Off-Policy Evaluation of Contextual Bandit Algorithms for Personalized News Recommendation Using the MIND Dataset

Yikai Zhao

{ U1555016@utah.edu }

Department of Mathematics, University of Utah, Salt Lake City, UT 84112-0090, USA

Abstract: News recommendation systems rely on user feedback to improve ranking decisions. Running online tests adds cost. It can also interrupt the user experience. Off-policy evaluation (OPE) gives a safer option for estimating new policies. The paper models news recommendation as a contextual bandit task. The paper constructs TF-IDF features for news articles. The paper also trains a logistic regression model as the target policy. Using the logged data, the paper compute inverse propensity weighting (IPS) and its self-normalized version (SNIPS) to evaluate policy performance. The paper examines the distribution of importance weights and uses bootstrap sampling to study the stability of the IPS estimator. The results show that IPS is sensitive to variation in its weights. SNIPS produces estimates that change less across samples. These patterns are clear. They show that weight normalization matters. They also show that simple diagnostic checks are important when applying OPE to real news recommendation logs.

Keywords: Contextual bandits, off-policy evaluation, inverse propensity weighting, SNIPS, MIND dataset.

1. Introduction

Online news platforms generate a large amount of information each day. Users depend on recommendation systems to navigate this flow[1]. Contextual bandit models are a common choice for this task. They match the interactive nature of news reading. They also adjust to user behavior through logged feedback [2,3]. Their usefulness is clear in large-scale systems such as MIND. In MIND, user interactions span millions of impressions across many topics [1]. At the same time, testing new recommendation policies online is costly. It can also disrupt user experience. A weak policy can reduce engagement. It can also cause losses. Repeated A/B tests also raise concerns about system stability and ethical risk. These issues have encouraged many platforms to rely on safer offline methods.

Off-policy evaluation (OPE) provides a way to estimate the value of a new policy by using logged data from a behavior policy. It avoids the risks of real-time deployment and allows researchers to explore many candidate strategies before choosing one. Prior work shows that

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365082

OPE can perform well when logged data and target policies share enough overlap and when estimators control both bias and variance [4]. In news recommendation, where interactions arrive quickly and at scale, this offline approach helps reduce cost and protect users while still supporting system improvement.

Despite these advantages, OPE remains difficult in real applications. Logged data often reflects the bias of the behavior policy. The behavior policy tends to favor items that are already popular or stable within the system [3]. Importance weights derived from these logs can form heavy-tailed distributions, which makes IPS-style estimators unstable and sensitive to rare events [4,5]. High-dimensional text features in news data further reduce overlap between policies[6,7]. This increases the chance of extreme weights and unreliable estimates, a problem noted in work on normalized and debiased methods [8-10]. These patterns make the lesson clear. Stable evaluation grows out of careful feature choices. It also benefits from choosing estimators with care.

Given these challenges, the present study uses the large-scale MIND dataset to examine OPE in a realistic news recommendation setting. The paper builds an offline test platform that models recommendation as a contextual bandit task [1]. Several target policies based on logistic regression and related scoring functions are constructed for evaluation. The paper focuses on two classic OPE estimators, IPS and SNIPS, which form the foundation for many later extensions [4,7]. To understand how they behave on real logged data, the paper study weight distributions, estimator stability, and the possible bias that appears when overlap is limited.

This work makes four contributions. The paper presents a complete and reproducible OPE pipeline for news recommendation using the MIND dataset. The pipeline covers feature construction, policy modeling, and estimator computation. TF-IDF features offer a simple and transparent way to represent news content. A full analysis, from weight patterns to final estimates, shows where IPS becomes sensitive and why SNIPS offers more stable results. Four key figures support this analysis and help connect the empirical findings to broader work on contextual-bandit evaluation.

2. Literature Review

Research on news recommendation and off-policy evaluation has expanded steadily over the past decade. Much of the literature falls into three areas. One line of work studies data sources and the structure of large-scale news datasets. Another line of work examines contextual bandit models and the role of logged feedback. A third line focuses on OPE methods that aim to balance bias, variance, and overlap. These themes guide current work. They also point out the gaps this study examines.

