Reward Optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search

Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, Maxime Vono

Reward optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search

Imad Aouali
imadaouali9@gmail.com
Criteo

Amine Benhalloum
ma.benhalloum@criteo.com
Criteo

Martin Bompaire
m.bompaire@criteo.com
Criteo

Achraf Ait Sidi Hammou
a.aitsidihammou@criteo.com
Criteo

Sergey Ivanov
s.ivanov@criteo.com
Criteo

Benjamin Heymann
b.heymann@criteo.com
Criteo

David Rohde
d.rohde@criteo.com
Criteo

Otmane Sakhi
o.sakhi@criteo.com
Criteo

Flavian Vasile
f.vasile@criteo.com
Criteo

Maxime Vono
m.vono@criteo.com
Criteo

ABSTRACT

How can we build and optimize a recommender system that must rapidly fill slates (i.e. banners) of personalized recommendations? The combination of deep learning stacks with fast maximum inner product search (MIPS) algorithms have shown it is possible to deploy flexible models in production that can rapidly deliver personalized recommendations to users. Albeit promising, this methodology is unfortunately not sufficient to build a recommender system which maximizes the reward, *e.g.* the probability of click. Usually instead a proxy loss is optimized and A/B testing is used to test if the system actually improved performance. This tutorial takes participants through the necessary steps to model the reward and directly optimize the reward of recommendation engines built upon fast search algorithms to produce high-performance reward-optimizing recommender systems.

CCS CONCEPTS

• **Computing methodologies** → *Maximum likelihood modeling*; *Neural networks*; • **Information systems** → **Recommender systems**.

ACM Reference Format:
Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, and Maxime Vono. 2022. Reward optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3534678.3542622

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08.
https://doi.org/10.1145/3534678.3542622

1 TARGET AUDIENCE AND PREREQUISITES FOR THE TUTORIAL

The intended audience is recommender systems researchers and practitioners who are familiar with supervised learning, offline experimentation and A/B testing.

2 TUTORS

2.1 Tutors' short bio and expertise

Flavian Vasile is part of the Criteo AI Lab where he works as the ML Recommendations Solutions Architect, with his main focus being on the development of Deep Learning-based Recommendation Systems and on introducing aspects of Causal Inference to Recommendation. Before joining Criteo, he worked as a Senior Researcher in the Twitter Advertising Science team; before that, in the Yahoo! Research Lab where he mostly focused on Content Understanding problems. His current research interests include Deep Sequential Models for Recommendation and understanding Recommendation as a decision-making system with reward uncertainty. Among his recent research publications, the work on Causal Embeddings for Recommendation [4] received the best paper award at RecSys 2018 and he is the co-organizer of the REVEAL Workshop series on Offline Evaluation and Bandit Learning for Recommender Systems in conjunction with the ACM RecSys. conference [12–14]. He also presented several tutorials on RecSys topics [22, 26, 29] Video ACM RecSys 2020, Video ACM UMAP 2020.

David Rohde is a research scientist at Criteo. His research interests are around Bayesian machine learning, offline evaluation and causal inference. He is known for RecoGym [21], the BLOB model [23] and promoting the idea that probability theory is sufficient for causal inference [16]. He co-organises the Laplace's Demon webinar series on Bayesian machine learning at scale. He was co-organiser of the Bayesian Causal Inference from Real World Interactive Systems at KDD 2021 [7] he also was co-organiser of SimURec at RecSys 2021 [9] and Laplace's Causal Demon webinar series. He also presented several tutorials on RecSys topics [22, 26, 29] Video ACM RecSys 2020, Video ACM UMAP 2020.

Amine Benhalloum is the lead of the DeepKNN Engineering team at Criteo, working on building large scale representation learning and retrieval systems for recommendation, applying Deep learning to personalize billions of daily display ads, reaching billions of users and connecting them with millions of products. His areas of expertise are: large scale machine learning, natural language processing, information retrieval and data intensive systems. Before joining Criteo, Amine worked on a variety of topics ranging from Natural Language processing to fraud detection. He holds a master's degree in Applied Mathematics.

Martin Bompaire is the lead of the Reco ML team at Criteo. His team handles the recommendation models design, scheduling and serving to select in real time the products chosen for each banner displayed by Criteo (4B a day). Before joining Criteo he defended a PhD on machine learning applied to point processes with a strong focus on Hawkes processes.

