Contribution à des problèmes statistiques d'ordonnancement et d'apprentissage par renforcement avec aversion au risque

Mastane Achab

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAT020

Thèse de doctorat

Ranking and Risk-Aware Reinforcement Learning

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

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

Thèse présentée et soutenue à Palaiseau, le 10 juillet 2020, par

MASTANE ACHAB

Composition du Jury :

Florence d'Alché-Buc Professeur, Télécom Paris (LTCI)	Présidente
Alexandra Carpentier Professeur, Otto-von-Guericke-Universität Magdeburg (IMST)	Rapporteure
Michal Valko Research Scientist, Google DeepMind / Inria (SequeL)	Rapporteur
Gergely Neu Research Assistant Professor, Universitat Pompeu Fabra (AI&ML)	Examinateur
Stéphan Clémençon Professeur, Télécom Paris (LTCI)	Directeur de thèse
Aurélien Garivier Professeur, École Normale Supérieure de Lyon (UMPA)	Co-directeur de thèse
Vianney Perchet Professeur, ENSAE Paris (CREST)	Invité
Bilal Piot Research Scientist, Google DeepMind	Invité

À ma mère Samia, mon père Ramdane, et mon grand frère Massil

Remerciements

Mes premières pensées vont à mes parents Samia et Ramdane et à mon frère Massil qui m'ont soutenu tout au long de mon cursus scolaire, ainsi qu'à ma famille aléatoirement dispersée entre l'Algérie, la France et le Canada. Plus spécifiquement concernant mon doctorat, je tiens en premier lieu à exprimer ma plus profonde gratitude envers mon directeur de thèse Stéphan Clémençon, notamment pour sa confiance, sa disponibilité, et la grande variété de sujets de recherche que j'ai eu la chance d'étudier à ses côtés. Je remercie ensuite mes deux co-directeurs de thèse : Aurélien Garivier, à qui j'ai plusieurs fois rendu visite à Toulouse et à Lyon à l'occasion d'intenses séances de travail, et Anne Sabourin, en particulier pour son précieux soutien durant mon stage de recherche à Télécom Paris puis en début de thèse. Merci au LabEx mathématique Hadamard d'avoir financé ma thèse, à mon laboratoire le LTCI et à Télécom Paris dans sa globalité : professeurs, personnels administratifs, logistiques, élèves ; et plus particulièrement, mes collègues doctorants avec qui j'ai eu le plaisir de partager ces quatre années. Je suis très reconnaissant envers Alexandra Carpentier et Michal Valko pour l'intérêt qu'ils ont porté à mon travail en acceptant de rapporter ma thèse. Je suis également très honoré de la présence de Florence d'Alché-Buc, Gergely Neu, Vianney Perchet et Bilal Piot dans mon jury de thèse, et ce malgré les contraintes dues à la COVID-19. Je remercie Bilal Piot, Claire Vernade et Olivier Pietquin pour leurs précieux conseils à propos de mon semestre de césure, au cours duquel j'ai réalisé un stage de recherche de cinq mois (de mars à juillet 2019) chez Google DeepMind à Londres. Bilal, merci encore une fois d'avoir accepté de m'encadrer pendant ce stage qui fut l'occasion d'élargir mon horizon scientifique et de prendre du recul sur ma thèse. Merci à ceux qui ont aidé à l'organisation de mon pot de thèse : mes parents, mon frère, tonton Saïd du Kremlin, le zin Yani, ainsi que mes acolytes du labo Robin Vogel et Hamid Jalalzai. Je rends aussi hommage à tous ceux (famille, amis) qui ont vibré durant la retransmission vidéo de ma soutenance sur Zoom. Enfin, grosses salutations à la cellule cachanaise, à l'équipe une du Val de Bièvre, à mes coéquipiers de la section 'Foot 2012' de Polytechnique, et bien sûr au premier rond-point de Tizi Ouzou.

Publications

- [ACG+17] M. Achab, S. Clémençon, A. Garivier, A. Sabourin, C. Vernade, *Max K-Armed Bandit: On the ExtremeHunter Algorithm and Beyond*, European Conference on Machine Learning, 2017.
- [CA17] S. Clémençon, M. Achab, *Ranking Data with Continuous Labels through Oriented Recursive Partitions*, Advances in Neural Information Processing Systems, 2017.
- [ACG18] M. Achab, S. Clémençon, A. Garivier, *Profitable Bandits*, Asian Conference on Machine Learning, 2018.
- [AKC18] M. Achab, A. Korba, S. Clémençon, *Dimensionality Reduction and (Bucket) Ranking: a Mass Transportation Approach*, Algorithmic Learning Theory, 2019.
- [VACT20] R. Vogel, M. Achab, S. Clémençon, C. Tillier, *Weighted Empirical Risk Minimization: Sample Selection Bias Correction based on Importance Sampling*, International Conference on Machine Learning, Artificial Intelligence and Applications, 2020.

Contents

List of Figures

List of Tables

CHAPTER I

Introduction

Outline

This thesis divides into two parts: the first part tackles various *offline* learning problems, with a focus on ranking tasks, based on empirical risk minimization, while the second part is about learning from *online* streams of data through the reinforcement learning framework. Each part is composed of three chapters.

- The introductory chapter I presents the different problems and frameworks considered along the following chapters, as well as the links between them. In addition, the contributions of each of the six chapters are summarized.

Part 1 focuses on empirical risk minimization and ranking.

- Chapter II tackles a specific transfer learning issue: when the training and testing distributions are different, the empirical risk minimization approach can be corrected by means of importance sampling weights. This chapter is based on the conference paper [VACT20] accepted for publication at ICMA 2020.
- Chapter III considers the continuous ranking problem, formulated as a continuum of bipartite ranking subproblems: optimal scoring rules are maximizing integrated versions of the usual ROC and AUC performance measures. It is based on the conference paper [CA17] (NIPS 2017).
- Chapter IV extends the ranking aggregation setting to the problem of dimensionality reduction for ranking data: a mass transportation approach is proposed to approximate a distribution on the symmetric group by a simpler distribution satisfying a bucket order structure. This work was published in the conference paper [AKC18] at ALT 2019.

Part 2 deals with risk-awareness in several reinforcement learning problems.

- Chapter V describes a variant of the classical multi-armed bandit problem inspired by default risk management applications. This chapter is based on the conference paper [ACG18] published at ACML 2018.

- Chapter VI contributes to the max K-armed bandit problem, motivated by risk-aware contexts where extreme rewards are more relevant than expected gains. It is based on the conference paper [ACG+17] (ECML 2017).

- Chapter VII considers the distributional reinforcement learning setting: we derive new 'atomic' Bellman equations by combining novel distributional Bellman operators with an atomic approximation scheme based on trimmed means. The development of this final chapter started during an internship at Google DeepMind (London) from March to July 2019. This chapter is the unique one containing unpublished content.

1 Ranking by Empirical Risk Minimization

We start by briefly recalling the *Empirical Risk Minimization* framework where one intends to minimize a loss function in expectation with respect to some *testing distribution* P, based on the observation of independent realizations of P. Then, we present our contributions to the sample selection bias correction problem, where the observations are sampled from a *training distribution* P' different from P.

Throughout the thesis, the notation $X \sim P$ means that P is the probability distribution of the random variable X.

1.1 Empirical Risk Minimization

The main paradigm of predictive learning is *Empirical Risk Minimization* (ERM in abbreviated form), see *e.g.* [DGL96]. In the standard setup, Z is a random variable (r.v. in short) that takes its values in a space $\mathcal{Z}$ with distribution P, Θ is a parameter space and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is a (measurable) loss function. The *risk* is then defined by: $\forall \theta \in \Theta$,

$$\mathcal{R}_P(\theta) = \mathbb{E}_P\left[\ell(\theta, Z)\right], \tag{I.1}$$

and more generally for any measure Q on $\mathcal{Z}$: $\mathcal{R}_Q(\theta) = \int_{\mathcal{Z}} \ell(\theta, z) dQ(z)$. In most practical situations, the distribution P involved in the definition of the risk is unknown and learning is based on the sole observation of an independent and identically distributed (i.i.d.) sample $Z_1, \ldots, Z_n$ drawn from P and the risk (I.1) must be replaced by an empirical counterpart, typically:

$$\widehat{\mathcal{R}}_P(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta), \tag{I.2}$$

where $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{Z_i}$ is the empirical measure of P and δ_z denotes the Dirac measure at any point z. Any empirical minimizer $\widehat{\theta}_n \in \arg\min_{\theta\in\Theta} \widehat{\mathcal{R}}_P(\theta)$ is then accessible to the learner — in fact, not necessarily true in practice as the minimization of the empirical risk may be intractable — and may be used as a substitute to the (unknown) optimal parameters $\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}_P(\theta)$ living in the hypothesis class Θ.

Example 1. (Binary Classification) *In* binary classification*, the flagship problem in machine learning (see e.g. [DGL96]), the goal is to find a* classifier $g : \mathcal{X} \to \{-1, +1\}$ *with classification risk* $\mathcal{R}_P(g) = \mathbb{P}\{g(X) \neq Y\}$ *as small as possible. The random pair* $Z = (X, Y)$*, with distribution* P*, is valued in* $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$*, and the* feature space $\mathcal{X}$ *is typically a subset of* $\mathbb{R}^d$ *(*$d \geq 1$*) ;* X *is called the* feature vector *and* Y *is the* label*. Denoting the* posterior probability *by* $\eta(x) = \mathbb{P}\{Y = +1 | X = x\}$ *for all* $x \in \mathcal{X}$*, the Bayes classifier* $g^*(x) = 2\mathbb{I}\{\eta(x) > \frac{1}{2}\} - 1$ *is the optimal classification rule as it minimizes* $\mathcal{R}_P$*: for any classifier* g*,* $\mathcal{R}_P(g) \geq \mathcal{R}_P(g^*)$ *(see Theorem 2.1 in [DGL96]). Empirically, the learner is given a* training dataset *composed of* n *i.i.d. copies* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *of* (X, Y)*. Using the notations introduced above,*

- *the parameter space* Θ *is a set* $\mathcal{G}$ *of classifiers* g*,*
- *the loss function* ℓ *is the zero-one loss function* $\ell_{0/1}$*:*

$$\forall (g, x, y) \in \mathcal{G} \times \mathcal{X} \times \{-1, +1\}, \quad \ell_{0/1}(g, (x, y)) = \mathbb{I}\{g(x) \neq y\}.$$

Binary classification belongs to the family of *supervised learning* problems as it attempts to learn how to label any new unlabeled observation X, on the basis of labeled examples (X_i, Y_i)'s provided by some 'teacher'. In contrast, the *clustering* task is an *unsupervised learning* problem. Indeed, it consists in finding similarity groups among the feature space without any label information.

Example 2. (k-Means Clustering) *Given a number of clusters* $k \geq 1$*, the* k-means clustering *approach (see e.g. [Bis06] or [HTF09]) solves the following minimization problem:*

$$\min_{(m_1, \ldots, m_k) \in \mathcal{Z}^k} \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||Z_i - m_j||_2^2, \tag{I.3}$$

with $\mathcal{Z} \subseteq \mathbb{R}^d$ *and* $|| \cdot ||_2$ *the Euclidean norm. This quantity measures the total squared Euclidean distance of each observation* Z_i *to its cluster center, namely the nearest point* $m_{j(Z_i)}$*, with* $j(Z_i) \in arg\ min_{1 \leq j \leq k} ||Z_i - m_j||_2$*, among the* k *centroids* $m_1, \ldots, m_k$ *(in fact here,* $m_{j(Z_i)}$ *is not necessarily unique). With our ERM notations, we have that:*

- *the parameter space is the set of* k*-tuples* $(m_1, \ldots, m_k)$*:* $\Theta = \mathcal{Z}^k$*,*
- *the* k*-means clustering loss function is:*

$$\forall ((m_1, \ldots, m_k), z) \in \mathcal{Z}^k \times \mathcal{Z}, \quad \ell((m_1, \ldots, m_k), z) = \min_{1 \leq j \leq k} ||z - m_j||_2^2.$$

In particular, the k*-means criterion in Eq. (I.3) is, up to normalization, equal to the empirical risk* $\widehat{\mathcal{R}}_P$*:*

$$\widehat{\mathcal{R}}_P((m_1, \ldots, m_k)) = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} ||Z_i - m_j||_2^2.$$

The performance of minimizers $\widehat{\theta}_n$ of (I.2) can be studied by controlling the *excess of risk* $\mathcal{R}(\widehat{\theta}_n) - \min_{\theta\in\Theta} \mathcal{R}(\theta)$, which satisfies the elementary inequality (see e.g. [BBL05])

$$\mathcal{R}_P(\widehat{\theta}_n) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) = \mathcal{R}_P(\widehat{\theta}_n) - \widehat{\mathcal{R}}_P(\widehat{\theta}_n) + \widehat{\mathcal{R}}_P(\widehat{\theta}_n) - \mathcal{R}_P(\theta^*)$$
$$\leq \mathcal{R}_P(\widehat{\theta}_n) - \widehat{\mathcal{R}}_P(\widehat{\theta}_n) + \widehat{\mathcal{R}}_P(\theta^*) - \mathcal{R}_P(\theta^*) \leq 2\sup_{\theta\in\Theta} |\widehat{\mathcal{R}}_P(\theta) - \mathcal{R}_P(\theta)|. \quad \text{(I.4)}$$

The fluctuations of the maximal deviations $\sup_{\theta\in\Theta} |\widehat{\mathcal{R}}_P(\theta) - \mathcal{R}_P(\theta)|$ in Eq. (I.4) can then be quantified by means of *concentration inequalities*, under various complexity assumptions for the functional class $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$ (*e.g.* VC dimension, metric entropies, Rademacher averages), see [BLM13] for instance.

Sometimes, the training sample is drawn from a training distribution P' different from the target distribution P of interest: this is called *sample selection bias*. Our approach to deal with this situation is called WERM for 'Weighted Empirical Risk Minimization', it relies on a reweighting step through the estimation of importance sampling weights for each observation of the training dataset.

1.2 Sample Selection Bias Correction

Bias selection issues in machine-learning, often resulting from errors during the data acquisition process, are now the subject of much attention in the literature, see [BCZ+16], [ZWY+17], [BHS+19], [LYLW16] or [HGB+07]. We consider the case where the i.i.d. sample $Z'_1, \ldots, Z'_n$ available for training is not drawn from P but from another distribution P', with respect to which P is absolutely continuous, and the goal pursued is to set theoretical grounds for the application of ideas behind Importance Sampling (IS in short) methodology to extend the ERM approach to this learning setup. IS methods are broadly used in machine learning, including in online learning contexts such as bandit problems (see [NB16]), which are presented in part 2. We highlight that the problem under study is a very particular case of *Transfer Learning* (see *e.g.* [PY10], [BDBC+10] and [Sto09]), a research area currently receiving much attention in the literature.

Figure I.1 depicts an example of such sample selection bias in a classification context: the training dataset is composed of pictures of four types of animals (dog, wolf, tiger and monkey), whereas the target population is simply a mixture of dogs and wolfs. In other words, the training labels Y' are valued in $\mathcal{Y} = \{\text{dog,wolf,tiger,monkey}\}$, while, for the testing/target distribution, Y takes its values in only a subset $\{\text{dog}, \text{wolf}\} \subset \mathcal{Y}$. The histogram levels represent the class probabilities: $\mathbb{P}\{Y' = y\}$ in blue for training, and $\mathbb{P}\{Y = y\}$ in green for testing, for each animal $y \in \mathcal{Y}$. We formulate below the WERM method to deal with these sample selection bias issues.

Weighted ERM (WERM). The *Weighted Empirical Risk Minimization* (WERM) approach that we propose in chapter II consists in minimizing a weighted version of the empirical risk. We investigate conditions guaranteeing that values for the parameter θ that nearly minimize (I.1) can be obtained through minimization of a weighted version of the empirical risk based on the Z'_i's, namely

$$\widetilde{\mathcal{R}}_{w,n}(\theta) = \mathcal{R}_{\widetilde{P}_{w,n}}(\theta), \quad \text{(I.5)}$$

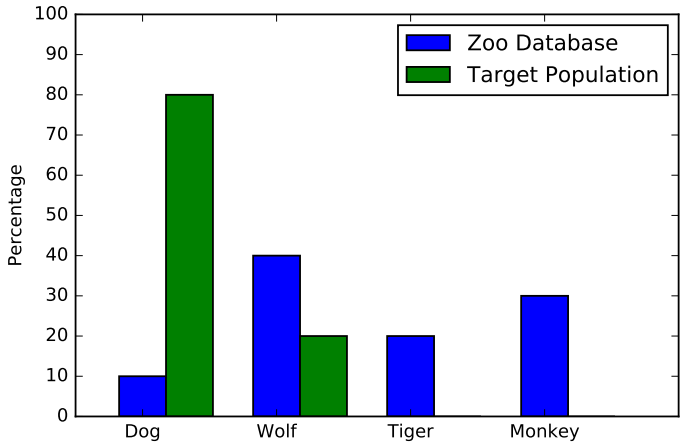


Figure I.1: Example of sample selection bias: for each type of animal in the zoo, the percentage of pictures in the image database is represented in blue, while its proportion in the target population of a video surveillance system, aiming at distinguishing wolfs from dogs (binary classification task), is in green.

where $\widetilde{P}_{w,n} = (1/n)\sum_{i=1}^{n} w_i \delta_{Z_i'}$ and $w = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ is a certain weight vector. The ideal weights w^* are given by the likelihood function $\Phi(z) = (dP/dP')(z)$: $w_i^* = \Phi(Z_i')$ for $i \in \{1, \ldots, n\}$. In this case, the quantity (I.5) is an unbiased estimate of the true risk (I.1):

$$\mathbb{E}_{P'}\left[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)\right] = \mathcal{R}_P(\theta), \tag{I.6}$$

and generalization bounds for the $\mathcal{R}_P$-risk excess of minimizers of $\widetilde{\mathcal{R}}_{w^*,n}$ can be directly established by studying the concentration properties of the empirical process related to the Z_i''s and the class of functions $\{\Phi(\cdot)\ell(\theta, \cdot) : \theta \in \Theta\}$. However, the *importance function* Φ is unknown in general, just like distribution P.

In Figure I.1, corresponding to a classification problem with training dataset $Z_1' = (X_1', Y_1'), \ldots, Z_n' = (X_n', Y_n')$, X_i' being a picture (vector of pixels in $[0,1]^d$ for instance) of an animal of type $Y_i' \in \mathcal{Y} = \{\text{dog}, \text{wolf}, \text{tiger}, \text{monkey}\}$, the likelihood function is given by:

$$\forall (x,y) \in [0,1]^d \times \mathcal{Y}, \quad \Phi((x,y)) = \frac{80\%}{10\%}\mathbb{I}\{y = \text{dog}\} + \frac{20\%}{40\%}\mathbb{I}\{y = \text{wolf}\},$$

which only depends on y, if we assume that the conditional distribution of $Z = (X, Y) \sim P$ given $Y = y$ is the same as that of $Z' = (X', Y') \sim P'$ given $Y' = y$, for all $y \in \mathcal{Y}$ (i.e. P and P' are both mixtures of the same 4 components/animals but with different weights).

Contributions. Our main contribution to this problem is to show that, in far from uncommon situations, the (ideal) weights w_i^* can be estimated from the Z_i's combined

with auxiliary information on the target population P. In particular, such favorable cases include:

- classification problems where class probabilities in the test stage differ from those in the training step (as in Figure I.1),
- risk minimization in stratified populations (see [BD18]), with strata statistically represented in a different manner in the test and training populations,
- positive-unlabeled learning (PU-learning, see *e.g.* [dPNS14]), which consists in solving a binary classification problem based on positive and unlabeled data solely.

In each of these cases, we show that the stochastic process obtained by plugging the weight estimates in the weighted empirical risk functional (I.5) is much more complex than a simple empirical process (*i.e.* a collection of i.i.d. averages) but can be however studied by means of *linearization techniques*, in the spirit of the ERM extensions established in [CLV08] or [CV09a]. Learning rate bounds for minimizers of the corresponding risk estimate are proved and, beyond these theoretical guarantees, the performance of the weighted ERM approach is supported by convincing numerical results.

1.3 The Bipartite Ranking Problem

We now introduce another ERM problem, of *global* nature contrary to binary classification: the *bipartite ranking* problem ([AGH$^+$05],[FISS03]), where one wants to order, by means of scoring methods, all the elements of the feature space $\mathcal{X}$, given a training dataset composed of i.i.d. copies of a random pair (X, Y) valued in $\mathcal{X} \times \{-1, +1\}$. Informally, good scoring rules are mappings $s : \mathcal{X} \to \mathbb{R}$ attributing large scores $s(x)$ to the elements $x \in \mathcal{X}$ with large posterior probability $\mathbb{P}\{Y = +1 | X = x\}$. Bipartite ranking finds many practical applications (see e.g. [CDV13b]), ranging from medical studies, where patients are ranked based on their probability of being ill, to recommendation systems ordering a catalogue of products based on some user's preferences. See for example [BK07] about some *movie recommendation* methods used during the 'Netflix Prize competition'. Another application of interest of ranking is the *credit-risk screening* problem, which will also serve as a motivation to our *profitable bandits* approach developed in chapter V in an online learning context. We recall below a few important notions for bipartite ranking before introducing our contributions to the *continuous ranking* problem, a generalization of bipartite ranking to the case of labels Y taking continuous values.

Formal Setup. The probabilistic framework of bipartite ranking is the same as in binary classification (example 1). Indeed, we consider a random variable (X, Y) valued in $\mathcal{X} \times \{-1, +1\}$, with feature space $\mathcal{X} \subseteq \mathbb{R}^d$ ($d \geq 1$), and with distribution P characterized by the pair (μ, η), where

- the *marginal distribution* of X denoted by μ,
- the *posterior probability* for all $x \in \mathcal{X}$ by

$$\eta(x) = \mathbb{P}\{Y = +1 | X = x\} = \frac{1}{2}(\mathbb{E}[Y | X = x] + 1).$$

Equivalently, P can be described by the triplet (p, G, H), where

- $p = \mathbb{P}\{Y = +1\}$ is the probability of occurence of a positive instance,
- G and H are respectively the conditional distributions of X given $Y = +1$ and of X given $Y = -1$.

The empirical problem of interest is the following: given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of $n \geq 1$ i.i.d. copies of (X, Y), a learning agent wants to select a scoring rule, i.e. a measurable map $s : \mathcal{X} \to \mathbb{R}$, ordering any new i.i.d. unlabeled sample $X'_1, \ldots, X'_{n'}$ (with common distribution μ) such that, with high probability, the observations X'_t with large scores $s(X'_t)$ have positive (unobserved) labels $Y'_t = +1$ more often than the observations with smaller scores. This framework arises in many applications, among which music recommendation systems (see [SDP12]).

Example 3. (Music Recommendation) *Bipartite ranking can be used to build a music recommendation system. Consider a collection $\mathcal{X}$ of songs, each song $x \in \mathcal{X}$ being modelled by d coordinates $x = (x_1, \ldots, x_d)$: $x_1 =$'title', $x_2 =$'artist', $x_3 =$'duration', etc., for instance. A user of the music platform produces a training dataset composed of n pairs song-feedback $(X_1, Y_1), \ldots, (X_n, Y_n)$ with Y_i a binary feedback produced by the user either equal to $+1$ if he enjoyed the song $X_i \in \mathcal{X}$, or else $Y_i = -1$ if he disliked it. Based on this partial information, the music platform may want to predict the preferences of the user over the whole catalogue of songs $\mathcal{X}$, by giving to each song $x \in \mathcal{X}$ a score $s_{user}(x)$. Then, a good score function s_{user} would give large scores to the songs that the user is likely to appreciate.*

Formally, for two real-valued random variables U and U', we recall that U is said to be *stochastically larger* than U' if $\mathbb{P}\{U \geq t\} \geq \mathbb{P}\{U' \geq t\}$ for all $t \in \mathbb{R}$. Then, the objective is to learn a scoring function s such that the r.v. $s(X)$ given $Y = +1$ is as stochastically larger as possible than the r.v. $s(X)$ given $Y = -1$. In other words, we want s to maximize the difference between $1 - G_s(t) = \bar{G}_s(t) = \mathbb{P}\{s(X) \geq t | Y = +1\}$ and $1 - H_s(t) = \bar{H}_s(t) = \mathbb{P}\{s(X) \geq t | Y = -1\}$ for all $t \in \mathbb{R}$. This functional criterion can also be expressed by means of the ROC curve of any scoring rule s, i.e. the parametrized curve $t \in \mathbb{R} \mapsto (\bar{H}_s(t), \bar{G}_s(t))$, or equivalently the graph of the mapping

$$\alpha \in (0, 1) \mapsto \mathrm{ROC}_s(\alpha) = \bar{G}_s \circ (1 - H_s^{-1})(1 - \alpha),$$

where possible discontinuity points are connected by linear segments. Indeed, the optimal elements s^* are those whose ROC curve $\mathrm{ROC}_{s^*} = \mathrm{ROC}^*$ dominates any other ROC curve ROC_s everywhere:

$$\forall \alpha \in (0, 1), \quad \mathrm{ROC}^*(\alpha) \geq \mathrm{ROC}_s(\alpha).$$

See Figure I.2 for an example. It is well known that optimal scoring functions s^* are strictly increasing transforms of the posterior probability function η (see e.g. [CLV05]).

Given its functional nature, the ROC curve ROC_s is often summarized by a simple scalar quantity, namely the area under it called the *Area Under the ROC Curve* (AUC in short):

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X) < s(X')|Y = -1, Y' = +1\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')|Y = -1, Y = +1\},$$

where (X', Y') is an i.i.d. copy of (X, Y). Indeed, when the ROC curves of two scoring functions s_1 and s_2 are crossing as in Figure I.2, neither can be considered better than the other, or even equal, from a ROC perspective. On the contrary, a global scalar criterion such as the AUC always allow to compare two scoring rules. Interestingly, the AUC comes with a probabilistic interpreation: it is the theoretical rate of concording pairs. The usual ERM approach for bipartite ranking consists in maximizing the empirical version of the AUC, given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$:

$$\widehat{\mathrm{AUC}}_n(s) = \frac{1}{n_+ \cdot n_-} \sum_{i:Y_i=-1} \sum_{j:Y_j=+1} \mathbb{I}\{s(X_i) < s(X_j)\} + \frac{1}{2}\mathbb{I}\{s(X_i) = s(X_j)\}, \qquad \text{(I.7)}$$

with $n_+ = \sum_{i=1}^n \mathbb{I}\{Y_i = +1\} = n - n_-$. Notice that the empirical AUC in Eq. (I.7) is a sum of dependent variables: more precisely, it is a U-statistic of degree 2 (see [CLV08]). Several algorithms based on the maximization of $\widehat{\mathrm{AUC}}_n$ have been proposed and studied in the literature, such as the TreeRank approach ([CV09b]). Extension to the case of label Y taking at least three ordinal values, called *multipartite ranking*, has also been investigated ([RA05], [SCV13]): we introduce next our contribution to the more general problem of *continuous ranking*, where Y is valued in the whole interval $[0, 1]$.

1.4 Ranking Data with Continuous Labels

In chapter III, we consider a ranking task akin to bipartite ranking, the difference lying in the nature of the label Y, whose support spreads over a continuum of scalar values: we call this problem *continuous ranking*. Depending on the context, Y may be represent a size, a biological measurement, or the cash flow of companies in quantitative finance. We describe below a potential application of continuous ranking for music recommendation, adapted from example 3 in the case of bipartite ranking.

Example 4. (Advanced Music Recommendation)
• *As in example 3, a music platform wants to smartly recommend the songs of playlist $\mathcal{X}$ to some user, by means of a scoring rule $s_{user} : \mathcal{X} \to \mathbb{R}$ specific to that user. Here also, the user generates a training dataset $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ after listening to n songs $X_i \in \mathcal{X}$. Nevertheless, each label Y_i now corresponds to the quantity of dopamine (a.k.a. the 'pleasure chemical') released by the user's brain — and measured by some sensor — while listening to the i-th song. Hence, these labels Y_i are not binary feedbacks/ratings as in bipartite ranking but are rather taking continuous values. Still, the objective for the recommender remains similar: giving large scores $s_{user}(x)$ to the songs $x \in \mathcal{X}$ that are likely to release a lot of dopamine in the user's brain.*

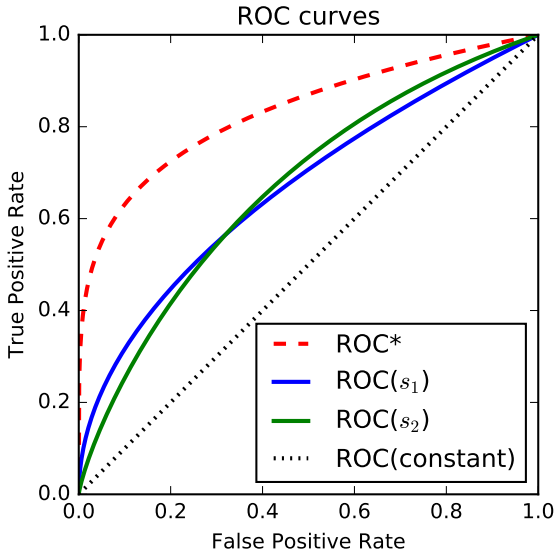


Figure I.2: The optimal ROC curve ROC* (dashed line) is uniformly above any other ROC curve. Both scores s_1 and s_2 perform better than any constant scoring rule, whose ROC curve ROC(constant) is simply the (dotted) line joining the points $(0,0)$ and $(1,1)$.

• *A more realistic example relies on* implicit feedbacks*, in particular the user's action 'skip the current song', that have received much attention in the literature recently ([RFGST12],[RJ05],[JGP+17]). In this case, a continuous label Y_i is defined by:*

$$Y_i = \frac{\textit{listening time of song } X_i \textit{ until skip}}{\textit{total duration of song } X_i} \in [0,1],$$

which implicitly interprets as a negative feedback when it is close to zero.

Formally, we assume that the random pair (X, Y) admits a density with respect to the Lebesgue measure on $\mathbb{R}^{d+1}$, and that the support of Y is compact, equal to $[0,1]$ for simplicity. The *regression function* is denoted by

$$m : x \in \mathcal{X} \mapsto \mathbb{E}[Y|X = x].$$

We formulate the continuous ranking problem as a continuum of nested bipartite ranking problems. Indeed for any threshold value $y \in (0,1)$, the bipartite ranking subproblem related to the pair (X, Z_y) with $Z_y = 2\mathbb{I}\{Y > y\} - 1$ can be viewed as a discrete approximation of the full problem: we respectively denote by $\mathrm{ROC}_{s,y}$ and $\mathrm{AUC}_{s,y}$ the corresponding ROC curve and AUC of any measurable scoring function $s : \mathcal{X} \to \mathbb{R}$. In other words, we want to solve simultaneously all these subproblems i.e. to identify a scoring rule s maximizing $\mathrm{ROC}_{s,y}$ and $\mathrm{AUC}_{s,y}$ for all $y \in (0,1)$.