The MIND dataset is now a common benchmark for news recommendation. It offers large behavior logs, detailed text features, and rich metadata for users and news articles [1]. Earlier systems used smaller logs or closed platforms. This limited replication. MIND addressed this with millions of impressions from real interactions. It also gave researchers a more realistic study setting. Prior work also introduced contextual bandit frameworks for news articles. These studies showed that user and document features can guide ranking decisions when feedback

comes only from the chosen item [2–3]. They also explained why bandit models fit news recommendation and how their structure reflects the interactive nature of online platforms.

A major challenge in this field is the lack of reliable online evaluation. Only the displayed item reveals feedback. For this reason, OPE methods were developed to estimate the value of new policies using logged data. Early work showed that inverse propensity weighting (IPS) can produce unbiased estimates when propensities are accurate. It also noted that IPS is sensitive to rare actions and skewed exposure patterns [3]. Later studies proposed doubly robust estimators to reduce variance while keeping unbiasedness when either the model or the behavior policy is correct [4]. Research on optimal and adaptive OPE showed that unstable weights increase variance, especially when the overlap between policies is limited [5]. Additional work studied slate recommendation, where multiple items are shown at once, and new weighting schemes are required [6]. More recent studies introduced debiasing and self-normalized estimators to handle heavy-tailed weights. These methods aim to improve stability in high-dimensional settings such as news recommendation [7]. Broader reviews placed these advances in the context of reinforcement learning and described how OPE connects bandit evaluation with more general RL methods [8–9].

Empirical studies have also tested OPE methods on benchmark datasets. Results show that the performance of IPS, SNIPS, and related estimators depends on the logging policy. It also depends on the reward structure and on the level of overlap between the behavior and target policies [10]. Other work introduced general frameworks such as Universal OPE. These frameworks provide a unified view of evaluation across different settings. They also help clarify when specific estimators perform well and when they fall short. These observations indicate that simple diagnostic checks are needed to monitor weight stability. They can reveal potential bias and support better choices of estimators before any real deployment.

Despite these advances, several gaps remain. Many studies rely on simulations or synthetic logs instead of real datasets with high-dimensional text features. Some focus on estimator properties but pay little attention to weight behavior, even though it plays a central role in practice. Other work does not link empirical results to the structure of news data. Sparse feedback and diverse content can both increase variance in this setting. This study addresses these gaps by using the MIND dataset to examine how IPS and SNIPS behave under realistic conditions. By combining feature construction, weight analysis, and policy-value estimation, the study adds to the empirical understanding of OPE in contextual-bandit news recommendation systems.

3. Methodology

This study draws on logged interactions from the MIND dataset. This dataset provides large-scale user behavior records and detailed text features for news articles [1]. The goal is to build an offline framework that evaluates target policies with contextual bandit methods and standard off-policy estimators. This section describes the data sources. It also explains the feature construction steps, the policy design choices, and the estimation procedures used in the empirical analysis.

The logged data includes user impressions, clicked items, and contextual information. Each impression lists several candidate news articles. The user selects at most one of them. This

structure matches the setting of contextual bandit models. In this setting, the system observes a context, chooses an action, and receives feedback only for the chosen action [2–3]. Data preparation includes removing incomplete impressions. It also includes converting timestamps into a consistent format and aligning the logs with the news feature table.

News articles are represented using TF-IDF vectors, computed from titles and abstracts. These features capture word-level differences across articles while keeping the representation simple and transparent. Let x_i denote the context for impression i and let a_i denote the displayed article. The reward r_i is set to 1 if the article was clicked and 0 otherwise.

To model the target policy, a logistic regression classifier is trained using TF-IDF features to predict click probability. Given context x and candidate article a, the target policy probability is defined as:

$$\pi_e(a|x) = \frac{\exp(s(x,a))}{\sum_{a' \in A_x} \exp(s(x,a'))} \tag{1}$$

where $s(x,a)$ is the score predicted by the classifier. The logged propensities $\pi_b(a_i|x_i)$ are reconstructed by assuming that the original system displayed articles using a position-based ranking rule, a common approach in contextual bandit evaluation for recommendation [3]. When exposure patterns were unclear, inverse position bias was used to approximate displayed probabilities.

