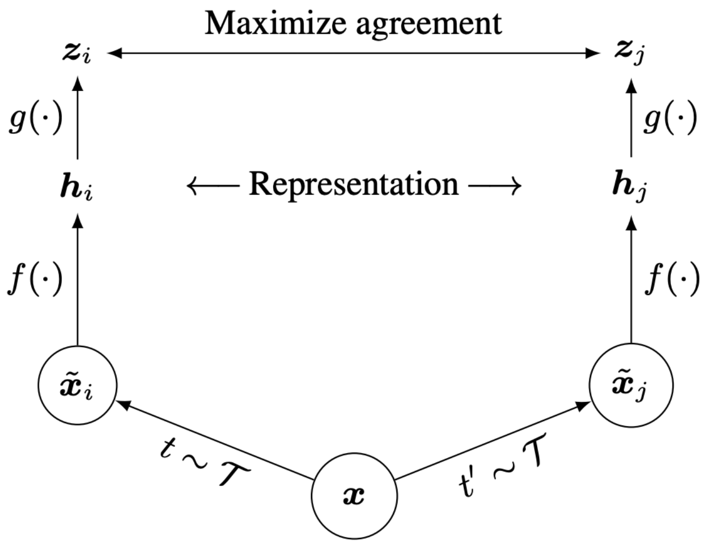
中国科学技术大学
University of Science and Technology of China

Understanding Contrastive Learning via Distributionally Robust Optimization

Authors: Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang & Xiangnan He
Paper: https://arxiv.org/pdf/2310.11048.pdf
Lab: USTC Lab for Data Science
Code: https://github.com/junkangwu/ADNCE
Date: Oct 31, 2023

Background and Motivation

- The core idea of CL is to learn representations that draw positive samples nearby and push away negative samples.



- Loss function : InfoNCE

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

[1] Chen, Ting, et al. A simple framework for contrastive learning of visual representations. ICML2020

Background and Motivation

❑ Sampling bias leads to performance drop:

- ➢ Negative counterparts are commonly drawn uniformly from the training data.
- ➢ True labels or true semantic similarity are typically **not available** …

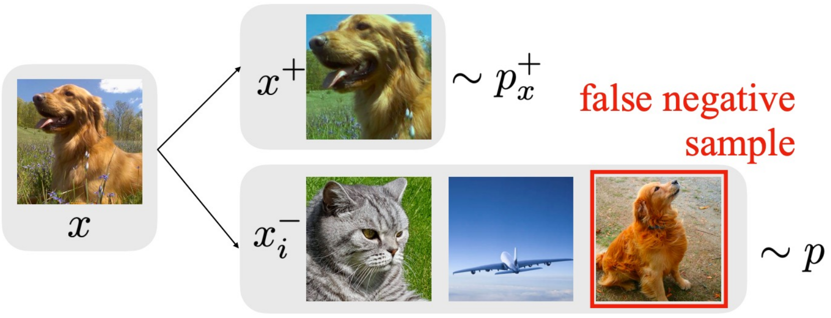


Figure 1: **"Sampling bias":** The common practice of drawing negative examples x_i^- from the data distribution $p(x)$ may result in x_i^- that are actually similar to x.

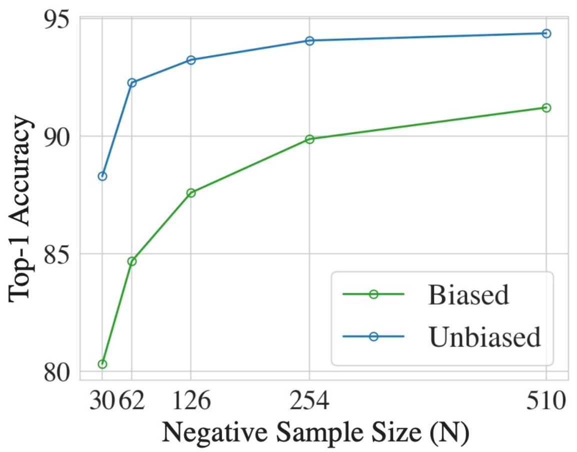


Figure 2: **Sampling bias leads to performance drop:** Results on CIFAR-10 for drawing x_i^- from $p(x)$ (biased) and from data with different labels, i.e., truly semantically different data (unbiased).

[1] Chuang, Ching-Yao, et al. "Debiased contrastive learning." NeurIPS2020

Background and Motivation

❑ Motivation: InfoNCE has the ability to mitigate sampling bias.

