ADS TRACK

Understanding the Ranking Loss for Recommendation with Sparse User Feedback

Zhutian Lin*, **Junwei Pan***, Shangyu Zhang, Ximei Wang, Xi Xiao, Shudong Huang, Lei Xiao, Jie Jiang

Tsinghua University, Tencent Inc.

KDD2024 BARCELONA, SPAIN

CTR Prediction: Binary Classification

- **BCE Method**

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(z_i)\right) + (1 - y_i)\log\left(\sigma(1 - z_i)\right)\right]$$

- **BCE-Ranking Combination Method**

$$L = \alpha L_{BCE} + (1 - \alpha) L_{rank}$$

- Combined-Pair [2015, Twitter], JRC [2023, Alibaba], RCR [2023,Google]
- They claim: introducing ranking loss is helpful to improve **Ranking Ability**
- **Classification Ability** is still unclear

KDD2024
BARCELONA, SPAIN

Classification Ability

- **BCE Method**

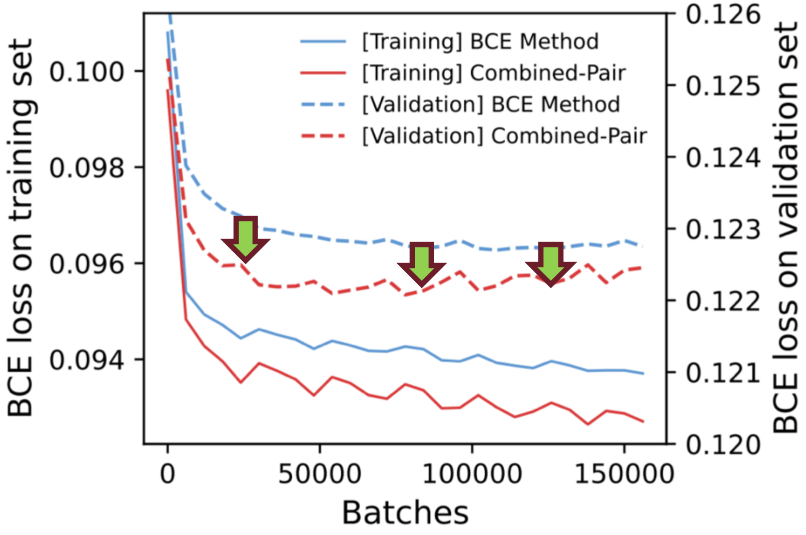
$$\boxed{L_{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\sigma(z_i)) + (1 - y_i)\log(\sigma(1 - z_i))\right]$$

- **Combined-Pair Method [2015, Twitter]**

$$L^{CP} = \boxed{\alpha L_{BCE}} + (1-\alpha)\boxed{L_{RankNet}}$$

$$L_{RankNet} = -\frac{1}{N_+ N_-}\sum_{i=1}^{N_+}\sum_{j=1}^{N_-}\log\left(\sigma\left(z_i^{(+)} - z_j^{(-)}\right)\right)$$

- **Can Combined-Pair gain better classification ability?**



Finding 1. Combined-Pair gets **a lower BCE loss** than the BCE method on the **validation set**, indicating that it **improves the classification ability** rather than only the ranking ability.