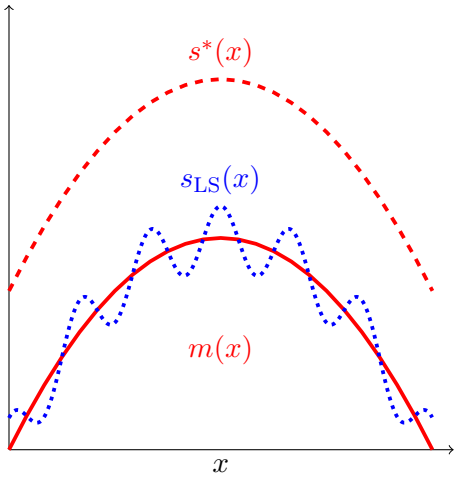


Figure I.3: The least squares regressor s_{LS} (dotted line) is a more accurate proxy than s^* (dashed line) of the regression function m (solid line), in terms of mean squared error. Still, s^* is optimal for the continuous ranking task as it is a strictly increasing transform of m, while s_{LS} is a very poor score function because of its undesirable oscillations.

Contributions. For that purpose, we introduce novel performance measures obtained by integrating $\mathrm{ROC}_{s,y}$ and $\mathrm{AUC}_{s,y}$ with respect to Y's marginal distribution F_Y:

$$\forall \alpha \in (0,1), \quad \mathrm{IROC}_s(\alpha) = \int_{y=0}^{1} \mathrm{ROC}_{s,y}(\alpha) F_Y(dy) \quad \text{and} \quad \mathrm{IAUC}(s) = \int_{\alpha=0}^{1} \mathrm{IROC}_s(\alpha) d\alpha.$$

Our theoretical analysis is twofold. We show that

(i) under some *monotone likelihood ratio* condition, the optimal scoring rules are strictly increasing transforms of the regression function m,

(ii) a scoring rule s^* is optimal if and only if its IROC curve dominates any other IROC curve IROC_s everywhere:

$$\forall \alpha \in (0,1), \quad \mathrm{IROC}_{s^*}(\alpha) = \mathbb{E}[\mathrm{ROC}^*_Y(\alpha)] \geq \mathrm{IROC}_s(\alpha),$$

and its IAUC is maximal:

$$\mathrm{IAUC}(s^*) = \mathbb{E}[\mathrm{AUC}^*_Y] \geq \mathrm{IAUC}(s) \quad \text{for any } s.$$

In addition, we provide a probabilistic expression of the IAUC:

$$\mathrm{IAUC}(s) = \mathbb{P}\{s(X) < s(X') | Y < Y'' < Y'\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X') | Y < Y'' < Y'\},$$

where (X', Y') is an i.i.d. copy of (X, Y) and Y'' is independently sampled from Y's marginal distribution F_Y. Based on this formula, we empirically estimate $\mathrm{IAUC}(s)$ from

an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ as follows:

$$\widehat{\mathrm{IAUC}}_n(s) = \frac{6}{n(n-1)(n-2)} \sum_{1 \leq i,j,k \leq n} \mathbb{I}\left\{s(X_i) < s(X_k), Y_i < Y_j < Y_k\right\} + \frac{3}{n(n-1)(n-2)} \sum_{1 \leq i,j,k \leq n} \mathbb{I}\left\{s(X_i) = s(X_k), Y_i < Y_j < Y_k\right\},$$

which is a U-statistic of degree 3. Finally, we provide a hierarchical algorithm, CRank, aiming at maximizing $\widehat{\mathrm{IAUC}}_n$: it returns a piecewise constant scoring rule obtained by recursively splitting the feature space.

In the next section, we focus on another ranking task, namely *ranking aggregation*, aiming at ordering a finite number of items, given complete or incomplete rankings as training data.

1.5 Ranking Aggregation by Empirical Risk Minimization

The bipartite (resp. continuous) ranking problem presented previously consisted in producing a scoring function, and consequently a ranking over some feature space $\mathcal{X}$, given (vectorial) observations of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$ valued in $\mathcal{X} \times \{-1, +1\}$ (resp. in $\mathcal{X} \times [0, 1]$). In the *ranking aggregation* problem, although the goal is still to *output* a ranking, their exist two principal differences with the earlier settings:

- the set of elements to rank is of finite cardinality N, contrary to the infinite feature spaces $\mathcal{X} \subseteq \mathbb{R}^d$ often considered in bipartite/continuous ranking,
- the *input* data themselves are rankings/comparisons, i.e. *relative* information contrary to labels representing *absolute* evaluations.

Historical Landmarks. Ranking data analysis dates from the 18-th century with the design of an election system for the French 'Académie des Sciences'. Different voting systems were proposed, each satisfying some desirable properties: in particular, the *Borda count* in 1781 ([Bor84]) and its contender, the *Condorcet method* in 1785 ([DC+14]), created the famous *Borda-Condorcet debate*. Later in 1951, Arrow proved an 'impossibility theorem' ([Arr12]) stating that no election rule can satisfy simulatneously some set of reasonable properties ; voting systems are also studied in social choice theory (see [Ris05]). Below, we focus on a specific problem arising in ranking data analysis, namely that of summarizing a set of rankings by a single permutation.

Ranking Aggregation. Given a list of $N \geq 2$ items indexed by $[\![N]\!] = \{1, \ldots, N\}$ and $n \geq 1$ permutations $\sigma_1, \ldots, \sigma_n$ in the *symmetric group* $\mathfrak{S}_N$ of the items, the ranking aggregation problem consists in identifying the single 'consensus' permutation $\widehat{\sigma}_n$ that best summarizes the σ_t's. In many applications using voting systems (e.g. recommendation systems), each ranking σ_t is obtained by asking an agent to order the N items by preference. Hence, the consensus $\widehat{\sigma}_n$ can be seen as the permutation maximizing the simultaneous agreement of the n agents, or equivalently minimizing their disagreement.

Figure I.4: $n = 4$ rankings of $N = 6$ sports teams: for a given ranking, $i \prec j$ means that team i is preferred over team j.

Given a permutation $\sigma \in \mathfrak{S}_N$ and two distinct items $(i, j) \in [\![N]\!]^2$, we will use the notation $i \prec j$ meaning that i is preferred over j i.e. that i is ranked lower than j in the ranking σ: $\sigma(i) < \sigma(j)$. A dataset of $n = 4$ rankings of $N = 6$ items is represented in Figure I.4. While several methods have been developed to solve this problem, we focus here on the Kemeny's consensus approach.

Kemeny's Consensus. The *Kemeny's consensus* method ([Kem]) defines the consensus ranking $\widehat{\sigma}_n$ as a minimizer of the sum of the distances $d(\widehat{\sigma}_n, \sigma_t)$ to the n permuations σ_t:

$$\widehat{\sigma}_n \in \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{t=1}^{n} d(\sigma, \sigma_t),$$

where d is some metric on the set of permutations $\mathfrak{S}_N$. More specifically, Kemeny's rule is based on the choice $d = d_\tau$ with the Kendall's τ distance d_τ defined by: for all $(\sigma, \sigma') \in \mathfrak{S}_N^2$,

$$d_\tau(\sigma, \sigma') = \sum_{1 \le i < j \le N} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

which is the number of pairwise disagreements between σ and σ'. We point out that computing Kemeny's consensus $\widehat{\sigma}_n$ is difficult in practice ([DKNS01]): we refer to [AM12] for discussion on tractable algorithms able to reasonably approximate $\widehat{\sigma}_n$.

Statistical Learning Framework. In [KCS17], a statistical learning formulation of ranking aggregation is introduced: the deterministic permutations σ_t are interpreted as i.i.d. random variables Σ_t with distribution P on $\mathfrak{S}_N$. In this probabilisitc setting, the ultimate goal is to identify a true median ranking σ^* of P characterized by

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_N} L_P(\sigma),$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\sigma, \Sigma)]$. Nevertheless, the distribution P being unknown to the learner having only access to a sample $\Sigma_1, \ldots, \Sigma_n$, the risk L_P and thus the median rank-

ing σ^* cannot be directly computed. Still, by following the ERM paradigm introduced earlier, the empirical consensus

$$\widehat{\sigma}_n \in \arg\min_{\sigma \in \mathfrak{S}_N} L_{\widehat{P}_n}(\sigma) = \frac{1}{n}\sum_{t=1}^{n} d(\sigma, \Sigma_t),$$

where $\widehat{P}_n = (1/n)\sum_{t=1}^{n} \delta_{\Sigma_t}$ denotes the empirical distribution, appears as the natural alternative to σ^*. In particular, [KCS17] established minimax bounds of order $O_{\mathbb{P}}(1/\sqrt{n})$ for the *excess of risk*

$$L_P(\widehat{\sigma}_n) - L_P(\sigma^*).$$

Hence, when the number of observations n grows to infinity, the performance of the empirical solution $\widehat{\sigma}_n$ converges to the minimal risk $L_P(\sigma^*) = \min_{\sigma \in \mathfrak{S}_N} L_P(\sigma)$. In addition, denoting by

$$p_{i,j} = \mathbb{P}_{\Sigma \sim P}\{\Sigma(i) < \Sigma(j)\}$$

the *pairwise probability* that the item $i \in [\![N]\!]$ is preferred over $j \in [\![N]\!] \setminus \{i\}$, the authors showed in the Kendall's τ case $d = d_\tau$ that if the distribution P satisfies the following *strict weak stochastic transitivity assumption*: for all $1 \leq i \neq j \leq N$, $p_{i,j} \neq \frac{1}{2}$ and

$$\forall k \in [\![N]\!] \setminus \{i, j\}, \quad \min(p_{i,j}, p_{j,k}) > \frac{1}{2} \Rightarrow p_{i,k} > \frac{1}{2},$$

then the Kemeny median σ^* is unique and equal to the Copeland ranking σ_{Cop} (see Theorem 5 in [KCS17]):

$$\sigma_{\text{Cop}}(i) = 1 + \sum_{j \neq i} \mathbb{I}\Big\{p_{i,j} < \frac{1}{2}\Big\}, \quad \forall i \in [\![N]\!].$$

Empirically, a similar result also holds: with overwhelming probability, the consensus $\widehat{\sigma}_n$ is equal to the plug-in Copeland ranking $\widehat{\sigma}_{\text{Cop}}$ defined for each item $i \in [\![N]\!]$ by

$$\widehat{\sigma}_{\text{Cop}}(i) = 1 + \sum_{j \neq i} \mathbb{I}\Big\{\widehat{p}_{i,j} < \frac{1}{2}\Big\},$$

with empirical pairwise probabilities

$$\widehat{p}_{i,j} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\{\Sigma_t(i) < \Sigma_t(j)\}, \quad \forall 1 \leq i \neq j \leq N.$$

It follows that this specific instance of the ranking aggregation problem can be solved efficiently based only on *pairwise comparisons* $\mathbb{I}\{\Sigma_t(i) < \Sigma_t(j)\}$, which are a particular case of *incomplete rankings*. In other words, one can avoid observing full rankings Σ_t that may be painful or expensive to obtain in practice, especially when the number of items N is large.

1.6 Dimensionality Reduction on $\mathfrak{S}_N$

In chapter IV, we propose a generalization of ranking aggregation and Kemeny's approach to the more general problem of dimensionality reduction on the symmetric group $\mathfrak{S}_N$. We first recall that the space of distributions on $\mathfrak{S}_N$ is of exploding dimensionality $N! - 1$ and we highlight that ranking aggregation can be seen as an extreme form of dimensionality reduction: indeed, it summarizes a whole distribution P on $\mathfrak{S}_N$ by a single median permutation σ^*. Nevertheless, this approach presents in its very formulation the drawback of hiding the complexity of the distribution P, which can for instance be multimodal and thus cannot be accurately represented by a single permutation. In contrast, we propose to relax Kemeny's consensus method by approximating the original distribution P by a simpler distribution P' in an *optimal transport* fashion (see [Vil08] or [PC$^+$19]). More precisely, our approach consists in choosing proxies P' in a set $\mathbf{P}_\mathcal{C}$ of *bucket distributions* i.e. such that the pairwise probabilities

$$p'_{i,j} = \mathbb{P}_{\Sigma' \sim P'}\{\Sigma'(i) < \Sigma'(j)\}$$

are equal to either zero or one as soon as the two items i and j belong to two distinct buckets $\mathcal{C}_k$ and $\mathcal{C}_l$ of the *bucket order* $\mathcal{C} = (\mathcal{C}_1, \dots, \mathcal{C}_K)$, which is an ordered partition of $[\![N]\!]$. Formally, $\bigcup_{k=1}^K \mathcal{C}_k = [\![N]\!]$, $\mathcal{C}_k \neq \emptyset$ for all $1 \leq k \leq K$, and if $1 \leq k < l \leq K$, $\mathcal{C}_k \cap \mathcal{C}_l = \emptyset$ and

$$(i,j) \in \mathcal{C}_k \times \mathcal{C}_l \Rightarrow p'_{j,i} = 1 - p'_{i,j} = 0.$$

Intuitively, a distribution $P' \in \mathbf{P}_\mathcal{C}$ is constrained in such a way that it cannot generate permutations that are hesistant about the relative rank of atoms in different buckets. For example with the four rankings $\sigma_1, \dots, \sigma_4$ from Figure I.4 and the three buckets orders $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ from Figure I.5,

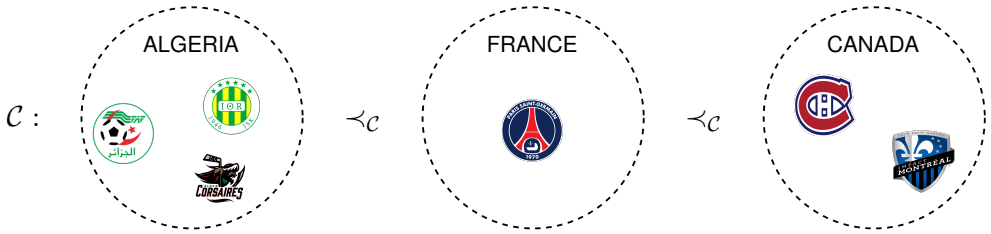
- the two rankings σ_1 and σ_2 both satisfy the structure of the bucket order $\mathcal{C}$: $\{\delta_{\sigma_1}, \delta_{\sigma_2}\} \subset \mathbf{P}_\mathcal{C}$,
- the ranking σ_3 satisfies $\mathcal{C}'$: $\delta_{\sigma_3} \in \mathbf{P}_{\mathcal{C}'}$,
- the last ranking σ_4 does not satisfy the constraints of any of the three bucket orders: $\delta_{\sigma_4} \notin \mathbf{P}_\mathcal{C} \cup \mathbf{P}_{\mathcal{C}'} \cup \mathbf{P}_{\mathcal{C}''}$.

Given a bucket order $\mathcal{C}$, we denote by $\lambda = (\#\mathcal{C}_1, \dots, \#\mathcal{C}_K)$ its *shape*: it describes the number of items contained in each of the K buckets and determines the dimensionality of $\mathbf{P}_\mathcal{C}$, namely $d_\mathcal{C} = \Pi_{1 \leq k \leq K} \#\mathcal{C}_k! - 1$.

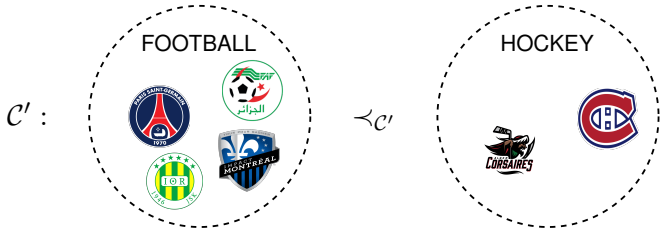
Then, the optimal proxy $P^*_\mathcal{C} \in \mathbf{P}_\mathcal{C}$ of some general distribution P is chosen by minimizing the *Wasserstein metric* $W_{d_\tau,1} = \inf_{\Sigma \sim P, \Sigma' \sim P'} \mathbb{E}[d_\tau(\Sigma, \Sigma')]$:

$$P^*_\mathcal{C} \in \operatorname*{arg\,min}_{P' \in \mathbf{P}_\mathcal{C}} W_{d_\tau,1}(P, P').$$

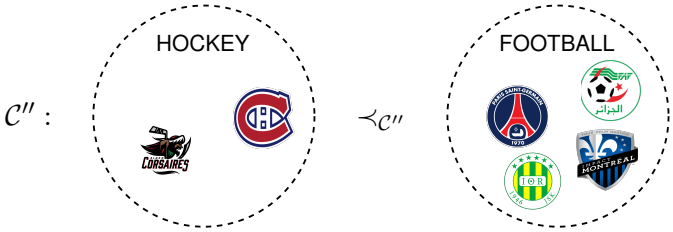
The unconstrained case $K = 1$, or equivalently $\lambda = N$ and $d_\mathcal{C} = N! - 1$, corresponds to no dimensionality reduction with $\mathbf{P}_\mathcal{C}$ equal to the whole space of distributions on $\mathfrak{S}_N$



(a) Bucket order $\mathcal{C}$ corresponding to a geographical hierarchy: Algerian teams are preferred over the French team, itself preferred over Canadian teams. $\mathcal{C}$ has size $K = 3$, shape $\lambda = (3, 1, 2)$ and dimension $d_{\mathcal{C}} = 3! \cdot 1! \cdot 2! - 1 = 11$.



(b) This bucket order constrains football teams to be preferred over hockey teams. $\mathcal{C}'$ has size $K = 2$, shape $\lambda = (4, 2)$ and dimension $d_{\mathcal{C}'} = 4! \cdot 2! - 1 = 47$.



(c) $\mathcal{C}''$ is based on the same partition as $\mathcal{C}'$ but their buckets are ordered differently. $\mathcal{C}''$ has size $K = 2$, shape $\lambda = (2, 4)$ and dimension $d_{\mathcal{C}''} = 2! \cdot 4! - 1 = 47$.

Figure I.5: Three bucket orders of the $N = 6$ sports teams.

and thus $P^*_{\mathcal{C}} = P$. In the opposite extreme case $K = N$, equivalent to $\lambda = (1, \ldots, 1)$ and $d_{\mathcal{C}} = 0$, the set $\mathbf{P}_{\mathcal{C}} = \{\delta_{\sigma_{\mathcal{C}}}\}$ is reduced to a singleton: as in ranking aggregation, the distribution P is simply approximated by a single ranking, here the unique permutation $\sigma_{\mathcal{C}}$ such that $i \in \mathcal{C}_{\sigma_{\mathcal{C}}(i)}$ for all $i \in [\![N]\!]$.

Contributions. Our analysis of this problem relies on the following results.

(i) We show that the *distortion* $\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_\tau,1}(P, P')$ of any bucket order $\mathcal{C}$ simply writes in terms of pairwise probabilities:

$$\Lambda_P(\mathcal{C}) = \sum_{1 \leq k < l \leq K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} p_{j,i},$$

which can be empirically estimated from pairwise comparisons, similarly to the risk in Kemeny's consensus method.

(ii) We derive and analyse the ERM version of the bucket order optimization problem:

$$\min_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C}),$$

where $\mathbf{C}_{K,\lambda}$ denotes the set of all bucket orders with same number of buckets K and shape λ.

We point out that for shape $\lambda = (1,\dots,1)$, problem (ii) coincides with Kemeny's consensus method: indeed, we have $\Lambda_P(\mathcal{C}) = L_P(\sigma_\mathcal{C})$ and $\{\sigma_\mathcal{C} : \mathcal{C} \in \mathbf{C}_{N,\lambda}\} = \mathfrak{S}_N$ in this case. Hence, our dimensionality reduction approach naturally extends ranking aggregation. We also provide a hierarchical algorithm, called BuMeRank, recursively merging adjacent buckets into coarser bucket orders.

2 Risk-Aware Reinforcement Learning

This section presents our contributions in two (nested) *online learning* frameworks: multi-armed bandits and reinforcement learning, the former being a particular case of the latter. From now on, as opposed to the offline ERM settings exposed in the previous section, the training datasets are not initially given to the learner. In contrast, the learner/decision-maker has to interact with an *environment* to simultaneously collect observations and design its own strategy.

2.1 The Stochastic Multi-Armed Bandit Problem

The *multi-armed bandit* (MAB) problem (see e.g. [BCB+12]) is a sequential decision making problem encountered by a gambler in a casino facing $K \geq 1$ slot machines: at each iteration $t \in \{1,\dots,T\}$ (with $T \geq 1$ the time horizon), he chooses a slot machine (a.k.a. 'one-armed bandit', or simply 'arm') $A_t \in \{1,\dots,K\}$ to play and receives a random reward $X_{A_t,t}$. Initially, the gambler/learner/decision-maker has no prior knowledge about the machines and his objective is to maximize his expected total reward across all iterations, namely $\mathbb{E}\left[\sum_{t=1}^T X_{A_t,t}\right]$.

In the *stochastic setting*, we assume that the rewards $X_{a,1},\dots,X_{a,T}$ generated by each machine $a \in \{1,\dots,K\}$ are i.i.d. sampled from some probability distribution ν_a on $\mathbb{R}$ with expectation μ_a. Then, the quantity to maximize becomes:

$$\mathbb{E}\left[\sum_{t=1}^T X_{A_t,t}\right] = \sum_{a=1}^K \mu_a \mathbb{E}[N_a(T)], \tag{I.8}$$

where $N_a(t') = \sum_{t=1}^{t'} \mathbb{I}\{A_t = a\}$ denotes the number of times the arm a has been pulled up to any time $t' \geq 1$. A MAB *model* $\mathcal{D}$ is a set of possible distributions ν_a with finite expectation: each K-tuple $(\nu_1,\dots,\nu_K) \in \mathcal{D}^K$ characterizes an instance of MAB

problem. The *optimal strategy in hindsight* thus consists in always pulling the optimal arm a^*, assumed to be unique:

$$a^* = \arg\max_{1\leq a\leq K} \mu_a,$$

where $\mu^* = \mu_{a^*} = \max_{1\leq a\leq K} \mu_a > \max_{a\neq a^*} \mu_a$. Formally, a bandit strategy is a mapping $h_t \mapsto (\mathbb{P}\{A_{t+1} = 1|h_t\}, \dots, \mathbb{P}\{A_{t+1} = K|h_t\})$, where the *history* of the arms pulled and of the rewards obtained up to the current iteration t is denoted by

$$h_t = (A_1, X_{A_1,1}, \dots, A_t, X_{A_t,t}).$$

Finding a strategy maximizing Eq. (I.8) can be equivalently reformulated through the minimization of the *expected regret*

$$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)],$$

with gaps $\Delta_a = \mu^* - \mu_a$. This regret interprets as the overall deficit of expected rewards generated by the bandit strategy compared to the best strategy in hindsight receiving reward $X_{a^*,t}$ at each step $1 \leq t \leq T$.

Exploration versus Exploitation. The MAB problem was originally introduced by Thompson in 1933 ([Tho33]), motivated by clinical trials comparing the effectiveness of several treatments by testing them over a sequence of patients. In this medical context, each reward corresponds to the observed effect of a treatment on a patient: hence, suboptimal treatments must be quickly identified and then discarded to save as many patients as possible. On the one hand, the range of all possible treatments has to be *explored* sufficiently in order to spot, with high confidence, the optimal treatment among them ; and on the other hand, the best treatment should be *exploited* as frequently as possible i.e. provided to the maximum number of patients, by avoiding unnecessary exploration. This example highlights the *exploration-exploitation trade-off* arising in bandit problems, including in modern applications such as ad placement (see [BCB+12]).

Asymptotic Distribution-Dependent Lower Bound. [LR85], [BK96], [CK15] and [GMS19] proved that, asymptotically, the regret of any *uniformly efficient* strategy is lower bounded by a logarithmic function of the time horizon T multiplied by a distribution-dependent constant involving Kullback-Leibler divergences.

Definition 1. *A MAB strategy is uniformly efficient for a model $\mathcal{D}$ if for all MAB problems $(\nu_a)_{1\leq a\leq K} \in \mathcal{D}^K$ and for all suboptimal arms $a \neq a^*$, it satisfies*

$$\mathbb{E}[N_a(T)] = o(T^\alpha), \quad \forall \alpha \in (0, 1].$$

Theorem 1 (Theorem 1 in [GMS19])**.** *For any model $\mathcal{D}$, uniformly efficient MAB strategy on $\mathcal{D}$, MAB problem $(\nu_1, \dots, \nu_K) \in \mathcal{D}^K$ and suboptimal arm a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})},$$

where

$$\mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}) = \inf\{KL(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D} \text{ and } \mathbb{E}_{X'\sim\nu'_a}[X'] > x\},$$

with KL the Kullback-Leibler divergence between two probability distributions.

In particular, if the model $\mathcal{D}$ is a *one-dimensional exponential family* (e.g. Bernoulli or Poisson distributions), the Kullback-Leibler divergence $KL(\nu, \nu')$ between two distributions ν, ν' in $\mathcal{D}$ is simply a function of their respective means $\mu = \mathbb{E}_{X\sim\nu}[X]$ and $\mu' = \mathbb{E}_{X'\sim\nu'}[X']$:

$$KL(\nu, \nu') = d(\mu, \mu').$$

Intuitively, any 'reasonable' MAB algorithm should at least produce a logarithmic regret: we next recall celebrated approaches guaranteeing regret with finite-time upper bounds asymptotically matching the lower bound in Theorem 1.

Asymptotically Optimal Algorithms. Several algorithms were proven to be *asymptotically optimal*, particularly in the case of distributions $\nu_1, \dots, \nu_K$ belonging to the same exponential family distribution, such as KL-UCB ([GC11]), Bayes-UCB ([Kau16]) and Thompson Sampling ([Tho33], [KKM12], [KKM13]). These strategies are all *index policies* i.e. they rely on some index $u_a(t)$ computed at each round $t \geq 1$ for each arm $a \in \{1, \dots, K\}$: a generic index policy is described in Algorithm 1.

Algorithm 1 MAB index policy

Require: time horizon T.
1: **Initialize:** Pull each arm once: $A_t = t, \quad \forall t \in \{1, \dots, K\}$.
2: **for** $t = K$ **to** $T - 1$ **do**
3: Compute index $u_a(t)$ for all arms $a \in \{1, \dots, K\}$.
4: Pull arm $A_{t+1} = \arg\max_{1\leq a\leq K} u_a(t)$.
5: **end for**

• **The KL-UCB Algorithm.** Based on the same *optimism in the face of uncertainty* principle used in the UCB1 algorithm ([ACBF02]) through the computation of confidence intervals for the empirical estimators of expectations μ_a's, the KL-UCB algorithm was introduced in [GC11]. It is an index policy characterized by the following index:

$$u_a(t) = \sup\Big\{q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t\Big\}, \tag{I.9}$$

where $\hat{\mu}_a(t) = (1/N_a(t)) \sum_{s=1}^{t} \mathbb{I}\{A_s = a\} X_{a,s}$ is the empirical average reward at time t, and c is a positive constant typically smaller than 3. This strategy is said to be 'optimistic' as it pulls the arm with highest upper confidence bound (UCB) $u_a(t)$ i.e. it (optimistically) considers that the true expected value μ_a is as large as its UCB.

• **The Bayes-UCB Algorithm.** The Bayes-UCB algorithm ([KCG12]) is a Bayesian index policy. It relies on the same intuition as KL-UCB but replaces the UCB by an upper quantile:

$$u_a(t) = Q(1 - 1/(t(\log t)^c), \pi_{a,t}), \tag{I.10}$$

where $Q(\alpha, \pi_{a,t})$ denotes the quantile of order α of the posterior distribution $\pi_{a,t}$ for arm a at time t.

• **The Thompson Sampling Algorithm.** The Thompson Sampling strategy (originally proposed in [Tho33], and analysed in [KKM12], [KKM13]) is a Bayesian approach consisting in sampling a natural parameter (of a one-dimensional exponential distribution) $\theta_a(t) \sim \pi_a(t)$ from the posterior distribution $\pi_a(t)$ updated with the $N_a(t)$ observations collected from arm a up to time t. Then, the index is given by:

$$u_a(t) = \mu(\theta_a(t)), \tag{I.11}$$

with $\mu(\theta)$ the expected value of the one-dimensional exponential distribution $\nu \in \mathcal{D}$ with natural parameter θ.

We present in the next subsection our study of a variant of MAB tailored to credit-risk management applications.

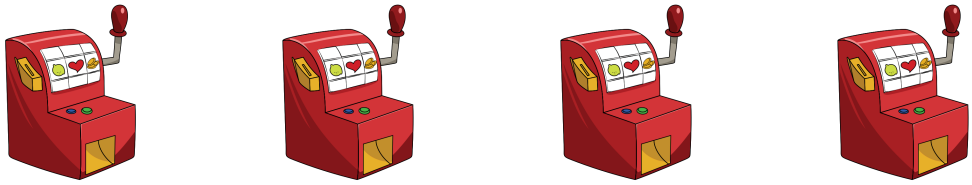
2.2 Bandits for Default Risk Management

In the default risk management problem, a loaner (typically a bank) is receiving credit requests — that he may either accept or reject — from individuals belonging to different populations. Each of the $K \geq 1$ populations is a category/arm, denoted by $a \in \{1, \dots, K\}$, predefined by the bank based on a features such as age, gender, salary or ethnicity for instance, combined with the average loan amount τ_a. Assuming that the bank has enough budget, it should maximize its total profit by loaning money to all profitable populations (if there is any), not only the most profitable arm. Hence, from a bandit point of view, the notion of single optimal arm is not relevant anymore. More formally, we consider in chapter V a variation of the MAB problem, that we call *profitable bandits*, where, at each iteration $t \in \{1, \dots, T\}$, the learner may pull a subset $\mathcal{A}_t \subseteq \{1, \dots, K\}$ of the arms, or possibly no arm at all (i.e. $\mathcal{A}_t = \emptyset$). To each population/arm $a \in \{1, \dots, K\}$ is associated an unknown distribution ν_a and a known threshold τ_a. The threshold τ_a corresponds to the average amount of money borrowed to the bank by each individual from population a. In addition, we assume that at each time step t, a (bounded) random number $n_a(t)$ of people from category a are asking for a credit. Then, the goal is to maximize the expected cumulative profit which sums, for each borrower $c \in \{1, \dots, n_a(t)\}$ from all chosen categories $a \in \mathcal{A}_t$, the difference between the average reimbursement $\mu_a = \mathbb{E}_{X \sim \nu_a}[X]$ and the average loan amount τ_a:

$$S_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{I}\{a \in \mathcal{A}_t\} \sum_{c=1}^{n_a(t)} X_{a,c,t} - L_{a,c,t}\right],$$

where the random variables $X_{a,c,t}$ are i.i.d. sampled from ν_a and the random loan amounts $L_{a,c,t}$ have expectation equal to τ_a. Here also, we reformulate the objective by means of the following expected regret:

$$R_T = \sum_{a \in \mathcal{A}^*} \Delta_a \widetilde{N}_a(T) - S_T = \sum_{a \in \mathcal{A}^*} \Delta_a \left(\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]\right) + \sum_{a \notin \mathcal{A}^*} |\Delta_a| \mathbb{E}[N_a(T)],$$



$\mu_1 = 2 > \tau_1 = 1$ $\mu_2 = 2 < \tau_2 = 3$ $\mu_3 = 5 > \tau_3 = 3$ $\mu_4 = 5 < \tau_4 = 6$

Figure I.6: $K = 4$ slot machines. In order to play with the a-th machine ($1 \leq a \leq K$), a gambler must insert a coin of value τ_a: he receives a random reward with unknown mean μ_a. The set $\mathcal{A}^* = \{1, 3\}$ contains the profitable arms a for which $\mu_a > \tau_a$.

where $\widetilde{N}_a(T) = \mathbb{E}\left[\sum_{t=1}^{T} n_a(t)\right]$ is the expected total number of clients from category a over the T rounds, $N_a(t) = \sum_{s=1}^{t} n_a(s)\mathbb{I}\{a \in \mathcal{A}_s\}$ is the number of observations obtained from category a up to time $t \geq 1$, $\Delta_a = \mu_a - \tau_a$ is the (unknown) expected profit provided by a client of category a and $\mathcal{A}^* = \big\{a \in \{1, \dots, K\}, \Delta_a > 0\big\}$ is the set of profitable arms.

