K+1}} \left\{ R_{\mathrm{cont}}(\mathbf{f}) - \mathbb{E}\log(\mathrm{Col}+1) \right\}$$

- τ_K : collision probability of positive class with negative classes
- v_{K+1} : coverage probability that negative classes contain every class

Issue | Disagreement of theory and practice!

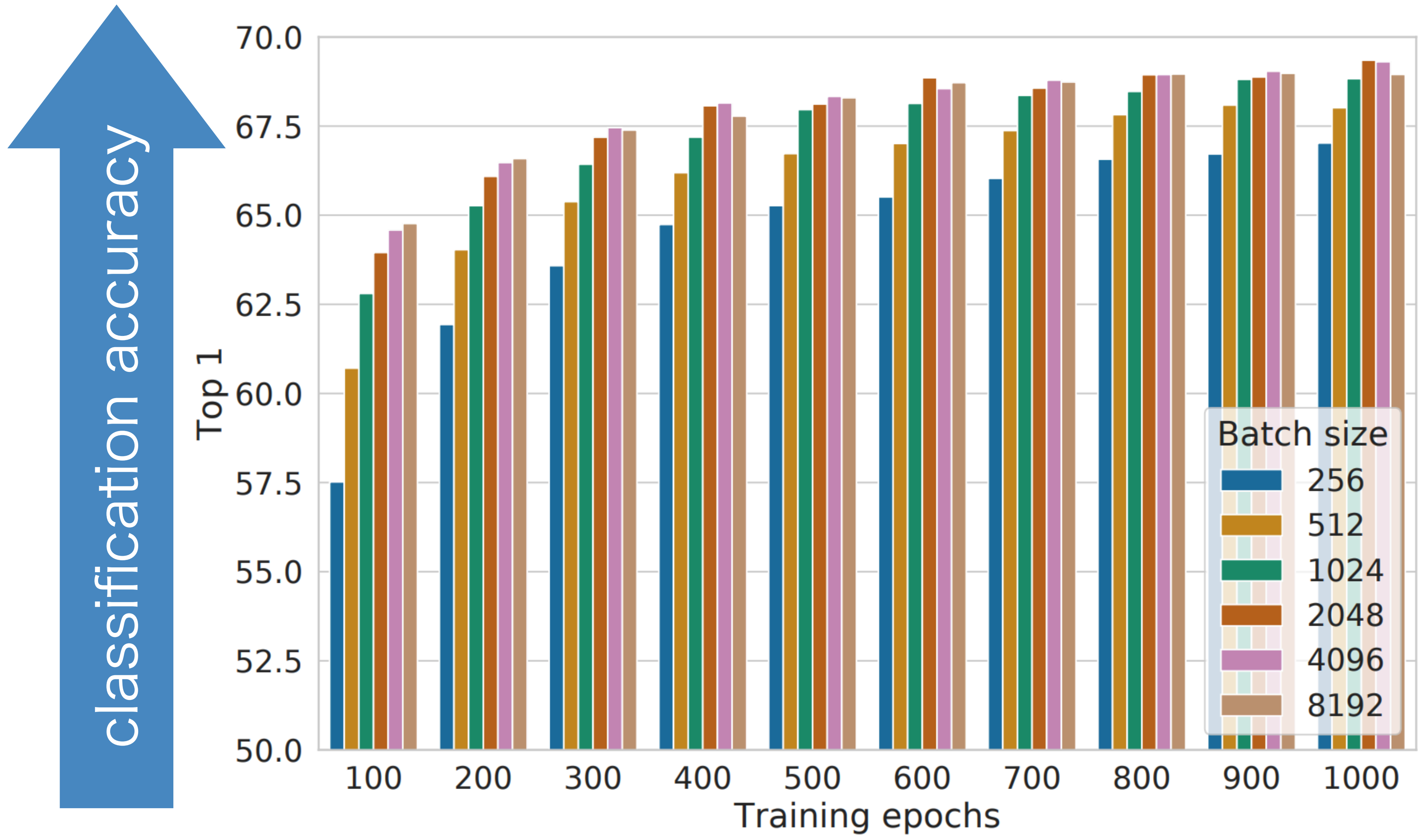
- Theory [Arora *et al.*, 2019]: larger K degrades downstream classification

$$R_{\mu-\text{supv}}(\mathbf{f}) \leq \boxed{\frac{1}{(1-\tau_K)v_{K+1}}} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$

$$O\left(e^K\right)$$

  - Upper bound becomes exponentially loose in K

- Practice [Chen *et al.*, 2020]: larger K improves downstream classification
  - Classification accuracy improves as K (= batch size) increases



Chen et al. "A simple framework for contrastive learning of visual representations" (ICML2020)
Arora et al. "A Theoretical Analysis of Contrastive Unsupervised Representation Learning" (ICML2019)