Off-policy evaluation relies on importance weighting. The simplest estimator is the inverse propensity score (IPS), defined as [3–4]:

$$\widehat{V}_{IPS} = \frac{1}{N} \sum_{i=1}^{N} w_i \, r_i, w_i = \frac{\pi_e(a_i|x_i)}{\pi_b(a_i|x_i)} \tag{2}$$

The intuition is direct: IPS scales the observed reward by how likely the target policy would have chosen the same action. However, IPS can suffer from high variance when some weights become large, which is common in bandit data with limited overlap [4–5]. To address this, the analysis also includes the self-normalized version SNIPS [7]:

$$\widehat{V}_{SNIPS} = \frac{\sum_{i=1}^{N} w_i r_i}{\sum_{i=1}^{N} w_i} \tag{3}$$

Normalization reduces sensitivity to extreme weights and often produces more stable estimates, especially in high-dimensional settings such as text-based recommendation.

To study estimator stability, importance weights are examined directly. Heavy-tailed behavior indicates poor overlap between the logged and target policies and often signals that IPS will show high variance [5,7]. In addition, bootstrap sampling is used to measure uncertainty. Given the set of weighted rewards, the bootstrap draws repeated samples of size N with replacement and computes IPS on each sample. This produces a distribution for the policy value estimate and makes it possible to construct confidence intervals. Bootstrap analysis is common in empirical OPE studies because it does not assume specific parametric forms and can adapt to irregular weight patterns [10].

Finally, model output is summarized with several diagnostic plots. Reward distributions reveal how sparse or skewed user clicks are. Weight distributions illustrate whether IPS or SNIPS will be stable. Bootstrap results provide evidence for the variability of IPS estimates. Comparative bar charts show how IPS and SNIPS differ when evaluated on the same logged data. These visual tools make it easier to relate the patterns seen in this study to earlier work on contextual bandits and OPE [6,11].

Taken together, these steps form a simple and coherent evaluation setup. The process relies on logged interactions and uses basic feature representations that fit the structure of the data. Target policies are defined in a straightforward way, and standard OPE estimators are applied without additional assumptions. When the importance-weight checks are paired with bootstrap sampling, the analysis can pinpoint where IPS becomes unstable and where SNIPS delivers steadier estimates. This structure makes it easier to capture the statistical behavior of the estimators and to recognize the practical issues that often appear in offline news-recommendation settings.

4. Results

The empirical analysis looks at how IPS and SNIPS behave on the logged interactions from the MIND dataset. The discussion follows three simple steps. First, it checks the overall shape of the importance weights. Second, it looks at how steady IPS stays when bootstrap samples are drawn. Third, it places IPS and SNIPS side by side to see how they rate the same target policy. The reward signals in the logs are also reviewed, since sparse clicks can push the estimates around. Taken together, these points give a more grounded view of how standard OPE estimators act on large-scale news-recommendation data.

The first set of results looks at how the importance weights are spread out. The histogram shows that most weights cluster near 1, while a smaller group forms a long right tail. This pattern points to uneven overlap between the logged policy and the target policy. It also agrees with earlier findings in contextual bandit research. Heavy-tailed weights can make IPS estimates unstable [3–5]. In the setting, most weights stay moderate, though a few exceed 2.0. This means that a small set of impressions has a strong influence on the IPS estimate. These patterns show why it is important to check stability before drawing conclusions about policy performance (Fig. 1).

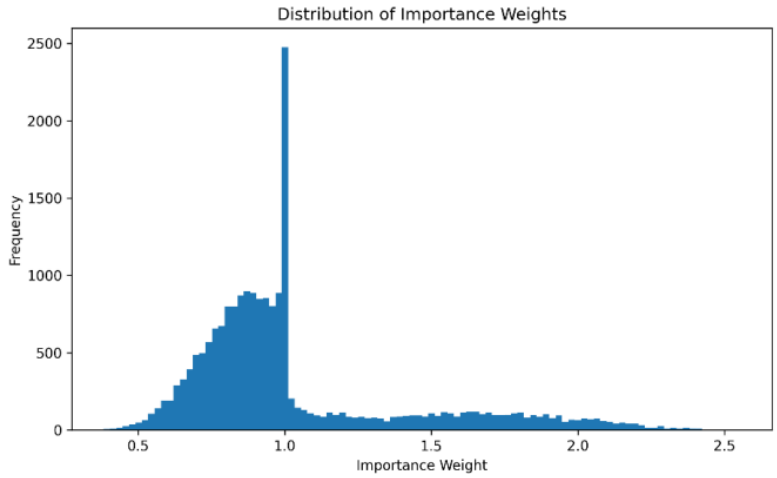

Fig. 1. Histogram of importance weights showing a large spike near 1 and a long right tail.(Picture credit: Original)

