Combating the Instability of Mutual Information-based Losses via Regularization

Kwanghee Choi* [1] juice500@sogang.ac.kr Siyeong Lee* [2] siyeong.lee@naverlabs.com *Equal contributions.

[1] Sogang University [2] NAVER LABS

Summary

We propose a novel dual representation of the KL divergence which regularizes the existing representations to mitigate the instability of MI estimators during optimization.

- We identify the symptoms of instability in the MI-based losses: (1) neural network's weights failing to converge even after the loss converges, and (2) saturating neural network outputs causing the loss to diverge.
- To evaluate MI-based losses on real-world settings, we present two novel benchmarks based on the concept of supervised and contrastive learning, verifying both MI estimation power and its capability on downstream tasks.
- We demonstrate that the proposed regularization works effectively for six different MI-based losses on multiple benchmarks.

Two Symptoms of Instability

Symptom 1: Drifting Problem

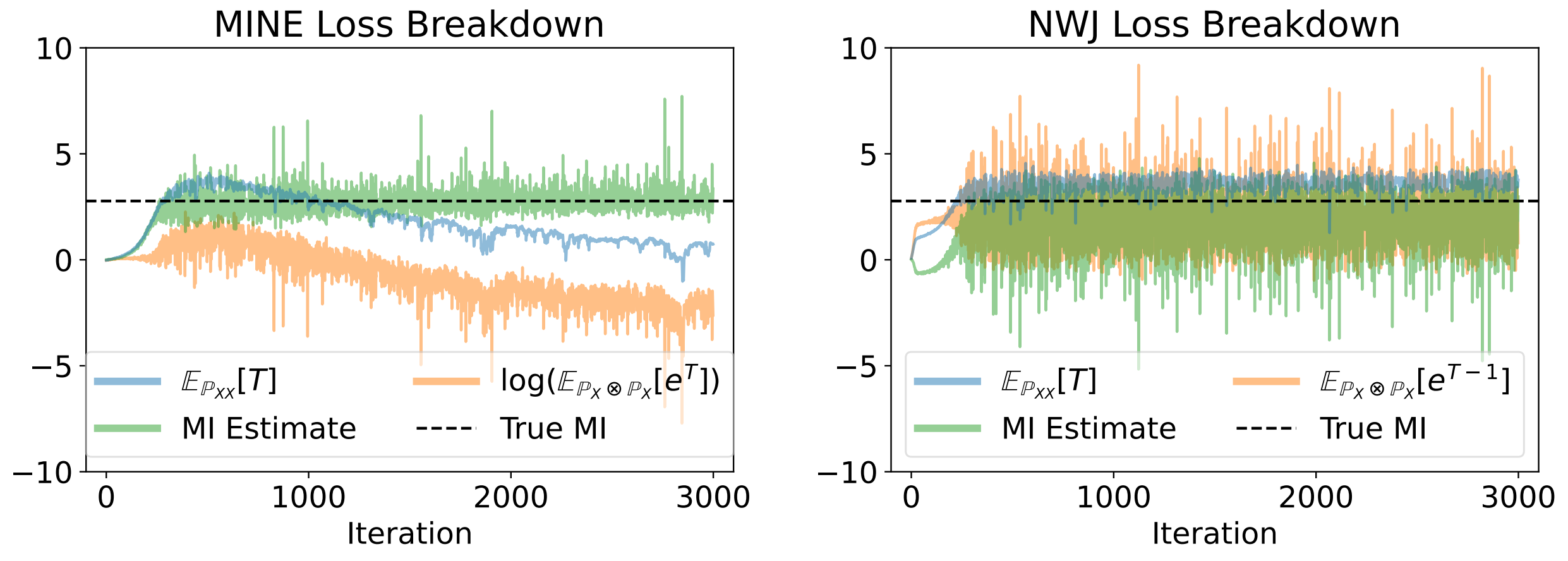

Figure 1. The MINE loss (green) seems to converge, but its constituents (green = blue - orange) are drifting in parallel. On the other hand, NWJ loss is self-regulating, avoiding the drifting problem.

