Safe, efficient and robust reinforcement learning for ranking and diffusion models

Gupta, S.

Publication date
2025

Document Version
Final published version

Reinforcement Learning for Ranking and Diffusion Models

Shashank Gupta

SAFE, EFFICIENT, AND ROBUST REINFORCEMENT LEARNING for RANKING AND DIFFUSION MODELS

SHASHANK GUPTA

Safe, Efficient, and Robust Reinforcement Learning for Ranking and Diffusion Models

Shashank Gupta

Safe, Efficient, and Robust Reinforcement Learning for Ranking and Diffusion Models

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in
de Agnietenkapel
op maandag 13 oktober 2025, te 16:00 uur

door

Shashank Gupta

geboren te Jaipur

Promotiecommissie

Promotor:	prof. dr. M. de Rijke	Universiteit van Amsterdam
Co-promotor:	dr. H. Oosterhuis	Radboud Universiteit
Overige leden:	dr. O. Jeunen	Aampe, Belgium
	prof. dr. T. Joachims	Cornell University
	prof. dr. E. Kanoulas	Universiteit van Amsterdam
	dr. S. Magliacane	Universiteit van Amsterdam
	prof. dr. C. Snoek	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research was carried out at the Information Retrieval Lab at the University of Amsterdam and in part during an internship at Meta AI. The research carried out at the University of Amsterdam was funded by DreamsLab.

Cover by Shashank Gupta
Printed by Proefschriftspecialist, Zaandam

ISBN: 978-94-93483-03-3

Acknowledgements

Looking back eight years, right after I finished my master's degree, I thought I was done with academia. I planned to stay in India and in industry for good. Then the pandemic happened, and suddenly I had more time on my hands than ever before. Things changed, and here I am, writing my PhD thesis. There are many people I want to thank for their support and kindness throughout this journey. But before that, I'd like to leave a small reflection—for anyone reading this (especially junior PhD students), or maybe just as a note to my future self.

Every PhD has its highs and lows. The highs: papers, internships, conferences etc. are documented in my CV. The lows are less visible, hidden in my own memory and in the patience of family and friends who shared them with me. Looking back, the biggest lessons I've learned are: Lesson $\#1$: Don't put too much pressure on yourself. For a long time, I treated the PhD as if it had to define my entire worth – learning everything, publishing endlessly, building networks, chasing milestones etc. That pressure came at the cost of my mental and physical health, and it left little room to celebrate small wins along the way. A PhD is just another degree, not your whole identity. Lesson $\#2$: Don't ignore your health. For most of my PhD years, I neglected exercise and let work consume me. Now, I feel the toll of that neglect; I am at my heaviest weight of my life as I am writing this thesis. Work can move slowly, but regaining lost health is far harder. Lesson $\#3$: Have a life outside of work. I never stuck with hobbies or creative pursuits, and that made me dependent on academic success for happiness, a shaky foundation, given how uncertain academia can be in general. A hobby gives balance and even helps professionally in the long run.

If these reflections are useful to someone else, that would make me glad. At the very least, I hope they'll remind the future me of what really matters.

Now I want to thank the people who helped me on this journey. First and foremost, I want to thank Maarten for agreeing to take me as his PhD student. Thank you, Maarten, for being so patient with me, thank you for supporting me initially when I was struggling with the move. I am also thankful to him for being so kind and caring, qualities I admire more than academic prowess. Thank you for always pushing me positively whenever needed and being patient when I was a bit down.

I would then like to thank my second advisor, Harrie. I am really glad I got to work with him eventually, given that I applied for a PhD position with him but eventually accepted one with Maarten. I learned academic rigor, discipline, and attention to detail by working with Harrie. I also learned how to do more theoretically grounded research and write theoretically oriented papers from Harrie. Thank you for being supportive, understanding, and optimistic outside of academics as well. Thank you for helping me become a better researcher.

I am honored to have Olivier, Thorsten, Sara, Cees, and Evangelos on my committee and for taking the time to read and critique my thesis.

Next, I would like to thank some people who helped me on my journey so far (in no particular order). I want to thank Maria for being a positive anchor and a great friend during my PhD. I tend to lean towards pessimism in general, so thank you for always being so positive. Thank you for always being there to listen to me and give

very practical advice in life in general. Seeing Ida grow up and playing with her always brought joy to me. I am also glad I met Mathijs through Maria, who is one of the most fun and funny people I know. I want to thank Philipp for being such a great friend throughout these four years. I am so glad you decided to pursue a PhD and move to Amsterdam. I don't know how I would have managed my PhD (and life in Amsterdam in general) if it were not for you. Thank you for always being there to talk, for always listening to me rant/complain, and just always being there for me. Thank you for always making the time to accompany me to restaurants, bike trips, and conference travels. I also want to thank Vaishali (and Prateek by extension) for always being there to listen to me, advising me, and accompanying me to Indian restaurants. Thank you also for bringing familiarity from India. Also thank you for always supporting me emotionally whenever I felt down. Thank you, Samarth, for bringing a calm energy, and thank you for visiting me in NYC. Thank you, Mariya, for being around in London. My time in London was much more fun because of you. And thank you for always taking the time to call me and meet me, despite your ever-busy schedule. Thank you, Norman, for inviting me whenever you visited Amsterdam. Thank you, Thilina and Clem, for being solid DreamsLab colleagues and for the fun chats at Thilina's dinners. Thank you, Clara, for the fun work-from-café days in Amsterdam. Thank you, Sharvaree, for cooking all the delicious Indian food and for inviting me for dinners.

I had the pleasure of visiting London and New York during my internships and working with the Meta recommender systems org. I want to thank Eric, Hanchao, and Aga for my internship in London. Outside my internship, I want to thank Palak for being a fun friend in London. For my internship in NYC, I want to thank my teammates: Satya, Tsung-Yu, Chaitanya, and Sreya for your mentorship during the internship. I also want to thank François, Paris, and Jun for your support during the internship. Outside my internship, I want to thank Krishna, Kinnari, Shiwangi, Sujal, and Pulkit for making my stay in NYC more fun. Last but not least, I want to especially thank Yiming for all your support during (and outside) my time at Meta. Thank you for your mentorship and guidance in navigating the internship during my first time in London, and thank you for helping me with getting the second internship and for always going out of your way (even when you did not have to) to help me with Meta-related stuff and for talking to me. You have been very kind, thank you.

I would also like to thank all my colleagues at IRLab with whom I have had the pleasure to learn so much during my PhD: Evangelos, Petra (thank you for supporting me initially during my move to Amsterdam), Pablo (thank you for your help during my PhD), Ivana (thank you for your help during my PhD), Ali (thank you for helping out with the tutorial), Amin, Amy, Ana, Andrew, Antonis, Arezoo, Barrie, Chang, Chen, Chuan, Dan, Daniel, David, Fen, Gabriel, Gabrielle, Georgios, Hongyi, Ilias, Jia-Hong, Jie, Jin (thank you for helping out with the tutorial), Jingwei (thank you for being an awesome officemate), Julien, Kidist (thank you for being an awesome officemate), Lu, Maarten Marx, Maartje, Mohammad, Maryam, Maurits, Maxime, Ming (thanks for being an awesome officemate), Mohanna, Mounia, Mozhdeh, Olivier, Panagiotis, Pooya (thank you for inspiring me with your insane fitness levels, though I never followed up on that), Romain, Roxana, Sami, Shaojie, Simon, Svitlana, Teng, Thilina, Thong, Siddharth Mehrotra, Siddharth Singh (thank you for the fun talks and your attempt to speak Hindi with me), Vera, Yangjun, Yibin, Yixing, Yuanna, Yongkang, Yougang,

Yubao, Yuyue, Zahra, Zhirui, Zihan, and Ziming. My apologies if I missed anyone here.

I also want to thank my friends from India for always being there and for always supporting me since our college days: Mitesh, Vivek, Subhadeep, Rohit, Siddharth, Pooja, Kirti, Nayan, and Sri Vidhya. I also want to thank my family for being my pillar of strength: my mom, dad, my sisters, my nephew, and everyone else.

Finally, I want thank you to the original scholar in my family: my grandfather – Gyan Swarop Gupta. I guess I learned how to do research by watching you do the same while growing up – on the history of Indus Valley civilization. I miss you dearly, Nana, I know you are watching me always.

Shashank Gupta
Jaipur
August 2025

Contents

1 Introduction

Reinforcement learning (RL) has established itself as a robust framework for enhancing decision-making systems across diverse application domains [96, 156]. In conventional reinforcement learning configurations, an agent aims to maximize cumulative rewards through sequential action selection while interacting with an environment. Traditionally, the concept of a "reinforcement learning agent" has been linked with robotics or strategic games like chess or Go. However, in recent years, reinforcement learning methodologies have increasingly been applied to user-facing applications, including web search engines, recommender systems, and fine-tuning foundation models for interactive use cases [2, 83, 173].

Among the most widely implemented forms of reinforcement learning in user-facing applications is the contextual bandit framework [9, 151]. Contextual bandits represent a simplified reinforcement learning scenario involving a single state (context) for each interaction with the environment. Within recommender systems, the context typically encompasses personalized user features, the action constitutes the recommended item, and the reward derives from user engagement metrics such as clicks or purchases [151]. For web search applications, the context is the user query (potentially enhanced with personalized features), actions comprise ranked document lists, and rewards stem from search engine result page (SERP) interaction metrics, including normalized discounted cumulative gain (NDCG) or click counts [83]. In foundation model fine-tuning, the context corresponds to an input prompt, the action is a generated sequence of words, and rewards typically originate from learned reward models approximating human feedback [7].

When compared to comprehensive multi-step reinforcement learning scenarios, contextual bandits offer computational simplicity and facilitate easier deployment, frequently using existing offline logged data in an efficient manner [127, 151]. Each context-action pair directly yields a reward, thus eliminating complexities associated with sequential decision-making. Given these practical advantages, this thesis specifically concentrates on contextual bandit methodologies.

Despite these benefits, implementing contextual bandits in ranking problems introduces notable challenges, particularly regarding biases in user feedback. Ranking policies gather data under specific display conditions, resulting in biased user interaction signals – such as position bias or trust bias – which inadequately represent true item relevance [78]. Items positioned lower in rankings receive fewer interactions, potentially

leading to their misclassification as irrelevant.

To address this bias, counterfactual learning-to-rank (LTR) approaches employ inverse propensity scoring techniques, adjusting observed interaction signals to approximate unbiased relevance estimates [4, 51, 78, 152]. However, inverse propensity scoring methods typically suffer from high variance, especially when working with limited data, resulting in unstable or suboptimal deployment [51, 106]. The first component of this thesis addresses safety concerns in counterfactual LTR. In this context, "safety" describes how the new ranking policy performs compared with the current production policy. When the new policy performs worse than the production policy, it is considered *unsafe*; when it performs better, it is considered safe. We propose a safe counterfactual LTR method that theoretically ensures a new ranking policy that performs at least as well as the currently deployed policy. While existing safe methods provide guarantees under assumed click behavior models, these guarantees fail if actual user behavior diverges [63, 109]. To tackle this issue, we introduce a robust safe counterfactual LTR approach that provides reliable guarantees even when user behavior deviates from assumptions.

In the second part of this thesis, we focus on enhancing sample efficiency in contextual bandit learning and evaluation, specifically, achieving lower error rate with limited data. In this setting, off-policy evaluation estimates how a new policy would perform using data collected under a different policy, while off-policy learning uses that same logged data to optimize the new policy itself. Both off-policy evaluation and off-policy learning typically exhibit high variance, causing instability in performance estimates. To reduce variance and improve sample efficiency, we propose an optimal baseline-correction method that significantly decreases the error in off-policy estimates while requiring fewer data points.

Beyond ranking and recommendation systems, diffusion models have recently achieved state-of-the-art results in generative tasks like text-to-image synthesis [170]. Denoising diffusion probabilistic models iteratively refine random noise into meaningful outputs guided by learned distributions [57]. However, these models do not inherently optimize custom objectives such as aesthetic quality or prompt alignment after training. By interpreting the denoising process as an RL action, diffusion models can incorporate user-defined reward functions [13]. While proximal policy optimization (PPO) is commonly used for reinforcement learning fine-tuning, it involves substantial computational costs and high variance, requiring multiple networks loaded simultaneously. In contrast, REINFORCE offers computational efficiency but suffers from high variance and poor sample efficiency [164]. We propose an efficient reinforcement learning fine-tuning method for text-to-image diffusion models, combining REINFORCE's computational efficiency with PPO's improved sample efficiency in a novel method – leave-one-out PPO (LOOP).

Collectively, the contributions made in this thesis highlight the shared challenges of safety, efficiency, and robustness in contextual bandit methods across ranking and diffusion modeling contexts within the reinforcement learning paradigm.

1.1 Research Outline and Questions

Counterfactual LTR corrects user interaction bias primarily using inverse propensity scoring (IPS), weighting clicks inversely proportional to their selection probability. As we pointed out above, while unbiased in expectation, IPS-based estimators are known to suffer from high variance, especially with limited logged interaction data, potentially yielding suboptimal ranking policies [51, 106]. Deploying such suboptimal policies poses significant risks to user experience and business metrics. Therefore, it is crucial to incorporate mechanisms ensuring safe deployment. This leads to the first research question:

RQ1 Can safety guarantees be provided for counterfactual LTR policies to ensure that the new policy is at least as good as the production policy?

To address RQ1, we derive a lower confidence bound for the counterfactual learning to rank (LTR) estimator, establishing a lower bound on the true ranking utility, the ideal target metric for optimization. In Chapter 2, we demonstrate that optimizing this lower bound ensures a ranking policy that is no worse than the current production policy. This property proves particularly valuable when click data is scarce, mitigating the substantial risk of deploying potentially harmful policies.

While RQ1 provides a probabilistic safety guarantee by optimizing the lower bound on the utility, these guarantees depend critically on assumptions regarding user behavior (click model). Deviations from these assumptions invalidate the guarantees, motivating the second research question:

RQ2 Can we provide robust safety guarantees for counterfactual LTR policies even under adversarial user behavior settings?

In Chapter 3, we introduce proximal ranking policy optimization (PRPO), a method ensuring safety for counterfactual LTR without reliance on user behavior assumptions, guaranteeing robust safety even under adversarial conditions.

Thus far, the discussion has focused on contextual bandits within ranking scenarios involving combinatorial action spaces. Next, we examine contextual bandits generating single actions, such as top-1 recommendations, with a focus on improving the sample efficiency in off-policy evaluation and learning. Standard methods like IPS are unbiased in expectation, but suffer from high variance. Alternative methods, including doubly robust (DR) estimators and self-normalized IPS (SNIPS), reduce variance using additive and multiplicative baseline corrections, respectively [79, 147], yet lack a unifying framework. This motivates our third research question:

RQ3 Can we unify variance reduction techniques using baseline corrections and a doubly robust estimator under a common framework?

Chapter 4 proposes the β-IPS estimator, integrating inverse propensity scoring (IPS), doubly robust methods, and self-normalized IPS under a unified baseline correction framework.

RQ4 Given a unified framework for variance reduction techniques under baseline corrections, can we derive a variance-optimal unbiased estimator?

Using the unified β-IPS estimator framework, we investigate whether a variance-optimal baseline correction (β^*) can be analytically derived. In Chapter 4, we confirm this possibility, presenting a closed-form solution for β^* that minimizes variance for both off-policy learning and evaluation tasks.

Contextual bandit theory as previously discussed emphasizes user interactions within ranking or recommendation systems. However, the framework has also been effectively employed in fine-tuning foundation models, such as large language models (LLMs) and diffusion models, typically using proximal policy optimization (PPO). Recent research highlights computational advantages of REINFORCE (policy gradient methods) over PPO for LLMs [5]. Given PPO's challenges with variance and sample inefficiency, we consider improvements through our final research question:

RQ5 Can we improve the sample efficiency of proximal policy optimization for fine-tuning text-to-image diffusion?

In Chapter 5, we systematically compare PPO and REINFORCE for diffusion model fine-tuning. We first demonstrate that REINFORCE exhibits inferior sample efficiency compared to PPO. Subsequently, we propose leave-one-out PPO (LOOP), an enhancement to PPO achieving superior performance with the same number of input prompts by generating multiple actions per prompt.

1.2 Main Contributions

The main contributions of this thesis are categorized into algorithmic and theoretical components, summarized as follows.

1.2.1 Algorithmic contributions

- A safe counterfactual LTR optimization framework that uses the REINFORCE policy gradient, guided by a derived generalization bound for the counterfactual LTR estimator (Chapter 2).
- A robust safe counterfactual LTR algorithm extending the REINFORCE policy gradient method with a clipping mechanism inspired by PPO from reinforcement learning literature (Chapter 3).
- A closed-form solution for the variance-optimal baseline correction term in off-policy evaluation and off-policy learning for contextual bandit methods, substantially reducing variance in practical applications (Chapter 4).
- An efficient reinforcement learning approach for fine-tuning text-to-image diffusion models that enhances sample efficiency by generating multiple diffusion trajectories per input prompt, thereby effectively reducing variance (Chapter 5).

1.2.2 Theoretical contributions

- A generalization bound for the counterfactual LTR estimator using a position-based click model combined with the inverse propensity scoring estimator (Theorem 2.4.2, Chapter 2).
- A generalization bound for the counterfactual LTR estimator employing a trust-bias click model with the doubly robust estimator (Theorem 3.4.1, Chapter 3).
- A proof extending the position-based inverse propensity scoring counterfactual LTR generalization bound to the trust-bias based counterfactual LTR and doubly robust estimator (Appendix 3.A, Chapter 3).
- A proof establishing a closed-form, variance-optimal baseline correction term applicable to off-policy evaluation estimates and learning gradients in contextual bandits (Section 4.3.3, Chapter 4).
- A theoretical demonstration of the sub-optimality of the REINFORCE estimator when samples are reused across iterations in reinforcement learning-based diffusion model fine-tuning (Theorem 5.3.2, Chapter 5).
- A proof demonstrating that the proposed LOOP algorithm achieves lower variance compared to traditional proximal policy optimization methods in diffusion model fine-tuning (Theorem 5.4.1, Chapter 5).

1.2.3 Empirical contributions

- **Safe counterfactual learning to rank.** Extensive simulations on three public learning to rank benchmarks (Yahoo! Webscope, MSLR-WEB30k, and Istella) show that the proposed exposure-based CRM method eliminates the long "unsafe" warm-up period of IPS, matching the production policy after $\sim$400 interactions and converging to the IPS performance asymptotically.
- **Robust safety guarantees in advanced CLTR.** Across the same public learning to rank datasets – even under an adversarial click model, the safe DR and PRPO algorithms reach logging policy performance within the first few hundred queries, more than three orders of magnitude earlier than doubly robust method, while still converging to the optimal ranking performance asymptotically.
- **Variance-optimal off-policy bandit learning.** On a synthetic benchmark that sweeps action-space sizes and logging-policy optimality, the proposed β-IPS estimator reduces gradient variance by up to two orders of magnitude and yields consistently higher policy value and lower MSE than IPS, SNIPS, and DR in both full-batch and mini-batch training settings.
- **Efficient RL fine-tuning of diffusion models.** On the T2I-CompBench, aesthetic, and the image-text-alignment tasks, the LOOP algorithm surpasses PPO and REINFORCE baselines: with $k{=}4$ trajectories per prompt it improves attribute-binding scores by 10–15 points and raises aesthetic quality by $+1.0$ reward points, while simultaneously lowering reward-variance throughout training.

1.2.4 Resource contributions

- **Safe Counterfactual LTR implementation.** All code and experiment scripts for the safe counterfactual LTR methods described in Chapter 2 and 3 are released at: `safe-cltr`.

- **Optimal baseline corrections for off-policy bandits.** The PyTorch implementation that accompanies Chapter 4 is released at: `optimal-baseline-cb`.

1.3 Thesis Overview

The dissertation begins with the introduction, which is the current chapter the reader is engaged with. The research chapters in this thesis are structured into two distinct parts, with the first part predominantly addressing the safety aspect of counterfactual LTR methods in Chapters 2 and 3. Chapter 3 should preferably be read following Chapter 2, as it extends the safety framework established in Chapter 2. Chapter 4 can be approached independently of other chapters in the thesis, as it primarily examines the top-1 action setting in contextual bandits. Similarly, Chapter 5 stands as a self-contained unit that can be read separately, focusing mainly on the post-training refinement of text-to-image foundation models.

1.4 Origins

In this section, we list the origins and the list of contributions for each chapter in the thesis.

Chapter 2 is based on the following paper:

- S. Gupta, H. Oosterhuis, and M. de Rijke. Safe deployment for counterfactual learning to rank with exposure-based risk minimization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 249–258, 2023.

  SG: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation. HO: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – Review & Editing. MdR: Conceptualization, Methodology, Supervision, Validation, Funding Acquisition, Writing – Review & Editing.

Chapter 3 is based on the following paper:

- S. Gupta, H. Oosterhuis, and M. de Rijke. Practical and robust safety guarantees for advanced counterfactual learning to rank. In *CIKM 2024: 33rd ACM International Conference on Information and Knowledge Management*, pages 737–747. ACM, October 2024.

SG: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation. HO: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – Review & Editing. MdR: Conceptualization, Methodology, Supervision, Validation, Funding Acquisition, Writing – Review & Editing.

Chapter 4 is based on the following paper:

- S. Gupta, O. Jeunen, H. Oosterhuis, and M. de Rijke. Optimal baseline corrections for off-policy contextual bandits. In *RecSys 2024: 18th ACM Conference on Recommender Systems*, pages 722–732. ACM, October 2024.

  SG: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation. OJ: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation. HO: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – Review & Editing. MdR: Conceptualization, Methodology, Supervision, Validation, Funding Acquisition, Writing – Review & Editing.

Chapter 5 is based on the following paper:

- S. Gupta, C. Ahuja, T.-Y. Lin, S. D. Roy, H. Oosterhuis, M. de Rijke, and S. N. Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025.

  SG: Conceptualization, Formal Analysis, Investigation, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation. CA: Formal Analysis, Writing – Review & Editing. TYL: Formal Analysis, Writing – Review & Editing. SDR: Formal Analysis, Writing – Review. HO: Formal Analysis, Writing – Review & Editing. MdR: Formal Analysis, Writing – Review & Editing. SNS: Formal Analysis, Funding Acquisition, Writing – Review & Editing

The writing of the thesis also benefited from work on the following publications:

- S. Gupta, H. Oosterhuis, and M. de Rijke. A deep generative recommendation method for unbiased learning from implicit feedback. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 87–93, 2023.
- S. Gupta, H. Oosterhuis, and M. de Rijke. A first look at selection bias in preference elicitation for recommendation (abstract). In *CONSEQUENCES Workshop at RecSys '23*. ACM, September 2023
- S. Gupta, P. Hager, J. Huang, A. Vardasbi, and H. Oosterhuis. Unbiased learning to rank: On recent advances and practical applications. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1118–1121, 2024

- S. Gupta, P. Hager, J. Huang, A. Vardasbi, and H. Oosterhuis. Recent advances in the foundations and applications of unbiased learning to rank. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3440–3443, 2023
- S. Gupta, P. K. Hager, and H. Oosterhuis. Recent advancements in unbiased learning to rank. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 145–148, 2023
- H. C. Bakker, S. Gupta, and H. Oosterhuis. A simpler alternative to variational regularized counterfactual risk minimization. In *CONSEQUENCES Workshop at ACM RecSys '24*, 2024
- S. Gupta, Y. Liao, and M. de Rijke. Towards two staged counterfactual learning to rank. In *Proceedings of the 2025 ACM SIGIR on International Conference on Innovative Concepts and Theories in Information Retrieval*, 2025

Part I

Safe Deployment in Learning-to-rank

2

Safe Deployment for Counterfactual Learning-to-Rank via Exposure-based Risk Minimization

The goal of counterfactual learning to rank (CLTR) is to correct for the selection bias in user interaction data (clicks). It does so by weighting click interactions by the inverse of the estimated effect of the selection bias. Mathematically, CLTR produces unbiased estimates of the "true" relevance signal from biased click signals, in expectation. However, it is well-known that CLTR suffers from high-variance problems, which is exacerbated in the limited logged interaction data size setting [51, 106]. This can lead to unsafe behavior during deployment, as the deployment of a sub-optimal ranking policy can have negative effects on the user experience, and subsequently on the business metrics. To avoid such scenarios, we need a safety mechanism to avoid such detrimental behavior. This brings us to the first research question:

RQ1 Can safety guarantees be provided for counterfactual LTR policies to ensure that the new policy is at least as good as the production policy?

In this chapter, we aim to answer this question by first deriving a generalization bound of the CLTR estimator for the position bias. We show that optimizing for the generalization bound results in a guarantee that the new ranking policy will be at least as good as the production/behavior policy.

2.1 Introduction

LTR methods optimize ranking systems so that the resulting ranking behavior maximizes a given ranking metric [91]. Traditionally, most LTR methods applied a supervised learning procedure based on manually-created relevance judgements. However, obtaining such judgements is time-consuming, expensive and does not scale [21, 116]. As an alternative, LTR methods have been developed that rely on clicks, as they are much cheaper to obtain in abundance in the form of user interaction logs [76].

This chapter was published as [50].

Despite its low costs, click data is generally strongly affected by different forms of interaction bias. Interactions with rankings often suffer from *position bias* [30]: the position at which an item was shown often affects its click through rate (CTR) more than its relevance. As a result, the clicks observed in interaction logs are often more reflective of where items were displayed during logging than how relevant users find them. Thus, naively using this data for LTR, without corrections, can result in heavily *biased* models with suboptimal ranking performance [78, 157].

To mitigate the bias problem in interaction data, the field of CLTR has proposed methods to mitigate bias with unbiased estimation [78]. CLTR mainly relies on exposure-based IPS [111, 158], a LTR specific adaptation of the IPS counterfactual estimation method [58, 77, 148]. Standard exposure-IPS weights clicks by the inverse effect of position-bias on the clicked item. This procedure thus gives more weight to clicks on items that are underrepresented due to position-bias, and vice versa. In expectation, this removes the effect of position-bias from the loss that is optimized.

Unsafe CLTR. Despite enabling unbiased optimization, IPS is also known to suffer from high variance [78, 107]. Specifically, in cases with a lack of click data or with large amounts of noise, high variance can make IPS-based CLTR unreliable and lead to very sub-optimal ranking models [63, 109]. This problem can be so severe that the learned ranking models can be worse than the model used to log the interaction data. Deploying such a learned model could thus result in a substantially degraded user experience. In other words, despite the improvements that IPS-based CLTR can bring, it is also an *unsafe* approach since it can lead to considerable deteriorations, under certain circumstances.

This (un)safety issue is not unique to IPS-based CLTR. Swaminathan and Joachims [148] address this issue for contextual bandit problems by applying a generalization bound. Such a bound can provide a high-confidence upper limit on the difference between the true and estimated performance of a bandit policy [138, 149]. This allows for safer *conservative* optimization. For instance, Wu and Wang [165] introduce a bound based on the divergence between the new policy and the logging policy. This bound avoids policies that stray away from the logging policy, unless there is strong evidence that they are actual improvements. This method might appear to be a great fit for CLTR, but, unfortunately, it is based on action propensities that do not generalize well to the very large action spaces in CLTR. Therefore, there is a need for a conservative generalization bound that is practical and effective in the CLTR setting.

Safe CLTR. To address this gap, in this chapter we propose an exposure-based counterfactual risk minimization (CRM) method that is specifically designed for safe CLTR. Similar to how exposure-based IPS deals with the large action spaces in ranking settings, our method is based on an exposure-based alternative to action-based generalization bounds. We first introduce a divergence measure based on differences between the distributions of exposure of a new policy and a safe logging policy. Then we provide a novel generalization bound and prove that it is a high-confidence lower-bound on the performance of a learned policy. When uncertain, this bound defaults to preferring the logging policy and thus avoids decreases in performance due to variance. In other words, with high-confidence, ranking models optimized with this bound are guaranteed to never deteriorate the user experience, even when little data is available.

Main contributions. We are the first to address CRM for CLTR and contribute a novel exposure-based CRM method for safe CLTR. Our experimental results show that our proposed method is effective at avoiding initial periods of bad performance when little date is available, while also maintaining high performance at convergence. Our novel exposure-based CRM method thus enables safe CLTR that can mitigate many of risks attached to previous methods.

Accordingly, we hope that our contribution makes the adoption of CLTR methods more attractive to practitioners working on real-world search and recommendation systems.

2.2 Related Work

In this section, we review related work on CLTR and CRM in off-policy learning.

2.2.1 Counterfactual learning to rank

LTR is a well-established area of research that deals with learning optimal rankings to maximize a pre-defined notion of utility [91]. Traditionally, LTR systems were optimized using supervised learning on manually-created relevance judgements [21]. However, the manual curation of relevance judgements is a time-consuming and costly process [21, 116]. Moreover, manually-graded relevance signals do not always align well with actual user preferences [133]. Due to these shortcomings, LTR from user interactions has become a popular alternative to supervised LTR [22, 75, 78, 141].

Learning from user interactions/click logs was introduced in the pioneering work of Joachims [76]. Click data is relatively cheap to collect and indicative of actual user preferences [119]. In spite of these advantages, click data is known to be a noisy and biased estimate of the true user preferences [30, 111]. Some of the common biases identified in the LTR literature are position bias [30]: trust bias [4], and item-selection bias [108].

To counter the effect of bias, Joachims et al. [78] introduced counterfactual learning in the context of LTR. They proposed the application of inverse propensity scoring (IPS), a causal inference technique that has prevalence in the offline bandit learning literature [77]. IPS models the probability of the user examining a document at a given displayed rank. The inverse of the examination probability, i.e., the inverse propensity, is used to correct for the position bias. As a result of the inverse weighing scheme, IPS-based LTR optimization is unaffected by position bias, in expectation [78]. Since its introduction, there has been an increasing interest in the area, with several application of IPS in the context of ranking [4, 108, 152, 158]. Recent work has also explored CLTR under a stochastic logging policy, where some exploration is introduced, as opposed to pure exploitation [108, 110, 169].

With regard to safety in learning from user interactions, Jagerman et al. [63] introduced the notation of safe exploration for offline contextual bandit algorithms. The authors introduced safe exploration algorithm (SEA), which applies high-confidence performance bounds to *safely* choose between the deployment of a logging policy and a learned policy. Oosterhuis and de Rijke [109] applied this context to LTR and

introduced a generalization and specialization framework to safely choose between a generalized feature-based LTR model, and a specialized tabular LTR model. The important difference between prior work and our work is that existing methods safely *choose* between policies, whereas our method safely *optimizes* a policy. To the best of our knowledge, we are the first to consider notion of safety for the *optimization* of LTR models.

2.2.2 Counterfactual risk minimization for offline learning from logs

A relevant area closely related to CLTR is off-policy learning, or offline learning from bandit feedback data [56, 77, 127, 148]. Off-policy learning tries to bridge the mismatch between the action distributions of a new policy and the logging policy [77]. The most common techniques used to achieve that goal are IPS and importance sampling [58]. However, as noted by Cortes et al. [29], the IPS estimator can have unbounded variance, which can lead to large errors in its estimation. Consequently, optimization with IPS can result in convergence problems and severely suboptimal policies.

To account for this high-variance problem, Swaminathan and Joachims [148] introduced counterfactual risk minimization (CRM), an off-policy method that explicitly controls for the variance during off-policy learning from bandit feedback data. Specifically, their learning objective consists of both the IPS loss and a variance regularization term, which minimizes the dissimilarity between the two policies. This variance regularization term represents the *risk* that stems from the variance of the IPS estimation, however, computing it requires a pass over the entire data which does not scale well. As a scalable alternative, Wu and Wang [165] introduced variational counterfactual risk minimization (VCRM), where the authors estimate the *risk* of the new policy by random sampling from the logged data. The objective function to be optimized in the VCRM method is derived from a generic theoretical analysis of learning from importance sampling [29]. The risk term in the VCRM method is defined in terms of a specific divergence between the logging policy and the new policy, known as the Rényi divergence [123]. To the best of our knowledge, there is no existing work on CRM in a LTR setting, making the work in this chapter the first to propose a CRM approach for the LTR task.

2.3 Background

2.3.1 Learning to rank

The objective of learning to rank methods is to optimize a ranking policy (π), so that for user-issued queries (q) it provides the optimal ranking of their pre-selected candidate document sets (D_q) [91]. Formally, this objective can be expressed as the maximization of the following utility function:

$$U(\pi) = \mathbb{E}_q \left[\sum_{d \in D_q} \rho(d \mid q, \pi) P(R = 1 \mid d, q) \right]. \tag{2.1}$$

where $\rho(d \mid q, \pi)$ is the weight π gives to document d for query q. The choice of ρ determines what metric is optimized, for instance, the well-known normalized discounted cumulative gain (NDCG) metric [65]:

$$\rho_{\text{DCG}}(d \mid q, \pi) = \mathbb{E}_{y \sim \pi(\cdot|q)}\big[(\log_2(\text{rank}(d \mid y) + 1))^{-1}\big]. \tag{2.2}$$

where y is a ranking sampled from the policy π. For this chapter, the aim is to optimize the expected number of clicks, the next subsection will explain how we choose ρ accordingly.

2.3.2 Counterfactual learning to rank

Position bias in clicks. Optimizing the LTR objective in Eq. 2.1 requires access to the true relevance labels ($P(R = 1 \mid d, q)$), which is often impossible in real-world ranking settings. As an alternative, CLTR uses clicks, since they are present in abundance as logged user interactions. However, clicks are a biased indicator of relevance; for this chapter, we will assume the relation between clicks and relevance is determined by a position-based click model [28, 78]. For a document d displayed in ranking y for query q, this means the click probability can be decomposed into a rank-based examination probability and a document-based relevance probability:

$$P(C = 1 \mid d, q, y) = P(E = 1 \mid \text{rank}(d \mid y))P(R = 1 \mid d, q). \tag{2.3}$$

The key characteristic of the position-based click model is that the probability of examination only depends on the rank at which a document is displayed: $P(E = 1 \mid d, q, y) = P(E = 1 \mid \text{rank}(d \mid y))$. Furthermore, this model assumes that clicks only take place when a document is both relevant to a user and examined by them. Consequently, the click signal is an indication of both the relevance and examination of documents. Thus, the position at which a document is displayed can have a stronger effect on its click probability than its actual relevance [30].

Inverse-propensity-scoring for CLTR. We assume a setting where N interactions have been logged using the logging policy π_0, for each interaction i the query q_i, the displayed ranking y_i, and the clicks c_i are logged:

$$\mathcal{D} = \big\{q_i, y_i, c_i\big\}_{i=1}^{N}. \tag{2.4}$$

We will use $c_i(d) \in \{0, 1\}$ to denote whether document d was clicked at interaction i. Furthermore, we choose ρ to match the examination probabilities under π:

$$\rho(d \mid q, \pi) = \mathbb{E}_{y \sim \pi(\cdot|q)}\big[P(E = 1 \mid \text{rank}(d \mid y))\big] = \rho(d). \tag{2.5}$$

Hence, our optimization objective $U(\pi)$ is equal to the expected number of clicks (cf. Eq. 2.1 and 2.3).

In order to apply IPS, we need the propensity of each document [78], following Oosterhuis and de Rijke [110] we use:

$$\begin{aligned}\rho(d \mid q, \pi_0) &= P(E = 1 \mid \pi_0, d, q) \\ &= \mathbb{E}_{y \sim \pi_0(\cdot|q)}\big[P(E = 1 \mid \text{rank}(d \mid y))\big] \\ &= \rho_0(d).\end{aligned}$$

Thus, the exposure of d represents how likely it is examined when using π_0 for logging. Thereby, it indicates how much the clicks on d underrepresent its relevance. For the sake of brevity, we drop q, π and π_0 from our notation when their values are clear from the context: i.e., $\rho(d \mid q, \pi) = \rho(d)$ and $\rho(d \mid q, \pi_0) = \rho_0(d)$.

The exposure-based IPS estimator takes each click in $\mathcal{D}$ and weights it inversely to $\rho_0(d)$ to correct for position-bias [78, 110]:

$$\hat{U}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D_{q_i}} \frac{\rho(d)}{\rho_0(d)} c_i(d). \tag{2.6}$$

In other words, to compensate that position bias lowers the click probability a document by a factor of $\rho_0(d)$, clicks are weighted by $1/\rho_0(d)$ to correct for this effect in expectation. As a result, clicks on documents that π_0 is likely to show at positions with low examination probabilities (i.e., the bottom of a ranking) receive a higher IPS weight to compensate.

Statistical properties of the IPS estimator. The IPS estimator $\hat{U}(\pi)$ (Eq. 2.6) is an unbiased and consistent estimate of our LTR objective $U(\pi)$ (Eq. 2.1) [106]. It is *unbiased* since its expected value is equal to our objective:

$$\mathbb{E}_{q,y,c}\Big[\hat{U}(\pi)\Big] = U(\pi), \tag{2.7}$$

and it is *consistent* because this equivalence also holds in the limit of infinite data:

$$\lim_{N \to \infty} \hat{U}(\pi) = U(\pi). \tag{2.8}$$

For proofs of these properties, we refer to previous work [78, 103, 108].