KDD2024 BARCELONA, SPAIN

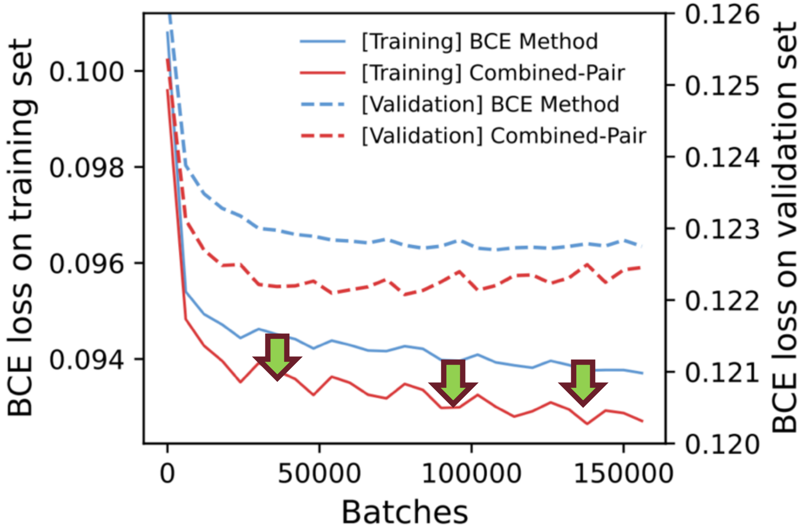
Optimization

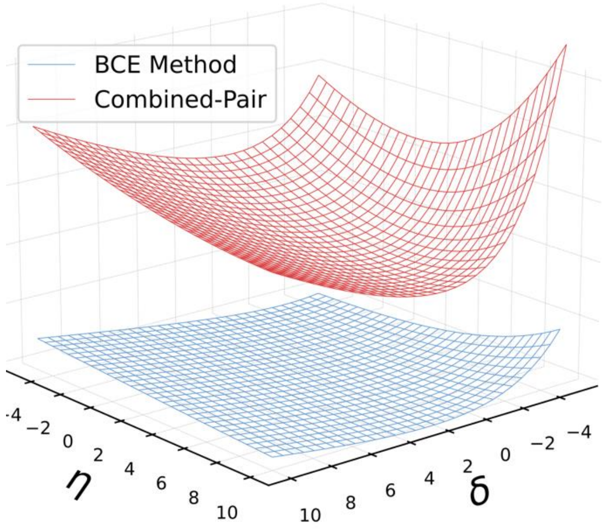
Finding 1. Combined-Pair gets **a lower BCE loss** than the BCE method on the **validation set**, indicating that it **improves the classification ability** rather than only the ranking ability.

- **Better generalization or better optimization?**

Finding 2. Combined-Pair gets a lower BCE loss than the BCE method on the **training set**, indicating that involving an auxiliary ranking loss **helps the optimization of the BCE loss.**

- **What is the optimization issue of BCE method?**
- *The BCE method has a flat loss landscape, indicating optimization.*





KDD2024
BARCELONA, SPAIN

Gradient Analysis of BCE method

- **For BCE Method**
- $L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(z_i)\right) + (1-y_i)\log\left(\sigma(1-z_i)\right)\right]$
- **Gradient to Logit:** Chain Rule

Negative Samples

$$\nabla_{z_j^{(-)}} \mathcal{L}_{\mathrm{BCE}} = \frac{1}{1-\sigma(z_j^{(-)})} \cdot \sigma(z_j^{(-)})(1-\sigma(z_j^{(-)}))$$

$$= \sigma(z_j^{(-)}) = \hat{p}_j.$$

Proportional to the estimated score.

Estimated score is low in scenarios with sparse positive feedback.

Finding 3. When positive feedback is sparse, the **gradients of negative samples vanish** since they are **proportional to the estimated scores**, which are small in an unbiased estimator.

An uncovered challenge

Gradient vanishing of negative samples under sparse positive feedback.

Gradient Analysis of the BCE method

- **For BCE Method**

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\sigma(z_i)\right) + (1 - y_i)\log\left(\sigma(1 - z_i)\right)\right]$$

Positive Samples

$$\nabla_{z_i^{(+)}} \mathcal{L}_{\mathrm{BCE}} = -\frac{1}{\sigma(z_i^{(+)})} \cdot \sigma(z_i^{(+)})(1 - \sigma(z_i^{(+)}))$$

$$= -(1 - \sigma(z_i^{(+)})) = -(1 - \hat{p}_i).$$

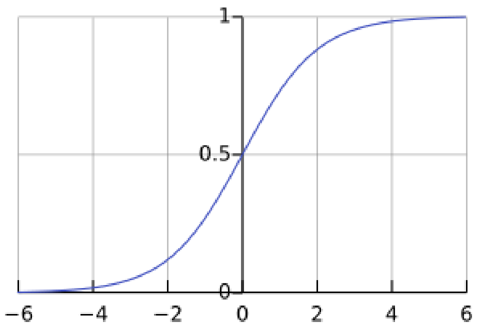
Proportional to (1-pCTR)