Otmane Sakhi is a PhD Student at ENSAE/Criteo. Prior to that, he obtained M.Sc degrees from both CentraleSupelec and ENS Paris Saclay in Applied Mathematics. He specialises in statistical approaches to recommendation including Bayesian value based approaches and counterfactual approaches.

Maxime Vono is a Senior Research Scientist at Criteo, working within the Recommendation research team. Before joining Criteo, he was a researcher at Huawei working with Eric Moulines and held a research visiting position at the Department of Statistics of the University of Oxford where he worked with Arnaud Doucet. His research interests include Bayesian statistics, federated and privacy-preserving learning, and recommender systems. He published papers in top-tier machine learning conferences (ICML, AISTATS) [19] and journals (JMLR) [28] and wrote a review paper on high-dimensional Gaussian sampling published in the prestigious SIAM Review journal [27]. He co-organises the Laplace's Demon webinar series on Bayesian machine learning at scale.

Imad Aouali is a recent graduate student from ENS Paris-Saclay in Applied Mathematics. He also holds an MEng degree in Data Science from Ecole Centrale de Lille, and an MRes degree in Mathematics from Lille University. He has gained experience in applied machine learning and learning theory through his internships at Amazon Science, Criteo AI Lab and Inria and his scholarship program at DeepMind. His research interests include Learning Theory, Bandits, Bayesian Statistics and Recommender Systems.

Achraf Ait Sidi Hammou is an undergraduate student from ENSTA Paris in Applied Mathematics. In his previous internships, he worked on image captioning with domain expertise at ENSTA Paris and Graph Neural Networks at Polytechnique.

Benjamin Heymann is a Senior Researcher at Criteo with expertise in marketplace design and e-commerce. He is a graduate from the Ecole polytechnique (Paris) and Columbia University (New York). He holds a PhD in applied mathematics. He is a recipient of the Siebel scholarship.

2.2 List of in-person presenters

At time of writing we expect the tutorial to be delivered in person by: *Imad Aouali, Amine Benhalloum, Martin Bompaire, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, Maxime Vono.*

The large team will allow us to give individual attention to participants. We have also proposed to present this tutorial at ECML 2022.

2.3 List of contributors

The content is created by: *Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammoua, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, Maxime Vono.*

2.4 Corresponding tutor

David Rohde: *d.rohde@criteo.com.*

3 TUTORIAL OUTLINE

Real world recommender systems identify interesting items to users from massive catalogues at very high speed. This tutorial covers state of the art methods for building recommender systems specifically building on the following technologies:

- Deep learning for flexible definitions of the objective to be optimized.
- Fast (approximate) maximum inner product search to allow very rapid large scale recommendation.
- Reward optimizing recommendation methods that align the optimization problem and the metrics of interest at A/B test time. This is either done using state of the art modelling approaches or the Horvitz-Thompson estimator. We are acutely aware that in real world settings multiple recommendations (i.e. slates) are typically shown simultaneously.

The task of recommendation involves finding a small number of relevant items for a user from a massive catalogue often at high speed. This tutorial covers how we can combine three new technologies in order to improve recommendation quality.

3.1 Deep Learning Combined with MIPS - A Winning Combination

In this section we outline the capability of combining Deep Learning with Maximum Inner Product Search in a production environment [15].

The recommendation engine relies on learning both a query function and P embeddings, one per item in the catalogue, which will later be indexed by the maximum inner product search. Randomized Singular Value Decomposition [18] can be used to produce embeddings of dimension d that can be used for further training just like NLP tasks often rely on pre-trained embeddings such as Word2Vec [17] or BERT [8]. Ranking loss's are shown to allow for extremely efficient training [20] and may be considered reward optimizing under strong assumptions, methods that are reward optimizing under more relaxed assumptions are considered in the later sections.