- By **fine-tuning** the temperature τ, basic SimCLR demonstrates significant improvement
- With an appropriately **selected** τ, the relative improvements realized by DCL[1] and HCL[2] are marginal.

Model	CIFAR10		STL10	
	Top-1	τ	Top-1	τ
SimCLR(τ_0)	91.10	0.5	81.05	0.5
SimCLR(τ^*)	92.19	0.3	87.91	0.2
DCL(τ_0)	92.00 (-0.2%)	0.5	84.26 (-4.2%)	0.5
DCL(τ^*)	92.09 (-0.1%)	0.3	88.20 (+1.0%)	0.2
HCL(τ_0)	92.12 (-0.0%)	0.5	87.44 (-0.5%)	0.5
HCL(τ^*)	92.10 (-0.0%)	0.3	87.46 (-0.5%)	0.2

[1] Ching-Yao Chuang et. al,Debiased contrastive learning. In NeurIPS 2020.
[2] Joshua David Robinson et al. Contrastive learning with hard negative samples. In ICLR 2021.

Background and Motivation

❑ Motivation: InfoNCE has the ability to mitigate sampling bias.

- By **fine-tuning** the temperature τ, basic SimCLR demonstrates significant improvement
- With an appropriately **selected** τ, the relative improvements realized by DCL[1] and HCL[2] are marginal.

Model	CIFAR10		STL10	
	Top-1	τ	Top-1	τ
SimCLR(τ_0)	91.10	0.5	81.05	0.5
SimCLR(τ^*)	92.19	0.3	87.91	0.2
DCL(τ_0)	92.00 (-0.2%)	0.5	84.26 (-4.2%)	0.5
DCL(τ^*)	92.09 (-0.1%)	0.3	88.20 (+1.0%)	0.2
HCL(τ_0)	92.12 (-0.0%)	0.5	87.44 (-0.5%)	0.5
HCL(τ^*)	92.10 (-0.0%)	0.3	87.46 (-0.5%)	0.2

[1] Ching-Yao Chuang et. al,Debiased contrastive learning. In NeurIPS 2020.
[2] Joshua David Robinson et al. Contrastive learning with hard negative samples. In ICLR 2021.

Background and Motivation

❑ Motivation: InfoNCE has the ability to mitigate sampling bias.

- By **fine-tuning** the temperature τ, basic SimCLR demonstrates significant improvement
- With an appropriately **selected** τ, the relative improvements realized by DCL[1] and HCL[2] are marginal.

1) Why does CL exhibit tolerance to sampling bias?
2) What role does τ play, and why is it so important?

[1] Ching-Yao Chuang et. al,Debiased contrastive learning. In NeurIPS 2020.
[2] Joshua David Robinson et al. Contrastive learning with hard negative samples. In ICLR 2021.

Understanding CL from DRO

❑ Preliminary

- DRO aims to minimize the worst-case expected loss over a set of potential distributions.

$$\mathcal{L}_{\text{DRO}} = \max_{Q} \mathbb{E}_Q[\mathcal{L}(x;\theta)] \qquad s.t.\ D_\phi(Q||Q_0) \le \eta,$$

- L_{basic} aims to increase the embedding similarity between the positive instances and decreases that of the negative ones.

$$L_{\text{basic}} \doteq -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \mathbb{E}_{Q_0}[f_\theta(x,y)]\right]$$

- CL-DRO improves L_{basic} by incorporating DRO on the negative side.

$$\mathcal{L}^{\phi}_{\text{CL-DRO}} = -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \max_{Q} \mathbb{E}_Q[f_\theta(x,y)]\right] \qquad s.t.\ D_\phi(Q||Q_0) \le \eta. \qquad (3)$$

Understanding CL from DRO

❑ Understanding CL from DRO

Theorem 3.2. *By choosing KL divergence $D_{KL}(Q||Q_0) = \int Q \log \frac{Q}{Q_0} dx$, optimizing CL-DRO* (cf. *Equation* (3)) *is equivalent to optimizing CL (InfoNCE,*cf. *Equation* (1)*):*

