In the appendix, we provide the results on Amazon book data set. **In summary, the experiment results of Amazon book data set are consistent with that of production data set.** But we recommend the readers to focus on the results of the production data set in the paper, which are more robust to the hyper-parameters and much closer to the online industrial recommender systems.

In the following experiments, the default batch size is 128, and the settings of other hyper-parameters are the same as the experiments of production data set.

A EXPERIMENTS OF ANALYSIS

A.1 Exploration of Model

A.1.1 Model Structure. Figure 1 shows that deep CTR model has the one epoch phenomenon, while LR CTR model does not. The LR is trained using Adam with learning rate 1e-2 for optimal performance. And we find that the LR CTR model never faces the one epoch phenomenon under various parameter settings.

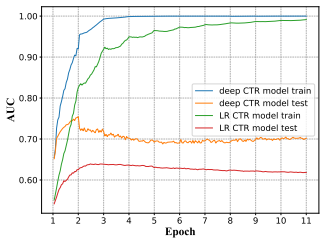
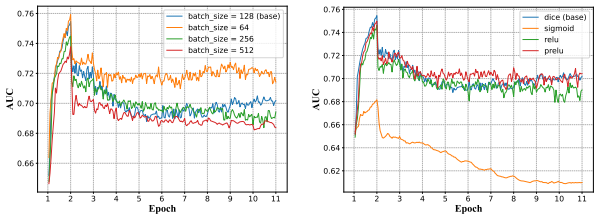


Figure 1: Comparison of convergence curves between deep CTR model and LR model.

A.1.2 Number of Model Parameters. For deep CTR model with Embedding and MLP architecture, the number of parameters of the model has no significant effect on the one epoch phenomenon, as shown in Figure 2.

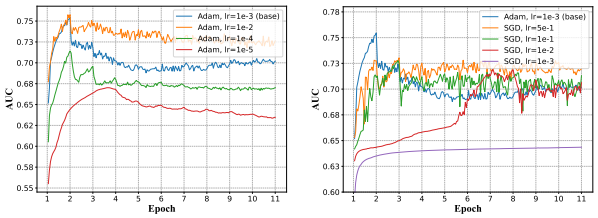
A.1.3 Batch Size and Activation Function. Figure 3 shows that the batch size and activation function are not the factors affecting the one epoch phenomenon.



(a) Batch size (b) Activation function

Figure 3: Test AUCs of models with different batch sizes and activation functions.

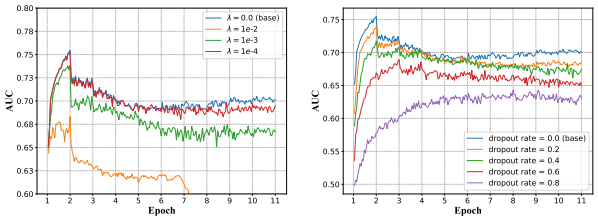
A.1.4 Optimizer and Learning Rate. Figure 4 reveals that models trained with a large and appropriate learning rate perform best, but at this point, the one epoch phenomenon occurs to some extent. In short, the fast model convergence speed, usually with a strong optimizer and a large learning rate, is necessary for the one epoch phenomenon.



(a) Adam (b) SGD

Figure 4: Test AUCs of models with different optimizers.

A.1.5 Techniques to Alleviate Overfitting. Figure 5 shows that weight decay and dropout have no obvious effect on alleviating overfitting and damage the model performance.

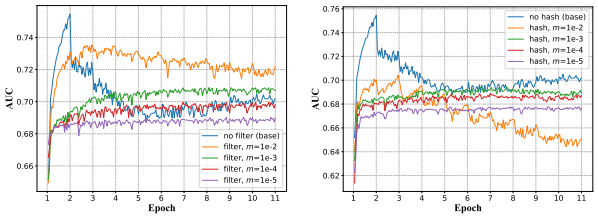


(a) Weight Decay (b) Dropout

Figure 5: Test AUCs of models with weight decay or dropout.

A.2 Data Sparsity

In Figure 6, m is the ratio of the number of IDs after and before compression. As m becomes smaller, the one epoch phenomenon is alleviated accordingly.



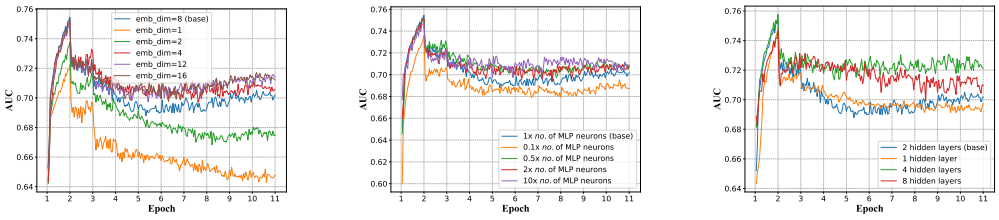
(a) Filter (b) Hash

Figure 6: Test AUCs of models with filter or hash.