Index Policies. Motivated by the success of the MAB index policies recalled earlier, we adapt them to profitable bandits: at each iteration t and for each category a, an index $u_a(t)$ is computed, and the arm a is pulled if $u_a(t)$ is larger than the known threshold τ_a (see Algorithm 2). Then, we propose three index policies for solving our profitable bandits problem:

- the kl-UCB-4P algorithm ('4P' means 'for profit') with same index as kl-UCB, namely $u_a(t)$ given by Equation (I.9),
- the Bayes-UCB-4P algorithm with same index as Bayes-UCB (see Eq. (I.10)),
- the TS-4P algorithm with same index as Thompson Sampling (see Eq. (I.11)).

Our analysis will show that these three strategies are all *asymptotically optimal* for the profitable bandits problem.

Algorithm 2 Profitable bandits index policy

Require: time horizon T, thresholds $(\tau_a)_{a \in \{1,\dots,K\}}$.
1: **Initialize:** Pull all arms: $\mathcal{A}_1 = \{1, \dots, K\}$.
2: **for** $t = 1$ **to** $T - 1$ **do**
3: Compute index $u_a(t)$ for all arms $a \in \{1, \dots, K\}$.
4: Pull arms in $\mathcal{A}_{t+1} = \{a \in \{1, \dots, K\} : u_a(t) \geq \tau_a\}$.
5: **end for**

Contributions. We extend the MAB analysis to our profitable bandits framework through two main results: lower and upper regret bounds respectively.

(i) First, we show that any *uniformly efficient* profitable bandits strategy produces a regret R_T asymptotically lower bounded as follows:

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq \sum_{a\notin\mathcal{A}^*} \frac{|\Delta_a|}{\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a)},$$

where

$$\mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}_a) = \inf\{\mathrm{KL}(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D}_a \text{ and } \mathbb{E}_{X'\sim\nu'_a}[X'] > x\}.$$

(ii) If $n_a(t) = n_a$ almost surely for all $1 \leq t \leq T$, with constant $n_a \geq 1$, then the three algorithms that we propose, namely KL-UCB-4P, BAYES-UCB-4P and TS-4P are all asymptotically optimal i.e. their regret matches the lower bound for T growing to infinity. Otherwise, a multiplicative gap exists between our lower and upper bounds. Indeed, we provide upper regret bounds with the following asymptotic order:

$$\limsup_{T\to\infty} \frac{R_T}{\log T} \leq \sum_{a\notin\mathcal{A}^*} \left(\frac{n_a^+}{n_a^-}\right) \frac{|\Delta_a|}{\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a)},$$

for constants $n_a^+, n_a^- \geq 1$ such that $n_a^- \leq n_a(t) \leq n_a^+$ almost surely for all t.

2.3 Bandits and Extreme Values

In various *risk-aware* contexts, the quantities of interest are not necessarily expectations. In environmental or financial applications for instance, a decision maker may be *risk-averse* by taking decisions to ensure sufficient protection against disastrous events such as flooding or financial crisis (see [BGST06], [Res07]). In other words, efficient strategies are designed by giving more importance to worst-case scenarios than to 'normal' observations. Mathematically, such pessimistic scenarios are often modeled as rare and extreme events. Hence, in many risk-aware problems, the learner has to optimize a criterion based on the tail of some distribution, which characterizes its extreme values.

Alternative Risk Measures. We first recall that the expectation of a real-valued random variable X with cumulative distribution function ('c.d.f.' in short) $F : x \mapsto \mathbb{P}\{X \leq x\}$ is equal to the integral of the *quantile function* (or *generalized inverse distribution function*) $F^{-1} : \tau \mapsto \inf\{x : F(x) \geq \tau\}$ over the interval $[0, 1]$ (see [Dev08]):

$$\mathbb{E}[X] = \mathbb{E}_{U\sim\mathcal{U}([0,1])}[F^{-1}(U)] = \int_{\tau=0}^{1} F^{-1}(\tau)d\tau,$$

where $\mathcal{U}([0, 1])$ is the uniform distribution on $[0, 1]$. Several risk measures have been proposed in the literature to replace the expectation:

- the *quantile* $F^{-1}(\tau)$ at some level $\tau \in (0, 1]$, also called 'value-at-risk' (VaR) (see e.g. [ADEH99]), is also a solution of the following asymmetric L_1 minimization problem:

$$F^{-1}(\tau) \in \arg\min_{\theta} \mathbb{E}[\ell_\tau^q(X - \theta)], \tag{I.12}$$

with quantile regression loss (a.k.a. 'pinball loss') $\ell_\tau^q(x) = x(\tau - \mathbb{I}\{x < 0\})$,

- the *conditional value-at-risk (CVaR)* $\mathrm{CVaR}_\alpha(X)$, also referred as 'expected shortfall' or 'superquantile' ([RU$^+$00]):

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha}\int_{\tau=1-\alpha}^{1} F^{-1}(\tau)d\tau, \tag{I.13}$$

  describes better the (right) tail of the distribution than the expectation, as the quantile function F^{-1} is only integrated over an upper part of the whole interval $(0,1)$,

- the *expectile* $e_\tau(X)$, sharing common properties with quantiles, solves an asymmetric L_2 minimization problem ([NP87]):

$$e_\tau(X) = \arg\min_\theta \mathbb{E}[\ell^e_\tau(X-\theta)], \tag{I.14}$$

  with expectile loss $\ell^e_\tau(x) = x^2|\tau - \mathbb{I}\{x<0\}|$.

We point out that at level $\tau = 1/2$, the quantile $F^{-1}(1/2)$ is a median of the distribution of X and the expectile its expectation: $e_{1/2}(X) = \mathbb{E}[X]$. Moreover if $\alpha = 1$, the CVaR coincides with the expectation: $\mathrm{CVaR}_1(X) = \mathbb{E}[X]$. Empirically, the CVaR_α of some distribution ν can be estimated from an i.i.d. sample $X_t \sim \nu$ for $1 \le t \le T$ by:

$$\widehat{\mathrm{CVaR}}_\alpha = \frac{1}{\lceil \alpha T \rceil}\sum_{t=\lfloor (1-\alpha)T \rfloor}^{T} X_{\sigma(t)}, \tag{I.15}$$

with permutation $\sigma \in \mathfrak{S}_T$ and the order statistics $X_{\sigma(1)} \le \cdots \le X_{\sigma(T)}$. We point out that $\widehat{\mathrm{CVaR}}_\alpha$ is an L-statistic (see [VdV00]), i.e. a linear combination of the order statistics.

Risk-Aware Bandits. Many variants of the classical MAB problem have been proposed for risk-aware applications. Basically, they consist in replacing the expectation by different risk measures. While [SBFWH15] focuses on quantiles, [GST13] and [KJ$^+$19] both propose strategies relying on the estimation of the CVaR. General bandits frameworks encompassing broad classes of risk criteria (including quantiles and CVaR) are studied in [TGP19] and [CMZ18]. In [SLM12], [VZ16] and [ZIJC14], the quality of an arm is assessed through a combination of its mean and its variance: for two arms with the same mean, the one with the smallest variance is deemed safer than the other one. In [Mai13], the risk-aversion is measured by means of the cumulant generating functions of the distributions of the arms; the mean-variance approach then appears as a particular case of this method in the Gaussian scenario (i.e. when ν_a is a normal distribution for each arm a). We introduce next the *max K-armed bandit problem*, also called 'extreme bandits', as an extreme form of the risk-aware problems discussed above. Indeed, while the CVaR_α (with $\alpha << 1$) of some distribution ν allows to study its right tail — empirically, by selecting the α-fraction of the largest order statistics $X_{\sigma(\lfloor (1-\alpha)T \rfloor)}, \ldots, X_{\sigma(T)}$ among an i.i.d. sample $(X_t)_{1\le t\le T}$ (see Eq. (I.15)) —, the max K-armed bandit problem defined below only focuses on the maximal observation of this sample, namely $\max_{1\le t\le T} X_t$.

Extreme Bandits. In some applications in medicine, insurance or finance, the quantity of interest is not the expected return, but rather the *extreme* observations ([BGST06]). From a multi-armed bandit point of view, the 'best' arm should not be defined as the one with highest expectation, but as that producing the maximal values. This setting, referred to as *extreme bandits* in [CV14], was originally introduced by [CS05] by the name of *max K-armed bandit* problem. In this framework, the goal pursued is to obtain the highest possible reward during the $T \geq 1$ steps. For a given arm $a \in \{1, \dots, K\}$, we denote by

$$G_T^{(a)} = \max_{1 \leq t \leq T} X_{a,t}$$

the maximal value up to round $T \geq 1$ and assume that, in expectation, there is a unique optimal arm

$$a^* = \arg\max_{1 \leq a \leq K} \mathbb{E}\left[G_T^{(a)}\right].$$

Then, *expected extreme regret* of a strategy, pulling the arm $A_t \in \{1, \dots, K\}$ at time t, is defined as

$$R_T = \mathbb{E}\left[G_T^{(a^*)}\right] - \mathbb{E}\left[\max_{1 \leq t \leq T} X_{A_t,t}\right], \tag{I.16}$$

where $\max_{1 \leq t \leq T} X_{A_t,t}$ is the maximal value observed by the learner up to the time horizon T. When the supports of the K reward distributions $\nu_1, \dots, \nu_K$ are bounded, no-regret is expected provided that every arm can be sufficiently explored, as shown in [NLB16] and [DS16]. If infinitely many arms are possibly involved in the learning strategy, the challenge is then to explore and exploit optimally the unknown reservoir of arms, see [CV15]. When, on the contrary, the rewards are unbounded, the situation is quite different: the best arm is that for which the maximum $G_T^{(a)}$ tends to infinity faster than the others. In [NLB16], it is shown that, for unbounded distributions, no policy can achieve no-regret without restrictive assumptions on the distributions. In accordance with the literature, we focus on a classical framework in extreme value analysis. Namely, we assume that the reward distributions are *heavy-tailed*.

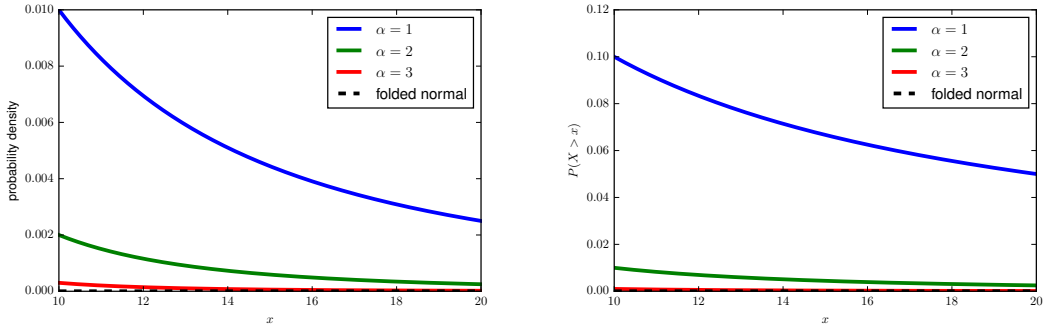
Heavy-tailed distributions are widely used to model extremes in many applications, where a conservative approach to risk assessment might be relevant (e.g. finance, environmental risks). Like in [CV14], we consider that the rewards are distributed as second order Pareto laws, which are similar to classical Pareto distributions. Formally, a probability law with c.d.f. $F(x)$ belongs to the (α, β, C, C')-second order Pareto family if, for every $x \geq 0$,

$$|1 - Cx^{-\alpha} - F(x)| \leq C' x^{-\alpha(1+\beta)}, \tag{I.17}$$

where α, β, C and C' are strictly positive constants, see *e.g.* [Res07]. Naturally, the *Pareto distribution* with *tail index* α and scale parameter C, whose c.d.f. is:

$$\forall x \geq C^{\frac{1}{\alpha}}, \quad F(x) = 1 - Cx^{-\alpha},$$

belongs to this family as it trivially verifies Eq. (I.17). These distributions have indeed 'heavy tails', see Figure I.7 for a comparison with a 'light tail' folded normal distribution. For each arm $a \in \{1, \dots, K\}$, the distribution ν_a is assumed to belong to the



(a) Probability density functions.

(b) Survival functions.

Figure I.7: Pareto laws for different tail indices α (and same scale parameter $C = 1$) compared to the folded normal distribution of $1 + |X|$ with X a standard normal variable (all distributions have same support).

$(\alpha_a, \beta_a, C_a, C')$-second order Pareto family with $\alpha_a > 1$, so that the expectation of the random variable $X_{a,t} \sim \nu_a$ is finite. In this context, [CV14] have proposed the ExtremeHunter algorithm to solve the *extreme bandit* problem and provided a regret analysis with the following upper bound

$$R_T = O\left(T^{\frac{1}{(1+b)\alpha_{a^*}}}\right),$$

where $b > 0$ is a known lower bound on the (unknown) β-coefficients: $b \leq \min_a \beta_a$.
Contributions. Our contribution to this problem is developed in chapter VI: it is twofold.

(i) First, we significantly improve the regret analysis of ExtremeHunter by a polynomial factor in the time horizon T, by proving that

$$R_T = O\left((\log T)^{2(2b+1)/b} T^{-(1-1/\alpha_{a^*})} + T^{-(b-1/\alpha_{a^*})}\right),$$

and we provide a matching lower bound in a specific case. This essentially relies on a finer bound for the difference between the expectation of the maximum among independent realizations $X_1, \ldots, X_T$ of a (α, β, C, C')-second order Pareto distribution, $\mathbb{E}[\max_{1 \leq i \leq T} X_i]$ namely, and its rough approximation $(TC)^{1/\alpha}\Gamma(1 - 1/\alpha)$ with Γ the Gamma function. As a by-product, we propose a more simple Explore-Then-Commit strategy that offers the same theoretical guarantees as ExtremeHunter.

(ii) Second, we explain how extreme bandit can be reduced to a classical bandit problem to a certain extent. We show that a MAB strategy such as Robust-UCB (see [BCL13]), applied on correctly left-censored rewards $X_{a,t}\mathbb{I}\{X_{a,t} > u\}$ with threshold

u large enough, may also reach a very good performance. This claim is supported by theoretical guarantees on the number of pulls of the best arm a^* and by numerical experiments both at the same time.

Next, we consider the general reinforcement learning setting, which includes the (*static*) multi-armed bandit problem that we studied. We point out that, halfway between these two problems, more *dynamic* MAB frameworks have also been considered in the literature: in particular, *contextual bandits* (see [Woo79], [Sli14], [PR$^+$13]), where the rewards depend on observable random covariates.

2.4 Reinforcement Learning

The multi-armed bandit problem discussed above can be seen as a very specific case of the more general *reinforcement learning* (RL) framework. In reinforcement learning, an agent seeks to maximize the expected sum of (discounted) future rewards by sequentially interacting with his environment. This total return defines policy-dependent value functions of the environment's state and of the agent's action. The objective is then to find an optimal policy maximizing these value functions in each state. If the environment is always in the same state, then RL is a bandit problem where the arms are the different actions. We introduce formally the RL setup below.

Mixtures. Here and in chapter VII, we denote by $\mathcal{P}(\mathcal{E})$ the set of probability distributions on a set $\mathcal{E}$ (either countable or $\mathbb{R}$). In addition, given a random variable Y valued in a countable set $\mathcal{Y}$ and a mapping $\nu : \mathcal{Y} \to \mathcal{P}(\mathcal{E})$, we denote by $\nu(Y) \in \mathcal{P}(\mathcal{E})$ the *mixture distribution* of the following random variable:

$$\sum_{y \in \mathcal{Y}} \mathbb{I}\{Y = y\} U_y,$$

where $U_y \sim \nu(y)$ and Y are independent for any $y \in \mathcal{Y}$.

Markov Decision Process. A *Markov decision process (MDP)* is described by a tuple $(\mathcal{X}, \mathcal{A}, P, R)$ with

- state space $\mathcal{X}$,
- action space $\mathcal{A}$,
- transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$,
- distributional reward function $R : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$.

For simplicity, we will assume that $\mathcal{X}$ and $\mathcal{A}$ are both countable. If the environment is in state $x \in \mathcal{X}$ and if the agent takes the action $a \in \mathcal{A}$, then he receives a reward $R_0 \sim R(x, a)$ and the next state X_1 is sampled from the distribution $P(\cdot|x, a) \in \mathcal{P}(\mathcal{X})$ such that R_0, X_1 are independent. See Figure I.9 for an example of a simple MDP with two states, two actions, and *deterministic rewards* (i.e. there exists a function $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that $R(x, a) = \delta_{r(x,a)}$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$).

A policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ maps any state $x \in \mathcal{X}$ to a distribution over the actions $\pi(\cdot|x) \in \mathcal{P}(\mathcal{A})$. Given a discount factor $\gamma \in [0, 1)$, we define the *distributional return* $Z^\pi(x, a)$ of a policy π after taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$ as the *probability distribution* of the random variable

$$\sum_{t=0}^{\infty} \gamma^t R_t \quad \text{given that } X_0 = x, A_0 = a,$$
$$\text{and for all } t \in \mathbb{N}, R_t \sim R(X_t, A_t), X_{t+1} \sim P(\cdot|X_t, A_t), A_{t+1} \sim \pi(\cdot|X_{t+1}). \quad \text{(I.18)}$$

The discount rate γ serves as both a mathematical device to ensure the convergence of the total return, and as a parameter determining the present value of future rewards: a small value of γ gives little importance to future rewards. An alternative to Eq. (I.18), that we will not consider here, is the sum of rewards $\sum_{t=0}^{T} R_t$, which only makes sense when there is a natural notion of final time step T (see [SB18]). Usually, RL focuses on expected returns through the *state-action value function*

$$Q^\pi(x, a) = \mathbb{E}_{Z_0 \sim Z^\pi(x,a)}[Z_0],$$

and the *value function*

$$V^\pi(x) = \mathbb{E}_{A_0 \sim \pi(\cdot|x)}[Q^\pi(x, A_0)],$$

verifying *Bellman's equation* ([Bel66]):

$$\forall (x, a), \quad Q^\pi(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q^\pi(X_1, A_1)],$$

where $R_0 \sim R(x, a)$, $X_1 \sim P(\cdot|x, a)$ and $A_1 \sim \pi(\cdot|X_1)$. The *optimal policies* can be characterized by means of the *optimal state-action value function* $Q^*(x, a)$, which verify *Bellman's optimality equation*:

$$\forall (x, a), \quad Q^*(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q^*(X_1, a')].$$

Then, denoting by $V^*(x) = \max_a Q^*(x, a)$ the *optimal value function*, a policy π^* is *optimal* if and only if for all x,

$$\mathbb{E}[Q^*(x, A_0)] = V^*(x), \quad \text{with } A_0 \sim \pi^*(\cdot|x).$$

Bellman Operators. In the *policy evaluation* task, one wants to compute Q^π for a given policy π, while in the *control* task, the goal it to approach Q^*. The usual dynamic programming way for solving these two tasks is based on two operators. First, the *Bellman operator* T^π ([Bel66]) defined by: for all $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$T^\pi Q(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q(X_1, A_1)], \quad \text{with } X_1 \sim P(\cdot|x, a), A_1 \sim \pi(\cdot|X_1).$$

Second, the *Bellman optimality operator* T defined by:

$$TQ(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q(X_1, a')], \quad \text{with } X_1 \sim P(\cdot|x, a).$$

In particular, the Bellman operator T^π (resp. Bellman optimality operator T) is known to be a γ-contraction[1] for the sup norm and its repeated application to an initial Q-function to converge exponentially fast to its unique fixed point Q^π (resp. Q^*) ([BT96]).

RL Algorithms. In RL, the transition kernel P is unknown and thus the Bellman operators cannot be computed exactly. Hence, practical RL methods such as *temporal-difference (TD) learning* ([Sut88]), SARSA ([RN94]), or Q-LEARNING ([Wat89]) consist in computing stochastic approximations of these operators based on trajectories composed of observed 'state-action-reward' sequences. Given a single *transition*

$$(X_t, A_t, R_t, X_{t+1}, A_{t+1}),$$

with $R_t \sim R(X_t, A_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $A_{t+1} \sim \pi(\cdot|X_{t+1})$, and learning rate $0 < \alpha \leq 1$, their update rules are the following:

- the TD(0) update:

$$V(X_t) \leftarrow (1-\alpha)V(X_t) + \alpha(R_t + \gamma V(X_{t+1})),$$

- the SARSA(0) update (using the next action A_{t+1}):

$$Q(X_t, A_t) \leftarrow (1-\alpha)Q(X_t, A_t) + \alpha(R_t + \gamma Q(X_{t+1}, A_{t+1})),$$

- the Q-LEARNING update:

$$Q(X_t, A_t) \leftarrow (1-\alpha)Q(X_t, A_t) + \alpha(R_t + \gamma \max_{a' \in \mathcal{A}} Q(X_{t+1}, a')).$$

Under technical conditions (tabular setting, states and actions are visited infinitely many times, either constant or decaying learning rate, etc.), TD methods were proved to converge to the value function V^π ([Sut88], [Day92]), while the SARSA(0) algorithm (combined with greedy policies, see [SJLS00]) as well as Q-LEARNING (see [WD92]) both converge to the optimal state-action value function Q^*.

As in the bandit case, many safe RL formulations have been proposed by replacing the expected returns by some risk-sensitive criterion, we refer to [GF15] for a survey of such methods. Our approach to this problem relies on the more challenging topic of distributional reinforcement learning, where the focus is not only on the value functions (i.e. expected returns) as in RL but on the whole distributions of the returns, which potentially allows risk-aware applications based on risk measures such as the CVaR for instance.

[1] A function mapping a metric space to itself is called a γ-contraction (resp. a non-expansion) if it is Lipschitz continuous with Lipschitz constant $\gamma < 1$ (resp. $\kappa \leq 1$).

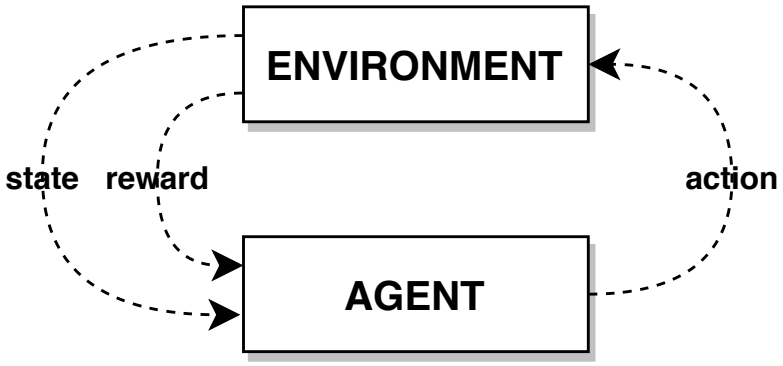


Figure I.8: The dynamics of reinforcement learning: the agent observes the current state of the environment, then takes an action, receives a reward, observes the new state, and so forth...

2.5 Beyond Value Functions: Atomic Bellman Equations

In *distributional reinforcement learning (DRL)*, the focus is on the distribution, denoted by $Z^\pi(x,a)$, of the random variable $\sum_{t\geq 0}\gamma^t R_t$ in Eq. (I.18), i.e. not only its expectation as it is the case in (non-distributional) RL. As shown in [BDM17], the usual RL tools such as Bellman's equations (for expected returns) can be generalized to distributions. Similarly, the authors proposed two distributional Bellman operators: while the first, denoted by $\mathcal{T}^\pi$, for distributional policy evaluation of a given policy π, is a contraction, the second, for the control task, is not (see respectively Lemma 3 and Proposition 1 in [BDM17]). Formally, the *distributional Bellman operator* $\mathcal{T}^\pi$ is defined by: for any *state-action distribution function*

$$Z : (x,a) \in \mathcal{X}\times\mathcal{A} \mapsto Z(x,a) \in \mathcal{P}(\mathbb{R}),$$

the image of Z by $\mathcal{T}^\pi$ is the state-action distribution function $\mathcal{T}^\pi Z$ given by:

$$\mathcal{T}^\pi Z : (x,a) \mapsto \text{ distribution of the r.v. } R_0 + \gamma Z_1, \text{ with } R_0 \sim R(x,a), Z_1 \sim Z(X_1,A_1),$$

where $X_1 \sim P(\cdot|x,a)$ and $A_1 \sim \pi(\cdot|X_1)$. The *distributional Bellman equation* then writes:

$$Z^\pi = \mathcal{T}^\pi Z^\pi.$$

In practical implementations, dealing with general distributions may be difficult from a computational point of view. Hence, existing DRL approaches have been developed by projecting the distributions into a simple parametric space of probability measures, thus leading to tractable computation. For instance, [DRBM18] and [RBD+18] both approximate distributional returns with atomic distributions but consider different metrics for evaluating the approximation errors: respectively the 1-Wasserstein distance[2] W_1 and the Cramér distance.

[2]For $p \in [1,+\infty)$, the p-Wasserstein distance between two distributions D_1 and D_2 on $\mathbb{R}$ (with c.d.f.'s F_1 and F_2) is $W_p(D_1,D_2) = \left(\int_{\tau=0}^1 \left|F_1^{-1}(\tau) - F_2^{-1}(\tau)\right|^p d\tau\right)^{\frac{1}{p}}$.

Atomic Projection. Our approach in chapter VII relies on the following *two design biases*.

(a) We approximate probability distributions D on $\mathbb{R}$ by atomic distributions $D_{\omega,\theta} = \sum_{i=1}^{N} \omega_i \delta_{\theta_i}$ with $\omega_i \geq 0$ and $\omega_1 + \cdots + \omega_N = 1$, and $\theta_1 \leq \cdots \leq \theta_N$.

(b) As in our dimensionality reduction problem on the symmetric group (part 1, chapter IV), we use a mass transportation metric to measure the approximation errors, namely the 2-Wasserstein distance W_2: the smaller

$$W_2(D, D_{\omega,\theta}) = \left(\sum_{i=1}^{N} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} \left(F^{-1}(\tau) - \theta_i \right)^2 d\tau \right)^{\frac{1}{2}}, \tag{I.19}$$

with cumulative probabilities $\overline{\omega}_i = \sum_{j \leq i} \omega_j$, the better $D_{\omega,\theta}$ approximates D.

Importantly, for fixed probability weights ω_i's, the approximation error in Eq. (I.19) is minimized with respect to the atoms θ_i's if and only if for all $1 \leq i \leq N$ such that $\omega_i \neq 0$, θ_i is equal to the following *trimmed mean* of the distribution D:

$$\theta^*_{\omega,i} = \frac{1}{\omega_i} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} F^{-1}(\tau) d\tau.$$

We point out that in the monoatomic case $N = 1$, the unique 'trimmed mean' is simply the expectation. Indeed, it also writes as the integral of the quantile function over the whole interval $(0, 1)$: $\mathbb{E}_{Y \sim D}[Y] = \int_{\tau=0}^{1} F^{-1}(\tau) d\tau$, which boils down to classical RL. In addition, these trimmed means may be used in a risk-aware context to compute risk measures such as the CVaR:

$$\mathrm{CVaR}_{1-\overline{\omega}_{i-1}}(Y) = \frac{\theta^*_{\omega,i} + \cdots + \theta^*_{\omega,N}}{N - i + 1}.$$

We provide below a use case with two policies having the same expected performance but different risk levels.

Use Case - Safe versus Risky Policies. For the MDP described in Figure I.9, combined with a discount factor $\gamma = \frac{1}{2}$, the two policies π, π' given by $\pi(a_1|\cdot) \equiv 1$ ('always choose action a_1') and $\pi'(a_2|\cdot) \equiv 1$ ('always choose action a_2') share the same value functions:

$$V^{\pi}(x_1) = Q^{\pi}(x_1, a_1) = \frac{1}{2} = V^{\pi'}(x_1) = Q^{\pi'}(x_1, a_2)$$
$$\text{and} \quad V^{\pi}(x_2) = Q^{\pi}(x_2, a_1) = \frac{3}{2} = V^{\pi'}(x_2) = Q^{\pi'}(x_2, a_2). \tag{I.20}$$

However, π' yields deterministic discounted returns, contrary to π, and is thus the safest of the two policies. More precisely, the distributional returns with expected values given

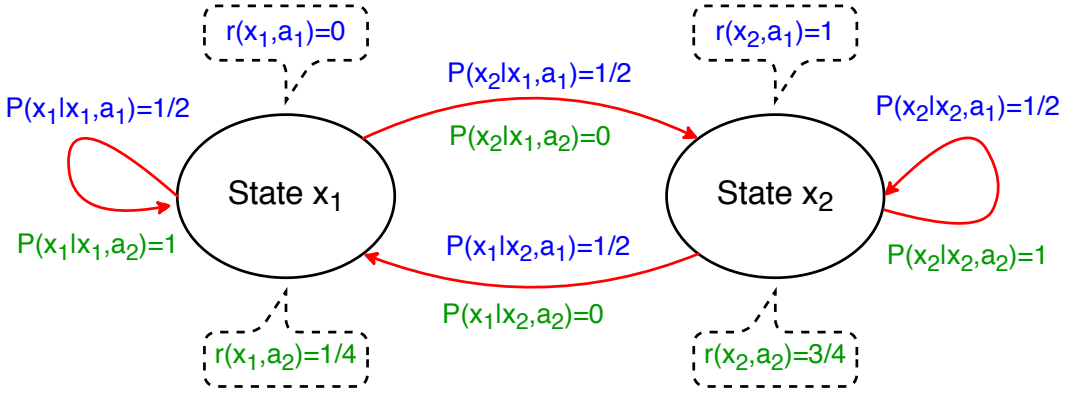


Figure I.9: Example of a Markov decision process with 2 states ($\mathcal{X} = \{x_1, x_2\}$), 2 actions ($\mathcal{A} = \{a_1, a_2\}$) and deterministic rewards ($R(x,a) = \delta_{r(x,a)}$).

in Eq. (I.20) are concentrated in Dirac masses in the case of the 'safe' policy π', while they are uniformly spread over some intervals for the 'risky' one π:

$$Z^\pi(x_1,a_1) = \mathcal{U}([0,1]) \neq \delta_{\frac{1}{2}} = Z^{\pi'}(x_1,a_2),$$
$$\text{and} \quad Z^\pi(x_2,a_1) = \mathcal{U}([1,2]) \neq \delta_{\frac{3}{2}} = Z^{\pi'}(x_2,a_2), \quad \text{(I.21)}$$

where $\mathcal{U}([\alpha,\beta])$ denotes the uniform distribution on any interval $[\alpha,\beta]$.
Contributions. Our contribution is threefold.