$$
\begin{aligned}
\mathcal{L}^{KL}_{\text{CL-DRO}} &= -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \min_{\alpha\geq 0,\beta}\max_{Q\in\mathbb{Q}}\{\mathbb{E}_Q[f_\theta(x,y)] - \alpha[D_{KL}(Q||Q_0)-\eta] + \beta(\mathbb{E}_{Q_0}[\frac{Q}{Q_0}]-1)\}\right] \\
&= -\mathbb{E}_{P_X}\mathbb{E}_{P_0}\left[\alpha^*(\eta)\log\frac{e^{f_\theta(x,y^+)/\alpha^*(\eta)}}{\mathbb{E}_{Q_0}[e^{f_\theta(x,y)/\alpha^*(\eta)}]}\right] + Constant \\
&= \alpha^*(\eta)\mathcal{L}_{\text{InfoNCE}} + Constant,
\end{aligned}
\tag{4}
$$

where α, β represent the Lagrange multipliers, and $\alpha^*(\eta)$ signifies the optimal value of α that minimizes the Equation (4)*, serving as the temperature τ in CL.*

The DRO enables CL to perform well across various potential distributions and thus equips it with the capacity to **alleviate sampling bias.**

Understanding CL from DRO

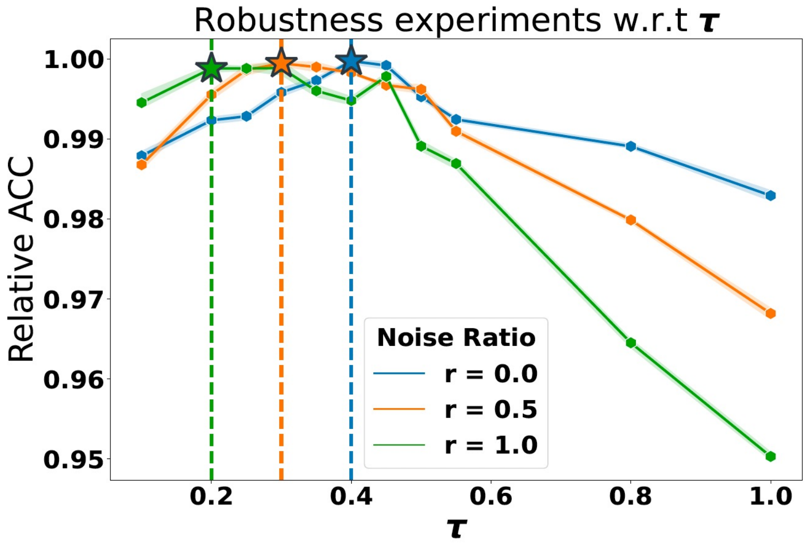
The role of τ

- Adjusting robust radius

Corollary 3.4. *[The optimal α - Lemma 5 of Faury et al. [46]] The value of the optimal α (i.e., τ) can be approximated as follow:*

$$\tau \approx \sqrt{\mathbb{V}_{Q_0}[f_\theta(x,y)]/2\eta}, \tag{6}$$

where $\mathbb{V}_{Q_0}[f_\theta(x,y)]$ denotes the variance of $f_\theta(x,y)$ under the distribution Q_0.



1. **There is an evident trade-off in the selection of τ.**
2. **As the ratio of false negative instances increases (r ranges from 0 to 1), the robustness radius increases and the optimal τ decreases.**

Understanding CL from DRO

❑ The role of τ

➢ Controlling variance of negative samples

Theorem 3.5. *Given any ϕ-divergence, the corresponding CL-DRO objective could be approximated as a mean-variance objective:*

$$\mathcal{L}^{\phi}_{CL\text{-}DRO} \approx -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \left(\mathbb{E}_{Q_0}[f_\theta(x,y)] + \frac{1}{2\tau}\frac{1}{\phi^{(2)}(1)} \cdot \mathbb{V}_{Q_0}[f_\theta(x,y)]\right)\right], \tag{7}$$

where $\phi^{(2)}(1)$ denotes the the second derivative value of $\phi(\cdot)$ at point 1, and $\mathbb{V}_{Q_0}[f_\theta]$ denotes the variance of f under the distribution Q_0.

