INSTITUT
POLYTECHNIQUE
DE PARIS

NNT : 2023IPPAG011

Thèse de doctorat

Offline Contextual Bandit : Theory and Large Scale Applications

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 18/12/2023, par

OTMANE SAKHI

Composition du Jury :

Olivier Catoni Directeur de Recherche, (CREST)	Président
Benjamin Guedj Chargé de recherche (UCL & Inria)	Rapporteur
Emilie Kauffman Chargée de recherche, Univ. Lille, Inria (CRIStAL)	Rapporteuse
Julyan Arbel Chargé de recherche, Inria Grenoble (Statify)	Examinateur
Anna Korba Maître de conférences, Institut Polytechnique de Paris (ENSAE)	Examinatrice
Nicolas Chopin Professeur, Institut Polytechnique de Paris (ENSAE)	Directeur de thèse
David Rohde Chercheur, Criteo AI Lab	Invité

FM
JH
FONDATION
MATHÉMATIQUE
JACQUES HADAMARD

CREST

CENTER FOR RESEARCH
IN ECONOMICS AND STATISTICS

CRITEO

AI Lab

Acknowledgements

First and foremost, I want to express my sincere and deep gratitude to my two thesis supervisors, David and Nicolas. Your guidance and enthusiasm for research have been immensely helpful throughout these intense years, both in the development of this thesis and in building my research abilities.

David, thank you for taking me under your wing in the middle of my internship and your trust to prolong the adventure into a doctorate. Thank you for the human qualities that you have and the patience and optimism that characterize you. I am grateful for all the laughter, the research discussions we engaged, your attention to detail, the freedom you gave me to explore different research topics and the unending support you showed at the beginning of each new project.

Nicolas, a big thanks for accepting to be part of the adventure and for your gentle supervision. Your experience, broad knowledge and rigour ignited my research curiosity and inspired me to strive for excellence. Thank you for all the scientific discussions, your open-mindedness and the infinitely-many research ideas that a single thesis cannot cover. Thank you for always finding great collaborators to push our work and for your never ending trust in my capabilities. I am confident that this thesis only represents the beginning of a beautiful friendship and many future research collaborations.

I also want to thank Criteo for providing a stimulating environment, rich with interesting problems that inspire impactful research. Thanks to the entire CAIL and to the Reco research team in particular for all the moments shared. Thanks to all my colleagues and friends with whom I collaborated, and with whoever I could share a discussion, a coffee or a beer. Numerous special moments within Criteo made this whole experience much more enjoyable.

Thanks to Julyan Arbel, Olivier Catoni, Benjamin Guedj, Émilie Kaufmann and Anna Korba for accepting to be part of my committee. Your enthusiasm and interest in my work have been truly heartwarming. A special thanks to Benjamin and Émilie for accepting to review my dissertation; it is a great honor for me.

Finally, I want to thank my family and loved ones. Special thanks to my parents and my brother, who have always supported and encouraged me unwaveringly. A huge thanks to Ayman, Oussama and El Ghanjaoui for turning this experience into a pleasant journey and to Iness for adding beauty to my days.

Abstract

Abstract. Modern interactive systems shape our internet experience. From search to recommendation engines, these systems organise vast amounts of content, allowing users to efficiently find answers to their needs. The quality of user experiences within these systems can vary significantly, and the ability to provide users with relevant options at the right moment can greatly enhance both user satisfaction and the profitability of the businesses operating these systems. In recent years, there has been a concerted effort to leverage machine learning techniques to improve interactive systems by combining various signals. This thesis focuses on harnessing a specific type of signal: *user interaction logs*. These logs are uniquely valuable as they directly capture successes and failures in previous interactions. Nonetheless, the interactive nature of the logs makes their analysis more challenging compared to classical supervised learning problems. The Offline Contextual Bandit formalises an idealized version of this learning problem. It reduces the interaction logs to triplets of an observed context, an action made by the system and a reward received. These triplets are core to analyse the problem and to learn improved interactive systems. Notwithstanding recent advances, there remains significant challenges to learn decision systems with performance certificates and scale current approaches to real world problems.

Our first concern is being able to measure how well an interactive system will perform before it engages with the environment. Statistical learning theory focuses on studying the generalization ability of algorithms, and presents itself as the perfect candidate to answer this question. Historically, its tools were used to improve our understanding of the supervised learning paradigm, resulting in Empirical and Structural Risk minimization principles. More recently, statistical learning theory was adapted to learning from interaction logs, and resulted in the Counterfactual Risk minimization principle. This new objective captures the difficulties of learning from contextual bandit logs, but its application is limited to simple scenarios. In particular, the learning objective is non-convex, it cannot be accelerated with stochastic gradient methods, it introduces new hyperparameters that are difficult to tune and fails to provide performance certificates on the newly trained interaction systems. The first part of the thesis focuses on developing new statistical learning ideas to address these challenges. We reframe the Counterfactual Risk minimization using Distributionally Robust Optimization. This change of perspective allows us to improve the optimization procedure, to automatically calibrate hyperparameters while enjoying the same guarantees. Furthermore, we explore PAC-Bayesian learning, a statistical learning framework that provides a finer analysis of the generalization ability of algorithms. Using this paradigm, we build new strategies that require no hyperparameter tuning, that enable fast optimization and can provide strong guarantees on the performance of our interactive systems.

Another concern is to efficiently learn decision systems operating on massive action spaces. The second part of the thesis addresses this challenge, focusing primarily on large scale recommendation. Efficient learning in this case can be achieved by exploiting different signals and speeding up the optimization routine. Existing methods rely solely on the bandit signal: the log of the past successes and failures. However, non-bandit signal, such as collaborative filtering, can be extremely valuable. Building on this observation, we dedicate a chapter to develop a Bayesian approach to recommendation that combines both signals. We give proper computational tools to scale the learning to large datasets and prove empirically that the resulting systems enjoy improved recommendation quality.

Large scale recommender systems are updated frequently to match the ever-shifting interests of the users. The ability to perform these updates regularly relies on the efficiency of the optimization routine. When confronted with exceedingly large action spaces, these systems are constrained to the maximum inner product search (MIPS) structure for rapid query responses. Despite their prevalence in the industry, optimizing these systems with common learning objectives tend to be slow. Indeed, every gradient iteration scales at least linearly with the catalog size. This complexity can be detrimental to learning recommender systems operating on billions of items. The last two chapters address this issue by proposing optimization routines with sublinear complexities; a first solution is based on a new importance sampling variant of the reinforce algorithm, and a second one introduces a novel architecture and method for optimizing MIPS-based interactive systems. The proposed solutions accelerate optimization without losing on the recommendation quality.

Résumé. Les systèmes interactifs modernes façonnent notre expérience de l'internet. Des moteurs de recherche aux moteurs de recommandation, ces systèmes organisent de vastes quantités de contenu, permettant aux utilisateurs de trouver efficacement des réponses à leurs besoins. La qualité de l'expérience utilisateur au sein de ces systèmes peut varier de manière significative, et la capacité à fournir aux utilisateurs des options pertinentes au bon moment peut, non seulement améliorer leur satisfaction, mais aussi la rentabilité des entreprises qui exploitent ces systèmes. Ces dernières années, des efforts concertés ont été déployés pour exploiter les techniques d'apprentissage automatique afin d'améliorer les systèmes interactifs en combinant différents signaux. Cette thèse se concentre sur l'exploitation d'un type spécifique de signaux : *les données d'interaction*. Ces données ont une valeur unique car ils enregistrent directement les succès et les échecs des interactions précédentes. Néanmoins, la nature interactive de ces données rend leur analyse plus difficile par rapport aux problèmes classiques d'apprentissage. Le bandit contextuel hors-ligne formalise une version idéalisée de ce problème d'apprentissage. Il réduit les données d'interaction à des triplets; un contexte observé, une action effectuée par le système et une récompense reçue. Ces triplets sont essentiels à l'analyse du problème et à l'apprentissage de systèmes interactifs améliorés. Malgré les progrès récents, il reste des défis importants à relever pour apprendre des systèmes de décision avec des certificats de performance et pour adapter les approches actuelles aux problèmes de grande échelle.

Notre première préoccupation est de pouvoir mesurer les performances de notre système avant qu'il intéragisse avec l'environnement. La théorie de l'apprentissage statistique se concentre sur l'étude de la capacité de généralisation des algorithmes et se présente comme le candidat idéal pour répondre à cette question. Historiquement, ses outils ont été utilisés pour améliorer notre compréhension du paradigme de l'apprentissage supervisé, donnant naissance aux principes de minimisation du risque empirique et structurel. Plus récemment, la théorie de l'apprentissage statistique a été adaptée à l'apprentissage à partir de données d'interactions, ce qui a donné nais-

sance au principe de minimisation du risque contrefactuel. Ce nouvel objectif tient compte des difficultés liées à l'apprentissage à partir de données de bandits contextuels, mais son application est limitée à des scénarios simples. En particulier, l'objectif d'apprentissage n'est pas convexe, il ne peut pas être accéléré avec des méthodes de gradient stochastique, il introduit de nouveaux hyperparamètres qui sont difficiles à régler et ne parvient pas à fournir des certificats de performance sur les systèmes d'interaction nouvellement formés. La première partie de la thèse se concentre sur le développement de nouvelles idées d'apprentissage statistique pour relever ces défis. Nous recadrons la minimisation du risque contrefactuel **(CRM)** en utilisant l'optimisation distributionnellement robuste **(DRO)**. Ce changement de perspective nous permet d'améliorer la procédure d'optimisation, de calibrer automatiquement les hyperparamètres tout en bénéficiant des mêmes garanties. En outre, nous nous intéressons à l'apprentissage PAC-Bayésien, un cadre d'apprentissage statistique capable de mieux analyser la capacité de généralisation des algorithmes. En utilisant ce paradigme, nous construisons de nouvelles stratégies qui ne nécessitent aucun réglage des hyperparamètres, qui permettent une optimisation rapide et qui peuvent fournir des garanties solides sur la performance de nos systèmes interactifs.

Une autre préoccupation est d'apprendre efficacement les systèmes de décision fonctionnant sur des espaces d'action massifs. La deuxième partie de la thèse aborde ce défi, en se concentrant principalement sur la recommandation à grande échelle. L'apprentissage efficace dans ce cas peut être réalisé en exploitant différents signaux et en accélérant la procédure d'optimisation. Les méthodes existantes s'appuient uniquement sur le signal de bandit : les données d'intéraction du système avec les utilisateurs. Cependant, les signaux autres que le signal bandit, tels que le comportement organique, peuvent s'avérer extrêmement précieux. Sur la base de cette observation, nous consacrons un chapitre au développement d'une approche bayésienne de la recommandation qui combine les deux signaux. Nous fournissons les outils d'optimisation appropriés pour étendre l'apprentissage à de grands ensembles de données et prouvons empiriquement que les systèmes résultants bénéficient d'une meilleure qualité de recommandation.

Les systèmes de recommandation à grande échelle sont fréquemment mis à jour pour s'adapter aux intérêts en constante évolution des utilisateurs. La capacité à effectuer ces mises à jour régulièrement dépend de l'efficacité de la procédure d'optimisation. Lorsqu'ils sont confrontés à des espaces d'action extrêmement vastes, ces systèmes sont contraints à la structure **(MIPS)**: recherche du produit scalaire maximal pour répondre rapidement aux requêtes. Malgré leur prévalence dans l'industrie, l'optimisation de ces systèmes avec des objectifs d'apprentissage communs tend à être lente. En effet, le calcul de chaque gradient a une complexité au moins linéaire par rapport à la taille du catalogue. Cette complexité peut être préjudiciable à l'apprentissage de systèmes de recommandation fonctionnant sur des milliards d'éléments. Les deux derniers chapitres abordent ce problème en proposant des procédures d'optimisation avec des complexités sous-linéaires ; une première solution est basée sur une nouvelle variante d'échantillonnage préférentiel, et une seconde introduit une nouvelle architecture et une méthode pour optimiser les systèmes interactifs de la structure **(MIPS)**. Les solutions proposées accélèrent l'optimisation sans nuire à la qualité de la recommandation.

Contents

Introduction en français

1 Présentation générale

Ce manuscrit présente des contributions récentes, allant de la théorie aux applications à grande échelle, à un formalisme hors ligne du problème de la prise de décision séquentielle. Il s'agit d'un problème important avec de nombreuses applications dans le monde réel où un décideur, chargé d'optimiser un objectif spécifique, intéragit avec un environnement inconnu, enregistre ces intéractions et les exploite afin de mieux résoudre la tâche. Dans ce contexte, nous souhaitons répondre à la question suivante :

Comment tirer parti des interactions antérieures du décideur pour améliorer ses performances?

La réponse à cette question peut avoir un impact important sur les problèmes pratiques du monde réel. Par exemple, elle peut aider une campagne de marketing en ligne à obtenir plus de dons pour une campagne caritative, elle peut rendre plus précise la prescription des médicaments, ou elle peut simplement améliorer la qualité de la recommandation de votre plateforme de streaming préférée. Dans cette introduction, nous présentons le problème de l'apprentissage des décideurs à l'aide de l'exemple de la recommandation, qui sera au centre d'une grande partie de cette thèse. Les systèmes de recommandation se présentent comme le plus grand pilier de l'expérience Internet moderne. Dans chaque interaction, ces systèmes naviguent silencieusement une quantité écrasante d'informations et la traitent pour répondre aux besoins spécifiques de l'utilisateur. Une seule interaction d'un moteur de recommandation peut être résumée comme suit : le système rencontre un utilisateur, il choisit un article (ou plusieurs articles) à recommander dans un catalogue potentiellement vaste, délivre la recommandation et observe un retour de l'utilisateur.

Le retour obtenu est précieux car il représente les succès et les échecs des interactions passées. Ces interactions sont enregistrées et sont ensuite utilisées pour améliorer la qualité des recommandations du système. La nature interactive de l'ensemble des données collectées fait que les paradigmes d'apprentissage courants, tels que l'apprentissage supervisé, ne sont pas adaptés à l'étude de ce problème. Récemment, on s'est intéressé à l'adaptation des formalismes de prise de décision séquentielle pour améliorer la recommandation à partir des interactions enregistrées. L'apprentissage par renforcement (RL) (Sutton and Barto, 2018) et les bandits contextuels (CB) (Lattimore and Szepesvári, 2020) commencent à s'imposer comme de bons candidats pour modéliser ce problème d'apprentissage. Le cadre RL repose sur l'idée que les actions effectuées peuvent avoir un impact sur l'environnement. Ce paradigme peut modéliser des problèmes de décision séquentielle complexes et permet la planification. Ses outils peuvent optimiser les systèmes de recommandation pour des objectifs long terme ; par exemple, augmenter l'engagement et la rétention des utilisateurs (Afsar et al., 2022). L'adoption de ce formalisme a toutefois

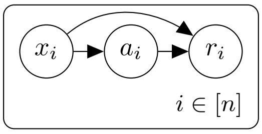

Figure 1: L'ensemble des données enregistrées $\mathcal{D}_n$ représentant n interactions du système de recommandation. Tous les triplets (contexte, action, récompense) sont indépendants.

un coût. La prise en compte des effets à long terme de la recommandation sur les utilisateurs rend l'analyse de cette approche plus difficile, ce qui nous incite à envisager un formalisme plus simple. Le bandit contextuel offre un compromis utile entre l'analyse formelle et l'impact pratique. Son hypothèse sous-jacente est que les actions effectuées par le système n'influencent pas les résultats futurs. Si cette formulation est moins convaincante lorsqu'il s'agit de récompenses différées (Afsar et al., 2022), son utilisation est raisonnable si nous voulons nous concentrer sur l'apprentissage de systèmes de recommandation qui optimisent des objectifs à court terme, limitées à l'action, telles que le taux de clics (Sakhi et al., 2020a) ou la durée de visionnage (Chen et al., 2019a). Dans cette thèse, nous adoptons la boîte à outils des bandits contextuels hors ligne (Bottou et al., 2013; Nguyen-Tang et al., 2022) pour formaliser l'apprentissage à partir des données d'interaction. Nous donnons de nouvelles approches fondées théoriquement pour apprendre des politiques avec de fortes garanties de performance et proposons de nouveaux algorithmes pour élargir l'impact de ce cadre à des applications à grande échelle du monde réel.

L'interaction d'un utilisateur avec un article recommandé peut être réduite à l'exemple suivant. Un utilisateur navigue sur un site web, le système de recommandation choisit un article dans un catalogue et le montre à l'utilisateur, l'utilisateur interagit avec l'article (clique ou non) et le résultat de cette interaction est encodé dans un retour d'information (présence/absence de clic) que le système enregistre. Dans le cadre du bandit contextuel, un utilisateur est représenté par un contexte x, généralement un vecteur réel vivant dans un espace à d dimensions $\mathcal{X} \subseteq \mathbb{R}^d$. Ces contextes, et donc les utilisateurs, sont échantillonnés *indépendamment* à partir de la même distribution inconnue $\nu(\mathcal{X})$. Après avoir vu un utilisateur, le moteur de recommandation lui fournit un article a issu d'un catalogue $\mathcal{A}$ de taille $|\mathcal{A}|\ dans\mathbb{N}$. Le système de recommandation est modélisé comme une politique $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, qui est une fonction qui prend un contexte x et produit une distribution $\pi(\cdot|x)$ sur l'espace des actions possibles $\mathcal{A}$. Recommander un article a pour le contexte x revient à échantillonner l'article à partir de la distribution produite $a \sim \pi(\cdot|x)$. Après avoir livré l'article a à l'utilisateur du contexte x, notre système reçoit un retour de l'utilisateur; une récompense stochastique $r \in \mathbb{R}^+$ provenant d'une distribution inconnue $p(\cdot|x, a)$. Cette récompense encode la performance de l'élément recommandé par rapport à la mesure souhaitée ; plus la récompense est élevée, plus la performance l'est aussi. Notre objectif est de trouver des politiques très performantes, en minimisant le risque, défini comme la récompense négative attendue en tirant des actions de notre politique. Le risque d'une politique donnée π peut être exprimé comme suit :

$$R(\pi) = -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\mathbb{E}_{r\sim p(\cdot|x,a)}[r]\right].$$

Ce risque est définie comme une espérance sous la distribution générée par la politique évaluée. Comme nous n'avons pas accès aux interactions de la nouvelle politique π avec l'environnement, un moyen simple d'estimer cette quantité est de laisser π interagir avec les utilisateurs en ligne. Dans la plupart des scénarios, cela n'est pas possible, car nous n'avons pas le luxe de déployer de mauvaises politiques. Dans les applications réelles, nous disposons déjà de la version actuelle de notre système de recommandation, représentée par la politique π_0, qui interagit avec l'environnement et enregistre ces intéractions. Notre objectif principal est d'évaluer dans quelle mesure une nouvelle itération du système améliorera la version actuellement déployée. Un moyen courant d'y parvenir est de réaliser des A/B-tests en ligne (Kohavi et al., 2012). Cette approche est considérée comme l'"étalon-or" pour estimer l'effet du remplacement de la politique actuelle π_0 par une politique potentiellement meilleure (Gupta et al., 2019). Les A/B-tests nécessitent toutefois un effort d'ingénierie important et un monitoring constant s'étalant sur plusieurs jours pour être correctement analysés. Idéalement, nous avons besoin d'outils d'évaluation et d'apprentissage hors ligne qui puissent nous trouver des politiques prometteuses afin de réduire le nombre d'A/B-tests inutiles. Lorsque les hypothèses du bandit contextuel sont satisfaites, nous pouvons utiliser la boîte à outils du cadre pour y parvenir. L'idée est d'exploiter les interactions existantes de π_0 pour trouver des politiques plus performantes. L'ensemble de données d'interaction est appelé dans la littérature "logged bandit feedback dataset" (Swaminathan and Joachims, 2015a) :

$$\mathcal{D}_n = \{x_i \sim \nu, a_i \sim \pi_0(\cdot|x_i), r_i \sim p(\cdot|x_i, a_i), \pi_0(a_i|x_i)\}_{i\in[n]}.$$

La figure 1 présente une représentation graphique des données. La principale difficulté rencontrée lors de l'apprentissage à partir de ces données est le biais potentiel créé par la procédure de collecte ; nous n'avons accès qu'aux résultats des actions échantillonnées à partir de π_0. Le cadre d'apprentissage hors ligne du bandit contextuel propose deux approches distinctes pour résoudre ce problème : l'approche de modélisation du coût et l'approche d'échantillonnage préférentiel.

l'approche de modélisation du coût ou *la méthode directe* exploite les données d'interaction $\mathcal{D}_n$ pour construire un modèle de la récompense (Sakhi et al., 2020a; Jeunen and Goethals, 2021). Une politique optimale est alors naturellement dérivée en jouant pour chaque contexte x, l'action avec la récompense la plus élevée selon le modèle. La méthode directe est simple à mettre en œuvre, car elle réduit l'apprentissage à un problème de régression (Brandfonbrener et al., 2021). Cette approche est théoriquement bien étudiée et bénéficie de solides garanties (Nguyen-Tang et al., 2022). Cependant, elle souffre d'un biais important et incontrôlé lorsque la récompense est complexe, ce qui rend son efficacité entièrement dépendante de notre capacité à modéliser la structure du problème. La méthode directe est efficace lorsque nous avons confiance en notre capacité à comprendre le problème. Lorsque le signal de récompense est complexe, nous pouvons préférer une autre approche qui ne dépend pas entièrement de notre effort de modélisation.

L'approche d'échantillonnage préférentiel (Horvitz and Thompson, 1952; Bottou et al., 2013; Dudík et al., 2014), souvent appelée *apprentissage hors politique*, ne nécessite pas de modèlisation. Elle apprend une nouvelle politique π directement à partir des interactions $\mathcal{D}_n$ en utilisant des estimateurs corrigés par échantillonnage préférentiel (Chopin and Papaspiliopoulos, 2020). Sous des hypothèses modérées (Horvitz and Thompson, 1952), cette méthode peut produire des estimateurs non biaisés, qui se présentent plus faciles à analyser et à optimiser (Ajalloeian and Stich, 2020). Ces estimateurs souffrent cependant d'une variance potentiellement importante dès que la politique apprise s'éloigne de la politique d'enregistrement π_0, ce qui les rend peu fiables pour l'apprentissage. Il est prouvé empiriquement que l'apprentissage avec ces estimateurs peut aboutir à des politiques peu performantes (Swaminathan and Joachims, 2015a,b), parfois même pires que π_0 (Chen et al., 2019b; London and Sandler, 2019). Cette observation motive l'utilisation d'outils de la théorie de l'apprentissage (Zhou, 2002; McAllester, 1998) pour proposer des objectifs avec un meilleur comportement, sans connaissance de la fonction de récompenses. L'objectif de cet effort de recherche est de produire de nouvelles politiques qui sont **théoriquement meilleures** que la politique d'enregistrement sans interactions additionelles avec l'environnement. Cela est utile dans les environnements de production où nous aimerions proposer un nouveau système qui améliorera le système de production actuel avec certitude.