Our result: much better upper & lower bounds

- Existing bound $$R_{\mu-\text{supv}}(\mathbf{f}) \leq \boxed{O\left(e^{K}\right)} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$

Our result: much better upper & lower bounds

- Existing bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le \boxed{O\left(e^K\right)} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$
- Our upper bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$

Our result: much better upper & lower bounds

- Existing bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le \boxed{O\left(e^{K}\right)} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$

- Our upper bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$

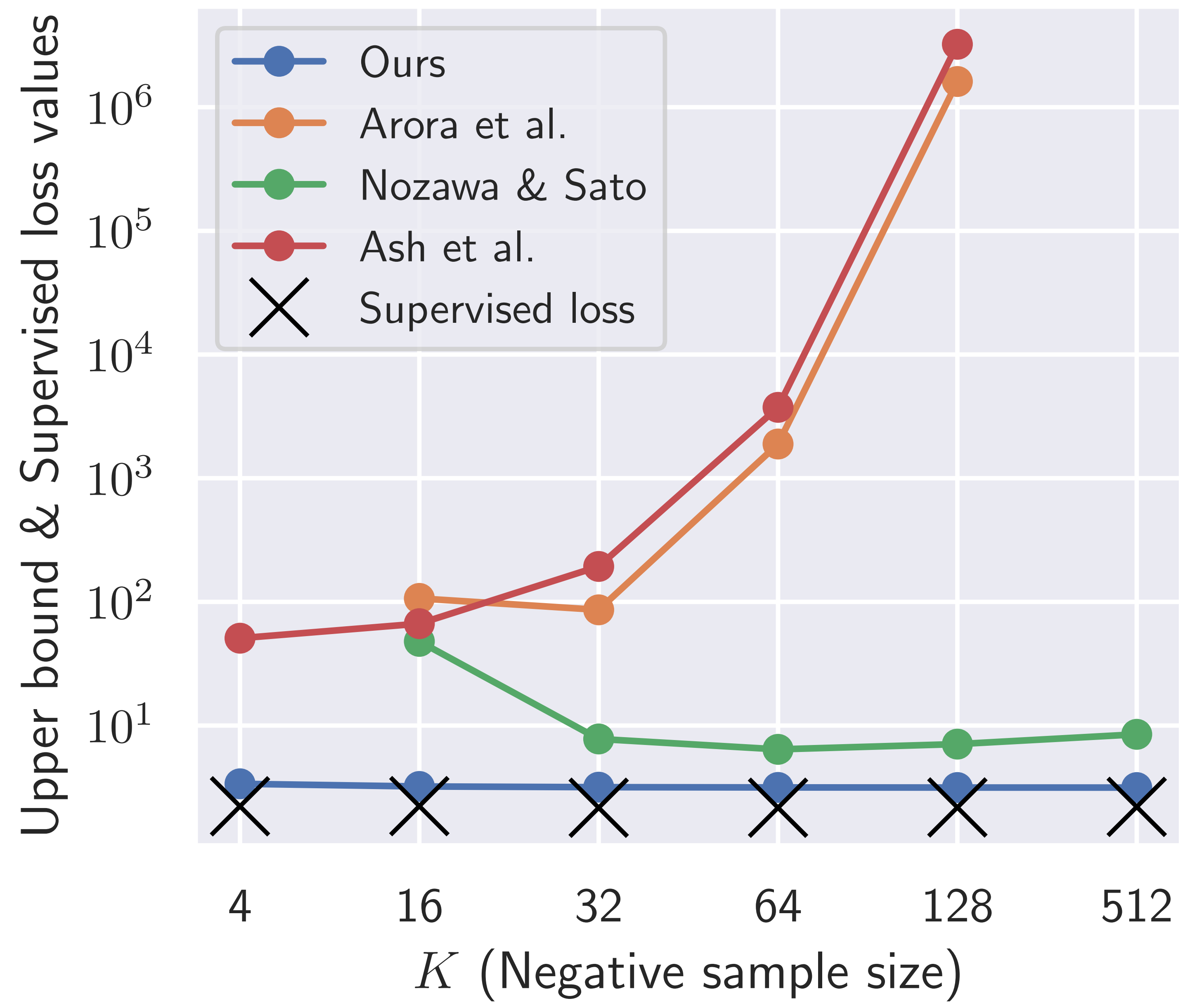
- Our lower bound $$R_{\mu-\text{supv}}(\mathbf{f}) \ge R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$

  - Proof sketch: linearize log-sum-exp function of both R_{cont} and $R_{\mu-\text{supv}}$

Our result: much better upper & lower bounds

- Existing bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le \boxed{O\left(e^K\right)} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$
- Our upper bound $$R_{\mu-\text{supv}}(\mathbf{f}) \le R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$
- Our lower bound $$R_{\mu-\text{supv}}(\mathbf{f}) \ge R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$