3.2 A Slate-Level Reward Model that Combines Reward and Rank

This module presents how we can leverage reward-optimizing recommendation to build an efficient and scalable slate recommender system that combines both reward information *i.e.* whether the user interacted with the banner, and rank signal *i.e.* the position of

the selected item in the banner [1, 2]. The benefits of the proposed methodology, *e.g.* recommendation performance and speed, in large-scale scenarios are illustrated by running A/B tests in a simulated environment. We compare our method with common and recently-proposed policy-learning approaches, such as inverse propensity scoring [25] and the top$-K$ heuristic proposed in [6]. We show that these baselines suffer from important caveats such as high variance, over-simplifying assumptions on the parametrised policy and poor scaling when the catalogue size becomes large. In contrast, by both combining reward and rank signals and by leveraging fast (approximate) MIPS techniques, the proposed framework shows promising recommendation results while meeting low-latency requirements.

3.3 Estimating Reward with the Horvitz-Thompson Estimator

The syllabus will be drawn from [3, 5, 11, 24, 25].

The direct way to use the Horvitz Thompson estimator embodies the assumptions of maximum inner product search but is extremely high variance:

$$E[c|\Xi, \boldsymbol{\beta}] \approx \sum_{n=0}^{N} \frac{c_n \pi_{\Xi,\beta,K}(a_1, \ldots, a_K|\Omega)}{\pi_0(a_1, \ldots, a_K|\Omega)}$$

where N is the number of data points, c_n is the reward, $\pi_{\Xi,\beta,K}(a_1, \ldots, a_K|\Omega)$ is the policy parameterized to be maximum inner product search friendly and $\pi_0(a_1, \ldots, a_K|\Omega)$ is the propensity score of the slate.

We investigate several proposals in the slate setting that reduce the variance at the expense of introducing bias to become managable in the recommender setting. We further show that by restricting the policy we are able to optimize maximum inner product search based algorithms.

3.4 Scaling REINFORCE to large catalogs with MIPS

Given a reward estimator $\hat{R}$ (a Reward model, Horwitz-Thompson estimator, Doubly Robust estimator..), Offline Policy based methods aim at learning a parametrised policy π_θ that maximizes the average reward on the logged data $\bar{R}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{a\sim\pi_\theta(.|x_i)}[\hat{R}(a, x_i)]$. We can achieve this by leveraging the REINFORCE gradient $\nabla_\theta \bar{R}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{a\sim\pi_\theta(.|x_i)}[\hat{R}(a, x_i)\nabla_\theta \log \pi_\theta(a|x_i)]$ that enables us to optimize our objective function to obtain reward maximizing policies. In the context of large scale recommender systems, this objective can be computationally demanding as it scales linearly with the size of the catalog. In this module, we want to shed light on a newly proposed method scaling logarithmically on the catalog size by leveraging Maximum Inner Product Search algorithms, allowing faster training time without losing in the quality of the policy learned. We will cover the intuition behind the approach and provide notebooks with toy and real world examples. We will also talk about how to naturally extend the method to slate recommendation with Plackett-Luce and the problems that can be faced when using such algorithms [10].

4 STRATEGIES TO ENCOURAGE AUDIENCE PARTICIPATION AND INTERACTIVITY

Our tutorial builds on Google Colaboratory to allow rapid exploration of ideas. We have a large number of in person tutors who can assist with people running the examples. Our team has extensive experience delivering these types of tutorials and is very much looking forward to doing them in person!

5 SOCIETAL IMPACTS

The goal of this work is to improve recommender systems and have more positive A/B tests of new recommendation algorithms. Achieving this is far from trivial, to the extent this work makes this task easier there may be positive and negative societal impacts. Firstly not everything of value can be measured even at A/B test time, this may result in reduced performance due to a miss-alignment in measurable metrics and those actually representing societal value. A further consideration is that the owner of the recommender system may have different values to the users of the system and society at large. This work prioritizes the concerns of the recommender system operator, in many situations the recommender systems operator has incentives to satisfy the users' interests but this may not always be the case and as a consequence there may be negative societal impacts.

REFERENCES

[1] Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, and Flavian Vasile. 2022. Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation. (2022).

[2] Imad Aouali, Sergey Ivanov, Mike Gartrell, David Rohde, Flavian Vasile, Victor Zaytsev, and Diego Legrand. 2021. Combining reward and rank signals for slate recommendation. *arXiv preprint arXiv:2107.12455* (2021).

[3] Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 19–26.

[4] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. 104–112.

[5] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14, 11 (2013).