It is important to note that the unbiasedness and consistency properties do not indicate that the actual IPS estimates will be reliable. The reason for this is that the estimates produced by IPS are also affected by its variance:

$$\mathrm{Var}_{y,c}\Big[\hat{U}(\pi) \mid q\Big] = \sum_{d \in D_q} \frac{\rho(d)^2}{\rho_0(d)^2} \mathrm{Var}_{y,c}[c(d) \mid \pi_0, q]. \tag{2.9}$$

As we can see, its variance is very large when some propensities are small, due to the $\rho_0(d)^{-2}$ term. As a result, the actual estimates that IPS produces can contain very large errors, especially when N is relatively small or clicks are very noisy. In other words, $\hat{U}(\pi)$ can be far removed from the true $U(\pi)$, and therefore, optimization with IPS can be very unsafe and lead to unpredictable results.

2.3.3 Counterfactual risk minimization for offline bandit learning

The foundational work by Swaminathan and Joachims [148] introduced the idea of counterfactual risk minimization (CRM) for off-policy learning in a contextual bandit setup. To avoid the negative effects of high-variance with IPS estimation during bandit optimization, they use a generalization bound through the addition of a risk term [98].

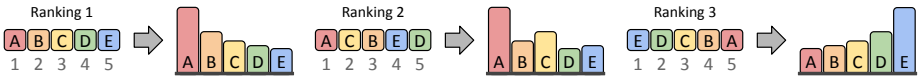

Figure 2.1: Three rankings and their normalized expected exposure distributions (Eq. 2.15) based on DCG weights (Eq. 2.2). According to our exposure-based divergence, ranking 1 and ranking 2 are quite similar despite only agreeing on the placing of document A. In contrast, ranking 1 and ranking 3 also agree on the placement of a single document (C) but have the highest possible dissimilarity, due to their highly mismatched exposure distributions.

With a probability of $1 - \delta$, the IPS estimate minus the risk term is a lower bound on the true utility of the policy:

$$P\big(U(\pi) \geq \hat{U}(\pi) - \text{Risk}(\delta)\big) > 1 - \delta. \tag{2.10}$$

Therefore, optimization of the lower bound can be more reliable than solely optimizing the IPS estimate ($\hat{U}(\pi)$), since it provides a high-confidence guarantee that a lower bound on the *true* utility of the policy is maximized.

In particular, Swaminathan and Joachims [148] propose using the sample variance as the risk factor:

$$\hat{U}_{\text{action-CRM}}(\pi) = \hat{U}_{\text{action}}(\pi) - \lambda \sqrt{\frac{1}{N} \text{Var}\big[\hat{U}_{\text{action}}(\pi)\big]}, \tag{2.11}$$

where $\lambda \in \mathbb{R}^{>0}$ is an alternative to the δ parameter that also determines how probable it provides a bound on the true utility. Importantly, this bound is based on an action-based IPS estimator. For our LTR setting this would translate to:

$$\hat{U}_{\text{action}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(y_i \mid q_i)}{\pi_0(y_i \mid q_i)} \sum_{d \in D_{q_i}} c_i(d). \tag{2.12}$$

However, action-based IPS estimation does not work well in the LTR setting because the very large number of possible rankings result in extremely small action propensities: $\pi_0(y_i \mid q_i)$, which creates a high-variance problem. As discussed in Section 2.3.2, for this reason CLTR uses exposure-based propensities instead (Eq. 2.6 and 2.6), as they effectively avoid extremely small values. As a result, the CRM approach by Swaminathan and Joachims [148] is not effective for CLTR, since the high-variance of its action-based IPS make the method impractical in the ranking setting.

Another downside of the CRM approach is that the computation of the sample-variance requires a full-pass over the training dataset, which is computationally costly for large-scale datasets. As a solution, Wu and Wang [165] introduce variational CRM (VCRM) which uses an upper bound on the variance term based on the Rényi divergence between the new policy and the logging policy [123]. This Rényi divergence is approximated via random sampling, thus making the VCRM method suitable for stochastic gradient descent-based training methods [102]. Nevertheless, this CRM approach still relies on action-based propensities, and therefore, does not provide an effective solution for the high-variance problem in CLTR.

2.4 A Novel Exposure-Based Generalization Bound for CLTR

In order to develop a CRM method for CLTR with safety gaurantees, we aim to find a risk term that gives us a generalization bound as in Eq. 2.10. Importantly, this bound has to be effective in the LTR setting, therefore, our approach should avoid action-based propensities.

We take inspiration from previous work by Wu and Wang [165], who use the fact that the Rényi divergence is an upper bound on the variance of an IPS estimator:

$$\mathrm{Var}\Big[\hat{U}_{\text{action}}(\pi)\Big] \leq d_2(\pi \,\|\, \pi_0), \tag{2.13}$$

where d_2 is the exponentiated Rényi divergence between the new policy and the logging policy [123]:

$$d_2(\pi \,\|\, \pi_0) = \mathbb{E}_q\left[\sum_y \left(\frac{\pi(y \mid q)}{\pi_0(y \mid q)}\right)^2 \pi_0(y \mid q)\right]. \tag{2.14}$$

In other words, the dissimilarity between the logging policy and a new policy can be used to bound the variance of the IPS estimate of the new policy's performance. However, because this divergence is based on action propensities, it is not effective in the LTR setting.

This section introduces a novel exposure-based measure of divergence that can produce a desired generalization bound for LTR optimization. Section 2.4.1 below introduces the concept of normalized exposure that treats rankings as exposure distributions. Subsequently, Section 2.4.2 proves that Rényi divergence based on normalized exposure can bound the variance of an exposure-based IPS estimator. Finally, Section 2.4.3 uses this novel variance bound to construct a proven generalization bound for CLTR.

2.4.1 Normalized expected exposure

Rényi divergence is only valid for probability distributions, e.g., $d_2(\pi \,\|\, \pi_0)$ with $\pi(y \mid q)$ and $\pi_0(y \mid q)$. However, expected exposure is not a probability distribution, i.e., the values of $\rho(d)$ (Eq. 2.5) or $\rho_0(d)$ (Eq. 2.6) do not necessarily sum up to one, over all documents to be ranked. This is because users generally examine more than a single item in a single displayed ranking [30], as a result, expected exposure can be seen as a distribution of multiple examinations. Our insight is that a valid probability distribution can be obtained by normalizing the expected exposure:

$$\rho'(d) = \frac{\rho(d)}{\sum_{d' \in D} \rho(d')} = \frac{\rho(d)}{\mathrm{Z}}, \tag{2.15}$$

where the normalization factor is a constant that only depends on K, the (truncated) ranking length:

$$\begin{aligned} \mathrm{Z} = \sum_{d \in D} \rho(d) &= \sum_{d \in D} \mathbb{E}_{y \sim \pi}\big[P\big(E = 1 \mid \text{rank}(d \mid y)\big)\big] \\ &= \mathbb{E}_{y \sim \pi}\Big[\sum_{d \in D} P\big(E = 1 \mid \text{rank}(d \mid y)\big)\Big] \\ &= \mathbb{E}_{y \sim \pi}\Big[\sum_{k=1}^{K} P\big(E = 1 \mid k\big)\Big] \\ &= \sum_{k=1}^{K} P\big(E = 1 \mid k\big). \end{aligned} \tag{2.16}$$

In this way, Z can be seen as the expected amount of examination that any ranking will receive, and ρ' as the probability distribution that indicates how it is expected to spread over documents.

An important property is that the ratio between two propensities is always equal to the ratio between their normalized counterparts:

$$\frac{\rho(d)}{\rho_0(d)} = \frac{\rho'(d)}{\rho'_0(d)}. \tag{2.17}$$

This is relevant to IPS estimation since it only requires the ratios between propensities, the proofs in the remainder of this chapter make use of this property.

Finally, using the normalized expected exposure, we can introduce the exponentiated exposure-based Rényi divergence:

$$d_2(\rho \,\|\, \rho_0) = \mathbb{E}_q\Big[\sum_{d \in D_q} \rho'_0(d)\left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2\Big]. \tag{2.18}$$

The key difference between our exposure-based divergence and action-based divergence is that it allows policies to be very similar, even when they have no overlap in the rankings they produce. As an intuitive example, Figure 2.1 displays three rankings and their associated normalized expected exposure distributions; these are the distributions for deterministic policies that give 100% probability to one of the rankings. Under action-based divergence, these policies would have the highest possible dissimilarity since they have no overlap in their possible actions, i.e., the rankings they give non-zero probability. In contrast, exposure-based divergence gives high similarity between ranking 1 and ranking 2, since the differences in their exposure distribution are minor. We note that these rankings still disagree on the placement of all documents except one. Conversely, for ranking 1 and ranking 3, which also only agree on a single document placement, exposure-based divergence gives the lowest possible similarity score because their exposure distributions are highly mismatched. Importantly, by solely considering differences in exposure distributions, exposure-based divergence naturally weighs differences at the bottom of rankings as less impactful than changes that affect the top. As a result, exposure-based divergence more closely corresponds with common ranking metrics (Eq. 2.1) than existing action-based divergences.

2.4.2 Exposure-divergence bound on variance

We now provide proof that exposure-based divergence is an upper bound on the variance of IPS estimators for CLTR.[1]

Theorem 2.4.1. *Given a ranking policy π and logging policy π_0, with the expected exposures $\rho(d)$ and $\rho_0(d)$ respectively, the variance of the exposure-based IPS estimate $\hat{U}(\pi)$ is upper-bounded by exposure-based divergence:*

$$\mathrm{Var}_{q,y,c}\Big[\hat{U}(\pi)\Big] \leq \frac{\mathrm{Z}}{N} d_2(\rho \,\|\, \rho_0) + \frac{1}{N}. \tag{2.19}$$

Proof. From the definition of $\hat{U}(\pi)$ (Eq. 2.6) and the assumption that queries q are independent and identically distributed (i.i.d), the variance of the counterfactual estimator can be expanded by applying the law of total variance as follows:

$$\mathrm{Var}_{q,y,c}\Big[\hat{U}(\pi)\Big] = \frac{1}{N}\Big(\mathbb{E}_q\Big[\mathrm{Var}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big] + \mathrm{Var}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big]\Big). \tag{2.20}$$

The second term (variance over queries) can be expanded as follows:

$$\begin{aligned}
\mathrm{Var}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big] &= \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]^2\Big] - \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big]^2 \\
&\leq \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]^2\Big] \\
&= \mathbb{E}_q\Big[[U(\pi) \mid q]^2\Big] \\
&\leq 1,
\end{aligned} \tag{2.21}$$

where in the second step, we use the unbiasedness property (Eq. 2.7) of the counterfactual estimator, and use the fact that the true utility is non-zero, i.e., $U(\pi) \geq 0$. In the last step, we make use of the fact that the true utility is bounded, and is upper bounded by 1. This is a safe assumption if the utility is normalized, as is the case for: normalized discounted cumulative gain or click-through rate.

Substituting it back in the counterfactual utility variance (Eq. 2.20), we get:

$$\mathrm{Var}_{q,y,c}\Big[\hat{U}(\pi)\Big] \leq \frac{1}{N}\Big(\mathbb{E}_q\Big[\mathrm{Var}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big] + 1\Big). \tag{2.22}$$

Next, since we have assumed a rank-based examination model (Section 2.3.2), the examinations of documents are independent. This allows us to rewrite the variance conditioned on a single query:

$$\mathrm{Var}_{y,c}\Big[\hat{U}(\pi \mid q)\Big] = \mathrm{Var}_{y,c}\left[\sum_{d \in D_q} \frac{\rho(d)}{\rho_0(d)} c(d,q)\right] \tag{2.23}$$

[1] The following proof differs slightly from the original proof published in [50], as we overlooked an additional constant term in the original paper.

$$= \sum_{d \in D_q} \mathrm{Var}_{y,c}\left[\frac{\rho(d)}{\rho_0(d)} c(d,q)\right]$$
$$\leq \sum_{d \in D_q} \mathbb{E}_{c,y}\left[\left(\frac{\rho(d)}{\rho_0(d)} c(d,q)\right)^2\right].$$

Since: $c(d,q)^2 = c(d,q)$, we can further rewrite to:

$$\sum_{d \in D_q} \mathbb{E}_{c,y}\left[\left(\frac{\rho(d)}{\rho_0(d)} c(d,q)\right)^2\right] = \sum_{d \in D_q} \mathbb{E}_{c,y}\left[\left(\frac{\rho(d)}{\rho_0(d)}\right)^2 c(d,q)\right] \tag{2.24}$$
$$= \sum_{d \in D_q} \left(\frac{\rho(d)}{\rho_0(d)}\right)^2 P(C = 1 \mid d, q, \pi_0).$$

Next, we use Eq. 2.3 and 2.6 to substitute the click probability; subsequently, we replace the examination propensities with normalized counterparts using Eq. 2.15 and 2.17; and lastly, we upper bound the result using the fact that $P(R = 1|d,q) \leq 1$:

$$\sum_{d \in D_q} \mathbb{E}_{c,y}\left[\left(\frac{\rho(d)}{\rho_0(d)} c(d,q)\right)^2\right] = \sum_{d \in D_q} \rho_0(d)\left(\frac{\rho(d)}{\rho_0(d)}\right)^2 P(R = 1 \mid d, q)$$
$$= \sum_{d \in D_q} \mathrm{Z}\, \rho'_0(d)\left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2 P(R = 1|d,q) \tag{2.25}$$
$$\leq \mathrm{Z} \sum_{d \in D_q} \rho'_0(d)\left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2.$$

Finally, we place this upper bound for a single query back into the expectation over all queries (Eq. 2.20):

$$\frac{1}{N}\mathbb{E}_q\Big[\mathrm{Var}_{y,c}\Big[\hat{U}(\pi) \mid q\Big]\Big] \leq \frac{\mathrm{Z}}{N}\mathbb{E}_q\Big[\sum_{d \in D_q} \rho'_0(d)\left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2\Big]. \tag{2.26}$$

Therefore, by Eq. 2.20, 2.26, 2.22, and the definition of exposure-based divergence in Eq. 2.18, it is a proven upper bound of the variance. □

2.4.3 Exposure-divergence bound on performance

Using the upper bound on the variance of an CLTR IPS estimator that was proven in Theorem 2.4.1, we can now introduce a generalization bound for the CLTR estimator.

Theorem 2.4.2. *Given the true utility $U(\pi)$ (Eq. 2.1) and its exposure-based IPS estimate $\hat{U}(\pi)$ (Eq. 2.6), for the ranking policy π and the logging policy π_0 with expected exposures $\rho(d)$ and $\rho_0(d)$, respectively, the following generalization bound*

holds with probability $1-\delta$*:*[2]

$$U(\pi) \geq \hat{U}(\pi) - \sqrt{\frac{\mathrm{Z}}{N}\Big(\frac{1-\delta}{\delta}\Big) d_2(\rho \,\|\, \rho_0)} - \sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}. \tag{2.27}$$

Proof. As per Cantelli's inequality [42], given an estimator $\hat{X}$ with the expected value $\mathbb{E}[\hat{X}]$ and variance $\mathrm{Var}[\hat{X}]$, the following tail-bound holds:

$$P(\hat{X} - \mathbb{E}[\hat{X}] \geq \lambda) \leq \frac{\mathrm{Var}[\hat{X}]}{\mathrm{Var}[\hat{X}] + \lambda^2}. \tag{2.28}$$

Since $\lambda > 0$ is a free parameter, we can define δ such that:

$$\delta = \frac{\mathrm{Var}[\hat{X}]}{\mathrm{Var}[\hat{X}] + \lambda^2}, \qquad \lambda = \sqrt{\frac{1-\delta}{\delta}\mathrm{Var}[\hat{X}]}. \tag{2.29}$$

Consequently, the following inequality holds:

$$P(\mathbb{E}[\hat{X}] \geq \hat{X} - \lambda) \geq 1 - \delta. \tag{2.30}$$

Building on this inequality, the following inequality must hold with probability $1-\delta$:

$$U(\pi) \geq \hat{U}(\pi) - \sqrt{\frac{1-\delta}{\delta}\mathrm{Var}_{q,y,c}\Big[\hat{U}(\pi)\Big]}. \tag{2.31}$$

Next, we replace the variance with the upper bound from Theorem 2.4.1, which results in the following bound:

$$U(\pi) \geq \hat{U}(\pi) - \sqrt{\frac{\mathrm{Z}}{N}\Big(\frac{1-\delta}{\delta}\Big) d_2(\rho \,\|\, \rho_0) + \Big(\frac{1-\delta}{\delta N}\Big)}. \tag{2.32}$$

By applying the Cauchy–Schwarz inequality, we get:

$$U(\pi) \geq \hat{U}(\pi) - \sqrt{\frac{\mathrm{Z}}{N}\Big(\frac{1-\delta}{\delta}\Big) d_2(\rho \,\|\, \rho_0)} - \sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}. \tag{2.33}$$

This completes the proof. □

Risk in CLTR. Based on the generalization bound proposed in Theorem 2.4.2, we see that it proposes the following measure of risk: $\mathrm{Risk}(\delta) = \sqrt{\frac{\mathrm{Z}}{N}\big(\frac{1-\delta}{\delta}\big) d_2(\rho \,\|\, \rho_0)}$ (cf. Eq. 2.10). Clearly, this risk is mostly determined by the exposure-based divergence between the new policy and the logging policy. Thereby, it states that the greater the difference between how exposure is spread over documents by the logging policy and the new policy, the higher the risk involved. Therefore, to optimize this lower bound, one has to balance the maximization of the estimated utility $\hat{U}(\pi)$ and the minimization of risk by not letting π differ too much from π_0 in terms of exposure.

[2] The following proof differs slightly from the original proof published in [50], as we overlooked an additional constant term in the original paper.

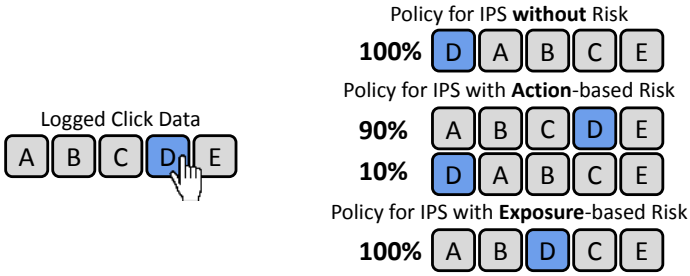

Figure 2.2: Example comparison of the optimal policy for a single logged click according to three different risk estimators.

Furthermore, we see that our measure of risk diminishes as N increases. As a result, the risk term will overwhelm the IPS term when N is very low, as there is much risk involved when estimating based on a few interactions. Conversely, when N is very large, the risk term mostly disappears, as the IPS estimate is more reliable when based on large numbers of interactions. Thus, during optimization, the generalization bound is expected to mostly help with avoiding initial decreases in performance, while still converging at the same place as the standard IPS estimator.

Lastly, the δ parameter determines the *safety* that is provided by the risk, where a lower δ makes it more likely that the generalization bound holds. Accordingly, as δ increases the risk term becomes smaller and will thus have less effect on optimization.

To the best of our knowledge, this is the first exposure-based generalization bound, which makes it the first method designed for safe optimization in the CLTR setting.

Illustrative comparison. To emphasize the working and novelty of our exposure-based risk, a comparison of the optimal policies for action-based risk, exposure-based risk, and no risk are shown in Figure 2.2. We see that IPS without a risk term places the once-clicked document at the first position, with 100% probability. This is very risky, as it greatly impacts the ranking while only being based on a single observation. The action-based risk tries to mitigate this risk with a probabilistic policy that gives most probability to the logging policy ranking (90%) and the remainder to the IPS ranking (10%). In contrast, with exposure-based risk, the optimal policy makes the risk and utility trade-off in a single ranking, that mostly follows the logging policy but places the clicked document slightly higher.

This example illustrates that because action-based risk does not have a similarity measure between rankings, it can only produce a probabilistic interpolation between the logging policy and IPS rankings. Alternatively, because exposure-based risk does have such a measure, it produces a ranking that is neither the logging ranking nor the IPS ranking, but one with an exposure distribution that is similar to both. Thereby, exposure-based risk has a more elegant and natural method of balancing utility maximization and risk minimization in the CLTR setting.

2.5 A Novel Counterfactual Risk Minimization Method for LTR

Now that we have the proven generalization bound described in Section 2.4.3 (Theorem 2.4.2), we can propose a novel risk-aware CLTR method for optimizing it. Accordingly, the aim of our method is to find the policy that maximizes this high-confidence lower bound on the true performance. In formal terms, we have the following optimization problem:

$$\max_{\pi} \hat{U}(\pi) - \sqrt{\frac{\mathrm{Z}}{N}\Big(\frac{1-\delta}{\delta}\Big) d_2(\rho \,\|\, \rho_0)}. \tag{2.34}$$

Note that we ignore the term $\sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}$, since it is constant with respect to the policy π, and therefore can be disregarded for optimization purposes. We propose to train a stochastic policy π via stochastic gradient descent, therefore, we need to derive the gradient and find a method of computing it. For the computation of the gradient w.r.t. the utility $\hat{U}(\pi)$, the first part of Eq. 2.34, we refer to several prior work that discusses this topic extensively [104, 108, 169]. Thus, we can focus our attention on the second part of Eq. 2.34:

$$\nabla_\pi \sqrt{\frac{\mathrm{Z}}{N}\Big(\frac{1-\delta}{\delta}\Big) d_2(\rho \,\|\, \rho_0)} = \sqrt{\frac{\mathrm{Z}(1-\delta)}{4N\delta d_2(\rho \,\|\, \rho_0)}} \nabla_\pi d_2(\rho \,\|\, \rho_0). \tag{2.35}$$

To derive the gradient of the exposure-based divergence function, we use the relation between ρ and ρ' from Eq. 2.16 and 2.17:

$$\begin{aligned} \nabla_\pi d_2(\rho \,\|\, \rho_0) &= \nabla_\pi \mathbb{E}_q \Bigg[\sum_{d \in D_q} \rho'_0(d) \left(\frac{\rho'(d)}{\rho'_0(d)} \right)^2 \Bigg] \\ &= \frac{2}{\mathrm{Z}} \mathbb{E}_q \Bigg[\sum_{d \in D_q} \frac{\rho(d)}{\rho_0(d)} \nabla_\pi \rho(d) \Bigg]. \end{aligned} \tag{2.36}$$

Thus, we only need the gradient w.r.t. the exposure of a document ($\nabla_\pi \rho(d)$) to complete our derivation. If π is a Plackett-Luce (PL) ranking model, one can make use of the specialized gradient computation algorithm from [104]. However, for this chapter, we will not make further assumptions about π and apply the more general log-derivate trick from the REINFORCE algorithm [164]:

$$\nabla_\pi \rho(d) = \mathbb{E}_{y \sim \pi} \big[P\big(E = 1 \mid \mathrm{rank}(d \mid y)\big) \big] \nabla_\pi \log \pi(y). \tag{2.37}$$

Putting all of the previous elements back together, gives us the gradient w.r.t. the exposure-based risk function:

$$\sqrt{\frac{1-\delta}{N\delta\, \mathrm{Z}\, d_2(\rho \,\|\, \rho_0)}} \mathbb{E}_{q, y \sim \pi} \left[\left(\sum_{k=1}^{K} \frac{\rho(y_k)}{\rho_0(y_k)} P(E = 1 | \, k) \right) \nabla_\pi \log \pi(y) \right], \tag{2.38}$$

where y_k is the document at rank k in ranking y. For a close approximation of this gradient, we substitute the gradient with the queries from the given dataset, and the rankings sampled from π during optimization [104, 164].

Similarly, since the exact computation of is $d_2(\rho \,\|\, \rho_0)$ infeasible in practice, we introduce a sample-based empirical divergence estimator:

$$\hat{d}_2(\rho \,||\, \rho_0) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D_{q_i}} \rho'_0(d) \left(\frac{\rho'(d)}{\rho'_0(d)} \right)^2. \tag{2.39}$$

This is an unbiased estimate of the true divergence given that the sampling process is truly Monte Carlo [64].

2.6 Experimental Setup

For our experiments, we follow the semi-synthetic experimental setup that is common in the CLTR literature [78, 109, 110, 152]. We make use of the three largest publicly available LTR datasets: Yahoo! Webscope [21], MSLR-WEB30k [115], and Istella [31]. The datasets consist of queries, a preselected list of documents per query, query-document feature vectors, and manually-graded relevance judgements for each query-document pair. To generate clicks, we follow previous work [109, 110, 152] and train a logging policy on a 3% fraction of the relevance judgements. This simulates a real-world setting, where a production ranker trained on manual judgements is used to collect click logs, which can then be used for subsequent click-based optimization. Typically, in real-world ranking settings, given that the production ranker is used on live-traffic, it is deemed as a safe policy that can be trusted with real users.

We simulate a top-K ranking setup [108] where five documents are presented at once. Clicks are generated with our assumed click model (Eq. 2.3) and the following rank-based position-bias:

$$P(E = 1 \mid q, d, y) = \begin{cases} \left(\frac{1}{\text{rank}(d|y)} \right)^2 & \text{if rank}(d \mid y) \leq 5, \\ 0 & \text{otherwise.} \end{cases} \tag{2.40}$$

In real-world click data, the observed CTR is typically very low [24, 87, 129]; hence, to simulate such a sparse click settings, we apply the following transformation from relevance judgements to relevance probabilities:

$$P(R = 1 \mid q, d) = 0.025 * rel(q, d) + 0.2, \tag{2.41}$$

where $rel(q, d) \in \{0, 1, 2, 3, 4\}$ is the relevance judgement for the query-document pair and 0.2 is added as click noise. During training, the only available data consists of clicks generated on the training and validation sets, no baseline method has access to the underlying relevance judgements (except the skyline).

Furthermore, we assume a setting where the exact logging policy is not available during training. As a result, the $\hat{\rho}_0$ propensities have to estimated, we use a simple

frequency estimate following [110]:

$$\hat{\rho}_0(d) = \sum_{i=1}^{N} \frac{\mathbb{1}\big[q = q_i\big]}{\sum_{j=1}^{N} \mathbb{1}\big[q = q_j\big]} P\big(E = 1 \mid \text{rank}(d \mid y_i)\big). \tag{2.42}$$

For the action-based baselines, the action propensities $\hat{\pi}_0(y \mid q)$ are similarly estimated based on observed frequencies:

$$\hat{\pi}_0(y \,|\, q) = \prod_{k=1}^{K-1} \hat{\pi}_0(y_k \,|\, q),\ \hat{\pi}_0(y_k \,|\, q) = \sum_{j=1}^{N} \frac{\mathbb{1}\big[y_k = y_j\big]}{\sum_{j=1}^{N} \mathbb{1}\big[q = q_j\big]}, \tag{2.43}$$

where $\hat{\pi}_0(y_k \,|\, q)$ is the estimated probability of d appearing at rank k for query q. As is common in CLTR [78, 103, 127], we clip propensities by $10/\sqrt{N}$ in the training set, to reduce variance, but not in the validation set.

We optimize neural PL ranking models [104] with early stopping based on validation clicks to prevent overfitting. For the REINFORCE policy-gradient, we follow [169] and use the average reward per query as a control-variate for variance reduction.

As our evaluation metric, we compute NDCG@5 metric using the relevance judgements on the test split of each dataset [65]. All reported results are averages over ten independent runs, significant testing is performed with a two-sided student-t test.

Finally, the following methods are included in our comparisons:

(i) *Naive*. As the most basic baseline, we train on the generated clicks without any correction (equivalent to $\forall d,\ \rho_0(d) = 1$).

(ii) *Skyline.* To compare with the highest possible performance, this baseline is trained on the actual relevance judgements.

(iii) *Action-based IPS.* Standard IPS estimation (Eq. 2.12) that is not designed for ranking and thus uses action-based propensities.

(iv) *Action-based CRM.* Standard CRM (Eq. 2.11) that is also not designed for ranking, for the risk function we use the action-based divergence function in Eq. 2.14.

(v) *Exposure-based IPS*. The IPS estimator designed for CLTR with exposure-based propensities (Eq. 2.6). The most important baseline, as it is the prevalent approach in the field [108, 110].

(vi) *Exposure-based CRM.* Our proposed CRM method (Eq. 2.34) using a risk function based on exposure-based divergence.

2.7 Results and Discussion

2.7.1 Comparison with baseline methods

The main results of our experimental comparison are presented in Figure 2.3 and 2.4, and Table 2.1 and 2.2. Figure 2.3 and 2.4 display the performance curves of the different

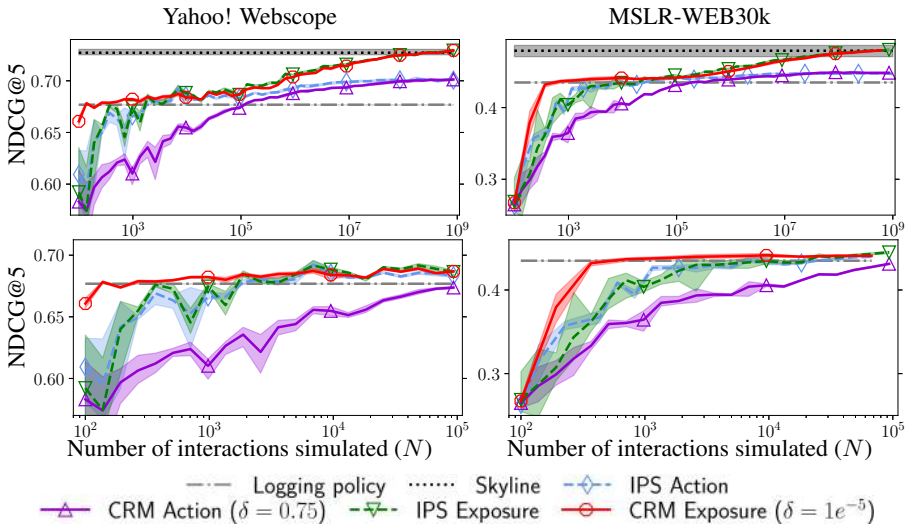

Figure 2.3: Performance in NDCG@5 of various IPS and CRM methods for CLTR on Yahoo! Webscope and MSLR-WEB30k datasets. The top-row presents the results when the size of the training data is varied from 10^2 to 10^9. The bottom-row is a zoomed-in view, focusing on the low-data region from 10^2 to 10^5. Results are averages over 10 runs; shaded areas indicate 80% confidence intervals.

methods as the number of logged interactions (N) increases. Table 2.1 and 2.2 present the performance at $N \in \{4 \cdot 10^2, 4 \cdot 10^7, 10^9\}$ and indicate whether the observed differences with our exposure-based CRM method are statistically significant.

We start by considering the performance curves in Figure 2.3 and 2.4. We see that both the action-based and exposure-based IPS baselines have an initial period of very similar performance that is far below the logging policy. Around $N \approx 10^4$ their performance is comparable to the logging policy, and finally at $N = 10^9$ the exposure-based IPS has reached optimal performance, while the performance of action-based IPS is still far from optimal. We can attribute this initial poor performance to the high variance problem of IPS estimation; when N is small, variance is at its highest, resulting in risky and sub-optimal optimization by the IPS estimators. However, even when $N = 10^9$, the variance of the action-based IPS estimator is too high to reach optimal performance, due to its extremely small propensities. This illustrates why the introduction of exposure-based propensities was so important to the CLTR field, and that even exposure-based IPS produces unsafe optimization when little data is available or variance from interactions is high.

Next, we consider whether action-based CRM is able to mitigate the high variance problem of action-based IPS. Despite being a proven generalization bound, Figure 2.3 and 2.4 clearly show us that action-based CRM only leads to decreases in performance compared to its IPS counterpart. It appears that this happens because the logging policy is not available in our setup, and the propensities have to be estimated from logged

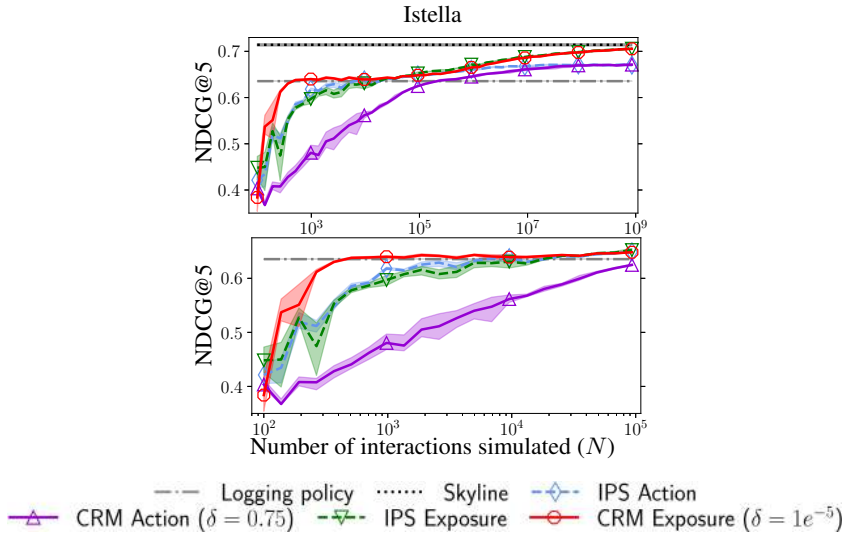

Figure 2.4: Performance in NDCG@5 of various IPS and CRM methods for CLTR on the Istella dataset. The top-row presents the results when the size of the training data is varied from 10^2 to 10^9. The bottom-row is a zoomed-in view, focusing on the low-data region from 10^2 to 10^5. Results are averages over 10 runs; shaded areas indicate 80% confidence intervals.

data. Consequently, the action-based risk pushes the optimization to mimic the exact rankings that were observed during logging. Thus, due to the variance introduced from the sampling of rankings from the logging policy, it appears that action-based CRM has an even higher variance problem than action-based IPS. As expected, our results thus clearly indicate that action-based CRM is also unsuited for the CLTR setting, to our surprise; it is substantially worse than its IPS counterpart.

Finally, we examine the performance of our novel exposure-based CRM method. Similar to the other methods, there is an initial period of low performance, but in stark contrast, this period ends very quickly; on Yahoo! logging policy performance is reached when $N \approx 125$, on MSLR-WEB30k when $N \approx 350$ and on Istella when $N \approx 400$. For comparison, exposure-based IPS needs $N \approx 1100$ on Yahoo!, $N \approx 10^4$ on MLSR-WEB30k and $N \approx 1.1 \cdot 10^4$ on Istella to do the same; meaning that our CRM method needs roughly 89%, 97% and 97% fewer interactions, respectively. In addition, Table 2.1 and 2.2 indicate that the logging policy performance is matched on all datasets when $N = 400$ by exposure-based CRM, where it also outperforms all baseline methods. We note that there is still an initial period of low performance, because the logging policy is unavailable at training, and thus, its behavior still has to be estimated from logged interactions. It is possible that in settings where the logging policy is fully known during training, this initial period is eliminated entirely. Nevertheless, our results show that exposure-based CRM reduces the initial periods of poor performance due to variance by an enormous magnitude.

Table 2.1: NDCG@5 performance for Yahoo! Webscope and MSLR-WEB30k datasets under different settings for several values of N, the number of logged interactions in the simulated training set. Values are averages over 10 independent runs on the held-out test sets; bold figures mark the highest score. Differences from the exposure-based CRM are assessed with a two-sided Student's *t*-test: $^{\blacktriangledown}$ denotes significantly lower ($p < 0.01$), while * indicates no significant difference.

	Yahoo! Webscope			MSLR-WEB30k		
	$N = 4{\cdot}10^2$	$N = 4{\cdot}10^7$	$N = 10^9$	$N = 4{\cdot}10^2$	$N = 4{\cdot}10^7$	$N = 10^9$
Logging	0.677	0.677	0.677	0.435	0.435	0.435
Skyline	0.727	0.727	0.727	0.479	0.479	0.479
Naive	0.652 (0.021)$^{\blacktriangledown}$	0.694 (0.000)$^{\blacktriangledown}$	0.695 (0.000)$^{\blacktriangledown}$	0.353 (0.003)$^{\blacktriangledown}$	0.448 (0.000)$^{\blacktriangledown}$	0.448 (0.001)$^{\blacktriangledown}$
Action IPS	0.656 (0.008)$^{\blacktriangledown}$	0.701 (0.001)$^{\blacktriangledown}$	0.701 (0.001)$^{\blacktriangledown}$	0.359 (0.007)$^{\blacktriangledown}$	0.448 (0.001)$^{\blacktriangledown}$	0.448 (0.001)$^{\blacktriangledown}$
Action CRM	0.617 (0.004)$^{\blacktriangledown}$	0.698 (0.001)$^{\blacktriangledown}$	0.700 (0.001)$^{\blacktriangledown}$	0.359 (0.005)$^{\blacktriangledown}$	0.448 (0.001)$^{\blacktriangledown}$	0.449 (0.001)$^{\blacktriangledown}$
Exp. IPS	0.659 (0.010)$^{\blacktriangledown}$	**0.723 (0.001)***	**0.730 (0.001)***	0.389 (0.014)$^{\blacktriangledown}$	**0.474 (0.001)***	**0.481 (0.001)***
Exp. CRM	**0.677 (0.001)**	**0.723 (0.001)**	**0.730 (0.000)**	**0.434 (0.001)**	**0.473 (0.001)**	**0.480 (0.001)**

Table 2.2: NDCG@5 performance for Istella dataset under different settings for several values of N, the number of logged interactions in the simulated training set. Reported numbers are averages over 10 independent runs evaluated on the held-out test-sets; bold numbers indicate the highest performance. Statistical significance for differences with the exposure-based CRM are measured via a two-sided student-t test: $^{\blacktriangledown}$ indicates methods with significantly lower NDCG ($p < 0.01$), and * no significant difference.