Symptom 2: Explosion Problem

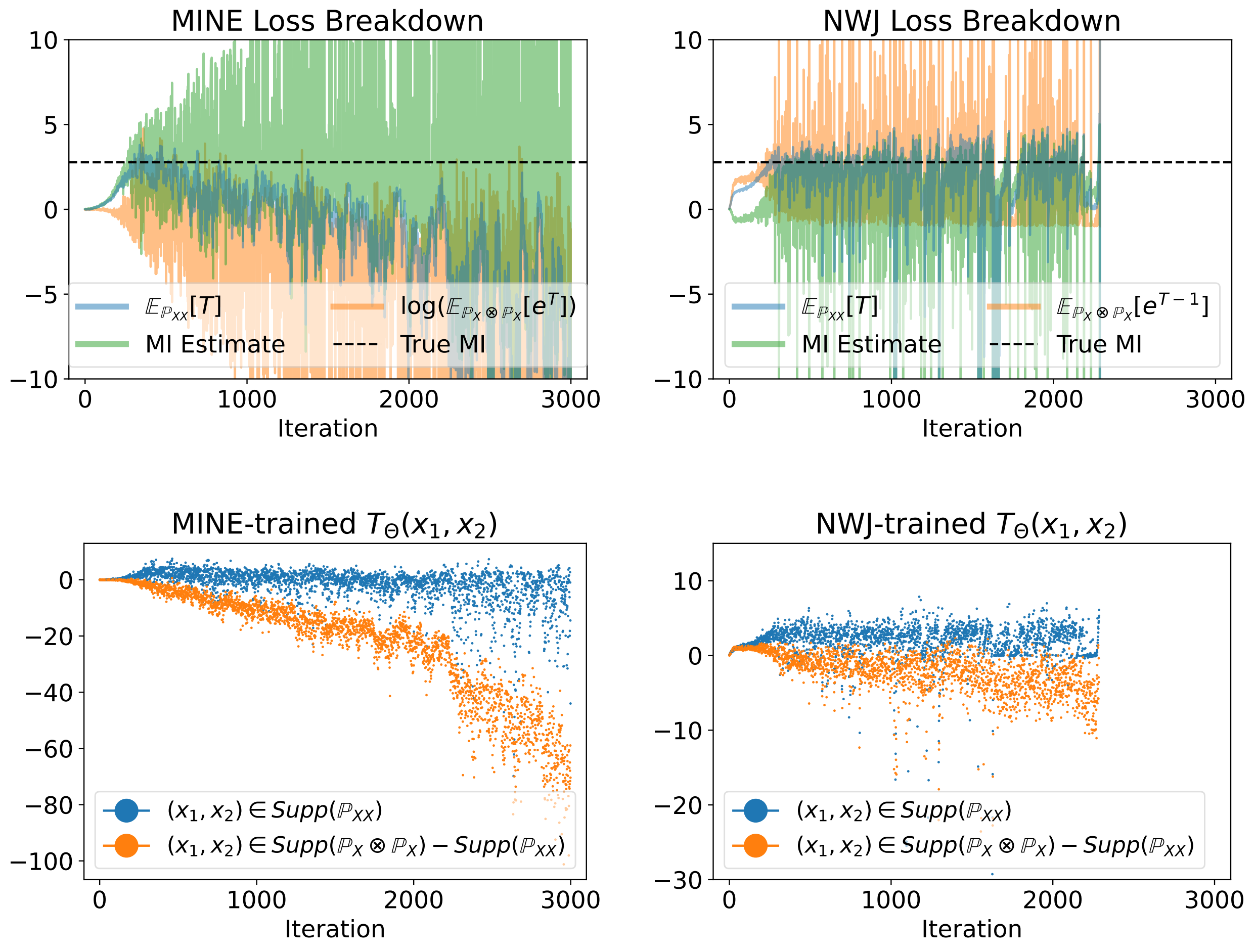

Figure 2. Same settings with Figure 1, but with reduced batch size. Even though both losses are diverging, the network is distinguishing samples based on their underlying distributions (blue and orange).

Regularizing Representations

To stabilize the two KL divergence dual representations D_{DV} [1] and D_{NWJ} [2], we introduce their regularized counterparts, D_{ReDV} and D_{ReNWj}.