Le premier effort dans ce sens a été mené par Swaminathan and Joachims (2015a) et a abouti au principe **CRM : Counterfactual Risk Minimisation** ou Minimisation du risque contrefactuel. Le principe **CRM** s'appuie sur les outils de la théorie de l'apprentissage statistique (Vapnik, 1998), un cadre qui permet d'étudier la capacité de généralisation des algorithmes d'apprentissage. Motivé par la construction d'une borne empirique de type Bernstein (Maurer and Pontil, 2009) sur le risque réel des politiques, et utilisant des arguments de nombre de couverture (Zhou, 2002), ce principe préconise de pénaliser les estimateurs de poids d'importance avec la *racine carrée de la variance empirique du risque*. Cette pénalité est contrôlée par un hyperparamètre λ, défini à l'aide d'une validation croisée sur une partie de validation. L'intuition sous-jacente est que pour améliorer la politique π_0, nous devrions rechercher des politiques qui ont un *petit risque empirique* tout en restant proches de π_0. Ce principe permet d'obtenir des politiques plus performantes que l'optimisation directe d'estimateurs d'échantillonnage préférentiel (Swaminathan and Joachims, 2015a,b). Toutefois, son paradigme d'apprentissage souffre de différentes limitations, ce qui réduit son application à des scénarios simples. En particulier, l'ajout de la pénalisation rend l'objectif d'apprentissage non convexe et non décomposable, ce qui interdit l'utilisation de méthodes de gradient stochastique. Cette pénalité est également contrôlée par un nouvel hyperparamètre λ qui est difficile à régler et qui ajoute à la complexité de l'approche. Enfin, le principe **CRM** ne fournit pas de certificats de performance sur la politique nouvellement formée. Ces limites seront examinées en détail plus loin dans l'introduction. Plus récemment, un nouveau principe a été introduit pour atténuer certaines de ces limitations. En analysant ce problème d'apprentissage sous l'angle PAC-Bayesien (McAllester, 1998; Alquier, 2021), London and Sandler (2019) développent une approche améliorée. Les auteurs fondent leur analyse sur la borne PAC-Bayesienne de McAllester (2003). Pour les politiques paramétriques, cela motive une régularisation L_2 du paramètre de la nouvelle politique vers le paramètre de la politique d'enregistrement π_0. La régularisation est également contrôlée par un hyperparamètre λ qui doit être réglé. Ce principe est basé sur la même intuition de rester proche de π_0, mais cette fois, il est effectué sur l'espace des paramètres. L'adoption d'une régularisation L_2 au lieu d'une pénalisation de la variance d'échantillon facilite le problème d'optimisation et permet l'utilisation de la descente de gradient stochastique. Cependant, le paramètre λ de la régularisation L_2 souffre des mêmes limitations et le principe ne peut pas produire de meilleures politiques. Les résultats empiriques démontrent que ces principes échouent parfois à améliorer la politique π_0 (Chen et al., 2019b). Ces limites seront développées dans la section suivante, avant que nous ne présentions les contributions de la première partie de la thèse. Le chapitre 3 recadre **CRM** en utilisant les outils de **Distributionnally Robust Optimisation** (Duchi et al., 2021), un cadre statistique conçu pour la prise de décision face à l'incertain. En outre, les chapitres 4 et 5 s'appuient sur les travaux de London and Sandler (2019) et poursuivent le développement des outils **PAC-Bayésien** (McAllester, 1998) pour le bandit contextuel hors ligne. L'analyse produit des principes qui sont plus faciles à optimiser, ne nécessitent pas d'hyperparamètres supplémentaires à régler et bénéficient, pour certains, de meilleures garanties de performance, ce qui nous rapproche de l'apprentissage de politiques améliorant π_0 hors ligne.

Dans le monde réel, les systèmes interactifs sont souvent confrontés à des scénarios à grande échelle, dans lesquels ils doivent apprendre à partir d'un nombre considérable d'interactions ($n \gg 1$) et opérer sur des catalogues massifs ($|\mathcal{A}| \gg 1$). Pour que ces systèmes puissent fournir des recommandations en quelques millisecondes, ils sont limités à une certaine structure (Shrivastava and Li, 2014; Aouali et al., 2022) afin de permettre une réponse rapide aux requêtes. Pendant longtemps, les systèmes de recommandation à grande échelle ont été formés à la prédiction des préférences (Harper and Konstan, 2015; Gomez-Uribe and Hunt, 2016) ou à la prédiction de l'élément suivant (Hidasi et al., 2015; Wu et al., 2019). Ces approches de modélisation sont généralement considérées comme de piètres substituts à la récompense que nous souhaitons optimiser (Jannach and Jugovac, 2019). L'adaptation de la boîte à outils de bandits contextuels hors ligne à l'apprentissage de systèmes de recommandation à grande échelle aura un impact considérable sur le secteur. Ces outils peuvent permettre d'aligner les recommandations sur des signaux de récompense complexes, améliorant ainsi la satisfaction des utilisateurs et la rentabilité des entreprises qui développent ces systèmes. Comme nous l'avons vu précédemment, nous pouvons soit adopter la **méthode directe** si nous savons comment modéliser la récompense, soit utiliser des **principes d'apprentissage avec des estimateurs d'échantillonnage préférentiel** pour apprendre une politique directement. Ces deux méthodes permettent d'apprendre de manière fiable un système de recommandation performant. Malheureusement, ces méthodes, dans leur forme simple, présentent des inconvénients lorsqu'elles traitent des problèmes à grande échelle. La deuxième partie de la thèse aborde ces limitations et permet un apprentissage efficace et rapide des systèmes de recommandation à grande échelle.

La méthode directe repose entièrement sur notre capacité à apprendre un modèle qui reflète les propriétés de la récompense. La compréhension parfaite du problème réduit le biais lié à la modélisation, mais il existe un autre problème, lié à l'apprentissage à partir de $\mathcal{D}_n$, qui devient plus prononcé dans les scénarios à grand catalogue. En effet, l'apprentissage naïf du modèle de récompense à partir de $\mathcal{D}_n$ souffre du déséquilibre présent dans les données collectées. Le modèle de récompense sera bien estimé pour les actions qui sont susceptibles d'être échantillonnées sous π_0, et mal estimé pour le reste. Cette différence dans la qualité de l'estimation peut rendre les décisions prises par la politique dérivée peu fiables (Smith and Winkler, 2006). Ce phénomène est accentué lorsqu'on a affaire à des catalogues de grande taille, car π_0 ne peut jamais collecter suffisamment d'échantillons pour couvrir l'ensemble de l'espace d'action. Nous consacrons le chapitre 6 à l'examen d'une solution bayésienne à ce problème. Nous introduisons une structure au modèle et utilisons une autre source de données pour apprendre efficacement le modèle de récompense. Plus de détails sur cette approche peuvent être trouvés dans la section contribution.

Les objectifs d'échantillonnage préférentiel deviennent intéressants lorsque le signal de récompense est complexe. Toutefois, dans les scénarios à grande échelle, ces objectifs d'apprentissage souffrent de deux problèmes majeurs. Le premier problème est lié à la variance de ces estimateurs, qui augmente avec la taille de l'espace d'action. En effet, la variance des estimateurs courants (Horvitz and Thompson, 1952; Ionides, 2008; Dudík et al., 2014) devient incontrôlable lorsque les politiques opèrent sur des catalogues massifs (Saito and Joachims, 2022b). Comme cette variance peut être très importante, l'ajout d'une pénalisation de la variance, par exemple, obligera la politique nouvellement apprise π à imiter le comportement de π_0. Ce phénomène rend nos principes d'apprentissage trop conservateurs, en renvoyant des politiques très proches de π_0. Cette observation a motivé la construction d'une nouvelle famille d'estimateurs (Saito and Joachims, 2022a; Saito et al., 2023) pour atténuer ce problème de variance. Ces contributions récentes traitent des limites statistiques des objectifs d'échantillonnage préférentiel dans les scénarios à grand catalogue, mais les problèmes de temps de calcul liés à l'optimisation de ces objectifs restent non résolus. Les systèmes à grande échelle sont fréquemment mis à jour, et des routines d'optimisation rapides sont hautement souhaitables dans ce contexte. Les méthodes existantes proposent des itérations de gradient dont l'échelle est au moins linéaire sur la taille du catalogue. Cette complexité peut être préjudiciable à l'apprentissage des systèmes de recommandation fonctionnant sur des milliards d'éléments. Les deux derniers chapitres (chapitres 7 et 8) se concentrent sur l'aspect computationnel et proposent des routines d'optimisation avec des complexités sous-linéaires. Ces solutions seront développées plus en détail dans la section contribution.

Dans cette thèse, nous couvrons différentes disciplines connectées, tout en équilibrant les outils théoriques et les algorithmes pratiques. Pour faciliter la présentation, nous souhaitons donner aux lecteurs un aperçu de l'avancement de chaque domaine de recherche. À cette fin, nous consacrons un chapitre à l'examen de la littérature existante, que nous jugeons utile pour tout chercheur.

Chapter 2. Literature Review. Ce chapitre présente une revue de la littérature couvrant ainsi les différents outils utilisés tout au long de cette thèse. Nous donnons un bref aperçu de la littérature sur le Bandit Contextuel, un formalisme pratique pour étudier la recommandation basée sur la récompense, en présentant à la fois ses formulations en ligne et hors ligne. En nous concentrant sur le cadre hors ligne, nous consacrons une section à la présentation des outils d'apprentissage statistique, nécessaires à l'étude des systèmes de décision d'apprentissage avec des garanties de performance. Nous présentons ensuite le développement des systèmes de recommandation et la manière dont la modélisation de la recommandation est passée de la prédiction des préférences à la maximisation de la récompense, et nous concluons par les considérations algorithmiques qui se posent dans le contexte de la prise de décision à grande échelle.

Part I - Offline Learning with Performance Guarantees. La première partie de la thèse se concentre sur les limites des principes d'apprentissage actuels. Ces principes ont été proposés pour améliorer la politique d'enregistrement π_0, en atténuant les problèmes liés à l'échantillonnage préférentiel. Sans perte de généralité, nous présentons le problème à l'aide du `IPS` : *Inverse Propensity Scoring* (Horvitz and Thompson, 1952), sans doute l'estimateur le plus simple et le plus étudié. Pour une politique π, nous rappelons son expression :

$$\hat{R}_n^{\texttt{IPS}}(\pi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i.$$

Lorsqu'elle est évaluée sur π_0, `IPS` donne la moyenne empirique des coûts collectés en tant qu'estimation du risque, ce qui est considéré comme un estimateur sans biais de $R(\pi_0)$. Toutefois, une simple analyse de la variance de cet estimateur montre que le fait de s'éloigner de π_0 entraîne une baisse de la qualité de l'estimation. Si la récompense observée est bornée (par exemple, $r \in [0,1]$), nous avons :

$$\begin{aligned}\mathbb{V}\left[\hat{R}_n^{\texttt{IPS}}(\pi)\right] &= \frac{1}{n}\left(\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x), r\sim p(\cdot|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2 r^2\right] - \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right]^2\right) \\ &\leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x), r\sim p(\cdot|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2\right] = \frac{1}{n}\left(\chi^2(\pi,\pi_0)+1\right),\end{aligned}$$

avec $\chi^2(\pi,\pi_0)$ la divergence χ-deux entre π et π_0. La variance a à peu près le même comportement que la divergence χ-deux, augmentant lorsque π s'éloigne de π_0. En particulier, la variance augmente avec les poids d'importance. Les poids d'importance sont très élevés lorsque la nouvelle politique π attribue une forte probabilité à des actions qui étaient très peu susceptibles d'être jouées sous π_0. Cela signifie que la qualité de l'estimation dépend fortement de la politique évaluée π, ce qui rend `IPS`[1] indigne de confiance pour les politiques éloignées

[1] Tous les estimateurs basés sur les poids d'importance souffrent de la même limitation.

du voisinage de π_0. Cette observation est confirmée dans la pratique, en particulier lorsque l'on utilise ces estimateurs comme objectif d'apprentissage. Par exemple, la minimisation de l'estimateur IPS par rapport à une classe de politiques peut conduire à des politiques ayant de mauvaises performances en ligne. Lorsque la politique apprise π est éloignée de π_0, l'estimation IPS du risque de π ne reflète pas son risque réel, car π se trouve dans une partie de l'espace qui induit un estimateur avec une grande variance. Pour contourner ces limitations, il faudrait restreindre l'optimisation aux politiques autour de la politique π_0. Le **CRM : Minimisation du risque contrefactuel** (Swaminathan and Joachims, 2015a) formalise cette idée en utilisant des arguments d'apprentissage statistique. Motivé par la construction d'une borne empirique de type Bernstein (Maurer and Pontil, 2009), le principe préconise la minimisation de l'estimateur IPS pénalisé par *sa variance*. Ce principe est ensuite utilisé pour produire un algorithme d'apprentissage de politiques "softmax" (Mei et al., 2020b) de la forme :

$$\begin{aligned}\forall (x,a) \quad \pi_\theta(a|x) &= \texttt{softmax}_{\mathcal{A}}\left(f_\theta(x,a)\right) \\ &= \frac{\exp(f_\theta(x,a)}{\sum_{a' \in \mathcal{A}} \exp(f_\theta(x,a'))}. \end{aligned} \tag{1}$$

avec θ un paramètre provenant d'un espace paramétrique $\theta \in \Theta$ et $f_\theta : \mathcal{X} \times \mathcal{A}$ une fonction qui encode la pertinence de l'action a par rapport au contexte x. L'algorithme proposé est appelé **POEM** : Policy Optimizer for Exponential Models (Swaminathan and Joachims, 2015a) et résout l'objectif suivant pour les politiques softmax :

$$\arg\min_{\theta \in \Theta} \left\{ \hat{R}_n^{\texttt{IPS}}(\pi_\theta) + \lambda \sqrt{\hat{V}^{\texttt{IPS}}(\pi_\theta)} \right\},$$

avec λ un hyperparamètre généralement défini à l'aide de données de validation, et $\hat{V}^{\texttt{IPS}}(\pi_\theta)$ la variance empirique induite par l'évaluation de π avec IPS :

$$\hat{V}^{\texttt{IPS}}(\pi_\theta) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i + \hat{R}_n^{\texttt{IPS}}(\pi_\theta) \right)^2.$$

Swaminathan and Joachims (2015a) a démontré empiriquement la supériorité de ce principe ; les politiques renvoyées par **POEM** présentent un risque beaucoup plus faible que celles obtenues en minimisant naïvement l'objectif IPS. L'ajout de la régularisation rend l'approche plus fondée, mais souffre encore de limitations qui réduisent son applicabilité dans les scénarios de la vie réelle:

(1) Mise à l'échelle. La plus grande limitation du principe CRM est sa capacité à s'adapter à de grands ensembles de données $\mathcal{D}_n$. La présence du terme de variance fait que le calcul du gradient de l'objectif CRM se fait en $\mathcal{O}(n)$, en termes de coûts de calcul et de mémoire, car il nécessite de parcourir l'entièreté des données. Dans un scénario de système de recommandation, des millions d'interactions sont enregistrées chaque jour. Ces applications traitent un très grand nombre d'échantillons n et ne peuvent se permettre ce coût de calcul. Ce problème est généralement résolu en recourant à un algorithme d'optimisation stochastique, qui ne nécessite qu'un accès aux gradients stochastiques non biaisés de l'objectif. Ceux-ci sont particulièrement faciles à obtenir lorsque l'objectif se décompose en une somme sur les entrées de l'ensemble de données, car il suffit de calculer la somme sur des lots pour obtenir des gradients non biaisés. Malheureusement, l'objectif CRM n'est pas adapté à l'optimisation stochastique, car le terme de pénalisation ne s'écrit pas comme une somme. Swaminathan and Joachims (2015a) a proposé une relaxation de l'objectif CRM, basée sur une stratégie de minimisation/majoration, qui peut bénéficier *partiellement* des gradients stochastiques. Leur approche nécessite toujours de passer par l'ensemble des données enregistrées de temps à autre, ce qui permet d'obtenir une procédure d'une complexité informatique identique.

London and Sandler (2019) propose un principe amélioré qui traite de la première limitation du CRM. Au lieu de s'appuyer sur la borne empirique de type Bernstein (Maurer and Pontil, 2009), London and Sandler (2019) adapte la borne PAC-Bayesienne de McAllester (2003) pour dériver des objectifs d'apprentissage pour ce problème. Le principe obtenu motive l'utilisation d'une régularisation L_2 vers le paramètre θ_0 de la politique π_0. Cette régularisation est contrôlée par un hyperparamètre λ, ce qui donne le problème d'optimisation suivant pour les politiques softmax paramétrées :

$$\arg\min_{\theta \in \Theta} \left\{ \hat{R}_n^{\texttt{IPS}}(\pi_\theta) + \lambda \|\theta - \theta_0\|^2 \right\},$$

L'objectif d'optimisation se prête à l'optimisation stochastique (décomposable en une somme), s'adapte à de grands ensembles de données et produit des politiques avec de meilleures performances empiriques. Cependant, ce principe, comme le CRM, souffre d'autres limitations, présentées ci-après :

(2) Pas de garanties de performance. Les deux principes dérivés sont motivés par la construction de bornes couvrant le risque réel des politiques. Ces bornes, dans leur forme brute, ne peuvent pas être utilisées directement comme objectif d'apprentissage. En effet, la borne dérivée dans Swaminathan and Joachims (2015a) contient des quantités théoriques et celle dérivée dans London and Sandler (2019) donne une couverture triviale. L'introduction de l'hyperparamètre λ permet d'obtenir des objectifs pratiques, qui perdent les garanties théoriques données par les bornes initiales. Ces objectifs ne couvrent pas nécessairement le risque réel, et leur optimisation peut conduire à des politiques pires que π_0. Des preuves empiriques peuvent être trouvées dans (Chen et al., 2019b) où le principe CRM ne parvient pas à améliorer π_0.

(3) Rajout d'hyperparamètre. Un autre problème majeur de ces principes est également causé par l'introduction de λ et sa sélection. Le paramètre libre λ nécessite un réglage minutieux, car son choix a un impact considérable sur les performances de la politique obtenue. Comme il n'existe pas de lignes directrices théoriques pour définir une bonne valeur de λ, la stratégie consiste à procéder à une validation croisée du paramètre sur une grille relativement fine en utilisant l'estimateur `IPS`. La validation croisée ajoute à la complexité de l'algorithme. Cette procédure nécessite également de disposer d'un ensemble de validation qui ne sera pas utilisé pour l'apprentissage, ce qui accentue le problème de la variance. En outre, comme l'estimateur `IPS` est toujours utilisé pour sélectionner la meilleure valeur de λ, la politique renvoyée à la fin est celle qui minimise le risque `IPS` sur l'ensemble de validation, ce qui rend l'ensemble du principe incohérent.

L'objectif de cette première partie est de fournir aux praticiens de meilleurs principes qui contournent complètement ces limitations, en bénéficiant de meilleures garanties statistiques et de performances empiriques.

Chapter 3. **Offline Learning with Distributionally Robust Optimization.** Dans ce chapitre, nous présentons une formulation alternative au principe CRM en recourant au cadre de l'optimisation distributionnellement robuste (DRO) (Duchi et al., 2021). Ces outils permettent de construire élégamment des intervalles de confiance sensibles à la variance sur le vrai risque en utilisant des ensembles d'ambiguïté basés sur la f-divergence. Nous appliquons ce principe au problème de l'évaluation et de l'optimisation des politiques hors ligne. L'objectif résultant traite des limites **(1)** et **(3)** ; il bénéficie des mêmes garanties statistiques que le CRM, peut être calibré automatiquement en utilisant des arguments de couverture asymptotique et se prête à l'optimisation stochastique. Nous présentons des expériences numériques solides montrant que l'approche proposée traite efficacement les lacunes de la CRM. Ce chapitre est adapté de la publication suivante :

- Otmane Sakhi, Louis Faury, and Flavian Vasile (2020b). Improving Offline Contextual Bandits with Distributional Robustness. *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems, 2020.*

Chapter 4. Offline Learning with PAC-Bayesian Theory. Dans ce chapitre, nous remettons complètement en question le paradigme de l'apprentissage hors politique et préconisons une stratégie théoriquement fondée pour améliorer avec certitude la politique déployée π_0. La méthode proposée consiste à créer des bornes inférieures de la quantité d'améliorations uivante $\mathcal{I}(\pi) = R(\pi_0) - R(\pi)$, et à déployer de nouvelles politiques uniquement lorsque nous sommes sûrs que $\mathcal{I}(\pi) > 0$. Nous basons notre approche sur la théorie de l'apprentissage PAC-Bayesien (Alquier, 2021) et démontrons que ses outils conviennent parfaitement au problème de l'apprentissage hors ligne. En particulier, en interprétant les politiques comme des mélanges de règles de décision, nous dérivons une borne PAC-Bayesienne étroite, de type Bernstein, qui rend notre stratégie viable. La stratégie résultante traite les trois limitations ; nous montrons que l'algorithme résultant peut donner des certificats d'amélioration, se prête à l'optimisation stochastique et ne nécessite aucun réglage d'hyperparamètre, ce qui constitue un grand pas en avant vers la réalisation d'un apprentissage hors politique pratique avec de véritables garanties de performance. Ce chapitre est basé sur la publication suivante :

- Otmane Sakhi, Pierre Alquier, and Nicolas Chopin (2023a). PAC-Bayesian Offline Contextual Bandits with Guarantees. *Proceedings of the 40th International Conference on Machine Learning, 2023.*

Chapter 5. A Better PAC-Bayesian Analysis of Offline Learning. Dans ce chapitre, nous poursuivons le développement de l'analyse PAC-Bayesienne du problème de l'apprentissage hors ligne des politiques. En exploitant la nature négative du risque, nous dérivons de nouvelles bornes plus étroites qui s'appliquent à une classe plus large d'estimateurs de risque. L'idée est basée sur un traitement raffiné de la fonction génératrice de moments du risque et étend les limites empiriques de Bernstein à des ordres supérieurs. La particularité de ces résultats est qu'ils sont entièrement empiriques ; nous ne supposons pas l'accès à π_0 contrairement aux bornes dérivées précédemment. Nous observons que nos résultats peuvent donner de meilleures garanties et nous permettent d'obtenir de nouvelles informations sur les estimateurs utilisés. Ce chapitre se concentre sur la fourniture de résultats techniques et est basé sur un travail non publié.