(i) First, we introduce two new '1-step' DRL operators, only dealing with the randomness induced by the first step. The first, for policy evaluation, is denoted by $\mathbb{T}^\pi$ and given by: for any state-action distribution function Z and $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\mathbb{T}^\pi Z(x,a)$ is the distribution of the r.v.

$$R_0 + \gamma \mathbb{E}[Z_1|X_1,A_1], \text{ with } R_0 \sim R(x,a), Z_1 \sim Z(X_1,A_1), X_1 \sim P(\cdot|x,a), A_1 \sim \pi(\cdot|X_1),$$

while our second DRL operator $\mathbb{T}$ (for the control task) is defined such that $\mathbb{T}Z(x,a)$ is the distribution of

$$R_0 + \gamma \max_{a'} \mathbb{E}[Z_{1,a'}|X_1], \text{ with } R_0 \sim R(x,a), X_1 \sim P(\cdot|x,a), Z_{1,a'} \sim Z(X_1,a') \forall a' \in \mathcal{A}.$$

Interestingly, $\mathbb{T}^\pi$ and $\mathbb{T}$ are both contraction mappings.

(ii) Then, we describe the *projected operators* resulting from choices (a) and (b) and prove that they are also contractions. In addition, we derive the *atomic Bellman equations*, that are the fixed-point equations of the projected operators: they generalize the usual (non-distributional) Bellman equations to the multiatomic case of several trimmed means.

(iii) Finally, we propose new DRL algorithms as multiatomic extensions of the TD learning and Q-LEARNING methods.

In a nutshell, the final chapter VII provides new theoretical DRL tools, namely the 1-step DRL operators and the atomic Bellman equations, that shall be used in risk-aware situations.

Conclusion - Perspectives

Many perspectives and lines of future research can be drawn from this thesis.

- Extending our WERM approach to solve the bipartite ranking task with positive-unlabeled data remains an open problem. We describe in section 6 (of chapter II) an incremental version of WERM for that purpose.
- The dimensionality reduction framework derived in chapter IV relies on the choice of the Kendall's τ distance to quantify the ranking approximation errors. In the section 7 of the same chapter, we propose an extension of our approach to another metric, namely the Spearman ρ distance. We provide an alternative formula for the distortion: interestingly, it shows that the triplet-wise probabilities $p_{i,j,k} = \mathbb{P}_{\Sigma\sim P}\{\Sigma(i) < \Sigma(j) < \Sigma(k)\}$ are playing a key role in the Spearman ρ case, similarly to the pairwise probabilities $p_{i,j} = \mathbb{P}_{\Sigma\sim P}\{\Sigma(i) < \Sigma(j)\}$ in the Kendall's τ case.
- The DRL approach proposed in the last chapter relies on an optimal spatial distribution of the atoms θ_i's, given predefined probability weights ω_i's. In the section 10 (of chapter VII), we introduce a notion of optimality for these probabilities: this paves the way for more sophisticated algorithms optimizing both the atoms and the probabilities.

Part 1

Ranking by Empirical Risk Minimization

CHAPTER II

Weighted Empirical Risk Minimization: Sample Selection Bias Correction based on Importance Sampling

Abstract

We consider statistical learning problems, when the distribution P' of the training observations $Z'_1, \ldots, Z'_n$ differs from the distribution P involved in the risk one seeks to minimize (referred to as the *test distribution*) but is still defined on the same measurable space as P and dominates it. In the unrealistic case where the likelihood ratio $\Phi(z) = dP/dP'(z)$ is known, one may straightforwardly extends the Empirical Risk Minimization (ERM) approach to this specific *transfer learning* setup using the same idea as that behind Importance Sampling, by minimizing a weighted version of the empirical risk functional computed from the 'biased' training data Z'_i with weights $\Phi(Z'_i)$. Although the *importance function* $\Phi(z)$ is generally unknown in practice, we show that, in various situations frequently encountered in practice, it takes a simple form and can be directly estimated from the Z'_i's and some auxiliary information on the statistical population P. By means of linearization techniques, we then prove that the generalization capacity of the approach aforementioned is preserved when plugging the resulting estimates of the $\Phi(Z'_i)$'s into the weighted empirical risk. Beyond these theoretical guarantees, numerical results provide strong empirical evidence of the relevance of the approach promoted in this chapter.

1 Introduction

Prediction problems are of major importance in statistical learning. The main paradigm of predictive learning is *Empirical Risk Minimization* (ERM in abbreviated form), see *e.g.* [DGL96]. In the standard setup, Z is a random variable (r.v. in short) that takes its values in a feature space $\mathcal{Z}$ with distribution P, Θ is a parameter space and $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is a (measurable) loss function. The risk is then defined by: $\forall \theta \in \Theta$,

$$\mathcal{R}_P(\theta) = \mathbb{E}_P\left[\ell(\theta, Z)\right], \tag{II.1}$$

and more generally for any measure Q on $\mathcal{Z}$: $\mathcal{R}_Q(\theta) = \int_{\mathcal{Z}} \ell(\theta, z) dQ(z)$. In most practical situations, the distribution P involved in the definition of the risk is unknown and learning is based on the sole observation of an independent and identically distributed (i.i.d.) sample $Z_1, \ldots, Z_n$ drawn from P and the risk (II.1) must be replaced by an empirical counterpart (or a possibly smoothed/penalized version of it), typically:

$$\widehat{\mathcal{R}}_P(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta), \tag{II.2}$$

where $\widehat{P}_n = (1/n) \sum_{i=1}^n \delta_{Z_i}$ is the empirical measure of P and δ_z denotes the Dirac measure at any point z. With the design of successful algorithms such as neural networks, support vector machines or boosting methods to perform ERM, the practice of predictive learning has recently received a significant attention and is now supported by a sound theory based on results in empirical process theory. The performance of minimizers of (II.2) can be indeed studied by means of concentration inequalities, quantifying the fluctuations of the maximal deviations $\sup_{\theta \in \Theta} |\widehat{\mathcal{R}}_P(\theta) - \mathcal{R}_P(\theta)|$ under various complexity assumptions for the functional class $\mathcal{F} = \{\ell(\theta, \cdot) : \theta \in \Theta\}$ (*e.g.* VC dimension, metric entropies, Rademacher averages), see [BLM13] for instance. Although, in the Big Data era, the availability of massive digitized information to train predictive rules is an undeniable opportunity for the widespread deployment of machine-learning solutions, the poor control of the data acquisition process one is confronted with in many applications puts practicioners at risk of jeopardizing the generalization ability of the rules produced by the algorithms implemented. Bias selection issues in machine-learning are now the subject of much attention in the literature, see [BCZ+16], [ZWY+17], [BHS+19], [LYLW16] or [HGB+07]. In the context of face analysis, a research area including a broad range of applications such as face detection, face recognition or face attribute detection, machine learning algorithms trained with baised training data, *e.g.* in terms of gender or ethnicity, raise concerns about fairness in machine learning. Unfair algorithms may induce systemic undesired disadvantages for specific social groups, see [DDB18] for further details. Several examples of bias in deep learning based face recognition systems are discussed in [NSS+19].

Throughout the present chapter, we consider the case where the i.i.d. sample $Z'_1, \ldots, Z'_n$ available for training is not drawn from P but from another distribution P', with respect to which P is absolutely continuous, and the goal pursued is to set theoretical grounds for the application of ideas behind Importance Sampling (IS in short) methodology to extend the ERM approach to this learning setup. We highlight that the problem under study is a very particular case of *Transfer Learning* (see *e.g.* [PY10], [BDBC+10] and [Sto09]), a research area currently receiving much attention in the literature and encompassing general situations where the information/knowledge one would like to transfer may take a form in the *target* space very different from that in the *source* space (referred to as *domain adaptation*).

Weighted ERM (WERM). In this chapter, we investigate conditions guaranteeing that values for the parameter θ that nearly minimize (II.1) can be obtained through

minimization of a weighted version of the empirical risk based on the Z_i''s, namely

$$\widetilde{\mathcal{R}}_{w,n}(\theta) = \mathcal{R}_{\widetilde{P}_{w,n}}(\theta), \tag{II.3}$$

where $\widetilde{P}_{w,n} = (1/n)\sum_{i=1}^n w_i \delta_{Z_i'}$ and $w = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ is a certain weight vector. Of course, ideal weights w^* are given by the likelihood function $\Phi(z) = (dP/dP')(z)$: $w_i^* = \Phi(Z_i')$ for $i \in \{1, \ldots, n\}$. In this case, the quantity (II.3) is obviously an unbiased estimate of the true risk (II.1):

$$\mathbb{E}_{P'}\left[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)\right] = \mathcal{R}_P(\theta), \tag{II.4}$$

and generalization bounds for the $\mathcal{R}_P$-risk excess of minimizers of $\widetilde{\mathcal{R}}_{w^*,n}$ can be directly established by studying the concentration properties of the empirical process related to the Z_i''s and the class of functions $\{\Phi(\cdot)\ell(\theta, \cdot) : \theta \in \Theta\}$ (see section 2 below). However, the *importance function* Φ is unknown in general, just like distribution P. It is the major purpose of this chapter to show that, in far from uncommon situations, the (ideal) weights w_i^* can be estimated from the Z_i's combined with auxiliary information on the target population P. As shall be seen below, such favorable cases include in particular classification problems where class probabilities in the test stage differ from those in the training step, risk minimization in stratified populations (see [BD18]), with strata statistically represented in a different manner in the test and training populations, positive-unlabeled learning (PU-learning, see *e.g.* [dPNS14]). In each of these cases, we show that the stochastic process obtained by plugging the weight estimates in the weighted empirical risk functional (II.3) is much more complex than a simple empirical process (*i.e.* a collection of i.i.d. averages) but can be however studied by means of *linearization techniques*, in the spirit of the ERM extensions established in [CLV08] or [CV09a]. Learning rate bounds for minimizers of the corresponding risk estimate are proved and, beyond these theoretical guarantees, the performance of the weighted ERM approach is supported by convincing numerical results.

The chapter is structured as follows. In section 2, the ideal case where the importance function Φ is known is preliminarily considered and a first basic example where the optimal weights can be easily inferred and plugged into the risk without deteriorating the learning rate is discussed. The main results of the chapter are stated in section 3, which shows that the methodology promoted can be applied to two important problems in practice, risk minimization in stratified populations and PU-learning, with generalization guarantees. Illustrative numerical experiments are displayed in section 4, while some concluding remarks are collected in section 5. Proofs are deferred to section 7.

2 Importance Sampling - Risk Minimization with Biased Data

Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the sup norm of any bounded function $h : \mathcal{Z} \to \mathbb{R}$ by $||h||_\infty$. We place ourselves in the framework

of statistical learning based on biased training data previously introduced. As a first go, we consider the unrealistic situation where the importance function Φ is known, insofar as we shall subsequently develop techniques aiming at mimicking the minimization of the ideally weighted empirical risk

$$\widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n}\sum_{i=1}^{n} w_i^* \ell(\theta, Z_i'), \tag{II.5}$$

namely the (unbiased) Importance Sampling estimator of (II.1) based on the instrumental data $Z_1', \ldots, Z_n'$. The following result describes the performance of minimizers $\widetilde{\theta}_n^*$ of (II.5). Since the goal of this chapter is to promote the main ideas of the approach rather than to state results with the highest level of generality, we assume throughout the chapter for simplicity that ℓ and Φ are both bounded functions. For $\sigma_1, \ldots, \sigma_n$ independent Rademacher random variables (*i.e.* symmetric $\{-1,1\}$-valued r.v.'s), independent from the Z_i''s, we define the Rademacher average associated to the class of function $\mathcal{F}$ as $R_n'(\mathcal{F}) := \mathbb{E}_\sigma\left[\sup_{\theta\in\Theta} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \ell(\theta, Z_i')\right|\right]$. This quantity can be bounded by metric entropy methods under appropriate complexity assumptions on the class $\mathcal{F}$, it is for instance of order $O_{\mathbb{P}}(1/\sqrt{n})$ when $\mathcal{F}$ is a VC major class with finite VC dimension, see *e.g.* [BBL05].

Lemma 1. *With probability at least* $1-\delta$, *we have:* $\forall n \geq 1$,

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 4||\Phi||_\infty \mathbb{E}\left[R_n'(\mathcal{F})\right] + 2||\Phi||_\infty \sup_{(\theta,z)\in\Theta\times\mathcal{Z}} \ell(\theta,z)\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Of course, when $P' = P$, we have $\Phi \equiv 1$ and the bound stated above simply describes the performance of standard empirical risk minimizers. The proof is based on the standard bound

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 2\sup_{\theta\in\Theta}\left|\widetilde{\mathcal{R}}_{w^*,n}(\theta) - \mathbb{E}\left[\widetilde{\mathcal{R}}_{w^*,n}(\theta)\right]\right|,$$

combined with basic concentration results for empirical processes, see section 7 for further details. Of course, the importance function Φ is generally unknown and must be estimated in practice. As illustrated by the elementary example below (related to binary classification, in the situation where the probability of occurence of a positive instance significantly differs in the training and test stages), in certain statistical learning problems with biased training distribution, Φ takes a simplistic form and can be easily estimated from the Z_i''s combined with auxiliary information on P.

Binary Classification with Varying Class Probabilities. The flagship problem in supervised learning corresponds to the simplest situation, where $Z = (X, Y)$, Y being a binary variable valued in $\{-1, +1\}$ say, and the r.v. X takes its values in a measurable space $\mathcal{X}$ and models some information hopefully useful to predict Y. The parameter space Θ is a set $\mathcal{G}$ of measurable mappings (*i.e.* classifiers) $g : \mathcal{X} \to \{-1, +1\}$ and the loss function is given by $\ell(g, (x,y)) = \mathbb{I}\{g(x) \neq y\}$ for all g in $\mathcal{G}$ and any $(x,y) \in$

$\mathcal{X} \times \{-1,\ +1\}$. The distribution P of the random pair (X, Y) can be either described by X's marginal distribution $\mu(dx)$ and the posterior probability $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ or else by the triplet (p, F_+, F_-) where $p = \mathbb{P}\{Y = +1\}$ and $F_\sigma(dx)$ is X's conditional distribution given $Y = \sigma 1$ with $\sigma \in \{-,\ +\}$. It is very common that the fraction of positive instances in the training dataset is significantly lower than the rate p expected in the test stage, supposed to be known here (see Remark 2 for the case where the rate p is only approximately known). We thus consider the case where the distribution P' of the training data $(X'_1, Y'_1),\ \ldots,\ (X'_n, Y'_n)$ is described by the triplet (p', F_+, F_-) with $p' < p$. The likelihood function takes the simple following form

$$\Phi(x, y) = \mathbb{I}\{y = +1\}\frac{p}{p'} + \mathbb{I}\{y = -1\}\frac{1-p}{1-p'} \stackrel{def}{=} \phi(y),$$

which reveals that it depends on the label y solely, and the ideally weighted empirical risk process is

$$\widetilde{\mathcal{R}}_{w^*,n}(g) = \frac{p}{p'}\frac{1}{n}\sum_{i:Y'_i=1} \mathbb{I}\{g(X'_i) = -1\} + \frac{1-p}{1-p'}\frac{1}{n}\sum_{i:Y'_i=-1} \mathbb{I}\{g(X'_i) = +1\}. \tag{II.6}$$

In general the theoretical rate p' is unknown and one replaces (II.6) with

$$\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) = \frac{p}{n'_+}\sum_{i:Y'_i=1} \mathbb{I}\{g(X'_i) = -1\} + \frac{1-p}{n'_-}\sum_{i:Y'_i=-1} \mathbb{I}\{g(X'_i) = +1\}, \tag{II.7}$$

where $n'_+ = \sum_{i=1}^n \mathbb{I}\{Y'_i = +1\} = n - n'_-$, $\widehat{w}^*_i = \widehat{\phi}(Y'_i)$ and $\widehat{\phi}(y) = \mathbb{I}\{y = +1\}np/n'_+ + \mathbb{I}\{y = -1\}n(1-p)/n'_-$. The stochastic process above is not a standard empirical process but a collection of sums of two ratios of basic averages. However, the following result provides a uniform control of the deviations between the ideally weighted empirical risk and that obtained by plugging the empirical weights into the latter.

Lemma 2. *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $p' \in (\varepsilon,\ 1-\varepsilon)$. For any $\delta \in (0, 1)$, we have with probability larger than $1 - \delta$:*

$$\sup_{g\in\mathcal{G}} \left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g)\right| \leq \frac{2}{\varepsilon^2}\sqrt{\frac{\log(2/\delta)}{2n}},$$

as soon as $n \geq 2\log(2/\delta)/\varepsilon^2$.

See section 7 for the technical proof. Consequently, minimizing (II.7) nearly boils down to minimizing (II.6). Combining Lemmas 2 and 1, we immediately get the generalization bound stated in the result below.

Corollary 1. *Suppose that the hypotheses of Lemma 2 are fulfilled. Let $\widetilde{g}_n$ be any minimizer of $\widetilde{\mathcal{R}}_{\widehat{w}^*,n}$ over class $\mathcal{G}$. We have with probability at least $1 - \delta$:*

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g\in\mathcal{G}} \mathcal{R}_P(g) \leq \frac{2\max(p, 1-p)}{\varepsilon}\left(2\mathbb{E}[R'_n(\mathcal{G})] + \sqrt{\frac{2\log(2/\delta)}{n}}\right) + \frac{4}{\varepsilon^2}\sqrt{\frac{\log(4/\delta)}{2n}},$$

as soon as $n \geq 2\log(4/\delta)/\varepsilon^2$*; where* $R'_n(\mathcal{G}) = (1/n)\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}} |\sum_{i=1}^n \sigma_i \mathbb{I}\{g(X'_i) \neq Y'_i\}|]$.

Hence, some side information (*i.e.* knowledge of parameter p) has permitted to weight the training data in order to build an empirical risk functional that approximates the target risk and to show that minimization of this risk estimate yields prediction rules with optimal (in the minimax sense) learning rates. The purpose of the subsequent analysis is to show that this remains true for more general problems. Observe in addition that the bound in Corollary 1 deteriorates as ε decays to zero: the method used here is not intended to solve the *few shot* learning problem, where almost no training data with positive labels is available (*i.e.* $p' \approx 0$). As shall be seen in subsection 3.2, alternative estimators of the importance function must be considered in this situation.

Remark 1. *Although the quantity* (II.7) *can be viewed as a cost-sensitive version of the empirical classification risk based on the* (X'_i, Y'_i)*'s (see e.g. [BHH06]), we point out that the goal pursued here is not to achieve an appropriate trade-off between type I and type II errors in the* P' *classification problem as in biometric applications for instance (i.e. optimization of the* (F_+, F_-)*-ROC curve at a specific point) but to transfer knowledge gained in analyzing the biased data drawn from* P' *to the classification problem related to distribution* P*.*

Remark 2. (Inaccurate Prior Information about the Test Distribution) *As noticed above, it may happen that the rate of positive instances in the target population is approximately known only. Suppose that our guess for* p *is* $\widetilde{p}$ *such that* $|p - \widetilde{p}| \leq \zeta$*, with* $\zeta \in (0, 1)$*. Denote by* $\widetilde{P}$ *the distribution over* $\mathcal{X} \times \{-1, +1\}$ *under which* X *is drawn from* $\widetilde{p}F_+ + (1-\widetilde{p})F_-$ *and such that* $\mathbb{P}_{(X,Y)\sim\widetilde{P}}\{Y = 1 \mid X = x\} = \mathbb{P}_{(X,Y)\sim P}\{Y = 1 \mid X = x\} = \eta(x)$*. By a change of measure we have,*

$$\mathbb{P}_{\widetilde{P}}(Y \neq g(X)) = \mathbb{P}_P(Y \neq g(X)) + \mathbb{E}_P\left[\left(\frac{d\widetilde{P}}{dP}(X,Y) - 1\right)\mathbb{I}\{Y \neq g(X)\}\right],$$

which allows to bound the difference of the classification risks of g *under* P *and* $\widetilde{P}$*:*

$$\left|\mathcal{R}_{\widetilde{P}}(g) - \mathcal{R}_P(g)\right| \leq \mathbb{E}_P\left[\left|\frac{d\widetilde{P}}{dP}(X,Y) - 1\right|\right] = 2|\widetilde{p} - p| \leq 2\zeta.$$

Related Work. We point out that the natural idea of using weights in ERM problems that mimic those induced by the importance function has already been used in [SNK+08] for *covariate shift adaptation* problems (*i.e.* supervised situations, where the conditional distribution of the output given the input information is the same in the training and test domains), when, in contrast to the framework considered here, a test sample is additionally available (a method for estimating directly the importance function based on Kullback-Leibler divergence minimization is proposed, avoiding estimation of the test density). Importance sampling estimators have been also considered in [GV14] in the setup of *inductive transfer learning* (the tasks between source and target are different, regardless of the similarities between source and target domains), where the authors have

proposed two methods to approximate the importance function, among which one is again based on minimizing the Kullback-Leibler divergence between the two distributions. In [CMRR08], the sample selection bias is assumed to be independent from the label, which is not true under our stratum-shift assumption or for the PU learning problem (see section 3). Lemma 1 assumes that the exact importance function is known, as does [CMM10]. The next section introduces new results for more realistic settings where it has to be learned from the data.

3 Weighted Empirical Risk Minimization - Generalization Guarantees

Through two important and generic examples, relevant for many applications, we show that the approach sketched above can be applied to general situations, where appropriate auxiliary information on the target distribution is available, with generalization guarantees.

3.1 Statistical Learning from Biased Data in a Stratified Population

A natural extension of the simplistic problem considered in section 2 is multiclass classification in a stratified population. The random labels Y and Y' are supposed to take their values in $\{1, \ldots, J\}$ say, with $J \geq 1$, and each labeled observation (X, Y) belongs to a certain random stratum S in $\{1, \ldots, K\}$ with $K \geq 1$. Again, the distribution P of a random element $Z = (X, Y, S)$ may be described by the parameters $\{(p_{j,k}, F_{j,k}) : 1 \leq j \leq J, 1 \leq k \leq K\}$ where $F_{j,k}$ is the conditional distribution of X given $(Y, S) = (j, k)$ and $p_{j,k} = \mathbb{P}_{(X,Y,S)\sim P}\{Y = j, S = k\}$. Then, we have

$$dP(x, y, s) = \sum_{j=1}^{J} \sum_{k=1}^{K} \mathbb{I}\{y = j, s = k\} p_{j,k} dF_{j,k}(x),$$

and considering a distribution P' with $F_{j,k} \equiv F'_{j,k}$ but possibly different class-stratum probabilities $p'_{j,k}$, the likelihood function becomes

$$\frac{dP}{dP'}(x, y, s) = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{p_{j,k}}{p'_{j,k}} \mathbb{I}\{y = j, s = k\} \stackrel{def}{=} \phi(y, s).$$

A more general framework can actually encompass this specific setup by defining 'meta-strata' in $\{1, \ldots, J\} \times \{1, \ldots, K\}$. Strata may often correspond to categorical input features in practice. The formalism introduced below is more general and includes the example considered in the preceding section, where strata are defined by labels.

Learning from Biased Stratified Data. Consider a general mixture model, where distributions P and P' are stratified over $K \geq 1$ strata. Namely, $Z = (X, S)$ and $Z' = (X', S')$ with auxiliary random variables S and S' (the strata) valued in $\{1, \ldots, K\}$. We place ourselves in a *stratum-shift* context, assuming that the conditional distribution

of X given $S = k$ is the same as that of X' given $S' = k$, denoted by $F_k(dx)$, for any $k \in \{1, \ldots, K\}$. However, stratum probabilities $p_k = \mathbb{P}(S = k)$ and $p'_k = \mathbb{P}(S' = k)$ may possibly be different. In this setup, the likelihood function depends only on the strata and can be expressed in a very simple form, as follows:

$$\frac{dP}{dP'}(x, s) = \sum_{k=1}^{K} \mathbb{I}\{s = k\} \frac{p_k}{p'_k} \stackrel{def}{=} \phi(s).$$

In this case, the ideally weighted empirical risk writes

$$\widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, Z'_i) \sum_{k=1}^{K} \mathbb{I}\{S'_i = k\} \frac{p_k}{p'_k}.$$

If the strata probabilities p_k's for the test distribution are known, an empirical counterpart of the ideal empirical risk above is obtained by simply plugging estimates of the p'_k's computed from the training data:

$$\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(\theta) = \sum_{i=1}^{n} \ell(\theta, Z'_i) \sum_{k=1}^{K} \mathbb{I}\{S'_i = k\} \frac{p_k}{n'_k}, \tag{II.8}$$

with $n'_k = \sum_{i=1}^{n} \mathbb{I}\{S'_i = k\}$, $\widehat{w}^*_i = \widehat{\phi}(S'_i)$ and $\widehat{\phi}(s) = \sum_{k=1}^{K} \mathbb{I}\{s = k\} n p_k / n'_k$.

A bound for the excess of risk is given in Theorem 1, that can be viewed as a generalization of Corollary 1.

Theorem 1. *Let $\varepsilon \in (0, 1/2)$ and assume that $p'_k \in (\varepsilon, 1 - \varepsilon)$ for $k = 1, \ldots, K$. Let $\widetilde{\theta}^*_n$ be any minimizer of $\widetilde{\mathcal{R}}_{\widehat{w}^*,n}$ as defined in* (II.8) *over class Θ. We have with probability at least $1 - \delta$:*

$$\mathcal{R}_P(\widetilde{\theta}^*_n) - \inf_{\theta \in \Theta} \mathcal{R}_P(\theta) \leq \frac{2 \max_k p_k}{\varepsilon} \left(2\mathbb{E}[R'_n(\mathcal{F})] + L\sqrt{\frac{2\log(2/\delta)}{n}} \right) + \frac{4L}{\varepsilon^2} \sqrt{\frac{\log(4K/\delta)}{2n}},$$

as soon as $n \geq 2\log(4K/\delta)/\epsilon^2$; where $R'_n(\mathcal{F}) = (1/n)\mathbb{E}_\sigma[\sup_{\theta \in \Theta} | \sum_{i=1}^{n} \sigma_i \ell(\theta, Z'_i)|]$, and the loss is bounded by $L = \sup_{(\theta,z) \in \Theta \times \mathcal{Z}} \ell(\theta, z)$.

Just like in Corollary 1, the bound in Theorem 1 explodes when ε vanishes, which corresponds to the situation where a stratum $k \in \{1, \ldots, K\}$ is very poorly represented in the training data, *i.e.* when $p'_k << p_k$. Again, as highlighted by the experiments carried out, reweighting the losses in a frequentist (ERM) approach guarantees good generalization properties in a specific setup only, where the training information, though biased, is sufficiently informative.

3.2 Positive-Unlabeled Learning

Relaxing the *stratum-shift* assumption made in the previous subsection, the importance function becomes more complex and writes:

$$\Phi(x,s) = \frac{dP}{dP'}(x,s) = \sum_{k=1}^{K} \mathbb{I}\{s=k\}\frac{p_k}{p'_k}\frac{dF_k}{dF'_k}(x),$$

where F_k and F'_k are respectively the conditional distributions of X given $S=k$ and of X' given $S'=k$. The Positive-Unlabeled (PU) learning problem, which has recently been the subject of much attention (see *e.g.* [dPNS14], [dPNS15], [KNdPS17]), provides a typical example of this situation. Re-using the notations introduced in section 2, in the PU problem, the testing and training distributions P and P' are respectively described by the triplets (p, F_+, F_-) and (q, F_+, F), where $F = pF_+ + (1-p)F_-$ is the marginal distribution of X. Hence, the objective pursued is to solve a binary classification task, based on the sole observation of a training sample pooling data with positive labels and unlabeled data, q denoting the theoretical fraction of positive data among the dataset. As noticed in [dPNS14] (see also [dPNS15], [KNdPS17]), the likelihood/importance function can be expressed in a simple manner, as follows:

$$\forall (x,y) \in \mathcal{X}\times\{-1,\ +1\},\ \ \Phi(x,y) = \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1}{1-q}\mathbb{I}\{y=-1\} - \frac{p}{1-q}\frac{dF_+}{dF}(x)\mathbb{I}\{y=-1\}. \tag{II.9}$$

Based on an i.i.d. sample $(X'_1, Y'_1),\ \ldots,\ (X'_n, Y'_n)$ drawn from P' combined with the knowledge of p (which can also be estimated from PU data, see e.g. [dPS14]) and using that $F_- = (1/(1-p))(F - pF_+)$, one may obtain estimators of q, F_+ and F by computing $n'_+/n = (1/n)\sum_{i=1}^n \mathbb{I}\{Y'_i = +1\}$, $\widehat{F}_+ = (1/n'_+)\sum_{i=1}^n \mathbb{I}\{Y'_i = +1\}\delta_{X'_i}$ and $\widehat{F} = (1/n'_-)\sum_{i=1}^n \mathbb{I}\{Y'_i = -1\}\delta_{X'_i}$. However, plugging these quantities into (II.9) do not permit to get a statistical version of the importance function, insofar as the probability measures $\widehat{F}_+$ and $\widehat{F}$ are mutually singular with probability one, as soon as F_+ is continuous. Of course, as proposed in [dPNS14], one may use statistical methods (*e.g.* kernel smoothing) to build distribution estimators, that ensures absolute continuity but are subject to the curse of dimensionality. However, WERM can still be applied in this case, by observing that: $\forall g \in \mathcal{G}$,

$$\mathcal{R}_P(g) = -p + \mathbb{E}_{P'}\left[\frac{2p}{q}\mathbb{I}\{g(X')=-1,\ Y'=+1\} + \frac{1}{1-q}\mathbb{I}\{g(X')=+1,\ Y'=-1\}\right], \tag{II.10}$$

which leads to the weighted empirical risk

$$\frac{2p}{n'_+}\sum_{i:Y'_i=+1}\mathbb{I}\{g(X'_i)=-1\} + \frac{1}{n'_-}\sum_{i:Y'_i=-1}\mathbb{I}\{g(X'_i)=+1\}. \tag{II.11}$$

Minimization of (II.11) yields rules $\widetilde{g}_n$ whose generalization ability regarding the binary problem related to (p, F_+, F_-) can be guaranteed, as shown by the following result, the form of the weighted empirical risk in this case being quite similar to (II.7).

Theorem 2. *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $q \in (\varepsilon,\ 1-\varepsilon)$. Let $\widetilde{g}_n$ be any minimizer of the weighted empirical risk* (II.11) *over class $\mathcal{G}$. We have with probability at least $1-\delta$:*

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g\in\mathcal{G}} \mathcal{R}_P(g) \leq \frac{2\max(2p,1)}{\varepsilon}\left(2\mathbb{E}[R'_n(\mathcal{G})] + \sqrt{\frac{2\log(2/\delta)}{n}}\right) + \frac{4(2p+1)}{\varepsilon^2}\sqrt{\frac{\log(4/\delta)}{2n}},$$

as soon as $n \geq 2\log(4/\delta)/\varepsilon^2$; where $R'_n(\mathcal{G}) = (1/n)\mathbb{E}_\sigma[\sup_{g\in\mathcal{G}} |\sum_{i=1}^n \sigma_i \mathbb{I}\{g(X'_i) \neq Y'_i\}|]$.