Regularized Repr.	Original Repr.	Our Regularizer
$D_{\text{ReDV}}(X,Y) :=$	$\sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \log(\mathbb{E}_{\mathbb{Q}}(e^T))$	$-d(\log(\mathbb{E}_{\mathbb{Q}}(e^T)), C^*)$
$D_{\text{ReNWJ}}(X,Y) :=$	$\sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}(T) - \mathbb{E}_{\mathbb{Q}}(e^{T-1})$	$-d(\mathbb{E}_{\mathbb{Q}}(e^{T-1})), 1)$

To demonstrate the wide applicability of our regularization scheme, we regularize six variational bounds, three for each representation. I_{MINE}, I_{SMILE}, and I_{InfoNCE} are based on D_{DV}. I_{NWJ}, I_{TUBA}, and I_{JS} are based on D_{NWJ}. We design their regularized counterparts to compare their performance.

Theoretical Properties

- **Novel KL dual representation** $D_{\text{KL}}(\mathbb{P}||\mathbb{Q}) = D_{\text{ReDV}}(X,Y) = D_{\text{ReNWJ}}(X,Y)$.
- **Consistency** I_{ReMINE} and I_{ReNWJ} are strongly consistent estimators.
- **Estimation variance** Variance of the second term $\mathbb{E}_{\mathbb{Q}}$ is affected by drifting.
- **Estimation bias** Both macro- and micro-averaging strategies produce a biased MI estimate when the drifting problem occurs.

Regularization Avoids Instability

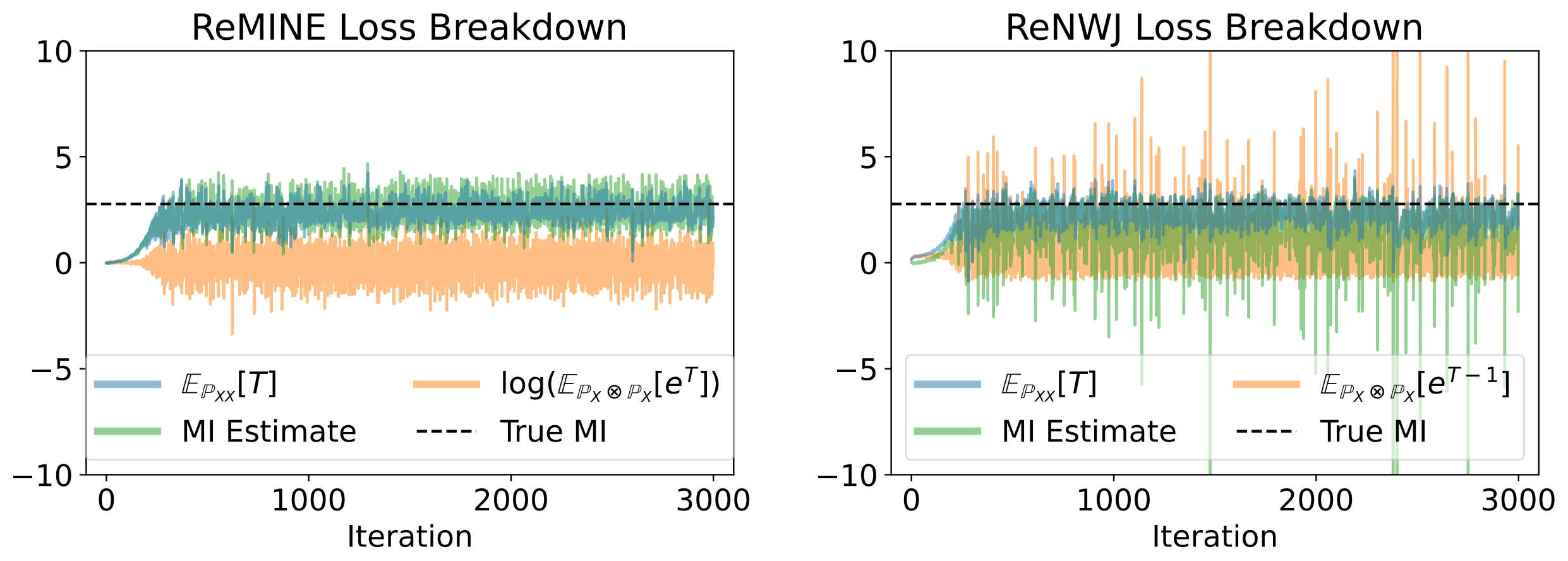

Figure 3. Same settings with Figure 2, but with regularized losses. Regularization mitigates both shifting and exploding symptoms.