	Istella		
	$N = 4 \cdot 10^2$	$N = 4 \cdot 10^7$	$N = 10^9$
Logging	0.635	0.635	0.635
Skyline	0.714	0.714	0.714
Naive	0.583 (0.007)$^{\blacktriangledown}$	0.661 (0.001)$^{\blacktriangledown}$	0.661 (0.001)$^{\blacktriangledown}$
Action IPS	0.578 (0.004)$^{\blacktriangledown}$	0.671 (0.001)$^{\blacktriangledown}$	0.671 (0.002)$^{\blacktriangledown}$
Action CRM	0.449 (0.013)$^{\blacktriangledown}$	0.668 (0.002)$^{\blacktriangledown}$	0.672 (0.001)$^{\blacktriangledown}$
Exp. IPS	0.576 (0.010)$^{\blacktriangledown}$	**0.696 (0.001)***	**0.706 (0.001)***
Exp. CRM	**0.635 (0.001)**	**0.695 (0.001)**	**0.706 (0.001)**

Furthermore, while the initial period is clearly improved, we should also consider whether there is a trade-off with the rate of convergence. Surprisingly, Figure 2.3 and 2.4 do not display any noticeable decrease in performance when compared with exposure-based IPS. Moreover, Table 2.1 and 2.2 show the differences between exposure-based IPS and CRM are barely measurable and not statistically significant when $N \in \{4 \cdot 10^7, 10^9\}$. We know from the risk formulation in Eq. 2.34 that the weight of the risk term decreases as N increases at a rate of $1/\sqrt{N}$. In other words, the more data is available, the more optimization is able to diverge from the logging policy. It appears that this balances utility maximization and risk minimization so well that we are unable to observe any downside of applying exposure-based CRM instead of IPS. Therefore, we conclude that, compared to all baseline methods and across all datasets, exposure-

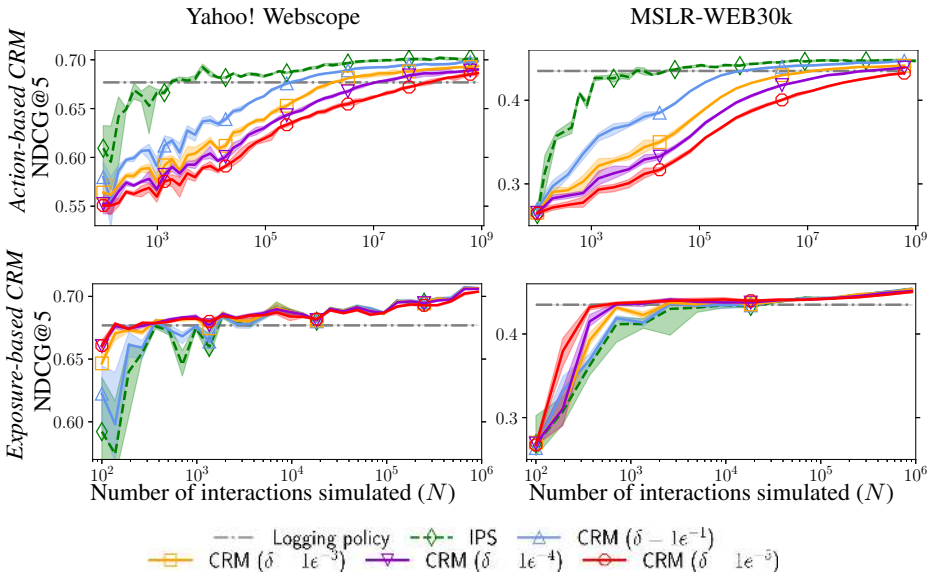

Figure 2.5: Performance of CRM methods with varying confidence parameters (δ) on Yahoo! Webscope and MSLR-WEB30k datasets. Top-row: action-based CRM baseline; bottom-row: our exposure-based CRM method. Results are averages of 10 runs; shaded areas indicate 80% confidence intervals.

based CRM drastically reduces the initial period of low performance, matches the best rate of convergence of all baseline, and has optimal performance at convergence.

2.7.2 Ablation study on the confidence parameter

To gain insights into how the confidence parameter δ affects the trade-off between safety and utility, an ablation study over various δ values was performed for both CRM methods.

The top-rows of Figure 2.5 and 2.6 show us the performance of action-based CRM, and contrary to expectation, a decrease in δ corresponds to a considerably worse performance. For the sake of clarity, in theory, δ is inversely tied to safety, a lower δ should result in less divergence from the safe logging policy [109]. Conversely, we see that action-based CRM displays the opposite trend. We think this further confirms our hypothesis that a frequency estimate of action-based divergence has an even higher variance problem than action-based IPS. Consequently, a higher weight to the risk function results in worse performance. This further confirms our previous conclusion that action-based CRM is unsuited for the CLTR setting, regardless of how the δ parameter is tuned.

In contrast, the bottom-rows of Figure 2.5 and 2.6 display the expected trend for exposure-based CRM; as δ decreases the resulting performance gets closer to the logging policy. With $\delta = 0.1$, CRM performs extremely close to its IPS counterpart, as optimization is less constrained to mimic the logging policy here. Decreasing δ

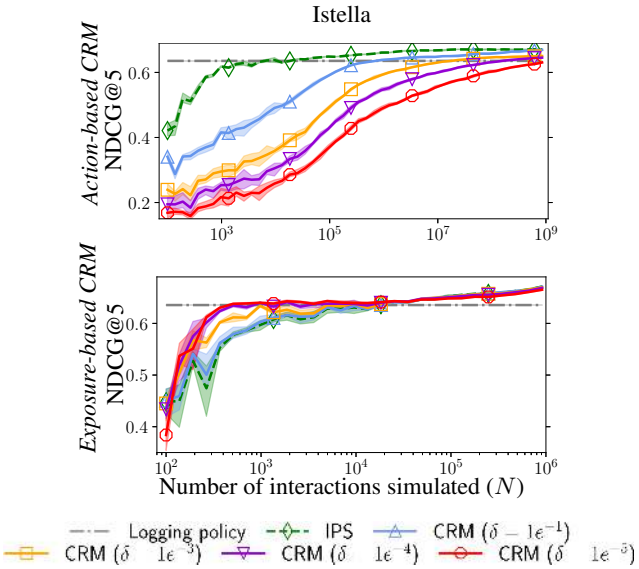

Figure 2.6: Performance of CRM methods with varying confidence parameters (δ) on the Istella dataset. Top-row: action-based CRM baseline; bottom-row: our exposure-based CRM method. Results are averages of 10 runs; shaded areas indicate 80% confidence intervals.

appears to have diminishing returns, as the difference between $\delta = 10^{-4}$ and $\delta = 10^{-5}$ is marginal. Importantly, we do not observe any downsides to setting $\delta = 10^{-5}$, thus we have not reached a point in our experiments where δ is set too conservatively. This suggests that exposure-based CRM is very robust to the setting of the δ parameter, and that a sufficiently low δ does not require fine-tuning. Therefore, this shows that the improvements we observed when comparing with baseline methods, did not stem from a fine-tuning of δ. Thus, we can conclude that this robustness further increases the safety that is provided by exposure-based CRM, as there is also little risk involved in the tuning of the δ parameter.

2.8 Conclusion

In this chapter, we introduced the first counterfactual risk minimization (CRM) method designed for CLTR, that relies on a novel exposure-based divergence function. In contrast with existing action-based CRM methods, exposure-based divergence avoids the problem of the enormous combinatorial action space when ranking, by measuring the dissimilarity between policies based on how they distribute exposure to documents. As a result, exposure-based CRM optimization produces policies that rank similar to the logging policy when it is risky to follow IPS, i.e., when little data is available or variance is very high. Consequently, our experimental results show that it almost completely removes initial periods of detrimental performance; to be precise, our method needed

89% to 97% fewer interactions than state-of-the-art IPS to match production system performance. Importantly, we observed no downsides in its application, as it maintained the same rate and point of convergence as IPS, in all tested experimental settings. Therefore, we conclude that our exposure-based CRM method provides the safest CLTR methods so far, as it almost completely alleviates the risk of decreasing the performance of a production system.

These improvements have big implications for practitioners who work on ranking systems in real-world settings, since the almost complete reduction of initial detrimental performance removes the main risks involved in applying CLTR. In other words, when applying our novel exposure-based CRM, practitioners can have significantly less worry that the resulting policy will perform worse than their production system and hurt user experience.

In this chapter, we answer the broad research question (RQ1) in affirmative. We derived a generalization bound for the counterfactual LTR estimator, establishing a lower bound on the true ranking utility. We then demonstrate that optimizing this lower bound ensures safety, i.e., the resulting ranking policy after optimizing the lower bound is no worse than the current production policy. In practice, this property is useful when click data is scarce, mitigating the risk of deploying potentially harmful policies, thereby ensuring safe deployment.

The safety method presented in this chapter depends on the assumed user model (position-based click model), and relies on the assumption that user interaction data will follow the click model. In settings where user interactions do not follow the assumed user model, the safety guarantees will not hold. In the next chapter, we will discuss a method that guarantees safety even under adversarial user behavior settings.

3

Practical and Robust Safety Guarantees for Advanced Counterfactual Learning-to-Rank

In Chapter 2, we presented a safe counterfactual LTR method that guarantees safe deployment by optimizing a lower confidence bound on the true ranking utility. By optimizing the lower confidence bound on the true ranking utility via exposure-based risk minimization, is it guaranteed that the new ranking policy will be at least as good as the production/logging ranking policy. However, these safety guarantees depend critically on assumptions regarding user behavior, i.e., the assumed click model. If the user behavior deviates from the assumed click model, the safety guarantees will be invalidated; which motivates the following research question:

RQ2 Can we provide robust safety guarantees for counterfactual LTR policies even under adversarial user behavior settings?

In this chapter, we introduce PRPO, a novel safe deployment method ensuring safety for counterfactual LTR without reliance on user behavior assumptions, guaranteeing robust safety even under adversarial conditions. Further, we extend the safety guarantees from position-based click model and IPS estimator introduced in Chapter 3 to trust-bias click model [4, 152] and doubly robust counterfactual estimator [107].

3.1 Introduction

CLTR [51, 78, 103, 157] concerns the optimization of ranking systems based on user interaction data using LTR methods [91]. A main advantage of CLTR is that it does not require manual relevance labels, which are costly to produce [21, 116] and often do not align with actual user preferences [132]. Nevertheless, CLTR also brings significant challenges since user interactions only provide a heavily biased form of implicit feedback [51]. User clicks are affected by many different factors, for example, the position at which an item is displayed in a ranking [30, 158]. Thus, click frequencies provide

This chapter was published as [53].

a biased indication of relevance, that is often more representative of how an item was displayed than actual user preferences [4, 157].

To correct for this bias, early CLTR applied IPS, which weights clicks inversely to the estimated effect of position bias [78, 157]. Later work expanded this approach to correct for other forms of bias, e.g., item-selection bias [108, 112] and trust bias [4, 152], and more advanced doubly robust (DR) estimation [107]. Using these methods, standard CLTR aims to create an unbiased estimate of relevance (or user preference) from click frequencies. In other words, their goal is to output an estimate per document with an expected value that is equal to their relevance.

However, unbiased estimates of CLTR have their limitations. Firstly, they assume a model of user behavior and require an accurate estimate of this model. If the assumed model is incorrect [108, 152] or its estimated parameters are inaccurate [78, 107], then their unbiasedness is not guaranteed. Secondly, even when unbiased, the estimates are subject to variance [106]. As a result, the actual estimated values are often erroneous, especially when the available data is sparse [50, 107]. Accordingly, unbiased CLTR does not guarantee that the ranking models it produces have optimal performance [106].

Safe counterfactual learning to rank. There are risks involved in applying CLTR in practice. In particular, there is a substantial risk that a learned ranking model is deployed that degrades performance compared to the previous production system [50, 63, 109]. This can have negative consequences to important business metrics, making CLTR less attractive to practitioners. To remedy this issue, a *safe* CLTR approach was proposed by Gupta et al. [50]; see Chapter 2. Their approach builds on IPS-based CLTR and adds exposure-based risk regularization, which keeps the learned model from deviating too much from a given safe model. Thereby, under the assumption of a position-biased user model, the safe CLTR approach can guarantee an upper bound on the probability of the model being worse than the safe model.

Limitations of the current safe CLTR method. Whilst safe CLTR is an important contribution to the field, it has two severe limitations – both are addressed by this chapter. Firstly, the existing approach is only applicable to IPS estimation, which is no longer the state-of-the-art in the field [51], and it assumes a rank-based position bias model [30, 157], the most basic user behavior model in the field. Secondly, because its guarantees rely on assumptions about user behavior, it can only provide a conditional notion of safety. Moreover, since user behavior can be extremely heterogeneous, it is unclear whether a practitioner could even determine whether the safety guarantees would apply to their application.

Main contributions. Our first contribution in this chapter addresses the mismatch between the existing safe CLTR approach and recent advances in CLTR. We propose a novel generalization of the exposure-based regularization term that provides safety guarantees for both IPS and DR estimation, also under more complex models of user behavior that cover both position and trust bias. Our experimental results show that our novel method reaches higher levels of performance significantly faster, while avoiding any notable decreases of performance. This is especially beneficial since DR is known to have detrimental performance when very little data is available [107].

Our second contribution in this chapter provides an unconditional notion of safety. We take inspiration from advances in reinforcement learning (RL) [89, 117, 137, 160,

161] and propose the novel *proximal ranking policy optimization* (PRPO) method. PRPO removes incentives for LTR methods to rank documents too much higher than a given safe ranking model would. Thereby, PRPO imposes a limit on the performance difference between a learned model and a safe model, in terms of standard ranking metrics. Importantly, PRPO is easily applicable to *any* gradient-descent-based LTR method, and makes *no assumptions* about user behavior. In our experiments, PRPO prevents any notable decrease in performance even under extremely adversarial circumstances, where other methods fail. Therefore, we believe PRPO is the first *unconditionally* safe LTR method.

Together, our contributions in this chapter bring important advances to the theory of safe CLTR, by proposing a significant generalization of the existing approach with theoretical guarantees, and the practical appeal of CLTR, with the first robustly safe LTR method: PRPO. All source code to reproduce our experimental results is available at: `https://github.com/shashankg7/cikm-safeultr`.

3.2 Related Work

Counterfactual learning to rank. Joachims et al. [78] introduced the first method for CLTR, a LTR specific adaptation of IPS from the bandit literature [48, 49, 52, 77, 127, 148] to correct for position bias. They weight each user interaction according to the inverse of its examination probability, i.e., its inverse propensity, during learning to correct for the position bias in the logged data. This weighting will remove the effect of position bias from the final ranking policy. Oosterhuis and de Rijke [108] extended this method for the top-K ranking setting with item-selection bias, where any item placed outside the top-K positions gets zero exposure probability, i.e., an extreme form of position bias. They proposed a policy-aware propensity estimator, where the propensity weights used in IPS are conditioned on the logging policy used to collect the data.

Agarwal et al. [4] introduced an extension of IPS, known as Bayes-IPS, to correct for *trust bias*, an extension of position-bias, with false-positive clicks at the higher ranks, because of the users' trust in the search engine. Vardasbi et al. [152] proved that Bayes-IPS cannot correct for trust bias and introduced an affine-correction method and unbiased estimator. Oosterhuis and de Rijke [110] combined the affine-correction with a policy-aware propensity estimator to correct for trust bias and item-selection bias simultaneously. Recently, Oosterhuis [107] introduced a DR-estimator for CLTR, which combines the existing IPS-estimator with a regression model to overcome some of the challenges with the IPS-estimator. The proposed DR-estimator corrects for item-selection and trust biases, with lower variance and improved sample complexity.

Safe policy learning from user interactions. In the context of offline evaluation for contextual bandits, Thomas et al. [149] introduced a high-confidence off-policy evaluation framework. A confidence interval is defined around the empirical off-policy estimates, and there is a high probability that the *true* utility can be found in the interval. Jagerman et al. [63] extended this framework for safe deployment in the contextual bandit learning setup. The authors introduce a SEA method that selects with high confidence between a safe behavior policy and the newly learned policy. In the context of LTR, Oosterhuis and de Rijke [109] introduced the generalization

and specialization (GENSPEC) method, which safely selects between a feature-based and tabular LTR model. For off-policy learning, Swaminathan and Joachims [148] introduced a CRM framework for the contextual bandit setup. They modify the IPS objective for bandits to include a regularization term, which explicitly controls for the variance of the IPS-estimator during learning, thereby overcoming some of the problems with the high-variance of IPS. Wu and Wang [165] extended the CRM framework by using a *risk* regularization, which penalizes mismatches in the action probabilities under the new policy and the behavior policy. In the previous chapter (Chapter 2) we made this general safe deployment framework effective in the LTR setting. We proposed an exposure-based risk regularization method where the difference in the document exposure distribution under the new and logging policies is penalized. When click data is limited, risk regularization ensures that the performance of the new policy is similar to the logging policy, ensuring safety.

To the best of our knowledge, the methodology proposed in Chapter 2 is the only method for safe policy learning in the LTR setting. While it guarantees safe ranking policy optimization, it has two main limitations:

(i) It is only applicable to the IPS estimator; and

(ii) It is only applicable under the position-based click model assumption, the most basic click model in the CLTR literature [51, 78, 103].

Proximal policy optimization. In the broader context of RL, *proximal policy optimization* (PPO) was introduced as a policy gradient method for training RL agents to maximize long-term rewards [89, 117, 137, 160, 161]. PPO clips the importance sampling ratio of action probability under the new policy and the current behavior policy, and thereby, it prevents the new policy to deviate from the behavior policy by more than a certain margin. PPO is not directly applicable to LTR, for the same reasons that the CRM framework is not: the combinatorial action space of LTR leads to extremely small propensities that PPO cannot effectively manage [50].

3.3 Background

3.3.1 Learning to rank

The goal in LTR is to find a ranking policy (π) that optimizes a given ranking metric [91]. Formally, given a set of documents (D), a distribution of queries Q, and the true relevance function ($P(R = 1 \mid d)$), LTR aims to maximize the following utility function:

$$U(\pi) = \sum_{q \in Q} P(q \mid Q) \sum_{d \in D} \omega(d \mid \pi)\ P(R = 1 \mid d), \tag{3.1}$$

where $\omega(d \mid \pi)$ is the weight of the document for a given policy π. The weight can be set accordingly to optimize for a given ranking objective, for example, setting the weight to:

$$\omega_{\text{DCG}}(d \mid q, \pi) = \mathbb{E}_{y \sim \pi(\cdot \mid q)} \left[(\log_2(\text{rank}(d \mid y) + 1))^{-1} \right], \tag{3.2}$$

optimizes discounted cumulative gain (DCG) [65]. For this chapter, we aim to optimize the expected number of clicks, so we set the weight accordingly [50, 107, 169].

3.3.2 Assumptions about user click behavior

The optimization of the true utility function (Eq. 3.1) requires access to the document relevance ($P(R = 1 \mid d)$). In the CLTR setting, the relevances of documents are not available, and instead, click interaction data is used to estimate them [78, 103, 157]. However, naively using clicks to optimize a ranking system can lead to sub-optimal ranking policies, as clicks are a biased indicator of relevance [28, 30, 76, 78]. CLTR work with theoretical guarantees starts by assuming a model of user behavior. The earliest CLTR works [78, 157] assume a basic model originally proposed by Craswell et al. [30]:

Assumption 3.3.1 (*The rank-based position bias model*)**.** The probability of a click on document d at position k is the product of the rank-based examination probability and document relevance:

$$P(C = 1 \mid d, k) = P(E = 1 \mid k)P(R = 1 \mid d) = \alpha_k P(R = 1 \mid d). \tag{3.3}$$

Later work has proposed more complex user models to build on [51]. Relevant to our work is the model proposed by Agarwal et al. [4], and its re-formulation by Vardasbi et al. [152]; it is a generalization of the above model to include a form of trust bias:

Assumption 3.3.2 (*The trust bias model*)**.** The probability of a click on document d at position k is an affine transformation of the relevance probability of d in the form:

$$P(C = 1 \mid d, k) = \alpha_k P(R = 1 \mid d) + \beta_k, \tag{3.4}$$

where $\forall k, \alpha_k \in [0, 1] \wedge \beta_k \in [0, 1] \wedge (\alpha_k + \beta_k) \in [0, 1]$.

Whilst it is named after trust bias, this model actually captures three forms of bias that were traditionally categorized separately: rank-based position bias, item-selection bias, and trust bias. Position bias was originally approached as the probability that a user would examine an item, which would decrease at lower positions in the ranking [30, 78, 157, 158]. In the trust bias model, this effect can be captured by decreasing $\alpha_k + \beta_k$ as k increases. Additionally, with $\forall k, \beta_k = 0$, the trust bias model is equivalent to the rank-based position bias model. Item-selection bias refers to users being unable to see documents outside a top-K, where they receive zero probability of being examined or interacted with [108]. This can be captured by the trust bias model by setting $\alpha_k + \beta_k = 0$ when $k > K$. Lastly, the key characteristic of trust bias is that users are more likely to click on non-relevant items when they are near the top of the ranking [4]. This can be captured by the model by making β_k larger as k decreases [152]. Thereby, the trust bias model is in fact a generalization of most of the user models assumed by earlier work [51, 110]. The following works all assume models that fit Assumption 3.3.2: [3, 4, 50, 107–110, 112, 152, 157, 158].

3.3.3 Counterfactual learning to rank

This section details the *policy-aware inverse propensity scoring* (IPS) estimator proposed by Oosterhuis and de Rijke [108] and the *doubly robust* (DR) estimator by Oosterhuis [107].

First, let $\mathcal{D}$ be a set of logged interaction data: $\mathcal{D} = \left\{q_i, y_i, c_i\right\}_{i=1}^{N}$, where each of the N interactions consists of a query q_i, a displayed ranking y_i, and click feedback $c_i(d) \in \{0, 1\}$ that indicates whether the user clicked on the document d or not. Both policies use propensities that are the expected α values for each document:

$$\rho_0(d \mid q_i, \pi_0) = \mathbb{E}_{y \sim \pi_0(q_i)}\left[\alpha_{k(d)}\right] = \rho_{i,0}(d). \tag{3.5}$$

Similarly, to keep our notation short, we also use $\omega(d \mid q_i, \pi) = \omega_i(d)$. Next, the policy-aware IPS estimator is defined as:

$$\hat{U}_{\text{IPS}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \frac{\omega_i(d)}{\rho_{i,0}(d)} c_i(d). \tag{3.6}$$

Oosterhuis and de Rijke [108] prove that under the rank-based position bias model (Assumption 3.3.1) and when $\forall(i, d), \rho_{i,0}(d) > 0$, this estimator is unbiased: $\mathbb{E}[\hat{U}_{\text{IPS}}(\pi)] = U(\pi)$.

The DR estimator improves over the policy-aware IPS estimator in terms of assuming the more general trust bias model (Assumption 3.3.2) and having lower variance. Oosterhuis [107] proposes the usage of the following ω values for the policy π:

$$\omega(d \mid q_i, \pi) = \mathbb{E}_{y \sim \pi(q_i)}\left[\alpha_{k(d)} + \beta_{k(d)}\right] = \omega_i(d), \tag{3.7}$$

since with these values U (Eq. 3.1) becomes the number of expected clicks on relevant items under the trust bias model; $U = (\alpha_k + \beta_k)P(R = 1 \mid d, q) = P(C = 1, R = 1 \mid k, d, q)$. We follow this approach and define the ω values for the logging policy π_0 as:

$$\omega_0(d \mid q_i, \pi_0) = \mathbb{E}_{y \sim \pi_0(q_i)}\left[\alpha_{k(d)} + \beta_{k(d)}\right] = \omega_{i,0}(d). \tag{3.8}$$

The DR estimator uses predicted relevances in its estimation, i.e., using predictions from a regression model. Let $\hat{R}_i(d) \approx P(R = 1 \mid d, q_i)$ indicate a predicted relevance; then the utility according to these predictions is:

$$\hat{U}_{\text{DM}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \omega_i(d) \hat{R}_i(d). \tag{3.9}$$

The DR estimator starts with this predicted utility and adds an IPS-based correction to remove its bias:

$$\hat{U}_{\text{DR}}(\pi) = \hat{U}_{\text{DM}}(\pi) + \frac{1}{N} \sum_{i=1}^{N} \sum_{d \in D} \frac{\omega_i(d)}{\rho_{i,0}(d)} \left(c_i(d) - \alpha_{k_i(d)} \hat{R}_i(d) - \beta_{k_i(d)}\right). \tag{3.10}$$

Thereby, the corrections of the IPS part of the DR estimator will be smaller if the predicted relevances are more accurate. Oosterhuis [107] proves that under the assumption of the trust bias model (Assumption 3.3.2), the DR estimator is unbiased

when $\forall(i,d), \rho_{i,0}(d) > 0 \vee \hat{R}_i(d) = P(R = 1 \mid d, q_i)$ and has less variance if $0 \leq \hat{R}_i(d) \leq 2P(R = 1 \mid d, q_i)$. The author also shows that the DR estimator needs less data to reach the same level of ranking performance as IPS, with especially large improvements when applied to top-K rankings [107].

3.3.4 Safety in counterfactual learning to rank

IPS-based CLTR methods, despite their unbiasedness and consistency, suffer from the problem of high-variance [51, 78, 107]. Specifically, if the logged click data is limited, training an IPS-based method can lead to an unreliable and unsafe ranking policy [50]. The problem of *safe* policy learning is well-studied in the bandit literature [63, 148, 149, 165]. Swaminathan and Joachims [148] proposed the first risk-aware off-policy learning method for bandits, with their risk term quantified as the variance of the IPS-estimator. Wu and Wang [165] proposed an alternative method for risk-aware off-policy learning, where the risk is quantified using a Renyi divergence between the action distribution of the new policy and the logging policy [123]. Thus, both consider it a risk for the new policy to be too dissimilar to the logging policy, which is presumed safe. Whilst effective at standard bandit problems, these risk-aware methods are not effective for ranking tasks due to their enormous combinatorial action spaces and correspondingly small propensities.

As a solution for CLTR, in Chapter 2 we introduced a risk-aware CLTR approach that uses divergence based on the exposure distributions of policies. They first introduce normalized propensities: $\rho'(d) = \rho/Z$, with a normalization factor Z based on K:

$$Z = \sum_{d \in D} \rho(d) = \sum_{d \in D} \mathbb{E}_{y \sim \pi}\left[\alpha_{k(d)}\right] = \mathbb{E}_{y \sim \pi}\left[\sum_{k=1}^{K} \alpha_{k(d)}\right] = \sum_{k=1}^{K} \alpha_k. \tag{3.11}$$

Since $\rho'(d) \in [0,1]$ and $\sum_d \rho'(d) = 1$, they can be treated as a probability distribution that indicates how exposure is spread over documents. In Chapter 2, we use Renyi divergence to quantify how dissimilar the new policy is from the logging policy:

$$d_2(\rho \,\|\, \rho_0) = \mathbb{E}_q\left[\sum_d \left(\frac{\rho'(d)}{\rho'_0(d)}\right)^2 \rho'_0(d)\right], \tag{3.12}$$

with the corresponding empirical estimate based on the log data ($\mathcal{D}$) defined as:

$$\hat{d}_2(\rho \,\|\, \rho_0) = \frac{1}{N} \sum_{i=1}^{N} \sum_d \left(\frac{\rho'_i(d)}{\rho'_{i,0}(d)}\right)^2 \rho'_{i,0}(d). \tag{3.13}$$

Based on this divergence term, they propose the following risk-aware CLTR objective, with parameter δ:

$$\max_{\pi} \hat{U}_{\text{IPS}}(\pi) - \sqrt{\frac{Z}{N}\left(\frac{1-\delta}{\delta}\right)\hat{d}_2(\rho \,\|\, \rho_0)}. \tag{3.14}$$

Thereby, the existing safe CLTR approach penalizes the optimization procedure from learning ranking behavior that is too dissimilar from the logging policy in terms of

the distribution of exposure. The weight of this penalty decreases as the number of datapoints N increases, thus it maintains the same point of convergence as standard IPS. Yet, initially when little data is available and the effect of variance is the greatest, it forces the learned policy to be very similar to the safe logging policy. Gupta et al. [50] prove that their objective bounds the real utility with a probability of $1 - \delta$:

$$P\left(U(\pi) \geq \hat{U}_{\text{IPS}}(\pi) - \sqrt{\frac{Z}{N}\left(\frac{1-\delta}{\delta}\right) d_2(\rho \,\|\, \rho_0)} - \sqrt{\frac{1}{N}\left(\frac{1-\delta}{\delta}\right)}\right) \geq 1 - \delta. \quad (3.15)$$

However, their proof of safety relies on the rank-based position bias model (Assumption 3.3.1) and their approach is limited to the basic IPS estimator for CLTR.

3.3.5 Proximal policy optimization

In the more general reinforcement learning (RL) field, *proximal policy optimization* (PPO) was introduced as a method to restrict a new policy π from deviating too much from a previously rolled-out policy π_0 [136, 137]. In contrast with the earlier discussed methods, PPO does not make use of a divergence term but uses a simple clipping operation in its optimization objective. Let s indicate a state, a an action and R a reward function, the PPO loss is:

$$U^{PPO}(s, a, \pi, \pi_0) = \mathbb{E}\left[\min\left(\frac{\pi(a \,|\, s)}{\pi_0(a \,|\, s)} R(a \,|\, s), g\big(\epsilon, R(a \,|\, s)\big)\right)\right], \quad (3.16)$$

where g creates a clipping threshold based on the sign of $R(a \,|\, s)$:

$$g(\epsilon, R(a \,|\, s)) = \begin{cases} (1+\epsilon)\, R(a \,|\, s) & \text{if } R(a \,|\, s) \geq 0, \\ (1-\epsilon)\, R(a \,|\, s) & \text{otherwise.} \end{cases} \quad (3.17)$$

The clipping operation removes incentives for the optimization to let π deviate too much from π_0, since there are no further increases in U^{PPO} when $\pi(a \,|\, s) > (1+\epsilon)\pi_0(a \,|\, s)$ or $\pi(a \,|\, s) < (1-\epsilon)\pi_0(a \,|\, s)$, depending on the sign of $R(a \,|\, s)$. Similar to the previously discussed general methods, PPO is not effective when directly applied to the CLTR setting due to the combinatorial action space and corresponding extremely small propensities (for most a and s: $\pi_0(a \,|\, s) \simeq 0$).

3.4 Extending Safety to Advanced CLTR

In this section, we introduce our first contribution of this chapter: our extension of the safe CLTR method to address trust bias and DR estimation.

3.4.1 Method: Safe doubly-robust CLTR

For the safe DR CLTR method, we extend the generalization bound from the existing IPS estimator and position bias [50, Eq. 26] to the DR estimator and trust bias.

Theorem 3.4.1. *Given the true utility $U(\pi)$ (Eq. 3.1) and its exposure-based DR estimate $\hat{U}_{DR}(\pi)$ (Eq. 3.10) of the ranking policy π with the logging policy π_0 and the metric weights ω and ω_0 (Eq. 3.7 and 3.8), assuming the trust bias click model (Assumption 3.3.2), the following generalization bound holds with probability $1-\delta$:* [1]

$$P\Bigg(U(\pi) \;\geq\; \hat{U}_{DR}(\pi) \;-\; \Big(1+\max_k \tfrac{\beta_k}{\alpha_k}\Big)\Big(\sqrt{\frac{2Z}{N}\Big(\frac{1-\delta}{\delta}\Big)\, d_2(\omega\|\omega_0)} + \sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}\Big)\Bigg) \;\geq\; 1-\delta. \tag{3.18}$$

Proof. For a proof, we refer to the appendix (Theorem 3.A.1). □

Given the novel generalization bound from Theorem 3.4.1, we define the safe DR CLTR objective as follows:

$$\max_{\pi} \hat{U}_{\text{DR}}(\pi) - \Big(1+\max_k \frac{\beta_k}{\alpha_k}\Big)\sqrt{\frac{2Z}{N}\Big(\frac{1-\delta}{\delta}\Big)\hat{d}_2(\omega \,\|\, \omega_0)}, \tag{3.19}$$

where $\hat{d}_2(\omega \,\|\, \omega_0)$ is defined analogously to Eq. 3.12. Note that we ignore the term $\sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}$, since it is constant with respect to the policy π, and therefore can be disregarded for optimization purposes. The objective optimizes the lower-bound on the true utility function, through a linear combination of the empirical DR estimator ($\hat{U}_{\text{DR}}(\pi)$) and the empirical risk regularization term ($\hat{d}_2(\omega \,\|\, \omega_0)$). In a setting where click data is limited, our safe DR objective will weight the risk regularization term higher, and as a result, the objective ensures that the new policy stays close to the safe logging policy. When a sufficiently high volume of click data is collected, and thus we have higher confidence in the DR estimate, the objective falls back to its DR objective counterpart.

For the choice of the ranking policy (π), we propose to optimize a stochastic ranking policy π with a gradient descent-based method. For the gradient calculation, we refer to previous work [50, 107, 169].

Conditions for safe DR CLTR. Finally, we note that besides the explicit assumption that user behavior follows the trust bias model (Assumption 3.3.2), there is also an important implicit assumption in this approach. Namely, the approach assumes that the bias parameters (i.e., α and β) are known, a common assumption in the CLTR literature [103, 107]. However, in practice, either of these assumptions could not hold, i.e., user behavior could not follow the trust bias model, or a model's bias parameters could be wrongly estimated. Additionally, in adversarial settings where clicks are intentionally misleading or incorrectly logged [20, 93, 118], the user behavior assumptions do not hold, and, the generalization bound of our DR CLTR is not guaranteed to hold. Thus, whilst it is an important advancement over the existing safe CLTR method [50], our approach is limited to only providing a *conditional* form of safety.

[1] The following proof differs slightly from the original proof published in [53], as we overlooked an additional constant term in the original paper.

3.5 Method: Proximal Ranking Policy Optimization (PRPO)

Inspired by the limitations of the method introduced in Section 3.4 and the PPO method from the RL field (Section 3.2), we propose the first *unconditionally* safe CLTR method: *proximal ranking policy optimization* (PRPO). Our novel PRPO method is designed for practical safety by making *no assumptions* about user behavior. Thereby, it provides the most robust safety guarantees for CLTR yet.

For safety, instead of relying on a high-confidence bound (e.g., Eq. 3.14 and 3.19), PRPO guarantees safety by removing the incentive for the new policy to rank documents too much higher than the safe logging policy. This is achieved by directly clipping the ratio of the metric weights for a given query q_i under the new policy $\omega_i(d)$, and the logging policy ($\omega_{i,0}(d)$), i.e., $\frac{\omega_i(d)}{\omega_{i,0}(d)}$ to be bounded in a fixed predefined range: $[\epsilon_-, \epsilon_+]$. As a result, the PRPO objective provides no incentive for the new policy to produce weights $\omega_i(d)$ outside of the range: $\epsilon_- \cdot \omega_{i,0}(d) \leq \omega_i(d) \leq \epsilon_+ \cdot \omega_{i,0}(d)$.