Specially, if we consider KL divergence, the approximation transforms:

$$\mathcal{L}^{KL}_{CL\text{-}DRO} \approx -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \left(\mathbb{E}_{Q_0}[f_\theta(x,y)] + \frac{1}{2\tau}\mathbb{V}_{Q_0}[f_\theta(x,y)]\right)\right]. \tag{8}$$

➢ Hard-mining.

$$\mathcal{L}^{KL}_{CL\text{-}DRO} = -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \min_{\alpha\geq 0,\beta}\max_{Q\in\mathbb{Q}}\{\mathbb{E}_Q[f_\theta(x,y)] - \alpha[D_{KL}(Q||Q_0) - \eta] + \beta(\mathbb{E}_{Q_0}[\frac{Q}{Q_0}] - 1)\}\right]$$

$$\Downarrow$$

$$Q^* = \frac{e^{\frac{f_\theta}{\alpha^*}}}{\mathbb{E}_{Q_0}[e^{\frac{f_\theta}{\alpha^*}}]}Q_0$$

Understanding CL from DRO

❑ The role of τ

- ➢ Controlling variance of negative samples

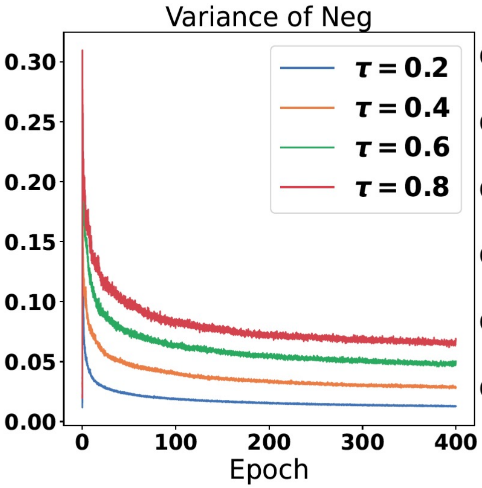
Theorem 3.5. *Given any ϕ-divergence, the corresponding CL-DRO objective could be approximated as a mean-variance objective:*

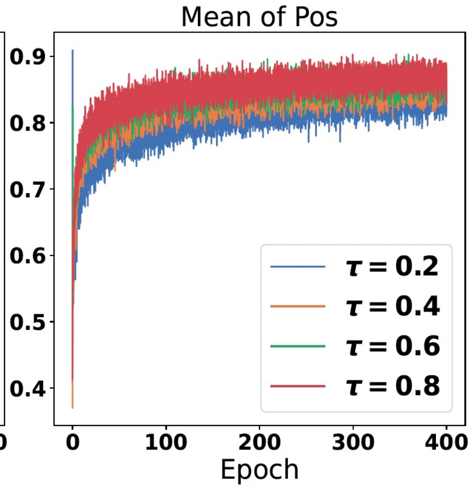
$$\mathcal{L}^{\phi}_{CL\text{-}DRO} \approx -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \left(\mathbb{E}_{Q_0}[f_\theta(x,y)] + \frac{1}{2\tau}\frac{1}{\phi^{(2)}(1)} \cdot \mathbb{V}_{Q_0}[f_\theta(x,y)]\right)\right], \quad (7)$$

where $\phi^{(2)}(1)$ denotes the the second derivative value of $\phi(\cdot)$ at point 1, and $\mathbb{V}_{Q_0}[f_\theta]$ denotes the variance of f under the distribution Q_0.

Specially, if we consider KL divergence, the approximation transforms:

$$\mathcal{L}^{KL}_{CL\text{-}DRO} \approx -\mathbb{E}_{P_X}\left[\mathbb{E}_{P_0}[f_\theta(x,y^+)] - \left(\mathbb{E}_{Q_0}[f_\theta(x,y)] + \frac{1}{2\tau}\mathbb{V}_{Q_0}[f_\theta(x,y)]\right)\right]. \quad (8)$$





Model	Cifar10		Stl10	
	Top-1	τ	Top-1	τ
SimCLR(τ_0)	91.10	0.5	81.05	0.5
SimCLR(τ^*)	**92.19**	0.3	**87.91**	0.2
MW	91.81	0.3	87.24	0.2

DRO, InfoNCE & Mutual Information

❑ Relations among DRO, InfoNCE and Mutual Information

Definition 4.1 (ϕ-MI)**.** The ϕ-divergence-based mutual information is defined as:

$$I_\phi(X;Y) = D_\phi(P(X,Y)||P(X)P(Y)) = \mathbb{E}_{P_X}[D_\phi(P_0||Q_0)]. \quad (9)$$

Theorem 4.2. *For distributions P, Q such that $P \ll Q$, let $\mathcal{F}$ be a set of bounded measurable functions. Let CL-DRO draw positive and negative instances from P and Q, marked as $\mathcal{L}^{\phi}_{\text{CL-DRO}}(P,Q)$. Then the CL-DRO objective is the tight variational estimation of ϕ-divergence. In fact, we have:*

$$D_\phi(P||Q) = \max_{f\in\mathcal{F}} -\mathcal{L}^{\phi}_{\text{CL-DRO}}(P,Q) = \max_{f\in\mathcal{F}} \mathbb{E}_P[f] - \min_{\lambda\in\mathbb{R}}\{\lambda + \mathbb{E}_Q[\phi^*(f-\lambda)]\}. \quad (10)$$

Here, the choice of ϕ in CL-DRO corresponds to the probability measures in $D_\phi(P||Q)$. And ϕ^* denotes the convex conjugate.