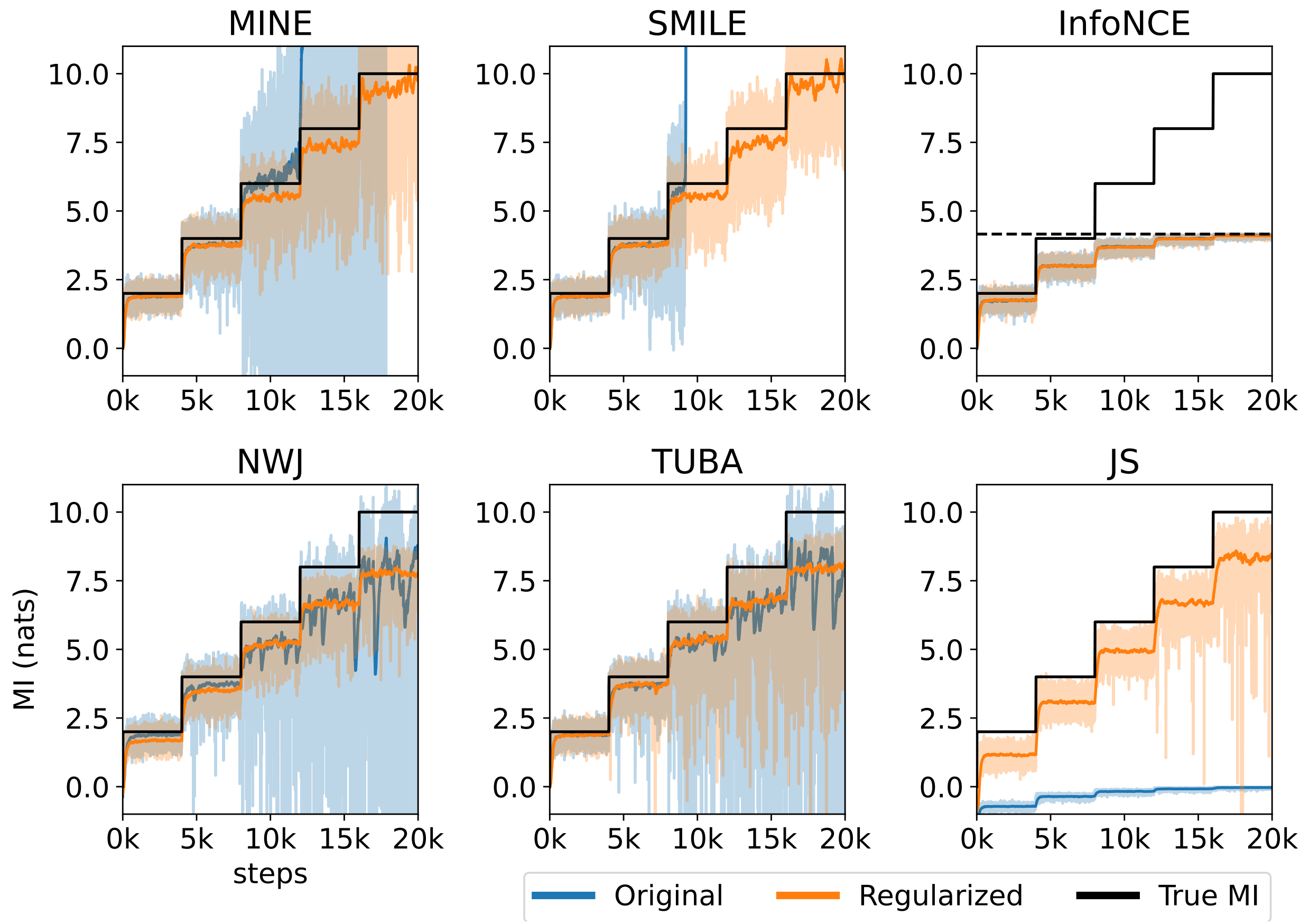

Figure 4. Evaluation on the standard Gaussian benchmark. Original losses' (blue) regularized counterparts (orange) successfully estimates the true MI.

Comparing MI-based Losses with Real-world Tasks

Consider a dataset $D = (X, Y)$ where Y is the label for the sample X, and $H(Y)$ is the entropy of Y. Under the assumption that Y implicitly determines X (i.e. $H(Y|X) = 0$) [3], we propose two types of MI estimation task.

