To REVIEWER 1: ***Q1***. *Uniform sampling is very unlikely to produce hard negatives with high scores.* **Reply.** Yes, this is also why uniform sampling is not able to generate high quality negative samples. Thus, previous works (IRGAN [37], AdvIR [28]) tried to fit the real negative sampling distribution with techniques of adversarial learning. However, by emphasizing hard negative samples with large scores, they overlook the risk of introducing false negative instances. To solve this problem, we propose to robustify negative sampling by favouring high-variance samples. Moreover, we simultaneously design a simplified memory-based solution for efficient sampling.

Q2. *Since new samples are likely to have lower scores, one either has to increase the temperature or leave Mu relatively static between iterations.* **Reply.** Since CF model can memorize easy training instances first and gradually adapt to hard instances, *a.k.a.* memorization effect [1] (See our experiment results in Fig. 3c/d.), it is unnecessary to avoid introducing new samples into negative sampling. After several training epochs, model is well trained and even new samples can have high scores.

Q3. *how the std can be accurately estimated in Equation 4? And estimating std is expansive.* **Reply.** Please check Appendix B.6 for details. For each candidate sample stored in memory Mu, we directly use its corresponding prediction probability in the latest 5 epochs to compute the std. These prediction results have already been logged even if this sample has just entered Mu. Without any extra forward or backward passes, the computation overhead is constant ($O(1)$) for each sampling operation.

To REVIEWER 2: ***Q1***. *Evaluation results on longer lists (@5, @10, @20 and beyond).*

Reply. In real applications, it is more important to rank the suitable items at top positions of a list. Therefore, a smaller value of K in evaluation emphasizes more on this capability. Previous works [19,23] set K as 1∼10 (out of 100 evaluated items) and 20 (out of 2000 items), respectively. Due to space limitation, we only report the results at $K = 1/3$. As suggested by reviewers, we list the rest results ($K = 5/10$) in the following table. It can be observed that the proposed SRNS still outperforms various baselines.

Method	Movielens-1m				Pinterest			
	N@5	N@10	R@5	R@10	N@5	N@10	R@5	R@10
ENMF	0.3507	0.4030	**0.5066**	0.6682	0.4777	0.5370	0.6824	**0.8643**
Uniform	0.3348	0.3932	0.4884	0.6689	0.4750	0.5323	0.6766	0.8524
NNCF	0.1835	0.2302	0.2840	0.4297	0.4309	0.4925	0.6218	0.8114
AOBPR	0.3428	0.4005	0.5002	**0.6780**	0.4790	0.5375	0.6837	0.8631
IRGAN	0.3372	0.3957	0.4912	0.6714	0.4750	0.5327	0.6758	0.8528
RNS-AS	0.3443	0.3993	0.4992	0.6684	0.4839	0.5390	0.6832	0.8523
AdvIR	0.3445	0.3973	0.5018	0.6644	0.4843	0.5393	0.6839	0.8527
SRNS	**0.3527**	**0.4093**	0.5025	0.6712	**0.4971**	**0.5505**	**0.6894**	0.8531
	0.57%	1.56%	-0.81%	-1.00%	2.64%	2.08%	0.80%	-1.30%

Q2. *Experimental results cannot be compared directly with published results due to different experimental conditions.* **Reply.** There is no standard experimental setting that is adopted by all previous CF works. By following [28,37], we regarded ratings with 4∼5 as positive labels and evaluated with similar list lengths. We will cover more experimental conditions in the final version.

To REVIEWER 3: ***Q1***. *The concept of "hard negative samples" is used without explanation.* **Reply.** They are negative samples with a high probability of being positive according to the model, which are hard for learning. We will elaborate more in the final version.

Q2. *The analysis based on synthetic data is relatively weak, hard to justify the observation.* **Reply.** 1) Variance-based criterion has been adopted in ML community, e.g., [8] improves stochastic optimization by emphasizing high variance samples, and similar technique is widely used in active learning for variance reduction (see "B. Settles. Active learning literature survey. 2010"). Here we introduce this into CF so as to filter out false negative samples. 2) The analysis on synthetic data is motivated by the needs of a reliable measure of sample quality. 3) Experiment results on both synthetic and real-world datasets demonstrate the effectiveness of our SRNS method.

Q3. *Experiment results on longer evaluation lists.* **Reply.** Please see *Q1* of REVIEWER 2.

To REVIEWER 4: ***Q1***. *why SRNS is much faster than existing sampling methods.* **Reply.** SRNS can converge to better performance (N@1) with less time (Fig. 4(a-c)). Moreover, it can be trained from scratch. For time complexity of std computation, please see *Q3* of REVIEWER 1.

Q2. *For experiment on changing scoring function r, better to compare SRNS with RNS-AS and AdvIR.* **Reply.** Original papers of RNS-AS and AdvIR does not consider using different r, thus, to be fair, we only compare SRNS with uniform sampling to demonstrate its generality on different choices of r.

Q3. *Performance gain seems to be marginal as the number of recommended items increases.* **Reply.** 1) As in Appendix B.4, results of both baselines and SRNS in Table 3 are tuned according to N@1 on validation set. 2) Generally learning difficulty increases for all methods as K increases.

Q4. *Needs to consider one candidate-based sampling method and another reinforced-based sampling method as baselines.* **Reply.** [Ding et al. WWW'18] is irrelevant, as it focuses on augmenting negative samples with additional view data, which is not available here. [Ding et al. IJCAI'19] is already compared in experiments, which is RNS-AS.