B EXPERIMENTS OF HYPOTHESIS

B.1 Distribution Change

The results are shown in Figure 7(a). For the train set, the first epoch corresponds to the untrained samples, and the second and subsequent epochs correspond to trained samples. We find that the $\mathcal{A}$-distance suddenly increases in the second epoch, which verifies



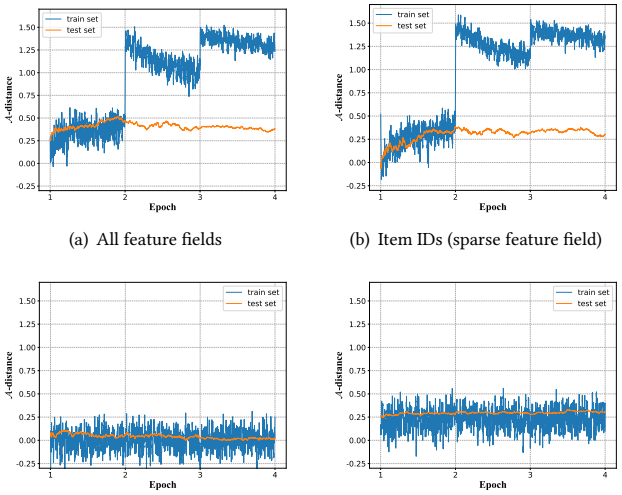
(a) Embedding dimension. (b) Number of MLP neurons. (c) Number of hidden layers.

Figure 2: Influence of number of model parameters of deep CTR model on the one epoch phenomenon.

that $\mathcal{D}\left(\text{EMB}(\text{X}_{\text{trained}}), y\right)$ is different from $\mathcal{D}\left(\text{EMB}(\text{X}_{\text{untrained}}), y\right)$. For the test set, all samples are untrained and the corresponding A-distance is stable, which shows that $\mathcal{D}(\text{EMB}(X_{\text{untrained}}, y)$ has no mutation during the training process.

Figure 7(b) and Figure 7(c) reveal that the variation of $\mathcal{D}(\text{EMB}(\text{X}, y)$ is related to the data sparsity. The difference between $\mathcal{D}(\text{EMB}(\text{X}_{\text{trained}}, y)$ and $\mathcal{D}(\text{EMB}(\text{X}_{\text{untrained}}, y)$ is mainly dominated by the fine-grained feature fields with large data sparsity.

As shown in Figure 7(d), when there is no the one epoch phenomenon, the $\mathcal{A}$-distance does not have sudden change between epochs, which means that there is no obvious difference between $\mathcal{D}(\text{EMB}(\text{X}_{\text{trained}}), y)$ and $\mathcal{D}(\text{EMB}(\text{X}_{\text{untrained}}), y)$ when 1 epoch does not occur.



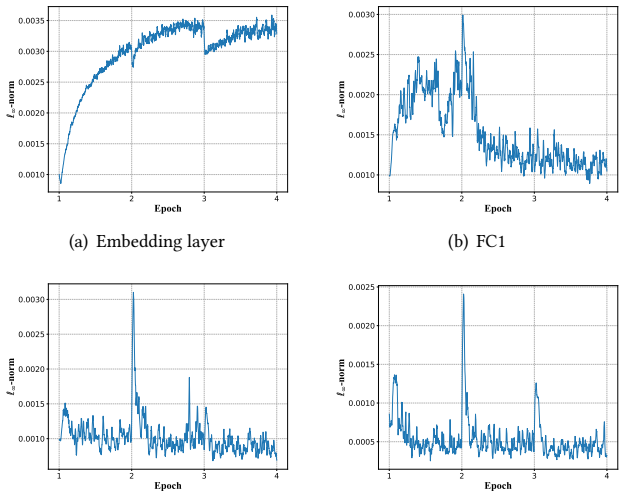
(a) All feature fields (b) Item IDs (sparse feature field)

(c) History clicked item categories (non-sparse feature field) (d) Filter m=1e-3. All feature fields

Figure 7: $\mathcal{A}$-distance between $\mathcal{D}(\textbf{EMB}(\text{X}) \mid y = 1)$ and $\mathcal{D}(\textbf{EMB}(\text{X}) \mid y = 0)$ in the train set and test set.

B.2 Rapid Changes of MLPs

Figure 8 shows that the variation of the MLP layers suddenly increases at the second epoch, while the embedding layer does not. The sudden increase in the parameter changes of the MLP layers supports the hypothesis that MLP layers fit $\mathcal{D}(\text{EMB}(\text{X}_{\text{untrained}}), y)$ at the first epoch but $\mathcal{D}(\text{EMB}(\text{X}_{\text{trained}}), y)$ at the second epoch.



(a) Embedding layer (b) FC1

(c) FC2 (d) FC3 (output layer)

Figure 8: Parameter changes of the Embedding and MLP layers during training. "FC" is short for "fully connected".

The same as the experiment of production data set, we pretrain the model with the default settings for 1 epoch, then fine-tune part of the model parameters with learning rate 1e-4 and freeze the others. Figure 9 shows that only fine-tuning MLP layers leads to the one epoch phenomenon, while only fine-tuning embedding layer greatly alleviates the one epoch phenomenon. It verifies that the rapid change of the MLP layer in the second epoch is the direct cause of the one epoch phenomenon.

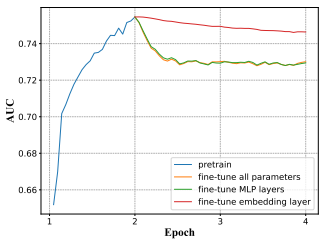


Figure 9: After 1 epoch of pretraining with default settings, fine-tune part of the model parameters and freeze the others. This figure shows the corresponding test AUCs.