[6] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[7] Nicolas Chopin, Mike Gartrell, Dawen Liang, Alberto Lumbreras, David Rohde, and Yixin Wang. 2021. Bayesian Causal Inference for Real World Interactive Systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4114–4115.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Michael D Ekstrand, Allison Chaney, Pablo Castells, Robin Burke, David Rohde, and Manel Slokom. 2021. SimuRec: Workshop on Synthetic Data and Simulation Methods for Recommender Systems Research. In *Fifteenth ACM Conference on Recommender Systems*. 803–805.

[10] Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, and Dmitry Vetrov. 2020. Low-variance black-box gradient estimates for the plackett-luce distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 10126–10135.

[11] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 198–206.

[12] Thorsten Joachims, Maria Dimakopoulou, Adith Swaminathan, Yves Raimond, Olivier Koch, and Flavian Vasile. 2019. REVEAL 2019: closing the loop with the real world: reinforcement and robust estimators for recommendation. In

Proceedings of the 13th ACM Conference on Recommender Systems. 568–569.

[13] Thorsten Joachims, Yves Raimond, Olivier Koch, Maria Dimakopoulou, Flavian Vasile, and Adith Swaminathan. 2020. REVEAL 2020: Bandit and Reinforcement Learning from User Interactions. In *Fourteenth ACM Conference on Recommender Systems*. 628–629.

[14] Thorsten Joachims, Adith Swaminathan, Yves Raimond, Olivier Koch, and Flavian Vasile. 2018. REVEAL 2018: offline evaluation for recommender systems. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 514–515.

[15] Olivier Koch, Amine Benhalloum, Guillaume Genthial, Denis Kuzin, and Dmitry Parfenchik. 2021. Scalable representation learning and retrieval for display advertising. *arXiv preprint arXiv:2101.00870* (2021).

[16] Finnian Lattimore and David Rohde. 2019. Replacing the do-calculus with Bayes rule. *arXiv preprint arXiv:1906.07125* (2019).

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[18] Tae-Hyun Oh, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. 2017. Fast randomized singular value thresholding for low-rank optimization. *IEEE transactions on pattern analysis and machine intelligence* 40, 2 (2017), 376–391.

[19] Vincent Plassier, Maxime Vono, Alain Durmus, and Eric Moulines. 2021. DG-LMC: A Turn-key and Scalable Synchronous Distributed MCMC Algorithm via Langevin Monte Carlo within Gibbs. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8577–8587.

[20] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[21] David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. 2018. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720* (2018).

[22] David Rohde, Flavian Vasile, Sergey Ivanov, and Otmane Sakhi. 2020. Bayesian Value Based Recommendation: A modelling based alternative to proxy and counterfactual policy based recommendation. In *Fourteenth ACM Conference on Recommender Systems*. 742–744.

[23] Otmane Sakhi, Stephen Bonner, David Rohde, and Flavian Vasile. 2020. Blob: A probabilistic model for recommendation that combines organic and bandit signals. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 783–793.

[24] Otmane Sakhi, Louis Faury, and Flavian Vasile. 2020. Improving Offline Contextual Bandits with Distributional Robustness. *arXiv preprint arXiv:2011.06835* (2020).

[25] Adith Swaminathan and Thorsten Joachims. 2015. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*. PMLR, 814–823.

[26] Flavian Vasile, David Rohde, Olivier Jeunen, and Amine Benhalloum. 2020. A gentle introduction to recommendation as counterfactual policy learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 392–393.

[27] Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. 2022. High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm. *SIAM Rev.* 64, 1 (2022), 3–56.

[28] Maxime Vono, Daniel Paulin, and Arnaud Doucet. 2022. Efficient MCMC Sampling with Dimension-Free Convergence Rate using ADMM-type Splitting. *Journal of Machine Learning Research* 23, 25 (2022), 1–69.

[29] Robert West, Smriti Bhagat, Paul Groth, Marinka Zitnik, Francisco M Couto, Pasquale Lisena, Albert Meroño-Peñuela, Xiangyu Zhao, Wenqi Fan, Dawei Yin, et al. 2021. Summary of Tutorials at The Web Conference 2021. In *Companion Proceedings of the Web Conference 2021*. 727–733.