Part II - Offline Learning of Large Scale Recommendation. L'apprentissage hors ligne offre des solutions pratiques pour aligner efficacement les systèmes de décision sur des signaux de récompense complexes. Si la communauté des chercheurs s'est concentrée sur l'amélioration des estimateurs et des paradigmes d'apprentissage existants, peu d'attention a été accordée à l'adaptation de ces approches au contexte des grands espaces d'action. Ceci est intéressant pour les moteurs de recherche d'apprentissage, les systèmes de recommandation et pratiquement toutes les applications où le nombre d'interactions n et la taille de l'espace d'action $|\mathcal{A}|$ sont massifs. Le principal défi dans ces applications est de concevoir des règles de décision qui satisfont aux contraintes d'ingénierie, tout en fournissant des algorithmes pratiques qui permettent leur alignement avec les signaux de récompense d'une manière rapide et fiable. Les moteurs de recherche doivent répondre aux requêtes en quelques millisecondes, et les systèmes de recommandation du monde réel (pensez à une plateforme de streaming vidéo) doivent remplir de manière rapide la page d'accueil avec du contenu. Ces contraintes de vitesse doivent être respectées même si le catalogue (espace d'action) contient des milliards d'éléments. Un autre aspect à prendre en considération est que ces systèmes sont fréquemment mis à jour, ce qui impose une contrainte considérable sur le temps d'apprentissage de ces systèmes. En effet, si nous devons mettre à jour notre système de décision *quotidiennement* sur la base de ses interactions, le temps d'apprentissage devrait être nettement inférieur à un*jour* car il faut collecter suffisamment d'interactions et mettre à jour le système dans le même laps de temps. Dans leurs implémentations naïves, la prise de décision et l'apprentissage de ces systèmes sont linéairement proportionnels à la taille de l'espace d'action $\mathcal{O}(|\mathcal{A}|)$, ce qui n'est pas possible dans les scénarios d'espace d'action massif. Dans ce qui suit, nous développons la discussion autour de ces deux aspects importants et présentons nos contributions dans ce domaine.

Prise de décision rapide. La politique déployée permet de répondre à une requête ou de fournir des recommandations précises. Quelle que soit la nature de la politique et de sa mise en œuvre, cette étape se résume généralement à l'identification *rapide* d'un sous-échantillon de taille $K \geq 1$ de bonnes actions (Chen et al., 2019a) à partir de l'espace d'action potentiellement massif. En règle générale, et pour un utilisateur x, la qualité des actions est encodée dans la fonction de score $f_\theta(\cdot, x) : \mathcal{A} \to \mathbb{R}$ tandis que les bonnes actions sont identifiées en trouvant les actions ayant le meilleur score, en résolvant le problème suivant :

$$[a_1, ..., a_K] = \underset{a' \in \mathcal{A}}{\overset{K}{\arg\operatorname{sort}}} \left\{ f_\theta(a', x) \right\}, \tag{2}$$

avec l'opérateur $\arg\operatorname{sort}^K_{a' \in \mathcal{A}}$ qui renvoie les K actions les mieux notées. Cette opération de tri a une complexité linéaire sur la taille de l'espace des actions $\mathcal{O}(|\mathcal{A}| \log K)$ et ne peut pas être adoptée dans un environnement de production à grande échelle. La solution courante pour réduire cette complexité consiste à imposer une structure à la fonction de score. En limitant l'espace de la fonction de score à ce qui suit :

$$\forall (x, a) \quad f_\theta(a, x) = h_\Xi(x)^\intercal \beta_a$$

avec $\theta = [\Xi, \beta]$, la fonction de score devient un produit scalaire entre une transformation du contexte $h_\Xi(x)$ et une transformation de l'action β_a, tous deux résidant dans un espace latent $\mathbb{R}^l$ de dimension $l \ll |\mathcal{A}|$. Avec cette structure, l'équation (2) peut être résolue en approximant MIPS : Maximum Inner Product Search (Shrivastava and Li, 2014) dans une complexité temporelle de $\mathcal{O}(\log |\mathcal{A}|)$ au lieu de $\mathcal{O}(|\mathcal{A}|)$, ce qui rend possible une prise de décision rapide sans considérations supplémentaires.

Apprentissage efficace pour la méthode directe. La méthode directe dérive une politique optimale, qui identifie pour chaque contexte les actions ayant le meilleur score selon le modèle de récompense $r_\mathcal{M}$. Cela signifie que si $r_\mathcal{M}$ est correctement paramétré, une prise de décision rapide est possible. L'apprentissage d'un bon modèle $r_\mathcal{M}$ nécessite une excellente compréhension du problème sous-jacent et est généralement obtenue par *maximum de vraisemblance* (Aouali et al., 2023b) ou des *heuristiques de classement* (Rendle et al., 2009), qui sont des méthodes dont l'apprentissage est indépendant de la taille de l'espace d'action. Ces algorithmes peuvent toutefois présenter d'autres lacunes si nous ne prenons pas garde aux particularités de ce cadre.

Dans les scénarios à grand espace d'action, il est impossible pour la politique déployée de collecter suffisamment d'interactions pour chaque action dans $\mathcal{A}$. Le signal de récompense est inégalement réparti, car la majorité des données collectées proviennent d'actions très probables sous π_0 et peu ou pas de données sont disponibles pour le reste des actions. Cela signifie que si nous utilisons le *principe du maximum de vraisemblance*, la qualité du modèle appris $r_{\mathcal{M}}$ dépendra de la paire contexte/action ; l'estimation est précise pour les paires action/contexte qui sont suffisamment présentes dans les données. Ce déséquilibre dans la qualité de l'estimation a un impact négatif sur la politique dérivée, car les décisions basées sur le *MLE* peuvent souffrir d'une déception post-décisionnelle (Smith and Winkler, 2006). L'un des moyens d'atténuer ce problème est de l'inscrire dans le cadre de la théorie de la décision *Bayesienne* (West et al., 2021). Par exemple, Jeunen and Goethals (2021) démontre que même une simple modélisation bayésienne de la récompense permet d'améliorer le comportement des politiques. Avec l'aide de distribution à priori bien choisis, cette formulation peut également intégrer des corrélations supplémentaires entre les contextes et les actions, ce qui rend l'apprentissage encore plus efficace (Aouali et al., 2023c). Cependant, le principal défi de la modélisation bayésienne est d'ordre computationnel ; l'approximation des distributions à posteriori sur des milliards d'interactions, à l'aide de modèles, est difficile et nécessite un soin particulier (Chopin and Papaspiliopoulos, 2020). Nous consacrons un chapitre à cette discussion et construisons un modèle de récompense bayésien pour la recommandation en utilisant des à priori bien construit, tout en fournissant des outils appropriés pour accélérer son apprentissage dans des applications à grande échelle.

Chapter 6. Scalable Bayesian Reward Modelling. Dans ce chapitre, nous empruntons la voie de la méthode directe et développons un modèle bayésien de la récompense dans le cas de la recommandation d'un seul article. Nous reconnaissons la présence de deux types de signaux dans les problèmes de recommandation : les signaux organiques et les signaux de bandits. Alors que nous conditionnons notre modèle au retour bandit, les interactions organiques entre les contextes et les actions nous aident à construire une distribution a priori qui incorpore trois similarités : la similarité contexte-action, la similarité action-action et la similarité contexte-contexte. Ces similarités nous permettent d'obtenir de bonnes estimations de la récompense dans toutes les régions de l'espace, même pour les actions et les contextes les moins explorés. Le modèle proposé est flexible, utilise efficacement les données existantes mais produit une distribution à posteriori intraitable. Nous fournissons des outils computationnels faciles à mettre en œuvre pour approximer sa solution en nous basant sur des approches variationnelles (Blei et al., 2017). L'algorithme résultant s'adapte à de grands ensembles de données, peut apprendre efficacement dans différents scénarios et bénéficie de la paramétrisation du produit scalaire, ce qui permet une prise de décision rapide. Ce chapitre est basé sur la publication suivante :

- Otmane Sakhi, Stephen Bonner, David Rohde and Flavian Vasile (2020a). BLOB: A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*

Apprentissage rapide avec les estimateurs d'échantillonnage préférentiel Dans notre quête d'algorithmes évolutifs d'apprentissage de politiques hors ligne, nous exprimons également notre intérêt pour les paradigmes d'apprentissage basés sur l'échantillonnage préférentiel. Cette approche apprend une politique directement, et même les opérations simples impliquent le calcul de sommes sur l'ensemble de l'espace d'action. En particulier, nous devons être très prudents lorsque nous calculons/approximons les gradients de nos objectifs, car cette opération se calcule linéairement dans $|\mathcal{A}|$, ce qui peut ralentir considérablement la routine d'optimisation. La question de la mise à l'échelle des objectifs généraux d'apprentissage hors politique a attiré peu d'attention ; Chen et al. (2019a) a appris une politique prête pour la production avec un objectif basé sur IPS sans se préoccuper de l'aspect computationnel. Nous nous intéressons à cette question et souhaitons fournir des méthodes d'accélération générales. Nous étudions la famille spécifique d'objectifs qui peuvent être écrits comme des espérances sous la politique évaluée. Cette famille comprend les estimateurs couramment adoptés (Horvitz and Thompson, 1952; Dudík et al., 2014; Wang et al., 2017; Saito and Joachims, 2022a; Saito et al., 2023; Aouali et al., 2023a), et les nouveaux objectifs d'apprentissage (London and Sandler, 2019; Sakhi et al., 2023a). Nous commençons par étudier l'accélération de cette famille spécifique d'objectifs d'apprentissage et fournissons des procédures d'optimisation en temps logarithmique pour les politiques à article unique, en nous concentrant particulièrement sur les politiques paramétrées avec la fonction de lien softmax.

Chapter 7. Fast Offline Learning for One-Item Recommendation. Dans ce chapitre, nous nous attachons à fournir une méthode pour accélérer l'apprentissage des politiques de softmax à produit scalaire pour un large panel d'objectifs. Nous identifions les problèmes posés par les gradients couramment adoptés et proposons une solution basée sur trois ingrédients : une nouvelle formule de gradient de covariance, l'exploitation de la structure MIPS : Maximum Inner Product Search dans la phase d'apprentissage et la conception d'outils Monte Carlo appropriés (Chopin and Papaspiliopoulos, 2020) pour obtenir des approximations accélérées. Il en résulte un algorithme d'apprentissage avec des mises à jour de gradient sous-linéaires (logarithmiques ou constantes). Nous menons des expériences approfondies sur des ensembles de données de recommandation à grande échelle et démontrons l'impact de notre approche ; la méthode proposée est jusqu'à 25 fois plus rapide que la méthode de base tout en produisant des politiques de qualité similaire. Ce chapitre est basé sur la publication suivante :

- Otmane Sakhi, David Rohde, and Alexandre Gilotte (2023c). Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023.*

Après avoir abordé le problème de l'apprentissage de systèmes de décision à grande échelle à un élément avec des objectifs linéaires, nous étendons notre analyse au cas plus difficile de l'apprentissage de systèmes de décision à ardoise. Au lieu de jouer une action, nos politiques doivent délivrer des ardoises, une liste ordonnée d'éléments de taille $K \geq 1$. Cela signifie que nos règles de décision et nos politiques sont construites pour agir sur l'espace combinatoire $\mathcal{S}_K$ de permutations tronquées à K. La taille de cet espace est $\mathcal{O}(|\mathcal{A}|^K)$ et rend les opérations de base, du calcul d'une moyenne à la recherche de la meilleure ardoise, infaisables. Nous nous concentrons sur une famille de systèmes de décision qui réduisent l'espace de recherche de l'ensemble combinatoirement grand des ardoises $\mathcal{S}_K$ à l'espace d'action original $\mathcal{A}$. Pour un contexte donné x, cette réduction consiste à attribuer un score $f_\theta(a, x)$ à chaque action a et à recommander une liste composée des K premiers éléments ayant les scores les plus élevés. Cela conduit à un temps de livraison de $\mathcal{O}(\log |\mathcal{A}|)$ lorsque nous adoptons la structure de produit scalaire pour $f_\theta(a, x)$. Aouali et al. (2023b) propose une méthode directe pour apprendre les systèmes de recommandation d'ardoises à grande échelle. Dans le chapitre suivant, nous présentons les défis posés par l'apprentissage des systèmes de décision en ardoise et proposons des solutions pour accélérer leur apprentissage.

Chapter 8. Fast Offline Learning for Slate Recommendation. Dans ce chapitre, nous nous concentrons sur l'accélération de l'apprentissage des politiques d'ardoise, un élément omniprésent des systèmes en ligne modernes. Nous commençons par présenter le problème et par analyser les algorithmes existants, leurs hypothèses communes et leurs limites. Nous proposons ensuite une nouvelle classe d'algorithmes, basée sur une nouvelle relaxation qui traite élégamment les contraintes à grande échelle. La méthode résultante fonctionne avec des récompenses arbitraires, possède de meilleures propriétés statistiques tout en réalisant des mises à jour d'apprentissage sous-linéaires. Nous menons des expériences à grande échelle et démontrons que l'approche proposée est plus rapide de plusieurs ordres de grandeur que les lignes de base, tout en produisant des politiques plus performantes. Ce chapitre est basé sur la publication suivante:

- Otmane Sakhi, David Rohde, and Nicolas Chopin (2023b). Fast Slate Policy Optimization: Going Beyond Plackett-Luce. *Transactions on Machine Learning Research.*

Chapter 1

Introduction

1.1 Overview

This manuscript presents recent contributions, ranging from theory to large scale applications, to an offline formalism of the problem of sequential decision-making under uncertainty. An important problem with numerous real-world applications where a decision maker, tasked with solving a specific goal, interacts with an unknown environment, log these interactions and leverage them in order to better solve the task. In this context, we want to answer the following:

How can we leverage previous interactions of the decision-maker to improve its performance?

Answering this question can have a big impact on real world practical problems. For example, it may help a charity online marketing campaign get more donations for a good cause, it may be of service to doctors improving the quality of drug prescription, or it may simply improve the recommendation quality of your favourite music streaming service making it easier to discover new artists. In this introduction, we showcase the problem of learning decision-makers using the example of recommendation, as it will be the focus of a big part of this thesis. Recommender systems are the backbone of the modern internet experience. In each interaction, these systems silently navigate an overwhelming amount of information and filter it to cater to the specific needs of the user. An interaction of a recommendation engine can be summarized in the following: the system encounters a user, the system chooses an item (or multiple items) to recommend from a potentially large catalogue and observes a feedback.

The feedback received is valuable as it represents successes and failures of past interactions. These interactions are logged and are later used to improve the recommendation quality of the system. The interactive nature of the collected dataset makes common learning paradigms, such as supervised learning, not adapted to study such problem. Recently, there has been an interest in adapting sequential decision-making framework to improve recommendation based on the log of interactions. Reinforcement learning (RL) (Sutton and Barto, 2018) and Contextual bandits (CB) (Lattimore and Szepesvári, 2020) start to take the spotlight as good candidates to model this learning problem. The RL framework builds on the idea that performed actions may impact the environment. This paradigm can model complex sequential decision problems, is versatile and allows for planning. Its tools can optimize recommender systems for long term metrics; for example, increase user engagement and retention (Afsar et al., 2022). This versatility however comes with a cost. Taking into account the long term effects of recommendation on users makes the analysis more difficult, prompting us to consider a simpler formalism. Contextual Bandit offers a useful compromise between principled analysis and practical impact. Its underlying assumption is that actions made by the system do not influence future outcomes. If this

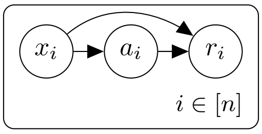

Figure 1.1: The logged dataset $\mathcal{D}_n$ representing n interactions of the recommender system. All (context, action, reward) triplets are independent.

formulation is less compelling when dealing with delayed rewards (Afsar et al., 2022), its use is reasonable if we want to focus on learning recommender systems that optimize short-term, action-bounded metrics, such as click-through rate (Sakhi et al., 2020a) or watch time (Chen et al., 2019a). In this thesis, we adopt the offline contextual bandits' (Bottou et al., 2013; Nguyen-Tang et al., 2022) toolbox to formalize learning from interaction logs. We give new principled approaches to learn policies with strong performance guarantees and propose new algorithms to widen the impact of this framework to large scale, real world applications.

An interaction of a user with a recommended item can be reduced to the following example. A user navigates a website, the recommender system chooses an item from a catalogue and shows it to the user, the user interacts with the item (either clicks or not) and the result of this interaction is encoded in a feedback (presence/absence of click) that the system logs. Within the Contextual Bandit framework, a user is represented by a context x, usually a real vector living in a d-dimensional space $\mathcal{X} \subseteq \mathbb{R}^d$. These contexts, and thus users, are sampled *independently* from the same, unknown distribution $\nu(\mathcal{X})$. After seeing a user, the recommendation engine delivers an item a from a catalogue $\mathcal{A}$ of size $|\mathcal{A}| \in \mathbb{N}$. The recommender system is modelled as a policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, which is a function that takes a context x and produces a distribution $\pi(\cdot|x)$ over the space of possible actions $\mathcal{A}$. Recommending an item a for the context x boils down to sampling the item from the produced distribution $a \sim \pi(\cdot|x)$. After delivering the item a to the user of context x, our system receives feedback; a stochastic reward $r \in \mathbb{R}^+$ coming from an unknown distribution $p(\cdot|x,a)$. This reward encodes how well the recommended item has performed on our desired metric; the higher the reward, the higher the performance. Our goal is to find policies of great performance, achieved by minimizing the risk, defined as the expected negative reward under the actions of the policy. The risk of any given policy π can be expressed as:

$$R(\pi) = -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\mathbb{E}_{r\sim p(\cdot|x,a)}[r]\right].$$

This risk is an expectation under actions taken by the policy evaluated. As we do not have access to interactions of the new policy π with the environment, a simple way to estimate this quantity is to let π interact with users online. In most scenarios, this is not possible, as we do not have the luxury to deploy bad policies. In real world applications, we already have the current version of our recommender system, represented by the policy π_0, that interacts with the environment and logs the feedbacks. Our primary focus is to assess how well a new iteration of the system will improve upon the currently deployed version. A common way to achieve this is by conducting online A/B-tests (Kohavi et al., 2012). This is considered the "gold standard" approach to estimate the effect of replacing the current policy π_0 by a potentially better one (Gupta et al., 2019). A/B-tests however require substantial engineering effort, constant monitoring and need several days to be properly analysed. Ideally, we want offline evaluation and learning tools that can give us promising policies to reduce the number of unnecessary A/B-tests. When the contextual bandit assumptions are satisfied, we can use the framework's toolbox to achieve this (Bottou et al., 2013). The idea is to leverage the existing interactions of π_0 to find policies of greater performance. The interaction dataset is called in the literature the logged bandit feedback dataset (Swaminathan and Joachims, 2015a):

$$\mathcal{D}_n = \{x_i \sim \nu, a_i \sim \pi_0(\cdot|x_i), r_i \sim p(\cdot|x_i, a_i), \pi_0(a_i|x_i)\}_{i \in [n]}.$$

A graphical representation of the data is shown in Figure 1.1. The main challenge encountered when learning from this data is the potential bias created by the collection procedure; we only have access to the outcome of actions sampled from π_0. The offline learning framework of contextual bandit offers two distinct approaches to solve this issue; the model-based approach and the importance weighting approach.

The model-based approach or *the direct method* leverages the interaction data $\mathcal{D}_n$ to construct a reward model (Sakhi et al., 2020a; Jeunen and Goethals, 2021). An optimal policy is then naturally derived by playing for each context x, the action with the highest reward according to the model. The direct method is straightforward to implement, as it reduces the learning to a regression problem (Brandfonbrener et al., 2021). This approach is theoretically well-studied and benefits from strong guarantees (Nguyen-Tang et al., 2022). However, it will suffer from a substantial, uncontrolled bias whenever the reward is complex, making its efficiency entirely dependent on our ability to model the problem's structure. The direct method is efficient when we are confident in our ability to understand the problem. When the reward signal is complex, we may prefer another approach that does not completely rely on our modelling effort.

The Importance-weighting approach (Horvitz and Thompson, 1952; Bottou et al., 2013; Dudík et al., 2014), often called *off-policy learning*, is agnostic to the reward model. It learns a new policy π directly from the interactions $\mathcal{D}_n$ using estimators corrected with importance sampling (Chopin and Papaspiliopoulos, 2020). Under mild assumptions (Horvitz and Thompson, 1952), this method can produce unbiased estimators, which are arguably easier to analyse and optimize (Ajalloeian and Stich, 2020). These estimators however suffer from a potentially large variance once the learned policy drifts away from the logging policy π_0, making them unreliable for learning. It is empirically proven that learning with these estimators can result in bad performing policies (Swaminathan and Joachims, 2015a,b), sometimes even worse than π_0 (Chen et al., 2019b; London and Sandler, 2019). This observation motivates the use of learning theory tools (Zhou, 2002; McAllester, 1998) to come up with principled objectives that are agnostic to the reward structure. The objective of this research effort is to produce new policies that are **provably better** than the logging policy without engaging with the environment. This is beneficial in production settings where we would like to propose a new system, that will improve on the current production system with high probability.

The first effort in this direction was driven by Swaminathan and Joachims (2015a) and resulted in the **CRM: Counterfactual Risk Minimization** principle. The **CRM** principle builds on tools from Statistical Learning Theory (Vapnik, 1998), a framework that has great success studying the generalization ability of learning algorithms. Motivated by the construction of an Empirical Bernstein Upper Bound (Maurer and Pontil, 2009) on the true risk of policies, and using covering number arguments (Zhou, 2002), this principle advocates for penalizing importance weights estimators with their *square-root sample variance.* This penalty is controlled by a hyperparameter λ that needs to be cross-validated on a hold-out set. The underlying intuition is that to improve on the logging policy, we should look for policies that have a *small empirical risk* while staying close to the logging policy π_0. This principle results in better performing policies compared to optimizing crude importance weighting estimators (Swaminathan and Joachims, 2015a,b). However, its learning paradigm suffers from different limitations, hindering its applicability to simple scenarios. In particular, adding the sample variance penalization makes the learning objective non-convex and non-decomposable, which forbids the use of stochastic gradient methods. This penalty is also controlled with a new hyperparameter λ that is difficult to tune and adds to the complexity of the approach. Finally, the **CRM** principle fails to provide performance certificates on the newly trained policy. These limitations will be discussed in details later in the introduction. More recently, a new principle was introduced to mitigate some of these limitations. By analysing this learning problem from the PAC-Bayesian lens (McAllester, 1998; Alquier, 2021), London and Sandler (2019) develop an improved approach. The authors build their analysis around McAllester (2003)'s PAC-Bayesian bound. For parametric policies, this motivates an L_2 regularization of the parameter of the new policy towards the parameter of the logging policy π_0. The regularization is also controlled by a hyperparameter λ that requires tuning. This principle is based on the same intuition of staying close to π_0, but this time, it is carried out on the parameter space. The adoption of an L_2 regularization instead of a *sample variance penalization* makes the optimization smoother and allows the use of stochastic gradient descent. However, the L_2 regularization parameter λ suffers from the same limitations and the principle cannot produce provably better policies. Empirical findings demonstrate that these principles sometimes fail at improving the logging policy π_0 (Chen et al., 2019b). These limitations will be developed even further in the next section, before we present the contributions of the first part of the thesis. Chapter 3 reframes **CRM** using tools from **Distributionally Robust Optimization** (Duchi et al., 2021), a statistical framework designed for decision-making under uncertainty. Furthermore, Chapters 4 and 5 build on London and Sandler (2019)'s work and continue the development of **PAC-Bayesian** tools (McAllester, 1998) for offline contextual bandit. The analysis yields principles that are easier to optimize, do not require additional hyperparameters to tune and enjoy, for some, even better performance guarantees, taking us a step closer to learn **provably better** policies offline.