Before defining the PRPO objective, we first introduce a term $r(d|q)$ that represents an unbiased DR relevance estimate, weighted by ω_0, for a single document-query pair (cf. Eq. 3.10):

$$r(d|q) = \omega_0(d|q)\hat{R}(d|q) + \frac{\omega_0(d|q)}{\rho_0(d|q)} \sum_{i \in \mathcal{D}: q_i = q} \Big(c_i(d) - \alpha_{k_i(d)} \hat{R}(d|q) - \beta_{k_i(d)} \Big). \tag{3.20}$$

For the sake of brevity, we drop π and π_0 from the notation when their corresponding value is clear from the context. This enables us to reformulate the DR estimator around the ratios between the metric weights ω and ω_0 (cf. Eq. 3.10):

$$\hat{U}_{\text{DR}}(\pi) = \sum_{q, d \in \mathcal{D}} \frac{\omega(d \mid q)}{\omega_0(d \mid q)} r(d \mid q). \tag{3.21}$$

Before defining the proposed PRPO objective, we first define the following clipping function:

$$f(x, \epsilon_-, \epsilon_+, r) = \begin{cases} \min(x, \epsilon_+) \cdot r & r \geq 0, \\ \max(x, \epsilon_-) \cdot r & \text{otherwise.} \end{cases} \tag{3.22}$$

Given the reformulated DR estimator (Eq. 3.21), and the clipping function (Eq. 3.22), the PRPO objective can be defined as follows:

$$\hat{U}_{\text{PRPO}}(\pi) = \sum_{q, d \in \mathcal{D}} f\left(\frac{\omega(d \mid q)}{\omega_0(d \mid q)}, \epsilon_-, \epsilon_+, r(d \mid q) \right). \tag{3.23}$$

Figure 3.1 visualizes the effect the clipping of PRPO has on the optimization incentives. We see how the clipped and unclipped weight ratios progress as documents are placed on different ranks. The unclipped weights keep increasing as documents are moved to the top of the ranking, when $r > 1$, or to the bottom, when $r < 1$. Consequently, optimization with unclipped weight ratios aims to place these documents at the absolute top or bottom positions. Conversely, the clipped weights do not increase beyond their

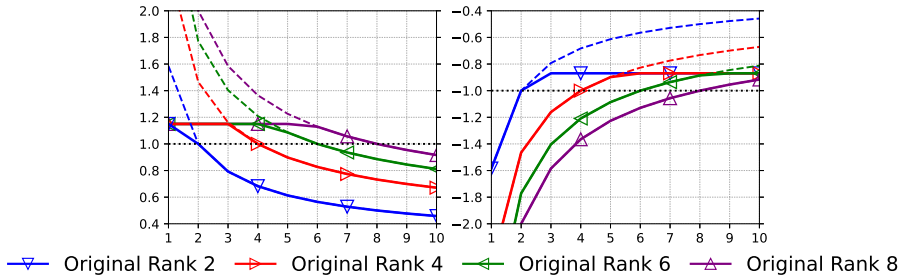

Figure 3.1: Weight ratios in the clipped PRPO objective (solid lines) and the unclipped counterparts (dashed lines), as documents are moved from four different original ranks. Left: positive relevance, $r = 1$; right: negative relevance, $r = -1$; x-axis: new rank for document; y-axis: unclipped weight ratios (dashed lines), $r \cdot \omega_i(d)/\omega_{i,0}(d)$; and clipped PRPO weight ratios (solid lines), $f\big(\omega_i(d)/\omega_{i,0}(d), \epsilon_- = 1.15^{-1}, \epsilon_+ = 1.15, r = \pm 1\big)$. DCG metric weights used: $\omega_i(d) = \log_2(\text{rank}(d \mid q_i, \pi) + 1)^{-1}$.

clipping threshold, which for most document is reached before being placed at the very top or bottom position. As a result, optimization with clipped weight ratios will not push these documents beyond these points in the ranking. For example, when $r > 0$, we see that there is no incentive to place a document at higher than rank 6, if it was placed at rank 8 by the logging policy. Similarly, placement higher than rank 4 leads to no gain if the original rank was 6, and higher than rank 3 leads to no improvement gain from an original rank of 4. Vice versa, when $r < 0$, each document has a rank, where placing it lower than that rank brings no increase in clipped weight ratio. Importantly, this behavior only depends on the metric and the logging policy; PRPO makes *no further assumptions*.

Whilst the clipping of PRPO is intuitive, we can prove that it provides the following formal form of unconditional safety:

Theorem 3.5.1. *Let q be a query, ω be metric weights, y_0 be a logging policy ranking, and $y^*(\epsilon_-, \epsilon_+)$ be the ranking that optimizes the PRPO objective in Eq. 3.23. Assume that $\forall d, \in \mathcal{D}, r(d \mid q) \neq 0$. Then, for any $\Delta \in \mathbb{R}_{\geq 0}$, there exist values for ϵ_- and ϵ_+ that guarantee that the difference between the utility of y_0 and $y^*(\epsilon_-, \epsilon_+)$ is bounded by Δ:*

$$\forall \Delta \in \mathbb{R}_{\geq 0}, \exists \epsilon_- \in \mathbb{R}_{\geq 0}, \epsilon_+ \in \mathbb{R}_{\geq 0} \;\; |U(y_0) - U(y^*(\epsilon_-, \epsilon_+))| \leq \Delta. \tag{3.24}$$

Proof. A proof is given in Appendix 3.A.2. □

Adaptive clipping. Theorem 3.5.1 describes a very robust sense of safety, as it shows PRPO can be used to prevent any given decrease in performance without assumptions. However, it also reveals that this safety comes at a cost; PRPO prevents both decreases and increases of performance. This is very common in safety approaches, as there is a generally a tradeoff between risks and rewards [50]. Existing safety methods, such as

the safe CLTR approach of Section 3.4, generally, loosen their safety measures as more data becomes available, and the risk is expected to have decreased [149].

We propose a similar strategy for PRPO through adaptive clipping, where the effect of clipping decreases as the number of datapoints N increases. Specifically, we suggest using a monotonically decreasing $\delta(N)$ function such that $\lim_{N\to\infty} \delta(N) = 0$. The ϵ parameters can then be obtained through the following transformation: $\epsilon_- = \delta(N)$ and $\epsilon_+ = \frac{1}{\delta(N)}$. This leads to a clipping range of $[\delta(N), \frac{1}{\delta(N)}]$, and in the limit: $\lim_{N\to\infty}$, it becomes: $[0, \infty]$. In other words, as more data is gathered, the effect of PRPO clipping eventually disappears, and the original objective is recovered. The exact choice of $\delta(N)$ determines how quickly this happens.

Gradient ascent with PRPO and possible extensions. Finally, we consider how the PRPO objective should be optimized. This turns out to be very straightforward when we look at its gradient. The clipping function f (Eq. 3.22) has a simpler gradient involving an indicator function on whether x is inside the bounded range:

$$\nabla_x f(x, \epsilon_-, \epsilon_+, r) = \mathbb{1}\big[(r > 0 \wedge x \leq \epsilon_+) \vee (r < 0 \wedge x \geq \epsilon_-)\big] r. \tag{3.25}$$

Applying the chain rule to the PRPO objective (Eq. 3.23) reveals:

$$\nabla_\pi \hat{U}_{\text{PRPO}}(\pi) = \sum_{q,d\in\mathcal{D}} \underbrace{\Big[\nabla_\pi \frac{\omega(d|q)}{\omega_0(d|q)}\Big]}_{\text{grad. for single doc.}} \underbrace{\nabla_\pi f\left(\frac{\omega(d|q)}{\omega_0(d|q)}, \epsilon_-, \epsilon_+, r(d|q)\right)}_{\text{indicator reward function}}. \tag{3.26}$$

Thus, the gradient of PRPO simply takes the importance weighted metric gradient per document, and multiplies it with the indicator function and reward. As a result, PRPO is simple to combine with existing LTR algorithms, especially ones that use policy-gradients [164], such as PL-Rank [104, 105] or StochasticRank [150]. For methods in the family of LambdaRank [18, 19, 159], it is a matter of replacing the $|\Delta DCG|$ term with an equivalent for the PRPO bounded metric.

Lastly, we note that whilst we introduced PRPO for DR estimation, it can be extended to virtually any relevance estimation by choosing a different r; e.g., one can easily adapt it for IPS [78, 110], or relevance estimates from a click model [28], etc. In this sense, we argue PRPO can be seen as a framework for robust safety in LTR.

3.6 Experimental Setup

For our experiments, we follow the semi-synthetic experimental setup that is prevalent in the CLTR literature [50, 107, 109, 152]. We make use of the three largest publicly available LTR datasets: Yahoo! Webscope [21], MSLR-WEB30k [115], and Istella [31]. The datasets consist of queries, a preselected list of documents per query, query-document feature vectors, and manually-graded relevance judgments for each query-document pair.

Following [50, 107, 152], we train a production ranker on a 3% fraction of the training queries and their corresponding relevance judgments. The goal is to simulate a real-world setting where a ranker trained on manual judgments is deployed in production and is used to collect click logs. The collected click logs can then be used for LTR. We

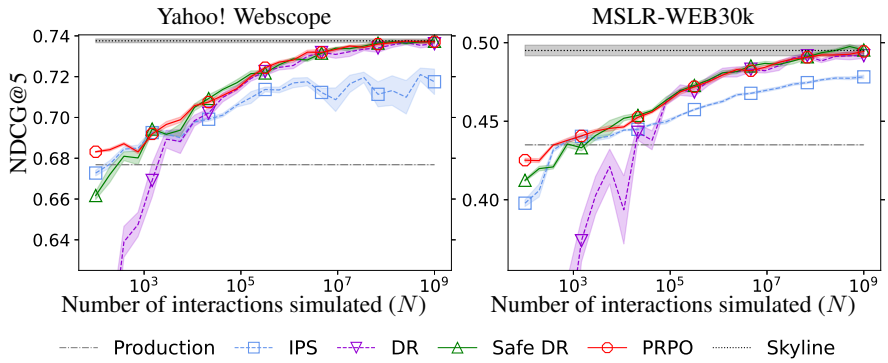

Figure 3.2: Performance in terms of NDCG@5 of the IPS, DR and proposed safe DR ($\delta = 0.95$) and PRPO ($\delta(N) = \frac{100}{N}$) methods for CLTR on Yahoo! Webscope and MSLR-WEB30k datasets. The results are presented varying size of training data (N), with number of simulated queries varying from 10^2 to 10^9. Results are averaged over 10 runs; the shaded areas indicate 80% prediction intervals.

assume the production ranker is safe, given that it would serve live traffic in a real-world setup.

We simulate a top-K ranking setup [108] where only $K = 5$ documents are displayed to the user for a given query, and any document beyond that gets zero exposure. To get the relevance probability, we apply the following transformation: $P(R = 1 \mid q, d) = 0.25 \cdot rel(q, d)$, where $rel(q, d) \in \{0, 1, 2, 3, 4\}$ is the relevance judgment for the given query-document pair. We generate clicks based on the trust bias click model (Assumption 3.3.2):

$$P(C = 1 \mid q, d, k) = \alpha_k P(R = 1 \mid q, d) + \beta_k. \tag{3.27}$$

The trust bias parameters are set based on the empirical observation in [4]: $\alpha = [0.35, 0.53, 0.55, 0.54, 0.52]$, and $\beta = [0.65, 0.26, 0.15, 0.11, 0.08]$. For CLTR training, we only use the training and validation clicks generated via the click simulation process (Eq. 3.27). To test the robustness of the safe CLTR methods in a setting where the click model assumptions do not hold, we simulate an *adversarial click model*, where the user clicks on the irrelevant document with a high probability and on a relevant document with a low click probability. We define the adversarial click model as:

$$P(C = 1 \mid q, d, k) = 1 - (\alpha_k P(R = 1 \mid q, d) + \beta_k). \tag{3.28}$$

Thereby, we simulate a maximally *adversarial* user who clicks on documents with a click probability that is inversely correlated with the assumed trust bias model (Assumption 3.3.2).

Further, we assume that the logging propensities have to be estimated. For the logging propensities ρ_0, and the logging metric weights (ω_0), we use a simple Monte Carlo estimate [50]:

$$\hat{\rho}_0(d) = \frac{1}{N} \sum_{i=1:y_i \sim \pi_0}^{N} \alpha_{k_i(d)}, \quad \hat{\omega}_0(d) = \frac{1}{N} \sum_{i=1:y_i \sim \pi_0}^{N} \big(\alpha_{k_i(d)} + \beta_{k_i(d)}\big). \tag{3.29}$$

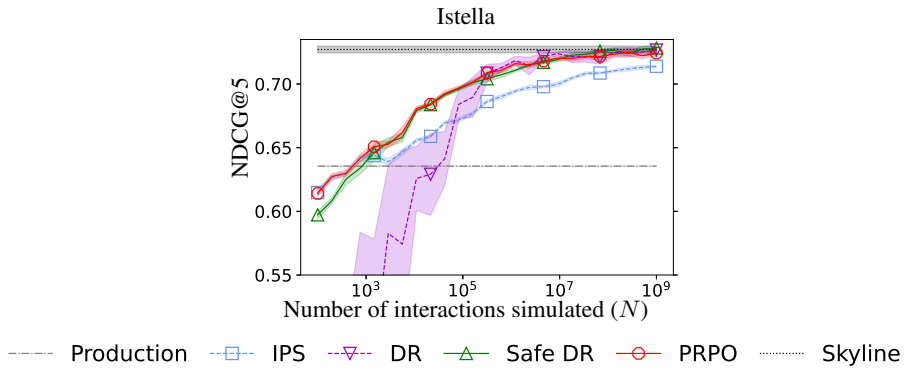

Figure 3.3: Performance in terms of NDCG@5 of the IPS, DR and proposed safe DR ($\delta = 0.95$) and PRPO ($\delta(N) = \frac{100}{N}$) methods for CLTR on the Istella dataset. The results are presented varying size of training data (N), with number of simulated queries varying from 10^2 to 10^9. Results are averaged over 10 runs; the shaded areas indicate 80% prediction intervals.

For the learned policies (π), we optimize PL ranking models [104] using the REINFORCE policy-gradient method [50, 169]. We perform clipping on the logging propensities (Eq. 3.5) only for the training clicks and not for the validation set. Following previous work, we set the clipping parameter to $10/\sqrt{N}$ [50, 110]. We do not apply the clipping operation for the logging metric weights (Eq. 3.8). To prevent overfitting, we apply early stopping based on the validation clicks. For variance reduction, we follow [50, 169] and use the average reward per query as a control-variate.

As our evaluation metric, we compute the NDCG@5 metric using the relevance judgments on the test split of each dataset [65]. Finally, the following methods are included in our comparisons:

(i) *IPS*. The IPS estimator with affine correction [110, 152] for CLTR with trust bias (Eq. 3.6).

(ii) *Doubly Robust*. The DR estimator for CLTR with trust bias (Eq. 3.10). This is the most important baseline for this chapter, given that the DR estimator is the state-of-the-art CLTR method [107].

(iii) *Safe DR*. Our proposed safe DR CLTR method (Eq. 3.19), which relies on the trust bias assumption (Assumption 3.3.2).

(iv) *PRPO*. Our proposed *proximal ranking policy optimization* (PRPO) method for safe DR CLTR (Eq. 3.23).

(v) *Skyline*. LTR method trained on the true relevance labels. Given that it is trained on the real relevance signal, the skyline performance is the upper bound on any CLTR methods performance.

3.7 Results and Discussion

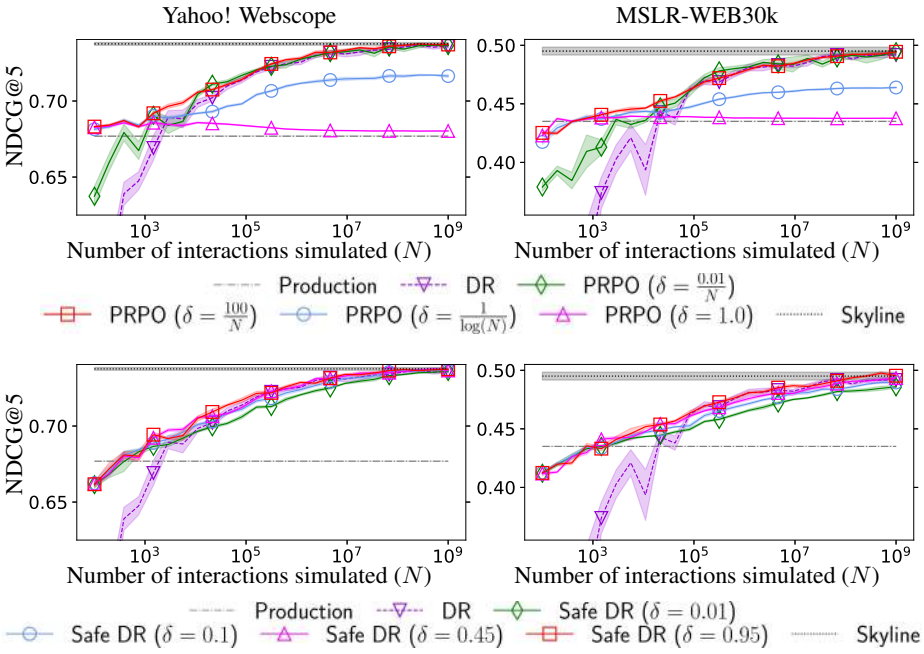

Figure 3.4: Performance of the safe DR and PRPO with varying safety parameter (δ) on Yahoo! Webscope and MSLR-WEB30k datasets. Top row: sensitivity analysis of PRPO with varying clipping parameter (δ) over varying dataset sizes N. Bottom row: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 runs; shaded areas indicate 80% prediction intervals.

Comparision with baseline methods. Figure 3.2 and 3.3 present the main results with different CLTR estimators with varying amounts of simulated click data. Amongst the baselines, we see that the DR estimator converges to the skyline much faster than the IPS estimator. The IPS estimator fails to reach the optimal performance even after training on 10^9 clicks, suggesting that it suffers from a high-variance problem. This aligns with the findings in [107]. As to safety, when the click data is limited ($N < 10^5$), the DR estimator performs much worse than the logging policy, i.e., it exhibits unsafe behavior, which can lead to a negative user experience if deployed online. A likely explanation is that when click data is limited, the regression estimates ($\hat{R}(d)$, Eq. 3.10) have high errors, resulting in a large performance degradation, compared to IPS.

Our proposed safety methods, safe DR and PRPO, reach the performance of the logging policy within ~500 queries on all datasets. For the safe DR method, we set the confidence parameter $\delta = 0.95$. For the PRPO method, we set $\delta(N) = \frac{100}{N}$. On the MSLR and the ISTELLA dataset, we see that PRPO reaches logging policy performance with almost 10^3 fewer queries than the DR method. Thus, our proposed methods, safe DR and PRPO, can be safely deployed, and avoid the initial period of bad performance

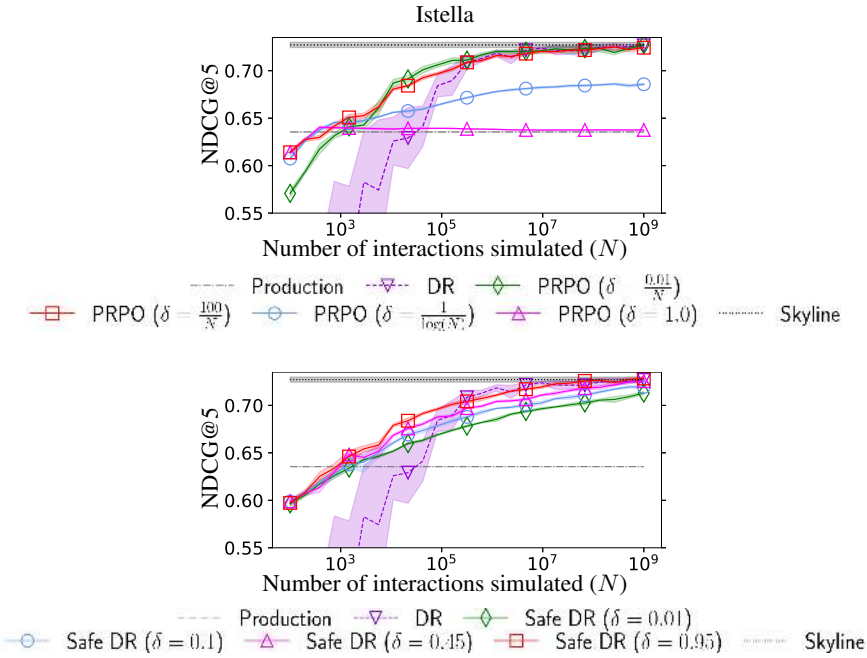

Figure 3.5: Performance of the safe DR and PRPO with varying safety parameter (δ) on ISTELLA dataset. Top row: sensitivity analysis of PRPO with varying clipping parameter (δ) over varying dataset sizes N. Bottom row: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 runs; shaded areas indicate 80% prediction intervals.

of DR, whilst providing the same state-of-the-art performance at convergence.

Sensitivity analysis of the safety parameter. To understand the tradeoff between safety and utility, we performed a sensitivity analysis by varying the safety parameter (δ) for the safe DR method and PRPO. The top rows of Figure 3.4 and 3.5 show us the performance of the PRPO method with different choices of the clipping parameter δ as a function of dataset size (N). We report results with the setting of the δ parameter, which results in different clipping widths. For the setting $\delta = \frac{0.01}{N}$ and $\delta = \frac{100}{N}$, the clipping range width grows linearly with the dataset size N. Hence, the resulting policy is safer at the start but converges to the DR estimator when N increases. With $\delta = \frac{0.01}{N}$, the clipping range is wider at the start. As a result, it is more unsafe than when $\delta = \frac{100}{N}$, which is the safest amongst all. For the case where the range grows logarithmically ($\delta = \frac{1}{\log(N)}$), the method is more conservative throughout, i.e., it is closer to the logging policy since the clipping window grows only logarithmically with N. For the extreme case where the clipping range is a constant ($\delta = 1$), PRPO avoids any change w.r.t. the logging policy, and as a result, it sticks closely to the logging policy.

The bottom rows of Figure 3.4 and 3.5 show the performance of the safe DR method with varying confidence parameter values (δ). Due to the nature of the generalization bound (Eq. 3.19), the confidence parameter is restricted to: $0 \leq \delta \leq 1$. We vary the

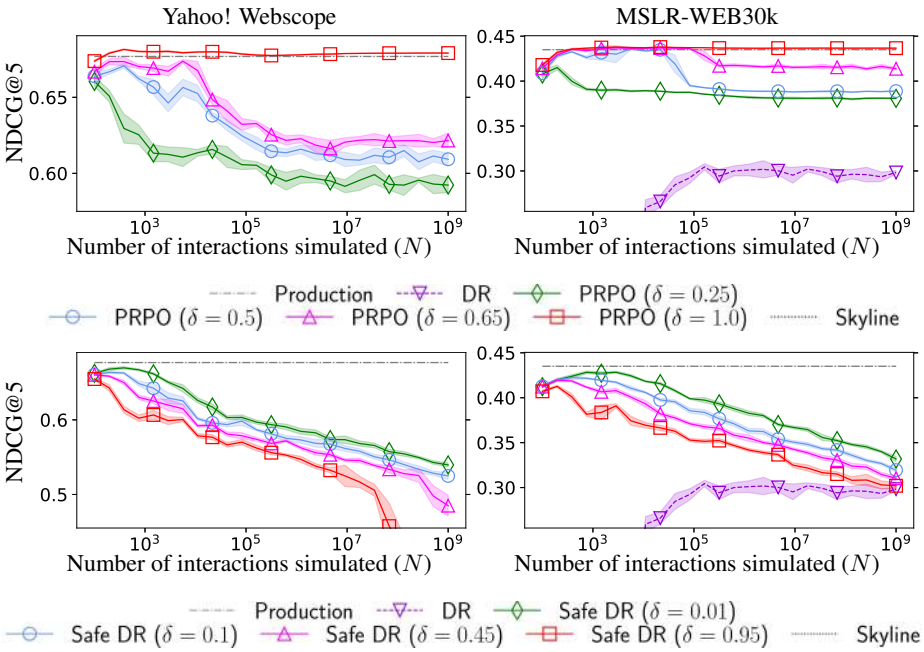

Figure 3.6: Performance of the proposed safe DR and PRPO with the adversarial click model on the Yahoo! and MSLR datasets. Top: sensitivity analysis results for the PRPO method with varying clipping parameter (δ). Bottom: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 independent runs; the shaded areas indicate 80% prediction intervals.

confidence parameters in the range $\delta \in \{0.01, 0.1, 0.45, 0.95\}$. We note that a lower δ value results in higher safety, and vice-versa. Until $N < 10^5$, there is no noticeable difference in performance. For the Yahoo! Webscope dataset, almost all settings result in a similar performance. For the MSLR and ISTELLA datasets, when $N < 10^5$, a lower δ value results in a more conservative policy, i.e., a policy closer to the logging policy. However, the performance difference with different setups is less drastic than with the PRPO method. Thus, we note that the safe DR method is *less flexible* in comparison to PRPO.

Therefore, compared to our safe DR method, we conclude that our PRPO method provides practitioners with greater flexibility and control when deciding between safety and utility.

Robustness analysis using an adversarial click model. To verify our initial claim that our proposed PRPO method provides safety guarantees *unconditionally*, we report results with clicks simulated via the adversarial click model (Eq. 3.28). With the adversarial click setup, the initial user behavior assumptions (Assumption 3.3.2) *do not hold*. The top rows of Figure 3.6 and 3.7 show the performance of the PRPO method with different safety parameters when applied to the data collected via the adversarial click model. We vary the δ parameter for PRPO in the range $\{0.25, 0.5, 0.65, 1.0\}$, e.g.,

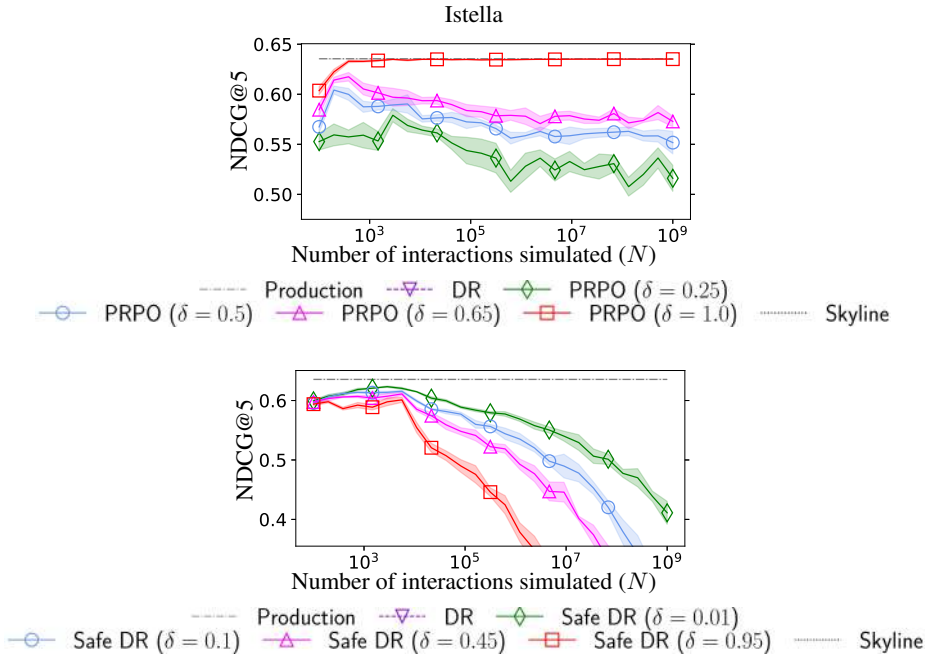

Figure 3.7: Performance of the proposed safe DR and PRPO with the adversarial click model on the ISTELLA dataset. Top: sensitivity analysis results for the PRPO method with varying clipping parameter (δ). Bottom: sensitivity analysis for the safe DR method with varying safety confidence parameter (δ). Results are averaged over 10 independent runs; the shaded areas indicate 80% prediction intervals.

$\delta = 0.5$ results in $\epsilon_- = 0.5$ and $\epsilon_+ = 2$. With the constant clipping range ($\delta = 1$), we notice that after $\sim$400 queries, the PRPO methods performance never drops below the safe logging policy performance. For greater values of δ, there are drops in performance but they are all bounded. For the Yahoo! Webscope dataset, the maximum drop in the performance is $\sim 12\%$; for the MSLR30K dataset, the maximum performance drop is $\sim 10\%$; and finally, for the Istella dataset, the maximum drop is $\sim 20\%$. Clearly, these observations show that PRPO provides robust safety guarantees, that are reliable even when user behavior assumptions are wrong.

In contrast, the generalization bound of our safe DR method (Theorem 3.4.1) holds only when the user behavior assumptions are true. This is not the case in the bottom rows of Figure 3.6 and 3.7, which show the performance of the safe DR method under the adversarial click model. Even with the setting where the safety parameters have a high weight ($\delta = 0.01$), as the click data size increases, the performance drops drastically. Regardless of the exact choice of δ, the effect of the regularization of safe DR disappears as N grows, thus in this adversarial setting, it is only a matter of time before the performance of safe DR degrades dramatically.

3.8 Conclusion

In this chapter, we have introduced the first safe CLTR method that uses state-of-the-art DR estimation and corrects trust bias. This is a significant extension of the existing safety method for CLTR that was restricted to position bias and IPS estimation. However, in spite of the importance of this extended safe CLTR approach, it heavily relies on user behavior assumptions. We argue that this means it only provides a *conditional* concept of safety, that may not apply to real-world settings. To address this limitation, we have made a second contribution: the *proximal ranking policy optimization* (PRPO) method. PRPO is the first LTR method that provides *unconditional* safety, that is applicable regardless of user behavior. It does so by removing incentives to stray too far away from a safe ranking policy. Our experimental results show that even in the extreme case of adversarial user behavior PRPO results in safe ranking behavior, unlike existing safe CLTR approaches.

PRPO easily works with existing LTR algorithms and relevance estimation techniques. We believe it provides a flexible and generic framework that enables practitioners to apply the state-of-the-art CLTR method with strong and robust safety guarantees. Future work may apply the proposed safety methods to exposure-based ranking fairness [104, 169] and to safe online LTR [110].

In this chapter, we answer the broad research question (RQ2) in affirmative. We introduce PRPO, a novel safe counterfactual LTR method that does not rely on any user behavior assumptions with robust safety guarantees, even under adversarial conditions.

So far, in the first part of the thesis, we have discussed safe deployment strategies for contextual bandits with combinatorial action space – for example, ranking for web search. In the second part of the thesis, we will switch the discussion to contextual bandits for traditional single-action recommender systems.

3.A Appendix: Extended Safety Proof

Lemma 3.A.1. *Under the trust bias click model (Assumption 3.3.2), and given the trust bias parameter α_k, β_k, the regression model estimates $\hat{R}_d$ and click indicator $c(d)$, the following holds:*

$$\mathrm{Cov}_{y,c}\Big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\Big] \geq 0. \tag{3.30}$$

Proof. The covariance term can be rewritten as:

$$\begin{aligned}
&\mathrm{Cov}_{y,c}\Big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\Big] \\
&= \mathbb{E}_{y,c}\Big[(c(d) - \beta_{k(d)})\alpha_{k(d)}\hat{R}_d\Big] - \mathbb{E}_{y,c}\big[c(d) - \beta_{k(d)}\big]\mathbb{E}_y\Big[\alpha_{k(d)}\hat{R}_d\Big] \\
&= \hat{R}_d\Big(\mathbb{E}_{y,c}\big[c(d)\alpha_{k(d)}\big] - \mathbb{E}_y\big[\beta_{k(d)}\alpha_{k(d)}\big] - R_d\ \rho_0(d)^2\Big),
\end{aligned} \tag{3.31}$$

where use $\rho_0(d) = \mathbb{E}_{y,c}\big[\alpha_{k(d)}\big]$ and $\mathbb{E}_{y,c}\big[(c_i(d) - \beta_{k_i(d)})/\rho_0(d)\big] = R_d$ [107]. Expanding the first expectation term in the expression:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y,c}\big[c(d)\alpha_{k(d)}\big] = \sum_{y\in\pi_0} \pi_0(y)\alpha_{k(d)}P(C=1 \mid d, y) \\
&= \sum_{y\in\pi_0} \pi_0(y)\alpha_{k(d)} \cdot \big(\alpha_{k(d)}R_d + \beta_{k(d)}\big) = R_d\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] + \mathbb{E}_y\big[\alpha_{k(d)}\beta_{k(d)}\big] \\
&= \sum_{y\in\pi_0} \pi_0(y)\Big[\alpha^2_{k(d)}R_d + \alpha_{k(d)}\beta_{k(d)}\Big],
\end{aligned} \tag{3.32}$$

where we substitute click model equation $P(C=1 \mid d, y)$ (Eq. 3.10). Substituting it back in Eq. 3.31, we get:

$$\begin{aligned}
\mathrm{Cov}_{y,c}\Big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}_d\Big] &= R_d\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] - R_d\ \mathbb{E}_y\big[\alpha_{k(d)}\big]^2 \\
R_d\Big(\mathbb{E}_y\Big[\alpha^2_{k(d)}\Big] - \mathbb{E}_y\big[\alpha_{k(d)}\big]^2\Big) &= R_d\mathrm{Var}_y[\alpha_k(d)] \geq 0.
\end{aligned}$$

□

3.A.1 Proof of Theorem 3.4.1

Theorem 3.4.1. *Given the true utility $U(\pi)$ (Eq. 3.1) and its exposure-based DR estimate $\hat{U}_{DR}(\pi)$ (Eq. 3.10) of the ranking policy π with the logging policy π_0 and the metric weights ω and ω_0 (Eq. 3.7 and 3.8), assuming the trust bias click model (Assumption 3.3.2), the following generalization bound holds with probability $1-\delta$:*[2]

$$\begin{aligned}
P\Bigg(U(\pi)\ \geq\ \hat{U}_{DR}(\pi)\ -\ \Big(1 + \max_k \tfrac{\beta_k}{\alpha_k}\Big)\Big(\sqrt{\frac{2Z}{N}\Big(\frac{1-\delta}{\delta}\Big)\,d_2(\omega\|\omega_0)} \\
+ \sqrt{\frac{1}{N}\Big(\frac{1-\delta}{\delta}\Big)}\Big)\Bigg)\ \geq\ 1-\delta.
\end{aligned} \tag{3.33}$$

[2]The following proof differs slightly from the original proof published in [53], as we overlooked an additional constant term in the original paper.