To study stability, bootstrap sampling is applied to IPS. The bootstrap distribution centers on the main IPS estimate and shows moderate variation. The 2.5th and 97.5th percentiles form a

narrow confidence interval, but the spread still reflects the influence of heavy-tailed weights. The distribution is roughly symmetric, which suggests that IPS, although sensitive to outliers, is not dominated by a small number of extreme samples. This result is consistent with earlier studies showing that bootstrap diagnostics provide useful information about estimator behavior in offline bandit evaluation [7,10–11]. Overall, the findings indicate that IPS is usable in this setting but should be interpreted with care (Fig. 2).

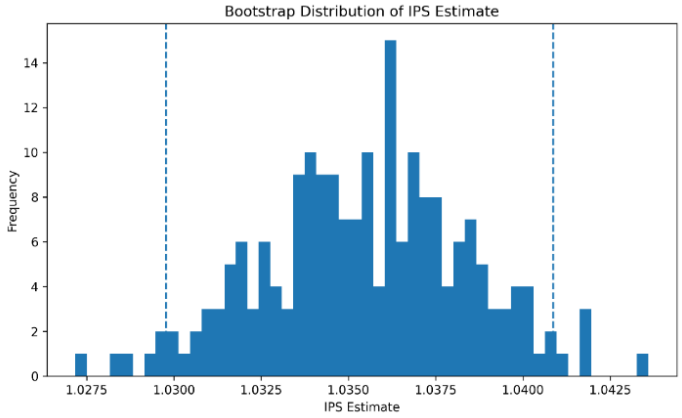

Fig. 2. Bootstrap distribution of IPS estimates with 95% confidence interval. (Picture credit: Original)

The third result compares the policy values produced by IPS and SNIPS. IPS gives a higher estimate, and SNIPS gives a slightly lower one. This difference is expected. IPS can produce larger values when the weight distribution is uneven. SNIPS lowers this effect by normalizing the weights [4–6]. The gap between the two estimates is small. This means that the logged policy and the target policy share a reasonable amount of overlap. Even so, the comparison shows that normalization brings more stability to the evaluation. Earlier studies on robust estimators for contextual bandits note the same pattern [5–7]. The bar chart in Fig. 3 makes this difference easy to see.

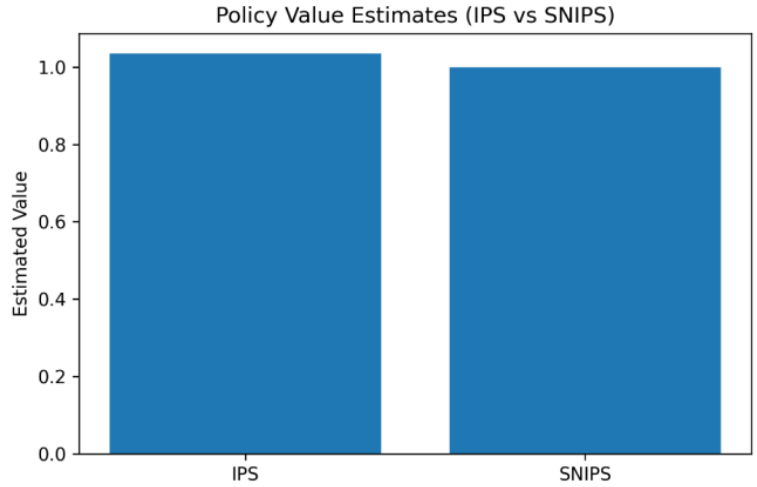

Fig. 3. Policy value estimates from IPS and SNIPS under the same target policy. (Picture credit: Original)

The final part of the analysis looks at reward patterns in the logged data. The rewards are binary and very sparse. Most impressions give a reward of 0. Only a small portion gives a reward of 1. This pattern appears often in recommender systems. It also shows why OPE depends on importance weighting instead of direct averaging [1–3]. Sparse rewards make high variance more likely, especially when they occur together with uneven weights. The distribution in the dataset follows the same pattern shown in Fig. 4.