In real world problems, interactive systems often deal with large scale scenarios, where they need to learn from enormous number of interactions ($n \gg 1$) and operate on massive catalogues ($|\mathcal{A}| \gg 1$). For these systems to deliver recommendations in a matter of milliseconds, they are restricted to a certain structure (Shrivastava and Li, 2014; Aouali et al., 2022) to allow for rapid query response. For a long time, large scale recommender systems were trained for preference prediction (Harper and Konstan, 2015; Gomez-Uribe and Hunt, 2016) or next-item prediction (Hidasi et al., 2015; Wu et al., 2019). These modelling approaches are usually considered poor proxies to the reward we are interested to optimize (Jannach and Jugovac, 2019). Adapting the offline contextual bandit toolbox to learn large scale recommender systems will have a great impact on the industry. These tools can enable the alignment of recommendation with complex reward signals, enhancing both user satisfaction and the profitability of the businesses operating these systems. As presented earlier, we can either adopt the **direct method** if we know how to model the reward, or **importance weighting estimators with learning principles** to learn a policy directly. Both can reliably learn a performing recommender system. Unfortunately, these methods in their simple form present some caveats when dealing with large scale problems. The second part of the thesis addresses these limitations and allows for efficient and fast training of reward optimizing, large scale recommender systems.

The direct method relies completely on our ability to learn a model that reflects the properties of the reward. Understanding perfectly the problem lowers the bias linked to modelling, but there is another problem, linked to learning from $\mathcal{D}_n$, that becomes more pronounced in large catalogue scenarios. Indeed, naively learning the reward model using $\mathcal{D}_n$ suffers from the unbalance present in the collected data. The reward model will be well estimated for actions that are likely to be sampled under π_0, and poorly estimated for the rest. This difference in the estimation quality can make the decisions taken by the derived policy unreliable (Smith and Winkler, 2006). This phenomenon is accentuated when dealing with large catalogue sizes, as π_0 can never collect enough samples to cover the whole action space. We dedicate Chapter 6 to discuss a Bayesian solution to this issue. We introduce structure to the model and use another valuable source of data to efficiently learn the reward model. More details about this approach can be found in the contribution section.

Importance-weighting objectives become interesting when the reward signal is complex. However, in large scale scenarios, these learning objectives suffer from two major caveats. The first issue is linked to the variance of these importance-weighting estimators, which grows with the size of the action space. Indeed, the variance of common importance weighting estimators (Horvitz and Thompson, 1952; Ionides, 2008; Dudík et al., 2014) become uncontrollable when the policies operate on massive catalogues (Saito and Joachims, 2022b). As this variance can be very large, adding a variance penalization for example will force the newly learned policy π to mimic the behaviour of π_0. This phenomenon makes our learning principles too conservative, returning policies very close to π_0. This observation motivated the construction of a new family of importance weighting estimators (Saito and Joachims, 2022a; Saito et al., 2023) to mitigate this variance problem. These recent contributions deal with the statistical limitations of importance-weighting objectives in large catalogue scenarios, but computational issues linked to optimizing these objectives remain unsolved. The importance weighting approach learns policies directly, and use gradient-based methods to computationally optimize the learning objectives. Large scale systems are updated frequently, and fast optimization routines are highly desirable in this context. Existing methods offer gradient iterations that scale at least linearly on the catalogue size. This complexity can be detrimental to learning recommender systems operating on billions of items. The last two chapters (Chapters 7 and 8) focus on the computational aspect and propose optimization routines with sublinear complexities. These solutions will be developed more in the contribution section.

We cover different, connected disciplines in this thesis while balancing between theoretical tools and practical algorithms. To ease the presentation, we want to give readers an overview of the advancement of each research field. To this end, we dedicate a chapter to review existing literature, that we deem valuable to researchers, whether they come from a theoretical or practical background.

Chapter 2. Literature Review. This chapter conducts a literature review to cover the different tools used throughout this thesis. We give a brief overview of the literature of Contextual Bandit, a practical formalism to study reward-driven recommendation, presenting both its online and offline formulations. With a focus on the offline setting, we dedicate a section to present statistical learning tools, necessary to study learning decision systems with online performance guarantees. We then present the development of recommender systems and how modelling recommendation shifted from predicting preferences to reward maximization, and conclude with the algorithmic considerations that arise in the context of large scale decision-making.

Part I - Offline Learning with Performance Guarantees. The first part of the thesis focuses on addressing the limitations of current learning principles. These principles were proposed to allow learning policies that improve on the logging policy π_0, mitigating the problems of importance weighting approaches. Without any loss of generality, we present the problem with the help of the IPS: *Inverse Propensity Scoring* estimator (Horvitz and Thompson, 1952), arguably the simplest and most studied estimator. For a policy π, we recall its expression:

$$\hat{R}_n^{\texttt{IPS}}(\pi) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i.$$

When evaluated on π_0, `IPS` gives the empirical mean of the collected costs as an estimation of the risk, which is considered to be a well-behaved, unbiased estimator of $R(\pi_0)$. However, a simple analysis of the variance of this estimator demonstrates that drifting away from π_0 leads to poorer estimation quality. If the observed reward is bounded (i.e. $r \in [0,1]$), we have:

$$\begin{aligned}\mathbb{V}\left[\hat{R}_n^{\texttt{IPS}}(\pi)\right] &= \frac{1}{n}\left(\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x), r\sim p(\cdot|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2 r^2\right] - \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right]^2\right)\\ &\leq \frac{1}{n}\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x), r\sim p(\cdot|x,a)}\left[\left(\frac{\pi(a|x)}{\pi_0(a|x)}\right)^2\right] = \frac{1}{n}\left(\chi^2(\pi,\pi_0)+1\right),\end{aligned}$$

with $\chi^2(\pi,\pi_0)$ the χ-Square divergence between π and π_0. The variance has roughly the same behaviour as the χ-Square divergence, growing when π is far from π_0. In particular, the variance grows with the importance weights. Importance weights are very large when the new policy π assigns high probability to actions that were very unlikely to be played under π_0. This means that the estimation quality is highly dependent on the policy evaluated π, making `IPS`[1] untrustworthy for policies far from the neighbourhood of π_0. This observation is confirmed in practice, especially when using importance weights-based estimators as a learning objective. For example, minimizing the `IPS` estimator with respect to a policy class can lead to policies with bad online performance. When the learned policy π is far from π_0, the `IPS` estimation of the risk of π will not reflect its true risk, as π will lie in a part of the space that induces an estimator with large variance. To circumvent these limitations, one would want to restrict the optimization to policies around the logging policy π_0. The **CRM: Counterfactual Risk Minimization** Principle (Swaminathan and Joachims, 2015a) formalizes this idea using statistical learning arguments. Motivated by the construction of an Empirical Bernstein Bound (Maurer and Pontil, 2009) covering the true risk of policies in a class of policies, the principle advocates for minimizing a *sample variance penalized* `IPS` estimator. This principle is then used to produce a tractable algorithm for learning, parametrized softmax policies (Mei et al., 2020b) of the form:

$$\begin{aligned}\forall(x,a) \quad \pi_\theta(a|x) &= \texttt{softmax}_{\mathcal{A}}\left(f_\theta(x,a)\right)\\ &= \frac{\exp(f_\theta(x,a)}{\sum_{a'\in\mathcal{A}}\exp(f_\theta(x,a'))}.\end{aligned} \tag{1.1}$$

with θ a parameter coming from a parametric space $\theta \in \Theta$ and $f_\theta : \mathcal{X}\times\mathcal{A}\to\mathbb{R}$ a function that encodes the relevance of action a to the context x. The algorithm proposed is named **POEM**: Policy Optimizer for Exponential Models (Swaminathan and Joachims, 2015a) and solves the following objective for softmax policies:

$$\arg\min_{\theta\in\Theta}\left\{\hat{R}_n^{\texttt{IPS}}(\pi_\theta) + \lambda\sqrt{\hat{V}^{\texttt{IPS}}(\pi_\theta)}\right\},$$

with λ a tuning parameter usually set with the help of a validation split, and $\hat{V}^{\texttt{IPS}}(\pi_\theta)$ the sample variance term induced by evaluating π with `IPS`:

$$\hat{V}^{\texttt{IPS}}(\pi_\theta) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i + \hat{R}_n^{\texttt{IPS}}(\pi_\theta)\right)^2.$$

[1] All importance weights based estimators suffer from the same caveat.

Swaminathan and Joachims (2015a) reported empirically the superiority of this principle; the policies returned by **POEM** have much lower risk than those obtained by naively minimizing the IPS objective. Adding the sample variance regularizer makes the approach more principled, but still suffers from limitations that reduce its applicability in real-life scenarios:

(1) Scalability. The biggest limitation of the CRM principle is its scalability to large logged datasets $\mathcal{D}_n$. The presence of the sample variance term makes computing the gradient of the CRM objective scale in $\mathcal{O}(n)$ in both computational and memory cost, as it requires going through the entire dataset. In a recommender system scenario, millions of interactions are logged daily. Such applications deal with extremely large number of samples n and cannot afford the cost of these computations. This issue is usually solved by resorting to stochastic optimization algorithm, which requires only access to unbiased stochastic gradients of the objective. Those are particularly easy to obtain when the objective decomposes into a sum over the dataset's entries, as computing the sum on batches of the dataset is enough to obtain unbiased gradients. Unfortunately, the CRM objective is not suited for stochastic optimization, as the square-root empirical variance term does not write as a sum. Swaminathan and Joachims (2015a) proposed a relaxation of the CRM objective, based on a majorization-minimization strategy, that can benefit *partially* from stochastic gradients. Their approach still requires passing through the whole logged dataset once in a while, obtaining a procedure of the same computational complexity.

London and Sandler (2019) propose an improved principle that deals with the first limitation of CRM. Instead of relying on Maurer and Pontil (2009)'s Empirical Bernstein Bound, London and Sandler (2019) adapts McAllester (2003)'s PAC-Bayesian bounds to derive learning objectives for this problem. The derived principle motivates the use of an L_2 regularization towards the parameter θ_0 of the logging policy π_0. This regularization is controlled by a hyperparameter λ, giving the following optimization problem for parametrized softmax policies:

$$\arg\min_{\theta \in \Theta} \left\{ \hat{R}_n^{\texttt{IPS}}(\pi_\theta) + \lambda \|\theta - \theta_0\|^2 \right\},$$

The optimization objective is amenable to stochastic optimization (decomposable into a sum), scales to large datasets and returns policies with better empirical performance. However, this principle, like CRM, suffers from other limitations, presented in the following:

(2) No Performance Guarantees. Both principles derived are motivated by the construction of bounds covering the true risk of policies. These bounds in their raw form cannot be used directly as a learning objective. Indeed, the bound derived in Swaminathan and Joachims (2015a) contains an intractable quantity and the one derived in London and Sandler (2019) is vacuous. Introducing the hyperparameter λ helps us obtain practical objectives, that lose the theoretical guarantees given by the initial bounds. These objectives do not necessarily cover the true risk, and optimizing them can lead to policies worse than the logging π_0. Empirical evidence can be found in (Chen et al., 2019b) where the CRM principle fails to improve on π_0.

(3) Hyper-parameter Selection. Another major problem of these principles is also caused by the introduction of λ and it is selected. The free-parameter λ requires careful tuning, as its choice drastically impacts the performance of the obtained policy. As there are no theoretical guidelines to define a good value of λ, the strategy consists of cross-validating the parameter over a relatively fine grid using the IPS estimator. Cross validation adds to the complexity of the algorithm. This procedure also requires having a hold-out set that will not be used for training, accentuating the variance problem. In addition, as the IPS estimator is still used to select the best value of λ, the policy returned in the end is the one that minimizes the IPS risk on the validation set, which renders the whole principle incoherent.

The goal in this first part is to provide practitioners with better principles that circumvent such limitations altogether, enjoying better statistical guarantees and empirical performance.

Chapter 3. **Offline Learning with Distributionally Robust Optimization.** In this chapter, we present an alternative formulation to the CRM principle by resorting to the distributionally robust optimization (DRO) framework (Duchi et al., 2021). These tools enable elegant construction of variance-sensitive confidence upper-bounds on the true risk by using f-divergence based ambiguity sets. We apply this principle to the problem of offline policy evaluation and optimization. The resulting objective deals with limitations **(1)** and **(3)**; it enjoys the same statistical guarantees than CRM, can be automatically calibrated using asymptotic coverage arguments and is amenable to stochastic optimization. We display strong numerical experiments showing that the proposed approach effectively deals with the shortcomings of CRM. This chapter is adapted from the following publication:

- Otmane Sakhi, Louis Faury, and Flavian Vasile (2020b). Improving Offline Contextual Bandits with Distributional Robustness. *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems, 2020.*

Chapter 4. **Offline Learning with PAC-Bayesian Theory.** In this chapter, we question the off-policy learning paradigm completely and advocate for a theoretically-grounded strategy to confidently improve on the deployed policy π_0. The proposed method revolves around creating tight, empirical lower bounds on the improvement $\mathcal{I}(\pi) = R(\pi_0) - R(\pi)$, and deploying new policies only when we are confident of $\mathcal{I}(\pi) > 0$. We base our approach on PAC-Bayesian learning theory (Alquier, 2021) and demonstrate that its tools suit perfectly the problem of off-policy learning. In particular, by interpreting policies as mixtures of decision rules, we derive a tight, Bernstein-type PAC-Bayes bound that makes our strategy viable. The resulting strategy deals with all three limitations; we show that the resulting algorithm can give improvement certificates, is amenable to stochastic optimization and does not require any hyperparameter tuning, making a big step towards achieving practical off-policy learning with true performance guarantees. This chapter is based on the following publication:

- Otmane Sakhi, Pierre Alquier, and Nicolas Chopin (2023a). PAC-Bayesian Offline Contextual Bandits with Guarantees. *Proceedings of the 40th International Conference on Machine Learning, 2023.*

Chapter 5. **A Better PAC-Bayesian Analysis of Offline Learning.** In this chapter, we continue the development of the PAC-Bayesian analysis of the problem of offline policy learning. By exploiting the negative nature of the risk, we derive new, tighter bounds that hold for a larger class of risk estimators. The idea is based on a refined treatment of the moment generating function of the risk and extend empirical Bernstein bounds to higher orders. The particularity of these results is that they are fully empirical; we do not assume access to π_0 contrary to previously derived bounds. We observe that our findings can give better guarantees and allow us to derive new insight about the estimators used. This chapter is focused on providing technical results and is based on new, unpublished work.

Part II - Offline Learning of Large Scale Recommendation. The offline learning setting provides practical solutions to efficiently align decision systems with complex reward signals. If the research community has focused on improving the existing estimators and learning paradigms, little attention was directed towards adapting these approaches to the large action space setting. This is of interest to learning search engines, recommender systems and practically any application where the number of interactions n and the size of the action space $|\mathcal{A}|$ are massive. The main challenge in these applications is to design decision rules that satisfy engineering constraints, while providing tractable algorithms that enable their alignment with reward signals in a fast and reliable manner. Search engines must answer queries in a matter of milliseconds, and real-world recommender systems (think of a video streaming platform) must seamlessly fill the landing page with content the user may like. These delivery speed constraints should be respected even if the catalogue (action space) contains billions of items. Another aspect to take into consideration is that these systems are updated frequently, putting a considerable constraint on the training time of such systems. Indeed, if we need to update our decision system *daily* based on its interactions, the training time should be substantially smaller than a *day* as you need to collect enough interactions and update the system in the same time frame. In their naive implementations, both the decision-making and training of these systems scale linearly in the size of the action space $\mathcal{O}(|\mathcal{A}|)$, which cannot be allowed in massive action space scenarios. In the following, we develop the discussion around these two important aspects and present our contributions in this field.

Fast Decision Making. Answering a query or delivering accurate recommendations is performed by the policy deployed. No matter the nature of the policy and its delivery, this step generally boils down to the *fast* identification of a sub-sample of size $K \geq 1$ of good actions (Chen et al., 2019a) from the potentially massive action space. As a general rule, and for a user x, the quality of the actions is encoded in the score function $f_\theta(\cdot, x) : \mathcal{A} \to \mathbb{R}$ while the good actions are identified by finding the best scoring actions, solving the following:

$$[a_1, ..., a_K] = \underset{a' \in \mathcal{A}}{\overset{K}{\arg\mathrm{sort}}} \left\{ f_\theta(a', x) \right\}, \tag{1.2}$$

with the operator $\arg\mathrm{sort}^K_{a' \in \mathcal{A}}$ returning the K highest scoring actions. This sorting operation has a linear complexity on the size of the action space $\mathcal{O}(|\mathcal{A}| \log K)$ and cannot be adopted in a large scale production environment. The common solution to reduce this complexity is to impose a structure for the score function. By restricting the score function space to the following:

$$\forall (x, a) \quad f_\theta(a, x) = h_\Xi(x)^\intercal \beta_a$$

with $\theta = [\Xi, \beta]$, the score function becomes an inner product between a context embedding $h_\Xi(x)$ and an action embedding β_a, both residing in a latent space $\mathbb{R}^l$ of dimension $l \ll |\mathcal{A}|$. With this structure, Equation (1.2) can be solved with *approximate* MIPS: Maximum Inner Product Search algorithms (Shrivastava and Li, 2014) in a time complexity of $\mathcal{O}(\log |\mathcal{A}|)$ instead of $\mathcal{O}(|\mathcal{A}|)$, rendering fast decision-making possible without additional considerations.

Efficient training with the direct method. The direct method derives an optimal policy that depends on identifying the best scoring actions according to the reward model $r_\mathcal{M}$. This means that if $r_\mathcal{M}$ has the proper parameterization, fast decision-making is possible. Training a good model $r_\mathcal{M}$ requires an excellent understanding of the underlying problem and is usually achieved through *maximum likelihood estimation* (Aouali et al., 2023b) or *ranking-based heuristics* (Rendle et al., 2009), which are methods that scale independently of the action space size.

These training algorithms however can have other shortfalls if we are not careful about the particularities of this setting. In large action space scenarios, it is impossible for the deployed policy to collect enough interactions for each action in $\mathcal{A}$. The reward signal is unevenly distributed, as the majority of data collected comes from actions that are highly likely under π_0 and little to no data is available for the rest of the actions. This means that if we use the *maximum likelihood principle*, the quality of the learned model $r_{\mathcal{M}}$ will depend on the context/action pair; the estimate is precise for action/context pairs that are present enough in the data. This unbalance in the estimate quality negatively impacts the policy derived, as acting based on the *MLE* might suffer from post-decision disappointment (Smith and Winkler, 2006). One principled way to mitigate this issue is to frame the whole problem within the lens of *Bayesian* decision theory (West et al., 2021). For example, Jeunen and Goethals (2021) demonstrate that even simple Bayesian modelling of the reward result in better behaved policies. With the help of well-chosen priors, this formulation can also incorporate additional correlations we have between contexts and actions, making learning even more efficient (Aouali et al., 2023c). However, the main challenge of Bayesian modelling is computational; approximating posteriors over billions of interactions, using complex models and priors is difficult and needs particular care (Chopin and Papaspiliopoulos, 2020). We dedicate a chapter to develop this discussion, and construct a Bayesian reward model for recommendation with strong, data-driven priors while giving proper tools to accelerate its training in large scale applications.

Chapter 6. Scalable Bayesian Reward Modelling. In this chapter, we take the path of the direct method and develop a bayesian model of the reward for the case of one-item recommendation. We acknowledge the presence of two types of signals in recommendation problems; the organic and bandit signals. While we condition our model on the the bandit feedback, the organic interactions between the contexts and actions help us construct a novel prior that incorporates three similarities: the context-action similarity, the action-action similarity and the context-context similarity. These similarities allow us to obtain good estimates of the reward on all regions of the space, even for less explored actions and contexts. The proposed model is flexible, efficiently uses the existing data but produces an intractable posterior. We provide easy-to-implement computational tools to approximate its solution based on ideas from Variational Bayes (Blei et al., 2017). The resulting algorithm scales to large datasets, can learn efficiently in different scenarios and benefits from the inner-product parametrization, allowing fast decision-making. This chapter is based on the following publication:

- Otmane Sakhi, Stephen Bonner, David Rohde and Flavian Vasile (2020a). BLOB: A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*

Fast training with importance-weighting methods. In our pursuit for scalable offline policy learning algorithm, we also express interest for importance weighting based learning paradigms. This approach learns a policy directly, and even simple operations involve computing sums over the whole action space. In particular, we need to be extra-careful when computing/approximating gradients of our objectives because this operation scales linearly in $|\mathcal{A}|$ which can drastically slow down the optimization routine. The question of scaling general off-policy learning objectives attracted little attention; Chen et al. (2019a) learned a production-ready policy with an `IPS`-based objective without any focus on the computational aspect. We are interested in this question and want to provide general acceleration methods. We study the specific family of objectives that can be written as expectations under the policy evaluated. This family include commonly adopted estimators (Horvitz and Thompson, 1952; Dudík et al., 2014; Wang et al., 2017; Saito and Joachims, 2022a; Saito et al., 2023; Aouali et al., 2023a), and principled learning objectives (London and Sandler, 2019; Sakhi et al., 2023a). We first study the acceleration of this specific family of learning objectives and provide logarithmic time optimization procedures for single-item policies, focusing particularly on policies parameterised with the softmax link function.