Remark 3. *Let $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ denote the posterior probability and recall that $(dF_+/dF_-)(x) = ((1-p)/p)(\eta(x)/(1-\eta(x)))$. Observing that*

$$\Phi(x,y) = \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1-\eta(x)}{1-q}\mathbb{I}\{y=-1\}, \tag{II.12}$$

in the case when an estimate $\widehat{\eta}(x)$ of $\eta(x)$ is available, one can perform WERM using the empirical weight function

$$\widehat{\Phi}(x,y) = \frac{np}{n'_+}\mathbb{I}\{y=+1\} + \frac{1-\widehat{\eta}(x)}{1-n'_+/n}\mathbb{I}\{y=-1\}. \tag{II.13}$$

A bound that describes how this approach generalizes, depending on the accuracy of estimate $\widehat{\eta}$, can be easily established.

3.3 Learning from Censored Data

Another important example of sample bias is the censorship setting where the learner has only access to (right) censored targets $\min(Y', C')$ instead of Y'. Intuitively, this situation occurs when Y' is a duration/date, e.g. the date of death of a patient modeled by covariates X', and the study happens at a (random) date C'. Hence if $C' \leq Y'$, then we know that the patient is still alive at time C' but the target time Y' remains unknown. This problem has been extensively studied (see e.g. [FH11], [ABGK12] and the references therein for the asymptotic theory and [ACP19] for finite-time guarantees): we show here that it is an instance of WERM. Formally, we respectively denote by P and P' the testing and training distributions of the r.v.'s $(X, \min(Y,C), \mathbb{I}\{Y \leq C\})$ and $(X', \min(Y',C'), \mathbb{I}\{Y' \leq C'\})$ both valued in $\mathbb{R}^d \times \mathbb{R}_+ \times \{0,1\}$ (with Y, Y', C, C' all non-negative r.v.'s) and such that the pairs (X,Y) and (X',Y') share the same distribution Q. Moreover, $C > Y$ with probability 1 (i.e. the testing data are never censored) and Y' and C' are assumed to be conditionally independent given X'. Hence, for all $(x,y,\delta) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0,1\}$:

$$dP(x,y,\delta) = \delta dQ(x,y)$$

and

$$\delta dP'(x,y,\delta) = \delta \mathbb{P}(C' \geq y) d\mathbb{P}(X' = x, Y' = y | C' \geq y) = \delta S_{C'}(y|x) dQ(x,y),$$

where $S_{C'}(y|x) = \mathbb{P}(C' \geq y|X' = x)$ denotes the conditional survival function of C' given X'. Then, the importance function is:

$$\forall (x, y, \delta) \in \mathbb{R}^d \times \mathbb{R}_+ \times \{0, 1\}, \quad \Phi(x, y, \delta) = \frac{dP}{dP'}(x, y, \delta) = \frac{\delta}{S_{C'}(y|x)}.$$

In survival analysis, the ratio $\delta/S_{C'}(y|x)$ is called IPCW (inverse of the probability of censoring weight) and $S_{C'}(y|x)$ can be estimated by using the Kaplan-Meier approach, see [KM58].

4 Numerical Experiments

This section illustrates the impact of reweighting by the likelihood ratio on classification performances, as a special case of the general strategy presented in section 2. A first simple illustration on known probability distributions highlights the impact of the shapes of the distributions on the importance of reweighting. This example illustrates in the infinite-sample case that separable or almost separable data do not require reweighting, in contrast to noisy data. Since the distribution shapes are unknown for real data, we infer that reweighting will have variable effectiveness, depending on the dataset. We detail here a second experiment that uses the structure of ImageNet to illustrate reweighting with a stratified population and strata distribution bias or *strata bias*. The code of the experiments can be found at `https://drive.google.com/drive/folders/1-tWJ4n4WyXuTza8dLPngyHSVprKUZFVJ?usp=sharing`.

We focus on the *learning from biased stratified data* setting introduced in section 3.1 by leveraging the ImageNet Large Scale Visual Recognition Challenge (ILSVRC); a well-known benchmark for the image classification task, see [RDS+14] for more details.

The challenge consists in learning a classifier from 1.3 million training images spread out over 1,000 classes. Performance is evaluated using the validation dataset of 50,000 images of ILSVRC as our test dataset. ImageNet is an image database organized according to the WordNet hierarchy, which groups nouns in sets of related words called synsets. In that context, images are examples of very precise nouns, e.g. *flamingo*, which are contained in a larger synset, e.g. *bird*.

The impact of reweighting in presence of strata bias is illustrated on the ILSVRC classification problem with broad significance synsets for strata. To do this, we encode the data using deep neural networks. Specifically our encoding is the flattened output of the last convolutional layer of the network ResNet50 introduced in [HZRS15]. It was trained for classification on the training dataset of ILSVRC. The encodings $X_1, \ldots, X_n$ belong to a 2,048-dimensional space.

A total of 33 strata are derived from a list of high-level categories provided by ImageNet[1]. By default, strata probabilities p_k and p'_k for $1 \leq k \leq K$ are equivalent between training and testing datasets, meaning that reweighting by Φ would have little to no effect. Since our testing data is the validation data of ILSVRC, we have around 25 times

[1] `http://www.image-net.org/about-stats`

Model	Reweighting	miss rate	top-5 error
Linear	Unif. $\widehat{\Phi} = 1$	0.344	0.130
	Strata $\widehat{\Phi}$	**0.329**	**0.120**
	Class $\widehat{\Phi}$	0.328	0.119
	No bias	0.297	0.102
MLP	Unif. $\widehat{\Phi} = 1$	0.371	0.143
	Strata $\widehat{\Phi}$	**0.364**	**0.138**
	Class $\widehat{\Phi}$	0.363	0.138
	No bias	0.316	0.111

Table II.1: Results for the strata reweighting experiment with ImageNet.

more training than testing data. Introducing a strata bias parameter $0 \leq \gamma \leq 1$, we set the strata train probabilities such that $p'_k = \gamma^{1-\lfloor K/2 \rfloor/k} p_k$ before renormalization and remove train instances so that the train set has the right distribution over strata. When γ is close to one, there is little to no strata bias. In contrast, when γ approaches 0, strata bias is extreme.

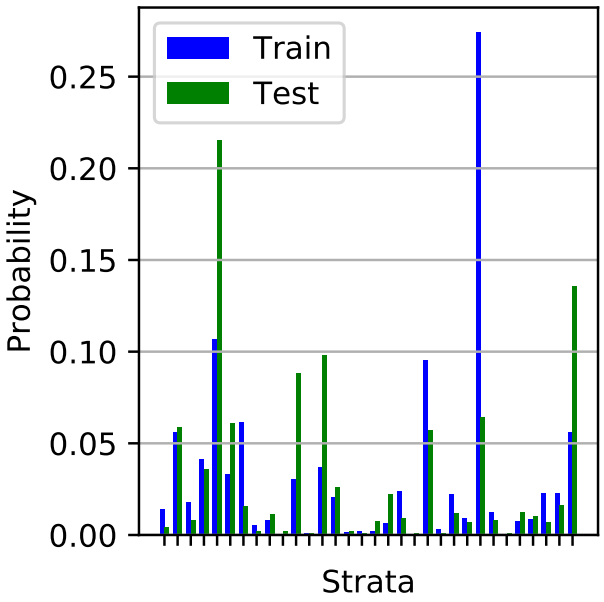
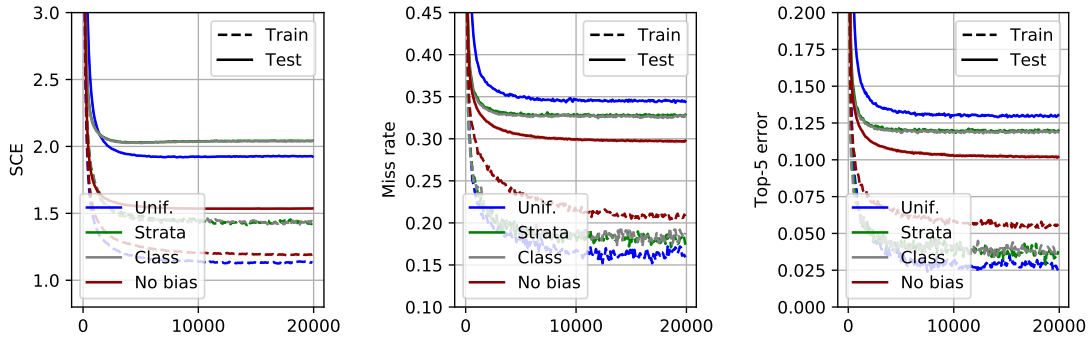
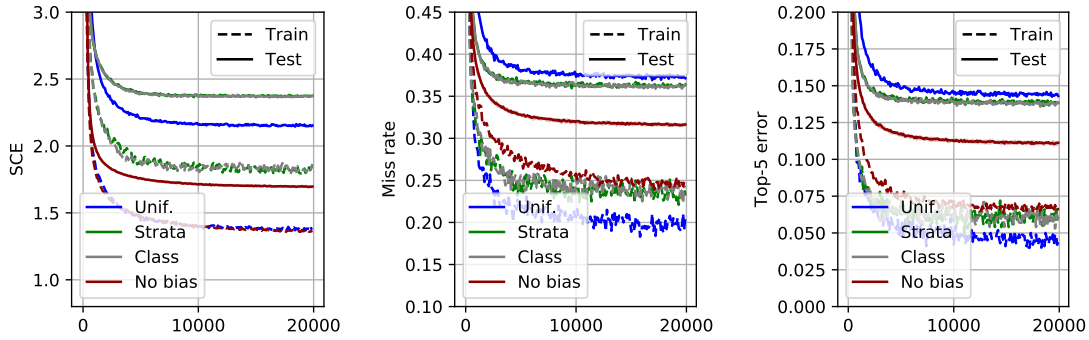


Figure II.1: Comparison of p_k's and p'_k's.



Dynamics for the SCE. Dynamics for the miss rate. Dynamics for the top-5 error.

Figure II.2: Dynamics for the linear model for the strata reweighting experiment with ImageNet.



Dynamics for the SCE. Dynamics for the miss rate. Dynamics for the top-5 error.

Figure II.3: Dynamics for the MLP model for the strata reweighting experiment with ImageNet.

The models used are a linear model and a multilayer perceptron (MLP) with one hidden layer. We report better performance when reweighting using the strata information, compared to the case where the strata information is ignored, see Figure II.1 and Table II.1. For comparison, we added two reference experiments: one which reweights the train instances by the class probabilities, which we do not know in a stratified population experiment, and one with more data and no strata bias because it uses all of the ILSVRC train data. The dominance of the linear model over the MLP can be justified by the much higher number of parameters to estimate

4.1 Importance of Reweighting for Simple Distributions

Introduce a random pair (X, Y) in $[0,1]\times\{-1,+1\}$ where $X \mid Y = +1$ has for probability density function (pdf) $f_+(x) = (1+\alpha)x^\alpha, \alpha > 0$ and $X \mid Y = -1$ has for pdf $f_-(x) = (1+\beta)(1-x)^\beta, \beta > 0$. As in section 2, the train and test datasets have different class probabilities p' and p for $Y = +1$. The loss ℓ is defined as $\ell(\theta, z) = \mathbb{I}\{(x-\theta)y \geq 0\}$ where $\theta > 0$ is a learnt parameter.

The true risk can be explicitely calculated. For $\theta > 0$, we have

$$R_P(\theta) = p\theta^{1+\alpha} + (1-p)(1-\theta)^{1+\beta},$$

and the optimal threshold θ_p^* can be found by derivating the risk $R_P(\theta)$. The derivative is zero when θ satisfies

$$p(1+\alpha)\theta^\alpha = (1-p)(1+\beta)(1-\theta)^\beta. \tag{II.14}$$

Solving Eq. (II.14) is straightforward for well-chosen values of α, β, which are detailed in Table II.2. The excess error $\mathcal{E}(p', p) = R_P(\theta_{p'}^*) - R_P(\theta_p^*)$ for the diagonal entries of Table II.2 are plotted in Figure II.4, in the infinite sample case.

	(α, β)			
	$(0,0)$	$(1/2, 1/2)$	$(1,1)$	$(2,2)$
θ_p^*	$[0,1]$	$\frac{(1-p)^2}{p^2+(1-p)^2}$	$1-p$	$\frac{\sqrt{1-p}}{\sqrt{p}+\sqrt{1-p}}$

Table II.2: Optimal parameters θ^* for different values of α, β.

The results of Figure II.4 show that the optimum for the train distribution is significantly different from the optimum for the test distribution when the problem involves Bayes noise.

4.2 On the Real Data Experiment

We provide further details about the real data experiment.

Strategy to Induce Bias in Balanced Datasets. In the real data experiment described above, a strategy is used to induce class distribution bias or strata bias, since the data is uniformly distributed on strata for the train and test set. Since the experiment involves a small test dataset, it is kept intact, while we discard elements of the train dataset to induce bias between the train and test datasets. The bias is parameterized by a single parameter γ, such that when γ is close to one, there is little strata or class bias, while when γ approaches 0, bias is extreme.

The bias we induce is inspired by a power law, which is often used to model unequal distributions. The distribution on the strata of the train set is modified so that the

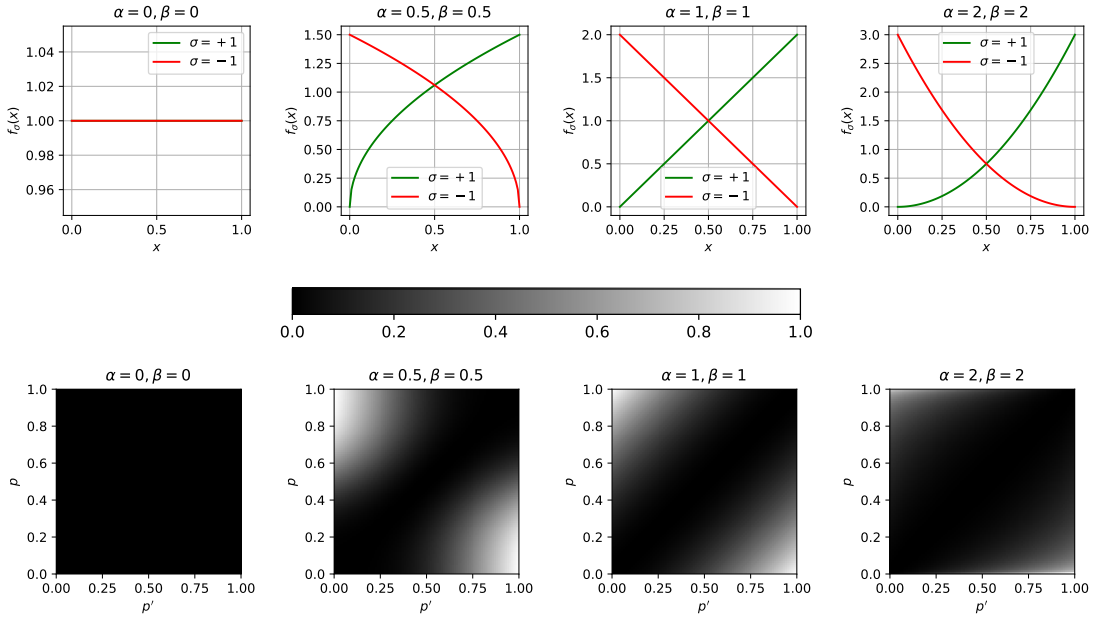


Figure II.4: Pdf's and values of the excess risk $\mathcal{E}(p', p)$ for different values of α, β.

generated train set follows a power law. Formally, the power law distribution $\{p'_k\}_{k=1}^K$ over $S \in \{1, \dots, K\}$, is defined for all $1 \leq k \leq K$ as

$$p'_k = \frac{\gamma^{-\frac{\lfloor K/2 \rfloor}{\sigma(k)}} p_k}{\sum_{l=1}^K \gamma^{-\frac{\lfloor K/2 \rfloor}{\sigma(k)}} p_k},$$

where σ is a random permutation in $\{1, \dots, K\}$.

To generate a train dataset with modality distribution $\{p'_k\}_{k=1}^K$, we sample instances from the original train data set $\mathcal{D}_n^\circ = \{(X'_i, Y'_i, S'_i)\}_{i=1}^n$, where Y'_i is the class, S'_i is the strata. The generated train dataset is noted $\mathcal{D}_n$. First, we define candidates $\mathcal{I}_k = \{i \mid 1 \leq i \leq n, S'_i = k\}$ for each strata $k \in \{1, \dots, K\}$. Then we select one of the candidate sets $\mathcal{I}_k$ with the probabilities p'_k's, to remove one of its elements, selected at random, and place it in the train dataset $\mathcal{D}_n$. We repeat this operation until one of the candidate sets is empty. A more efficient implementation of this process was used in the provided code.

Models. We compare two models: a linear model and a multilayer perceptron (MLP) with one hidden layer of size 1,524. Given a classification problem of input x of dimension d with K classes, precisely with $d = 2048, K = 1000$, a linear model simply learns the weights matrix $W \in \mathbb{R}^{d \times K}$ and the bias vector $b \in \mathbb{R}^K$ and outputs logits $l = W^\top x + b$. On the other hand, the MLP has a hidden layer of dimension $h = \lfloor (d+K)/2 \rfloor$ and learns the weights matrices $W_1 \in \mathbb{R}^{d,h}, W_2 \in \{h, K\}$ and bias vectors $b_1 \in \mathbb{R}^h, b_2 \in \mathbb{R}^K$ and outputs logits $l = W_2^\top h(W_1^\top x + b_1) + b_2$ where h is the ReLU function, i.e. $h : x \mapsto$

$\max(x, 0)$. The MLP model involves approximatively 5M (million) parameters, while the MLP model uses only 2M. The weight decay or l2 penalization for the linear model and MLP model are written, respectively

$$\mathcal{P} = \frac{1}{2}\|W\| \quad \text{and} \quad \mathcal{P} = \frac{1}{2}\|W_1\| + \frac{1}{2}\|W_1\|.$$

Cost Function. The cost function is the Softmax Cross-Entropy (SCE), which is the most used classification loss in deep learning. Specifically, given logits $l = (l_1, \ldots, l_K) \in \mathbb{R}^K$, the softmax function is $\gamma : \mathbb{R}^k \to [0, 1]^K$ with $\gamma = (\gamma_1, \ldots, \gamma_K)$ and for all $k \in \{1, \ldots, K\}$,

$$\gamma_k : l \mapsto \frac{\exp(l_k)}{\sum_{j=0}^{K} \exp(l_j)}.$$

Given an instance with logits l and ground truth class value y, the expression of the softmax cross-entropy $c(l, y)$ is

$$c(l, y) = \sum_{k=1}^{K} \mathbb{I}\{y = k\} \log\left(\gamma_k(l)\right).$$

The loss that is reweighted depending on the cases as described in section 3 is this quantity $c(l, y)$. The loss on the test set is never reweighted, since the test set is the target distribution. The weights and bias of the model that yield the logits are tuned using backpropagation on this loss averaged on random batches of B elements of the training data summed with the regularization term $\lambda \cdot \mathcal{P}$ where λ is a hyperparameter that controls the strength of the regularization.

Preprocessing, Optimization, Parameters. The images of ILSVRC were encoded using the implementation of ResNet50 provided by the library *keras*, see [C+15], by taking the flattened output of the last convolutional layer.

Optimization is performed using a momentum batch gradient descent algorithm on batches of size 1,000, which updates the parameters θ_t at timestep t with an update vector v_t by performing the following operations:

$$v_t = \gamma v_{t-1} + \eta \nabla C(\theta_{t-1}),$$
$$\theta_t = \theta_{t-1} - v_t,$$

where $\eta = 0.001$ is the learning rate and $\gamma = 0.9$ is the momentum, as explained in [Rud16]. The weight decay parameters λ were cross-validated by trying values on the logarithmic scale $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and then we tried more fine-grained values between the two best results, in practice 10^{-3} was best and 10^{-2} was second best so we tried $\{0.002, 0.003, 0.004, 0.005\}$. The standard deviation initialization of the weights $\sigma_0 = 0.01$ was chosen by trial-and-error to avoid overflows. The learning rate was fixed after trying different values to have fast convergence while keeping good convergence properties.

Stratified Information for ImageNet. In this section, we detail the data preprocessing necessary to assign strata to the ILSVRC data. These were constructed using a list of 27 high-level categories found on the ImageNet website[2].

Each ILSVRC image has a ground truth low level synset, either from the name of the training instance, or in the validation textfile for the validation dataset, that is provided by the ImageNet website. The ImageNet API [3] provides the hierarchy of synsets in the form of *is-a* relationships, e.g. *a flamingo is a bird*. Using this information, for each synset in the validation and training database, we gathered all of its ancestors in the hierarchy that were high-level categories. Most of the synsets had only one ancestor, which then accounts for one stratum. Some of the synsets had no ancestors, or even several ancestors in the table, which creates extra strata, either a *no-category* stratum or a strata composed of the union of several ancestors. The final distribution of the dataset over the created strata is summarized by Figure II.5. Observe the presence of a *no_strata* stratum and of unions of two high-level synsets strata, e.g. *n00015388_n01905661*. The definitions of the strata can be requested to the API, see Table II.3 for examples.

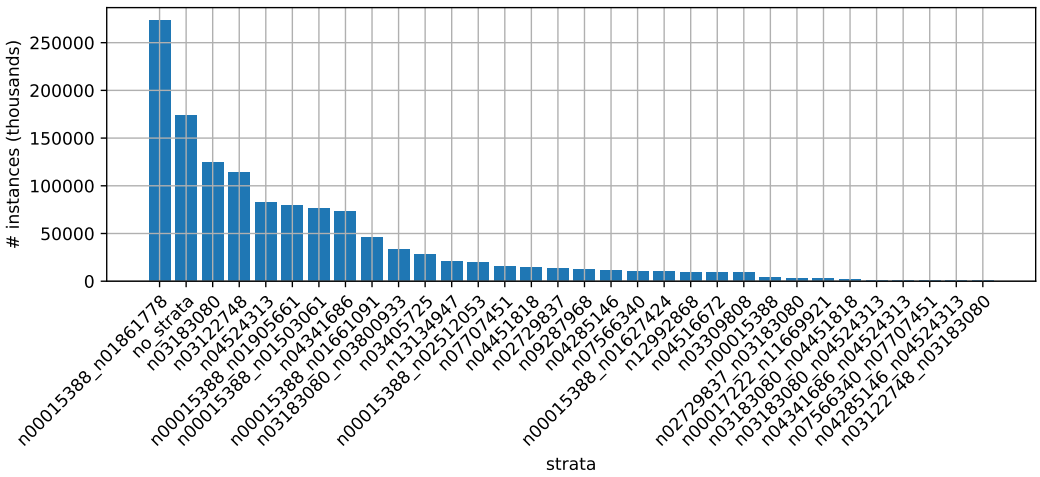


Figure II.5: Distribution of the ImageNet train dataset over the created strata, with examples of definitions in Table II.3.

5 Conclusion

In this chapter, we have considered specific transfer learning problems, where the distribution of the test data P differs from that of the training data, P', and is absolutely continuous with respect to the latter. This setup encompasses many situations in practice, where the data acquisition process is not perfectly controlled. In this situation, a simple change of measure shows that the target risk may be viewed as the expectation of

[2] `http://www.image-net.org/about-stats`
[3] `http://image-net.org/download-API`

Strata name		Definition	
n00015388	n01861778	animal, animate being, beast (...)	mammal, mammalian
n04524313		vehicle	
n13134947		fruit	
n00015388	n02512053	animal, animate being, beast (...)	fish
n00017222	n11669921	plant, flora, plant life	flower
n07566340	n07707451	foodstuff, food product	vegetable, veggie, veg

Table II.3: Examples of definitions of the strata created for the experiments.

a weighted version of the basic empirical risk, with ideal weights given by the importance function $\Phi = dP/dP'$, unknown in practice. Throughout this chapter, we have shown that, in statistical learning problems corresponding to a wide variety of practical applications, these ideal weights can be replaced by statistical versions based solely on the training data combined with very simple information about the target distribution. The generalisation capacity of rules learnt from biased training data by minimization of the weighted empirical risk has been established, with learning bounds. These theoretical results are also illustrated with several numerical experiments.

As a direction of future research, we propose to extend our WERM approach to an iterative procedure for the challenging problem of bipartite ranking based on PU data. We motivate this perspective in the next section.

6 Perspective - Extension to Iterative WERM

As highlighted in Remark 3, the importance function can be expressed as a function of the ideal decision function in certain situations: Eq. (II.12) involves the regression function $\eta(x)$, that defines the optimal (Bayes) classifier $g^*(x) = 2\mathbb{I}\{\eta(x) \geq 1/2\} - 1$. This simple observation paves the way for a possible incremental application of the WERM approach: in the case where the solution of the WERM problem considered outputs an estimate of the optimal decision function, it can be next re-used for defining and solving a novel WERM problem. Whereas binary classification based on PU data only aims at recovering a single level set of the posterior probability $\eta(x)$, it is not the case of a more ambitious statistical learning problem, referred to as *bipartite ranking*, for which such an incremental version of WERM can be described.

Bipartite Ranking Based on PU Data. In bipartite ranking, the statistical challenge consists of ranking all the instances $x \in \mathcal{X}$ through a *scoring function* $s : \mathcal{X} \to \mathbb{R}$ in the same order as the likelihood ratio $\Psi(X) = (dF_+/dF_-)(X)$, or, equivalently, as the regression function $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$: the higher the score $s(X)$, the more likely one should observe $Y = +1$. Let $\mathcal{S} = \{s : \mathcal{X} \to \mathbb{R} \text{ measurable}\}$ denotes the set of all scoring functions on the input space $\mathcal{X}$. A classical way of measuring "how much stochastically larger" a distribution $\mathcal{G}$ on $\mathbb{R}$ than another one, $\mathcal{H}$ say, consists in

drawing the "probability-probability plot":

$$t \in \mathbb{R} \mapsto (1 - \mathcal{H}(t),\ 1 - \mathcal{G}(t)),$$

with the convention that possible jumps are connected by line segments (in order to guarantee the continuity of the curve). Equipped with this convention, this boils down to plot the graph of the mapping

$$\mathrm{ROC}_{\mathcal{H},\mathcal{G}} : \alpha \in (0,1) \mapsto \mathrm{ROC}_{\mathcal{H},\mathcal{G}} = 1 - \mathcal{G} \circ \mathcal{H}^{-1}(1-\alpha),$$

where $\Gamma^{-1}(u) = \inf\{t \in \mathbb{R} :\ \Gamma(t) \geq u\}$ denotes the pseudo-inverse of any cumulative distribution function $\Gamma(t)$ on $\mathbb{R}$. The closer to the left upper corner of the unit square $[0,1]^2$, the larger the distribution $\mathcal{G}$ is compared to $\mathcal{H}$ in a stochastic sense. This approach is known as ROC analysis. The gold standard for evaluating the ranking performance of a scoring function s is thus the ROC curve:

$$\mathrm{ROC}_s \stackrel{def}{=} \mathrm{ROC}_{F_{s,-},F_{s,+}},$$

where $F_{s,+}$ and $F_{s,-}$ denote the conditional distributions of $s(X)$ given $Y = +1$ and given $Y = -1$ respectively, *i.e.* the images of class distributions F_+ and F_- by the mapping $s(x)$. Indeed, it follows from a standard Neyman-Pearson argument that the ROC curve ROC^* of strictly increasing transforms of $\eta(x)$ is optimal with respect to this criterion in the sense that:

$$\forall \alpha \in (0,1),\ \ \mathrm{ROC}_s(\alpha) \leq \mathrm{ROC}^*(\alpha),$$

for any scoring function s. We set $\mathcal{S}^* = \{T \circ \eta :\ T : (0,1) :\to \mathbb{R}\}$. A summary quantity of this functional criterion that is widely used in practice is the *Area Under the ROC Curve* (AUC in short), given by:

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}_s(\alpha)\ d\alpha,$$

for $s \in \mathcal{S}$. Beyond its scalar nature, an attractive property of this criterion lies in the fact that it can be interpreted in a probabilistic manner, insofar as we have the relation: $\forall s \in \mathcal{S}$,

$$\mathrm{AUC}(s) = \mathbb{P}\left\{s(X) < s(X') \mid (Y,Y') = (-1,+1)\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X') \mid (Y,Y') = (-1,+1)\right\}.$$

Denoting by (X_i, Y_i), $i \in \{1,\ 2\}$, independent copies of the pair (X,Y) and placing ourselves in the situation where $s(X)$'s distribution is continuous, as observed in [CLV08], we have $\mathrm{AUC}(s) = 1 - L_P(s)/(2p(1-p))$, where

$$L_P(s) \stackrel{def}{=} \mathbb{P}\left\{(s(X_1) - s(X_2))(Y_1 - Y_2)) < 0\right\},$$

is the *ranking risk*, the theoretical rate of discording pairs namely, that can be viewed as a pairwise classification risk. Hence, bipartite ranking can be formulated as the problem of learning a scoring function s that minimizes the ranking risk

$$L_P(s) = \mathbb{E}_{P' \otimes P'}\left[\frac{dP}{dP'}(X_1', Y_1')\frac{dP}{dP'}(X_2', Y_2') \times \mathbb{I}\left\{(s(X_1') - s(X_2'))(Y_1' - Y_2')) < 0\right\}\right].$$

Now, using Eq. (II.12) and the fact that $\eta = p\Psi/(1-p+p\Psi)$, we have:

$$\frac{dP}{dP'}(x,y) = \Phi(x,y) = \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1-\eta(x)}{1-q}\mathbb{I}\{y=-1\}$$
$$= \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1-p}{(1-q)(1-p+p\Psi(x))}\mathbb{I}\{y=-1\}.$$

Therefore, it has been shown in [CV09b] (see Corollary 5 therein) that for any s^* in $\mathcal{S}^*$,

$$\frac{dF_+}{dF_-}(X) = \frac{dF_{s^*,+}}{dF_{s^*,-}}(s^*(X)) \text{ almost-surely.}$$

For any s candidate, setting $\Psi_s(x) = dF_{s,+}/dF_{s,-}(s(x))$, one can define

$$\Phi_s(x,y) = \frac{p}{q}\mathbb{I}\{y=+1\} + \frac{1-p}{(1-q)(1-p+p\Psi_s(s(x)))}\mathbb{I}\{y=-1\}.$$

From this formula, it is the easy to see how an incremental use of the WERM could be implemented.

- Start from an initial guess s for the optimal scoring functions (*e.g.* solve the empirical ranking risk minimization problem ignoring the bias issue)
- Estimate Φ_s from the (X_i', Y_i')'s and the knowledge of p, observing that one is not confronted with the curse of dimensionality in this case
- Solve the Weighted Empirical Ranking Risk Minimization problem using the weight function
$$\widehat{\Phi}_s(x_1,y_1)\widehat{\Phi}_s(x_2,y_2),$$
which produces a new scoring function s and iterate.