Proof. As per Cantelli's inequality [42], the following inequality must hold with probability $1-\delta$:

$$U(\pi) \geq \hat{U}_{\text{DR}}(\pi) - \sqrt{\frac{1-\delta}{\delta}\text{Var}_{q,y,c}\Big[\hat{U}_{\text{DR}}(\pi)\Big]}. \tag{3.34}$$

Following a similar approach as previous works [50, 165], we look for an upper-bound on the variance of the DR estimator. From the definition of $\hat{U}_{\text{DR}}(\pi)$ (Eq. 2.6) and the assumption that queries q are i.i.d, the variance of the counterfactual estimator can be expanded by applying the law of total variance as follows:

$$\text{Var}_{q,y,c}\Big[\hat{U}_{\text{DR}}(\pi)\Big] = \frac{1}{N}\Big(\mathbb{E}_q\Big[\text{Var}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]\Big] + \text{Var}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]\Big]\Big). \tag{3.35}$$

The second term (variance over queries) can be expanded as follows:

$$\begin{aligned}
\text{Var}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]\Big] &= \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]^2\Big] - \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]\Big]^2 \\
&\leq \mathbb{E}_q\Big[\mathbb{E}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]^2\Big] &(3.36)\\
&= \mathbb{E}_q\Big[[U(\pi) \mid q]^2\Big] &(3.37)\\
&\leq 1,
\end{aligned}$$

where in the second step, we use the unbiasedness property of the doubly robust estimator [107, Eq. 37], and used the fact that the true utility is non-zero, i.e. $U(\pi) \geq 0$. In the last step, we made use of the fact that the true utility is bounded, and is upper bounded by 1 (safe to assume if the utility is normalized, for ex: normalized discounted cumulative gain, or click-through rate.), which results in the following bound for the doubly-robust variance:

$$\text{Var}_{q,y,c}\Big[\hat{U}_{\text{DR}}(\pi)\Big] \leq \frac{1}{N}\Big(\mathbb{E}_q\Big[\text{Var}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi) \mid q\Big]\Big] + 1\Big). \tag{3.38}$$

Now, focusing on the first part of the doubly-robust variance, from the definition of $\hat{U}_{\text{DR}}(\pi)$ (Eq. 3.10), the variance of the DR estimator (for a single query) can be expressed as the variance of the second term (Eq. 3.10):

$$\text{Var}_{y,c}\Big[\hat{U}_{\text{DR}}(\pi)\Big] = \frac{1}{N}\underset{y,c}{\text{Var}}\left[\sum_{d\in D}\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right]. \tag{3.39}$$

Using Assumption 3.3.2 and assuming that document examinations are independent from each other [50], we rewrite further:

$$\begin{aligned}
N \cdot \underset{y,c}{\text{Var}}\Big[\hat{U}_{\text{DR}}(\pi)\Big] &= \sum_{d\in D_q}\underset{y,c}{\text{Var}}\left[\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right] \\
&= \sum_{d\in D_q}\left(\frac{\omega(d)}{\rho_0(d)}\right)^2\underset{y,c}{\text{Var}}\Big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\Big].
\end{aligned} \tag{3.40}$$

The total variance can be split into the following:

$$\begin{aligned}\mathrm{Var}_{y,c}\Big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}_i(d)\Big] &= \mathrm{Var}_{y,c}\Big[\alpha_{k(d)}\hat{R}(d)\Big] \\ &+ \mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big] - 2\mathrm{Cov}_{y,c}\Big[c(d) - \beta_{k(d)}, \alpha_{k(d)}\hat{R}(d)\Big].\end{aligned} \tag{3.41}$$

Using Lemma 3.A.1, we upper-bound the total variance term to:

$$\begin{aligned}&\mathrm{Var}_{y,c}\Big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\Big] \\ &\leq \mathrm{Var}_{y,c}\Big[\alpha_{k(d)}\hat{R}(d)\Big] + \mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big].\end{aligned} \tag{3.42}$$

Next, we consider the two variance terms separately; with the variance of the first term following:

$$\underset{y,c}{\mathrm{Var}}\Big[\alpha_{k(d)}\hat{R}(d)\Big] = \underset{y,c}{\mathrm{Var}}\big[\alpha_{k(d)}\big]\hat{R}(d)^2 \leq \mathbb{E}_{y,c}\Big[\alpha_{k(d)}^2\Big] \leq \mathbb{E}_y\big[\alpha_{k(d)}\big],$$

where we make use of the fact that $\hat{R}_d^2 \leq 1$, and $\alpha \in [0,1] \rightarrow \alpha_k^2 \leq \alpha_k$. Next, we consider the second term:

$$\begin{aligned}\mathrm{Var}_{y,c}\big[c(d) - \beta_{k(d)}\big] &\leq \mathbb{E}_{y,c}\Big[\big(c(d) - \beta_{k(d)}\big)^2\Big] \\ &= \mathbb{E}_{y,c}\Big[c(d)^2 + \beta_{k(d)}^2 - 2c(d)\beta_{k(d)}\Big] \leq \mathbb{E}_{y,c}[c(d)] + \mathbb{E}_y\big[\beta_{k(d)}\big],\end{aligned} \tag{3.43}$$

since $c(d)^2 = c(d)$, $\beta_k^2 \leq \beta_k$, and $\mathbb{E}_{y,c}\big[c(d)\beta_{k(d)}\big] \geq 0$. Substituting the click probabilities with Eq. 3.4, we get:

$$\begin{aligned}&\mathbb{E}_{y,c}[c(d)] + \mathbb{E}_{y,c}[\beta_{k(d)}] = \mathbb{E}_{y,c}[\alpha_{k(d)}]P(R=1|\,d) + 2\,\mathbb{E}_{y,c}[\beta_{k(d)}] \\ &\leq \mathbb{E}_y\big[\alpha_{k(d)}\big] + 2\,\mathbb{E}_y[\beta_{k(d)}],\end{aligned} \tag{3.44}$$

where we use the fact that $P(R = 1 \mid d) \leq 1$. Putting together the bounds on both parts of Eq. 3.42, we have:

$$\mathrm{Var}_{y,c}\Big[c(d) - \beta_{k(d)} - \alpha_{k(d)}\hat{R}(d)\Big] \leq 2\omega_0(d), \tag{3.45}$$

where $\omega_0(d) = \mathbb{E}_y\big[\alpha_{k(d)}\big] + \mathbb{E}_y\big[\beta_{k(d)}\big]$. Substituting the final variance upper bound in Eq. 3.40, we get:

$$\begin{aligned}&\underset{y,c}{\mathrm{Var}}\left[\sum_{d\in D}\frac{\omega(d)}{\rho_0(d)}\big(c(d) - \alpha_{k(d)}\hat{R}(d) - \beta_{k(d)}\big)\right] \\ &\leq 2\sum_{d\in D_q}\left(\frac{\omega(d)}{\rho_0(d)}\right)^2\omega_0(d) \\ &= 2\sum_{d\in D_q}\left(\frac{\omega(d)}{\rho_0(d)}\right)^2\omega_0(d)\left(\frac{\omega_0(d)}{\omega_0(d)}\right)^2\end{aligned}$$

$$= 2 \sum_{d \in D_q} \left(\frac{\omega(d)}{\omega_0(d)} \right)^2 \omega_0(d) \left(\frac{\omega_0(d)}{\rho_0(d)} \right)^2, \tag{3.46}$$

where we multiply and divide by $\omega_0(d)^2$ in the third step. Finally, we make use of the fact: $\frac{\omega_0(d)}{\rho_0(d)} \leq \max_{\pi_0} \frac{\omega_0(d)}{\rho_0(d)} \leq 1 + \max_k \frac{\beta_k}{\alpha_k}$, and put everything back together:

$$\begin{aligned} N \cdot \mathrm{Var}_{y,c}\Big[\hat{U}_{\mathrm{DR}}(\pi)\Big] &\leq 2Z \left(1 + \max_k \frac{\beta_k}{\alpha_k}\right)^2 \sum_{d \in D_q} \left(\frac{\omega'(d)}{\omega'_0(d)} \right)^2 \omega'_0(d) \\ &= 2Z \left(1 + \max_k \frac{\beta_k}{\alpha_k}\right)^2 d_2(\omega \,\|\, \omega_0). \end{aligned} \tag{3.47}$$

where $d_2(\omega \,\|\, \omega_0)$ is the Renyi divergence between the normalized expected exposure $\omega'(d)$ and $\omega'_0(d)$ (cf. Eq. 3.12). Next, we replace the variance with the Renyi divergence-based term, and substituting back into the upper-bound on variance in Eq. 3.34 results in the following:

$$U(\pi) \geq \hat{U}_{\mathrm{DR}}(\pi) - (1 + \max_k \tfrac{\beta_k}{\alpha_k}) \Big(\sqrt{\tfrac{2Z}{N} \left(\tfrac{1-\delta}{\delta}\right) d_2(\omega \| \omega_0) + \tfrac{1}{N} \left(\tfrac{1-\delta}{\delta}\right)} \Big) \geq 1 - \delta. \tag{3.48}$$

By applying the Cauchy–Schwarz inequality, we get:

$$U(\pi) \geq \hat{U}_{\mathrm{DR}}(\pi) - (1 + \max_k \tfrac{\beta_k}{\alpha_k}) \Big(\sqrt{\tfrac{2Z}{N} \left(\tfrac{1-\delta}{\delta}\right) d_2(\omega \| \omega_0)} + \sqrt{\tfrac{1}{N} \left(\tfrac{1-\delta}{\delta}\right)} \Big) \geq 1 - \delta. \tag{3.49}$$

This completes the proof. □

3.A.2 Proof of Theorem 3.5.1

Proof. Given a logging policy ranking y_0, a user defined metric weight ω, and non-zero $r(d \mid q)$, for the choice of the clipping parameters $\epsilon_- = \epsilon_+ = 1$, the ranking $y^*(\epsilon_-, \epsilon_+)$ that maximizes the PRPO objective (Eq. 3.23) will be the same as the logging ranking y_0, i.e. $y^*(\epsilon_-, \epsilon_+) = y_0$. This is trivial to prove since any change in ranking can only lead in a decrease in the clipped ratio weights, and thus, a decrease in the PRPO objective. Therefore, $y^*(\epsilon_- = 1, \epsilon_+ = 1) = y_0$ when $\epsilon_- = \epsilon_+ = 1$. Accordingly: $|U(y_0) - U(y^*(\epsilon_- = 1, \epsilon_+ = 1))| = 0$ directly implies Eq. 3.24. This completes our proof. □

Whilst the above proof is performed in the extreme case where $\epsilon_- = \epsilon_+ = 1$ and the optimal ranking has the same utility as the logging policy ranking, other choices of ϵ_- and ϵ_+ bound the difference in utility to a lesser degree and allow for more deviation. As our experimental results show, the power of PRPO is that it gives practitioners direct control over this maximum deviation.

Part II

Robust and Efficient Reinforcement Learning for Recommendation and Diffusion Models

4

Optimal Baseline Corrections for Off-policy Contextual Bandits

So far, the first part of this thesis has examined contextual bandits with *combinatorial* action spaces – for example, web-search ranking and slate recommendation. The second part shifts attention to contextual bandits that select *a single action*, such as top-1 recommendations or reinforcement-learning-based fine-tuning of foundation models. This chapter zooms in on the top-1 recommendation case.

Our aim is to increase sample efficiency in off-policy evaluation and learning from logged user interactions. Although inverse propensity scoring (IPS) is unbiased in expectation, it suffers from high variance [50, 127]. Variance-reduction techniques – most notably the doubly robust (DR) estimator [107] and self-normalized IPS (SNIPS) [147] – mitigate this problem through additive and multiplicative baseline corrections, respectively [79, 147]. Yet, the literature still lacks a unifying lens on these approaches, which motivates the following research question:

RQ3 Can we unify variance reduction techniques using baseline corrections and a doubly robust estimator under a common framework?

To address **RQ3**, we introduce the β-IPS estimator, which places IPS, DR, and SNIPS inside a single baseline-correction framework. This, in turn, raises a second question:

RQ4 Given a unified framework for variance reduction techniques under baseline corrections, can we derive a variance-optimal unbiased estimator?

Within the unified β-IPS framework, we ask whether a variance-optimal baseline β^* can be derived analytically. In the latter part of this chapter we show that it can, presenting a closed-form expression for β^* that minimizes variance for both off-policy evaluation and learning.

4.1 Introduction & Motivation

Recommender systems have undergone a paradigm shift in the last few decades, moving their focus from *rating* prediction in the days of the Netflix Prize [12], to *item* prediction

This chapter was published as [52].

from implicit feedback [122] and *ranking* applications gaining practical importance [74, 142]. Recently, work that applies ideas from the algorithmic *decision-making* literature to recommendation problems has become more prominent [51, 72, 126, 153]. While this line of research is not inherently new [87, 139], methods based on contextual bandits (or reinforcement learning by extension) have now become widespread in the recommendation field [11, 16, 69, 99, 100, 145, 171]. The *off-policy* setting is particularly attractive for practitioners [151], as it allows models to be trained and evaluated in an offline manner [25–27, 34, 48–50, 53, 68, 70, 71, 92, 97]. Indeed, methods exist to obtain unbiased *offline* estimators of *online* reward metrics, which can then be optimized directly [66].

Research at the forefront of this area typically aims to find Pareto-optimal solutions to the bias-variance trade-off that arises when choosing an estimator: reducing variance by accepting a small bias [62, 144], by introducing control variates [35, 147], or both [143]. Control variates are especially attractive as they (asymptotically) preserve the unbiasedness of the widespread inverse propensity scoring (IPS) estimator. Additive control variates give rise to baseline corrections [43], regression adjustments [40], and doubly robust estimators [35]. Multiplicative control variates lead to self-normalised estimators [81, 147]. Previous work has proven that for off-policy *learning* tasks, the multiplicative control variates can be re-framed using an equivalent additive variate [17, 79], enabling mini-batch optimization methods to be used. We note that the self-normalised estimator is only *asymptotically* unbiased: a clear disadvantage for evaluation with finite samples. The common problem which most existing methods tackle is that of *variance reduction* in offline value estimation, either for learning or for evaluation. The common solution is the application of a control variate, either multiplicative or additive [113]. However, to the best of our knowledge, there is no work that attempts to unify these methods. Our work in this chapter addresses this gap by presenting these methods in a unifying framework of baseline corrections which, in turn, allows us to find the optimal baseline correction for variance reduction.

In the context of off-policy learning, adding to the well-known equivalence between reward-translation and self-normalisation described by Joachims et al. [79], we demonstrate that the equivalence extends to baseline corrections, regression adjustments, and doubly robust estimators with a constant reward model. Further, we derive a novel baseline correction method for off-policy learning that minimizes the variance of the gradient of the (unbiased) estimator. We further show that the baseline correction can be estimated in a closed-form fashion, allowing for easy practical implementation.

In line with recent work on off-policy evaluation/learning for recommendation [68, 69, 73, 124, 130], we adopt an off-policy simulation environment to emulate real-world recommendation scenarios, such as stochastic rewards, large action spaces, and controlled randomisation. This choice also encourages future reproducibility [129]. Our experimental results indicate that our proposed baseline correction for gradient variance reduction enables substantially faster convergence and lower gradient variance during learning.

In addition, we derive a closed-form solution to the optimal baseline correction for off-policy evaluation, i.e., the one that minimizes the variance of the estimator itself. Importantly, since our framework only considers unbiased estimators, the variance-optimality implies overall optimality. Our experimental results show that this leads

to lower errors in policy value estimation than widely used doubly-robust and SNIPS estimators [35, 147].

All source code to reproduce our experimental results is available at: `https://github.com/shashankg7/recsys2024_optimal_baseline`.

4.2 Background and Related Work

The goal of this section is to introduce common contextual bandit setups for recommendation, both on-policy and off-policy.

4.2.1 On-policy contextual bandits

We address a general contextual bandit setup [77, 128] with contexts X, actions A, and rewards R. The context typically describes *user* features, actions are the *items* to recommend, and rewards can be any type of *interaction* logged by the platform. A policy π defines a conditional probability distribution over actions x: $\mathsf{P}(A = a \mid X = x, \Pi = \pi) \equiv \pi(a \mid x)$. Its *value* is the expected reward it yields:

$$V(\pi) = \mathop{\mathbb{E}}_{x \sim \mathsf{P}(X)} \Big[\mathop{\mathbb{E}}_{a \sim \pi(\cdot|x)} [R] \Big]. \tag{4.1}$$

When the policy π is deployed, we can estimate this quantity by averaging the rewards we observe. We denote the expected reward for action a and context r as $r(a, x) \coloneqq \mathbb{E}[R \mid X = x; A = a]$.

In the field of contextual bandits (and reinforcement learning (RL) by extension), one often wants to learn π to maximise $V(\pi)$ [84, 146]. This is typically achieved through gradient ascent. Assuming π_θ is parameterised by θ, we iteratively update with learning rate η:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta (V(\pi_\theta)). \tag{4.2}$$

Using the well-known REINFORCE "log-trick" [163], the above gradient can be formulated as an expectation over sampled actions, whereby tractable Monte Carlo estimation is made possible:

$$\begin{aligned}
\nabla_\theta (V(\pi_\theta)) &= \nabla_\theta \left(\mathop{\mathbb{E}}_{x \sim \mathsf{P}(X)} \left[\mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot|x)} [R] \right] \right) \\
&= \nabla_\theta \left(\int \sum_{a \in \mathcal{A}} \pi_\theta(a \mid x) r(a, x) \mathsf{P}(X = x) \mathrm{d}x \right) \\
&= \int \sum_{a \in \mathcal{A}} \nabla_\theta (\pi_\theta(a \mid x) r(a, x)) \mathsf{P}(X = x) \mathrm{d}x \\
&= \int \sum_{a \in \mathcal{A}} \pi_\theta(a \mid x) \nabla_\theta (\log(\pi_\theta(a \mid x)) r(a, x)) \mathsf{P}(X = x) \mathrm{d}x \\
&= \mathop{\mathbb{E}}_{x \sim \mathsf{P}(X)} \left[\mathop{\mathbb{E}}_{a \sim \pi_\theta(\cdot|x)} [\nabla_\theta (\log(\pi_\theta(a \mid x)) R)] \right].
\end{aligned} \tag{4.3}$$

This provides an unbiased estimate of the gradient of $V(\pi_\theta)$. However, it may be subject to high variance due to the inherent variance of R. Several techniques have been proposed in the literature that aim to alleviate this, mostly using additive *control variates*.

Control variates are random variables with a known expectation [113, §8.9]. If the control variate is correlated with the original estimand – in our case $V(\pi_\theta)$ – they can be used to reduce the estimator's variance. A natural way to apply control variates to a sample average estimate for Eq. 4.1 is to estimate a model of the reward $\widehat{r}(a, x) \approx \mathbb{E}[R|X = x; A = a]$ and subtract it from the observed rewards [40]. This is at the heart of key RL techniques (i.a., generalised advantage estimation [136]), and it underpins widely used methods to increase sensitivity in online controlled experiments [10, 17, 33, 114]. As such, it applies to both *evaluation* and *learning* tasks. We note that if the model $\widehat{r}(a, x)$ is biased, this bias propagates to the resulting estimator for $V(\pi_\theta)$.

Alternatively, instead of focusing on reducing the variance of $V(\pi_\theta)$ directly, other often-used approaches tackle the variance of its gradient estimates $\nabla_\theta(V(\pi_\theta))$ instead.

Observe that $\mathbb{E}_{a\sim\pi_\theta(\cdot|x)}[\nabla_\theta(\log(\pi_\theta(a \mid x)))] = 0$ [101, Eq. 12]. This implies that a translation on the rewards in Eq. 4.3 does not affect the unbiasedness of the gradient estimate. Nevertheless, as such a translation can be framed as an additive control variate, it will affect its variance. Indeed, "*baseline corrections*" are a well-known variance reduction method for on-policy RL methods [43]. For a dataset consisting of logged contexts, actions and rewards $\mathcal{D} = \{(x_i, a_i, r_i)_{i=1}^N\}$, we apply a *baseline* control variate β to the estimate of the final gradient to obtain:

$$\begin{aligned}\nabla_\theta(V(\pi_\theta)) &\approx \nabla_\theta\widehat{(V_\beta(\pi_\theta))} \\ &= \frac{1}{|\mathcal{D}|}\sum_{(x,a,r)\in\mathcal{D}} (r-\beta)\nabla_\theta \log \pi_\theta(a \mid x).\end{aligned} \tag{4.4}$$

Williams [162] originally proposed to use the average observed reward for β. Subsequent work has derived optimal baselines for general on-policy RL scenarios [32, 43]. However, to the best of our knowledge, *optimal baselines for on-policy contextual bandits have not been considered in previous work.*

Optimal baseline for on-policy bandits. The optimal baseline β for the on-policy gradient estimate in Eq. 4.4 is the one that minimizes the variance of the gradient estimate. In accordance with earlier work [43], we define the variance of a vector random variable as the sum of the variance of its individual components. Therefore, the optimal baseline is given by:

$$\begin{aligned}&\arg\min_\beta \operatorname{Var}\Big(\nabla_\theta\widehat{(V_\beta(\pi_\theta))}\Big) \\ &= \arg\min_\beta \frac{1}{|\mathcal{D}|}\operatorname{Var}[\nabla_\theta(\log(\pi_\theta(a \mid x))(r-\beta))] &(4.5)\\ &= \arg\min_\beta \frac{1}{|\mathcal{D}|}\mathbb{E}\Big[\nabla_\theta \log(\pi_\theta(a \mid x))^\top \nabla_\theta \log(\pi_\theta(a \mid x))(r-\beta)^2\Big] &(4.6)\\ &\quad - \frac{1}{|\mathcal{D}|}\mathbb{E}[\nabla_\theta \log(\pi_\theta(a \mid x))(r-\beta)\,]^\top\,\mathbb{E}[\nabla_\theta \log(\pi_\theta(a \mid x))(r-\beta)]\end{aligned}$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{D}|} \mathbb{E}\Big[|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2 \, (r-\beta)^2 \Big], \tag{4.7}$$

where we ignore the second term in Eq. 4.6, since it is independent of β [101, Eq. 12]. The result from this derivation (Eq. 4.7) reveals that the optimal baseline can be obtained by solving the following equation:

$$\frac{\partial \mathrm{Var}\big(\nabla_\theta \widehat{(V_\beta(\pi_\theta))}\big)}{\partial \beta} = \frac{2}{|\mathcal{D}|} \mathbb{E}\big[|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2 (\beta - r)\big] = 0, \tag{4.8}$$

which results in the following optimal baseline correction:

$$\beta^* = \frac{\mathbb{E}\big[|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2 r(a,x)\big]}{\mathbb{E}[|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2]}, \tag{4.9}$$

and the empirical estimate of the optimal baseline correction:

$$\widehat{\beta^*} = \frac{\sum_{(x,a,r)\in\mathcal{D}} \big[|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2 r(a,x)\big]}{\sum_{(x,a,r)\in\mathcal{D}} [|\nabla_\theta \log(\pi_\theta(a \mid x))|_2^2]}. \tag{4.10}$$

This derivation follows the more general derivation from Greensmith et al. [43] for partially observable Markov decision processes (POMDPs). We have not encountered its use in the existing bandit literature applied to recommendation problems. In Section 4.3.2, we show that a similar line of reasoning can be applied to derive a variance-optimal gradient for the off-policy contextual bandit setup.

4.2.2 Off-policy estimation for general bandits

Deploying π is a costly prerequisite for estimating $V(\pi)$, that comes with the risk of deploying a possible poorly valued π. Therefore, commonly in real-world model validation pipelines, practitioners wish to estimate $V(\pi)$ *before* deployment. Accordingly, we will address this *counterfactual* evaluation scenario that falls inside the field of off-policy estimation (OPE) [130, 153].

The expectation $V(\pi)$ can be unbiasedly estimated using samples from a *different* policy π_0 through *importance sampling*, also known as inverse propensity score weighting (IPS) [113, §9]:

$$\mathop{\mathbb{E}}_{x\sim\mathsf{P}(X)}\left[\mathop{\mathbb{E}}_{a\sim\pi(\cdot|x)}[R]\right] = \mathop{\mathbb{E}}_{x\sim\mathsf{P}(X)}\left[\mathop{\mathbb{E}}_{a\sim\pi_0(\cdot|x)}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)} R\right]\right]. \tag{4.11}$$

To ensure that the so-called *importance weights* $\frac{\pi(a|x)}{\pi_0(a|x)}$ are well-defined, we assume "*common support*" by the logging policy: $\forall a \in \mathcal{A}, x \in \mathcal{X} : \pi(a \mid x) > 0 \implies \pi_0(a \mid x) > 0$.

From Eq. 4.11, we can derive an unbiased estimator for $V(\pi)$ using contexts, actions and rewards logged under π_0, denoted by $\mathcal{D}$:

$$\widehat{V}_{\mathrm{IPS}}(\pi, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a \mid x)}{\pi_0(a \mid x)} r. \tag{4.12}$$

To keep our notation brief, we suppress subscripts when they are clear from the context. In the context of gradient-based optimization methods, we often refer to a minibatch $\mathcal{B} \subset \mathcal{D}$ instead of the whole dataset, as is typical for, e.g., stochastic gradient descent (SGD).

If we wish to learn a policy that maximises this estimator, we need to estimate its gradient for a batch $\mathcal{B}$. Whilst some previous work has applied a REINFORCE estimator [25, 27, 97], we use a straightforward Monte Carlo estimate for the gradient:

$$\nabla \widehat{V}_{\mathrm{IPS}}(\pi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,a,r)\in\mathcal{B}} \frac{\nabla \pi(a \mid x)}{\pi_0(a \mid x)} r. \tag{4.13}$$

Importance sampling – the bread and butter of unbiased off-policy estimation – often leads to increased variance compared to on-policy estimators. Several variance reduction techniques have been proposed specifically to combat the excessive variance of $\widehat{V}_{\mathrm{IPS}}$ [35, 62, 147]. Within the scope of this chapter, we only consider techniques that reduce variance *without* introducing bias.

Self-normalised importance sampling. The key idea behind *self-normalisation* [113, §9.2] is to use a *multiplicative* control variate to rescale $\widehat{V}_{\mathrm{IPS}}(\pi, \mathcal{D})$. An important observation for this approach is that for any policy π and a dataset $\mathcal{D}$ logged under π_0, the expected average of importance weights should equal 1 [147, §5]:

$$\mathop{\mathbb{E}}_{\mathcal{D}\sim\mathsf{P}(\mathcal{D})} \left[\frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a \mid x)}{\pi_0(a \mid x)} \right] = 1. \tag{4.14}$$

Furthermore, as this random variable (Eq. 4.14) is likely to be correlated with the IPS estimates, we can expect that its use as a control variate will lead to reduced variance (see [e.g., 81]). This gives rise to the asymptotically unbiased and parameter-free self-normalised IPS (SNIPS) estimator, with $S \coloneqq \frac{1}{D} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a|x)}{\pi_0(a|x)}$ as its normalization term:

$$\widehat{V}_{\mathrm{SNIPS}}(\pi, \mathcal{D}) = \frac{\sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a|x)}{\pi_0(a|x)} r}{\sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a|x)}{\pi_0(a|x)}} = \frac{\widehat{V}_{\mathrm{IPS}}(\pi, \mathcal{D})}{S}. \tag{4.15}$$

Given the properties of being asymptotically unbiased and parameter-free, this estimator is often a go-to method for off-policy *evaluation* use-cases [130]. An additional advantage is that the SNIPS estimator is invariant to translations in the reward, which cannot be said for $\widehat{V}_{\mathrm{IPS}}$. Whilst the formulation in Eq. 4.15 is not obvious in this regard, it becomes clear when we consider its gradient:

$$\begin{aligned}
\nabla \widehat{V}_{\mathrm{SNIPS}}(\pi, \mathcal{D}) &= \nabla \left(\frac{\sum_{(x,a,r)} \frac{\pi(a|x)}{\pi_0(a|x)} r}{\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)}} \right) \\
&= \frac{\left(\sum_{(x,a,r)} \frac{\nabla \pi(a|x)}{\pi_0(a|x)} r \right) \left(\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)} \right)}{\left(\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)} \right)^2}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{\left(\sum_{(x,a,r)} \frac{\pi(a|x)}{\pi_0(a|x)} r\right)\left(\sum_{(x,a)} \frac{\nabla\pi(a|x)}{\pi_0(a|x)}\right)}{\left(\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)}\right)^2} \\
&= \frac{\sum_{(x_i,a_i,r_i)} \sum_{(x_j,a_j,r_j)} \frac{\pi(a_i|x_i)\nabla\pi(a_j|x_j)}{\pi_0(a_i|x_i)\pi_0(a_j|x_j)}(r_j - r_i)}{\left(\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)}\right)^2} \\
&= \frac{\sum\limits_{(x_i,a_i,r_i)} \sum\limits_{(x_j,a_j,r_j)} \frac{\pi(a_i|x_i)\pi(a_j|x_j)}{\pi_0(a_i|x_i)\pi_0(a_j|x_j)} \nabla \log \pi(a_j \mid x_j)(r_j - r_i)}{\left(\sum_{(x,a)} \frac{\pi(a|x)}{\pi_0(a|x)}\right)^2}.
\end{aligned}
\tag{4.16}
$$

Indeed, as the SNIPS gradient relies on the *relative difference* in observed reward between two samples, a constant correction would not affect it (i.e., if $\overline{r} = r - \beta$, then $r_j - r_i \equiv \overline{r}_j - \overline{r}_i$).

Swaminathan and Joachims [147] effectively apply the SNIPS estimator (with a variance regularisation term [148]) to off-policy *learning* scenarios. Note that while $\widehat{V}_{\text{IPS}}$ neatly decomposes into a single sum over samples, $\widehat{V}_{\text{SNIPS}}$ no longer does. Whilst this may be clear from the gradient formulation in Eq. 4.16, a formal proof can be found in [79, App. C]. This implies that mini-batch optimization methods (which are often necessary to support learning from large datasets) are no longer directly applicable to $\widehat{V}_{\text{SNIPS}}$.

Joachims et al. [79] solve this by re-framing the task of maximising $\widehat{V}_{\text{SNIPS}}$ as an optimization problem on $\widehat{V}_{\text{IPS}}$ with a constraint on the self-normalisation term. That is, if we define:

$$
\pi^\star = \arg\max_{\pi \in \Pi} \widehat{V}_{\text{SNIPS}}(\pi, \mathcal{D}), \text{ with } S^\star = \frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi^\star(a \mid x)}{\pi_0(a \mid x)},
\tag{4.17}
$$

then, we can equivalently state this as:

$$
\pi^\star = \arg\max_{\pi \in \Pi} \widehat{V}_{\text{IPS}}(\pi, \mathcal{D}), \text{ s.th. } \frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a \mid x)}{\pi_0(a \mid x)} = S^\star.
\tag{4.18}
$$

Joachims et al. [79] show via the Lagrange multiplier method that this optimization problem can be solved by optimising for $\widehat{V}_{\text{IPS}}$ with a translation on the reward:

$$
\begin{aligned}
\pi^\star &= \arg\max_{\pi \in \Pi} \widehat{V}_{\lambda^\star\text{-IPS}}(\pi, \mathcal{D}), \text{where} \\
\widehat{V}_{\lambda\text{-IPS}}(\pi, \mathcal{D}) &= \frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r - \lambda).
\end{aligned}
\tag{4.19}
$$

This approach is called BanditNet [79]. Naturally, we do not know $\lambda^\star$ beforehand (because we do not know $S^\star$), but we do know that $S^\star$ should concentrate around 1 for large datasets (see Eq. 4.14). Joachims et al. [79] essentially propose to treat λ as a hyper-parameter to be tuned in order to find $S^\star$.

Doubly robust estimation. Another way to reduce the variance of $\widehat{V}_{\mathrm{IPS}}$ is to use a model of the reward $\widehat{r}(a,x) \approx \mathbb{E}[R|X=x; A=a]$. Including it as an additive control variate in Eq. 4.12 gives rise to the doubly robust (DR) estimator, deriving its name from its unbiasedness if *either* the logging propensities π_0 or the reward model $\widehat{r}$ is unbiased [35]:

$$\widehat{V}_{\mathrm{DR}}(\pi, \mathcal{D}) = \tag{4.20}$$
$$\frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \left(\frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r - \widehat{r}(a,x)) + \sum_{a' \in \mathcal{A}} \pi(a' \mid x)\widehat{r}(a',x) \right).$$

Several further extensions have been proposed in the literature: one can optimize the reward model $\widehat{r}(a,x)$ to minimize the resulting variance of $\widehat{V}_{\mathrm{DR}}$ [38], further parameterise the trade-off relying on $\widehat{V}_{\mathrm{IPS}}$ or $\widehat{r}(a,x)$ [143], or shrink the IPS weights to minimize a bound on the MSE of the resulting estimator [144]. One disadvantage of this method, is that practitioners are required to fit the secondary reward model $\widehat{r}(a,x)$, which might be costly and sample inefficient. Furthermore, variance reduction is generally not guaranteed, and stand-alone $\widehat{V}_{\mathrm{IPS}}$ can be empirically superior in some scenarios [67].

4.3 Unifying Off-Policy Estimators

Section 4.2 provides an overview of (asymptotically) unbiased estimators for the value of a policy. We have introduced the contextual bandit setting, detailing often used variance reduction techniques for both on-policy (i.e., regression adjustments and baseline corrections) and off-policy estimation (i.e., self-normalisation and doubly robust estimation). In this section, we demonstrate that they perform equivalent optimization as baseline-corrected estimation. Subsequently, we characterize the baseline corrections that either minimize the variance of the estimator, or that of its gradient.

4.3.1 A unified off-policy estimator

Baseline corrections for $\nabla\widehat{V}_{\mathrm{IPS}}(\pi, \mathcal{D})$. Baseline corrections are common in on-policy estimation, but occur less often in the off-policy literature. The estimator is obtained by removing a baseline control variate $\beta \in \mathbb{R}$ from the reward of each action, while also adding it to the estimator:

$$\widehat{V}_{\beta\text{-IPS}} = \beta + \frac{1}{|\mathcal{D}|} \sum_{(x,a,r)\in\mathcal{D}} \frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r - \beta). \tag{4.21}$$

Its unbiasedness is easily verified:

$$\begin{aligned} \mathbb{E}\left[\widehat{V}_{\beta\text{-IPS}}\right] &= \mathbb{E}[\beta] + \mathbb{E}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r - \beta)\right] \\ &= \beta + \mathbb{E}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)} r\right] - \beta \qquad = V(\pi). \end{aligned} \tag{4.22}$$

From an optimization perspective, we are mainly interested in the gradient of the $\widehat{V}_{\beta\text{-IPS}}$ objective:

$$\nabla \widehat{V}_{\beta\text{-IPS}}(\pi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,a,r)\in\mathcal{B}} \frac{\nabla \pi(a \mid x)}{\pi_0(a \mid x)}(r - \beta). \tag{4.23}$$

Our key insight is that SNIPS and certain doubly-robust estimators have an equivalent gradient to the proposed β-IPS estimator. As a result, optimizing them is equivalent to optimizing $\widehat{V}_{\beta\text{-IPS}}$ for a specific β value.

Self-normalisation through BanditNet and $\widehat{V}_{\lambda\text{-IPS}}(\pi, \mathcal{D})$. If we consider the optimization problem for SNIPS that is solved by BanditNet in Eq. 4.19 [79], we see that its gradient is given by:

$$\nabla \widehat{V}_{\lambda\text{-IPS}}(\pi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,a,r)\in\mathcal{B}} \frac{\nabla \pi(a \mid x)}{\pi_0(a \mid x)}(r - \lambda). \tag{4.24}$$

Doubly robust estimation via $\widehat{V}_{\text{DR}}(\pi, \mathcal{D})$. As mentioned, a nuisance of doubly robust estimators is the requirement of fitting a regression model $\widehat{r}(a, x)$. Suppose that we instead treat $\widehat{r}$ as a single scalar hyper-parameter, akin to the BanditNet approach. Then, the gradient of such an estimator would be given by:

$$\nabla \widehat{V}_{\widehat{r}\text{-DR}}(\pi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(x,a,r)\in\mathcal{B}} \frac{\nabla \pi(a \mid x)}{\pi_0(a \mid x)}(r - \widehat{r}). \tag{4.25}$$

Importantly, these three approaches are motivated through entirely different lenses: minimizing gradient variance, applying a multiplicative control variate to reduce estimation variance, and applying an additive control variate to improve robustness. But they result in equivalent gradients, and thus, in equivalent optima. Specifically, for optimization, the estimators are equivalent when $\beta \equiv \lambda \equiv \widehat{r}$.

This equivalence implies that the choice between these three approaches is not important. Since the simple baseline correction estimator $\widehat{V}_{\beta\text{-IPS}}$ (Eq. 4.21) has an equivalence with all SNIPS estimators and all doubly-robust estimators with a constant reward, we propose that $\widehat{V}_{\beta\text{-IPS}}$ should be seen as an estimator that unifies all three approaches. Accordingly, we argue that the real task is to find the optimal β value for $\widehat{V}_{\beta\text{-IPS}}$, since this results in an estimator that is at least as optimal as any estimator in the underlying families of estimators, and possibly superior to them.

The remainder of this section describes the optimal β values for minimizing gradient variance and estimation value variance.

4.3.2 Minimizing gradient variance

Similar to the on-policy variant derived in Eq. 4.7, we can derive the optimal baseline in the off-policy case as the one which results in the minimum variance for the gradient estimate given by Eq. 4.13:

$$\arg\min_{\beta} \text{Var}\Big(\nabla_\theta(\widehat{V}_{\beta\text{-IPS}}(\pi_\theta, \mathcal{B}))\Big) \tag{4.26}$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{B}|} \mathrm{Var}\left[\frac{\nabla\pi(a \mid x)}{\pi_0(a \mid x)}(r-\beta)\right] \tag{4.27}$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{B}|} \mathbb{E}\left[\mid \nabla\pi(a \mid x)) \mid_2^2 \left(\frac{r-\beta}{\pi_0(a \mid x)}\right)^2\right] \tag{4.28}$$

$$- \frac{1}{|\mathcal{B}|} \mid \mathbb{E}\left[\frac{\nabla\pi(a \mid x)}{\pi_0(a \mid x)}(r-\beta)\right] \mid_2^2$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{B}|} \mathbb{E}\left[\frac{\mid \nabla\pi(a \mid x)) \mid_2^2}{\pi_0(a \mid x)^2}(r-\beta)^2\right], \tag{4.29}$$

where we can ignore the second term of the variance in Eq. 4.28, since it is independent of β [101, Eq. 12]. The optimal baseline can be obtained by solving for:

$$\frac{\partial \mathrm{Var}\Big(\nabla(\widehat{V}_{\beta\text{-IPS}}(\pi, \mathcal{B}))\Big)}{\partial\beta} = \frac{2}{|\mathcal{B}|} \mathbb{E}\left[\frac{\mid \nabla\pi(a \mid x)) \mid_2^2}{\pi_0(a \mid x)^2}(\beta - r)\right] = 0, \tag{4.30}$$

which results in the following optimal baseline:

$$\beta^* = \frac{\mathbb{E}_{x,a\sim\pi_0,r}\left[\frac{|\nabla\pi(a|x))|_2^2}{\pi_0(a|x)^2} r(a,x)\right]}{\mathbb{E}_{x,a\sim\pi_0,r}\left[\frac{|\nabla\pi(a|x))|_2^2}{\pi_0(a|x)^2}\right]}, \tag{4.31}$$

with its empirical estimate given by:

$$\widehat{\beta^*} = \frac{\sum_{(x,a,r)\in\mathcal{B}}\left[\frac{|\nabla\pi(a|x))|_2^2}{\pi_0(a|x)^2} r\right]}{\sum_{(x,a,r)\in\mathcal{B}}\left[\frac{|\nabla\pi(a|x))|_2^2}{\pi_0(a|x)^2}\right]}. \tag{4.32}$$

Note that this expectation is over actions sampled by the *logging* policy. As a result, we can obtain Monte Carlo estimates of the corresponding expectations. The derivation has high similarity with the on-policy case (cf. Section 4.2.1) [43]. Nevertheless, we are unaware of any work on off-policy learning that uses it. Joachims et al. [79] refer to the on-policy variant with: "*we cannot sample new roll-outs from the current policy under consideration, which means we cannot use the standard variance-optimal estimator used in REINFORCE*." Since the expectation is over actions sampled by the *logging* policy and not the *target* policy, we have shown that we do not need new roll-outs. Thereby, our estimation strategy is a novel off-policy approach that estimates the variance-optimal baseline.