Chapter 7. Fast Offline Learning for One-Item Recommendation. In this chapter, we focus on providing a principled way to accelerate the learning of inner-product softmax policies for a large panel of off-policy objectives. We identify the problems of commonly adopted gradients and propose a solution based on three ingredients; a new covariance gradient formula, exploiting the `MIPS`: Maximum Inner Product Search structure in the training phase and designing proper Monte Carlo tools (Chopin and Papaspiliopoulos, 2020) to achieve accelerated approximations. This results in a training algorithm with sub-linear (logarithmic or constant) gradient updates. We conduct extensive experiments on large scale recommendation datasets and demonstrate the impact of our approach; the proposed method is up to 25 times faster than the baseline while producing trained policies of similar quality. This chapter is based on the following publication:

- Otmane Sakhi, David Rohde, and Alexandre Gilotte (2023c). Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023.*

After attacking the problem of training large scale one-item decision systems with linear objectives, we extend our analysis to the more challenging case of training slate decision systems. Instead of playing one action, our policies need to deliver slates; an ordered list of items of size $K \geq 1$. This means that our decision rules and policies are constructed to act on the combinatorial space $\mathcal{S}_K$ of K-truncated permutation. The size of this space is $\mathcal{O}(|\mathcal{A}|^K)$ and makes basic operations, from computing an average to searching for the best slate infeasible. We focus on a family of decision systems that reduce the search space from the combinatorially large set of slates $\mathcal{S}_K$ to the original action space $\mathcal{A}$. For a given context x, this reduction consists of assigning a score $f_\theta(a, x)$ to each action a and recommend a slate composed of the top-K items with the highest scores. This leads to a $\mathcal{O}(\log |\mathcal{A}|)$ delivery time when we adopt the inner-product structure for $f_\theta(a, x)$. Aouali et al. (2023b) suggest a direct method approach to learn large scale slate recommendation systems. In the next chapter, we present the challenges of learning slate decision systems and propose solutions to accelerate their training.

Chapter 8. Fast Offline Learning for Slate Recommendation. In this chapter, we focus on accelerating the learning of slate policies, a ubiquitous building block of modern online systems. We begin by introducing the problem and analysing the existing algorithms, their common assumptions and limitations. We then propose a new class of algorithms, based on a novel relaxation that deals elegantly with the large scale constraints. The resulting method works with arbitrary rewards, has better statistical properties while achieving sub-linear training updates. We conduct large scale experiments and demonstrate that the proposed approach is orders of magnitude faster than the baselines while resulting in better performing policies. This chapter is based on the following publication:

- Otmane Sakhi, David Rohde, and Nicolas Chopin (2023b). Fast Slate Policy Optimization: Going Beyond Plackett-Luce. *Transactions on Machine Learning Research.*

1.2 List of Publications

Journal/Conference Articles

1. Otmane Sakhi, David Rohde and Nicolas Chopin (2023b). Fast Slate Policy Optimization: Going Beyond Plackett-Luce. *Transactions on Machine Learning Research.*

2. Otmane Sakhi, Pierre Alquier and Nicolas Chopin (2023a). PAC-Bayesian Offline Contextual Bandits with Guarantees. *Proceedings of the 40th International Conference on Machine Learning, 2023.*

3. Otmane Sakhi, David Rohde and Alexandre Gilotte (2023c). Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023.*

4. Otmane Sakhi, Stephen Bonner, David Rohde and Flavian Vasile (2020a). BLOB: A Probabilistic Model for Recommendation that Combines Organic and Bandit Signals. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.*

Workshop Articles

1. Otmane Sakhi, Louis Faury and Flavian Vasile (2020b). Improving Offline Contextual Bandits with Distributional Robustness. *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems, 2020.*

2. Otmane Sakhi, Stephen Bonner, David Rohde, Flavian Vasile (2019). Reconsidering Analytical Variational Bounds for Output Layers of Deep Networks. *4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.*

Preprints

1. Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, Flavian Vasile (2023b). Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation. *arXiv preprint.*

Tutorials

1. Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, Maxime Vono (2022). Reward Optimizing Recommendation using Deep Learning and Fast Maximum Inner Product Search. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

2. Flavian Vasile, David Rohde, Olivier Jeunen, Amine Benhalloum, and Otmane Sakhi (2021). Recommender Systems Through the Lens of Decision Theory. *Companion Proceedings of the Web Conference 2021.*

3. David Rohde, Flavian Vasile, Sergey Ivanov, Otmane Sakhi (2020). Bayesian Value Based Recommendation: A modelling based alternative to proxy and counterfactual policy based recommendation. *Proceedings of the 14th ACM Conference on Recommender Systems.*

Chapter 2

Literature Review

2.1 The Landscape of Contextual Bandit

2.1.1 The Online Setting

A (stochastic)[1] contextual bandit is a powerful sequential decision-making framework where an agent interacts with an unknown environment for $T \in \mathbb{N}^*$ rounds. This environment provides contexts (user information, web page, etc) and a set of available actions $\mathcal{A}$ that our agent can make. In each round, the agent observes a context $x \in \mathcal{X}$, acts by taking an action a and receives a feedback; a reward $r \in \mathbb{R}^+$ that depends on both the action and the context observed, coming from a *fixed, but unknown distribution*. The particularity of this setting compared to classical supervised learning is that we observe partial feedback; we get access to the reward associated with the context and the action made by the agent and nothing more. Formally, for each round $t \in [T]$:

- The environment reveals a context $x_t \in \mathcal{X}$ coming from an unknown distribution ν.
- The agent acts on the context x_t by making action a_t. The agent is represented by a stochastic policy $\pi_t : \mathcal{X} \to \mathcal{P}(\mathcal{A})$, that given the context x_t, defines a probability distribution $\pi_t(\cdot|x_t) \in \mathcal{P}(\mathcal{A})$ over the space of available actions $\mathcal{A}$. Acting boils down to sampling from the policy given the context x_t; $a_t \sim \pi_t(\cdot|x_t)$.
- Making action a_t for the context x_t reveals a reward $r_t \in \mathbb{R}^+$ coming from an unknown distribution $r_t \sim p(\cdot|a_t, x_t)$.
- The feedback received r_t updates the policy π_t.

Every interaction helps the agent learn about the environment and improves it *online* to better act in the future. The contextual bandit framework is flexible and can model various problems. However, it is noteworthy to point out that it relies on the fundamental assumption that **the problem is stateless**: actions made by the agent do not affect the environment; for each round t, both contexts and rewards are drawn i.i.d. as the action a_t does not influence ν. This makes contextual bandit not suitable for problems that require long-term planning, for which we can use the more general framework of Reinforcement Learning. We direct the reader to Sutton and Barto (2018) for a great introduction to the field. With these assumptions in mind, we want our agent to achieve a goal that the practitioner is interested in. Depending on the application, we are interested in either maximising the expected cumulative reward after

[1] Different from the adversarial setting.

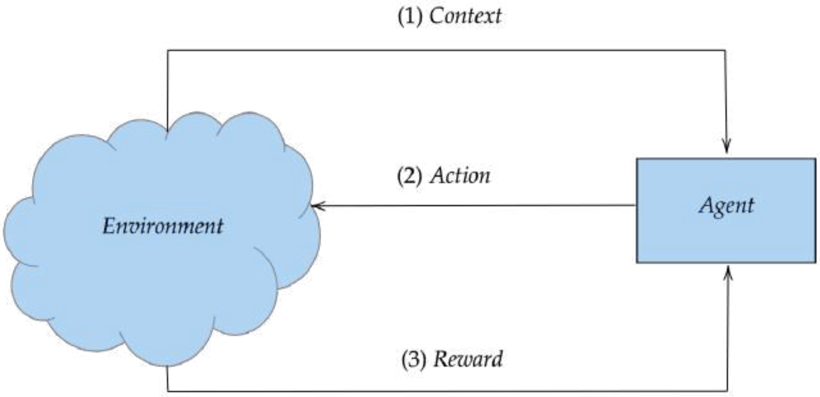

Figure 2.1: A Simple Illustration of the contextual bandit framework. One interaction consists of the environment revealing a context, the agent acting on the context and receiving a reward.

T rounds (when playing actions is costly, think about running Ad campaigns, when one does not own the display space and needs to pay for it) or identify the best arms given a confidence tolerance or a fixed interactions budget (when playing actions has little to no cost, think about a *casino owner* wanting to identify slot machines with high payouts to get rid of them).

Regret Minimisation. The regret of the agent (Auer et al., 2002) is defined as the gap between the highest expected cumulative reward (achieved by an optimal policy) and the cumulative reward the agent actually obtains after T round. Maximising the cumulative reward is equivalent to minimising the regret, the latter quantity however is better suited to theoretically compare the strategies to the best attainable outcome. Figure 2.2 illustrates the regret of a bandit strategy. Algorithms achieving optimal regret need to carefully balance between two conflicting objectives: increase their knowledge by playing new actions (exploration) and leverage the information acquired so far to enhance their performance (exploitation), giving rise to the well known explore-exploit dilemma (Lattimore and Szepesvári, 2020). Optimal strategies for regret minimisation are based on the *optimism in the face of uncertainty* principle, with the most notable strategies being **UCB**: Upper Confidence Bounds (Chu et al., 2011) and **TS**: Thomson Sampling (Agrawal and Goyal, 2013). In each round, these strategies construct (or update) a confidence interval around the true reward[2] and play the action with the highest "potential" outcome. We illustrate in Figure 2.3 a simplified view of the idea behind these algorithms.

Pure Exploration. Pure Exploration is a paradigm used within the contextual bandit framework to identify the best policy under practical constraints. This is suitable for applications where we do not necessarily need to exploit or gather reward to counterbalance the cost of playing actions. The constraints can be split into two types:

- **Fixed Confidence**: Given a tolerance δ, we want to identify the optimal policy with confidence at least $1 - \delta$ while reducing the number of interactions T as much as possible. Some algorithms used for this type of problem are variants of confidence interval strategies (Kalyanakrishnan et al., 2012; Degenne et al., 2019) and Track-and-Stop strategies (Garivier and Kaufmann, 2016).

[2] Thomson Sampling can also be cast within this framework (Abeille and Lazaric, 2017).

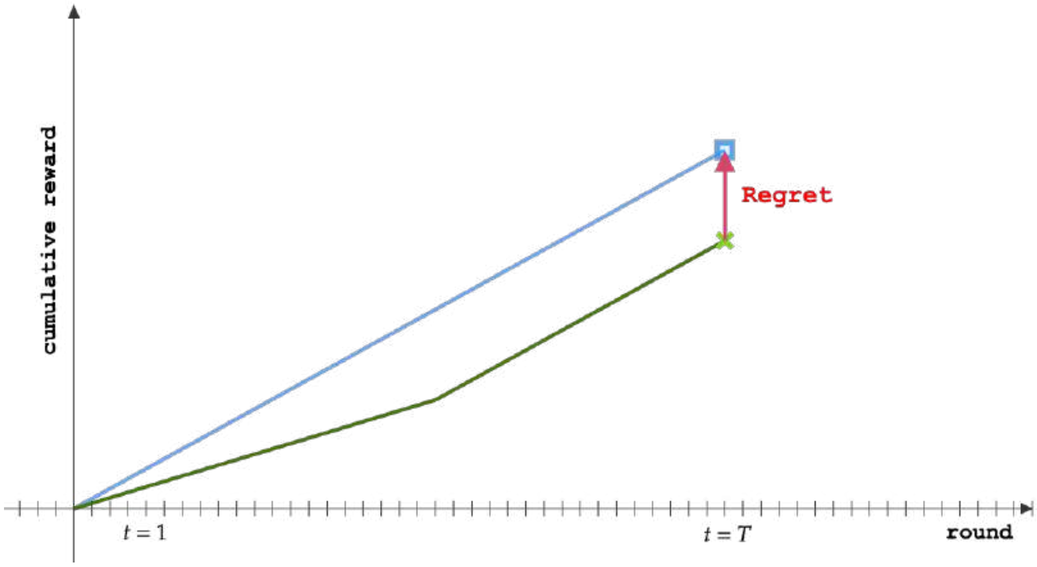

Figure 2.2: A Simple example of the regret of a strategy after T rounds: the difference (red line) between the cumulative reward of the optimal policy (blue curve) and the cumulative reward of our bandit strategy (the green curve). One can observe that starting a certain time, our strategy begins to play optimal actions (both the blue and green line start having the same slope).

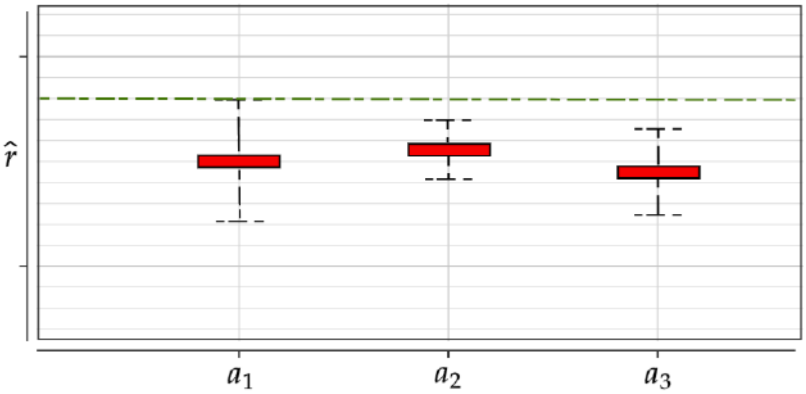

Figure 2.3: An Illustration of the principle of "optimism in face of uncertainty" with an example of a contextual bandit problem with $|\mathcal{A}| = 3$. At round t, we construct a confidence interval around the reward of each arm, and choose the arm with the highest potential payout. In this case, even if a_2 has the highest empirical mean, we choose the arm a_1 as it can have the best reward in the most "optimistic" case.

- **Fixed Interactions budget**: Given a number of interactions T, we want to maximise the probability of returning the optimal policy. One of the algorithms that deal with this type of constraint is the sequential halving algorithm developped in Karnin et al. (2013).

If it is by no mean our ambition to cover the rich literature of the bandit framework in this introduction, the reader can already imagine the endless applications and practical impact this modelling approach might have. All the strategies devised for these different applications benefit from strong theoretical guarantees (achieving low regret, finding optimal policies) while letting the agent learn by its own; besides setting parameters for the strategy, both acting and learning is done online, automatically by the agent. As attractive this online learning setting can be, there are some practical considerations that limit its viability, and motivate us to think about the problem differently:

- **Robust Infrastructure**: Deploying a bandit algorithm online, especially for large scale applications, requires a scalable and robust infrastructure, that is capable of handling hard engineering constraints (asynchronous and automatic updates, monitoring capabilities, etc) requiring a full rethinking of the model deployment pipeline. This can represent a big engineering cost that few companies are willing to pay.
- **Slower experimentation**: The same decision-making problem can be attacked by different bandit strategies, built on different assumptions, while having different hyperparameters to tune. Testing one strategy online requires the deployment of an agent that will learn by interacting with some traffic for enough rounds before convergence. If we can collect n interactions per day, and let us suppose that our bandit strategy need $7n$ interactions to converge, then we can only test out a bandit strategy per week (7 days) which renders experimentation really slow and costly.
- **Might be too costly**: Evaluating a bandit strategy offline before deployment is hard to do, making practitioners deploy agents "blindly'. This can result in unreasonable losses especially in the case of high risk applications. In addition, even if we choose the best suited bandit strategy to our problem, the level of exploration recommended by theory is often costly in the short-mid term. While a good level of exploration is beneficial for the long-term, it can result in immediate loss of revenue that might be detrimental to the business operated as it needs to comply with short-term revenue constraints.

With all these limitations taken into consideration, we want to adapt this framework to better answer the needs of industrial applications. In practice, we usually want to have full control of the amount of exploration done by the systems and prefer being able to manipulate it easily. In addition, businesses rarely face 'cold-start' problems; for the majority of the problems faced, one can leverage expert knowledge combined with non-bandit signal (information about contexts and actions) to design reasonable strategies even before the first interactions with the environment. The main challenge then shifts to the improvement of such strategies with data-driven approaches. It is highly desirable to being able to train the next strategy *offline* (as it reduces drastically the infrastructure prerequisites and accelerates experimentation) while having guarantees on its performance; before deploying the brand-new recommendation engine, one would like to make sure that it will generate at least as much revenue as its predecessor. This requires the development of a counterfactual reasoning, and the construction of specific estimators that allow us to answer the question: "What revenue would I have generated if I had acted differently?". In the hope of answering this question, the offline formulation of the contextual bandit framework was developed with the idea of leveraging logged interactions of an already deployed strategy, to confidently evaluate (What is the revenue generated of a given

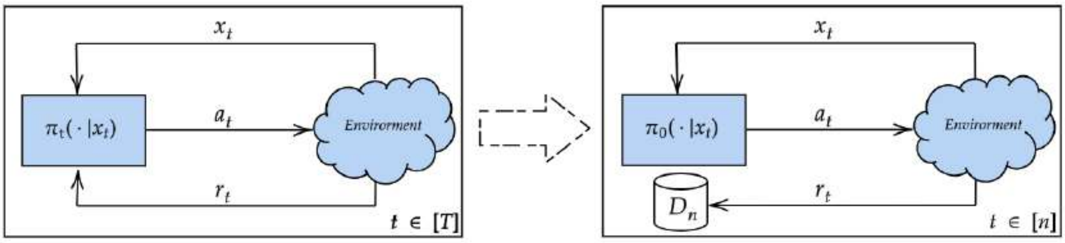

Figure 2.4: The difference between (Online) Contextual bandit and its offline formulation. The online approach (on the left) updates the model every time we observe a reward on an action. The objective is to minimize the regret in the long run. The offline setting (on the right) updates the model once based on the logged interactions of the policy π_0 with the environment. This update is done offline and the new strategy, if better, will be deployed in the future.

policy?) and learn (Find the policy that will maximize revenue) newly constructed strategies *offline*.

2.1.2 The Offline Setting

The offline contextual bandit setting is particularly interesting for industrial applications. It provides more control to practitioners, as they can evaluate and learn new policies, and fully decide on whether to deploy them online or not. In this formulation, the agent gathers data and is not updated after each interaction. Instead, this data is logged and is used by practitioners to design better performing agents for the next deployment. The current agent is represented by the policy π_0 which, in each round $t \in [n]$, acts on the context x_t by performing the action a_t and receives the feedback r_t. Figure 2.4 represents the difference between this formulation and the classical contextual bandit. All the n interactions are logged in the so-called bandit feedback dataset:

$$\mathcal{D}_n = \{x_i \sim \nu, a_i \sim \pi_0(\cdot|x_i), r_i \sim p(\cdot|x_i, a_i), \pi_0(a_i|x_i)\}_{i \in [n]}.$$

The goal in this formulation is often performance driven, as we want to find policies that minimize the risk; defined as the expected negative reward under the actions of the policy. For a given policy π, the risk is expressed as:

$$\begin{aligned} R(\pi) &= -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\mathbb{E}_{r\sim p(\cdot|x,a)}[r]\right] \\ &= -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &= \mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[c(a,x)\right]. \end{aligned}$$

with the cost $c(a, x)$ defined as $-\bar{r}(a, x)$. These notations produce the same definition and will be used exchangeably in the rest of the manuscript. As we cannot have access to the true expected risk, we proceed by building an estimator of this quantity to first evaluate the risk of any policy offline and learn reward maximizing policies in a second time.

Policy Evaluation. We want to be able to evaluate the performance of any given policy π and be able to compare it to the performance of the policy acting in production. The most reliable way to achieve this is to actually deploy the policy π and gather interactions under it to estimate its expected risk. Modern online decision systems rely on A/B tests (Kohavi et al., 2012), considered the "gold-standard" of evaluation practices. When conducted properly (Gupta et al., 2019), an A/B test can accurately estimate the effect of replacing the current policy “A” with the new candidate “B”. The common protocol begins by choosing a promising policy with the help of extensive offline experiments. The new policy is then deployed, alongside the current system and the A/B test is conducted to decide, whether the chosen candidate “B” improves and should replace the current system “A”. In large scale production systems, A/B tests require substantial engineering effort, constant monitoring and need several days to be properly analysed. Ideally, the offline selection process should produce excellent candidates that align with the online metrics, to avoid unnecessary A/B tests. Aligning offline and online performance is the goal of the research literature on policy evaluation. The challenge that arises from this approach is that we can only use data collected under the policy π_0 to evaluate, any, possibly different policy π. A common idea is to correct the bias of the estimation of the risk of new policies π with importance weighting (Chopin and Papaspiliopoulos, 2020), as we have:

$$\begin{aligned} R(\pi) &= -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &= -\mathbb{E}_{x\sim\nu, a\sim\pi_0(\cdot|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)}\bar{r}(a,x)\right], \end{aligned}$$

The expectation becomes computed under π_0 and thus can be approximated by the collected interactions, giving the well known `IPS`: Inverse Propensity Scoring estimator (Horvitz and Thompson, 1952) as a result:

$$\hat{R}_n^{\texttt{IPS}}(\pi) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}r_i.$$

This estimator of the risk of π is unbiased when the support[3] of π is included in the support of π_0. This is a desirable property as it means that the estimator is easy to analyse, consistent and will converge to the true risk with enough samples. However, as this estimator relies on importance weighting, its variance depends on the disparity between the policy that we want to evaluate and the policy that gathered the data (Bottou et al., 2013), its use can be problematic when the new policy π differs drastically from π_0. In these cases, one would prefer an estimator that do not suffer from large variance problems. A common way to achieve this is to learn a model $r_{\mathcal{M}} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^+$ of the reward mean $\bar{r}$. Once we have a model $r_{\mathcal{M}}$, we can build a simple estimator of the risk of any policy from the following observation:

$$\begin{aligned} R(\pi) &= -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &\approx -\mathbb{E}_{x\sim\nu, a\sim\pi(\cdot|x)}\left[r_{\mathcal{M}}(a,x)\right]. \end{aligned}$$

This produces the `DM`: Direct Method estimator that writes:

$$\hat{R}_n^{\texttt{DM}}(\pi) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi(a|x_i)r_{\mathcal{M}}(a,x_i).$$

The `DM` estimator does not suffer from variance problems coming from the mismatch of both policies as it does not rely on importance weighting. It can evaluate any policy π, even if π and π_0 do not share the same support. The efficiency of this estimator however depends entirely on our ability to model the problem. If this estimator enjoys a well-behaved variance, its limitation comes from a potentially substantial bias, as it is generally hard to model the reward perfectly. As both estimators have complementary properties, we can mitigate their limitations by combining them. The `DR`: Doubly Robust estimator (Dudík et al., 2014) does

[3] $\mathrm{supp}(\pi) = \{(x,a) \in \mathcal{X}\times\mathcal{A}, \pi(a|x) > 0\}$

that and results in an improved estimator. The idea behind the construction of this estimator stems from the following identity:

$$\begin{aligned} R(\pi) &= -\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &= -\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[\bar{r}(a,x) - r_{\mathcal{M}}(a,x)\right] - \mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[r_{\mathcal{M}}(a,x)\right] \\ &= -\mathbb{E}_{x\sim\nu,a\sim\pi_0(\cdot|x)}\left[\frac{\pi(a|x)}{\pi_0(a|x)}\left(\bar{r}(a,x) - r_{\mathcal{M}}(a,x)\right)\right] - \mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[r_{\mathcal{M}}(a,x)\right]. \end{aligned}$$

Which combines both the importance weighting technique and the use of a reward model $r_{\mathcal{M}}$, resulting in the DR estimator:

$$\hat{R}_n^{\mathtt{DR}}(\pi) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}\left(r_i - r_{\mathcal{M}}(a_i,x_i)\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\pi(a|x_i)r_{\mathcal{M}}(a,x_i).$$

The estimator obtained is unbiased under the same common support condition and enjoys a better variance (Nguyen-Tang et al., 2022). Research in the area of offline (also called off-policy) evaluation focuses on deriving estimators with an improved bias-variance trade-off, either by using different importance weighting techniques (Ionides, 2008; Swaminathan and Joachims, 2015b; Wang et al., 2017; Metelli et al., 2021) or by assuming a certain structure on the reward (Swaminathan et al., 2017; Saito and Joachims, 2022a; Saito et al., 2023). Building these estimators help us evaluate the performance of any policy π, thus they can be used as a training objective to find the best policy π offline.