Investigating the performance of such an incremental procedure will be the subject of future research.

7 Technical Proofs

Here we detail the proofs of the results stated in the present chapter and discuss their connection with related work.

Proof of Lemma 1

Let $\delta \in (0,1)$. Applying the classic maximal deviation bound stated in Theorem 3.2 of [BBL05] to the bounded class $\mathcal{K} = \{z \in \mathcal{Z} \mapsto \Phi(z)l(\theta,z) : \ \theta \in \Theta\}$, we obtain that, with probability at least $1-\delta$:

$$\sup_{\theta\in\Theta}\left|\widetilde{\mathcal{R}}_{w^*,n}(\theta) - \mathbb{E}\left[\widetilde{\mathcal{R}}_{w^*,n}(\theta)\right]\right| \leq 2\mathbb{E}\left[R_n'(\mathcal{K})\right] + ||\Phi||_\infty \sup_{(\theta,z)\in\Theta\times\mathcal{Z}}|\ell(\theta,z)|\sqrt{\frac{2\log(1/\delta)}{n}}.$$

In addition, by virtue of the contraction principle, we have $R'_n(\mathcal{K}) \leq ||\Phi||_\infty R'_n(\mathcal{F})$ almost-surely. The desired result can be thus deduced from the bound above combined with the classic bound

$$\mathcal{R}_P(\tilde{\theta}_n^*) - \min_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 2 \sup_{\theta\in\Theta} \left| \widetilde{\mathcal{R}}_{w^*,n}(\theta) - \mathbb{E}\left[\widetilde{\mathcal{R}}_{w^*,n}(\theta)\right] \right|.$$

Proof of Lemma 2

Apply twice the Taylor expansion

$$\frac{1}{x} = \frac{1}{a} - \frac{x-a}{a^2} + \frac{(x-a)^2}{xa^2},$$

so as to get

$$\begin{aligned}
\frac{1}{n'_+/n} &= \frac{1}{p'} - \frac{n'_+/n - p'}{p'^2} + \frac{(n'_+/n - p')^2}{p'^2 n'_+/n}, \\
\frac{1}{n'_-/n} &= \frac{1}{1-p'} - \frac{n'_-/n - 1 + p'}{(1-p')^2} + \frac{(n'_-/n - 1 + p')^2}{(1-p')^2 n'_-/n}.
\end{aligned}$$

This yields the decomposition

$$\begin{aligned}
\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g) = &- \frac{p}{p'^2}\left(\frac{n'_+}{n} - p'\right)\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{g(X'_i) = -1,\ Y'_i = +1\} \\
&- \frac{1-p}{(1-p')^2}\left(\frac{n'_-}{n} - 1 + p'\right)\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{g(X'_i) = +1,\ Y'_i = -1\} \\
&+ \frac{p(n'_+/n - p')^2}{p'^2 n'_+/n}\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{g(X'_i) = -1,\ Y'_i = +1\} \\
&+ \frac{(1-p)(n'_-/n - 1 + p')^2}{(1-p')^2 n'_-/n}\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{g(X'_i) = +1,\ Y'_i = -1\}.
\end{aligned}$$

We deduce that

$$\left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g)\right| \leq \frac{|n'_+/n - p'|}{\varepsilon^2}\left(1 + |n'_+/n - p'|\left(\frac{p}{n'_+/n} + \frac{1-p}{1-n'_+/n}\right)\right).$$

By virtue of Hoeffding inequality, we obtain that, for any $\delta \in (0,1)$, we have with probability larger than $1-\delta$:

$$\left|n'_+/n - p'\right| \leq \sqrt{\frac{\log(2/\delta)}{2n}},$$

so that, in particular, $\min\{n'_+/n,\ 1 - n'_+/n\} \geq \varepsilon - \sqrt{\log(2/\delta)/(2n)}$. This yields the desired result.

Proof of Corollary 1

Observe first that $||\Phi||_\infty \leq \max(p,\ 1-p)/\varepsilon$ and

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g\in\mathcal{G}} \mathcal{R}_P(g) \leq 2\sup_{g\in\mathcal{G}} \left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g)\right| + 2\sup_{g\in\mathcal{G}} \left|\widetilde{\mathcal{R}}_{w^*,n}(g) - \mathcal{R}_P(g)\right|.$$

The result then directly follows from the application of Lemmas 1-2 combined with the union bound.

Proof of Theorem 1

Observe first that $||\Phi||_\infty \leq \max_k p_k/\varepsilon$ and

$$\mathcal{R}_P(\widetilde{\theta}_n^*) - \inf_{\theta\in\Theta} \mathcal{R}_P(\theta) \leq 2\sup_{\theta\in\Theta} \left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(\theta) - \widetilde{\mathcal{R}}_{w^*,n}(\theta)\right| + 2\sup_{\theta\in\Theta} \left|\widetilde{\mathcal{R}}_{w^*,n}(\theta) - \mathcal{R}_P(\theta)\right|.$$

The result then directly follows from the application of Lemmas 1-3 combined with the union bound.

Lemma 3. *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $p'_k \in (\varepsilon,\ 1-\varepsilon)$ for $k \in \{1,\ \ldots,\ K\}$. For any $\delta \in (0,1)$, we have with probability larger than $1-\delta$:*

$$\sup_{\theta\in\Theta} \left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(\theta) - \widetilde{\mathcal{R}}_{w^*,n}(\theta)\right| \leq \frac{2L}{\varepsilon^2}\sqrt{\frac{\log(2K/\delta)}{2n}},$$

as soon as $n \geq 2\log(2K/\delta)/\varepsilon^2$, where $L = \sup_{(\theta,z)\in\Theta\times\mathcal{Z}} \ell(\theta,z)$.

PROOF.

Apply the Taylor expansion

$$\frac{1}{x} = \frac{1}{a} - \frac{x-a}{a^2} + \frac{(x-a)^2}{xa^2},$$

so as to get for all $k \in \{1,\ldots,K\}$

$$\frac{1}{n'_k/n} = \frac{1}{p'_k} - \frac{n'_k/n - p'_k}{p'^2_k} + \frac{(n'_k/n - p'_k)^2}{p'^2_k n'_k/n}.$$

This yields the decomposition

$$\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(\theta) - \widetilde{\mathcal{R}}_{w^*,n}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(\theta, Z'_i) \sum_{k=1}^K \mathbb{I}\{S'_i = k\} \left(-\frac{p_k}{p'^2_k}\left(\frac{n'_k}{n} - p'_k\right) + \frac{p_k(n'_k/n - p'_k)^2}{p'^2_k n'_k/n}\right).$$

We deduce that

$$\left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(\theta) - \widetilde{\mathcal{R}}_{w^*,n}(\theta)\right| \leq \frac{L}{\varepsilon^2}\sum_{k=1}^K |n'_k/n - p'_k| p_k \left(1 + \frac{|n'_k/n - p'_k|}{n'_k/n}\right).$$

By virtue of Hoeffding inequality, we obtain that, for any $k \in \{1, \ldots, K\}$ and $\delta \in (0, 1)$, we have with probability larger than $1 - \delta$:

$$\left| n'_k/n - p'_k \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}},$$

so that, by a union bound, $\max_k \{n'_k/n\} \geq \varepsilon - \sqrt{\log(2K/\delta)/(2n)}$. This yields the desired result.

Proof of Theorem 2

Observe first that $||\Phi||_\infty \leq \max(2p,\ 1)/\varepsilon$ and

$$\mathcal{R}_P(\widetilde{g}_n) - \inf_{g \in \mathcal{G}} \mathcal{R}_P(g) \leq 2 \sup_{g \in \mathcal{G}} \left| \widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g) \right| + 2 \sup_{g \in \mathcal{G}} \left| \widetilde{\mathcal{R}}_{w^*,n}(g) - \mathcal{R}_P(g) \right|,$$

with weighted empirical risk $\widetilde{\mathcal{R}}_{w^*,n}(g)$ defined in (II.11). The result then directly follows from the application of Lemmas 1-4 combined with the union bound.

Lemma 4. *Let $\varepsilon \in (0,\ 1/2)$. Suppose that $q \in (\varepsilon,\ 1 - \varepsilon)$. For any $\delta \in (0, 1)$, we have with probability larger than $1 - \delta$:*

$$\sup_{g \in \mathcal{G}} \left| \widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g) \right| \leq \frac{2(2p+1)}{\varepsilon^2} \sqrt{\frac{\log(2/\delta)}{2n}},$$

as soon as $n \geq 2 \log(2/\delta)/\varepsilon^2$.

PROOF. Apply twice the Taylor expansion

$$\frac{1}{x} = \frac{1}{a} - \frac{x - a}{a^2} + \frac{(x-a)^2}{xa^2},$$

so as to get

$$\begin{aligned} \frac{1}{n'_+/n} &= \frac{1}{q} - \frac{n'_+/n - q}{q^2} + \frac{(n'_+/n - q)^2}{q^2 n'_+/n}, \\ \frac{1}{n'_-/n} &= \frac{1}{1-q} - \frac{n'_-/n - 1 + q}{(1-q)^2} + \frac{(n'_-/n - 1 + q)^2}{(1-q)^2 n'_-/n}. \end{aligned}$$

This yields the decomposition

$$\begin{aligned}\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g) = &-\frac{2p}{q^2}\left(\frac{n'_+}{n} - q\right)\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{g(X'_i) = -1,\ Y'_i = +1\} \\ &- \frac{1}{(1-q)^2}\left(\frac{n'_-}{n} - 1 + q\right)\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{g(X'_i) = +1,\ Y'_i = -1\} \\ &+ \frac{2p(n'_+/n - q)^2}{q^2 n'_+/n}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{g(X'_i) = -1,\ Y'_i = +1\} \\ &+ \frac{(n'_-/n - 1 + q)^2}{(1-q)^2 n'_-/n}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{g(X'_i) = +1,\ Y'_i = -1\}.\end{aligned}$$

We deduce that

$$\left|\widetilde{\mathcal{R}}_{\widehat{w}^*,n}(g) - \widetilde{\mathcal{R}}_{w^*,n}(g)\right| \leq \frac{|n'_+/n - q|}{\varepsilon^2}\left(2p + 1 + |n'_+/n - q|\left(\frac{2p}{n'_+/n} + \frac{1}{1 - n'_+/n}\right)\right).$$

By virtue of Hoeffding inequality, we obtain that, for any $\delta \in (0,1)$, we have with probability larger than $1-\delta$:

$$|n'_+/n - q| \leq \sqrt{\frac{\log(2/\delta)}{2n}},$$

so that, in particular, $\min\{n'_+/n,\ 1 - n'_+/n\} \geq \varepsilon - \sqrt{\log(2/\delta)/(2n)}$. This yields the desired result.

CHAPTER III

Ranking Data with Continuous Labels through Oriented Recursive Partitions

Abstract

We formulate a supervised learning problem, referred to as *continuous ranking*, where a continuous real-valued label Y is assigned to an observable r.v. X taking its values in a feature space $\mathcal{X}$ and the goal is to order all possible observations x in $\mathcal{X}$ by means of a scoring function $s : \mathcal{X} \to \mathbb{R}$ so that $s(X)$ and Y tend to increase or decrease together with highest probability. This problem generalizes *bi/multi-partite ranking* to a certain extent and the task of finding optimal scoring functions $s(x)$ can be naturally cast as optimization of a dedicated functional criterion, called the IROC curve here, or as maximization of the Kendall τ related to the pair $(s(X), Y)$. From the theoretical side, we describe the optimal elements of this problem and provide statistical guarantees for empirical Kendall τ maximization under appropriate conditions for the class of scoring function candidates. We also propose a recursive statistical learning algorithm tailored to empirical IROC curve optimization and producing a piecewise constant scoring function that is fully described by an oriented binary tree. Preliminary numerical experiments highlight the difference in nature between *regression* and *continuous ranking* and provide strong empirical evidence of the performance of empirical optimizers of the criteria proposed.

1 Introduction

The predictive learning problem considered in this chapter can be easily stated in an informal fashion, as follows. Given a collection of objects of arbitrary cardinality, $N \geq 1$ say, respectively described by characteristics $x_1, \ldots, x_N$ in a feature space $\mathcal{X}$, the goal is to learn how to order them by increasing order of magnitude of a certain unknown continuous variable y. To fix ideas, the attribute y can represent the 'size' of the object and be difficult to measure, as for the physical measurement of microscopic bodies in chemistry and biology or the cash flow of companies in quantitative finance and the features

x may then correspond to *indirect measurements*. The most convenient way to define a preorder on a feature space $\mathcal{X}$ is to transport the natural order on the real line onto it by means of a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}$: an object with charcateristics x is then said to be 'larger' ('strictly larger', respectively) than an object described by x' according to the scoring rule s when $s(x') \leq s(x)$ (when $s(x) < s(x')$). Statistical learning boils down here to build a scoring function $s(x)$, based on a *training* data set $\mathcal{D}_n = \{(X_1, Y_1), \ \ldots, \ (X_n, Y_n)\}$ of objects for which the values of all variables (direct and indirect measurements) have been jointly observed, such that $s(X)$ and Y tend to increase or decrease together with highest probability or, in other words, such that the ordering of new objects induced by $s(x)$ matches that defined by their true measures as well as possible. This problem, that shall be referred to as *continuous ranking* throughout the chapter can be viewed as an extension of *bipartite ranking*, where the output variable Y is assumed to be binary and the objective can be naturally formulated as a functional M-estimation problem by means of the concept of ROC curve, see [CV09b]. Refer also to [CLV05], [FISS03], [AGH$^+$05] for approaches based on the optimization of summary performance measures such as the AUC criterion in the binary context. Generalization to the situation where the random label is ordinal and may take a finite number $K \geq 3$ of values is referred to as *multipartite ranking* and has been recently investigated in [SCV13] (see also *e.g.* [RA05]), where distributional conditions guaranteeing that ROC surface and the VUS criterion can be used to determine optimal scoring functions are exhibited in particular.

It is the major purpose of this chapter to formulate the *continuous ranking* problem in a quantitative manner and explore the connection between the latter and bi/multi-partite ranking. Intuitively, optimal scoring rules would be also optimal for any bipartite subproblem defined by thresholding the continuous variable Y with cut-off $t > 0$, separating the observations X such that $Y < t$ from those such that $Y > t$. Viewing this way *continuous ranking* as a continuum of nested bipartite ranking problems, we provide here sufficient conditions for the existence of such (optimal) scoring rules and we introduce a concept of *integrated ROC curve* (IROC curve in abbreviated form) that may serve as a natural performance measure for continuous ranking, as well as the related notion of *integrated AUC criterion*, a summary scalar criterion, akin to Kendall tau. Generalization properties of empirical Kendall tau maximizers are discussed in subsection 4.2. The chapter also introduces a novel recursive algorithm that solves a discretized version of the empirical *integrated ROC curve* optimization problem, producing a scoring function that can be computed by means of a hierarchical combination of binary classification rules. Numerical experiments providing strong empirical evidence of the relevance of the approach promoted in this chapter are also presented.

The chapter is structured as follows. The probabilistic framework we consider is described and key concepts of bi/multi-partite ranking are briefly recalled in section 2. Conditions under which optimal solutions of the problem of ranking data with continuous labels exist are next investigated in section 3, while section 4 introduces a dedicated quantitative (functional) performance measure, the IROC curve. The algorithmic approach we propose in order to learn scoring functions with nearly optimal IROC curves

is presented at length in section 5. Numerical results are displayed in section 6. Some technical proofs are deferred to section 8.

2 Notations and Preliminaries

Throughout the chapter, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$. The pseudo-inverse of any cdf $F(t)$ on $\mathbb{R}$ is denoted by $F^{-1}(u) = \inf\{s \in \mathbb{R} : \ F(s) \geq u\}$, while $\mathcal{U}([0,1])$ denotes the uniform distribution on the unit interval $[0,1]$.

2.1 The Probabilistic Framework

Given a continuous real valued r.v. Y representing an attribute of an object, its 'size' say, and a random vector X taking its values in a (typically high dimensional euclidian) feature space $\mathcal{X}$ modelling other observable characteristics of the object (*e.g.* 'indirect measurements' of the size of the object), hopefully useful for predicting Y, the statistical learning problem considered here is to learn from $n \geq 1$ training independent observations $\mathcal{D}_n = \{(X_1, Y_1), \ \ldots, \ (X_n, Y_n)\}$, drawn as the pair (X, Y), a measurable mapping $s : \mathcal{X} \to \mathbb{R}$, that shall be referred to as a *scoring function* throughout the chapter, so that the variables $s(X)$ and Y tend to increase or decrease together: ideally, the larger the score $s(X)$, the higher the size Y. For simplicity, we assume throughout the chapter that $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$ and that the support of Y's distribution is compact, equal to $[0,1]$ say. For any $q \geq 1$, we denote by λ_q the Lebesgue measure on $\mathbb{R}^q$ equipped with its Borelian σ-algebra and suppose that the joint distribution $F_{X,Y}(dxdy)$ of the pair (X, Y) has a density $f_{X,Y}(x, y)$ w.r.t. the tensor product measure $\lambda_d \otimes \lambda_1$. We also introduces the marginal distributions $F_Y(dy) = f_Y(y)\lambda_1(dy)$ and $F_X(dx) = f_X(x)\lambda_d(dx)$, where $f_Y(y) = \int_{x \in \mathcal{X}} f_{X,Y}(x, y)\lambda_d(dx)$ and $f_X(x) = \int_{y \in [0,1]} f_{X,Y}(x, y)\lambda_1(dy)$ as well as the conditional densities $f_{X|Y=y}(x) = f_{X,Y}(x, y)/f_Y(y)$ and $f_{Y|X=x}(y) = f_{X,Y}(x, y)/f_X(x)$. Observe incidentally that the probabilistic framework of the continuous ranking problem is quite similar to that of *distribution-free regression*. However, as shall be seen in the subsequent analysis, even if the regression function $m(x) = \mathbb{E}[Y \mid X = x]$ can be optimal under appropriate conditions, just like for regression, measuring ranking performance involves criteria that are of different nature than the expected least square error and plug-in rules may not be relevant for the goal pursued here, as depicted by Fig. I.3 in the introductory chapter.

Scoring Functions. The set of all scoring functions is denoted by $\mathcal{S}$ here. Any scoring function $s \in \mathcal{S}$ defines a total preorder on the space $\mathcal{X}$: $\forall (x, x') \in \mathcal{X}^2$, $x \preceq_s x' \Leftrightarrow s(x) \leq s(x')$. We also set $x \prec_s x'$ when $s(x) < s(x')$ and $x =_s x'$ when $s(x) = s(x')$ for $(x, x') \in \mathcal{X}^2$.

2.2 Bi/Multi-partite Ranking

Suppose that Z is a binary label, taking its values in $\{-1, +1\}$ say, assigned to the r.v. X. In bipartite ranking, the goal is to pick s in $\mathcal{S}$ so that the larger $s(X)$, the greater

the probability that Y is equal to 1 ideally. In other words, the objective is to learn $s(x)$ such that the r.v. $s(X)$ given $Y = +1$ is as *stochastically larger*[1] as possible than the r.v. $s(X)$ given $Y = -1$: the difference between $\bar{G}_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = +1\}$ and $\bar{H}_s(t) = \mathbb{P}\{s(X) \geq t \mid Y = -1\}$ should be thus maximal for all $t \in \mathbb{R}$. This can be naturally quantified by means of the notion of ROC curve of a candidate $s \in \mathcal{S}$, *i.e.* the parametrized curve $t \in \mathbb{R} \mapsto (\bar{H}_s(t), \bar{G}_s(t))$, which can be viewed as the graph of a mapping $\mathrm{ROC}_s : \alpha \in (0,1) \mapsto \mathrm{ROC}_s(\alpha)$, connecting possible discontinuity points by linear segments (so that $\mathrm{ROC}_s(\alpha) = \bar{G}_s \circ (1 - H_s^{-1})(1-\alpha)$ when H_s has no flat part in $H_s^{-1}(1-\alpha)$, where $H_s = 1 - \bar{H}_s$). A basic Neyman Pearson's theory argument shows that the optimal elements $s^*(x)$ related to this natural (functional) bipartite ranking criterion (*i.e.* scoring functions whose ROC curve dominates any other ROC curve everywhere on $(0,1)$) are transforms $(T \circ \eta)(x)$ of the posterior probability $\eta(x) = \mathbb{P}\{Z = +1 \mid X = x\}$, where $T : \mathrm{SUPP}(\eta(X)) \to \mathbb{R}$ is any strictly increasing borelian mapping. Optimization of the curve in sup norm has been considered in [CV09b] or in [CV10] for instance. However, given its functional nature, in practice the ROC curve of any $s \in \mathcal{S}$ is often summarized by the area under it, which performance measure can be interpreted in a probabilistic manner, as the theoretical rate of *concording pairs*

$$\mathrm{AUC}(s) = \mathbb{P}\left\{s(X) < s(X') \mid Z = -1,\ Z' = +1\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X') \mid Z = -1,\ Z' = +1\right\}, \tag{III.1}$$

where (X', Z') denoted an independent copy of (X, Z). A variety of algorithms aiming at maximizing the AUC criterion or surrogate pairwise criteria have been proposed and studied in the literature, among which [FISS03], [Rak04] or [CDV13a], whereas generalization properties of empirical AUC maximizers have been studied in [CLV08], [AGH$^+$05] and [MW16]. An analysis of the relationship between the AUC and the error rate is given in [CM04].

Extension to the situation where the label Y takes at least three ordinal values (*i.e.* multipartite ranking) has been also investigated, see *e.g.* [RA05] or [CR14]. In [SCV13], it is shown that, in contrast to the bipartite setup, the existence of optimal solutions cannot be guaranteed in general and conditions on (X, Y)'s distribution ensuring that optimal solutions do exist and that extensions of bipartite ranking criteria such as the ROC manifold and the volume under it can be used for learning optimal scoring rules have been exhibited. An analogous analysis in the context of continuous ranking is carried out in the next section.

3 Optimal Elements in Ranking Data with Continuous Labels

In this section, a natural definition of the set of optimal elements for continuous ranking is first proposed. Existence and characterization of such optimal scoring functions are next discussed.

[1] Given two real-valued r.v.'s U and U', recall that U is said to be *stochastically larger* than U' when $\mathbb{P}\{U \geq t\} \geq \mathbb{P}\{U' \geq t\}$ for all $t \in \mathbb{R}$.

3.1 Optimal Scoring Rules for Continuous Ranking

Considering a threshold value $y \in [0,1]$, a considerably weakened (and discretized) version of the problem stated informally above would consist in finding s so that the r.v. $s(X)$ given $Y > y$ is as stochastically larger than $s(X)$ given $Y < y$ as possible. This *sub-problem* coincides with the *bipartite ranking* problem related to the pair (X, Z_y), where $Z_y = 2\mathbb{I}\{Y > y\} - 1$. As briefly recalled in subsection 2.2, the optimal set $\mathcal{S}_y^*$ is composed of the scoring functions that induce the same ordering as

$$\eta_y(X) = \mathbb{P}\{Y > y \mid X\} = 1 - (1 - p_y)/(1 - p_y + p_y \Phi_y(X)),$$

where $p_y = 1 - F_Y(y) = \mathbb{P}\{Y > y\}$ and $\Phi_y(X) = (dF_{X|Y>y}/dF_{X|Y<y})(X)$.

A Continuum of Bipartite Ranking Problems. The rationale behind the definition of the set $\mathcal{S}^*$ of optimal scoring rules for continuous ranking is that any element s^* should score observations x in the same order as η_y (or equivalently as Φ_y).

Definition 1. (Optimal scoring rule) *An optimal scoring rule for the continuous ranking problem related to the random pair (X, Y) is any element s^* that fulfills: $\forall y \in (0,1)$,*

$$\forall (x, x') \in \mathcal{X}^2, \;\; \eta_y(x) < \eta_y(x') \Rightarrow s^*(x) < s^*(x'). \tag{III.2}$$

In other words, the set of optimal rules is defined as $\mathcal{S}^* = \bigcap_{y \in (0,1)} \mathcal{S}_y^*$.

It is noteworthy that, although the definition above is natural, the set $\mathcal{S}^*$ can be empty in absence of any distributional assumption, as shown by the following example.

Example 1. *As a counter-example, consider the distributions $F_{X,Y}$ such that $F_Y = \mathcal{U}([0,1])$ and $F_{X|Y=y} = \mathcal{N}(|2y-1|, (2y-1)^2)$. Observe that $(X, 1-Y) \overset{d}{=} (X, Y)$, so that $\Phi_{1-t} = \Phi_t^{-1}$ for all $t \in (0,1)$ and there exists $t \neq 0$ s.t. Φ_t is not constant. Hence, there exists no s^* in $\mathcal{S}$ such that* (III.2) *holds true for all $t \in (0,1)$.*

Remark 1. (Invariance) *We point out that the class $\mathcal{S}^*$ of optimal elements for continuous ranking thus defined is invariant by strictly increasing transform of the 'size' variable Y (in particular, a change of unit has no impact on the definition of $\mathcal{S}^*$): for any borelian and strictly increasing mapping $H : (0,1) \to (0,1)$, any scoring function $s^*(x)$ that is optimal for the continuous ranking problem related to the pair (X, Y) is still optimal for that related to $(X, H(Y))$ (since, under these hypotheses, for any $y \in (0,1)$: $Y > y \Leftrightarrow H(Y) > H(y)$).*

3.2 Existence and Characterization of Optimal Scoring Rules

We now investigate conditions guaranteeing the existence of optimal scoring functions for the continuous ranking problem.

Proposition 1. *The following assertions are equivalent.*

1. *For all $0 < y < y' < 1$, for all $(x, x') \in \mathcal{X}^2$: $\Phi_y(x) < \Phi_y(x') \Rightarrow \Phi_{y'}(x) \leq \Phi_{y'}(x')$.*

2. *There exists an optimal scoring rule* s^* *(i.e.* $\mathcal{S}^* \neq \emptyset$*).*

3. *The regression function* $m(x) = \mathbb{E}[Y \mid X = x]$ *is an optimal scoring rule.*

4. *The collection of probability distributions* $F_{X|Y=y}(dx) = f_{X|Y=y}(x)\lambda_d(dx)$*,* $y \in (0,1)$ *satisfies the monotone likelihood ratio property: there exist* $s^* \in \mathcal{S}$ *and, for all* $0 < y < y' < 1$*, an increasing function* $\varphi_{y,y'} : \mathbb{R} \to \mathbb{R}_+$ *such that:* $\forall x \in \mathbb{R}^d$*,*

$$\frac{f_{X|Y=y'}}{f_{X|Y=y}}(x) = \varphi_{y,y'}(s^*(x)).$$

Refer to section 8 for the technical proof. Truth should be said, assessing that Assertion 1. is a very challenging statistical task. However, through important examples, we now describe (not uncommon) situations where the conditions stated in Proposition 1 are fulfilled.

Example 2. *We give a few important examples of probabilistic models fulfilling the properties listed in Proposition 1.*

• ***Regression Model.*** *Suppose that* $Y = m(X) + \epsilon$*, where* $m : \mathcal{X} \to \mathbb{R}$ *is a borelian function and* ϵ *is a centered r.v. independent from* X*. One may easily check that* $m \in \mathcal{S}^*$*.*

• ***Exponential Families.*** *Suppose that* $f_{X|Y=y}(x) = \exp(\kappa(y)T(x) - \psi(y))f(x)$ *for all* $x \in \mathbb{R}^d$*, where* $f : \mathbb{R}^d \to \mathbb{R}_+$ *is borelian,* $\kappa : [0,1] \to \mathbb{R}$ *is a borelian strictly increasing function and* $T : \mathbb{R}^d \to \mathbb{R}$ *is a borelian mapping such that*

$$\psi(y) = \log \int_{x \in \mathbb{R}^d} \exp(\kappa(y)T(x))f(x)dx < +\infty.$$

We point out that, although the regression function $m(x)$ is an optimal scoring function when $\mathcal{S}^* \neq \emptyset$, the *continuous ranking* problem does not coincide with *distribution-free regression* (notice incidentally that, in this case, any strictly increasing transform of $m(x)$ belongs to $\mathcal{S}^*$ as well). As depicted by Fig. I.3 the least-squares criterion is not relevant to evaluate continuous ranking performance and naive plug-in strategies should be avoided, see Remark 2 below. Dedicated performance criteria are proposed in the next section.

4 Performance Measures for Continuous Ranking

We now investigate quantitative criteria for assessing the performance in the continuous ranking problem, which practical machine-learning algorithms may rely on. We place ourselves in the situation where the set $\mathcal{S}^*$ is not empty, see Proposition 1 above.

A Functional Performance Measure. It follows from the view developped in the previous section that, for any $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$ and for all $y \in (0,1)$, we have:

$$\forall \alpha \in (0,1), \quad \mathrm{ROC}_{s,y}(\alpha) \leq \mathrm{ROC}_{s^*,y}(\alpha) = \mathrm{ROC}^*_y(\alpha), \tag{III.3}$$

denoting by $\mathrm{ROC}_{s,y}$ the ROC curve of any $s \in \mathcal{S}$ related to the bipartite ranking subproblem (X, Z_y) and by ROC^*_y the corresponding optimal ROC curve, *i.e.* the ROC curve of strictly increasing transforms of $\eta_y(x)$. Based on this observation, it is natural to design a dedicated performance measure by aggregating these 'sub-criteria'. Integrating over y w.r.t. a σ-finite measure μ with support equal to $[0, 1]$, this leads to the following definition $\mathrm{IROC}_{\mu,s}(\alpha) = \int \mathrm{ROC}_{s,y}(\alpha)\mu(dy)$. The functional criterion thus defined inherits properties from the $\mathrm{ROC}_{s,y}$'s (*e.g.* monotonicity, concavity). In addition, the curve IROC_{μ,s^*} with $s^* \in \mathcal{S}^*$ dominates everywhere on $(0, 1)$ any other curve $\mathrm{IROC}_{\mu,s}$ for $s \in \mathcal{S}$. However, except in pathologic situations (*e.g.* when $s(x)$ is constant), the curve $\mathrm{IROC}_{\mu,s}$ is not invariant when replacing Y's distribution by that of a strictly increasing transform $H(Y)$. In order to guarantee that this desirable property is fulfilled (see Remark 1), one should integrate w.r.t. Y's distribution (which boils down to replacing Y by the uniformly distributed r.v. $F_Y(Y)$).

Definition 2. (Integrated ROC/AUC criteria) *The integrated ROC curve of any scoring rule $s \in \mathcal{S}$ is defined as: $\forall \alpha \in (0, 1)$,*

$$IROC_s(\alpha) = \int_{y=0}^{1} ROC_{s,y}(\alpha) F_Y(dy) = \mathbb{E}\left[ROC_{s,Y}(\alpha)\right]. \tag{III.4}$$

The integrated AUC criterion is defined as the area under the integrated ROC curve: $\forall s \in \mathcal{S}$,

$$IAUC(s) = \int_{\alpha=0}^{1} IROC_s(\alpha) d\alpha. \tag{III.5}$$

Additional properties of IROC curves are listed below.