Theorem 4.3.1. *Within the family of gradient estimators with a global additive control variate, i.e., β-IPS (Eq. 4.23), IPS (Eq. 4.13), BanditNet (Eq. 4.24), and DR with a constant correction (Eq. 4.25), β-IPS with our proposed choice of β in Eq. 4.31 has minimal gradient variance.*

Proof. Eq. 4.30 shows that the β value in Eq. 4.31 attains a minimum. Because the variance of the gradient estimate (Eq. 4.28) is a quadratic function of β, and hence a convex function (Eq. 4.29), it must be the global minimum for the gradient variance. □

4.3.3 Minimizing estimation variance

Besides minimizing gradient variance, one can also aim to minimize the variance of estimation, i.e., the variance of the estimated value. We note that the β value for minimizing estimation does not need to be the same value that minimizes gradient variance. Furthermore, since $\widehat{V}_{\beta\text{-IPS}}$ is unbiased, any estimation error will entirely be driven by variance. As a result, the value for β that results in minimal variance will also result in minimal estimation error:

$$\arg\min_{\beta} \operatorname{Var}\Big(\widehat{V}_{\beta\text{-IPS}}(\pi, \mathcal{D})\Big) \tag{4.33}$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{D}|} \operatorname{Var}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r-\beta)\right] \tag{4.34}$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{D}|} \mathbb{E}\left[\Big(\frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r-\beta)\Big)^2\right] \tag{4.35}$$
$$- \frac{1}{|\mathcal{D}|}\Big(\mathbb{E}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)}(r-\beta)\right]\Big)^2$$

$$= \arg\min_{\beta} \frac{1}{|\mathcal{D}|} \mathbb{E}\left[\left(\frac{\pi(a \mid x)}{\pi_0(a \mid x)}\right)^2 (r-\beta)^2\right] \tag{4.36}$$
$$- \frac{1}{|\mathcal{D}|}\Big(\mathbb{E}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)} r\right] - \beta\Big)^2.$$

The minimum is obtained by solving for the following equation:

$$\frac{\partial\Big(\operatorname{Var}\Big(\widehat{V}_{\beta\text{-IPS}}(\pi, \mathcal{D})\Big)\Big)}{\partial\beta} \tag{4.37}$$
$$= \frac{2}{|\mathcal{D}|} \mathbb{E}\left[\left(\frac{\pi(a \mid x)}{\pi_0(a \mid x)}\right)^2 (\beta - r)\right] - \frac{2}{|\mathcal{D}|}\left(\beta - \mathbb{E}\left[\frac{\pi(a \mid x)}{\pi_0(a \mid x)} r\right]\right) = 0,$$

which results in the following optimal baseline:

$$\beta^* = \frac{\mathbb{E}\left[\left(\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2 - \frac{\pi(a|x)}{\pi_0(a|x)}\right) r(a,x)\right]}{\mathbb{E}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2 - \left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)\right]}. \tag{4.38}$$

We can estimate β^* using logged data, resulting in a Monte Carlo estimate of the optimal baseline. Such a sample estimate will not be unbiased (because it is a ratio of expectations), but the bias will vanish asymptotically (similar to the bias of the $\widehat{V}_{\text{SNIPS}}$ estimator).

Next, we formally prove that optimal estimator variance leads to overall optimality (in terms of the MSE of the estimator).

Theorem 4.3.2. *Within the family of offline estimators with a global additive control variate, i.e., β-IPS (Eq. 4.21), IPS (Eq. 4.12), and DR with a constant correction*

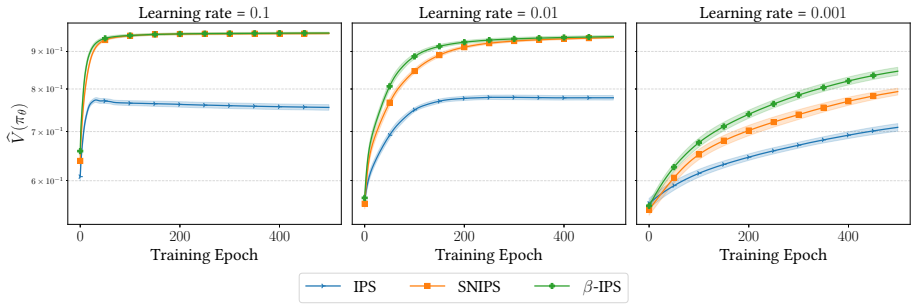

Figure 4.1: Performance of different off-policy learning methods trained in a full-batch gradient descent fashion in terms of the policy value on the test set. x-axis corresponds to the training epoch during the optimization (we use a maximum of 500 epochs for all methods), and y-axis corresponds to the policy value. A decaying learning rate is used. Reported results are averages over 32 independent runs with 95% confidence interval.

(Eq. 4.25), β-IPS with our proposed β in Eq. 4.38 has the minimum mean squared error (MSE):

$$\text{MSE}(\hat{V}(\pi)) = \mathbb{E}_{\mathcal{D}}\left[(\hat{V}(\pi, \mathcal{D}) - V(\pi))^2\right]. \tag{4.39}$$

Proof. The MSE of any off-policy estimator $\hat{V}(\pi, \mathcal{D})$ can be decomposed in terms of the bias and variance of the estimator [144]:

$$\text{MSE}(\hat{V}(\pi)) = \text{Bias}(\hat{V}(\pi), \mathcal{D})^2 + \text{Variance}(\hat{V}(\pi), \mathcal{D}), \tag{4.40}$$

where the bias of the estimator is defined as:

$$\text{Bias}(\hat{V}(\pi), \mathcal{D}) = \left|\mathbb{E}_{\mathcal{D}}\left[\hat{V}(\pi, \mathcal{D}) - V(\pi)\right]\right|, \tag{4.41}$$

and the variance of the estimator is defined previously (see Section 4.3.3). Eq. 4.22 proves that β-IPS is unbiased: $\text{Bias}(\hat{V}(\pi), \mathcal{D}) = 0$. Thus, the minimum variance (Eq. 4.37) implies minimum MSE. □

We note that SNIPS is not covered by this theorem, as it is only asymptotically unbiased. As a result, the variance reduction brought on by SNIPS might be higher than that by β-IPS, but as it introduces bias, its estimation error (MSE) is not guaranteed to be better. Our experimental results below indicate that our method is always at least as good as SNIPS, and outperforms it in most cases, in both learning and evaluation tasks.

4.4 Experimental Setup

In order to evaluate off-policy learning and evaluation methods, we need access to logged data sampled from a stochastic policy involving logging propensities (exact or estimated) along with the corresponding context and action pairs. Recent work that focuses on off-policy learning or evaluation for contextual bandits in recommender

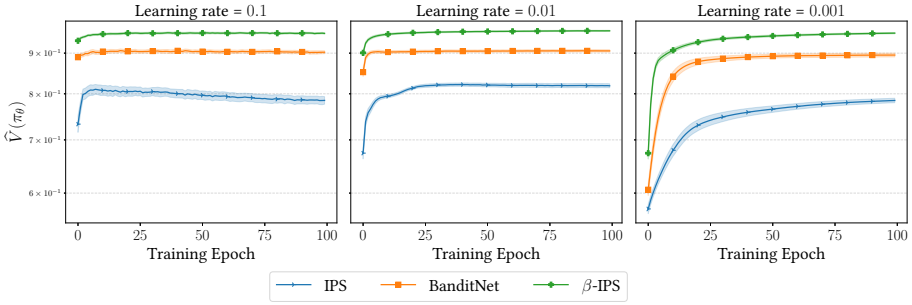

Figure 4.2: Performance of different off-policy learning methods trained in a mini-batch gradient descent fashion in terms of policy value on the test set. The axis labels are similar to Figure 4.1.

systems follows a supervised-to-bandit conversion process to simulate a real-world bandit feedback dataset [68, 69, 73, 130, 143, 144], or conducts a live experiment on actual user traffic to evaluate the policy in an *on-policy* or *online* fashion [25, 27]. In this work, we adopt the Open Bandit Pipeline (OBP) to simulate, in a reproducible manner, real-world recommendation setups with stochastic rewards, large action spaces, and controlled randomization [124]. Although the Open Bandit Pipeline simulates a generic offline contextual bandit setup, there is a strong correspondence to real-world recommendation setups where the environment context vector corresponds to the user context and the actions correspond to the items recommended to the user. Finally, the reward corresponds to the user feedback received on the item (click, purchase, etc.). As an added advantage, the simulator allows us to conduct experiments in a *realistic* setting where the logging policy is sub-optimal to a controlled extent, the logged data size is limited, and the action space is large. In addition, we conduct experiments with real-world recommendation logs from the OBP for off-policy evaluation.[1]

The research questions we answer with our experimental results in this chapter are:

RQC1 Does the proposed *estimator-variance-minimizing* baseline correction (Eq. 4.38) improve off-policy learning (OPL) in a full-batch setting?

RQC2 Does the proposed *gradient-variance-minimizing* baseline correction (Eq. 4.31) improve OPL in a mini-batch setting?

RQC3 How does the proposed *gradient-variance-minimizing* baseline correction (Eq. 4.31) affect gradient variance during OPL?

RQC4 Does the proposed *estimator-variance-minimizing* baseline correction (Eq. 4.38) improve off-policy evaluation (OPE) performance?

[1] `https://research.zozo.com/data.html`

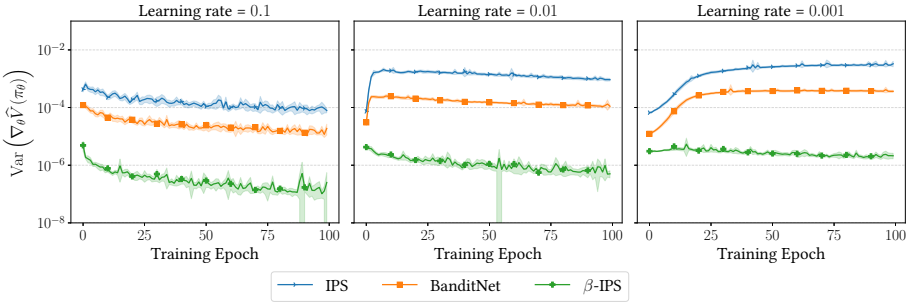

Figure 4.3: Empirical variance of the gradient of different off-policy learning estimators in a mini-batch optimization setup with varying learning rates (in title). We compute gradient variance for each mini-batch during training and then report the average value across all mini-batches in a training epoch. Results are averaged across 32 independent runs with 95% confidence interval.

4.5 Results and Discussion

4.5.1 Off-policy learning performance (RQC1–3)

To evaluate the performance of the proposed β-IPS method on an OPL task, we consider two learning setups:

1. *Full-batch*. In this setup, we directly optimize the β-IPS policy value estimator (Eq. 4.21) with the optimal baseline correction, which minimizes the variance of the value (Eq. 4.38). Given that the optimal baseline correction involves a ratio of two expectations, optimizing the value function directly via a mini-batch stochastic optimization is not possible for the same reason as the SNIPS estimator, i.e., it is not possible to get an unbiased gradient estimate with a ratio function [79]. Therefore, for this particular setting, we use a *full-batch* gradient descent method for the optimization, where the gradient is computed over the entire training dataset.

2. *Mini-batch*. In this setup, we focus on optimizing the β-IPS policy value estimator with the baseline correction, which minimizes the gradient estimate (Eq. 4.31). This setup translates to a traditional machine learning training setup where the model is optimized in a stochastic mini-batch fashion.

Full-batch. The results for the full-batch training in terms of the policy value on the test set are reported in Figure 4.1, over the number of training epochs. To minimize the impact of external factors, we use a linear model without bias, followed by a softmax to generate a distribution over all actions, given a context vector x (this is a common setup, see [e.g., 67, 71, 131]). We note that the goal of this chapter is not to get the maximum possible policy value on the test set but rather to evaluate the effect of baseline corrections on gradient and estimation variance. The simple model setup allows us to easily track the empirical gradient variance, given that we have only one parameter vector.

An advantage of the full-batch setup is that we can compute the gradient of the SNIPS estimator directly [147]. SNIPS is a natural baseline method to consider, along with the traditional IPS estimator. Because of practical concerns, we only consider 500 epochs of optimization. Additionally, we use the state-of-the-art and widely used Adam optimizer [80].

The IPS method converges to a lower test policy value in comparison to the SNIPS and the proposed β-IPS methods, even after 500 epochs. A likely reason is the high-variance of the IPS estimator [35], which can cause it to get stuck in bad local minima.

The methods with a control variate, i.e., SNIPS (with multiplicative control variate) and β-IPS (with additive control variate) converge to substantially better test policy values. In terms of the convergence speed, β-IPS converges to the optimal value faster than the SNIPS estimator, most likely because it has lower estimator variance than SNIPS. With this, we can answer RQC1 as follows: in the full-batch setting, our proposed optimal baseline correction enables β-IPS to converge faster than SNIPS at similar performance.

Mini-batch. The results for mini-batch training in terms of the test policy value are reported in Figure 4.2. Different from the full-batch setup, where the focus is on reducing the variance of the *estimator* value (Section 4.3.3), in the mini-batch mode, the focus is on reducing the variance of the gradient estimate (Section 4.3.2). The model and training setup are similar to the full-batch mode, except that we fixed the batch size to 1024 for the mini-batch experiments. Preliminary results indicated that the batch size hyper-parameter has a limited effect.

Analogous to the full-batch setup, the IPS estimator results in a lower test policy value, most likely because of the high gradient variance which prevents convergence to high performance. In contrast, due to their baseline corrections, BanditNet (Eq. 4.24) and β-IPS have a lower gradient variance. Accordingly, they also converge to better performance [15], i.e., resulting in superior test policy values.

Amongst these baseline-corrected gradient-based methods (BanditNet and β-IPS), our proposed β-IPS estimator outperforms BanditNet as it provides a policy with substantially higher value. The differences are observed over different choices of learning rates. Thus we answer RQC2 accordingly: in the mini-batch setting, our proposed gradient-minimizing baseline method results in considerably higher policy value compared to both IPS and BanditNet.

Next, we directly consider the empirical gradient variance of different estimators; Figure 4.3 reports the average mini-batch gradient variance per epoch. As expected, the IPS estimator has the highest gradient variance by a large margin. For BanditNet, we observe a lower gradient variance, which is the desired result of the additive baseline it employs. Finally, we observe that our proposed method β-IPS has the lowest gradient variance. This result corroborates the theoretical claim (Theorem 4.3.1) that the β-IPS estimator has the lowest gradient variance amongst all global additive control variates (including IPS and BanditNet). Our answer to RQC3 is thus clear: our proposed β-IPS results in considerably lower gradient variance compared to BanditNet and IPS.

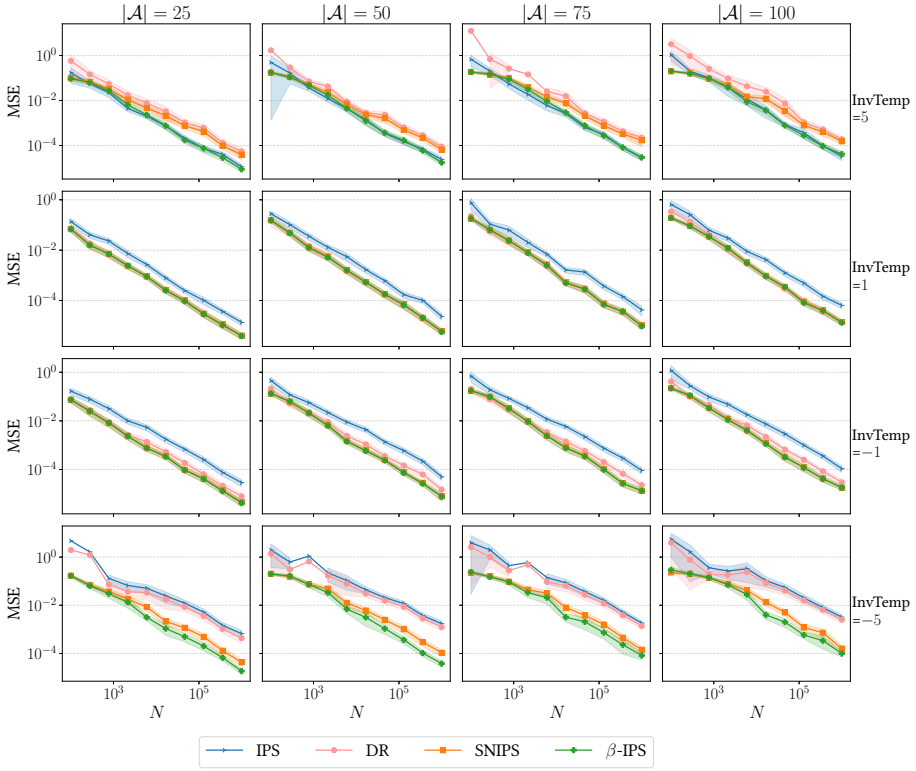

Figure 4.4: Mean Squared Error (MSE) of different off-policy estimators with varying action space (from left to right), and varying inverse temperature parameter of the softmax logging policy (from top to bottom). X-axis corresponds to the size of the logged data simulated (ranging from 10^2 to 10^6), and the y-axis corresponds to the MSE (evaluated over 100 independent samples of the synthetic data) along with 95% confidence interval. Each row corresponds to a different setting of inverse temperature of the softmax logging policy. We only consider unbiased (asymptotically or otherwise) estimators.

4.5.2 Off-policy evaluation performance (RQC4)

To evaluate the performance of the proposed β-IPS method, which minimizes the estimated policy value (Eq. 4.38), in an OPE task, results are presented in Figure 4.4. The target policy (to be evaluated) is a logistic regression model trained via the IPS objective (Eq. 4.13) on logged data and evaluated on a separate full-information test set. We evaluate the MSE of the estimated policy value against the *true* policy value (Eq. 4.39). To evaluate the MSE of different estimators realistically, we report results with varying degrees of the optimality of the behavior policy (decided by the inverse temperature parameter of the softmax) and with a varying cardinality of the action space. A positive (and higher) inverse softmax temperature results in a increasingly optimal

Table 4.1: Comparison of different OPE methods on real-world recommender system logs of ZOZOTOWN from a campaign targeted towards men with a uniformly random production policy. We report the mean relative absolute error (with std).

OPE estimator	**Abs. relative error** $\downarrow$
IPS	0.1277 (0.0142)
SNIPS	0.1113 (0.0372)
DR	0.1144 (0.0366)
β-IPS	**0.1078 (0.0383)**

behavior policy (selects action with highest reward probability), and a negative (and lower) inverse softmax temperature parameter results in an increasingly sub-optimal behavior policy (selects actions with lowest reward probability). Our proposed β-IPS method has the lowest MSE in all simulated settings. Interestingly, the proposed β-IPS has a lower MSE than the DR method, which has a regression model-based control variate, arguably more powerful than the constant control variate from the proposed β-IPS method. Similar observations have been made in previous work, e.g., Jeunen and Goethals [67] reported that the DR estimator's performance heavily depends on the logging policy.

Depending on the setting, we see that β-IPS either has performance comparable to the SNIPS estimator, i.e., when inverse temperature $\in \{-1, 1\}$; or noticeably higher performance than SNIPS, i.e., when inverse temperature $\in \{-5, 5\}$.

Real-world evaluation. To evaluate different estimators in a real-world recommender systems setup, we report the results of OPE from the production logs of a real-world recommender system in Table 4.1. Similar to the simulation setup, the proposed β-IPS has the lowest absoluate relative error amongst all estimators in the comparison. In conclusion, we answer RQC4: our proposed policy-value variance minimizing baseline method results in substantially improved MSE, compared to IPS, SNIPS and DR, in offline evaluation tasks that are typical recommender system use-cases.

4.6 Conclusion and Future Work

In this chapter, we have proposed to unify different off-policy estimators as equivalent additive baseline corrections. We look at off-policy evaluation and learning settings and propose baseline corrections that minimize the variance in the estimated policy value and the empirical gradient of the off-policy learning objective. Extensive experimental comparisons on a synthetic benchmark with realistic settings show that our proposed methods improve performance in the off-policy estimation (OPE) and off-policy learning (OPL) tasks.

We believe our work in this chapter represents a significant step forward in the understanding and use of off-policy estimation methods (for both evaluation and learning use-cases), since we show that the prevalent SNIPS estimator can be improved upon with essentially no cost, as our proposed method is parameter-free and – in contrast with SNIPS – it retains the unbiasedness that comes with IPS. Future work may apply a

similar approach to offline reinforcement learning setups [86], or consider extensions of our approach for ranking applications [94].

In this chapter, we answer the RQ4, and RQ3 in the affirmative. First, we present a unifying framework for off-policy evaluation and learning tasks :β-IPS. The new estimator β-IPS combines most commonly estimators such as: inverse propensity scoring (IPS), doubly robust (DR), and self-normalized IPS (SNIPS). Further, we presented a closed-form solution baseline correction term – β that minimizes variance for both off-policy learning and evaluation tasks.

Broadly, in this chapter, we explored off-policy evaluation and learning estimators derived from logged user interactions in recommender systems. In the next chapter, we turn to contextual-bandit learning within a diffusion-model framework, aiming to optimize arbitrary user-defined objectives.

4.A Appendix: Off-policy Estimator Variance

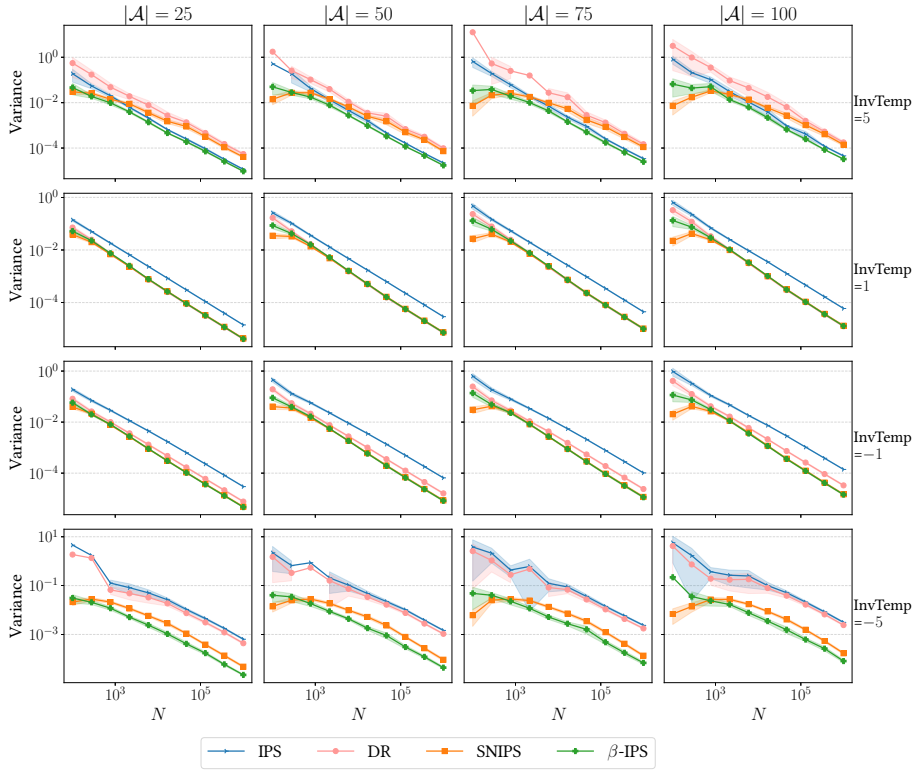

Figure 4.A.1: Empirical variance of different off-policy estimators with varying action space (from left to right), and varying sub-optimality of a temperature-based softmax behavior policy (from top to bottom). The x-axis corresponds to the size of the logged data simulated (ranging from 10^2 to 10^6), and the y-axis corresponds to the variance of different estimators (evaluated over 100 independent samples of the synthetic data) along with 95% confidence interval. Each row corresponds to a different optimality level of the logging policy, decided by the inverse temperature parameter. We only consider unbiased (asymptotically or otherwise) estimators.

In this chapter appendix, we report additional results from the experimental section (Section 4.5) from the main chapter), answering RQC4. Specifically, we look the the empirical variance of various offline estimators for the task of off-policy evaluation. The mean squared error (MSE) of different offline estimators are reported in Figure 4.4. In this appendix, we report the empirical variance of various offline estimators in Figure 4.A.1.

From the figure, it is clear that our proposed β-IPS estimator with estimator variance

minimizing β value (Eq. 4.38) results in the lowest empirical variance in most of the cases. It is interesting to note that when the logged data is limited ($N < 10^3$), sometimes the SNIPS estimator has lower estimator variance. We suspect that the reason could be a bias in the estimate of the variance-optimal β estimate (Eq. 4.38), when the dataset size is small, given that it is a ratio estimate of expectations. For practical settings, i.e., when $N > 10^3$, the proposed estimator β-IPS results in a minimum sample variance, thereby empirically validating the effectiveness of our proposed β-IPS estimator for the task of OPE.

5

A Simple and Effective Reinforcement Learning Method for Text-to-Image Diffusion Models

So far in this thesis, we focused on learning from user interactions via contextual bandits within ranking or recommendation systems. However, the contextual bandit framework has recently also been effectively employed in fine-tuning foundation models, such as textual large language models (LLMs) and diffusion models. The typical choice of method for model optimization is proximal policy optimization (PPO) [137]. While effective, recent research has highlighted computational advantages of REINFORCE (policy gradient methods) over PPO for text-based LLMs [5]. Given PPO's ongoing challenges with variance and sample inefficiency, we consider improvements through our final research question:

RQ5 Can we improve the sample efficiency of proximal policy optimization for fine-tuning text-to-image diffusion?

In this chapter, we systematically compare PPO and REINFORCE for diffusion model fine-tuning. In the first part of this chapter, we demonstrate that REINFORCE exhibits inferior sample efficiency compared to PPO. Subsequently, we propose LOOP, an enhancement to PPO achieving superior performance with the same number of input prompts by generating multiple actions per prompt.

5.1 Introduction

Diffusion models have emerged as a powerful tool for generative modeling [57, 140], with a strong capacity to model complex data distributions from various modalities, like images [125], text [6], natural molecules [168], and videos [14].

Diffusion models are typically pre-trained on a large-scale dataset, such that they can subsequently generate samples from the same data distribution. The training objective typically involves maximizing the data distribution likelihood. This pre-training stage helps generate high-quality samples from the model. However, some applications might

This chapter was published as [54].

require optimizing a custom reward function, for example, optimizing for generating aesthetically pleasing images [167], semantic alignment of image-text pairs based on human feedback [134], or generating molecules with specific properties [155].

To optimize for such black-box objectives, RL-based fine-tuning has been successfully applied to diffusion models [13, 37, 44, 88, 154]. For RL-based fine-tuning, the reverse diffusion process is treated as a Markov decision process (MDP), wherein prompts are treated as part of the input state, the generated image at each time-step is mapped to an action, which receives a reward from a fixed reward model (environment in standard MDP), and finally the diffusion model is treated as a policy, which we optimize to maximize rewards. For optimization, typically PPO is applied [13, 37]. In applications where getting a reward model is infeasible or undesirable, "RL-free" fine-tuning (typically offline) can also be applied [154]. For this chapter, we only focus on diffusion model fine-tuning using "online" RL methods, specifically PPO [137].

An advantage of PPO is that it removes the incentive for the new policy to deviate too much from the previous reference policy, via importance sampling and clipping operation [137]. While effective, PPO can have a significant computational overhead. In practice, RL fine-tuning for diffusion models via PPO requires concurrently loading three models in memory:

(i) The **reference policy**: The base policy, which is usually initialized with the pre-trained diffusion model.

(ii) The **current policy**: The policy that is RL fine-tuned, and also initialized with the pre-trained diffusion model.

(iii) The **reward model**: Typically, a large vision-language model, trained via supervised fine-tuning objective [85], which assigns a scalar reward to the final generated image during the online optimization stage.

This can result in a considerable computational burden, given that each policy can potentially have millions of parameters. In addition to its computational overhead, PPO is also known to be sensitive to hyper-parameters [36, 61, 174].

Simpler approaches, like REINFORCE [164] avoid such complexities, and could theoretically be more efficient. However, in practice, they often suffer from high variance and instability. Recently, a variant of REINFORCE: reinforce leave-one-out (RLOO) [82] was proposed which samples multiple sequences per input prompt, and a baseline correction term to reduce the variance, however, it still suffers from sample inefficiency.

This raises a fundamental question about the **efficiency-effectiveness** trade-off in RL-based diffusion fine-tuning. In this chapter, first we systematically explore this trade-off between *efficiency* – a lower computational cost, and reduced implementation complexity (i.e., fewer hyper-parameters) – and *effectiveness* – stable training, and final performance. We compare a simple REINFORCE approach with the standard PPO framework, demonstrating that while REINFORCE greatly reduces computational complexity, it comes at the cost of reduced performance.

Motivated by this finding, we propose a novel RL for diffusion fine-tuning method, LOOP, which combines the best of the both worlds. To reduce the variance during policy

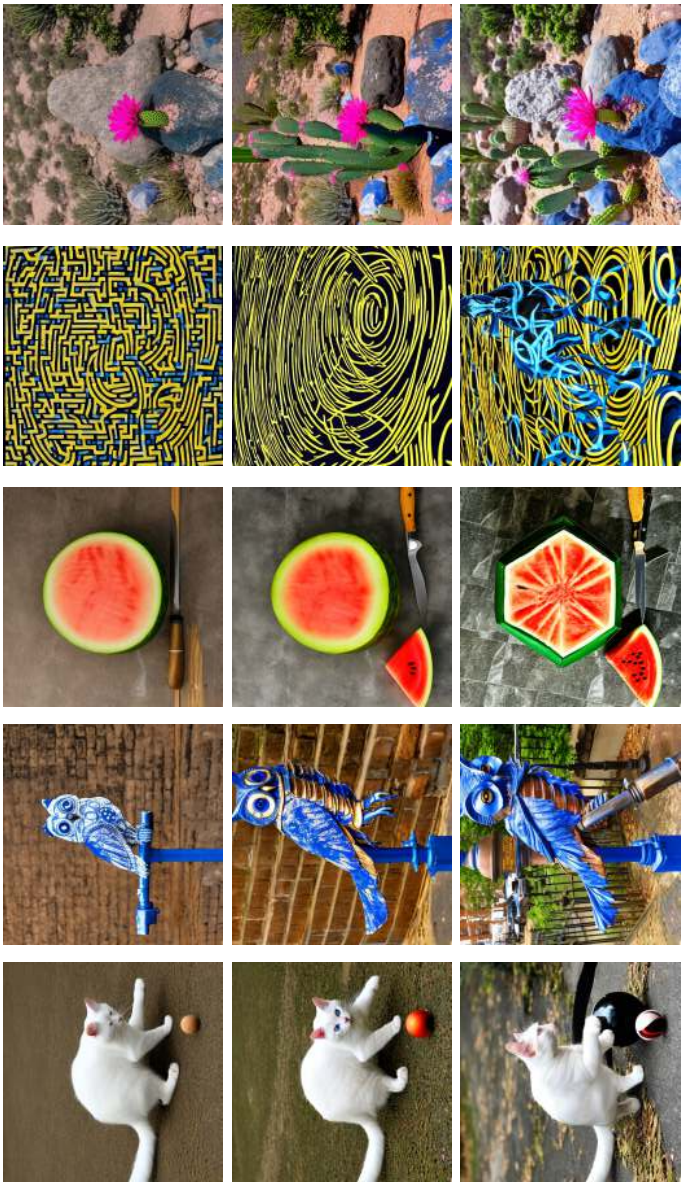

Figure 5.1: **LOOP improves attribute binding**. Qualitative examples generated with SD 2.0 (first row), PPO (second row) and LOOP k=4 (third row). In each prompt SD and PPO miss at least one listed attribute, whereas LOOP succeeds.

optimization, LOOP uses multiple actions (diffusion trajectories) and a (REINFORCE) baseline correction term per input prompt. To maintain the stability and robustness of PPO, LOOP uses clipping and importance sampling.

We choose the text-to-image compositionality benchmark (T2I-CompBench; Huang et al., 2023) as our primary evaluation benchmark. Text-to-image models often fail to satisfy an essential reasoning ability of attribute binding, i.e., the generated image often fails to *bind* certain *attributes* specified in the instruction prompt [41, 60, 121]. As illustrated in Figure 5.1, LOOP outperforms previous diffusion methods on attribute binding. As attribute binding is a key skill necessary for real-world applications, we choose the T2I-CompBench benchmark alongside two other common tasks: aesthetic image generation and image-text semantic alignment.

To summarize, our main contributions are as follows:

PPO vs. REINFORCE efficiency-effectiveness trade-off. We systematically study how design elements like clipping, reference policy, value function in PPO compare to a simple REINFORCE method, highlighting the efficiency-effectiveness trade-off in diffusion fine-tuning. To the best of our knowledge, we are the first ones to present such a systematic study, highlighting the trade-offs in diffusion fine-tuning.

Introducing LOOP. We propose LOOP, a novel RL for diffusion fine-tuning method combining the best of REINFORCE and PPO. LOOP uses multiple diffusion trajectories and a REINFORCE baseline correction term for variance reduction, as well as clipping and importance sampling from PPO for robustness and sample efficiency.

Empirical validation. To validate our claims empirically, we conduct experiments on the T2I-CompBench benchmark image compositionality benchmark. The benchmark evaluates the attribute binding capabilities of the text-to-image generative models and shows that LOOP succeeds where previous text-to-image generative models often fail. We also evaluate LOOP on two common objectives from the RL for diffusion literature: image aesthetics, and text-image semantic alignment [13].

5.2 Background and Related Work

5.2.1 Diffusion models

We focus on denoising diffusion probabilistic models (DDPM) as the base model for text-to-image generative modeling [57, 140]. Briefly, given a conditioning context variable $\mathbf{c}$ (text prompt in our case), and the data sample $\mathbf{x}_0$, DDPM models $p(\mathbf{x}_0 \mid \mathbf{c})$ via a Markov chain of length T, with the following dynamics:

$$p_\theta(\mathbf{x}_{0:T} \mid \mathbf{c}) = p(\mathbf{x}_T \mid \mathbf{c}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}). \tag{5.1}$$

Image generation in diffusion model is achieved via the following ancestral sampling scheme, which is a reverse diffusion process:

$$\begin{aligned} &\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \\ &\mathbf{x}_t \sim N\left(\mathbf{x}_t \mid \mu_\theta(\mathbf{x}_t, \mathbf{c}, t), \sigma_\theta^2 I\right), \forall t \in [0, T-1], \end{aligned} \tag{5.2}$$

where the distribution at time-step t is assumed to be a multivariate normal distribution with the predicted mean $\mu_\theta(\mathbf{x}_t, \mathbf{c}, t)$, and a constant variance.

5.2.2 Proximal policy optimization for RL

PPO was introduced for optimizing a policy with the objective of maximizing the overall reward in the RL paradigm [137]. PPO removes the incentive for the current policy π_t to diverge from the previous policy π_{t-1} outside the range $[1-\epsilon, 1+\epsilon]$, where ϵ is a hyper-parameter. As long as the subsequent policies are closer to each other in the action space, the monotonic policy improvement bound guarantees a monotonic improvement in the policy's performance as the optimization progresses. This property justifies the clipping term in the mathematical formulation of the PPO objective function [1, 117, 135]. Formally, PPO the objective function is:

$$J(\theta) = \mathbb{E}\Big[\min\Big(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\Big)\Big], \tag{5.3}$$

where $r_t(\theta) = \frac{\pi_t(a|c)}{\pi_{t-1}(a|c)}$ is the importance sampling ratio between the current policy $\pi_t(a \mid c)$ and the previous reference policy $\pi_{t-1}(a \mid c)$, $\hat{A}_t$ is the advantage function [146], and the clip operator restricts the importance sampling ratio in the range $[1-\epsilon, 1+\epsilon]$.

5.2.3 RL for text-to-image diffusion models

The diffusion process can be viewed as an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \rho_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the state transition kernel, $\mathcal{R}$ is the reward function, and ρ_0 is the distribution of initial state $\mathbf{s_0}$. In the context of text-to-image diffusion models, the MDP is defined as:

$$\begin{aligned}
\mathbf{s_t} = (\mathbf{c}, t, \mathbf{x_t}), \;\; \pi_\theta(\mathbf{a_t} \mid \mathbf{s_t}) &= p_\theta(\mathbf{x_{t-1}} \mid \mathbf{x_t}, \mathbf{c}), \\
\mathcal{P}(\mathbf{s_{t+1}} \mid \mathbf{s_t}, \mathbf{a_t}) &= \delta\big(\mathbf{c}, \mathbf{a_t}\big), \quad \mathbf{a_t} = \mathbf{x_{t-1}}, \\
\rho_0(\mathbf{s_0}) &= \big(p(\mathbf{c}), \delta_T, \mathcal{N}(0, \mathbf{I})\big), \\
\mathcal{R}(\mathbf{s_t}, \mathbf{a_t}) &= \begin{cases} r(\mathbf{x_0}, \mathbf{c}) & \text{if } t = 0, \\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{5.4}$$

The input state $\mathbf{s_t}$ is defined in terms of the context (prompt features), sampled image at the given time-step t. The policy π_θ is the diffusion model itself. The state transition kernel is a dirac delta function δ with the current sampled action $\mathbf{x}_t$ as the input. The reward is assigned only at the last step in the reverse diffusion process, when the final image is generated. The initial state ρ_0 corresponds to the last state in the forward diffusion process: $\mathbf{x}_T$.