Policy Learning. The ingredients to learn a policy are to choose an objective function; often a *regularized* off-policy estimator (Swaminathan and Joachims, 2015a; Ahmed et al., 2019; London and Sandler, 2019), and a policy class on which to optimize it. These choices dictate the approach that will be adopted and often result in different policy learning algorithms. Let $\Pi = \{\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})\}$ be the space of policies, and let us begin by introducing one of the simplest approaches. If we are confident about our ability to model the problem, and have built a reward model $r_{\mathcal{M}}$, we can proceed and learn a policy using the Direct Method. The idea stems from the following:

$$\begin{aligned} \operatorname*{arg\,min}_{\pi\in\Pi} R(\pi) &= \operatorname*{arg\,min}_{\pi\in\Pi} -\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &\approx \operatorname*{arg\,min}_{\pi\in\Pi} -\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[r_{\mathcal{M}}(a,x)\right]. \end{aligned}$$

By replacing the unknown mean reward $\bar{r}$ by our model $r_{\mathcal{M}}$, we can solve the unconstrained policy optimization problem and obtain the DM solution:

$$\forall(x,a) \quad \pi_{\mathtt{DM}}(a|x) = \mathbb{1}\left[\operatorname*{arg\,max}_{a'\in\mathcal{A}} r_{\mathcal{M}}(a',x) = a\right].$$

For each context x, the DM policy chooses the action a that has the biggest reward according to our model $r_{\mathcal{M}}$. This approach is called the Direct Method because we can directly derive the optimal policy from the reward model. Sometimes, we want to enforce some constraint on the policy deployed. For example, some applications require policies that diversify the actions played for the same context, others need some exploration to better identify the best actions. This constraint is often encoded by adding a regularization to the learning objective. To achieve better diversification, we add an entropy regularization (Ahmed et al., 2019) and modify our optimization problem to solve the following:

$$\operatorname*{arg\,min}_{\pi\in\Pi}\left\{R(\pi) + \gamma\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[\log\pi(a|x)\right]\right\} \approx \operatorname*{arg\,min}_{\pi\in\Pi}\left\{\mathbb{E}_{x\sim\nu,a\sim\pi(\cdot|x)}\left[-r_{\mathcal{M}}(a,x) + \gamma\log\pi(a|x)\right]\right\}$$

with γ a positive parameter that controls the diversity level of the policy. The solution of this optimization problem can be obtained analytically and is expressed as:

$$\begin{aligned}\forall(x,a) \quad \pi^{\gamma}_{\mathtt{DM}}(a|x) &= \mathtt{softmax}_{\mathcal{A}}\left(r_{\mathcal{M}}(a,x)/\gamma\right) \\ &= \frac{\exp(r_{\mathcal{M}}(x,a)/\gamma)}{\sum_{a'\in\mathcal{A}}\exp(r_{\mathcal{M}}(x,a')/\gamma)}.\end{aligned}$$

This policy has a positive probability mass on all actions, interpolating between a uniform distribution ($\gamma \to +\infty$) and $\pi_{\mathtt{DM}}$ ($\gamma \to 0$). The policies derived with the direct method depend on the reward model, directing all our efforts towards building a $r_{\mathcal{M}}$ that reflects the properties of the true rewards and from which the policy derived fits our engineering constraints. Sometimes, our reward model $r_{\mathcal{M}}$ produces an optimal policy $\pi_{\mathtt{DM}}$ that cannot be deployed due to application-dependent constraints (in low latency applications, finding the action with maximum reward for a particular context x can take more time than allowed). In these cases, we restrict our optimization problem to a space of policies that fits the requirements of our problem. Building a space of policies is usually done through the definition of:

- A *parametric* space of score functions $\mathcal{F}(\Theta) = \{f_\theta : \mathcal{X}\times\mathcal{A} \to \mathbb{R}, \theta \in \Theta \subset \mathbb{R}^d\}$ with d the dimension of the parameters. Given a $\theta \in \Theta$ and for a particular context x and action a, the value of $f_\theta(x,a)$ reflects the relevance of action a to the context x.

- A link function L that takes a score function f_θ and transforms it in order to define a policy π_θ. If we want to write:

  $$\forall(x,a), \quad \pi_\theta(a|x) = L(f_\theta(a,x)).$$

  L needs to be a positive, real valued function $L : \mathbb{R} \to \mathbb{R}^+$ that verifies the following condition:

  $$\forall(\theta,x), \quad \sum_{a'\in\mathcal{A}} L(f_\theta(a',x)) = 1.$$

  The space of functions verifying these conditions will be denoted by $\mathcal{L}$. Different link functions produce policies with different properties (Mei et al., 2020a,b; Sakhi et al., 2023a). We already saw from the DM example that the link function defining our policy can be an indicator or a softmax function depending on the objective we aim for. In general, we want smooth link functions that facilitate optimization making the softmax function (Mei et al., 2020b) a commonly adopted option.

The choice of the couple $(\mathcal{F}(\Theta), L \in \mathcal{L})$ is enough to define a parametric policy space $\Pi(\Theta)$ on which the optimization of our objective function can be done. Getting back to the Direct Method approach, we shift our focus to solving the following constrained optimization problem:

$$\begin{aligned}\operatorname*{arg\,min}_{\pi_\theta\in\Pi(\Theta)} R(\pi_\theta) &= \operatorname*{arg\,min}_{\pi_\theta\in\Pi(\Theta)} -\mathbb{E}_{x\sim\nu, a\sim\pi_\theta(\cdot|x)}\left[\bar{r}(a,x)\right] \\ &\approx \operatorname*{arg\,min}_{\pi_\theta\in\Pi(\Theta)} -\mathbb{E}_{x\sim\nu, a\sim\pi_\theta(\cdot|x)}\left[r_{\mathcal{M}}(a,x)\right].\end{aligned}$$

As we do not know if $\pi_{\mathtt{DM}} \in \Pi(\Theta)$, we proceed by computationally solving the empirical counterpart of the objective:

$$\operatorname*{arg\,min}_{\pi_\theta\in\Pi(\Theta)} \left\{ -\frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi_\theta(a|x_i) r_{\mathcal{M}}(a,x_i) \right\} = \operatorname*{arg\,min}_{\pi_\theta\in\Pi(\Theta)} \left\{ \hat{R}_n^{\mathtt{DM}}(\pi_\theta) \right\}.$$

Which can be interpreted as distilling the potentially complicated reward model $r_{\mathcal{M}}$ into a policy that fits our constraints. In this example, our learning objective was the `DM` estimator. As a general rule, off-policy learning objectives rely on optimizing a regularized risk estimator on a parametric policy class:

$$\arg\min_{\pi_\theta \in \Pi(\Theta)} \left\{ \hat{R}_n(\pi_\theta) + \lambda C(\pi_\theta) \right\}, \tag{2.1}$$

with $\hat{R}_n$ a risk estimator, λ a tunable parameter and $C(\pi_\theta)$ a regularization term that is either motivated by additional constraints we want to enforce (Ahmed et al., 2019; Schulman et al., 2015) or statistical learning theory arguments making the learning of these policies more principled (Swaminathan and Joachims, 2015a; Ma et al., 2019; London and Sandler, 2019). In the next section, we will develop the policy learning discussion more, with a particular focus on statistical learning tools that enable us to learn systems with performance certificates.

2.2 Performance Guarantees with Statistical Learning

Statistical learning theory (Vapnik, 1998) studies the problem of inference; that is, of gaining knowledge and making predictions based on a set of data. In particular, we are interested in the **PAC**: Probably Approximately Correct framework (Valiant, 1984), a branch of learning theory that investigates the problem of generalisation, answering the question of how well a predictor (or a family of predictors) can perform on unseen data. Developments of this branch improved our understanding of common learning paradigms, with contributions in supervised learning (Vapnik, 1991; Cortes and Vapnik, 1995; McAllester, 1998; Catoni, 2007; Germain et al., 2009), unsupervised learning (Bengio et al., 2013; Saunshi et al., 2019; Nozawa et al., 2020) and online learning (Even-Dar et al., 2002; Seldin et al., 2011; Haddouche and Guedj, 2022; Tirinzoni et al., 2023; Al-Marjani et al., 2023). Historically, supervised learning had attracted most attention and is best understood from this perspective. It is only natural to choose this learning paradigm to present some of the tools used by the PAC framework. In this setting, we are given a data set, and a loss to measure performance. We fix a set of predictors and look for a good predictor in this set, w.r.t to the loss defined. Formally, we have:

- A dataset $S_n = \{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}\}_{i\in[n]}$ composed of n i.i.d. observations coming from an unknown joint distribution $p(\mathcal{X}, \mathcal{Y})$. $\mathcal{X}$ is the object set (text, image) and $\mathcal{Y}$ the label set (sentiment of the text, class of the image).

- A loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ measuring the quality of the predictions, with the convention that $l(y, y) = 0$.

- We look for good predictors in $\mathcal{H}_\Theta = \{h_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ a class of predictors, parameterized by θ coming from the parameter set Θ.

- We are interested in finding a predictor h from $\mathcal{H}_\Theta$ that will minimize the expected loss $l(h) = \mathbb{E}_{(X,Y)\sim\nu}[l(h(X), Y)]$. We denote by $l_n(h)$ the empirical loss estimate.

As it is usually impossible to have access to the true, expected loss, the PAC toolbox provide us with bounds that control this quantity for any predictor $h_\theta \in \mathcal{H}_\Theta$. PAC bounds give us the following result, holding with high probability over the data:

$$\forall \theta \in \Theta : l(h_\theta) \leq l_n(h_\theta) + \mathcal{O}\left(C_n(\mathcal{H}_\Theta)\right),$$

with $C_n(\mathcal{H}_\Theta)$ a measure of complexity (Vapnik, 1998; Zhou, 2002; Bartlett and Mendelson, 2003) of the class of predictors used. In our applications, we want to obtain a performance guarantee on our predictors with the help of these bounds. For a predictor $h_{\hat{\theta}}$, we want to control its true expected loss with high probability. We aim at obtaining a *performance guarantee/certificate*, which is a result of the form:

$$l(h_{\hat{\theta}}) \leq 0.12. \tag{2.2}$$

This result **guarantees us** (with high probability) that our predictor, will suffer a loss of at most 0.12. To obtain the smallest guarantees, we seek bounds that are tight and advocate for minimizing the right-hand side over all $\theta \in \Theta$ in order to control and minimize expected loss. For example, Maurer and Pontil (2009) derived an empirical Bernstein-type bound, and used the notion of covering number (Zhou, 2002) to control the loss of a class of predictors. For a tolerance $\delta \in]0,1]$, Their main result is a bound holding with probability $1-\delta$:

$$\forall \theta \in \Theta : l(h_\theta) \leq l_n(h_\theta) + \sqrt{\frac{18 v_n(h_\theta) \ln\left(\mathcal{M}_n(\mathcal{H}_\Theta)/\delta\right)}{n}} + \frac{15 \ln\left(\mathcal{M}_n(\mathcal{H}_\Theta)/\delta\right)}{n-1}, \tag{2.3}$$

with $v_n(h_\theta)$ the empirical variance of the loss estimate and $\mathcal{M}_n(\mathcal{H}_\Theta)$ a complexity measure defined in (Maurer and Pontil, 2009, Theorem 6). This complexity is intractable even for simple predictor classes, which means that its presence makes the bound unusable as-is for learning purposes. $\mathcal{M}_n(\mathcal{H}_\Theta)$ can be upper bounded by empirical quantities (Zhang, 2002) but this often results in loose and overly conservative bounds. In particular, this complexity is known to be very large for rich predictor classes (i.e. neural networks), making the bound vacuous. To circumvent this limitation, the usual approach is to identify useful quantities from the bound and propose a learning principle, replacing intractable quantities with tunable hyperparametes. This approach motivated numerous learning principles, such as Empirical Risk Minimization (Vapnik, 1991) and Structural Risk Minimization (Cortes and Vapnik, 1995). In this example, the **SVP** principle was derived from Equation (2.3) proposing to solve the following optimization problem:

$$\operatorname*{arg\,min}_{\theta \in \Theta} \left\{ l_n(h_\theta) + \lambda \sqrt{\frac{v_n(h_\theta)}{n}} \right\},$$

with λ a hyperparameter selected with cross-validation. Using these learning principles provide practitioners with tractable optimization objectives, but does not result in performance certificate like in Equation (2.2). Swaminathan and Joachims (2015a) adapted these results to the offline contextual bandit framework. See (Swaminathan and Joachims, 2015a, Table 1) for the differences between the supervised learning problem and the offline contextual bandit problem. Particularly, they based their analysis on `cIPS`: clipped IPS (Bottou et al., 2013) in order to respect the bounded assumption of the loss. This risk estimate was used to derive a bound similar to Equation (2.3) holding for policies π_θ in a policy class $\Pi(\Theta)$. The obtained bound (Swaminathan and Joachims, 2015a, Theorem 1) is intractable, and motivated the use of a similar learning principle.

The Distributionally Robust Optimization framework (Duchi et al., 2021) provides an intuitive approach to control the loss of our predictors. After observing the samples S_n, it treats the induced empirical distribution with scepticism and seeks a solution that minimizes the worst-case expected cost over a family of distributions, described in terms of an uncertainty ball (around the observed, empirical distribution). These tools were proven to be powerful for decision theory (Duchi and Namkoong, 2019) and in training robust classifiers (Madry et al., 2018). Let $\mathcal{U}_\epsilon(\hat{p}_n)$ be the uncertainty ball of radius ϵ, around the empirical distribution $\hat{p}_n$, and let h_θ be a predictor from $\mathcal{H}_\Theta$. Instead of studying the empirical loss $l_n(h_\theta)$, the DRO formulation focuses on the following, worst-case empirical estimator (Duchi et al., 2021):

$$l_n(h_\theta, \mathcal{U}_\epsilon(\hat{p}_n)) = \max_{q \in \mathcal{U}_\epsilon(\hat{p}_n)} \mathbb{E}_{(x,y)\sim q}\left[l(h_\theta(x), y)\right].$$

This framework is also called generalized, empirical likelihood as a well-chosen uncertainty ball recovers the empirical likelihood approach of Owen (2001). For a particular choice of the uncertainty set $\mathcal{U}_\epsilon^{\chi^2}(\hat{p}_n)$ (using the χ^2 divergence to quantify the distance from the empirical distribution), Duchi and Namkoong (2019) prove that the DRO, worst-case empirical estimator is equivalent to a *variance-regularized* empirical loss:

$$l_n(h_\theta, \mathcal{U}_\epsilon^{\chi^2}(\hat{p}_n)) = l_n(h_\theta) + \sqrt{\epsilon v_n(h_\theta)}.$$

This result means that minimizing the worst-case empirical loss is equivalent to solving the **SVP** principle. These tools were adapted to the problem of off-policy learning (Faury et al., 2020; Dai et al., 2020) and we develop them further in Chapter 3. Their use is motivated by asymptotic-coverage arguments (in the limit, the worst-case risk will cover the true risk) and their finite-sample analysis is loose, failing to produce satisfying performance guarantees (a result similar to Equation (2.2)) in practical scenarios (Dai et al., 2020).

If our objective is to know how a policy will perform before it interacts with the environment, deriving a learning principle is not enough. We are interested in obtaining **performance guarantees**; results similar to Equation (2.2), where we control with high confidence the risk of a trained policy $\pi_{\hat{\theta}}$:

$$R(\pi_{\hat{\theta}}) \leq -0.81. \quad \text{(The risk is in } [-1, 0]) \tag{2.4}$$

This result certifies that, in the worst case, our policy $\pi_{\hat{\theta}}$ will have a risk of -0.81. This performance guarantee give practitioners a way to identify promising policies that are worth A/B testing. For this same example, if the logging policy π_0 have a risk of -0.71, then Equation (2.4) alone guarantee us that the new learned policy improves on π_0. These results are desired in the offline policy learning context and can have a substantial impact on real world problems. Obtaining these results rely on the derivation of **tight** and **tractable** PAC bounds. Recently, PAC-Bayes bounds (McAllester, 1998; Catoni, 2007), a family of PAC bounds, promise the delivery of performance guarantees for difficult problems (Dziugaite and Roy, 2017) and present themselves as good candidates to answer this question. If the notion of complexity in PAC bounds limited their application to simple predictor classes, PAC-Bayes techniques can deal elegantly with any predictor class $\mathcal{H}_\Theta$ and are proven to provide performance guarantees even for well-known, over-parameterized neural networks (Dziugaite and Roy, 2017). However, an artefact of these bounds is that we need to change the quantities of interest. PAC bounds study the performance of a predictor h_θ in $\mathcal{H}_\Theta$ by controlling its loss $l(h_\theta)$. PAC-Bayes bounds however study randomized predictors; obtained by sampling in a set of basic predictors, according to some prescribed probability distribution. Formally, let $\mathcal{P}(\Theta)$ be the set of all probability distributions on Θ (equipped with its σ-algebra). let $Q \in \mathcal{P}(\Theta)$ a probability distribution over Θ (and thus $\mathcal{H}_\Theta$), PAC-Bayes bounds control the loss of randomized predictors, computed as :

$$\mathbb{E}_{\theta\sim Q}\left[l(h_\theta)\right].$$

In a supervised learning setting, this quantity can be interpreted as adopting the following procedure: for each sample $(X, Y) \sim \nu(\mathcal{X}, \mathcal{Y})$, we sample a predictor h_θ from Q, predict the label $Y^p = h_\theta(X)$ and then compute the loss $l(Y^p, Y)$. This is different from studying aggregated predictors (Breiman, 2001) where for each sample, we aggregate (by either voting or averaging) all predictor's results to predict the label. We present the differences introduced with the PAC-Bayesian approach for supervised learning. We have:

- A dataset $S_n = \{X_i \in \mathcal{X}, Y_i \in \mathcal{Y}\}_{i\in[n]}$ composed of n i.i.d. observations coming from an unknown joint distribution $p(\mathcal{X}, \mathcal{Y})$. $\mathcal{X}$ is the object set (text, image) and $\mathcal{Y}$ the label set (sentiment of the text, class of the image).
- A loss function $l : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ measuring the quality of the predictions, with the convention that $l(y, y) = 0$.
- We define $\mathcal{H}_\Theta = \{h_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ a class of predictors, parameterized by θ coming from the parameter set Θ.
- **(PAC-Bayes)** We define a a set of probability distribution $\mathcal{M}(\Theta) \subseteq \mathcal{P}(\Theta)$ over Θ.
- **(PAC-Bayes)** We are interested in finding a good distribution $Q \in \mathcal{M}(\Theta)$ that will minimize the expected loss of the randomized predictor $\mathbb{E}_{\theta\sim Q}\left[l(h_\theta)\right]$.

This is achieved through the derivation of bounds holding for all distributions $Q \in \mathcal{M}(\Theta)$. Let $P \in \mathcal{P}(\Theta)$ a reference distribution that does not depend on the data S_n. The general form of PAC-Bayesian bounds is an inequality holding with high probability:

$$\boxed{\forall Q \in \mathcal{M}(\Theta) : \mathbb{E}_{\theta\sim Q}\left[l(h_\theta)\right] \le \mathbb{E}_{\theta\sim Q}\left[l_n(h_\theta)\right] + \mathcal{O}\left(\mathcal{KL}\left(Q||P\right)\right),}$$

with $\mathcal{KL}$ the KL-divergence defined by:

$$\mathcal{KL}(Q||P) = \begin{cases} \int \ln\left\{\frac{dQ}{dP}\right\} dQ & \text{if } Q \text{ is } P\text{-continuous,} \\ +\infty & \text{otherwise.} \end{cases}$$

For example, we give below a simple PAC-Bayes bound (Catoni, 2007) to showcase the versatility of this framework. Let $P \in \mathcal{P}(\Theta)$ a reference distribution, $\delta \in]0, 1]$ a tolerance and $\lambda > 0$, we have with probability at least $1 - \delta$:

$$\boxed{\forall Q \in \mathcal{P}(\Theta) : \mathbb{E}_{\theta\sim Q}\left[l(h_\theta)\right] \le \mathbb{E}_{\theta\sim Q}\left[l_n(h_\theta)\right] + \frac{\mathcal{KL}\left(Q||P\right) + \log 1/\delta}{\lambda} + \frac{\lambda}{8n} \qquad (2.5)}$$

Minimizing the r.h.s gives the smallest guarantees on aggregated predictors. We get to solve the following optimization problem:

$$\underset{Q\in\mathcal{P}(\Theta)}{\arg\min} \quad \mathbb{E}_{\theta\sim Q}\left[l_n(h_\theta)\right] + \frac{\mathcal{KL}\left(Q||P\right)}{\lambda},$$

which happens to be analytically tractable and we obtain the Gibbs distribution (Guedj, 2019) as a solution:

$$\forall \theta \in \Theta, \quad d\hat{Q}(\theta) \propto \exp\left(-\lambda l_n(h_\theta)\right) \times dP(\theta).$$

The bound also gives us an idea on the worst case performance of the randomized predictor according to $\hat{Q}$:

$$\boxed{\mathbb{E}_{\theta \sim \hat{Q}}\left[l(h_\theta)\right] \leq \log\left(\mathbb{E}_{\theta \sim P}\left[\exp\left(-\lambda l_n(h_\theta)\right)\right]\right) + \frac{\log 1/\delta}{\lambda} + \frac{\lambda}{8n}. \qquad (2.6)}$$

Sampling from the distribution $\hat{Q}$ is mandatory if we want to use the randomized predictor, and computing $L(P) = \log\left(\mathbb{E}_{\theta \sim P}\left[\exp\left(-\lambda l_n(h_\theta)\right)\right]\right)$ is desired to have an idea on its performance. Both sampling from $\hat{Q}$ and approximating $L(P)$ can be done with (Markov Chain/Sequential) Monte Carlo (Chopin and Papaspiliopoulos, 2020). If these methods fail to scale to our problem, a usual solution is to restrict the probability set $\mathcal{M}(\Theta)$ to simple distributions. If $\Theta = \mathbb{R}^d$, a common choice is to set $\mathcal{M}(\Theta) = \left\{\mu \in \mathbb{R}^d, \mathcal{N}(\mu, I_d)\right\}$ to unit-variance, isotropic Gaussians. By fixing the reference distirbution to $P = \mathcal{N}(\mu_0, I_d)$, the previous bound becomes:

$$\boxed{\forall \mu \in \mathbb{R}^d : \mathbb{E}_{\theta \sim \mathcal{N}(\mu, I_d)}\left[l(h_\theta)\right] \leq \mathbb{E}_{\theta \sim \mathcal{N}(\mu, I_d)}\left[l_n(h_\theta)\right] + \frac{||\mu - \mu_0||^2 + 2\log 1/\delta}{2\lambda} + \frac{\lambda}{8n}. \qquad (2.7)}$$

Obtaining a good randomized predictor boils down to computationally solving the optimization problem:

$$\operatorname*{arg\,min}_{\mu \in \mathbb{R}^d} \mathbb{E}_{\theta \sim \mathcal{N}(\mu, I_d)}\left[l_n(h_\theta)\right] + \frac{||\mu - \mu_0||^2}{2\lambda}.$$

This optimization problem looks like Variational Bayes objectives (Blei et al., 2017) for which a multitude of solutions were proposed to solve it efficiently (Xu et al., 2019). Once we have our solution, obtaining the worst case loss for the randomized predictor can be done by evaluating the bound. We can observe that, contrary to classical PAC bounds, PAC-Bayesian bounds are tractable and benefit from various computational tools to find their minimizers. They are also tight enough to be valuable in learning both simple (Germain et al., 2009) and complex predictors (Dziugaite and Roy, 2017) with guarantees. To increase the impact of these bounds, research in this area is focused on deriving new, tighter bounds (Mhammedi et al., 2019; Jang et al., 2023), loosening assumptions (Alquier and Guedj, 2018; Kuzborskij and Szepesvári, 2020; Haddouche and Guedj, 2023) and adapting them to various learning problems (Seldin et al., 2011; London and Sandler, 2019; Haddouche and Guedj, 2022). To make them even more viable, new disintegration techniques (Viallard et al., 2023) were also developed to allow these bounds to give strong performance guarantees on single predictors drawn from the learned distribution. If working with randomized predictors can be seen as a "bug" in most settings, it is considered a "feature" in policy optimization as both policies and randomized predictors are closely related.