4.1 Properties of IROC Curves

For any scoring function $s \in \mathcal{S}$ and $y \in (0, 1)$, we define the conditional cdfs of $s(X)$ as follows:

$$H_{s,y}(v) = \mathbb{P}(s(X) \leq v \mid Y < y),$$

$$G_{s,y}(v) = \mathbb{P}(s(X) \leq v \mid Y > y).$$

Now we give some properties of the IROC curve which are easily derived from ROC curve properties by integration over bipartite ranking subproblems.

Theorem 1. *For any scoring function $s \in \mathcal{S}$, the following properties hold:*

- ***Limit values.*** *We have $IROC_s(0) = 0$ and $IROC_s(1) = 1$.*
- ***Invariance.*** *For any strictly increasing function $T : \mathbb{R} \to \mathbb{R}$, we have for all $\alpha \in (0, 1)$, $IROC_{T \circ s}(\alpha) = IROC_s(\alpha)$.*
- ***Concavity.*** *If for all $y \in (0, 1)$ the likelihood ratio $dG_{s,y}/dH_{s,y}$ is a monotone function, then the IROC curve is concave.*

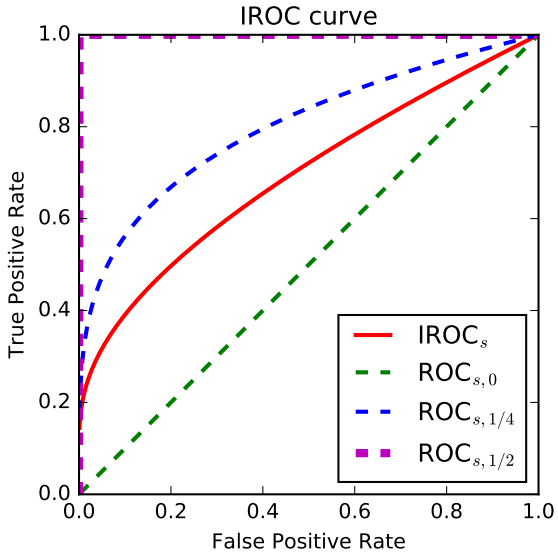


Figure III.1: IROC curve $\mathrm{IROC}_s(\alpha) = \frac{1-\alpha}{\log \frac{1}{\alpha}}$ for score s and Y's distribution such that for all $0 \leq y \leq \frac{1}{2}$, the ROC curve of s at sublevels y and $1-y$ is $\mathrm{ROC}_{s,y}(\alpha) = \mathrm{ROC}_{s,1-y}(\alpha) = \alpha^{1-4y(1-y)}$ and $f_Y(y) = f_Y(1-y) = 2(1-2y)$.

PROOF. Use Proposition 24 in [CV09b] for each bipartite ranking subproblem at level $y \in (0,1)$. Then integrate over y w.r.t. F_Y.

The following result reveals the relevance of the functional/summary criteria defined above for the continuous ranking problem.

Theorem 2. *Let $s^* \in \mathcal{S}$. The following assertions are equivalent.*

1. *The assertions of Proposition 1 are fulfilled and s^* is an optimal scoring function in the sense given by Definition 1.*
2. *For all $\alpha \in (0,1)$, $IROC_{s^*}(\alpha) = \mathbb{E}\left[ROC^*_Y(\alpha)\right]$.*
3. *We have $IAUC_{s^*} = \mathbb{E}\left[AUC^*_Y\right]$, where $AUC^*_y = \int_{\alpha=0}^{1} ROC^*_y(\alpha)d\alpha$ for all $y \in (0,1)$.*

If $\mathcal{S}^* \neq \emptyset$, then we have: $\forall s \in \mathcal{S}$,

$$\begin{aligned} IROC_s(\alpha) &\leq IROC^*(\alpha) \overset{def}{=} \mathbb{E}\left[ROC^*_Y(\alpha)\right], \quad \text{for any } \alpha \in (0,1), \\ IAUC(s) &\leq IAUC^* \overset{def}{=} \mathbb{E}\left[AUC^*_Y\right]. \end{aligned}$$

In addition, for any borelian and strictly increasing mapping $H : (0,1) \to (0,1)$, replacing Y by $H(Y)$ leaves the curves $IROC_s$, $s \in \mathcal{S}$, unchanged.

Equipped with the notion defined above, a scoring rule s_1 is said to be more accurate than another one s_2 if $\mathrm{IROC}_{s_2}(\alpha) \leq \mathrm{IROC}_{s_1}(\alpha)$ for all $\alpha \in (0,1)$.The IROC curve criterion thus provides a partial preorder on $\mathcal{S}$. Observe also that, by virtue of Fubini's theorem, we have $\mathrm{IAUC}(s) = \int \mathrm{AUC}_y(s) F_Y(dy)$ for all $s \in \mathcal{S}$, denoting by $\mathrm{AUC}_y(s)$ the AUC of s related to the bipartite ranking subproblem (X, Z_y). Just like the AUC for bipartite ranking, the scalar IAUC criterion defines a full preorder on $\mathcal{S}$ for continuous ranking. Based on a training dataset $\mathcal{D}_n$ of independent copies of (X, Y), statistical versions of the IROC/IAUC criteria can be straightforwardly computed by replacing the distributions F_Y, $F_{X|Y>t}$ and $F_{X|Y<t}$ by their empirical counterparts in (III.3)-(III.5). The lemma below provides a probabilistic interpretation of the IAUC criterion.

Lemma 1. *Let (X', Y') be a copy of the random pair (X, Y) and Y'' a copy of the r.v. Y. Suppose that (X, Y), (X', Y') and Y'' are defined on the same probability space and are independent. Denote by $\mu(dy) = 6\mathbb{P}\{Y \leq y\}\mathbb{P}\{Y > y\}F_Y(dy)$ the probability measure used to integrate the ROC curves. For all $s \in \mathcal{S}$, we have:*

$$\begin{aligned} IAUC(s) &:= \int_{\alpha=0}^{1} IROC_{\mu,s}(\alpha) d\alpha \\ &= \mathbb{P}\left\{s(X) < s(X') \mid Y < Y'' < Y'\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X') \mid Y < Y'' < Y'\right\}. \quad \text{(III.6)} \end{aligned}$$

This result shows in particular that a natural statistical estimate of $\mathrm{IAUC}(s)$ based on $\mathcal{D}_n$ involves U-statistics of degree 3. Its proof is given in section 8.

Remark 2. (Connection to distribution-free regression) *Consider the non-parametric regression model $Y = m(X) + \epsilon$, where ϵ is a centered r.v. independent from X. In this case, it is well-known that the regression function $m(X) = \mathbb{E}[Y \mid X]$ is the (unique) solution of the expected least squares minimization. However, although $m \in \mathcal{S}^*$, the least squares criterion is far from appropriate to evaluate ranking performance, as depicted by Fig. I.3. Observe additionally that, in contrast to the criteria introduced above, increasing transformation of the output variable Y may have a strong impact on the least squares minimizer: except for linear stransforms, $\mathbb{E}[H(Y) \mid X]$ is not an increasing transform of $m(X)$.*

Remark 3. (On discretization) *Bi/multi-partite algorithms are not directly applicable to the continuous ranking problem. Indeed a discretization of the interval [0, 1] would be first required but this would raise a difficult question outside our scope: how to choose this discretization based on the training data? We believe that this approach is less efficient than ours which reveals problem-specific criteria, namely IROC and IAUC.*

Before describing a practical algorithm for recursive maximization of the IROC curve, we discuss the Kendall τ criterion.

4.2 The Kendall τ Statistic.

The quantity (III.6) is akin to another popular way to measure the tendency to define the same ordering on the statistical population in a summary fashion:

$$\begin{aligned} d_\tau(s) &\stackrel{def}{=} \mathbb{P}\left\{\big(s(X)-s(X')\big)\cdot\big(Y-Y'\big)>0\right\}+\frac{1}{2}\mathbb{P}\left\{s(X)=s(X')\right\} \quad \text{(III.7)}\\ &= \mathbb{P}\{s(X)<s(X')\mid Y<Y'\}+\frac{1}{2}\mathbb{P}\left\{X=_s X'\right\}, \end{aligned}$$

where (X', Y') denotes an independent copy of (X, Y), observing that $\mathbb{P}\{Y < Y'\} = 1/2$. The empirical counterpart of (III.7) based on the sample $\mathcal{D}_n$, given by

$$\widehat{d}_n(s)=\frac{2}{n(n-1)}\sum_{i<j}\mathbb{I}\left\{(s(X_i)-s(X_j))\cdot(Y_i-Y_j)>0\right\}+\frac{1}{n(n-1)}\sum_{i<j}\mathbb{I}\left\{s(X_i)=s(X_j)\right\} \quad \text{(III.8)}$$

is known as the *Kendall τ statistic* and is widely used in the context of statistical hypothesis testing. The quantity (III.7) shall be thus referred to as the (theoretical or true) *Kendall τ*. Notice that $d_\tau(s)$ is invariant by strictly increasing transformation of $s(x)$ and thus describes properties of the order it defines. The following result reveals that the class $\mathcal{S}^*$, when non empty, is the set of maximizers of the theoretical Kendall τ. Refer to section 8 for the technical proof.

Proposition 2. *Suppose that $\mathcal{S}^* \neq \emptyset$. For any $(s, s^*) \in \mathcal{S} \times \mathcal{S}^*$, we have: $d_\tau(s) \leq d_\tau(s^*)$.*

Equipped with these criteria, the objective expressed above in an informal manner can be now formulated in a quantitative manner as a (possibly functional) M-estimation problem. In practice, the goal pursued is to find a reasonable approximation of a solution to the optimization problem $\max_{s\in\mathcal{S}} d_\tau(s)$ (respectively $\max_{s\in\mathcal{S}} \mathrm{IAUC}(s)$), where the supremum is taken over the set of all scoring functions $s : \mathcal{X} \to \mathbb{R}$. Of course, these criteria are unknown in general, just like (X, Y)'s probability distribution, and the empirical risk minimization (ERM in abbreviated form) paradigm (see [DGL96]) invites for maximizing the statistical version (III.8) over a class $\mathcal{S}_0 \subset \mathcal{S}$ of controlled complexity when considering the criterion $d_\tau(s)$ for instance. The generalization capacity of empirical maximizers of the Kendall τ can be straightforwardly established using results in [CLV08].

On Empirical Kendall τ Maximization Here we state a result describing the performance of scoring rules obtained through maximization of the empirical Kendall τ over a class $\mathcal{S}_0 \subset \mathcal{S}$ of controlled complexity. An empirical *Kendall τ* maximizer over $\mathcal{S}_0$ is any scoring function $\widehat{s}_n \in \mathcal{S}_0$ s.t.

$$\widehat{d}_n(\widehat{s}_n)=\max_{s\in\mathcal{S}_0}\widehat{d}_n(s). \quad \text{(III.9)}$$

Theorem 3. *Suppose that $\mathcal{S}^* \neq \emptyset$ and set $d_\tau^* = d_\tau(s^*)$ for $s^* \in \mathcal{S}^*$. Assume that $\mathcal{S}_0$ is a VC major class of functions with VC dimension $V < +\infty$. Let $\delta \in (0, 1)$. With probability at least $1 - \delta$, we have:*

$$d_\tau^*-d_\tau(\widehat{s}_n)\leq c\sqrt{\frac{V}{n}}+4\sqrt{\frac{\log(1/\delta)}{n-1}}+\left\{d_\tau^*-\max_{s\in\mathcal{S}_0}d_\tau(s)\right\}. \quad \text{(III.10)}$$

PROOF. The argument is based on the simple bound

$$d_\tau^* - d_\tau(\widehat{s}_n) \leq 2 \sup_{s \in \mathcal{S}_0} \left| \widehat{d}_n(s) - d_\tau(s) \right| + \left\{ d_\tau^* - \max_{s \in \mathcal{S}_0} d_\tau(s) \right\},$$

combined with the use of concentration results for the U-process $\{\widehat{d}_n(s) - d_\tau(s)\}_{s \in \mathcal{S}_0}$. The proof is finished by mimicking that of Corollary 3 in [CLV08].

From a computational perspective, maximizing $\widehat{d}_n$ is a challenge, the optimization problem being NP-hard due to the absence of convexity/smoothness of the pairwise loss function $\mathbb{I}\{(s(x) - s(x'))(y - y') > 0\}$. Whereas replacing this loss by a surrogate loss, more suited to continuous optimization, is a possible strategy, using greedy algorithms in the spirit of the popular CART method can also be considered for this purpose. A slight modification of CART based on recursive maximization of the empirical Kendall τ criterion (rather than the Gini index or the least squares criterion) permit to build an *oriented ranking tree* in a top down manner, see subsection 5.1. Just like for classification/regression, the procedure can be followed by a pruning stage (model selection), based here on (*e.g.* cross-validation based) estimates of Kendall τ.

Remark 4. (ON KENDALL τ AND AUC) *We point out that, in the bipartite ranking problem (i.e. when the output variable Z takes its values in $\{-1, +1\}$, see subsection 2.2) as well, the AUC criterion can be expressed as a function of the Kendall τ related to the pair $(s(X), Z)$ when the r.v. $s(X)$ is continuous. Indeed, we have in this case $2p(1-p)AUC(s) = d_\tau(s)$, where $p = \mathbb{P}\{Z = +1\}$ and $d_\tau(s) = \mathbb{P}\{(s(X)-s(X'))\cdot(Z-Z') > 0\}$, denoting by (X', Z') an independent copy of (X, Z).*

5 Continuous Ranking through Oriented Recursive Partitioning

It is the purpose of this section to introduce the algorithm CRANK, a specific tree-structured learning algorithm for continuous ranking.

5.1 Ranking Trees and Oriented Recursive Partitions

Decision trees undeniably figure among the most popular techniques, in supervised and unsupervised settings, refer to [BFOS84] or [Qui86] for instance. This is essentially due to the visual model summary they provide, in the form of a binary tree graphic that permits to describe predictions by means of a hierachichal combination of elementary rules of the type '$X^{(j)} \leq \kappa$' or '$X^{(j)} > \kappa$', comparing the value taken by a (quantitative) component of the input vector X (the *split variable*) to a certain threshold (the *split value*). In contrast to local learning problems such as classification or regression, predictive rules for a global problem such as *ranking* cannot be described by a (tree-structured) partition of the feature space: cells (corresponding to the terminal leaves of the binary decision tree) must be ordered so as to define a scoring function. This leads to the definition of

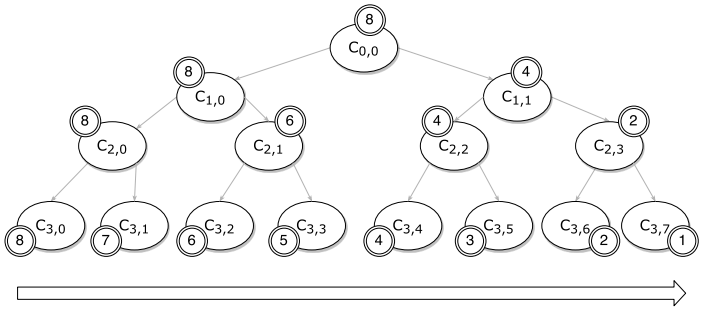


Figure III.2: A scoring function described by an oriented binary subtree $\mathcal{T}$. For any element $x \in \mathcal{X}$, one may compute the quantity $s_{\mathcal{T}}(x)$ very fast in a top-down fashion by means of the heap structure: starting from the initial value 2^J at the root node, at each internal node $\mathcal{C}_{j,k}$, the score remains unchanged if x moves down to the left sibling, whereas one subtracts $2^{J-(j+1)}$ from it if x moves down to the right.

ranking trees as binary trees equipped with a 'left-to-right' orientation, defining a tree-structured collection of anomaly scoring functions, as depicted by Fig. III.2. Binary ranking trees have been in the context of bipartite ranking in [CV09b] or in [CDV13a] and in [SCV13] in the context of multipartite ranking. The root node of a ranking tree $\mathcal{T}_J$ of depth $J \geq 0$ represents the whole feature space $\mathcal{X}$: $\mathcal{C}_{0,0} = \mathcal{X}$, while each internal node (j,k) with $j < J$ and $k \in \{0, \ldots, 2^j - 1\}$ corresponds to a subset $\mathcal{C}_{j,k} \subset \mathcal{X}$, whose left and right siblings respectively correspond to disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. Equipped with the left-to-right orientation, any subtree $\mathcal{T} \subset \mathcal{T}_J$ defines a preorder on $\mathcal{X}$: elements lying in the same terminal cell of $\mathcal{T}$ being equally ranked. The scoring function related to the oriented tree $\mathcal{T}$ can be written as:

$$s_{\mathcal{T}}(x) = \sum_{\mathcal{C}_{j,k} \text{: terminal leaf of } \mathcal{T}} 2^J \left(1 - \frac{k}{2^j}\right) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,k}\}. \tag{III.11}$$

5.2 The CRank Algorithm

Based on Proposition 2, one can try to build from the training dataset $\mathcal{D}_n$ a ranking tree by recursive empirical Kendall τ maximization. We propose below an alternative tree-structured recursive algorithm, relying on a (dyadic) discretization of the 'size' variable Y. At each iteration, the local sample (*i.e.* the data lying in the cell described by the current node) is split into two halves (the highest/smallest halves, depending on Y) and the algorithm calls a binary classification algorithm $\mathcal{A}$ to learn how to divide the node into right/left children. The theoretical analysis of this algorithm and its connection with approximation of IROC* are difficult questions left as open problems. Indeed we found out that the IROC cannot be represented as a parametric curve contrary to the ROC, which renders proofs much more difficult than in the bipartite case.

THE CRANK ALGORITHM

1. **Input.** Training data $\mathcal{D}_n$, depth $J \geq 1$, binary classification algorithm $\mathcal{A}$.
2. **Initialization.** Set $\mathcal{C}_{0,0} = \mathcal{X}$.
3. **Iterations.** For $j = 0, \ldots, J-1$ and $k = 0, \ldots, 2^J - 1$,
   a) Compute a median $y_{j,k}$ of the dataset $\{Y_1, \ldots, , Y_n\} \cap \mathcal{C}_{j,k}$ and assign the binary label $Z_i = 2\mathbb{I}\{Y_i > y_{j,k}\} - 1$ to any data point i lying in $\mathcal{C}_{j,k}$, *i.e.* such that $X_i \in \mathcal{C}_{j,k}$.
   b) Solve the binary classification problem related to the input space $\mathcal{C}_{j,k}$ and the training set $\{(X_i, Y_i) : \ 1 \leq i \leq n, \ X_i \in \mathcal{C}_{j,k}\}$, producing a classifier $g_{j,k} : \mathcal{C}_{j,k} \to \{-1, +1\}$.
   c) Set $\mathcal{C}_{j+1,2k} = \{x \in \mathcal{C}_{j,k}, \ g_{j,k}(x) = +1\} = \mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k+1}$.
4. **Output.** Ranking tree $\mathcal{T}_J = \{\mathcal{C}_{j,k} : \ 0 \leq j \leq J, \ 0 \leq k < D\}$.

Of course, the depth J should be chosen such that $2^J \leq n$. One may also consider continuing to split the nodes until the number of data points within a cell has reached a minimum specified in advance. In addition, it is well known that recursive partitioning methods fragment the data and the unstability of splits increases with the depth. For this reason, a ranking subtree must be selected. The growing procedure above should be classically followed by a pruning stage, where children of a same parent are progressively merged until the root $\mathcal{T}_0$ is reached and a subtree among the sequence $\mathcal{T}_0 \subset \ldots \subset \mathcal{T}_J$ with nearly maximal IAUC should be chosen using cross-validation. Issues related to the implementation of the CRANK algorithm and variants (*e.g.* exploiting randomization/aggregation) are beyond the scope of the present chapter.

6 Numerical Experiments

In order to illustrate the idea conveyed by Fig. I.3 that the least squares criterion is not appropriate for the continuous ranking problem we compared on a toy example CRANK with CART. Recall that the latter is a regression decision tree algorithm which minimizes the MSE (Mean Squared Error). We also runned an alternative version of CRANK which maximizes the empirical Kendall τ instead of the empirical IAUC: this method is refered to as KENDALL from now on.

Experimental Setting. The experimental setting is composed of a unidimensional feature space $\mathcal{X} = [0, 1]$ (for visualization reasons) and a simple regression model without any noise: $Y = m(X)$. Intuitively, a least squares strategy can miss slight oscillations of the regression function, which are critical in ranking when they occur in high probability regions as they affect the order among the feature space. We considered a polynomial

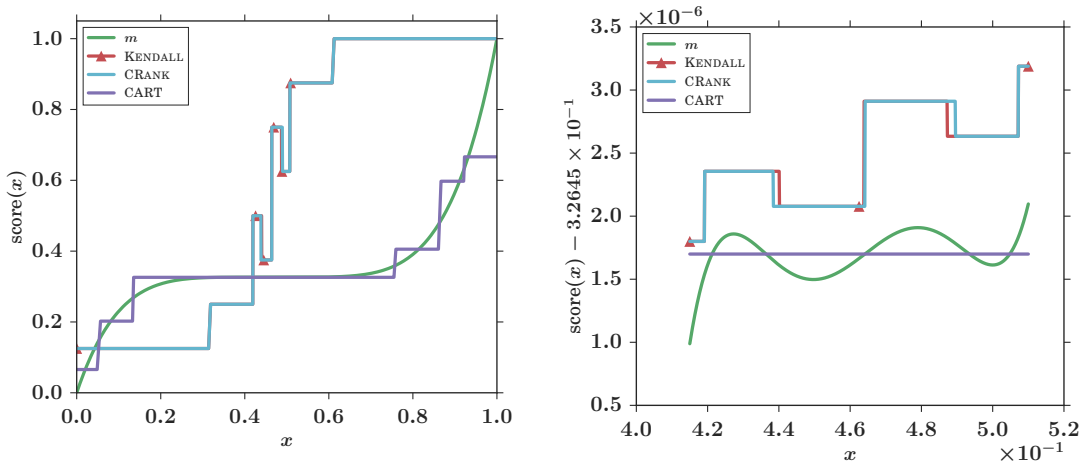
regression function m over $[0,1]$ and valued in $[0,1]$, namely:

$$m(x) = \frac{P(x) - P(0)}{P(1) - P(0)},$$

where the polynomial function P is given by:

$$P(x) = z^2 \cdot (z+1) \cdot (z+1.5) \cdot (z+2), \quad \text{with} \quad z = 25 \cdot (x - 0.5).$$

Observe that m slightly oscillates in the interval $I_2 = [0.415, 0.51]$ (see III.3b). With respective probabilities $p_1 = 0.1$, $p_2 = 0.8$ and $p_3 = 0.1$, X is uniformly sampled in one of the three intervals $I_1 = [0, 0.415]$, I_2 and $I_3 = [0.51, 1]$: the critical window I_2 is then a high probability region. The three algorithms (CRank, Kendall and CART) where trained on the same dataset $(X_1, Y_1), \ldots, (X_{n_{\text{train}}}, Y_{n_{\text{train}}})$ with $Y_i = m(X_i)$ and $n_{\text{train}} = 100$ with the same constraint on the depth of the tree: at most $D = 3$. Then we tested them on $n_{\text{test}} = 2000$ new iid copies of X. In Fig. III.3 we plot the polynomial function m and piecewise constant scoring functions provided by the three approaches. We observe in Fig. III.3 that CRank and Kendall almost provide the same ranking functions ($s_{\text{CRank}} \approx s_{\text{Kendall}}$) and achieve similar performance (see Fig. III.1). Also notice in Fig. III.1 that CRank, Kendall and CART respectively achieve maximum IAUC, Kendall τ and MSE. As expected, CART misses the critical oscillations that is why its IAUC and Kendall τ are considerably lower than for its concurrents.



(a) (b)

Figure III.3: Polynomial regression function m and scoring functions provided by CRank, Kendall and CART. For visualization reasons, s_{CRank} and s_{Kendall} have been renormalized by $2^D = 8$ to take values in $[0,1]$ and, in Fig. III.3b, affine functions have been applied to the three scoring functions.

Results. The results are presented in Table III.1. As expected, the CRank and Kendall continuous ranking methods both ouperform the CART regression approach.

	IAUC	Kendall τ	MSE
CRank	**0.95**	0.92	0.10
Kendall	0.94	**0.93**	0.10
CART	0.61	0.58	$\mathbf{7.4 \times 10^{-4}}$

Table III.1: IAUC, Kendall τ and MSE empirical measures

7 Conclusion

This chapter considers the problem of learning how to order objects by increasing 'size', modeled as a continuous r.v. Y, based on *indirect measurements* X. We provided a rigorous mathematical formulation of this problem that finds many applications (*e.g.* quality control, chemistry) and is referred to as *continuous ranking*. In particular, necessary and sufficient conditions on (X, Y)'s distribution for the existence of optimal solutions are exhibited and appropriate criteria have been proposed for evaluating the performance of scoring rules in these situations. In contrast to distribution-free regression where the goal is to recover the local values taken by the regression function, *continuous ranking* aims at reproducing the preorder it defines on the feature space as accurately as possible. The numerical results obtained via the algorithmic approaches we proposed for optimizing the criteria aforementioned highlight the difference in nature between these two statistical learning tasks.

8 Technical Proofs

We provide below the proofs of the theoretical results stated in the chapter.

Proof of Proposition 1

Observe first that 3. $\Rightarrow$ 2. and 1. $\Leftrightarrow$ 4. are obvious.
2. $\Rightarrow$ 1.: Let us assume that assertion 2. is true. Let $(x, x') \in \mathcal{X}^2$ and $y \in (0, 1)$ such that $\Phi_y(x) < \Phi_y(x')$. Then, from assumption 2., $s^*(x) < s^*(x')$. For $t' \in (y, 1)$, if $\Phi_{y'}(x) > \Phi_{y'}(x')$, it leads to the following contradiction: $s^*(x) > s^*(x')$. Hence $\Phi_{y'}(x) \leq \Phi_{y'}(x')$.
1. $\Rightarrow$ 3.: Let us assume that assertion 1. is true. Let $(x, x') \in \mathcal{X}^2$ and $y \in (0, 1)$ such that $\eta_y(x) < \eta_y(x')$. Observe that $(x, y') \mapsto \eta_{y'}(x)$ is continuous. It follows from assumption 1. that for $y' \in (0, 1)$, $\eta_{y'}(x) \leq \eta_{y'}(x')$ with strict inequality on a nonempty interval by continuity of $(x, y') \mapsto \eta_{y'}(x)$. Integrating the latter inequality against the uniform distribution over $(0, 1)$ leads to $m(x) < m(x')$.

Proof of Theorem 2

The implications 1. $\Rightarrow$ 2. and 2. $\Rightarrow$ 3. are obvious.
3. $\Rightarrow$ 1.: Let us assume that assertion 3. is true. Assume ad absurdum that 1. is false. Then there exists $y \in (0,1)$ s.t. $\mathrm{AUC}_y(s^*) < \mathrm{AUC}_y(\eta_y)$. Notice that $(x, y') \mapsto \eta_{y'}(x)$ and, for any scoring function s, $y' \mapsto \mathrm{AUC}_{y'}(s)$ are continuous. By integration w.r.t. F_Y we obtain $\mathrm{IAUC}(s^*) < \mathbb{E}\left[\mathrm{AUC}_Y^*\right]$, which contradicts assertion 3. Hence 1. is true.

Proof of Lemma 1

Recall that, for any $s \in \mathcal{S}$ and all $y \in (0,1)$, we have:

$$\mathrm{AUC}_y(s) = \mathbb{P}\left\{s(X) < s(X') \mid Y < y < Y'\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X') \mid Y < y < Y'\right\}.$$

Integrating the terms in the equation above w.r.t. $\mu(dy)$ leads to the desired formula. Then, a natural empirical version of $\mathrm{IAUC}(s)$ is:

$$\begin{aligned}\widehat{\mathrm{IAUC}}_n(s) = &\frac{6}{n(n-1)(n-2)} \sum_{1\le i,j,k\le n} \mathbb{I}\left\{s(X_i) < s(X_k), Y_i < Y_j < Y_k\right\} \\ &+ \frac{3}{n(n-1)(n-2)} \sum_{1\le i,j,k\le n} \mathbb{I}\left\{s(X_i) = s(X_k), Y_i < Y_j < Y_k\right\}.\end{aligned}$$

Proof of Proposition 2

We assume that $s(X)$ is a continuous r.v. for simplicity, the slight modifications needed to extend the argument to the general framework being left to the reader. As a first go, observe that

$$d_\tau(s) \;=\; \mathbb{P}\{s(X') > s(X) \mid Y' > Y\} = \int_{y'=0}^{1} \mathbb{P}\left\{s(X') > s(X) \mid Y' = y',\ Y < y'\right\} F_Y(dy')$$

Notice next that, for any $y' \in (0,1)$, $\mathbb{P}\left\{s(X') > s(X) \mid Y' = y',\ Y < y'\right\}$ is nothing else than the AUC criterion of $s(x)$ related to the distribution of X given $Y < y'$ (negative distribution) and $F_{X|Y=y'}$ (positive distribution). Since we assumed $\mathcal{S}^* \neq \emptyset$, the collection $\{F_{X|Y=y} :\ y \in (0,1)\}$ is of increasing likelihood ratio and according to Theorem 1, any $s^* \in \mathcal{S}^*$ is a Neyman Pearson test statistic and thus defines uniformly most powerful tests (among unbiased tests) of $\mathcal{H}_0 : Y < y$ against $\mathcal{H}_1 : Y = y$. Hence, for any $y' \in (0,1)$, $\mathbb{P}\left\{s(X') > s(X) \mid Y' = y',\ Y < y'\right\} \le \mathbb{P}\left\{s^*(X') > s^*(X) \mid Y' = y',\ Y < y'\right\}$. Integrating over y' w.r.t. F_Y yields the desired result.

CHAPTER IV

Dimensionality Reduction and (Bucket) Ranking: a Mass Transportation Approach

Abstract

Whereas most dimensionality reduction techniques (*e.g.* PCA, ICA, NMF) for multivariate data essentially rely on linear algebra to a certain extent, summarizing ranking data, viewed as realizations of a random permutation Σ on a set of items indexed by $i \in \{1, \ldots, N\}$, is a great statistical challenge, due to the absence of vector space structure for the set of permutations $\mathfrak{S}_N$. It is the goal of this chapter to develop an original framework for possibly reducing the number of parameters required to describe the distribution of a statistical population composed of rankings/permutations, on the premise that the collection of items under study can be partitioned into subsets/buckets, such that, with high probability, items in a certain bucket are either all ranked higher or else all ranked lower than items in another bucket. In this context, Σ's distribution can be hopefully represented in a sparse manner by a *bucket distribution*, *i.e.* a bucket ordering plus the ranking distributions within each bucket. More precisely, we introduce a dedicated distortion measure, based on a mass transportation metric, in order to quantify the accuracy of such representations. The performance of buckets minimizing an empirical version of the distortion is investigated through a rate bound analysis. Complexity penalization techniques are also considered to select the shape of a bucket order with minimum expected distortion. Beyond theoretical concepts and results, numerical experiments on real ranking data are displayed in order to provide empirical evidence of the relevance of the approach promoted.