DRO, InfoNCE & Mutual Information

❑ Relations among DRO, InfoNCE and Mutual Information

- InfoNCE is a tighter MI estimation.[1]

$$D_\phi(P||Q) := \max_{f\in\mathcal{F}}\{\mathbb{E}_P[f] - \mathbb{E}_Q[\phi^*(f)]\}.$$

$$\Downarrow$$

$$D_\phi(P||Q) := \max_{f\in\mathcal{F}}\{\mathbb{E}_P[f] - \min_{\lambda\in\mathbb{R}}\{\lambda + \mathbb{E}_Q[\phi^*(f-\lambda)]\}\}$$

- DRO bridges the gap between MI and InfoNCE

"MINE uses a critic in Donsker-Varadhan target to derive a bound that is **neither an upper nor lower bound** on MI, while CPC relies on **unnecessary approximations** in its proof, resulting in some redundant approximations"

- DRO provides general MI estimation.

[1] Avraham Ruderman, et al. Tighter variational representations of f-divergences via restriction to probability measures. In ICML 2012.
[2] Ben Poole, etl al. On variational bounds of mutual information. In ICML 2019.

Method

❑ Shortcomings of InfoNCE

- Too conservative: overemphasizing on the hardest negative samples.
- Sensitive to outliers: DRO's weakness.

❑ Adjusted InfoNCE (ADNCE)

Our goal is to refine the worst-case distribution, aiming to assign more **reasonable** weights to negative instances.

$$w(f_\theta(x,y),\mu,\sigma) \propto \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{f_\theta(x,y)-\mu}{\sigma})^2], \tag{11}$$

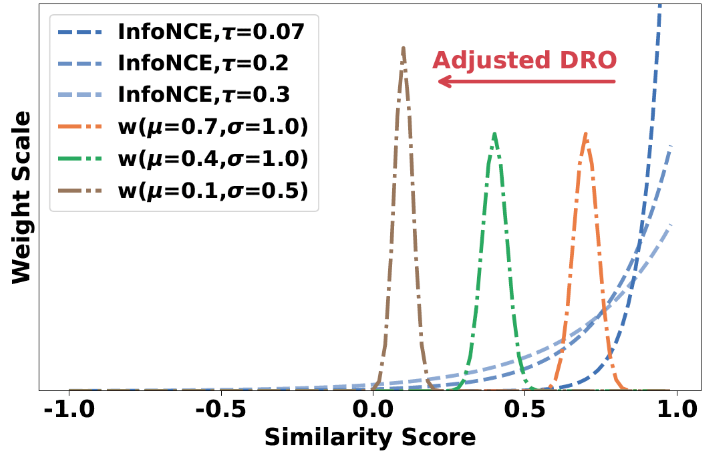


Figure 3: We visualize the training weight of negative samples *w.r.t.* similarity score. InfoNCE (in BlUE) over-emphasize hard negative samples, while ADNCE utilizes the weight w (in ORANGE, GREEN, BROWN) to adjust the distribution.

$$\mathcal{L}_{\text{ADNCE}} = -\mathbb{E}_P[f_\theta(x,y^+)/\tau] + \log \mathbb{E}_{Q_0}[w(f_\theta(x,y),\mu,\sigma)e^{f_\theta(x,y)/\tau}/Z_{\mu,\sigma}],$$

Experiments

❑ Images

Model	CIFAR10				STL10				CIFAR100			
	100	200	300	400	100	200	300	400	100	200	300	400
InfoNCE (τ_0)	85.70	89.21	90.33	91.01	75.95	78.47	80.39	81.67	59.10	63.96	66.03	66.53
InfoNCE (τ^*)	86.54	89.82	91.18	91.64	81.20	84.77	86.27	87.69	62.32	66.85	68.31	69.03
α-CL-direct	87.65	90.11	90.88	91.24	80.91	84.71	87.01	87.96	62.75	66.27	67.35	68.54
ADNCE	**87.67**	**90.65**	**91.42**	**91.88**	**81.77**	**85.10**	**87.01**	**88.00**	**62.79**	**66.89**	**68.65**	**69.35**