- The procedure of a randomized predictor is the following: for each sample $(X, Y) \sim \nu(\mathcal{X}, \mathcal{Y})$, we sample a predictor h_θ from Q, predict the label $Y^p = h_\theta(X)$ and then suffer the loss $l(Y^p, Y)$.
- The procedure of a policy is the following: for each context $x \sim \nu(\mathcal{X})$, we sample an action a from $\pi(\cdot|x)$, and receive the reward $r \sim p(\cdot|x, a)$.

The procedures are similar and both objects can be related if we work with class of predictors that map contexts x to actions in $\mathcal{A}$. Indeed, instead of sampling directly from a distribution on the action set, we sample from a distribution on a predictor space, $\mathcal{H}_\Theta \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$. As such, for a distribution, Q over $\mathcal{H}_\Theta$, the probability of an action, $a \in \mathcal{A}$, given a context, $x \in \mathcal{X}$, is the probability that a random predictor, $h_\theta \sim Q$, maps x to a; that is:

$$\pi_Q(a|x) = \mathbb{E}_{h\sim Q}\left[\mathbb{1}[h(x) = a]\right].$$

This result shows that a policy is a randomized predictor in disguise. This perspective was developed in Seldin et al. (2012), adopted by London and Sandler (2019) and later formalized in Sakhi et al. (2023a). This result is key to the analysis of London and Sandler (2019), that adapted McAllester (2003)'s bound to the offline contextual bandit setting. To achieve this, they clipped the propensity score and used the following risk estimator:

$$\hat{R}^\tau_n(\pi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} r_i,$$

with $\tau \in]0, 1]$. Their analysis resulted in the bound below. Given a reference distribution P and a tolerance parameter $\delta \in]0, 1]$, the following holds with probability at least $1 - \delta$, uniformly over all distributions $Q \in \mathcal{P}(\Theta)$:

$$R(\pi_Q) \leq \hat{R}^\tau_n(\pi_Q) + \frac{2(\mathcal{KL}(Q||P) + \ln\frac{2\sqrt{n}}{\delta})}{\tau n} + \sqrt{\frac{2[\hat{R}^\tau_n(\pi_Q) + \frac{1}{\tau}](\mathcal{KL}(Q||P) + \ln\frac{2\sqrt{n}}{\delta})}{\tau n}}.$$

One can see that for offline contextual bandits, PAC-Bayes bounds control the quantity of interest, which is the risk of the policy directly. Working with randomized predictors for this problem matches perfectly our needs. Another connection between the PAC-Bayes framework and offline contextual bandits is that the reference distribution can be set naturally to match the logging policy π_0. Indeed, P can be chosen such as $\pi_0 = \pi_P$ to obtain a bound that encourages policies with low empirical risk that stay close to the logging policy π_0. All of these connections make PAC-Bayes the perfect candidate for guaranteed performance. The bound proposed in London and Sandler (2019) however, is not tight enough and produces vacuous results in practical scenarios (Sakhi et al., 2023a). London and Sandler (2019) avoided using the bound and derived a learning principle for parametrized softmax policies. This principle advocates for a L_2 regularization towards the parameter of π_0:

$$\arg\min_{\theta\in\Theta}\left\{\hat{R}^\tau_n(\pi_\theta) + \lambda||\theta - \theta_0||^2\right\}.$$

If this principle improves on **CRM** (Swaminathan and Joachims, 2015a), these results are far from being satisfying if we want to have guarantees on the learned policies. To this end, we continue the development of PAC-Bayes bounds for this problem in Chapters 4 and 5 to finally obtain tight bounds, that certify the performance of the new policies and can confidently improve on the logging policy π_0. These results are desired in production settings where we would like to propose a new system that will improve on the current production system with high probability. This is the case of online decision systems, in particular, recommender systems, where our goal is to always improve the quality of recommendation to better answer the needs of the users. In the next section, we cover the history of recommendation and present how the offline contextual bandit tools fit in the picture, playing a crucial role in redefining the modern internet experience.

2.3 Online Decision Systems: History of Recommendation

Online decision systems have revolutionized the way we interact with the vast ocean of content present on the internet. From search engines to recommender systems, they offer a personalized experience by efficiently exploring the overwhelming amount of information and filtering it to cater to the specific needs of the users. Although these systems are now ubiquitous, it was not always the case during the emergence of the internet. Democratizing the access to web-based information resulted in an exponential increase in the quantity of available data. This increase alone did not upgrade the internet experience, as having access to non structured, vast amount of information is not beneficial unless we have tools to efficiently explore it. This issue attracted research interest which gave birth to the field of IR: Information Retrieval (Rijsbergen, 1979). A natural application of IR is web search engines, now considered an integral part of the internet experience. In their simplest form, these engines take in queries like “Is it normal to be depressed during COVID?” and produce an ordering of, hopefully relevant web pages as a result. If we are more ambitious, we would like to know what happens when our query is incomplete as we need implicit information to better answer it? What happens when we do not have an explicit query at all? What happens when we do not know which musical artist can be interesting to listen to or which movie we would like to watch? In such scenarios, the field of Recommender Systems comes into play, providing a needed solution to these challenges. The concept of filtering and recommending information to users has been around for some time, with early examples dating back to the 1990s. Belkin and Croft (1992) analyzed the two notions of Information Filtering and Information Retrieval, arguing that the latter constitutes the fundamental technology behind Search Engines, while Recommender Systems are built with ideas rooted in the former. In the same year, Goldberg et al. (1992) proposed the "Tapestry" system allowing users, through a graphical interface, to explicitly rate items and view recommendations based on their preferences and the ratings of other users with similar tastes. The term "Collaborative Filtering" first appeared in this work to denote that the information extracted from other users preferences, combined with your preferences (explaining the collaborative part) would be used to infer what the system should recommend (explaining the filtering part) to you. During the same period, content-based filtering also emerged, where recommendations are made based on item features or attributes. Despite its simplicity, creating an operational content-based recommender system, even for basic applications, was a significant challenge, as it required a deep (not in a machine learning sense) understanding of the topic under consideration and the factors influencing the relationship between users and the topics themselves. While modern machine learning tools, emerging from the combination of accurate modelling and powerful computations (Blei et al., 2003; Vaswani et al., 2017b), can now extract valuable factors from the content being recommended, this was not the case in the 1990s. One of the earliest successful real-world projects in this area was the Music Genome Project, which aimed to capture the essence of music through its properties. This project represents any song with over 450 properties and describes the interplay between each one of them. Once we obtain the song’s representations, the recommendation procedure follows a natural design. When a user likes a song, the procedure attributes positive values to its specific properties, promoting similar songs (with similar properties) and bringing them to the user’s attention. Collaborative filtering and content-based filtering are built on distinct principles, each with its own strengths and weaknesses. Content-based filtering relies on a comprehensive understanding of the recommended content, and therefore does not necessarily require input from other users. In contrast, collaborative filtering depends heavily on user interactions to identify individuals with similar preferences. Content-based filtering may, however, have limitations when it comes to generating novel or diverse recommendations, since it is primarily based on an understanding of the properties of the content. Fortunately, both

$$R = \left.\underbrace{\begin{bmatrix} ? & \mathbf{1} & \dots & \mathbf{5} & ? \\ \mathbf{4} & \mathbf{2} & \dots & ? & ? \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{5} & ? & \dots & \mathbf{4} & ? \\ ? & ? & \dots & \mathbf{5} & \mathbf{2} \end{bmatrix}}_{I}\right\} U \implies \hat{f}(R, M) = \begin{bmatrix} 4 & \mathbf{1} & \dots & \mathbf{5} & 2 \\ \mathbf{4} & \mathbf{2} & \dots & 3 & 3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{5} & 1 & \dots & \mathbf{4} & 1 \\ 3 & 2 & \dots & \mathbf{5} & \mathbf{2} \end{bmatrix}$$

Figure 2.5: An example of a rating matrix completion problem. The recommendation procedure $\hat{f}$ is tasked to predict the missing ratings of the incomplete user-item matrix R using the information provided by the matrix R and some metadata M about the items, if available.

$$L = \left\{ \begin{matrix} u_1 & t^1_{u_1} & I_1 \\ u_1 & t^2_{u_1} & I_1 \\ u_1 & t^3_{u_1} & I_2 \\ u_2 & t^1_{u_2} & I_2 \\ \vdots & \vdots & \vdots \\ u_n & t^j_{u_n} & I_n \end{matrix} \right\} \implies IF = \begin{bmatrix} 1 & 1 & \dots & ? & ? \\ ? & 1 & \dots & ? & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ ? & ? & \dots & 1 & ? \\ 1 & ? & \dots & ? & ? \end{bmatrix}$$

Figure 2.6: Typical logs of views of items by users transformed into an implicit feedback matrix. The user u_1 viewed the item I_1 twice, its row is highlighted with orange in the IF matrix where these views were deduplicated and transformed to a positive signal for the item I_1 (column highlighted with yellow).

approaches offer distinct benefits, and the most successful recommender systems in use combine the strengths of both methods (Vasile et al., 2016; Jeunen et al., 2020). The early recommender systems were often limited by the availability of data, as well as by the computational resources needed to process that data. However, the ideas behind them laid the groundwork for more sophisticated recommendation paradigms that would emerge in the years to come. Even if it is by no means our ambition to provide an exhaustive covering of the recommendation research landscape in this introduction, we want to give the reader in the next paragraphs different perspectives on how recommendation is modelled.

2.3.1 Recommendation as Preference Prediction

The "Tapestry" system, which was introduced earlier, approached the recommendation problem as predicting the rating that a user would give to an item. This approach gained further popularity with the work of Resnick et al. (1994) within the GroupLens Research Lab, which provided a complete architecture to support research in this area. The idea is to have different users rate items and gather this information in a dataset. Since asking each user to rate all items is not feasible (think about massive movie catalogues, for example), users are randomly exposed to a few items for which they give a rating as shown in Harper and Konstan (2015). These ratings are then compiled and represented in a user-item rating matrix R of size $U \times I$ with U and I respectively the number of users and the number of items. Each entry in R, $R_{u,i}$, represents the given or missing rating of user u to item i. The goal is to learn a procedure $\hat{f}$ that can predict the missing ratings and complete the matrix R. In addition to the ratings' dataset, some metadata about the items M (relevant properties of the items) is often available (Harper and Konstan, 2015). This allows practitioners to explore different ways to combine the users ratings data (collaborative filtering) and item specific data (content-based filtering) to obtain a procedure that produces the most accurate predictions. The quality of the predictions is typically measured by assessing the difference between the true and estimated ratings on a separate test set (Salakhutdinov and Mnih, 2007). Figure 2.5 visualises an example of an incomplete rating dataset and the expected output of the procedure $\hat{f}$. Given R and/or M as input, the learned procedure $\hat{f}$ generates potential ratings for every user item pair. These complete ratings are then used to identify items that may be of interest to each user by selecting the ones with high predicted ratings. The underlying assumption of this approach is that items with higher ratings are considered suitable candidates for recommendation. This framework is referred to in the literature as the "explicit feedback" setting, as it requires users to explicitly provide ratings on a predefined scale. Recommendation based on explicit feedback has had a huge practical success, and were responsible for the success of many tech companies. For example, Netflix, a *DVD* rental service at the time, launched a competition rewarding a million US dollars to whoever achieves the smallest reconstruction error of their rating dataset. Despite its success, the "explicit feedback" paradigm suffers from significant limitations. The fundamental premise of the approach is to build a method that, gathers in an unbiased manner, genuine ratings that accurately reflect the user's true appreciation of items. However, obtaining this data requires the system to explicitly ask users to rate items, which can be costly and may be detrimental to the user's experience. To deal with this issue, these systems provide a non-intrusive way to rate an item; a like button to express if they loved the content they interacted with, or a rating system for the product bought from a retail store. These methods are integrated in the system and do not necessarily harm the user experience, but they give the entire freedom to the user to rate an item or not. This introduces an additional bias to the rating matrix as the presence of a rating is influenced by the decision of the user, making ratings "Missing Not At Random" (MNAR) (Yang et al., 2018). For example, once a user watches a movie, how much they enjoyed the movie influences directly the likelihood that they will leave a rating. Additionally, new ratings of an item tend to be biased by all the previous ratings that item received, making it hard to measure how much a new user really likes an item. Actually, in depth studies have shown that most ratings collected by these systems are biased, and correlate poorly with the true interest of users, as evidenced by Zhang et al. (2017). A potentially better signal to consider is the organic behaviour of users. By exploiting the information that is inherent to a user interacting with an item, we can avoid the need for explicit ratings. Indeed, we can reasonably assume that a user will mostly view retail product pages of items they are interested in, or movies and series that they think they will enjoy. This information is denoted in the literature by the "implicit feedback" as it is not asked directly from the user but reflects to a certain extent its interests through his organic interactions with the system. Implicit-feedback recommendation took the industry by storm and dominated the industrial applications in recent years. For example, Gomez-Uribe and Hunt (2016) describe the recommender system recently used by Netflix and show that they moved from their heavy dependence on rating feedback to a simpler feedback mechanism, focusing primarily on signal acquired from interaction data. This interaction data can come in different forms depending on the nature of the service provided. For online recommender systems, the most common form of interaction is a view/visit (or multiple views/visits) of an item by a user. In general, these views are logged, processed and deduplicated to build a matrix of binary, positive-only signal as shown in Figure 2.6. This simple organic signal differs from the explicit rating given by users, as it cannot encode negative information. When a user interacts with an item, we assume that the user is interested in the item. In the other hand, when an interaction is missing from the data, we do not know whether this means that the user is simply unaware of the item, or whether it is irrelevant to the user. The absence of negative signal motivated new collaborative filtering algorithms, sometimes augmented with item-related data,

Bibliography

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/abeille17a.html.

Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6):734–749, 2005.

M. Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Comput. Surv.*, 55(7), dec 2022. ISSN 0360-0300. doi: 10.1145/3543846. URL https://doi.org/10.1145/3543846.

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431, 2017.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *International Conference on Machine Learning*, pages 1638–1646, 2014.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/agrawal13.html.

Zafarali Ahmed, Nicolas Le Roux, Mohammad Norouzi, and Dale Schuurmans. Understanding the impact of entropy on policy optimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 151–160. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ahmed19a.html.

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients, 2020. URL https://api.semanticscholar.org/CorpusID:234358812.

Aymen Al-Marjani, Andrea Tirinzoni, and Emilie Kaufmann. Active coverage for pac reinforcement learning. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5044–5109. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/al-marjani23a.html.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *ArXiv*, abs/2110.11216, 2021. URL https://api.semanticscholar.org/CorpusID:239049660.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018. doi: 10.1007/s10994-017-5690-0. URL https://doi.org/10.1007/s10994-017-5690-0.

Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, and Maxime Vono. Reward optimizing recommendation using deep learning and fast maximum inner product search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4772–4773, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3542622. URL https://doi.org/10.1145/3534678.3542622.

Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 984–1017. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/aouali23a.html.

Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, and Flavian Vasile. Probabilistic Rank and Reward: A Scalable Model for Slate Recommendation. working paper or preprint, January 2023b. URL https://hal.science/hal-03959643.

Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect thompson sampling. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2087–2115. PMLR, 25–27 Apr 2023c. URL https://proceedings.mlr.press/v206/aouali23a.html.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40 – 79, 2010. doi: 10.1214/09-SS054. URL https://doi.org/10.1214/09-SS054.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Ishan Bajaj, Akhil Arora, and M. M. Faruque Hasan. *Black-Box Optimization: Methods and Applications*, pages 35–65. Springer International Publishing, Cham, 2021. ISBN 978-3-030-66515-9. doi: 10.1007/978-3-030-66515-9_2. URL https://doi.org/10.1007/978-3-030-66515-9_2.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3(null):463–482, mar 2003. ISSN 1532-4435.

Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, dec 1992. ISSN 0001-0782. doi: 10.1145/138859.138861. URL https://doi.org/10.1145/138859.138861.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Yoshua Bengio and Jean-Sébastien Senecal. Quick training of probabilistic neural nets by importance sampling. In Christopher M. Bishop and Brendan J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pages 17–24. PMLR, 03–06 Jan 2003. URL https://proceedings.mlr.press/r4/bengio03a.html. Reissued by PMLR on 01 April 2021.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: a review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798—1828, August 2013. ISSN 0162-8828. doi: 10.1109/tpami.2013.50. URL http://arxiv.org/pdf/1206.5538.

James O Berger, Robert L Wolpert, MJ Bayarri, MH DeGroot, Bruce M Hill, David A Lane, and Lucien LeCam. The likelihood principle. *Lecture notes-Monograph series*, 6:iii–199, 1988.

H.G. Beyer and HP. Schwefel. Evolution strategies - a comprehensive introduction. *Natural Computing*, 1(1):3–52, March 2002. doi: 10.1023/A:1015059928466.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138. ACM, 2009.

Guy Blanc and Steffen Rendle. Adaptive sampled softmax with kernel based sampling. In *International Conference on Machine Learning*, pages 590–599. PMLR, 2018.

David M. Blei and John D. Lafferty. Correlated topic models. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pages 147–154, 2005. URL http://papers.nips.cc/paper/2906-correlated-topic-models.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar 2003. ISSN 1532-4435.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, apr 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080%2F01621459.2017.1285773.

Fedor Borisyuk, Krishnaram Kenthapadi, David Stein, and Bo Zhao. Casmos: A framework for learning candidate selection models over structured queries and documents. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 441–450, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939718. URL https://doi.org/10.1145/2939672.2939718.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013. URL http://jmlr.org/papers/v14/bottou13a.html.

Guillaume Bouchard. Efficient bounds for the softmax function, applications to inference in hybrid models, 2007.

Nicolas Boulle and Alex Townsend. A generalization of the randomized singular value decomposition. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hgKtwSb4S2.

David Brandfonbrener, William Whitney, Rajesh Ranganath, and Joan Bruna. Offline contextual bandits with overparameterized models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1049–1058. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/brandfonbrener21a.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

Alexander Buchholz, Jan Malte Lichtenberg, Giuseppe Di Benedetto, Yannik Stein, Vito Bellini, and Matteo Ruffini. Low-variance estimation in the plackett-luce model via quasi-monte carlo sampling, 2022. URL https://arxiv.org/abs/2205.06024.

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 593–613, 2009.

Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *IMS Lecture Notes Monograph Series*, page 1–163, 2007. ISSN 0749-2170. doi: 10.1214/074921707000000391. URL http://dx.doi.org/10.1214/074921707000000391.

Chong Chen, Weizhi Ma, Min Zhang, Chenyang Wang, Yiqun Liu, and Shaoping Ma. Revisiting negative sampling vs. non-sampling in implicit recommendation. *ACM Trans. Inf. Syst.*, 41(1), feb 2023. ISSN 1046-8188. doi: 10.1145/3522672. URL https://doi.org/10.1145/3522672.

Jiawei Chen, Can Wang, Sheng Zhou, Qihao Shi, Jingbang Chen, Yan Feng, and Chun Chen. Fast adaptively weighted matrix factorization for recommendation with implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3470–3477, Apr. 2020. doi: 10.1609/aaai.v34i04.5751. URL https://ojs.aaai.org/index.php/AAAI/article/view/5751.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed 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*, WSDM '19, page 456–464, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450359405. doi: 10.1145/3289600.3290999. URL https://doi.org/10.1145/3289600.3290999.

Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/84899ae725ba49884f4c85c086f1b340-Paper.pdf.

Minmin Chen, Can Xu, Vince Gatto, Devanshu Jain, Aviral Kumar, and Ed Chi. Off-Policy Actor-Critic for Recommender Systems. In *Proceedings of the 16th ACM Conference on*

Recommender Systems, RecSys '22, page 338–349, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3546758. URL https://doi.org/10.1145/3523227.3546758.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014. URL http://aclweb.org/anthology/D/D14/D14-1179.pdf.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to Sequential Monte Carlo / Nicolas Chopin, Omiros Papaspiliopoulos.* Springer Series in Statistics. Springer, Cham, Switzerland, 1st ed. 2020. edition, 2020. ISBN 3-030-47845-9.

Konstantina Christakopoulou, Can Xu, Sai Zhang, Sriraj Badam, Trevor Potter, Daniel Li, Hao Wan, Xinyang Yi, Ya Le, Chris Berg, Eric Bencomo Dixon, Ed H. Chi, and Minmin Chen. Reward shaping for user satisfaction in a reinforce recommender, 2022. URL https://arxiv.org/abs/2209.15166.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584805. doi: 10.1145/1646396.1646452. URL https://doi.org/10.1145/1646396.1646452.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

Patrick L Combettes. Perspective Functions: Properties, Constructions, and Examples. *Set-Valued and Variational Analysis*, 26(2):247–264, 2018.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. Coindice: Off-policy confidence interval estimation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9398–9411. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6aaba9a124857622930ca4e50f5afed2-Paper.pdf.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996. PMLR, 2019.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019. URL http://jmlr.org/papers/v20/17-750.html.