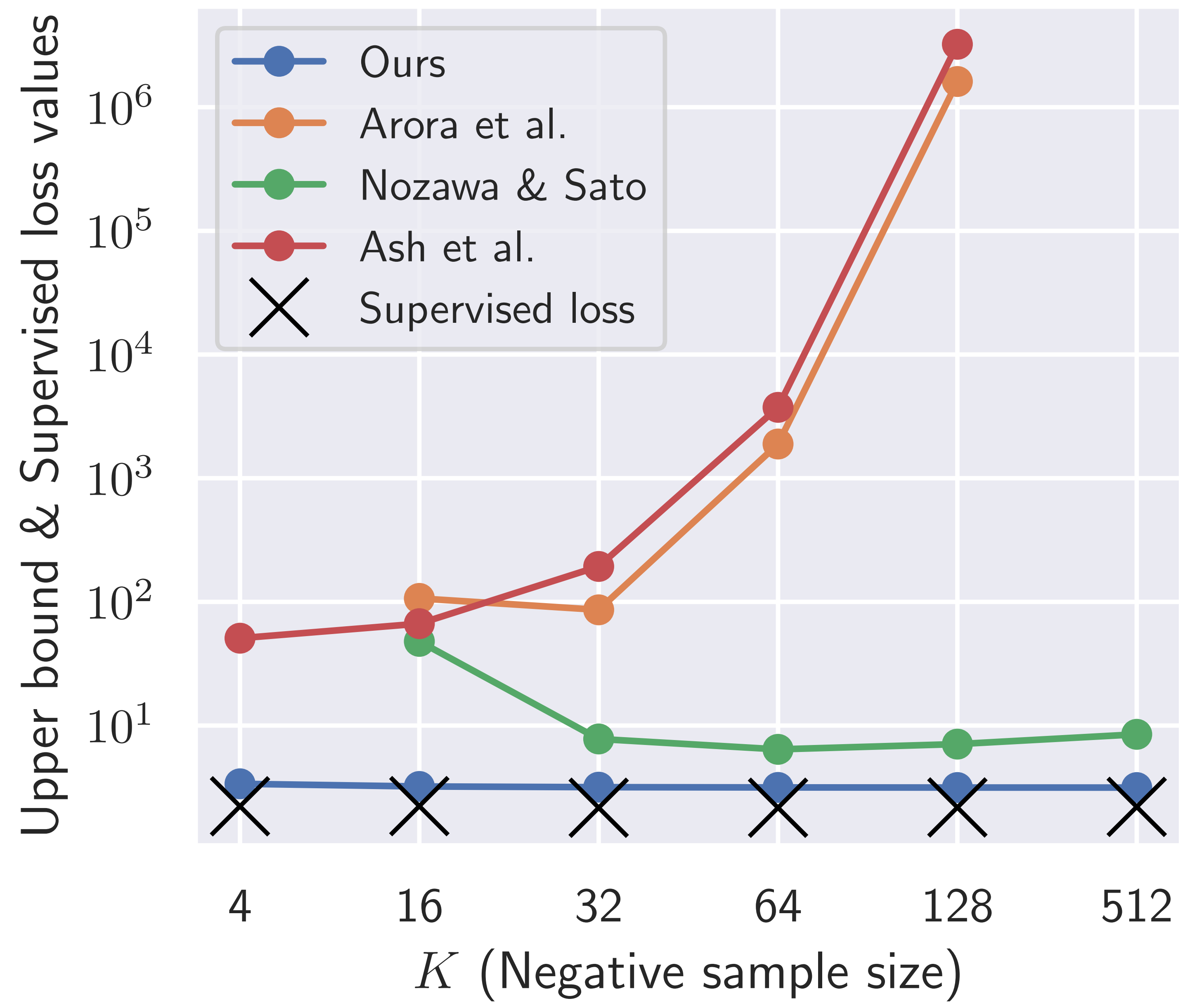
❖ Proof sketch: linearize log-sum-exp function of both R_{cont} and $R_{\mu-\text{supv}}$



Our result: much better upper & lower bounds

- Existing bound $$R_{\mu-\text{supv}}(\mathbf{f}) \leq \boxed{O\left(e^K\right)} \left\{R_{\text{cont}}(\mathbf{f}) - \mathbb{E}\log(\text{Col}+1)\right\}$$
- Our upper bound $$R_{\mu-\text{supv}}(\mathbf{f}) \leq R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$
- Our lower bound $$R_{\mu-\text{supv}}(\mathbf{f}) \geq R_{\text{cont}}(\mathbf{f}) + O\left(\ln\frac{1}{K}\right)$$

❖ Proof sketch: linearize log-sum-exp function of both R_{cont} and $R_{\mu-\text{supv}}$



Message: our bounds suggest that larger K is indeed good even in theory!