Gradient Analysis of Combined-Pair

For Combined-Pair

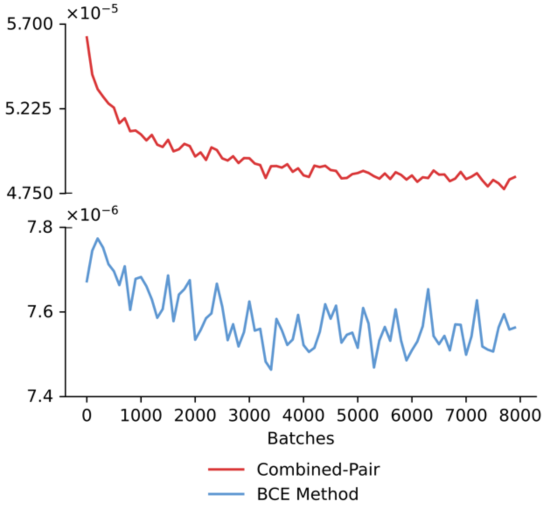
Usually a negative value

$$
\begin{aligned}
\nabla_{z_j^{(-)}} \mathcal{L}_{\text{Rank}}^{\text{CP}} &= \frac{1}{N_+}\sum_{i=1}^{N_+}\sigma(z_j^{(-)} - \boxed{z_i^{(+)}}) \\
&> \frac{1}{N_+}\cdot N_+ \cdot \sigma(z_j^{(-)}) \\
&= \sigma(z_j^{(-)}) = \nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}}.
\end{aligned}
$$



$$
\begin{aligned}
\nabla_{z_j^{(-)}} \mathcal{L}^{\text{CP}} &= \alpha\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} + (1-\alpha)\nabla_{z_j^{(-)}}\mathcal{L}_{\text{Rank}}^{\text{CP}} \\
&> \alpha\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} + (1-\alpha)\nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}} \\
&= \nabla_{z_j^{(-)}}\mathcal{L}_{\text{BCE}}.
\end{aligned}
$$

Gradients Norm



Finding 4. When positive feedback is sparse, Combined-Pair has **larger gradients** for negative samples than the BCE method.

KDD2024
BARCELONA, SPAIN

[1] Le Yan,Zhen Qin,Xuanhui Wang,Michael Bendersky,andMarc Najork.2022. Scale calibration of deep ranking models. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 4300–4309.

Findings so far

Observation
Combined-Pair gets lower BCE loss on validation and training set.

Finding 1. Combined-Pair gets a lower BCE loss than the BCE method on **the validation set**, indicating that it improves the classification ability rather than only the ranking ability. x

Finding 2. Combined-Pair gets a lower BCE loss than the BCE method on the **training set**, indicating that involving an auxiliary ranking loss helps the optimization of the BCE loss.

Perspective
BCE method suffer from the **gradient vanishing of negative samples**, while Combined-Pair mitigate this with larger gradients.

Finding 3. When positive feedback is sparse, the **gradients of negative samples vanish** since they are **proportional to the estimated positive rates**, which are small in an unbiased estimator.

Finding 4. When positive feedback is sparse, Combined-Pair has **larger gradients** for negative samples than the BCE method.

KDD2024
BARCELONA, SPAIN

Experimental Analysis

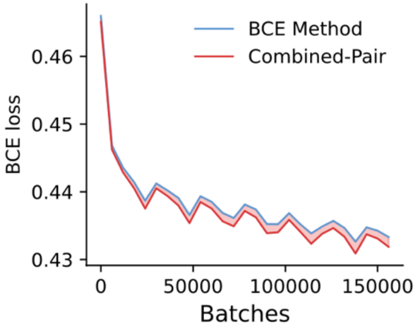
Setting

- **Backbone:** DCN V2
- **Implementation:** FuxiCTR with the same settings as BARS.
- **Dataset:** Criteo_x1
- **Metrics:** (a) BCE loss (i.e., binary-cross entropy loss) to measure the classification ability. (b) AUC to measure the ranking ability.
- **Artificial Datasets with varying Sparsity Degree**: based on the Criteo dataset, by assigning a weight $0 < \beta_{pos} \leq 1$for all its positive samples.