5.2.4 PPO for diffusion fine-tuning

The objective function of RL fine-tuning for a diffusion policy π_θ can be defined as follows:

$$\begin{aligned} J_\theta(\pi) &= \mathbb{E}_{\tau\sim p(\tau|\pi_\theta)}\left[\sum_{t=0}^{T}\mathcal{R}(\mathbf{s}_t,\mathbf{a}_t)\right] \\ &= \mathbb{E}_{\tau\sim p(\tau|\pi_\theta)}\left[r(\mathbf{x}_0,\mathbf{c})\right], \end{aligned} \tag{5.5}$$

where the trajectory $\tau = \{\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0\}$ refers to the reverse diffusion process (Eq. 5.1), and the total reward of the trajectory is the reward of the final generated image $\mathbf{x}_0$ (Eq. 5.4). We ignore the KL-regularized version of the equation, which is commonly applied in the RLHF for LLM literature [120, 172, 175], and proposed by Fan et al. [37] in the context of RL for diffusion models. As shown by Black et al. [13], adding the KL-regularization term makes no empirical difference in terms of the final performance. The PPO objective is given as:

$$J_\theta^{\mathrm{PPO}}(\pi) = \mathbb{E}\left[\sum_{t=0}^{T}\mathrm{clip}\left(\frac{\pi_\theta(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{c})}{\pi_{\text{old}}(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{c})},1-\epsilon,1+\epsilon\right)r(\mathbf{x}_0,\mathbf{c})\right], \tag{5.6}$$

where the clipping operation removes the incentive for the new policy π_θ to differ from the previous round policy π_{old} [13, 137].

5.3 REINFORCE vs. PPO: An Efficiency-Effectiveness Trade-Off

In this section, we explore the efficiency-effectiveness trade-off between two prominent reinforcement learning methods for diffusion fine-tuning: REINFORCE and PPO. Understanding this trade-off is crucial for selecting the appropriate algorithm given constraints on computational resources and desired performance outcomes.

In the context of text-to-image diffusion models, we aim to optimize the policy π to maximize the expected reward $\mathcal{R}(x_{0:T}, c) = r(x_0, c)$. Our objective function is defined as:

$$J_\theta(\pi) = \mathbb{E}_{c\sim p(C),x_{0:T}\sim p_\theta(x_{0:T}|c)}\left[r(x_0,c)\right]. \tag{5.7}$$

REINFORCE for gradient calculation. For optimizing this objective, the REINFORCE policy gradient (also known as score function (SF)) [164] provides the following gradient estimate:

$$\begin{aligned} \nabla_\theta J_\theta^{\mathrm{SF}}(\pi) &= \mathbb{E}_{\mathbf{x}_{0:T}}\left[\nabla_\theta\log\left(\prod_{t=1}^{T}p_\theta\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{c}\right)\right)r\left(\mathbf{x}_0,\mathbf{c}\right)\right] \\ &= \mathbb{E}_{\mathbf{x}_{0:T}}\left[\sum_{t=0}^{T}\nabla_\theta\log p_\theta\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{c}\right)r\left(\mathbf{x}_0,\mathbf{c}\right)\right], \end{aligned} \tag{5.8}$$

where the second step follows from the reverse diffusion policy decomposition (Eq. 5.1).

In practice, a batch of trajectories is sampled from the reverse diffusion distribution, i.e., $\mathbf{x}_{0:T} \sim p_\theta(\mathbf{x}_{0:T})$, and a Monte Carlo estimate of the REINFORCE policy gradient (Eq. 5.8) is calculated for the model update.

REINFORCE with baseline correction. To reduce variance of the REINFORCE estimator, a common trick is to subtract a constant baseline correction term from the reward function [43, 101]:

$$\nabla_\theta J_\theta^{\mathrm{SFB}}(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \nabla_\theta \log p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})(r(\mathbf{x}_0, \mathbf{c}) - b_t)\right]. \tag{5.9}$$

REINFORCE Leave-one-out (RLOO). To further reduce the variance of the REINFORCE estimator, RLOO samples K diffusion trajectories per prompt ($\{\mathbf{x}_{0:T}^k\} \sim \pi(. \mid \mathbf{c})$), for a better Monte Carlo estimate of the expectation [5, 82]. The RLOO estimator is:

$$\nabla_\theta J_\theta^{\mathrm{RLOO}}(\pi) = \mathbb{E}\left[K^{-1} \sum_{k=0}^{K} \sum_{t=0}^{T} \nabla_\theta \log p_\theta\left(\mathbf{x}_{t-1}^k \mid \mathbf{x}_t^k, \mathbf{c}\right)\left(r\left(\mathbf{x}_0^k, \mathbf{c}\right) - b_t\right)\right]. \tag{5.10}$$

However, REINFORCE-based estimators have a significant disadvantage: they do not allow sample reuse (i.e., reusing trajectories collected from previous policies) due to a distribution shift between policy gradient updates during training. Sampled trajectories can only be used once, prohibiting mini-batch updates. This makes it *sample inefficient*.

To allow for sample reuse, the importance sampling (IS) trick can be applied [113, 135]:

$$J_\theta^{\mathrm{IS}}(\pi) = \mathbb{E}_{c_t \sim p(C), a_t \sim \pi_{\mathrm{old}}(a_t \mid c_t)}\left[\frac{\pi_\theta(a_t \mid c_t)}{\pi_{\mathrm{old}}(a_t \mid c_t)} \mathcal{R}_t\right], \tag{5.11}$$

where π_θ is the *current* policy to be optimized, and π_{old} is the policy from the previous update round. With the IS trick, we can sample trajectories from the current policy in a batch, store it in a temporary buffer, and re-use them to apply mini-batch optimization [137].

Motivation for PPO. With the IS trick, the samples from the old policy can be used to estimate the policy gradient under the current policy π_θ (Eq. 5.8) in a statistically unbiased fashion [113], i.e., in expectation the IS and REINFORCE gradients are equivalent (Eq. 5.11, Eq. 5.8). Thus, potentially, we can improve the sample efficiency of REINFORCE gradient estimation with IS.

While unbiased, the IS estimator can exhibit high variance [113]. This high variance may lead to unstable training dynamics. Additionally, significant divergence between the current policy π_θ and the previous policy π_{old} can result in the updated diffusion policy performing worse than the previous one [1, 135]. Next, we will prove this formally. We note that this result has previously been established by [1] for the more general RL setting. In this chapter, we extend this finding to the context of diffusion model fine-tuning.

A key component of the proof relies on the distribution of states under the current policy, i.e., $d^\pi(s)$. In the case of diffusion models, the state transition kernel $P(s_{t+1} \mid s_t, a_t)$ is deterministic, because the next state consists of the action sampled from the previous state (Eq. 5.4), i.e., $P(s_{t+1} \mid s_t, a_t) = 1$. While the state transition kernel is

deterministic, the distribution of states is stochastic, given that it depends on the action at time t, which is sampled from the policy (Eq. 5.4). We define the state distribution as:

Definition 5.3.1. Given the distribution over contexts $\mathbf{c} \sim p(\mathbf{C})$, the (deterministic) distribution over time $t = \delta(t)$, and the diffusion policy π, the state distribution at time t is:

$$p(\mathbf{s}_t \mid \pi) = p(\mathbf{c})\delta(t)\int_{\mathbf{x}_{t+1}} \pi(\mathbf{x}_t \mid \mathbf{x}_{t+1}, \mathbf{c}, t)\pi(\mathbf{x}_{t+1} \mid \mathbf{c}, t)\, \mathbf{dx}_{t+1}.$$

Subsequently, the normalized discounted state visitation distribution can be defined as:

$$d^{\pi}(\mathbf{s}) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p(\mathbf{s}_t = \mathbf{s} \mid \pi). \tag{5.12}$$

The advantage function is defined as: $A^{\pi_k}(\mathbf{s}, \mathbf{a}) = Q^{\pi_k}(\mathbf{s}, \mathbf{a}) - V^{\pi_k}(\mathbf{s})$ [146]. Given this, the monotonic policy improvement bound can be derived:

Theorem 5.3.2. *Consider a current policy π_k. For any future policy π, we have:*

$$\begin{aligned} J(\pi) - J(\pi_k) \geq \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^{\pi_k}}\left[\frac{\pi(a \mid s)}{\pi_k(a \mid s)} A^{\pi_k}(s,a)\right] \\ - \frac{2\gamma C^{\pi,\pi_k}}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi_k}}\left[\mathrm{TV}(\pi(\cdot \mid s), \pi_k(\cdot \mid s))\right], \end{aligned}$$

where $C^{\pi,\pi_k} = \max_{s\in S} \left|\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi_k}(s,a)\right]\right|$ and $\mathrm{TV}(\pi(\cdot \mid s), \pi_k(\cdot \mid s))$ represents the total variation distance between the policies $\pi(\cdot \mid s)$ and $\pi_k(\cdot \mid s)$ [1].

A direct consequence of this theorem is that when optimizing a policy with the IS objective (Eq. 5.11), to guarantee that the new policy will improve upon the previous policy, the policies should not diverge too much. Therefore, we need to apply a constraint on the current policy. This can be achieved by applying the clipping operator in the PPO objective (Eq. 5.6) [1, 53, 117, 137].

This gives rise to an *efficiency-effectiveness trade-off* between REINFORCE and PPO. REINFORCE offers greater computational and implementation efficiency due to its simplicity, but it comes at the cost of lower sample efficiency and potential suboptimal performance. In contrast, PPO is more computationally demanding and involves more complex hyper-parameter tuning, yet it achieves higher performance and reliable policy improvements during training.

We note that a similar trade-off analysis was performed in the context of RL fine-tuning for large language models (LLM) [5]. However, their analysis was limited to an empirical study, whereas we present a theoretical analysis in addition to the empirical analysis. To the best of our knowledge, we are the first to conduct such a study for diffusion methods.

5.4 Method: Leave-One-Out PPO (LOOP) for Diffusion Fine-tuning

We demonstrated the importance of PPO in enhancing sample efficiency and achieving stable improvements during training for diffusion fine-tuning. Additionally, we showcased the RLOO method's effectiveness in reducing the variance of the REINFORCE method. In this section, we introduce our proposed method, **LOOP**, a novel RL for diffusion fine-tuning method. We start with highlighting the potential high-variance in the PPO objective.

The expectation in the PPO loss (Eq. 5.6) is typically estimated by sampling a single trajectory for a given prompt c:

$$\sum_{t=0}^{T} \text{clip}\left(\frac{\pi_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})}{\pi_{\text{old}}(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})}, 1-\epsilon, 1+\epsilon\right) r(\mathbf{x}_0, \mathbf{c}), \tag{5.13}$$

where $\mathbf{x}_{0:T} \sim \pi_{old}$. Even though the single sample estimate is an unbiased Monte-Carlo approximation of the expectation, it has high-variance [113]. Additionally, the IS term ($\frac{\pi_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{c})}{\pi_{\text{old}}(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{c})}$) can also contribute to high-variance of the PPO objective [147, 166]. Both factors combined, can lead to high-variance, and unstable training of the PPO.

Taking inspiration from RLOO (Eq. 5.10), we sample K independent trajectories from the previous policy for a given prompt c, and apply a baseline correction term from each trajectories reward, to reduce the variance of the estimator:

$$\hat{J}_\theta^{\text{LOOP}}(\pi) = \frac{1}{K}\sum_{i=1}^{K}\sum_{t=0}^{T} \text{clip}\left(\frac{\pi_\theta(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c)}{\pi_{\text{old}}(\mathbf{x}_{t-1}^i \mid \mathbf{x}_t^i, c)}, 1-\epsilon, 1+\epsilon\right) \cdot \left(r(\mathbf{x}_0^i, \mathbf{c}) - b^i\right), \tag{5.14}$$

where $\mathbf{x}_{0:T}^i \sim \pi_{old}, \forall i \in [1, K]$. The baseline correction term b^i reduces the variance of the gradient estimate, while being unbiased in expectation [52, 101]. A simple choice of baseline correction can be the average reward across the K trajectories, i.e.,

$$b^i = \frac{1}{k}\sum_{i=0}^{K} r(\mathbf{x}_0^i). \tag{5.15}$$

However, we choose the leave-one-out average baseline, with average taken across all samples in the trajectory, except the current sample i, i.e.,

$$b^i = \frac{1}{k-1}\sum_{j \neq i} r(\mathbf{x}_0^j). \tag{5.16}$$

Originally RLOO sampling and baseline corrections were proposed in the context of REINFORCE, with a focus on on-policy optimization [5, 82], whereas we are applying these in the off-policy step of PPO. We call this method *leave-one-out PPO* (LOOP). Provenly, LOOP has lower variance than PPO:

Proposition 5.4.1. *The LOOP estimator $\hat{J}_\theta^{\text{LOOP}}(\pi)$ (Eq. 5.14) has lower variance than the PPO estimator $\hat{J}_\theta^{\text{PPO}}(\pi)$ (Eq. 5.13):*

$$\text{Var}\left[\hat{J}_\theta^{\text{LOOP}}(\pi)\right] < \text{Var}\left[\hat{J}_\theta^{\text{PPO}}(\pi)\right]. \tag{5.17}$$

Proof. Since the sampled trajectories are independent:

$$\mathrm{Var}\Big[\hat{J}_\theta^{\mathrm{LOOP}}(\pi)\Big] = \frac{1}{K^2}\mathrm{Var}\Big[\hat{J}_\theta^{\mathrm{PPO}}(\pi)\Big] < \mathrm{Var}\Big[\hat{J}_\theta^{\mathrm{PPO}}(\pi)\Big]. \ \square$$

5.5 Experimental Setup

Benchmark. Text-to-image diffusion and language models often fail to satisfy an essential reasoning skill of attribute binding. Attribute binding reasoning capability refers to the ability of a model to generate images with attributes such as color, shape, texture, spatial alignment, (and others) specified in the input prompt. In other words, generated images often fail to *bind* certain *attributes* specified in the instruction prompt [41, 60, 121]. Since attribute binding seems to be a basic requirement for useful real-world applications, we choose the T2I-CompBench benchmark [60], which contains multiple attribute binding/image compositionality tasks, and its corresponding reward metric to benchmark text-to-image generative models. We also select two common tasks from RL for diffusion works: improving aesthetic quality of generation, and image-text semantic alignment [13, 37]. To summarize, we choose the following tasks for the RL optimization: (i) Color, (ii) Shape, (iii) Texture, (iv) 2D Spatial, (v) Numeracy, (vi) Aesthetic, and (vii) Image-text Alignment. For all tasks, the prompts are split into training/validation prompts. We report the average reward on both training and validation split.

Model. As the base diffusion model, we use Stable diffusion V2 [125], which is a latent diffusion model. For optimization, we fully update the UNet model, with a learning rate of $1e^{-5}$. We also tried LORA fine-tuning [59], but the results were not satisfactory, so we update the entire model instead. The hyper-parameters are reported in Appendix 5.A.

5.6 Results and Discussion

5.6.1 REINFORCE vs. PPO efficiency-effectiveness trade-off

We discuss our empirical results for the REINFORCE vs. PPO efficiency-effectiveness trade-off. Our empirical validation of the trade-off compares the following methods:

REINFORCE. The REINFORCE policy gradient for diffusion fine-tuning (Eq. 5.8).
REINFORCE with baseline correction. We compare the REINFORCE policy gradient with a baseline correction (BC) term (Eq. 5.9). For the baseline term, we choose the average reward for the given prompt [13].
PPO. The PPO objective for diffusion fine-tuning with importance sampling and clipping (Eq. 5.6).

Figure 5.2 shows the training reward over epochs for the attributes: Color, Shape, and Texture from the T2I-CompBench benchmark, and training reward from optimizing the aesthetic model. It is clear that REINFORCE policy gradient is not effective in terms of performance, as compared to other variants. Adding a baseline correction term

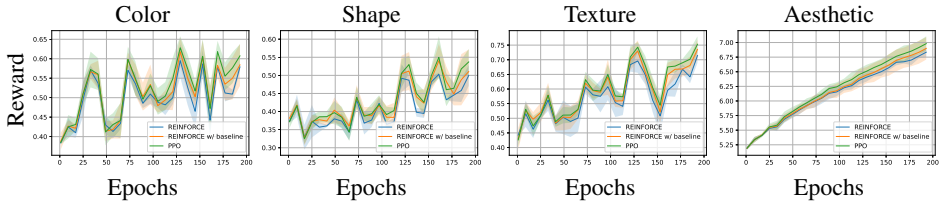

Figure 5.2: Evaluating REINFORCE vs. PPO trade-off by comparing: REINFORCE (Eq. 5.8), REINFORCE with baseline correction term (Eq. 5.9), and PPO (Eq. 5.6). We evaluate on the T2I-CompBench benchmark over three image attributes: Color, Shape, and Texture. We also compare on the aesthetic task. Y-axis corresponds to the training reward, x-axis corresponds to the training epoch. Results are averaged over 3 runs; shaded areas indicate 80% prediction intervals.

Table 5.1: Comparing REINFORCE with PPO on the T2I-CompBench benchmark over three image attributes: Color, Shape, and Texture. The metrics in this table are average reward on an unseen test set (higher is better). For each prompt, average rewards over 10 independent generated images are calculated.

Method	**Color** ↑	**Shape** ↑	**Texture** ↑
REINFORCE	0.6438	0.5330	0.6359
REINFORCE w/ BC	0.6351	0.5347	0.6656
PPO	**0.6821**	**0.5655**	**0.6909**

indeed improves the training performance, validating the effectiveness of baseline in terms of training performance, possibly because of reduced variance. PPO achieves the highest training reward, validating the effectiveness of importance sampling and clipping for diffusion fine-tuning.

We also evaluate the performance on a separate validation set. For each validation prompt, we generate 10 independent images from the diffusion policy, and average the reward, finally averaging over all evaluation prompts. The validation results are reported in Table 5.1. The results are consistent with the pattern observed with the training rewards, i.e., REINFORCE with baseline provides a better performance than plain REINFORCE, suggesting that baseline correction indeed helps with the final performance. Nevertheless, PPO still performs better than REINFORCE.

We now have empirical evidence supporting the *efficiency-effectiveness trade-off* discussed in Section 5.3. From these results, we can conclude that fine-tuning text-to-image diffusion models is more effective with IS and clipping from PPO, or baseline corrections from REINFORCE. This bolsters our motivation for proposing LOOP as an approach to effectively combine these methods.

Table 5.2: Comparing the performance of the proposed LOOP method with state-of-the-art baselines on the T2I-CompBench benchmark over image attributes such as Color, Shape, Texture, Spatial relation, and Numeracy. The metrics in this table are average reward on an unseen test set (higher is better). For each prompt we generate and average rewards across 10 different generated images.

Model	**Color** ↑	**Shape** ↑	**Texture** ↑	**Spatial** ↑	**Numeracy** ↑
Stable v1.4 [125]	0.3765	0.3576	0.4156	0.1246	0.4461
Stable v2 [125]	0.5065	0.4221	0.4922	0.1342	0.4579
Composable v2 [90]	0.4063	0.3299	0.3645	0.0800	0.4261
Structured v2 [39]	0.4990	0.4218	0.4900	0.1386	0.4550
Attn-Exct v2 [23]	0.6400	0.4517	0.5963	0.1455	0.4767
GORS unbiased [60]	0.6414	0.4546	0.6025	0.1725	–
GORS [60]	0.6603	0.4785	0.6287	0.1815	0.4841
PPO [13]	0.6821	0.5655	0.6909	0.1961	0.5102
LOOP ($k = 2$)	0.6785	0.5746	0.6937	0.1800	0.5072
LOOP ($k = 3$)	0.7515	0.6220	0.7353	0.1966	0.5242
LOOP ($k = 4$)	**0.7859**	**0.6676**	**0.7518**	**0.2136**	**0.5422**

Table 5.3: Comparing the performance of LOOP with PPO on the aesthetic and image-text alignment tasks. Higher values are better.

Method	**Aesthetic** ↑	**Image Alignment** ↑
PPO [13]	6.8135	20.466
LOOP ($k = 2$)	6.8617	20.788
LOOP ($k = 3$)	7.0772	20.619
LOOP ($k = 4$)	**7.8606**	**20.909**

5.6.2 Evaluating LOOP

Next we discuss the results from our proposed RL for diffusion fine-tuning method, LOOP.

Performance during training. Figure 5.3 shows the training reward curves for different tasks, against number of epochs. LOOP outperforms PPO across all seven tasks consistently throughout training. This establishes the effectiveness of sampling multiple diffusion trajectories per input prompt, and the leave-one-out baseline correction term (Eq. 5.10) during training. The training reward curve is smoother for the aesthetic task, as compared to tasks from the T2I-CompBench benchmark. We hypothesise that improving the attribute binding property of the diffusion model is a harder task than improving the aesthetic quality of generated images.

Table 5.2 reports the validation rewards across different tasks from the T2I-CompBench benchmark. LOOP outperforms PPO and other strong supervised learning based baseline significantly across all tasks. It shows that PPO improves the attribute-binding

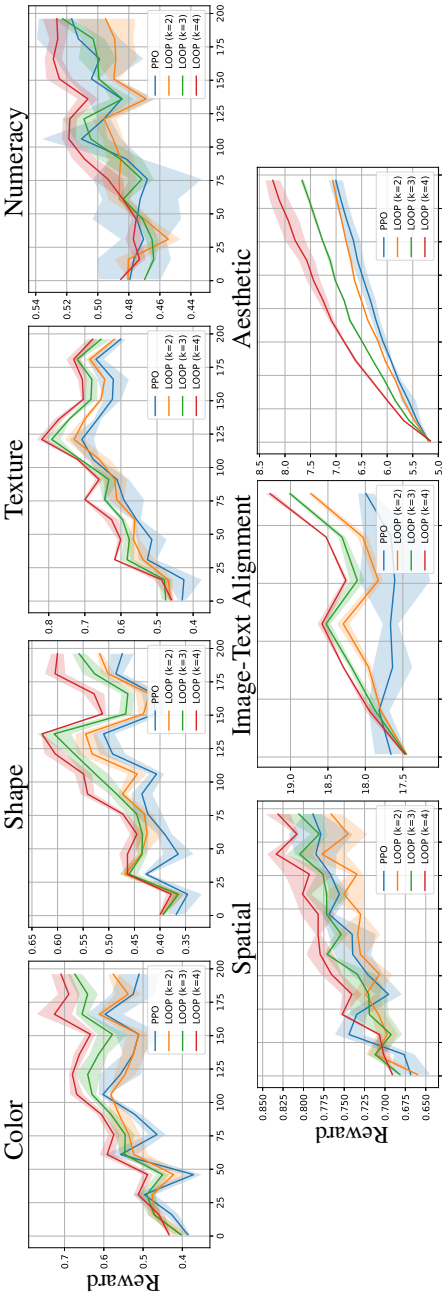

Figure 5.3: Comparing PPO with proposed LOOP on the T2I-CompBench benchmark with respect to image attributes: Color, Shape, Texture, Numeracy, and Spatial relationship. We also compare against aesthetic and image–text alignment tasks [13]. The y-axis is the training reward; the x-axis is the training epoch. Results are averaged over three independent runs; shaded areas denote 80 % prediction intervals.

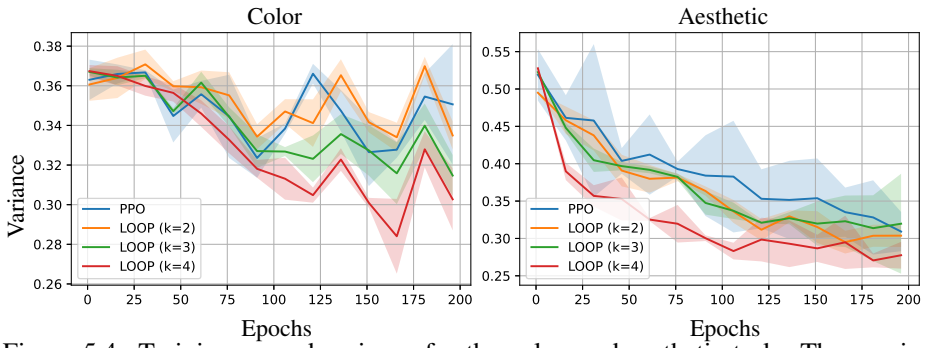

Figure 5.4: Training reward variance for the color, and aesthetic task. The y-axis corresponds to the training reward variance; the x-axis indicates the number of training epochs. Results are averaged over 3 runs; shaded areas indicate 80% prediction intervals. We observe that higher values of samples reused (i.e., k) produce lower reward variance during training.

reasoning ability of the diffusion model compared to other supervised learning based methods.

For the aesthetic and image-text alignment objectives, the validation rewards are reported in Table 5.3. LOOP results in a **15.37**% relative improvement over PPO for the aesthetic task, and a **2.16**% improvement over PPO for the image-text alignment task.

Impact of the number of independent trajectories (k). The LOOP variant with the number of independent trajectories where K set to 4 performs the best across all tasks, followed by the variant $K = 3$. This is intuitive given that Monte-Carlo estimates get better with more number of samples [113]. Surprisingly, the performance of the variant with $K = 2$ is comparable to PPO.

Impact on training variance. We evaluate whether LOOP results in a lower empirical variance than PPO, as proved theoretically in Lemma 5.4.1. Figure 5.4 reports the empirical reward variance during training for the color attribute and aesthetic objective. LOOP results in a lower empirical variance than PPO, thereby empirically validating our claim that LOOP has lower variance than PPO.

Qualitative results. For a qualitative evaluation of the attribute-binding reasoning ability, we present some example image generations from SD, PPO, and LOOP in Figure 5.1. In the first example, the input prompt specifies a black colored ball with a white cat. Stable diffusion and PPO fail to bind the color black with the generated ball, whereas LOOP successfully binds that attribute. Similarly, in the third example, SD and PPO fail to bind the hexagon shape attribute to the watermelon, whereas LOOP manages to do that. In the fourth example, SD and PPO fail to add the horse object itself, whereas LOOP adds the horse with the specified black color, and flowing cyan patterns.

5.7 Conclusion

We have studied the **efficiency-effectiveness** trade-off between two fundamental RL for diffusion methods: REINFORCE, and PPO. REINFORCE, while computationally

efficient and easier to implement, is subpar to PPO in terms of sample efficiency and performance. Building on these insights, we have introduced a simple and effective RL for diffusion method, LOOP, which builds on the variance reduction techniques from REINFORCE and the effectiveness and robustness of PPO. We have found that LOOP improves over diffusion models on multiple black-box objectives. A limitation of LOOP is that sampling multiple diffusion trajectories per prompt can lead to more computational overhead and an increase in training time. A potential future direction would be to keep the effectiveness of LOOP while maintaining the computational complexity of PPO.

In this chapter, we answer the broad research question (RQ5) in the affirmative. We systematically compare PPO and REINFORCE for diffusion model fine-tuning, where we demonstrate that REINFORCE exhibits inferior sample efficiency compared to PPO. Building on top of PPO, we propose LOOP, which achieves superior performance with the same number of input prompts by generating multiple actions per prompt.

Appendices

5.A Hyperparameter and Implementation Details

For REINFORCE (including REINFORCE with baseline correction term), PPO, and LOOP the number of denoising steps (T) is set to 50. The diffusion guidance weight is set to 5.0. For optimization, we use AdamW [95] with a learning rate of $1e^{-5}$, and the weight decay of $1e{-}4$, with other parameters kept at the default value. We clip the gradient norm to 1.0. We train all models using 8 A100 GPUs with a batch size of 4 per GPU. The clipping parameter ϵ for PPO, and LOOP is set to $1e^{-4}$.

5.B Additional Qualitative Examples

We present some additional qualitative examples in this section.

Figure 5.B.1: **LOOP improves aesthetic quality**. Qualitative examples are presented from images generated via Stable Diffusion 2.0 (first row), PPO (second row), and LOOP $k = 4$ (third row). LOOP consistently generates more aesthetic images than PPO and SD.

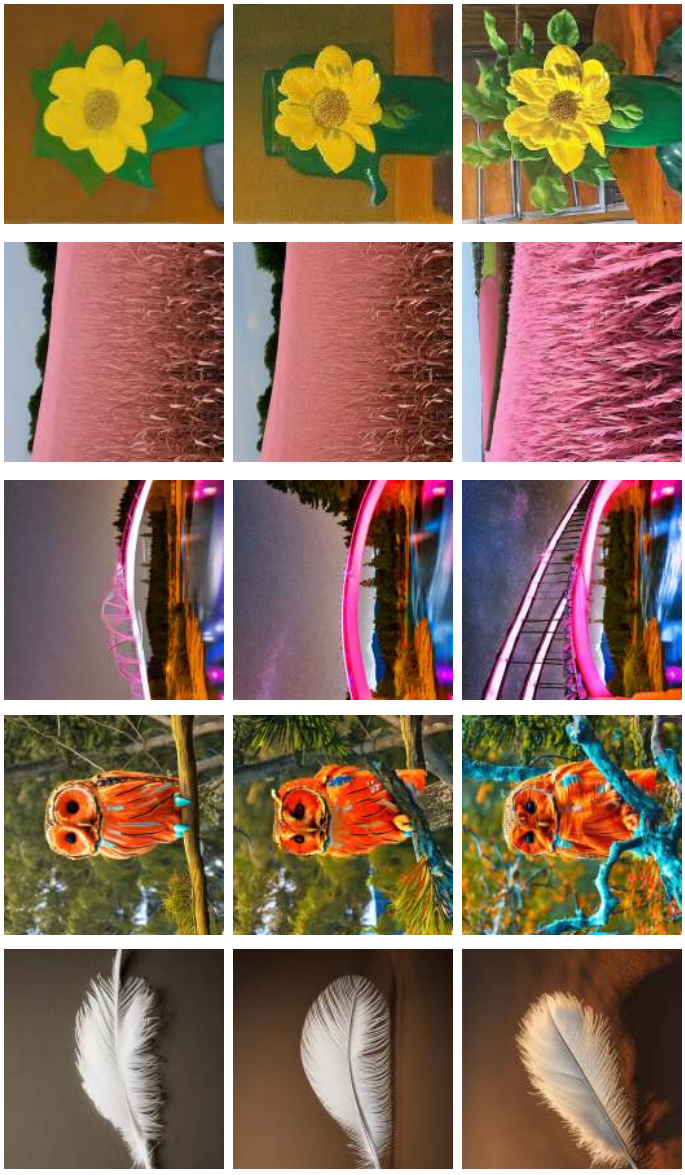

Figure 5.B.2: Additional qualitative examples presented from images generated via Stable Diffusion 2.0 (first row), PPO (second row), and LOOP $k = 4$ (third row). LOOP consistently generates more aesthetic images than PPO and SD (first, third, and fifth prompt). LOOP also binds the color attribute (teal branch in the second example and pink cornfield in the fourth) where SD and PPO fail.

6 Conclusions

In this thesis, we have investigated approaches to develop safe, robust and efficient reinforcement learning methods for real-world applications. Specifically, in the four chapters preceding this conclusion, we have demonstrated:

(i) A safe counterfactual LTR method which guarantees that the new ranking policy will be at least as good as the production/logging policy, presented in Chapter 2.

(ii) A robust safe counterfactual learning to rank (LTR) method where the safety guarantees are agnostic to the user behavior model, and hold even under adversarial click behavior settings, presented in Chapter 3.

(iii) A closed-form baseline correction method for off-policy evaluation and learning for contextual bandits with guaranteed minimum variance, presented in Chapter 4.

(iv) An efficient reinforcement learning method for text-to-image diffusion fine-tuning, based on a simple and practical extension of the popular Proximal Policy Optimization (PPO) algorithm, with significantly improved performance, presented in Chapter 5.

6.1 Main Findings

In this section we revisit the research questions presented in Chapter 1 followed by a summary of the most important findings.

RQ1 Can safety guarantees be provided for counterfactual LTR policies to ensure that the new policy is at least as good as the production policy?

The answer to this question is in the affirmative. In Chapter 2, we derive a generalization bound for the counterfactual LTR estimator, establishing a lower bound on the true ranking utility, the ideal target metric for optimization. We demonstrated that optimizing this lower bound ensures a ranking policy no worse than the current production policy. This property proves especially valuable when click data is scarce, mitigating the risk of deploying potentially harmful policies, thereby ensuring safe deployment.

The broader implication of this work is in the practical *deployment* stage for all modern search and recommender ranking systems. Using the presented safety techniques, search and recommendation teams can reliably deploy a ranking policy without risking deploying a policy with sub-optimal user experience. An example use-case is deploying a ranking policy in a new geographic region, with limited user interactions. The safety method presented will help ensure that the new policy is never worse than the safe production policy.

A limitation of our work is that we assume the ranking policy to be stochastic in nature, i.e., for a given query context, the ranking policy generates different ranked lists at each time. In certain real-world applications, stochastic rankings might be unfeasible, or not preferred because of external reasons.

While RQ1 provides a probabilistic safety guarantee by optimizing the lower bound, these guarantees depend critically on assumptions regarding user behavior (click model). Deviations from these assumptions invalidate the guarantees, which motivated the second research question:

RQ2 Can we provide robust safety guarantees for counterfactual LTR policies even under adversarial user behavior settings?

The answer to this question is in the affirmative; in Chapter 3, we introduced proximal ranking policy optimization (PRPO), a method ensuring safety for counterfactual LTR without reliance on user behavior assumptions, guaranteeing robust safety even under adversarial conditions.

The broader implication of this work is in providing a robust safe deployment framework. In practice, the proposed method in this chapter provides reliability in the wild. Search engines or recommender systems that serve heterogeneous markets (or fast-shifting verticals like news) have shifting user preferences, and relying on a single user behavior assumption can have detrimental effects. The robust safety method presented in this chapter can ensure safe deployment even under such dynamic heterogeneous markets.

Similar to the previous chapter, a limitation of this work is that we assume the ranking policy to be stochastic in nature, i.e., for a given query context, the ranking policy generates different ranked lists at each time. A robust safety method for deterministic ranking policy might be preferred.

In the context of off-policy evaluation and learning with single action contextual bandits, standard methods like IPS are unbiased but suffer from high variance. Alternative methods, including doubly robust (DR) estimators and self-normalized IPS (SNIPS), reduce variance using additive and multiplicative baseline corrections respectively, yet lack a unifying framework. This motivated our third research question:

RQ3 Can we unify variance reduction techniques using baseline corrections and a doubly robust estimator under a common framework?

The answer to this question is in the affirmative; in Chapter 4, we proposed the β-IPS estimator, integrating IPS, doubly robust methods, and Self-Normalized IPS under a unified baseline correction framework.

RQ4 Given a unified framework for variance reduction techniques under baseline corrections, can we derive a variance-optimal unbiased estimator?

The answer to this question is in the affirmative; in Chapter 4, we presented a closed-form solution for β that minimizes variance for both learning and evaluation tasks. Empirical evidence under different scenarios validates the effectiveness of our approach.

A broader implication of this work is in providing a *unified vocabulary/framework* for off-policy evaluation and learning tasks. A common framework for off-policy evaluation and learning tasks reduces the burden of choosing an estimator in practice for practitioners. Further, a closed-form solution for the baseline correction term makes the practical implementation of the estimator easier.

Contextual bandit theory previously discussed emphasizes user interactions within ranking or recommendation systems. However, the framework has also been effectively employed in fine-tuning foundation models, such as large language models (LLMs) and diffusion models, typically using proximal policy optimization (PPO). Recent research highlights computational advantages of REINFORCE (policy gradient methods) over PPO for LLMs [5]. Given PPO's ongoing challenges with variance and sample inefficiency, we consider improvements through our final research question:

RQ5 Can we improve the sample efficiency of proximal policy optimization for fine-tuning text-to-image diffusion?

The answer to this question is in the affirmative; in Chapter 5, we systematically compare PPO and REINFORCE for diffusion model fine-tuning. Initially, we demonstrate that REINFORCE exhibits inferior sample efficiency compared to PPO. Subsequently, we propose leave-one-out PPO (LOOP), an enhancement to PPO achieving superior performance with the same number of input prompts by generating multiple actions per prompt.

6.2 Future Work

Finally, this section addresses some limitations with the existing work and potential future directions of the research presented in this thesis.

6.2.1 Safety with real-world constraints

First, the safe counterfactual LTR methods presented in Chapter 2 and Chapter 3 are designed with a stochastic ranking policy in mind. In many real-world applications, deploying a stochastic policy might not be feasible, necessitating safety methods for deterministic ranking policies [45]. A future direction along this line would be to add safety regularization to the top-K LambdaLoss LTR method with deterministic ranking policy [108].