John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Gabriel Dulac-Arnold, Richard Evans, Hado van Hasselt, Peter Sunehag, Timothy Lillicrap, Jonathan Hunt, Timothy Mann, Theophane Weber, Thomas Degris, and Ben Coppin. Deep reinforcement learning in large discrete action spaces. *arXiv preprint arXiv:1512.07679*, 2015.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In Jyrki Kivinen and Robert H. Sloan, editors, *Computational Learning Theory*, pages 255–270, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45435-9.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

Louis Faury, Ugo Tanielian, Flavian Vasile, Elena Smirnova, and Elvis Dohmatob. Distributionally robust counterfactual risk minimization. In *AAAI*, 2020.

Artyom Gadetsky, Kirill Struminsky, Christopher Robinson, Novi Quadrianto, and Dmitry Vetrov. Low-variance black-box gradient estimates for the plackett-luce distribution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06):10126–10135, Apr. 2020. doi: 10.1609/aaai.v34i06.6572. URL https://ojs.aaai.org/index.php/AAAI/article/view/6572.

F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proc. of the 8th ACM Conference on Recommender Systems*, RecSys '14, pages 169–176, 2014.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/garivier16a.html.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian Learning of Linear Classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 353–360, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553419. URL https://doi.org/10.1145/1553374.1553419.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, page 518–529, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606157.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, dec 1992. ISSN 0001-0782. doi: 10.1145/138859.138867. URL https://doi.org/10.1145/138859.138867.

Carlos A. Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4), dec 2016. ISSN 2158-656X. doi: 10.1145/2843948. URL https://doi.org/10.1145/2843948.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyzKd1bCW.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1eSS3CcKX.

Benjamin Guedj. A Primer on PAC-Bayesian Learning, 2019.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/1908.10396.

Somit Gupta, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, Sumita Chandran, Nanyu Chen, Dominic Coey, Mike Curtis, Alex Deng, Weitao Duan, Peter Forbes, Brian Frasca, Tommy Guy, Guido W. Imbens, Guillaume Saint Jacques, Pranav Kantawala, Ilya Katsev, Moshe Katzwer, Mikael Konutgan, Elena Kunakova, Minyong Lee, MJ Lee, Joseph Liu, James McQueen, Amir Najmi, Brent Smith, Vivek Trehan, Lukas Vermeer, Toby Walker, Jeffrey Wong, and Igor Yashkov. Top challenges from the first practical online controlled experiments summit. *SIGKDD Explor. Newsl.*, 21(1):20–35, may 2019. ISSN 1931-0145. doi: 10.1145/3331651.3331655. URL https://doi.org/10.1145/3331651.3331655.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Y.W. Teh and M. Titterington, editors, *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR W&CP*, pages 297–304, 2010.

Maxime Haddouche and Benjamin Guedj. Online PAC-bayes learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4pwCvvel8or.

Maxime Haddouche and Benjamin Guedj. PAC-bayes generalisation bounds for heavy-tailed losses through supermartingales. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=qxrwt6F3sf.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. ISSN 2160-6455. doi: 10.1145/2827872. URL https://doi.org/10.1145/2827872.

Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402, jan 2015. ISSN 1532-4435.

Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 549–558, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340694. doi: 10.1145/2911451.2911489. URL https://doi.org/10.1145/2911451.2911489.

David A. Hensher and William H. Greene. The mixed logit model: The state of practice. *Transportation*, 30(2):133–176, 2003. doi: 10.1023/A:1022558715350. URL https://doi.org/10.1023/A:1022558715350.

Miguel A Hernan and James M Robins. Causal inference, 2010.

Balázs Hidasi and Alexandros Karatzoglou. Recurrent neural networks with top-k gains for session-based recommendations. In Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang, editors, *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*, pages 843–852. ACM, 2018. ISBN 978-1-4503-6014-2. doi: 10.1145/3269206.3271761. URL https://doi.org/10.1145/3269206.3271761.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015.

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. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4337–4348. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/hsieh21a.html.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2333–2338, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450322638. doi: 10.1145/2505515.2505665. URL https://doi.org/10.1145/2505515.2505665.

Iris AM Huijben, Wouter Kool, Max B Paulus, and Ruud JG van Sloun. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning. *arXiv preprint arXiv:2110.01515*, 2021.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. doi: 10.1198/106186008X320456. URL https://doi.org/10.1198/106186008X320456.

Tommi Jaakkola and Michael Jordan. A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, volume 82, page 4, 1997.

Darko Janeković and Dario Bojanjac. Randomized algorithms for singular value decomposition: Implementation and application perspective. In *2021 International Symposium ELMAR*, pages 165–168, 2021. doi: 10.1109/ELMAR52657.2021.9550979.

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2240–2264. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/jang23a.html.

Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Trans. Manage. Inf. Syst.*, 10(4), dec 2019. ISSN 2158-656X. doi: 10.1145/3370082. URL https://doi.org/10.1145/3370082.

Olivier Jeunen and Bart Goethals. *Pessimistic Reward Models for Off-Policy Learning in Recommendation*, page 63–74. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450384582. URL https://doi.org/10.1145/3460231.3474247.

Olivier Jeunen, Jan Van Balen, and Bart Goethals. Closed-form models for collaborative filtering with side-information. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 651–656, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3418480. URL https://doi.org/10.1145/3383313.3418480.

Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJaP_-xAb.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *AISTATS*, 2018.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 227–234, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1238–1246, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/karnin13.html.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL https://openreview.net/group?id=ICLR.cc/2014.

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980. doi: 10.1109/TAC.1980.1102314.

David A. Knowles and Tom Minka. Non-conjugate variational message passing for multinomial and binary regression. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709. Curran Associates, Inc., 2011.

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

Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Ya Xu. Trustworthy online controlled experiments: Five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 786–794, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339653. URL https://doi.org/10.1145/2339530.2339653.

Yehuda Koren and Robert Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

Ilja Kuzborskij and Csaba Szepesvári. Efron-Stein PAC-Bayesian Inequalities, 2020.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvari. Confident off-policy evaluation and selection through self-normalized importance weighting. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 640–648. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/kuzborskij21a.html.

John D. Lafferty and David M. Blei. Correlated topic models. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 147–154. MIT Press, 2006. URL http://papers.nips.cc/paper/2906-correlated-topic-models.pdf.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, page 28, 2018.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Pierre L'Ecuyer. Randomized Quasi-Monte Carlo: An Introduction for Practitioners. In *12th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (MCQMC 2016)*, Stanford, United States, August 2016. URL https://hal.inria.fr/hal-01561550.

Gaël Letarte, Pascal Germain, Benjamin Guedj, and Francois Laviolette. Dichotomize and generalize: PAC-Bayesian binary activated deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7ec3b3cf674f4f1d23e9d30c89426cce-Paper.pdf.

Lihong Li, Wei Chu, John Langford, and Robert 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.

Xiangyang Li, Bo Chen, Huifeng Guo, Jingjie Li, Chenxu Zhu, Xiang Long, Sujian Li, Yichao Wang, Wei Guo, Longxia Mao, Jinxing Liu, Zhenhua Dong, and Ruiming Tang. Inttower: The next generation of two-tower model for pre-ranking system. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 3292–3301, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557072. URL https://doi.org/10.1145/3511808.3557072.

Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356398. doi: 10.1145/3178876.3186150. URL https://doi.org/10.1145/3178876.3186150.

Tie-Yan Liu. Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.*, 3(3): 225–331, mar 2009. ISSN 1554-0669. doi: 10.1561/1500000016. URL https://doi.org/10.1561/1500000016.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4125–4133. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/london19a.html.

Malte Ludewig and Dietmar Jannach. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 28(4-5):331–390, oct 2018. doi: 10.1007/s11257-018-9209-6. URL https://doi.org/10.1007%2Fs11257-018-9209-6.

Alberto Lumbreras, Louis Filstroff, and Cédric Févotte. Bayesian mean-parameterized nonnegative binary matrix factorization. *arXiv preprint arXiv:1812.06866*, 2018.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H. Chi. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web*

Conference 2020, WWW '20, page 463–473, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370233. doi: 10.1145/3366423.3380130. URL https://doi.org/10.1145/3366423.3380130.

Jiaqi Ma, Xinyang Yi, Weijing Tang, Zhe Zhao, Lichan Hong, Ed Chi, and Qiaozhu Mei. Learning-to-Rank with Partitioned Preference: Fast Estimation for the Plackett-Luce Model. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 928–936. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/ma21a.html.

Yifei Ma, Yu-Xiang Wang, and Balakrishnan Narayanaswamy. Imitation-regularized offline learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2956–2965. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/ma19b.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Yu A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, apr 2020. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2889473. URL https://doi.org/10.1109/TPAMI.2018.2889473.

Yury A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Vaden Masrani, Tuan Anh Le, and Frank Wood. The thermodynamic variational objective. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/618faa1728eb2ef6e3733645273ab145-Paper.pdf.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Annual Conference Computational Learning Theory*, 2009. URL https://api.semanticscholar.org/CorpusID:17090214.

David McAllester. Simplified PAC-Bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279989. URL https://doi.org/10.1145/279943.279989.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvari, and Dale Schuurmans. Escaping the gravitational pull of softmax. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21130–21140. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f1cf2a082126bf02de0b307778ce73a7-Paper.pdf.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6820–6829. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/mei20b.html.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8119–8132. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4476b929e30dd0c4e8bdbcc82c6ba23a-Paper.pdf.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes Un-Expected Bernstein Inequality. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3dea6b598a16b334a53145e78701fa87-Paper.pdf.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Thanh Nguyen-Tang, Sunil Gupta, A. Tuan Nguyen, and Svetha Venkatesh. Offline neural contextual bandits: Pessimism, optimization and generalization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=sPIFuucA3F.

Tui H Nolan and Matt P Wand. Accurate logistic variational message passing: algebraic and numerical details. *Stat*, 6(1):102–112, 2017.

Kento Nozawa, Pascal Germain, and Benjamin Guedj. PAC-Bayesian Contrastive Unsupervised Representation Learning. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 21–30. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/nozawa20a.html.

Harrie 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*, SIGIR '22, page 2266–2271, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531842. URL https://doi.org/10.1145/3477495.3531842.

Art Owen and Yi Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000. ISSN 01621459. URL http://www.jstor.org/stable/2669533.

Art B Owen. *Empirical likelihood.* CRC press, 2001.

Art B. Owen. *Monte Carlo theory, methods and examples.* https://artowen.su.domains/mc/, 2013.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. ISSN 00359254, 14679876. URL http://www.jstor.org/stable/2346567.

Sebastian Prillo and Julian Martin Eisenschlos. Softsort: A continuous relaxation for the argsort operator. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Jérémie Rappaz, Julian McAuley, and Karl Aberer. *Recommendation on Live-Streaming Platforms: Dynamic Availability and Repeat Consumption*, page 390–399. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450384582. URL https://doi.org/10.1145/3460231.3474267.

Ankit Singh Rawat, Jiecao Chen, Felix Xinnan X Yu, Ananda Theertha Suresh, and Sanjiv Kumar. Sampled softmax with random fourier features. *Advances in Neural Information Processing Systems*, 32, 2019.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 452–461, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, page 175–186, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916891. doi: 10.1145/192844.192905. URL https://doi.org/10.1145/192844.192905.

C. J. Van Rijsbergen. *Information Retrieval.* Butterworth-Heinemann, 2nd edition, 1979.

Ya'acov Ritov, Peter J Bickel, Anthony C Gamst, Bastiaan Jan Korneel Kleijn, et al. The bayesian analysis of complex, high-dimensional models: Can it be coda? *Statistical Science*, 29(4):619–639, 2014.

Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951a. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, 22(3):400–407, 1951b.

David Rohde and Matt P Wand. Semiparametric mean field variational bayes: General principles and numerical issues. *The Journal of Machine Learning Research*, 17(1):5975–6021, 2016.

David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. In *REVEAL workshop, ACM Conference on Recommender Systems 2018*, 2018.

David Rohde, Flavian Vasile, Sergey Ivanov, and Otmane Sakhi. Bayesian value based recommendation: A modelling based alternative to proxy and counterfactual policy based recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 742–744, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3411544. URL https://doi.org/10.1145/3383313.3411544.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, oct 2013. doi: 10.1613/jair.3987. URL https://doi.org/10.1613%2Fjair.3987.

Nicolas Le Roux. Tighter bounds lead to improved classifiers. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyAbMKwxe.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Francisco JR Ruiz, Michalis K Titsias, Adji B Dieng, and David M Blei. Augment and reduce: Stochastic inference for large categorical distributions. *arXiv preprint arXiv:1802.04220*, 2018.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19089–19122. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/saito22a.html.

Yuta Saito and Thorsten 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*, KDD '22, page 4824–4825, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3542601. URL https://doi.org/10.1145/3534678.3542601.

Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-Policy Evaluation for Large Action Spaces via Conjunct Effect Modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29734–29759. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/saito23b.html.

Otmane Sakhi, Stephen Bonner, David Rohde, and Flavian Vasile. Reconsidering analytical variational bounds for output layers of deep networks, 2019.

Otmane Sakhi, Stephen Bonner, David Rohde, and Flavian 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, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403121. URL https://doi.org/10.1145/3394486.3403121.

Otmane Sakhi, Louis Faury, and Flavian Vasile. Improving Offline Contextual Bandits with Distributional Robustness. In *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems (REVEAL '20)*, 2020b. URL https://arxiv.org/abs/2011.06835.

Otmane Sakhi, Pierre Alquier, and Nicolas Chopin. PAC-Bayesian Offline Contextual Bandits With Guarantees. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29777–29799. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/sakhi23a.html.

Otmane Sakhi, David Rohde, and Nicolas Chopin. Fast Slate Policy Optimization: Going Beyond Plackett-Luce. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856. URL https://openreview.net/forum?id=f7a8XCRtUu.

Otmane Sakhi, David Rohde, and Alexandre Gilotte. Fast Offline Policy Optimization for Large Scale Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8): 9686–9694, Jun. 2023c. doi: 10.1609/aaai.v37i8.26158. URL https://ojs.aaai.org/index.php/AAAI/article/view/26158.

Ruslan Salakhutdinov and Andriy Mnih. Probabilistic matrix factorization. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 1257–1264, Red Hook, NY, USA, 2007. Curran Associates Inc. ISBN 9781605603520.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 1670–1679, 2016.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

Yevgeny Seldin, Peter Auer, John Shawe-taylor, Ronald Ortner, and François Laviolette. PAC-Bayesian analysis of contextual bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/58e4d44e550d0f7ee0a23d6b02d9b0db-Paper.pdf.

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits. In Dorota Glowacka, Louis Dorard, and John Shawe-Taylor, editors, *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 98–111, Bellevue, Washington, USA, 02 Jul 2012. PMLR. URL https://proceedings.mlr.press/v26/seldin12a.html.

Li Shen, Yan Sun, Zhiyuan Yu, Liang Ding, Xinmei Tian, and Dacheng Tao. On efficient training of large-scale deep learning models: A literature review, 2023.

Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/310ce61c90f3a46e340ee8257bc70e93-Paper.pdf.

James E. Smith and Robert L. Winkler. The optimizers curse: Skepticism and postdecision surprise in decision analysis. *Manage. Sci.*, 52(3):311–322, mar 2006. ISSN 0025-1909. doi: 10.1287/mnsc.1050.0451. URL https://doi.org/10.1287/mnsc.1050.0451.

Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, page 421–430, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934472. doi: 10.1145/1180639.1180727. URL https://doi.org/10.1145/1180639.1180727.

Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, Ioannis Yiannis Kompatsiaris, Grigorios Tsoumakas, and Ioannis Vlahavas. A comprehensive study over VLAD and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 16(6):1713–1728, 2014. doi: 10.1109/TMM.2014.2329648.

Joe Staines and David Barber. Variational Optimization, 2012. URL https://arxiv.org/abs/1212.4507.

Harald Steck. Autoencoders that don't overfit towards the identity. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19598–19608. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/e33d974aae13e4d877477d51d8bafdc4-Paper.pdf.

A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 814–823, Lille, France, 07–09 Jul 2015a. PMLR. URL https://proceedings.mlr.press/v37/swaminathan15.html.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NIPS*, 2015b.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3635–3645, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Ugo Tanielian and Flavian Vasile. Relaxed Softmax for PU Learning. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 119–127, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347034. URL https://doi.org/10.1145/3298689.3347034.

Ambuj Tewari and Susan A Murphy. From Ads to Interventions: Contextual Bandits in Mobile Health. In *Mobile Health*, pages 495–517. Springer, 2017.

Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-Confidence Off-Policy Evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Optimistic pac reinforcement learning: the instance-dependent view. In Shipra Agrawal and Francesco Orabona, editors, *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pages 1460–1480. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/tirinzoni23a.html.

Michalis Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*, pages 4161–4169, 2016.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, nov 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral Bandits for Smooth Graph Functions. In *International Conference on Machine Learning*, pages 46–54, 2014.

V. Vapnik. Principles of risk minimization for learning theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, page 831–838, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc. ISBN 1558602224.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Flavian Vasile, Elena Smirnova, and Alexis Conneau. Meta-prod2vec: Product embeddings using side-information for recommendation. In *Proceedings of the 10th ACM Conference on*

Recommender Systems, RecSys '16, page 225–232, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359. doi: 10.1145/2959100.2959160. URL https://doi.org/10.1145/2959100.2959160.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

Paul Viallard, Pascal Germain, Amaury Habrard, and Emilie Morvant. A General Framework for the Practical Disintegration of PAC-Bayesian Bounds, 2023.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Mengting Wan and Julian J. McAuley. Item recommendation on monotonic behavior chains. In Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan, editors, *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM, 2018. doi: 10.1145/3240323.3240369. URL https://doi.org/10.1145/3240323.3240369.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1248. URL https://doi.org/10.18653/v1/p19-1248.

Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *Proc. VLDB Endow.*, 14(11):1964–1978, jul 2021a. ISSN 2150-8097. doi: 10.14778/3476249.3476255. URL https://doi.org/10.14778/3476249.3476255.

Shoujin Wang, Longbing Cao, Yan Wang, Quan Z. Sheng, Mehmet A. Orgun, and Defu Lian. A survey on session-based recommender systems. *ACM Comput. Surv.*, 54(7), jul 2021b. ISSN 0360-0300. doi: 10.1145/3465401. URL https://doi.org/10.1145/3465401.

Xuanhui Wang, Cheng Li, Nadav Golbandi, Mike Bendersky, and Marc Najork. The LambdaLoss Framework for Ranking Metric Optimization. In *Proceedings of The 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, pages 1313–1322, 2018.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H. Chi, and Minmin Chen. Surrogate for Long-Term User Experience in Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4100–4109, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539073. URL https://doi.org/10.1145/3534678.3539073.

Robert West, Smriti Bhagat, Paul Groth, Marinka Zitnik, Francisco M. Couto, Pasquale Lisena, Albert Meroño Peñuela, Xiangyu Zhao, Wenqi Fan, Dawei Yin, Jiliang Tang, Linjun Shou, Ming Gong, Jian Pei, Xiubo Geng, Xingjie Zhou, Daxin Jiang, Benjamin Ricaud, Nicolas Aspert, Volodymyr Miz, Jennifer Dy, Stratis Ioannidis, undefinedlkay Yıldız, Rezvaneh Rezapour, Samin Aref, Ly Dinh, Jana Diesner, Alexey Drutsa, Dmitry Ustalov, Nikita Popov, Daria Baidakova, Shubhanshu Mishra, Arjun Gopalan, Da-Cheng Juan, Cesar Ilharco Magalhaes, Chun-Sung Ferng, Allan Heydon, Chun-Ta Lu, Philip Pham, George Yu, Yicheng Fan, Yueqi Wang, Florian Laurent, Yanick Schraner, Christian Scheller, Sharada Mohanty, Jiawei Chen, Xiang Wang, Fuli Feng, Xiangnan He, Irene Teinemaa, Javier Albert, Dmitri Goldenberg, Flavian Vasile, David Rohde, Olivier Jeunen, Amine Benhalloum, Otmane Sakhi, Yu Rong, Wenbing Huang, Tingyang Xu, Yatao Bian, Hong Cheng, Fuchun Sun, Junzhou Huang, Shobeir Fakhraei, Christos Faloutsos, Onur Çelebi, Martin Müller, Manuel Schneider, Olesia Altunina, Wolfram Wingerath, Benjamin Wollmer, Felix Gessert, Stephan Succo, Norbert Ritter, Evann Courdier, Tudor Mihai Avram, Dragan Cvetinovic, Levan Tsinadze, Johny Jose, Rose Howell, Mario Koenig, Michaël Defferrard, Krishnaram Kenthapadi, Ben Packer, Mehrnoosh Sameki, and Nashlie Sephus. Summary of tutorials at the web conference 2021. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 727–733, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383134. doi: 10.1145/3442442.3453701. URL https://doi.org/10.1145/3442442.3453701.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992. doi: 10.1007/BF00992696.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):346–353, jul 2019. doi: 10.1609/aaai.v33i01.3301346. URL https://doi.org/10.1609%2Faaai.v33i01.3301346.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank: Theory and Algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1192–1199, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390306. URL https://doi.org/10.1145/1390156.1390306.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Bolin Ding, and Bin Cui. Contrastive learning for sequential recommendation, 2021.

Ming Xu, Matias Quiroz, Robert Kohn, and Scott A. Sisson. Variance reduction properties of the reparameterization trick. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2711–2720. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/xu19a.html.

Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 279–287, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450359016. doi: 10.1145/3240323.3240355. URL https://doi.org/10.1145/3240323.3240355.

Amir Zandieh, Insu Han, Majid Daliri, and Amin Karbasi. KDEformer: Accelerating transformers via kernel density estimation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 40605–40623. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zandieh23a.html.

Eva Zangerle and Christine Bauer. Evaluating recommender systems: Survey and framework. *ACM Comput. Surv.*, 55(8), dec 2022. ISSN 0360-0300. doi: 10.1145/3556536. URL https://doi.org/10.1145/3556536.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res.*, 2:527–550, 2002. URL http://dblp.uni-trier.de/db/journals/jmlr/jmlr2.html#Zhang02.

Xiaoying Zhang, Junzhou Zhao, and John C.S. Lui. Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 98–106, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346528. doi: 10.1145/3109859.3109885. URL https://doi.org/10.1145/3109859.3109885.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002. ISSN 0885-064X. doi: https://doi.org/10.1006/jcom.2002.0635. URL https://www.sciencedirect.com/science/article/pii/S0885064X02906357.