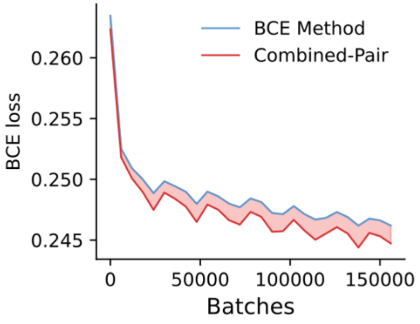
1 https://github.com/reczoo/Datasets/tree/main/Criteo/Criteo_x1
2 https://github.com/reczoo/FuxiCTR/tree/main/model_zoo/DCNv2
3 https://github.com/reczoo/BARS/tree/main/ranking/ctr/DCNv2/DCNv2_criteo_x1

RQ1: Performance Evaluation with various Positive Sparsity Rates

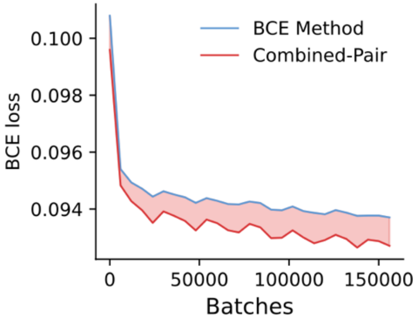
Training BCE Loss



(a) $\beta_{\text{pos}} = 1.0$