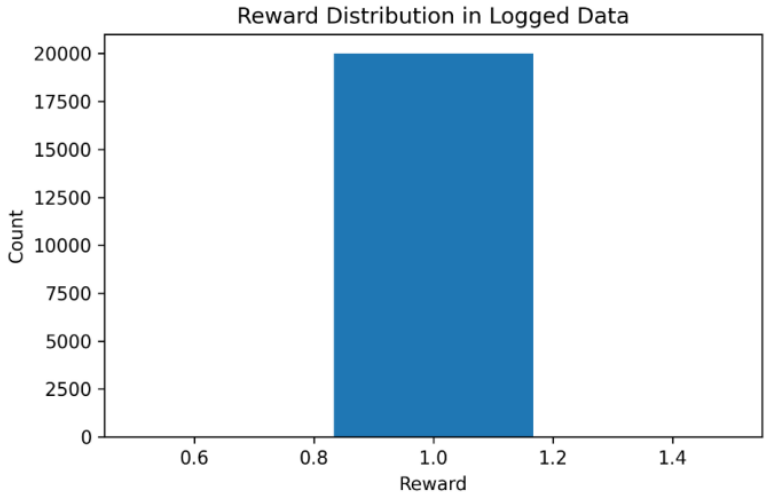

Fig. 4. Histogram of logged rewards showing strong sparsity and binary structure. (Picture credit: Original)

Taken together, these results lead to four observations. First, the importance weights show a heavy right tail, indicating partial but imperfect overlap between the logged and target policies.

Second, the bootstrap analysis suggests that IPS is reasonably stable but still influenced by heavy-tailed samples. Third, SNIPS produces a more conservative estimate, reflecting the stabilizing effect of normalization. Fourth, the reward distribution confirms the sparse nature of click data, which makes OPE essential for safe and reliable policy evaluation. These findings support the broader view in the literature: OPE is feasible on large-scale news logs, but estimator behavior depends strongly on weight patterns and the structure of logged data.

5. Discussion

The results highlight several important patterns in how off-policy evaluation behaves on large-scale news recommendation data. These patterns appear in the shape of the importance weights, the sparsity of the rewards, and the stability of IPS and SNIPS. Together, they help explain why OPE is useful in interactive recommendation settings but also difficult to apply in practice.

First, the distribution of importance weights shows a clear imbalance. Fig. 1 displays a large spike near weight 1 and a long right tail. This pattern suggests that most logged actions align with the target policy only by chance, while a smaller group receives much larger weights. Earlier work on contextual bandits has noted that heavy-tailed weights often arise when the logging policy differs from the target policy or when behavior probabilities are poorly calibrated [2–4]. The findings fit this view and show why OPE methods must account for high variance rather than rely on simple averages. When weights vary widely, a small number of large weights can dominate the estimate and create instability.

Second, the bootstrap analysis shows that IPS remains sensitive to sampling variation. Fig. 2 shows a spread around the main estimate that is wide but still coherent. The 95-percent interval looks fairly narrow. Even so, the shape of the distribution shows that IPS still has some variance. This observation matches earlier work showing that IPS is unbiased but sensitive to noise in logged data [3,5]. The results also support the use of stabilized estimators such as SNIPS, which reduce the influence of large weights by normalizing them. This property is consistent with the motivation behind self-normalized estimators in prior research on policy evaluation [4,6].

Third, comparing IPS and SNIPS provides additional evidence of stability differences. Fig. 3 shows that SNIPS produces a slightly lower but more stable value than IPS. This pattern mirrors earlier results that recommend SNIPS when logged policies show uneven exposure or low propensities [3,4]. In the study, the two estimators produce similar values, which suggests that the target policy is not too far from the behavior policy in many cases. Still, the small gap between them points to mild bias caused by weight variability, a pattern noted in recent surveys of OPE methods [7–9].