Regarding experiments, all of our evaluations are based on semi-synthetic simulations with click signal derived from the manual relevance judgments. Real-world experiments are typically conducted via A/B tests on actual users. As part of future

work, applying the proposed safety methods to real-world user interaction data, followed by comprehensive A/B testing, would be particularly valuable.

Modern recommendation systems increasingly include a LLM that both selects content and generates natural-language explanations. Extending safe counterfactual LTR to such LLM-based ranking policies will require handling high-dimensional textual actions, possibly by constraining the language model output with exposure-based bounds handing the high-dimensional nature of action space.

6.2.2 Extending optimal baseline corrections to reinforcement learning

Next, the optimal-baseline correction method presented in Chapter 4 reduces the variance of existing off-policy evaluation and learning estimators for contextual bandits. Extending the proposed optimal variance baseline to offline RL scenarios would be an interesting future research direction [86].

In moving from contextual bandits to full reinforcement learning settings, where decisions involve trajectories (sequences of actions), it would be interesting to derive an optimal scalar (or state-dependent) baseline that minimises variance while preserving unbiasedness.

Further, the optimal baseline correction presented for off-policy learning involves calculating and storing gradients of all parameters for each example separately. With large language models involving billions of parameters, storing and calculating gradients separately for each example could be practically challenging, presenting another potential avenue for future exploration. Compression techniques such as low-rank adapters, or selective checkpointing could bring the memory cost, opening up an interesting practical future direction.

6.2.3 RL-based diffusion fine-tuning

The diffusion fine-tuning setup presented in Chapter 5 is treated as a contextual bandit framework, with the entire reverse diffusion process treated as a single action. A scalar reward is generated for the final image in the generation. Extending the model to a more traditional reinforcement learning setup, where we have a scalar reward at each step of the reverse diffusion process, could be an interesting future direction.

Further, the reward signal for diffusion fine-tuning comes from an external reward model, which roughly represents the average population score for the corresponding task, and is not personalized. An interesting future direction would be to fine-tune foundational models directly with user interaction data to enable personalized generative models. An example would be a personalized email writing assistant, which learns to generate tailored email text based on user interactions, such as edits, binary feedback, etc. In the context of image generation, an example use-case could be personalized music playlist cover generation, based on the user's interactions and preferences.

6.2.4 Personalised generative models

Finally, the reward model used for diffusion fine-tuning in Chapter 5 reflects an average user preference. Personalising generation to individual users raises new challenges: privacy, fairness and extreme data sparsity. Promising future research directions in this area are:

- *Privacy-preserving on-device fine-tuning*. Employ federated RL or secure aggregation to learn user-specific adapters without sharing raw images or prompts.
- *Meta-learning reward models*. Train a global model that can be rapidly adapted to a new user with a handful of interaction signals (edits, binary feedback).
- *Fairness and calibration*. Ensure that personalised models do not amplify sensitive-attribute biases by incorporating fairness constraints into the safe-bandit objective.

Bibliography

[1] J. Achiam, D. Held, A. Tamar, and P. Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017. (Cited on pages 83, 85, and 86.)

[2] M. M. Afsar, T. Crump, and B. Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022. (Cited on page 1.)

[3] A. Agarwal, K. Takatsu, I. Zaitsev, and T. Joachims. A general framework for counterfactual learning-to-rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5–14, 2019. (Cited on page 37.)

[4] A. Agarwal, X. Wang, C. Li, M. Bendersky, and M. Najork. Addressing trust bias for unbiased learning-to-rank. In *The World Wide Web Conference*, pages 4–14, 2019. (Cited on pages 2, 13, 33, 34, 35, 37, and 45.)

[5] A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting REINFORCE style optimization for learning from human feedback in LLMs. *arXiv preprint arXiv:2402.14740*, 2024. (Cited on pages 4, 79, 85, 86, 87, and 101.)

[6] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021. (Cited on page 79.)

[7] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022. (Cited on page 1.)

[8] H. C. Bakker, S. Gupta, and H. Oosterhuis. A simpler alternative to variational regularized counterfactual risk minimization. In *CONSEQUENCES Workshop at ACM RecSys '24*, 2024.

[9] A. Barraza-Urbina and D. Glowacka. Introduction to bandits in recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 748–750, 2020. (Cited on page 1.)

[10] S. Baweja, N. Pokharna, A. Ustimenko, and O. Jeunen. Variance reduction in ratio metrics for efficient online experiments. In *Proc. of the 46th European Conference on Information Retrieval*, ECIR '24. Springer, 2024. (Cited on page 62.)

[11] W. Bendada, G. Salha, and T. Bontempelli. Carousel personalization in music streaming apps with contextual bandits. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 420–425. ACM, 2020. doi: 10.1145/3383313.3412217. (Cited on page 60.)

[12] J. Bennett, S. Lanning, et al. The Netflix Prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007. (Cited on page 59.)

[13] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. (Cited on pages 2, 80, 82, 84, 88, 90, and 91.)

[14] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. (Cited on page 79.)

[15] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018. (Cited on page 73.)

[16] L. Briand, T. Bontempelli, W. Bendada, M. Morlon, F. Rigaud, B. Chapus, T. Bouabça, and G. Salha-Galvan. Let's get it started: Fostering the discoverability of new releases on Deezer. In *Proc. of the 46th European Conference on Information Retrieval*, ECIR '24. Springer, 2024. (Cited on page 60.)

[17] R. Budylin, A. Drutsa, I. Katsev, and V. Tsoy. Consistent transformation of ratio metrics for efficient online controlled experiments. In *Proc. of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 55–63. ACM, 2018. doi: 10.1145/3159652.3159699. (Cited on pages 60 and 62.)

[18] C. Burges, R. Ragno, and Q. Le. Learning to rank with nonsmooth cost functions. *Advances in Neural Information Processing Systems*, 19, 2006. (Cited on page 44.)

[19] C. J. Burges. From RankNet to LambdaRank to LambdaMART. *Learning*, 11(23-581):81, 2010. (Cited on page 44.)

[20] C. Castillo and B. D. Davison. Adversarial web search. *Foundations and Trends in Information Retrieval*, 4(5):377–486, 2011. (Cited on page 41.)

[21] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning*

to Rank Challenge, pages 1–24. PMLR, 2011. (Cited on pages 11, 13, 25, 33, and 44.)

[22] O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, pages 1–10, 2009. (Cited on page 13.)

[23] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10, 2023. (Cited on page 90.)

[24] J. Chen, J. Mao, Y. Liu, M. Zhang, and S. Ma. Tiangong-st: A new dataset with large-scale refined real-world web search sessions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2485–2488, 2019. (Cited on page 25.)

[25] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi. Top-k off-policy correction for a reinforce recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 456–464, 2019. (Cited on pages 60, 64, and 71.)

[26] M. Chen, B. Chang, C. Xu, and E. H. Chi. User response models to improve a reinforce recommender system. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, WSDM '21, page 121–129, New York, NY, USA, 2021. ACM. doi: 10.1145/3437963.3441764.

[27] M. Chen, C. Xu, V. Gatto, D. Jain, A. Kumar, and E. Chi. Off-policy actor-critic for recommender systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 338–349. ACM, 2022. doi: 10.1145/3523227.3546758. (Cited on pages 60, 64, and 71.)

[28] A. Chuklin, I. Markov, and M. de Rijke. *Click Models for Web Search*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2015. (Cited on pages 15, 37, and 44.)

[29] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pages 442–450, 2010. (Cited on page 14.)

[30] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 87–94, 2008. (Cited on pages 12, 13, 15, 18, 33, 34, and 37.)

[31] D. Dato, C. Lucchese, F. M. Nardini, S. Orlando, R. Perego, N. Tonellotto, and R. Venturini. Fast ranking with additive ensembles of oblivious and non-oblivious regression trees. *ACM Transactions on Information Systems (TOIS)*, 35(2):1–31, 2016. (Cited on pages 25 and 44.)

[32] P. Dayan. Reinforcement comparison. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, and G. E. Hinton, editors, *Connectionist Models*, pages 45–51. Morgan Kaufmann, 1991. doi: https://doi.org/10.1016/B978-1-4832-1448-1.50011-1. (Cited on page 62.)

[33] A. Deng, Y. Xu, R. Kohavi, and T. Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proc. of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, page 123–132. ACM, 2013. doi: 10.1145/2433396.2433413. (Cited on page 62.)

[34] Z. Dong, H. Zhu, P. Cheng, X. Feng, G. Cai, X. He, J. Xu, and J. Wen. Counterfactual learning for recommender system. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 568–569. ACM, 2020. doi: 10.1145/3383313.3411552. (Cited on page 60.)

[35] M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485 – 511, 2014. (Cited on pages 60, 61, 64, 66, and 73.)

[36] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2019. (Cited on page 80.)

[37] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on pages 80, 84, and 88.)

[38] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456. PMLR, 10–15 Jul 2018. (Cited on page 66.)

[39] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. (Cited on page 90.)

[40] D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008. ISSN 0196-8858. doi: https://doi.org/10.1016/j.aam.2006.12.003. (Cited on

pages 60 and 62.)
[41] H. Fu and G. Cheng. Enhancing semantic mapping in text-to-image diffusion via gather-and-bind. *Computers & Graphics*, 125:104118, 2024. (Cited on pages 82 and 88.)
[42] B. K. Ghosh. Probability inequalities related to Markov's theorem. *The American Statistician*, 56(3): 186–190, 2002. (Cited on pages 22 and 54.)
[43] E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004. (Cited on pages 60, 62, 63, 68, and 85.)
[44] Y. Gu, Z. Wang, Y. Yin, Y. Xie, and M. Zhou. Diffusion-RPO: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024. (Cited on page 80.)
[45] R. Guo, J.-F. Ton, Y. Liu, and H. Li. Inference-time stochastic ranking with risk control. *arXiv preprint arXiv:2306.07188*, 2023. (Cited on page 101.)
[46] S. Gupta, P. Hager, J. Huang, A. Vardasbi, and H. Oosterhuis. Recent advances in the foundations and applications of unbiased learning to rank. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3440–3443, 2023.
[47] S. Gupta, P. K. Hager, and H. Oosterhuis. Recent advancements in unbiased learning to rank. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 145–148, 2023.
[48] S. Gupta, H. Oosterhuis, and M. de Rijke. A first look at selection bias in preference elicitation for recommendation (abstract). In *CONSEQUENCES Workshop at RecSys '23*. ACM, September 2023. (Cited on pages 35 and 60.)
[49] S. Gupta, H. Oosterhuis, and M. de Rijke. A deep generative recommendation method for unbiased learning from implicit feedback. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 87–93, 2023. (Cited on page 35.)
[50] S. Gupta, H. Oosterhuis, and M. de Rijke. Safe deployment for counterfactual learning to rank with exposure-based risk minimization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 249–258, 2023. (Cited on pages 11, 20, 22, 34, 36, 37, 39, 40, 41, 43, 44, 45, 46, 54, 59, and 60.)
[51] S. Gupta, P. Hager, J. Huang, A. Vardasbi, and H. Oosterhuis. Unbiased learning to rank: On recent advances and practical applications. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1118–1121, 2024. (Cited on pages 2, 3, 11, 33, 34, 36, 37, 39, and 60.)
[52] S. Gupta, O. Jeunen, H. Oosterhuis, and M. de Rijke. Optimal baseline corrections for off-policy contextual bandits. In *RecSys 2024: 18th ACM Conference on Recommender Systems*, pages 722–732. ACM, October 2024. (Cited on pages 35, 59, and 87.)
[53] S. Gupta, H. Oosterhuis, and M. de Rijke. Practical and robust safety guarantees for advanced counterfactual learning to rank. In *CIKM 2024: 33rd ACM International Conference on Information and Knowledge Management*, pages 737–747. ACM, October 2024. (Cited on pages 33, 41, 53, 60, and 86.)
[54] S. Gupta, C. Ahuja, T.-Y. Lin, S. D. Roy, H. Oosterhuis, M. de Rijke, and S. N. Shukla. A simple and effective reinforcement learning method for text-to-image diffusion fine-tuning. *arXiv preprint arXiv:2503.00897*, 2025. (Cited on page 79.)
[55] S. Gupta, Y. Liao, and M. de Rijke. Towards two staged counterfactual learning to rank. In *Proceedings of the 2025 ACM SIGIR on International Conference on Innovative Concepts and Theories in Information Retrieval*, 2025.
[56] L. He, L. Xia, W. Zeng, Z.-M. Ma, Y. Zhao, and D. Yin. Off-policy learning for multiple loggers. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1184–1193, 2019. (Cited on page 14.)
[57] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. (Cited on pages 2, 79, and 82.)
[58] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. (Cited on pages 12 and 14.)
[59] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. (Cited on page 88.)
[60] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. (Cited on pages 82, 88, and 90.)
[61] S. Huang, M. Noukhovitch, A. Hosseini, K. Rasul, W. Wang, and L. Tunstall. The N+ implementation

details of RLHF with PPO: A case study on TL;DR summarization. *arXiv preprint arXiv:2403.17031*, 2024. (Cited on page 80.)

[62] E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17 (2):295–311, 2008. doi: 10.1198/106186008X320456. (Cited on pages 60 and 64.)

[63] R. Jagerman, I. Markov, and M. de Rijke. Safe exploration for optimizing contextual bandits. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–23, 2020. (Cited on pages 2, 12, 13, 34, 35, and 39.)

[64] F. James. Monte Carlo theory and practice. *Reports on Progress in Physics*, 43(9):1145, 1980. (Cited on page 25.)

[65] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. (Cited on pages 15, 26, 37, and 46.)

[66] O. Jeunen. *Offline Approaches to Recommendation with Online Success*. PhD thesis, University of Antwerp, 2021. (Cited on page 60.)

[67] O. Jeunen and B. Goethals. An empirical evaluation of doubly robust learning for recommendation. In *Proc. of the ACM RecSys Workshop on Bandit Learning from User Interactions*, REVEAL '20, 2020. (Cited on pages 66, 72, and 75.)

[68] O. Jeunen and B. Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 63–74. ACM, 2021. ISBN 9781450384582. doi: 10.1145/3460231.3474247. (Cited on pages 60 and 71.)

[69] O. Jeunen and B. Goethals. Top-k contextual bandits with equity of exposure. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 310–320. ACM, 2021. ISBN 9781450384582. doi: 10.1145/3460231.3474248. (Cited on pages 60 and 71.)

[70] O. Jeunen and B. Goethals. Pessimistic decision-making for recommender systems. *ACM Trans. Recomm. Syst.*, 1(1), feb 2023. doi: 10.1145/3568029. (Cited on page 60.)

[71] O. Jeunen, D. Rohde, F. Vasile, and M. Bompaire. Joint policy-value learning for recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 1223–1233. ACM, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3403175. (Cited on pages 60 and 72.)

[72] O. Jeunen, T. Joachims, H. Oosterhuis, Y. Saito, and F. Vasile. Consequences—causality, counterfactuals and sequential decision-making for recommender systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 654–657, 2022. (Cited on page 60.)

[73] O. Jeunen, S. Murphy, and B. Allison. Off-policy learning-to-bid with AuctionGym. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4219–4228. ACM, 2023. (Cited on pages 60 and 71.)

[74] O. Jeunen, I. Potapov, and A. Ustimenko. On (normalised) discounted cumulative gain as an offline evaluation metric for top-n recommendation. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2024. (Cited on page 60.)

[75] S. Jiang, Y. Hu, C. Kang, T. Daly Jr, D. Yin, Y. Chang, and C. Zhai. Learning query and document relevance from a web-scale click graph. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 185–194, 2016. (Cited on page 13.)

[76] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002. (Cited on pages 11, 13, and 37.)

[77] T. Joachims and A. Swaminathan. Counterfactual evaluation and learning for search, recommendation and ad placement. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1199–1201, 2016. (Cited on pages 12, 13, 14, 35, and 61.)

[78] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017. (Cited on pages 1, 2, 12, 13, 15, 16, 25, 26, 33, 34, 35, 36, 37, 39, and 44.)

[79] T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018. (Cited on pages 3, 59, 60, 65, 67, 68, and 72.)

[80] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. (Cited on page 73.)

[81] A. Kong. A note on importance sampling using standardized weights. Technical Report 348, University of Chicago, Dept. of Statistics, 1992. (Cited on pages 60 and 64.)

[82] W. Kool, H. van Hoof, and M. Welling. Buy 4 REINFORCE samples, get a baseline for free! *ICLR 2019 Deep Reinforcement Learning meets Structured Prediction Workshop*, 2019. (Cited on pages 80, 85, and 87.)

[83] A. Kuhnle, M. Aroca-Ouellette, A. Basu, M. Sensoy, J. Reid, and D. Zhang. Reinforcement learning for information retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2669–2672, 2021. (Cited on page 1.)

[84] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. (Cited on page 61.)

[85] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. (Cited on page 80.)

[86] S. Levine, A. Kumar, G. Tucker, and J. Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. (Cited on pages 76 and 102.)

[87] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010. (Cited on pages 25 and 60.)

[88] S. Li, K. Kallidromitis, A. Gokul, Y. Kato, and K. Kozuka. Aligning diffusion models by optimizing human utility. *arXiv preprint arXiv:2404.04465*, 2024. (Cited on page 80.)

[89] B. Liu, Q. Cai, Z. Yang, and Z. Wang. Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems*, 32, 2019. (Cited on pages 34 and 36.)

[90] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. (Cited on page 90.)

[91] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. (Cited on pages 11, 13, 14, 33, and 36.)

[92] Y. Liu, J.-N. Yen, B. Yuan, R. Shi, P. Yan, and C.-J. Lin. Practical counterfactual policy learning for top-k recommendations. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1141–1151. ACM, 2022. doi: 10.1145/3534678.3539295. (Cited on page 60.)

[93] Y.-A. Liu, R. Zhang, J. Guo, M. de Rijke, W. Chen, Y. Fan, and X. Cheng. Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1647–1656, 2023. (Cited on page 41.)

[94] B. London, A. Buchholz, G. Di Benedetto, J. M. Lichtenberg, Y. Stein, and T. Joachims. Self-normalized off-policy estimators for ranking. In *CONSEQUENCES Workshop at ACM RecSys '23*, CONSEQUENCES '23, 2023. (Cited on page 76.)

[95] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. (Cited on page 95.)

[96] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim. Applications of deep reinforcement learning in communications and networking: A survey. *IEEE communications surveys & tutorials*, 21(4):3133–3174, 2019. (Cited on page 1.)

[97] J. Ma, Z. Zhao, X. Yi, J. Yang, M. Chen, J. Tang, L. Hong, and E. H. Chi. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*, WWW '20, page 463–473. ACM, 2020. ISBN 9781450370233. doi: 10.1145/3366423.3380130. (Cited on pages 60 and 64.)

[98] A. Maurer and M. Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Annual Conference Computational Learning Theory*, 2009. (Cited on page 16.)

[99] J. McInerney, B. Lacker, S. Hansen, K. Higley, H. Bouchard, A. Gruson, and R. Mehrotra. Explore, exploit, and explain: Personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 31–39. ACM, 2018. doi: 10.1145/3240323.3240354. (Cited on page 60.)

[100] R. Mehrotra, N. Xue, and M. Lalmas. Bandit based optimization of multiple objectives on a music streaming platform. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3224–3233. ACM, 2020. doi: 10.1145/3394486.3403374. (Cited on page 60.)

[101] S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte Carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020. (Cited on pages 62, 63, 68, 85,

and 87.)

[102] S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems*, 29, 2016. (Cited on page 17.)

[103] H. Oosterhuis. *Learning from User Interactions with Rankings: A Unification of the Field*. PhD thesis, Informatics Institute, University of Amsterdam, 2020. (Cited on pages 16, 26, 33, 36, 37, and 41.)

[104] H. Oosterhuis. Computationally efficient optimization of Plackett-Luce ranking models for relevance and fairness. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1023–1032, 2021. (Cited on pages 24, 25, 26, 44, 46, and 51.)

[105] H. Oosterhuis. Learning-to-rank at the speed of sampling: Plackett-Luce gradient estimation with minimal computational complexity. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2266–2271, 2022. (Cited on page 44.)

[106] H. Oosterhuis. Reaching the end of unbiasedness: Uncovering implicit limitations of click-based learning to rank. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval*, 2022. (Cited on pages 2, 3, 11, 16, and 34.)

[107] H. Oosterhuis. Doubly robust estimation for correcting position bias in click feedback for unbiased learning to rank. *ACM Transactions on Information Systems*, 41(3):1–33, 2023. (Cited on pages 12, 33, 34, 35, 37, 38, 39, 41, 44, 46, 47, 53, 54, and 59.)

[108] H. Oosterhuis and M. de Rijke. Policy-aware unbiased learning to rank for top-k rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 489–498, 2020. (Cited on pages 13, 16, 24, 25, 26, 34, 35, 37, 38, 45, and 101.)

[109] H. Oosterhuis and M. de Rijke. Robust generalization and safe query-specialization in counterfactual learning to rank. In *Proceedings of the Web Conference 2021*, pages 158–170, 2021. (Cited on pages 2, 12, 13, 25, 30, 34, 35, and 44.)

[110] H. Oosterhuis and M. de Rijke. Unifying online and counterfactual learning to rank: A novel counterfactual estimator that effectively utilizes online interventions. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 463–471, 2021. (Cited on pages 13, 15, 16, 25, 26, 35, 37, 44, 46, and 51.)

[111] H. Oosterhuis, R. Jagerman, and M. de Rijke. Unbiased learning to rank: Counterfactual and online approaches. In *Companion Proceedings of the Web Conference 2020*, pages 299–300, 2020. (Cited on pages 12 and 13.)

[112] Z. Ovaisi, R. Ahsan, Y. Zhang, K. Vasilaky, and E. Zheleva. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*, pages 1863–1873, 2020. (Cited on pages 34 and 37.)

[113] A. B. Owen. *Monte Carlo Theory, Methods and Examples*. `https://artowen.su.domains/mc/`, 2013. (Cited on pages 60, 62, 63, 64, 85, 87, and 92.)

[114] A. Poyarkov, A. Drutsa, A. Khalyavin, G. Gusev, and P. Serdyukov. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 235–244. ACM, 2016. doi: 10.1145/2939672.2939688. (Cited on page 62.)

[115] T. Qin and T.-Y. Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013. (Cited on pages 25 and 44.)

[116] T. Qin, T.-Y. Liu, J. Xu, and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010. (Cited on pages 11, 13, and 33.)

[117] J. Queeney, Y. Paschalidis, and C. G. Cassandras. Generalized proximal policy optimization with sample reuse. In *Advances in Neural Information Processing Systems*, volume 34, pages 11909–11919, 2021. (Cited on pages 34, 36, 83, and 86.)

[118] F. Radlinski. Addressing malicious noise in clickthrough data. In *LR4IR 2007: Learning to Rank for Information Retrieval Workshop at SIGIR*, volume 2007, 2007. (Cited on page 41.)

[119] F. Radlinski, M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 43–52, 2008. (Cited on page 13.)

[120] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on page 84.)

[121] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image

generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. (Cited on pages 82 and 88.)

[122] S. Rendle. *Item Recommendation from Implicit Feedback*, pages 143–171. Springer US, New York, NY, 2022. ISBN 978-1-0716-2197-4. doi: 10.1007/978-1-0716-2197-4_4. (Cited on page 60.)

[123] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1. Berkeley, California, USA, 1961. (Cited on pages 14, 17, 18, and 39.)

[124] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. RecoGym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720*, 2018. (Cited on pages 60 and 71.)

[125] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. (Cited on pages 79, 88, and 90.)

[126] Y. Saito and T. Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proc. of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 828–830. ACM, 2021. ISBN 9781450384582. doi: 10.1145/3460231.3473320. (Cited on page 60.)

[127] Y. Saito and T. Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Fifteenth ACM Conference on Recommender Systems*, pages 828–830, 2021. (Cited on pages 1, 14, 26, 35, and 59.)

[128] Y. Saito and T. Joachims. Counterfactual evaluation and learning for interactive systems: Foundations, implementations, and recent advances. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4824–4825, 2022. (Cited on page 61.)

[129] Y. Saito, S. Aihara, M. Matsutani, and Y. Narita. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020. (Cited on pages 25 and 60.)

[130] Y. Saito, T. Udagawa, H. Kiyohara, K. Mogi, Y. Narita, and K. Tateno. Evaluating the robustness of off-policy evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 114–123. ACM, 2021. ISBN 9781450384582. doi: 10.1145/3460231.3474245. (Cited on pages 60, 63, 64, and 71.)

[131] O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. 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*, KDD '20, page 783–793. ACM, 2020. ISBN 9781450379984. doi: 10.1145/3394486.3403121. (Cited on page 72.)

[132] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. (Cited on page 33.)

[133] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 555–562, 2010. (Cited on page 13.)

[134] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. (Cited on page 80.)

[135] J. Schulman. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015. (Cited on pages 83 and 85.)

[136] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. (Cited on pages 40 and 62.)

[137] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. (Cited on pages 34, 36, 40, 79, 80, 83, 84, 85, and 86.)

[138] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. (Cited on page 12.)

[139] G. Shani, R. I. Brafman, and D. Heckerman. An MDP-based recommender system. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 453–460, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608974. (Cited on page 60.)

[140] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265.

PMLR, 2015. (Cited on pages 79 and 82.)

[141] M. Speretta and S. Gauch. Personalized search based on user search histories. In *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*, pages 622–628. IEEE, 2005. (Cited on page 13.)

[142] H. Steck. Evaluation of recommendations: Rating-prediction and ranking. In *Proc. of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 213–220. ACM, 2013. ISBN 9781450324090. doi: 10.1145/2507157.2507160. (Cited on page 60.)

[143] Y. Su, L. Wang, M. Santacatterina, and T. Joachims. CAB: Continuous adaptive blending for policy evaluation and learning. In *Proc. of the 36th International Conference on Machine Learning*, volume 97 of *ICML '19*, pages 6005–6014. PMLR, 09–15 Jun 2019. (Cited on pages 60, 66, and 71.)

[144] Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020. (Cited on pages 60, 66, 70, and 71.)

[145] Y. Su, X. Wang, E. Y. Le, L. Liu, Y. Li, H. Lu, B. Lipshitz, S. Badam, L. Heldt, S. Bi, E. H. Chi, C. Goodrow, S.-L. Wu, L. Baugher, and M. Chen. Long-term value of exploration: Measurements, findings and algorithms. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 636–644. ACM, 2024. doi: 10.1145/3616855.3635833. (Cited on page 60.)

[146] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018. (Cited on pages 61, 83, and 86.)

[147] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015. (Cited on pages 3, 59, 60, 61, 64, 65, 73, and 87.)

[148] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015. (Cited on pages 12, 14, 16, 17, 35, 36, 39, and 65.)

[149] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. (Cited on pages 12, 35, 39, and 44.)

[150] A. Ustimenko and L. Prokhorenkova. Stochasticrank: Global optimization of scale-free discrete functions. In *International Conference on Machine Learning*, pages 9669–9679. PMLR, 2020. (Cited on page 44.)

[151] B. van den Akker, O. Jeunen, Y. Li, B. London, Z. Nazari, and D. Parekh. Practical bandits: An industry perspective. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1132–1135. ACM, 2024. (Cited on pages 1 and 60.)

[152] A. Vardasbi, H. Oosterhuis, and M. de Rijke. When inverse propensity scoring does not work: Affine corrections for unbiased learning to rank. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1475–1484, 2020. (Cited on pages 2, 13, 25, 33, 34, 35, 37, 44, and 46.)

[153] F. Vasile, D. Rohde, O. Jeunen, and A. Benhalloum. A gentle introduction to recommendation as counterfactual policy learning. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 392–393. ACM, 2020. ISBN 9781450368612. doi: 10.1145/3340631.3398666. (Cited on pages 60 and 63.)

[154] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. (Cited on page 80.)

[155] C. Wang, M. Uehara, Y. He, A. Wang, T. Biancalani, A. Lal, T. Jaakkola, S. Levine, H. Wang, and A. Regev. Fine-tuning discrete diffusion models via reward optimization with applications to dna and protein design. *arXiv preprint arXiv:2410.13643*, 2024. (Cited on page 80.)

[156] H.-n. Wang, N. Liu, Y.-y. Zhang, D.-w. Feng, F. Huang, D.-s. Li, and Y.-m. Zhang. Deep reinforcement learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 21(12):1726–1744, 2020. (Cited on page 1.)

[157] X. Wang, M. Bendersky, D. Metzler, and M. Najork. Learning to rank with selection bias in personal search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 115–124. ACM, 2016. (Cited on pages 12, 33, 34, and 37.)

[158] X. Wang, N. Golbandi, M. Bendersky, D. Metzler, and M. Najork. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on*

Web Search and Data Mining, pages 610–618, 2018. (Cited on pages 12, 13, 33, and 37.)

[159] X. Wang, C. Li, N. Golbandi, M. Bendersky, and M. Najork. The LambdaLoss framework for ranking metric optimization. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1313–1322, 2018. (Cited on page 44.)

[160] Y. Wang, H. He, X. Tan, and Y. Gan. Trust region-guided proximal policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. (Cited on pages 34 and 36.)

[161] Y. Wang, H. He, and X. Tan. Truly proximal policy optimization. In *Uncertainty in Artificial Intelligence*, pages 113–122. PMLR, 2020. (Cited on pages 35 and 36.)

[162] R. J. Williams. Toward a theory of reinforcement-learning connectionist systems. Technical Report NU-CCS-88-3, Northeastern University, 1988. (Cited on page 62.)

[163] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992696. (Cited on page 61.)

[164] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. (Cited on pages 2, 24, 25, 44, 80, and 84.)

[165] H. Wu and M. Wang. Variance regularized counterfactual risk minimization via variational divergence minimization. In *International Conference on Machine Learning*, pages 5353–5362. PMLR, 2018. (Cited on pages 12, 14, 17, 18, 36, 39, and 54.)

[166] Z. Xie, C. Yu, and W. Qiao. Dropout strategy in reinforcement learning: Limiting the surrogate objective variance in policy optimization methods. *arXiv preprint arXiv:2310.20380*, 2023. (Cited on page 87.)

[167] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on page 80.)

[168] M. Xu, A. S. Powers, R. O. Dror, S. Ermon, and J. Leskovec. Geometric latent diffusion models for 3D molecule generation. In *International Conference on Machine Learning*, pages 38592–38610. PMLR, 2023. (Cited on page 79.)

[169] H. Yadav, Z. Du, and T. Joachims. Policy-gradient training of fair and unbiased ranking functions. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1044–1053, 2021. (Cited on pages 13, 24, 26, 37, 41, 46, and 51.)

[170] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023. (Cited on page 2.)

[171] X. Yi, S.-C. Wang, R. He, H. Chandrasekaran, C. Wu, L. Heldt, L. Hong, M. Chen, and E. H. Chi. Online matching: A real-time bandit system for large-scale recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, page 403–414, New York, NY, USA, 2023. ACM. doi: 10.1145/3604915.3608792. (Cited on page 60.)

[172] Y. Zeng, G. Liu, W. Ma, N. Yang, H. Zhang, and J. Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024. (Cited on page 84.)

[173] W. Zhang, X. Zhao, L. Zhao, D. Yin, G. H. Yang, and A. Beutel. Deep reinforcement learning for information retrieval: Fundamentals and advances. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2468–2471, 2020. (Cited on page 1.)

[174] R. Zheng, S. Dou, S. Gao, Y. Hua, W. Shen, B. Wang, Y. Liu, S. Jin, Y. Zhou, L. Xiong, et al. Delve into PPO: Implementation matters for stable RLHF. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. (Cited on page 80.)

[175] H. Zhong, G. Feng, W. Xiong, X. Cheng, L. Zhao, D. He, J. Bian, and L. Wang. DPO meets PPO: Reinforced token optimization for RLHF. In *arXiv preprint arXiv:2404.18922*, 2024. (Cited on page 84.)

Summary

This dissertation investigates how reinforcement learning methods can be made simultaneously safe, sample-efficient, and robust when trained only from logged user interactions. Under the unifying lens of contextual-bandit reinforcement learning, the work spans two application families: web-search ranking/recommendation and text-to-image diffusion models. In this thesis, we pair new theory with practical algorithms that (i) guarantee safe deployment against the production system, (ii) extract more signal from limited logs in an off-policy evaluation and learning setup, and (iii) scale to modern large-scale generative models.

In the first part of the thesis, we derive an exposure-based generalisation bound that upper-bounds the true ranking utility. Optimising the bound yields a counterfactual risk-minimisation (CRM) objective whose solution is provably no worse than the logging policy even with few clicks, resulting in safe deployment. Further, we proposed a robust safe deployment method that extends safety to doubly-robust estimators, and retains guarantees under adversarial or mis-specified behaviour models. The proposed method offers practitioners direct control over the maximum allowed utility drop.

In the second part of the thesis, shifting to single-action bandits, our contribution unifies IPS, self-normalised IPS and doubly robust estimators inside an unifying baseline-correction framework. We propose a closed-form optimal baseline term that is proved to minimise both evaluation and policy-gradient variance.

In the final chapter we revisit the efficiency-effectiveness trade-off in a generative reinforcement learning setup. A systematic PPO-vs-REINFORCE study reveals an "efficiency–effectiveness" trade-off, inspiring leave-one-out PPO (LOOP). LOOP generates several diffusion trajectories per prompt and inserts a REINFORCE-style baseline inside PPO's clipped objective, matching PPO quality while binding textual attributes more faithfully on text-to-image diffusion benchmark.

Finally, the thesis gives the following answers: (i) safety can be guaranteed for ranking – with or without click-model assumptions; (ii) a single baseline parameter can unify and optimise bandit variance reduction; and (iii) lightly modified reinforcement learning algorithms can fine-tune large diffusion models efficiently. Together these advances demonstrate a path toward reliable, safe, and data-efficient reinforcement learning pipelines for real-world information access and generative AI and open avenues for extending safe-bandit theory to multitask and multi-objective foundation models.

Samenvatting

Dit proefschrift onderzoekt hoe *reinforcement learning*-methoden tegelijkertijd veilig, sample-efficiënt en robuust kunnen worden gemaakt wanneer ze uitsluitend worden getraind op basis van geregistreerde gebruikersinteracties.

Onder de overkoepelende lens van *contextual-bandit* reinforcement learning omvat het werk twee families van toepassingen: webzoekrangschikking/-aanbeveling en tekst-naar-afbeelding diffusiemodellen.

In dit proefschrift combineren we nieuwe theorie met praktische algoritmen die (i) veilige implementatie tegen het productiesysteem garanderen, (ii) meer signaal extraheren uit beperkte logs in een off-policy evaluatie- en leeropstelling, en (iii) opschalen naar moderne grootschalige generatieve modellen.

In het eerste deel van het proefschrift leiden we een *exposure*gebaseerde generalisatiegrens af die de werkelijke bruikbaarheid van de rangschikking begrenst. Het optimaliseren van de grens levert een contrafactische risicominimalisatiedoelstelling op waarvan de oplossing aantoonbaar niet slechter is dan het logbeleid, zelfs met weinig klikken, wat resulteert in een veilige implementatie. Verder hebben we een robuuste, veilige implementatiemethode voorgesteld die de veiligheid uitbreidt naar dubbelrobuuste schatters en garanties behoudt onder vijandige of verkeerd gespecificeerde gedragsmodellen. De voorgestelde methode biedt professionals directe controle over de maximaal toegestane utiliteitsdaling.

In het tweede deel van het proefschrift, overgaand op *single-action* bandits, verenigt onze bijdrage IPS, zelfgenormaliseerde IPS en dubbelrobuuste schatters binnen een uniform basislijncorrectiekader. We stellen een gesloten, optimale basislijnterm voor waarvan bewezen is dat deze zowel de variantie in evaluatie als in beleidsgradiënt minimaliseert.

In het laatste hoofdstuk bekijken we de afweging tussen efficiëntie en effectiviteit in een generatieve reinforcement learning-opzet opnieuw. Een systematische PPO-versus-REINFORCE-studie onthult een afweging tussen efficiëntie en effectiviteit, wat inspireert tot leave-one-out PPO (LOOP). LOOP genereert meerdere diffusietrajecten per prompt en voegt een REINFORCE-achtige basislijn in binnen de afgeknipte doelstelling van PPO, die overeenkomt met de PPO-kwaliteit en tegelijkertijd tekstuele kenmerken getrouwer koppelt aan de tekst-naar-afbeelding diffusiebenchmark.

Ten slotte geeft het proefschrift de volgende antwoorden: (i) veiligheid kan worden gegarandeerd voor rangschikking – met of zonder aannames voor het klikmodel; (ii) één basislijnparameter kan de reductie van banditvariantie verenigen en optimaliseren; (iii) licht aangepaste *reinforcement learning* algoritmen kunnen grote diffusiemodellen efficiënt verfijnen.

Samen tonen deze ontwikkelingen een pad naar betrouwbare, veilige en data-efficiënte *reinforcement learning* pipelines voor real-world informatietoegang en generatieve AI, en openen ze mogelijkheden om de safe-banditheorie uit te breiden naar multitask- en multi-objectieve basismodellen.