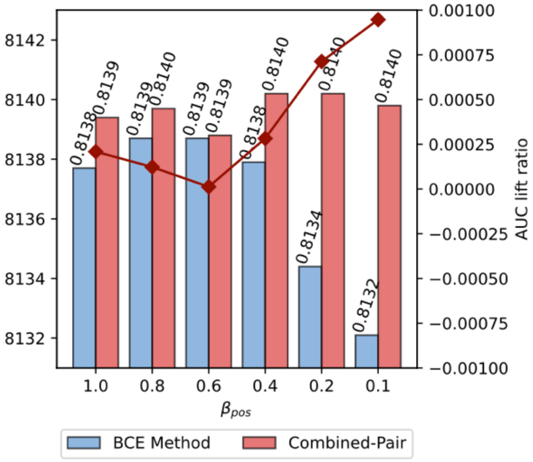


(b) $\beta_{\text{pos}} = 0.4$

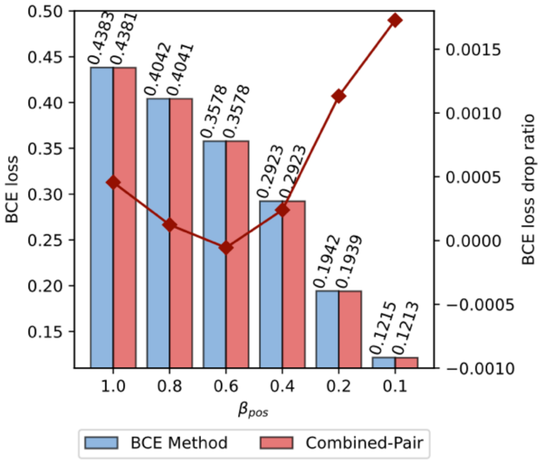


(c) $\beta_{\text{pos}} = 0.1$

Test AUC & BCE Loss



(a) AUC



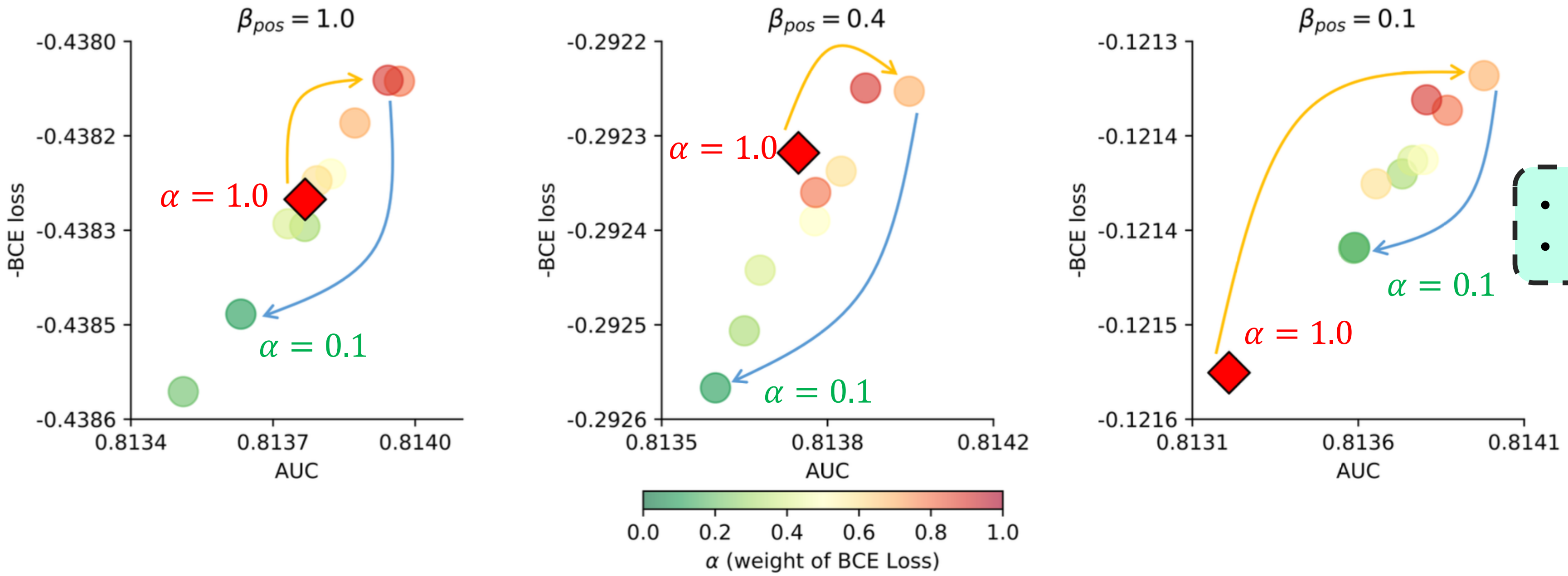
(b) BCE loss

- A smaller β_{pos} indicates **sparser** positive feedback.
- **When positives becomes more sparse** by reducing the β_{pos} , the **more severe of gradient vanishing issue** of negative samples:

RQ2: Impact of Loss Weight

$$L = \alpha L_{BCE} + (1 - \alpha) L_{rank}$$

The most sparse dataset

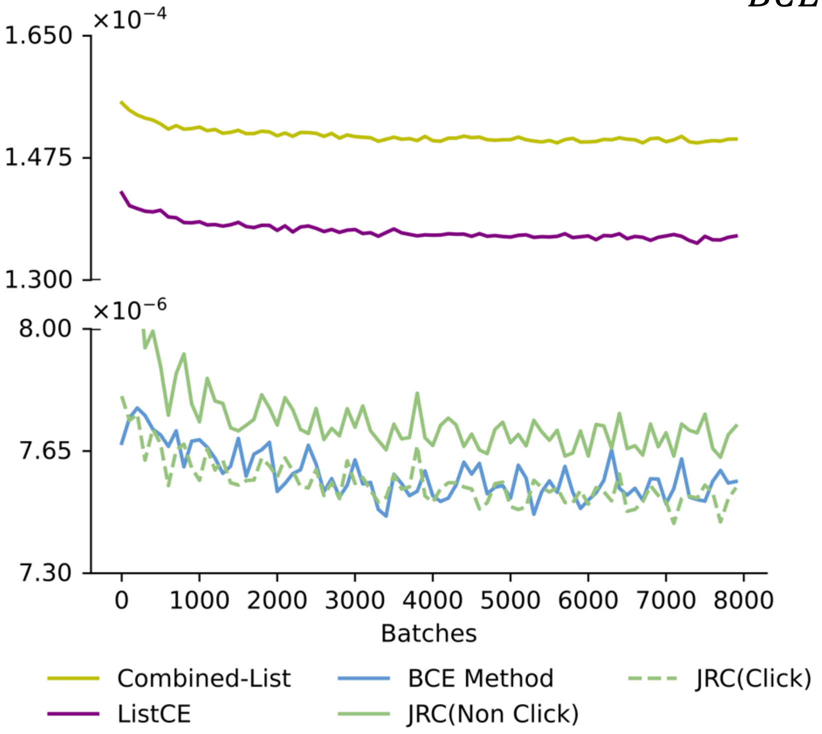


- 0.1 weight for BCE loss
- 0.9 weight for Ranking loss

KDD2024
BARCELONA, SPAIN

RQ3: Evaluation of Different Ranking Losses

$$L = \alpha L_{BCE} + (1 - \alpha) L_{rank}$$



Metric	BCE	BCE+Pairwise	BCE+Listwise		
		Combined-Pair	JRC	Combined-List	RCR
AUC↑	0.81321	**0.81398**↑	0.81355↑	0.81351↑	0.81349↑
BCE loss↓	0.12152	**0.12131**↓	0.12146↓	0.12152	0.12141↓

Observations:

All auxiliary ranking losses get:

- Higher AUC
- **Lower BCE loss**
- **Larger gradients on negative samples**

KDD2024
BARCELONA, SPAIN

RQ4: Beyond Ranking Loss

New Method: Combined-Contrastive

- Combine classification loss with contrastive loss

$$\mathcal{L}^{\mathrm{CC}} = \alpha \mathcal{L}_{\mathrm{BCE}} + (1-\alpha)\mathcal{L}_{\mathrm{Contr}},$$

$$\mathcal{L}_{\mathrm{Contr}} = \frac{1}{|N|}\sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \mathbf{z}_p/\tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \mathbf{z}_a/\tau)}$$

Stage	Metrics	BCE Method	Combined-Contrastive
Training	Gradient Norm	4.9×10^{-6}	7.5×10^{-6}
	BCE loss ↓	0.09667	**0.09428**↓
Testing	AUC↑	0.81321	**0.81340**↑
	BCE loss↓	0.12152	**0.12147**↓

KDD2024 BARCELONA, SPAIN

Online Deployment

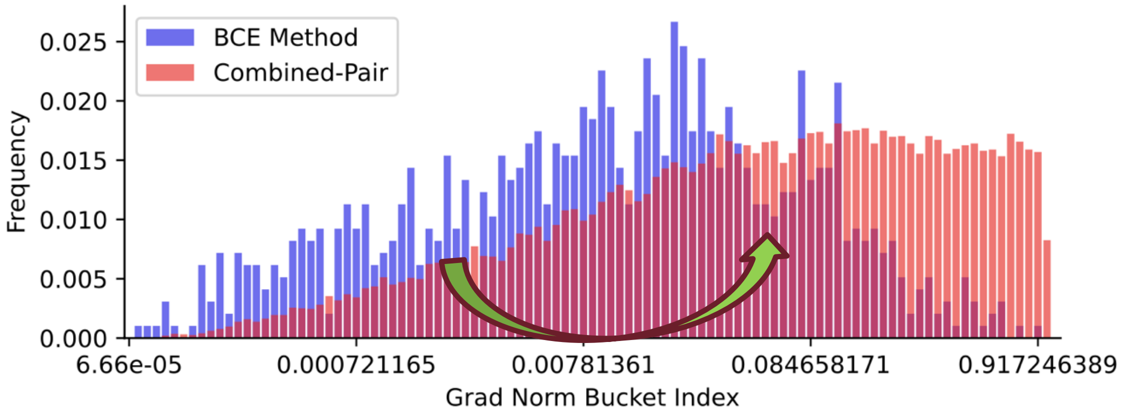
Background:

- Online A/B testing from early July 2023 to August 2023.
- Streaming training with α = 0.9.
- The CTR varies from 0.1% to 2.0% in different scenarios.

A/ B Test Results :

Ad Scenario	CTR	GMV	Cost
WeChat Channels	+0.91%	+1.08%	+0.29%
WeChat Moment	+0.16%	+0.70%	+0.59%
DSP	-0.04%	+0.55%	+0.15%

Gradient Norm Distribution of Negative Samples:



- Right skewed
- More negative samples with **larger gradient norm**

New ads:

Launch Date	GMV	Cost
T	+1.04%	+0.27%
T-1	+1.04%	+0.27%
T-2	+0.83%	+0.47%
T-3	+0.81%	+0.17%
Total	+1.26%	+0.34%

KDD2024
BARCELONA, SPAIN

Q & A

QR of Github repo

KDD2024
BARCELONA, SPAIN