- **Supervised Learning Benchmark** $I(X,Y) = H(Y)$.
- **Contrastive Learning Benchmark** $I(X_1, X_2) = I(X_1, Y) = I(X_2, Y) = H(Y)$, where X_1 be a sample drawn from the dataset with the label Y and X_2 be another sample with the same label Y.

Performance Comparison: Original vs. Regularized

Task		Loss	MI Estimation		Test Accuracy	
			Original	Regularized	Original	Regularized
Supervised Learning Benchmark	CIFAR-10	MINE	**2.300 ± 0.003**	2.298 ± 0.005	0.850 ± 0.009	**0.856 ± 0.004**
		SMILE	2.297 ± 0.009	**2.300 ± 0.003**	**0.854 ± 0.008**	0.853 ± 0.009
		InfoNCE	2.301 ± 0.002	**2.302 ± 0.001**	0.845 ± 0.006	0.845 ± 0.005
		NWJ	**2.297 ± 0.009**	2.294 ± 0.013	0.859 ± 0.003	**0.862 ± 0.004**
		TUBA	2.297 ± 0.008	**2.300 ± 0.003**	**0.862 ± 0.008**	0.859 ± 0.003
		JS	1.944 ± 0.039	**2.000 ± 0.049**	0.838 ± 0.012	**0.842 ± 0.004**
	CIFAR-100	MINE	4.597 ± 0.011	**4.603 ± 0.001**	0.610 ± 0.007	0.610 ± 0.006
		SMILE	4.595 ± 0.015	**4.602 ± 0.002**	0.601 ± 0.015	**0.606 ± 0.007**
		InfoNCE	4.594 ± 0.017	**4.599 ± 0.005**	0.589 ± 0.010	**0.593 ± 0.005**
		NWJ	4.572 ± 0.055	**4.586 ± 0.034**	0.558 ± 0.042	**0.599 ± 0.009**
		TUBA	4.495 ± 0.207	**4.603 ± 0.002**	0.543 ± 0.055	**0.611 ± 0.007**
		JS	4.088 ± 0.430	**4.240 ± 0.116**	0.591 ± 0.026	**0.598 ± 0.010**
Contrastive Learning Benchmark	CIFAR-10	MINE	2.233 ± 0.674	**2.240 ± 0.657**	0.812 ± 0.026	**0.823 ± 0.012**
		SMILE	0.000 ± 0.000	**2.065 ± 0.842**	0.100 ± 0.001	**0.830 ± 0.008**
		InfoNCE	1.705 ± 0.462	**1.739 ± 0.431**	**0.830 ± 0.008**	0.826 ± 0.006
		NWJ	0.000 ± 0.000	**1.910 ± 0.662**	0.100 ± 0.000	**0.831 ± 0.005**
		TUBA	0.000 ± 0.000	**1.358 ± 0.590**	0.100 ± 0.000	**0.830 ± 0.009**
		JS	1.552 ± 0.485	**1.556 ± 0.546**	**0.837 ± 0.003**	0.832 ± 0.009
	CIFAR-100	MINE	**4.634 ± 0.186**	4.563 ± 0.162	0.522 ± 0.026	**0.540 ± 0.020**
		SMILE	0.000 ± 0.000	**4.677 ± 0.162**	0.012 ± 0.003	**0.585 ± 0.007**
		InfoNCE	4.112 ± 0.147	**4.115 ± 0.145**	0.576 ± 0.019	**0.585 ± 0.014**
		NWJ	0.000 ± 0.000	**4.065 ± 0.255**	0.010 ± 0.000	**0.521 ± 0.025**
		TUBA	0.000 ± 0.000	**2.731 ± 0.786**	0.010 ± 0.000	**0.490 ± 0.023**
		JS	3.253 ± 0.368	**3.393 ± 0.124**	0.451 ± 0.020	**0.463 ± 0.031**

Table 1. We compare both MI estimation task and downstream task performance of original and regularized loss using CIFAR10 and CIFAR100 datasets. **Bold text** and blue text indicates the better performance with overlapping and non-overlapping confidence interval, respectively.

References

[1] M. Donsker et al. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 1975.

[2] X. Nguyen et al. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory*, 2010.

[3] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015.