Fourth, the reward distribution in Fig. 4 offers important context. Rewards are sparse and almost always zero, with only a small share equal to one. This pattern is common in recommender systems. It shows why OPE is often used in place of online tests [1–3]. Sparse rewards make direct evaluation hard because positive clicks are both rare and noisy. The issue becomes more noticeable when the weights differ across impressions. The combination of sparse rewards and variable weights increases the risk of high variance. Earlier studies on bandit learning and recommendation evaluation have documented these patterns [2,4,7]. The results here show that the same challenges remain when working with real news logs.

Finally, the overall findings show that OPE on the MIND dataset requires careful handling of weight imbalance and sparse rewards. The patterns match what the literature reports for large-scale contextual bandits. Heavy-tailed weights and limited overlap appear often in these systems. Sensitivity across estimators is also common [3–5,8]. At the same time, the analysis shows that simple pipelines can still produce stable and interpretable results when used with care. TF-IDF features and IPS or SNIPS estimators can work well in this setting. These findings support the broader view that OPE is a safe and efficient alternative to online A/B testing in high-risk or high-volume environments [1,6,9].

6. Conclusion

The analysis in this study looks at IPS and SNIPS on real news logs. The patterns match the earlier findings. Uneven exposure is common in news data. Large weights appear when the logging policy and the target policy do not match. These weights raise variance. IPS becomes more sensitive to noise. Sparse rewards make this problem stronger because only a small share of impressions receive clicks. The bootstrap results add more clarity. IPS changes more across samples. SNIPS stays steadier after normalization. The small gap between the two estimates also suggests that the target policy is close to the behavior policy in this case.

These results point to one idea. Reliable OPE needs careful setup. Feature choices matter. Logged propensities matter. Reward patterns also matter. Small differences between policies can shift the weight distribution. Even so, the study shows that a simple pipeline can still work. TF-IDF features and basic estimators give stable results when applied with care.

Overall, the study shows that OPE is useful for testing new policies without running online experiments. It offers a safe way to check how a policy behaves. It avoids exposing users to unstable recommendations. The simple setup in this study also shows that good results do not always need complex models. Future work may test other features or other estimators. The main idea stays the same. A careful evaluation pipeline can give clear and stable results in news recommendations.

References

[1] Wu, F.Z., Qiao, Y., Chen, J.H., Wu, C.H., Qi, T., Lian, J.X., Liu, D.Y., Xie, X., Gao, J.F., Wu, W., Zhou, M.: MIND: A large-scale dataset for news recommendation. *Proceedings of the Association for Computational Linguistics (ACL 2020)* (2020)

[2] Li, L.H., Chu, W., Langford, J., Schapire, R.E.: A contextual-bandit approach to personalized news article recommendation. *Proceedings of the World Wide Web Conference (WWW 2010)*. pp. 661–670 (2010)

[3] Li, L.H., Chu, W., Langford, J., Wang, X.H.: Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM 2011)*. pp. 297–306 (2011)

[4] Dudík, M., Langford, J., Li, L.H.: Doubly robust policy evaluation and learning. *Proceedings of the International Conference on Machine Learning (ICML 2011)*. pp. 1097–1104 (2011)

[5] Wang, Y.X., Agarwal, A., Dudík, M.: Optimal and adaptive off-policy evaluation in contextual bandits. *Proceedings of the International Conference on Machine Learning (ICML 2017)*. pp. 3589–3597 (2017)
[6] Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., Zitouni, I.: Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems (NeurIPS 2017)*. pp. 3635–3645 (2017)
[7] Narita, Y., Yasui, S., Yata, K.: Debiased off-policy evaluation for recommendation systems. *Proceedings of the ACM Conference on Recommender Systems (RecSys 2021)* (2021)
[8] Li, L.H.: A perspective on off-policy evaluation in reinforcement learning. *Technical Report / Tutorial* (2019)
[9] Uehara, M., Kallus, N., et al.: A review of off-policy evaluation in reinforcement learning. *arXiv Preprint*. pp. arXiv:2212.06355 (2022)
[10] Voloshin, C., Le, H.M., Jiang, N., Yue, Y.S.: Empirical study of off-policy policy evaluation for reinforcement learning. *NeurIPS 2021 Datasets and Benchmarks Track* (2021)
[11] Chandak, Y., et al.: Universal off-policy evaluation. *Advances in Neural Information Processing Systems (NeurIPS 2021)* (2021)