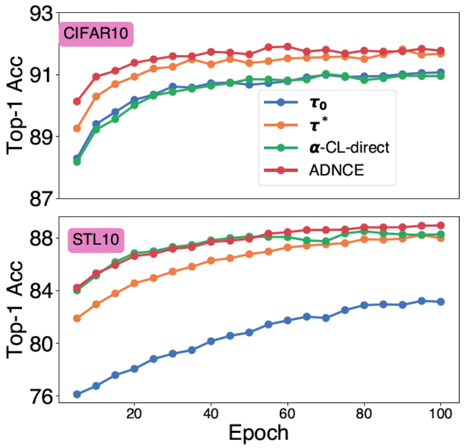


Figure 4: Learning curve for Top-1 accuracy by linear evaluation on CIFAR10 and STL10.

1. ADNCE exhibits sustained improvement and notably enhances performance in the early stages of training.
2. Training curve to further illustrate the stable superiority of ADNCE.

Experiments

❑ Sentences and Graphs

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\mathrm{BERT}_{\mathrm{base}}$-flow♣	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\mathrm{BERT}_{\mathrm{base}}$-whitening♣	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
CT-$\mathrm{BERT}_{\mathrm{base}}$♣	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
SimCSE-$\mathrm{BERT}_{\mathrm{base}}(\tau_0)$	68.40	**82.41**	74.38	80.91	78.56	76.85	**72.23**	76.25
SimCSE-$\mathrm{BERT}_{\mathrm{base}}(\tau^*)$	<u>71.37</u>	81.18	<u>74.41</u>	**82.51**	<u>79.24</u>	<u>78.26</u>	70.65	76.81
ADNCE-$\mathrm{BERT}_{\mathrm{base}}$	**71.38**	<u>81.58</u>	**74.43**	<u>82.37</u>	**79.31**	**78.45**	<u>71.69</u>	**77.03**
SimCSE-$\mathrm{RoBERTa}_{\mathrm{base}}(\tau_0)$	**70.16**	81.77	73.24	81.36	80.65	80.22	68.56	76.57
SimCSE-$\mathrm{RoBERTa}_{\mathrm{base}}(\tau^*)$	68.20	**81.95**	<u>73.63</u>	<u>81.83</u>	<u>81.55</u>	<u>80.96</u>	<u>69.56</u>	<u>76.81</u>
ADNCE-$\mathrm{RoBERTa}_{\mathrm{base}}$	<u>69.22</u>	<u>81.86</u>	**73.75**	**82.88**	**81.88**	**81.13**	**69.57**	**77.10**

Table 5: **Self-supervised representation learning on TUDataset**: The baseline results are excerpted from the published papers.

Methods	RDT-B	NCI1	PROTEINS	DD
node2vec	-	54.9±1.6	57.5±3.6	-
sub2vec	71.5±0.4	52.8±1.5	53.0±5.6	-
graph2vec	75.8±1.0	73.2±1.8	73.3±2.1	-
InfoGraph	82.5±1.4	76.2±1.1	74.4±0.3	72.9±1.8
JOAO	85.3±1.4	78.1±0.5	74.6±0.4	77.3±0.5
JOAOv2	86.4±1.5	78.4±0.5	74.1±1.1	77.4±1.2
RINCE	90.9±0.6	78.6±0.4	74.7±0.8	78.7±0.4
GraphCL (τ_0)	89.5±0.8	77.9±0.4	74.4±0.5	78.6±0.4
GraphCL (τ^*)	90.7±0.6	79.2±0.3	74.7±0.6	78.5±1.0
ADNCE	**91.4±0.3**	**79.3±0.7**	**75.1±0.6**	**79.2±0.6**

1. The improvements of τ^* over τ_0 emphasize the significance of selecting a **proper** robustness radius.
2. ADNCE **outperforms** all baselines with a significant margin on four datasets

Summary

- We provide a novel perspective on contrastive learning (CL) via the lens of Distributionally Robust Optimization (DRO)
  - key insights about the tolerance to sampling bias
  - the role of τ
  - the theoretical connection between DRO and MI
- We propose a novel CL loss— ADNCE
  - alleviate over-conservatism and sensitivity to outliers.

Thanks