1 Introduction

Recommendation systems and search engines are becoming ubiquitous in modern technological tools. Operating continuously on still more content, use of such tools generate or take as input more and more data. The scientific challenge relies on the nature of the data feeding or being produced by such algorithms: input or/and output information

generally consists of rankings/orderings, expressing *preferences*. Because the number of possible rankings explodes with the number of instances, it is of crucial importance to elaborate dedicated dimensionality reduction methods in order to represent ranking data efficiently. Whatever the type of task considered (supervised, unsupervised), machine-learning algorithms generally rest upon the computation of statistical quantities such as averages or linear combinations of the observed features, representing efficiently the data. However, summarizing ranking variability is far from straightforward and extending simple concepts such as that of an average or median in the context of preference data raises a certain number of deep mathematical and computational problems. For instance, whereas it is always possible to define a barycentric permutation (*i.e.* a consensus ranking) given a set of rankings and a metric on the symmetric group, its computation can be very challenging, as evidenced by the increasing number of contributions devoted to the ranking aggregation problem in the machine-learning literature, see *e.g.* [DKNS01], [PS16], [JKSO16] or [JKS16] among others. Regarding dimensionality reduction, it is far from straightforward to adapt traditional techniques such as Principal Component Analysis and its numerous variants to the ranking setup, the main barrier being the absence of a vector space structure on the set of permutations. Even if one can embed permutations into the Birkhoff polytope (which is the convex hull of the set of permutation matrices, see [CJ10],[LMC$^+$17]), the coordinates of the embeddings are highly correlated, and a low-dimensional representation of the original distribution over rankings could not be interpreted in a straightforward manner. In this chapter, we develop a novel framework for representing the distribution of ranking data in a simple manner, that is shown to extend, remarkably, consensus ranking in some sense. The rationale behind the approach we promote is that, in many situations encountered in practice, the set of instances may be partitioned into subsets/buckets, such that, with high probability, objects belonging to a certain bucket are either all ranked higher or else all ranked lower than objects lying in another bucket. In such a case, the ranking distribution can be described in a sparse fashion by: 1) a gross ordering structure (related to the buckets) and 2) the marginal ranking distributions associated to each bucket. Precisely, optimal representations are defined here as those associated to a bucket order minimizing a certain distortion measure we introduce, the latter being based on a mass transportation metric on the set of ranking distributions. Noticeably, this distortion measure is shown to admit a very simple closed-form expression, based on the marginal pairwise probabilities solely, when the cost of the mass transportation metric considered is the Kendall's τ distance and can be thus straightforwardly estimated. In the Kendall's τ case, we also highlight the fact that distortion minimization over bucket orders, when buckets are singletons, reduces to Kemeny consensus ranking. We establish rate bounds describing the generalization capacity of bucket order representations obtained by minimizing an empirical version of the distortion over collections of bucket orders and address model selection issues related to the choice of the bucket order size/shape. Numerical results are also displayed, providing in particular strong empirical evidence of the relevance of the notion of sparsity considered, which the dimensionality reduction technique introduced is based on.

The chapter is organized as follows. In section 2, a few concepts and results per-

taining to (Kemeny) consensus ranking are briefly recalled and the extended framework we consider for dimensionality reduction in the ranking context is described at length. Statistical results guaranteeing that optimal representations of reduced dimension can be learnt from ranking observations are established in section 3, while numerical experiments are presented in section 5 for illustration purpose. Some concluding remarks are collected in section 6. Technical details are deferred to section 8.

2 Preliminaries - Background

It this section, we introduce the main concepts and definitions that shall be used in the subsequent analysis. The indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$, the Dirac mass at any point a by δ_a, the cardinality of any finite subset A by $\#A$. Here and throughout, a full ranking on a set of items indexed by $[\![N]\!] = \{1, \ldots, N\}$ is seen as the permutation $\sigma \in \mathfrak{S}_N$ that maps any item i to its rank $\sigma(i)$. For any non empty subset $\mathcal{I} \subset [\![N]\!]$, any ranking σ on $[\![N]\!]$ naturally defines a ranking on $\mathcal{I}$, denoted by $\Pi_{\mathcal{I}}(\sigma)$ (*i.e.* $\forall i \in \mathcal{I}$, $\Pi_{\mathcal{I}}(\sigma)(i) = 1 + \sum_{j \in \mathcal{I} \setminus \{i\}} \mathbb{I}\{\sigma(j) < \sigma(i)\}$). If Σ is a random permutation on $\mathfrak{S}_N$ with distribution P, the distribution of $\Pi_{\mathcal{I}}(\Sigma)$ will be referred to as the marginal of P related to the subset $\mathcal{I}$. In particular, for a pair of items $(i, j) \in [\![N]\!]$, the quantity $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$ for $\Sigma \sim P$ is referred to as the pairwise marginal of P and indicates the probability that item i is preferred to (ranked lower than) item j (so $p_{i,j} + p_{j,i} = 1$). A bucket order $\mathcal{C}$ (also referred to as a partial ranking in the literature) is a strict partial order defined by an ordered partition of $[\![N]\!]$, *i.e* a sequence $(\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of $K \geq 1$ pairwise disjoint non empty subsets (buckets) of $[\![N]\!]$ such that: (1) $\cup_{k=1}^{K} \mathcal{C}_k = [\![N]\!]$, (2) $\forall (i, j) \in [\![N]\!]^2$, we have: $i \prec_{\mathcal{C}} j$ (i is ranked lower than j in $\mathcal{C}$) iff $\exists k < l$ s.t. $(i, j) \in \mathcal{C}_k \times \mathcal{C}_l$. We write $i \sim_{\mathcal{C}} j$ to mean that i and j belong to the same bucket (and cannot be compared/ordered by means of $\mathcal{C}$). The items in $\mathcal{C}_1$ have thus the lowest ranks (i.e. they are the most preferred items), whereas those in $\mathcal{C}_K$ have the highest ranks. For any bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$, its number of buckets K is referred to as its *size*, while its *shape* is the vector $\lambda = (\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$, i.e the sequence of sizes of buckets in $\mathcal{C}$ (verifying $\sum_{k=1}^{K} \#\mathcal{C}_k = N$). Hence, any bucket order $\mathcal{C}$ of size N corresponds to a full ranking/permutation $\sigma \in \mathfrak{S}_N$, whereas the set of all items $[\![N]\!]$ is the unique bucket order of size 1.

2.1 Background on Consensus Ranking

Given a collection of $n \geq 1$ rankings $\sigma_1, \ldots, \sigma_n$, consensus ranking, also referred to as ranking aggregation, aims at finding a ranking $\sigma^* \in \mathfrak{S}_N$ that best summarizes it. A popular way of tackling this problem, the metric-based consensus approach, consists in solving:

$$\min_{\sigma \in \mathfrak{S}_N} \sum_{s=1}^{n} d(\sigma, \sigma_s), \tag{IV.1}$$

where $d(.\,, .)$ is a certain metric on $\mathfrak{S}_N$. As the set $\mathfrak{S}_N$ is of finite cardinality, though not necessarily unique, such a barycentric permutation, called *consensus/median ranking*,

always exists. In Kemeny ranking aggregation, the most widely documented version in the literature, one considers the number of pairwise disagreements as metric, namely the Kendall's τ distance, see [Kem59]:

$$\forall(\sigma,\sigma') \in \mathfrak{S}_N^2, \;\; d_\tau(\sigma,\sigma') = \sum_{i<j} \mathbb{I}\{(\sigma(i)-\sigma(j))(\sigma'(i)-\sigma'(j)) < 0\}. \tag{IV.2}$$

Remark 1. *Many other distances are considered in the literature (see e.g. Chapter 11 in [DD09]). In particular, the following distances, originally introduced in the context of nonparametric hypothesis testing, are also widely used.*

- ***The Spearman ρ distance.*** $\forall(\sigma,\sigma') \in \mathfrak{S}_N^2$, $d_2(\sigma,\sigma') = \left(\sum_{i=1}^N (\sigma(i)-\sigma'(i))^2\right)^{1/2}$
- ***The Spearman footrule distance.*** $\forall(\sigma,\sigma') \in \mathfrak{S}_N^2$, $d_1(\sigma,\sigma') = \sum_{i=1}^N |\sigma(i)-\sigma'(i)|$
- ***The Hamming distance.*** $\forall(\sigma,\sigma') \in \mathfrak{S}_N^2$, $d_H(\sigma,\sigma') = \sum_{i=1}^N \mathbb{I}\{\sigma(i) \neq \sigma'(i)\}$

The problem (IV.1) can be viewed as a M-estimation problem in the probabilistic framework stipulating that the collection of rankings to be aggregated/summarized is composed of $n \geq 1$ independent copies $\Sigma_1, \ldots, \Sigma_n$ of a generic r.v. Σ, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and drawn from an unknown probability distribution P on $\mathfrak{S}_N$ (*i.e.* $P(\sigma) = \mathbb{P}\{\Sigma = \sigma\}$ for any $\sigma \in \mathfrak{S}_N$). Just like a median of a real valued r.v. Z is any scalar closest to Z in the L_1 sense, a (true) median of distribution P w.r.t. a certain metric d on $\mathfrak{S}_N$ is any solution of the minimization problem:

$$\min_{\sigma \in \mathfrak{S}_N} L_P(\sigma), \tag{IV.3}$$

where $L_P(\sigma) = \mathbb{E}_{\Sigma\sim P}[d(\Sigma,\sigma)]$ denotes the expected distance between any permutation σ and Σ. In this framework, statistical ranking aggregation consists in recovering a solution σ^* of this minimization problem, plus an estimate of this minimum $L_P^* = L_P(\sigma^*)$, as accurate as possible, based on the observations $\Sigma_1, \ldots, \Sigma_n$. A median permutation σ^* can be interpreted as a central value for distribution P, while the quantity L_P^* may be viewed as a dispersion measure. Like problem (IV.1), the minimization problem (IV.3) has always a solution but can be multimodal. However, the functional $L_P(.)$ is unknown in practice, just like distribution P. Suppose that we would like to avoid rigid parametric assumptions on P and only have access to the dataset $(\Sigma_1, \ldots, \Sigma_n)$ to find a reasonable approximant of a median. The Empirical Risk Minimization (ERM) paradigm, see [Vap00], encourages us to substitute in (IV.3) the quantity $L_P(\sigma)$ with its statistical version

$$\widehat{L}_n(\sigma) = \frac{1}{n}\sum_{s=1}^n d(\Sigma_s,\sigma) = L_{\widehat{P}_n}(\sigma), \tag{IV.4}$$

where $\widehat{P}_n = (1/n)\sum_{s=1}^n \delta_{\Sigma_s}$ denotes the empirical measure. The performance of empirical consensus rules, solutions $\widehat{\sigma}_n$ of $\min_{\sigma\in\mathfrak{S}_N} \widehat{L}_n(\sigma)$, has been investigated in [KCS17].

Precisely, rate bounds of order $O_{\mathbb{P}}(1/\sqrt{n})$ for the excess of risk $L_P(\widehat{\sigma}_n) - L_P^*$ in probability/expectation have been established and proved to be sharp in the minimax sense, when d is the Kendall's τ distance. Whereas problem (IV.1) is NP-hard in general (see *e.g.* [Hud08]), in the Kendall's τ case, exact solutions, referred to as *Kemeny medians*, can be explicitly derived when the pairwise probabilities $p_{i,j} = \mathbb{P}\{\Sigma(i) < \Sigma(j)\}$, $1 \leq i \neq j \leq N$, fulfill the following property, referred to as *stochastic transitivity*.

Definition 1. *Let P be a probability distribution on $\mathfrak{S}_N$.*
(i) Distribution P is said to be (weakly) stochastically transitive iff

$$\forall (i,j,k) \in [\![N]\!]^3 : \ \ p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq 1/2.$$

If, in addition, $p_{i,j} \neq 1/2$ for all $i < j$, one says that P is strictly stochastically transitive.
(ii) Distribution P is said to be strongly stochastically transitive iff

$$\forall (i,j,k) \in [\![N]\!]^3 : \ \ p_{i,j} \geq 1/2 \text{ and } p_{j,k} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq max(p_{i,j}, p_{j,k}).$$

This is equivalent to the following condition (see [DM59]):

$$\forall (i,j) \in [\![N]\!]^2 : \ \ p_{i,j} \geq 1/2 \ \Rightarrow \ p_{i,k} \geq p_{j,k} \text{ for all } k \in [\![N]\!] \setminus \{i,j\}.$$

These conditions were firstly introduced in the psychology literature ([Fis73], [DM59]) and were used recently for the estimation of pairwise probabilities and ranking from pairwise comparisons ([SBGW15], [SW15]). Examples of stochastically transitive distributions on $\mathfrak{S}_N$ are far from uncommon and include most popular parametric models such as Mallows or Bradley-Terry-Luce-Plackett models, see *e.g.* [Mal57] or [Pla75]. When stochastic transitivity holds true, the set of Kemeny medians (see Theorem 5 in [KCS17]) is the set $\{\sigma \in \mathfrak{S}_N : \ (p_{i,j} - 1/2)(\sigma(j) - \sigma(i)) > 0 \text{ for all } i < j \text{ s.t. } p_{i,j} \neq 1/2\}$, and the minimum is given by

$$L_P^* = \sum_{i<j} \min\{p_{i,j}, 1 - p_{i,j}\} = \sum_{i<j} \{1/2 - |p_{i,j} - 1/2|\}. \tag{IV.5}$$

If a strict version of stochastic transitivity is fulfilled, we denote by σ_P^* the Kemeny median which is unique and given by the Copeland ranking, that assigns for each i its rank as:

$$\sigma_P^*(i) = 1 + \sum_{j \neq i} \mathbb{I}\{p_{i,j} < 1/2\} \text{ for } 1 \leq i \leq N. \tag{IV.6}$$

Assume that the underlying distribution P is strictly stochastically transitive and verifies additionally a certain low-noise condition $\mathbf{NA}(h)$, defined for $h > 0$ by:

$$\min_{i<j} |p_{i,j} - 1/2| \geq h. \tag{IV.7}$$

This condition is checked in many situations, including most conditional parametric models (see Remark 13 in [KCS17]) under simple assumptions on their parameters. It may be considered as analogous to that introduced in [KB05] in binary classification,

and was used to prove fast rates also in ranking, for the estimation of the matrix of pairwise probabilities (see [SBGW15]) or ranking aggregation (see [KCS17]). Indeed it is shown in [KCS17] that under condition (IV.7), the empirical distribution $\widehat{P}_n$ is also strictly stochastically transitive with overwhelming probability, and that the expectation of the excess of risk of empirical Kemeny medians decays at an exponential rate, see Proposition 14 therein. In this case, the nearly optimal solution $\sigma^*_{\widehat{P}_n}$ can be made explicit and straightforwardly computed using Eq. (IV.6) based on the empirical pairwise probabilities:

$$\widehat{p}_{i,j} = \frac{1}{n}\sum_{s=1}^{n} \mathbb{I}\{\Sigma_s(i) < \Sigma_s(j)\}.$$

As shall be shown below, the quantity $L_P(\sigma)$ can be seen as a Wasserstein distance between P and the Dirac mass δ_σ, so that Kemeny consensus ranking can thus be viewed as a radical dimensionality reduction procedure, summarizing P by its closest Dirac measure w.r.t. the distance on the set of probability distributions on $\mathfrak{S}_N$ aforementioned. The general framework for dimensionality reduction developed in the next subsection can be viewed as an extension of consensus ranking.

2.2 A Mass Transportation Approach to Dimensionality Reduction on $\mathfrak{S}_N$

We now develop a framework, that is shown to extend consensus ranking, for *dimensionality reduction* fully tailored to ranking data exhibiting a specific type of *sparsity*. For this purpose, we consider the so-termed *mass transportation* approach to defining metrics on the set of probability distributions on $\mathfrak{S}_N$ as follows, see [Rac91] (incidentally, this approach is also used in [CJ10] to introduce a specific relaxation of the consensus ranking problem).

Definition 2. *Let $d : \mathfrak{S}_N^2 \to \mathbb{R}_+$ be a metric on $\mathfrak{S}_N$ and $q \geq 1$. The q-th Wasserstein metric with d as cost function between two probability distributions P and P' on $\mathfrak{S}_N$ is given by:*

$$W_{d,q}\left(P, P'\right) = \inf_{\Sigma \sim P,\, \Sigma' \sim P'} \mathbb{E}\left[d^q(\Sigma, \Sigma')\right], \tag{IV.8}$$

where the infimum is taken over all possible couplings[1] *(Σ, Σ') of (P, P').*

As revealed by the following result, when the cost function d is equal to the Kendall's τ distance, which case the subsequent analysis focuses on, the Wasserstein metric is bounded by below by the l_1 distance between the pairwise probabilities.

Lemma 1. *For any probability distributions P and P' on $\mathfrak{S}_N$:*

$$W_{d_\tau,1}\left(P, P'\right) \geq \sum_{i<j} |p_{i,j} - p'_{i,j}|. \tag{IV.9}$$

[1] Recall that a coupling of two probability distributions Q and Q' is a pair (U, U') of random variables defined on the same probability space such that the marginal distributions of U and U' are Q and Q'.

The equality holds true when the distribution P' is deterministic (i.e. when $\exists \sigma \in \mathfrak{S}_N$ s.t. $P' = \delta_\sigma$).

The proof of Lemma 1 as well as discussions on alternative cost functions (the Spearman ρ distance) are deferred to section 8. As shown below, (IV.9) is actually an equality for various distributions P' built from P that are of special interest regarding dimensionality reduction.

Sparsity and Bucket Orders. Here, we propose a way of describing a distribution P on $\mathfrak{S}_N$, originally described by $N!-1$ parameters, by finding a much simpler distribution that approximates P in the sense of the Wasserstein metric introduced above under specific assumptions, extending somehow the consensus ranking concept. Let $2 \leq K \leq N$ and $\mathcal{C} = (\mathcal{C}_1, \ \ldots, \ \mathcal{C}_K)$ be a *bucket order* of $[\![N]\!]$ with K buckets. In order to gain insight into the rationale behind the approach we promote, observe that a distribution P' can be naturally said to be *sparse* if, for all $1 \leq k < l \leq K$ and all $(i, j) \in \mathcal{C}_k \times \mathcal{C}_l$ (i.e. $i \prec_{\mathcal{C}} j$), we have $p'_{j,i} = 0$, which means that with probability one $\Sigma'(i) < \Sigma'(j)$, when $\Sigma' \sim P'$. In other words, the relative order of two items belonging to two different buckets is deterministic. Throughout the chapter, such a probability distribution is referred to as a *bucket distribution* associated to $\mathcal{C}$. Since the variability of a bucket distribution corresponds to the variability of its marginals within the buckets $\mathcal{C}_k$'s, the set $\mathbf{P}_\mathcal{C}$ of all bucket distributions associated to $\mathcal{C}$ is of dimension $d_\mathcal{C} = \prod_{k \leq K} \#\mathcal{C}_k! - 1 \leq N! - 1$. A best summary in $\mathbf{P}_\mathcal{C}$ of a distribution P on $\mathfrak{S}_N$, in the sense of the Wasserstein metric (IV.8), is then given by any solution $P^*_\mathcal{C}$ of the minimization problem

$$\min_{P' \in \mathbf{P}_\mathcal{C}} W_{d_\tau,1}(P, P'). \tag{IV.10}$$

Set $\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_\mathcal{C}} W_{d_\tau,1}(P, P')$ for any bucket order $\mathcal{C}$.

Dimensionality Reduction. Let $K \leq N$. We denote by $\mathbf{C}_K$ the set of all bucket orders $\mathcal{C}$ of $[\![N]\!]$ with K buckets. If P can be accurately approximated by a probability distribution associated to a bucket order with K buckets, a natural dimensionality reduction approach consists in finding a solution $\mathcal{C}^{*(K)}$ of

$$\min_{\mathcal{C} \in \mathbf{C}_K} \Lambda_P(\mathcal{C}), \tag{IV.11}$$

as well as a solution $P^*_{\mathcal{C}^{*(K)}}$ of (IV.10) for $\mathcal{C} = \mathcal{C}^{*(K)}$ and a coupling $(\Sigma, \Sigma_{\mathcal{C}^{*(K)}})$ s.t. $\mathbb{E}[d_\tau(\Sigma, \Sigma_{\mathcal{C}^{*(K)}})] = \Lambda_P(\mathcal{C}^{*(K)})$.

Connection with Consensus Ranking. Observe that $\cup_{\mathcal{C} \in \mathbf{C}_N} \mathbf{P}_\mathcal{C}$ is the set of all Dirac distributions δ_σ, $\sigma \in \mathfrak{S}_N$. Hence, in the case $K = N$, dimensionality reduction as formulated above boils down to solve Kemeny consensus ranking. Indeed, we have: $\forall \sigma \in \mathfrak{S}_N$, $W_{d_\tau,1}(P, \delta_\sigma) = L_P(\sigma)$. Hence, medians σ^* of a probability distribution P (*i.e.* solutions of (IV.3)) correspond to the Dirac distributions δ_{σ^*} closest to P in the sense of the Wasserstein metric (IV.8): $P^*_{\mathcal{C}^{*(N)}} = \delta_{\sigma^*}$ and $\Sigma_{\mathcal{C}^{*(N)}} = \sigma^*$. Whereas the space

of probability measures on $\mathfrak{S}_N$ is of explosive dimension $N! - 1$, consensus ranking can be thus somehow viewed as a radical dimension reduction technique, where the original distribution is summarized by a median permutation σ^*. In constrast, the other extreme case $K = 1$ corresponds to no dimensionality reduction at all, *i.e.* $\Sigma_{\mathcal{C}^{*(1)}} = \Sigma$.

2.3 Optimal Couplings and Minimal Distortion

Fix a bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$. A simple way of building a distribution in $\mathbf{P}_\mathcal{C}$ based on P consists in considering the random ranking $\Sigma_\mathcal{C}$ coupled with Σ, that ranks the elements of any bucket $\mathcal{C}_k$ in the same order as Σ and whose distribution $P_\mathcal{C}$ belongs to $\mathbf{P}_\mathcal{C}$:

$$\forall k \in \{1, \ldots, K\},\ \forall i \in \mathcal{C}_k,\ \ \Sigma_\mathcal{C}(i) = 1 + \sum_{l<k} \#\mathcal{C}_l + \sum_{j \in \mathcal{C}_k} \mathbb{I}\{\Sigma(j) < \Sigma(i)\}, \tag{IV.12}$$

which defines a permutation. Distributions P and $P_\mathcal{C}$ share the same marginals within the $\mathcal{C}_k$'s and thus have the same intra-bucket pairwise probabilities $(p_{i,j})_{(i,j)\in\mathcal{C}_k^2}$, for all $k \in \{1, \ldots, K\}$. Observe that the expected Kendall's τ distance between Σ and $\Sigma_\mathcal{C}$ is given by:

$$\mathbb{E}\left[d_\tau\left(\Sigma, \Sigma_\mathcal{C}\right)\right] = \sum_{i \prec_\mathcal{C} j} p_{j,i} = \sum_{1 \leq k < l \leq K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} p_{j,i}, \tag{IV.13}$$

which can be interpreted as the expected number of pairs for which Σ violates the (partial) strict order defined by the bucket order $\mathcal{C}$. The result stated below shows that $(\Sigma, \Sigma_\mathcal{C})$ is *optimal* among all couplings between P and distributions in $\mathbf{P}_\mathcal{C}$ in the sense where (IV.13) is equal to the minimum of (IV.10), namely $\Lambda_P(\mathcal{C})$.

Proposition 1. *Let P be any distribution on $\mathfrak{S}_N$. For any bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$, we have:*

$$\Lambda_P(\mathcal{C}) = \sum_{i \prec_\mathcal{C} j} p_{j,i}. \tag{IV.14}$$

The proof, given in section 8, reveals that (IV.9) in Lemma 1 is actually an equality when $P' = P_\mathcal{C}$ and that $\Lambda_P(\mathcal{C}) = W_{d_\tau,1}\left(P, P_\mathcal{C}\right) = \mathbb{E}\left[d_\tau\left(\Sigma, \Sigma_\mathcal{C}\right)\right]$. Attention must be paid that it is quite remarkable that, when the Kendall's τ distance is chosen as cost function, the distortion measure introduced admits a simple closed-analytical form, depending on elementary marginals solely, the pairwise probabilities namely. Hence, the distortion of any bucket order can be straightforwardly estimated from independent copies of Σ, opening up to the design of practical dimensionality reduction techniques based on empirical distortion minimization, as investigated in the next section. The case where the cost is the Spearman ρ distance is also discussed in section 7: it is worth noticing that, in this situation as well, the distortion can be expressed in a simple manner, as a function of triplet-wise probabilities namely.

Property 1. *Let P be stochastically transitive. A bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ is said to agree with Kemeny consensus iff we have: $i \prec_\mathcal{C} j$ (i.e. $\exists k < l$, $(i,j) \in \mathcal{C}_k \times \mathcal{C}_l$) $\Rightarrow p_{j,i} \leq 1/2$.*

As recalled in the previous subsection, the quantity L_P^* can be viewed as a natural dispersion measure of distribution P and can be expressed as a function of the $p_{i,j}$'s as soon as P is stochastically transitive. The remarkable result stated below shows that, in this case and for any bucket order $\mathcal{C}$ satisfying Property 1, P's dispersion can be decomposed as the sum of the (reduced) dispersion of the simplified distribution $P_{\mathcal{C}}$ and the minimum distortion $\Lambda_P(\mathcal{C})$.

Corollary 1. *Suppose that P is stochastically transitive. Then, for any bucket order $\mathcal{C}$ that agrees with Kemeny consensus, we have:*

$$L_P^* = L_{P_{\mathcal{C}}}^* + \Lambda_P(\mathcal{C}). \tag{IV.15}$$

In the case where P is strictly stochastically transitive, the Kemeny median σ_P^* of P is unique (see [KCS17]). If $\mathcal{C}$ fulfills Property 1, it is also obviously the Kemeny median of the bucket distribution $P_{\mathcal{C}}$. As shall be seen in the next section, when P fulfills a strong version of the stochastic transitivity property, optimal bucket orders $\mathcal{C}^{*(K)}$ necessarily agree with the Kemeny consensus, which may greatly facilitates their statistical recovery.

2.4 Related Work

The dimensionality reduction approach developed in this chapter is connected with the *optimal bucket order* (OBO) problem considered in the literature, see *e.g.* [AGR17], [AGR18], [FFN08], [GMPU06], [UPGM09]. Given the pairwise probabilities $(p_{i,j})_{1\leq i\neq j\leq N}$ of a distribution P over $\mathfrak{S}_N$, solving the OBO problem consists in finding a bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ that minimizes the following cost:

$$\widetilde{\Lambda}_P(\mathcal{C}) = \sum_{i\neq j} |p_{i,j} - \widetilde{p}_{i,j}|, \tag{IV.16}$$

where $\widetilde{p}_{i,j} = 1$ if $i \prec_{\mathcal{C}} j$, $\widetilde{p}_{i,j} = 0$ if $j \prec_{\mathcal{C}} i$ and $\widetilde{p}_{i,j} = 1/2$ if $i \sim_{\mathcal{C}} j$. In other words, the $\widetilde{p}_{i,j}$'s are the pairwise marginals of the bucket distribution $\widetilde{P}_{\mathcal{C}}$ related to $\mathcal{C}$ with independent and uniformly distributed partial rankings $\Pi_{\mathcal{C}_k}(\widetilde{\Sigma}_{\mathcal{C}})$'s for $\widetilde{\Sigma}_{\mathcal{C}} \sim \widetilde{P}_{\mathcal{C}}$. Moreover, this cost verifies:

$$\widetilde{\Lambda}_P(\mathcal{C}) = 2\Lambda_P(\mathcal{C}) + \sum_{k=1}^{K} \sum_{(i,j)\in\mathcal{C}_k^2} |p_{i,j} - 1/2|. \tag{IV.17}$$

Observe that solving the OBO problem is much more restrictive than the framework we developed, insofar as no constraint is set about the intra-bucket marginals of the summary distributions solutions of (IV.11). Another related work is documented in [SBW16, PMM$^+$17] and develops the concept of *indifference sets*. Formally, a family of pairwise probabilities $(\widetilde{p}_{i,j})$ is said to satisfy the indifference set partition (or bucket order) $\mathcal{C}$ when:

$$\widetilde{p}_{i,j} = \widetilde{p}_{i',j'} \text{ for all quadruples } (i,j,i',j') \text{ such that } i \sim_{\mathcal{C}} i' \text{ and } j \sim_{\mathcal{C}} j', \tag{IV.18}$$

which condition also implies that the intra-bucket marginals are s.t. $\widetilde{p}_{i,j} = 1/2$ for $i \sim_{\mathcal{C}} j$ (take $i' = j$ and $j' = i$ in (IV.18)). Though related, our approach significantly differs from these works, since it avoids stipulating arbitrary distributional assumptions. For instance, it permits in contrast to test *a posteriori*, once the best bucket order $\mathcal{C}^{*(K)}$ is determined for a fixed K, statistical hypotheses such as the independence of the bucket marginal components (*i.e.* $\Pi_{\mathcal{C}_k^{*(K)}}(\Sigma)$'s) or the uniformity of certain bucket marginal distributions. A summary distribution, often very informative and of small dimension both at the same time, is the marginal of the first bucket $\mathcal{C}_1^{*(K)}$ (the top-m rankings where $m = |\mathcal{C}_1^{*(K)}|$).

3 Empirical Distortion Minimization - Rate Bounds and Model Selection

In order to recover optimal bucket orders, based on the observation of a training sample $\Sigma_1, \ldots, \Sigma_n$ of independent copies of Σ, Empirical Risk Minimization, the major paradigm of statistical learning, naturally suggests to consider bucket orders $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ minimizing the empirical version of the distortion (IV.14)

$$\widehat{\Lambda}_n(\mathcal{C}) = \sum_{i \prec_{\mathcal{C}} j} \widehat{p}_{j,i} = \Lambda_{\widehat{P}_n}(\mathcal{C}), \tag{IV.19}$$

where the $\widehat{p}_{i,j}$'s are the pairwise probabilities of the empirical distribution. For a given shape λ, we define the Rademacher average

$$\mathcal{R}_n(\lambda) = \mathbb{E}_{\epsilon_1,\ldots,\epsilon_n}\left[\max_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \frac{1}{n}\left|\sum_{s=1}^{n} \epsilon_s \sum_{i\prec_{\mathcal{C}} j} \mathbb{I}\{\Sigma_s(j) < \Sigma_s(i)\}\right|\right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher r.v.'s (*i.e.* symmetric sign random variables), independent from the Σ_s's. Fix the number of buckets $K \in \{1, \ldots, N\}$, as well as the bucket order shape $\lambda = (\lambda_1, \ldots, \lambda_K) \in \mathbb{N}^{*K}$ such that $\sum_{k=1}^{K} \lambda_k = N$. We recall that $\mathbf{C}_K = \cup_{\lambda'=(\lambda'_1,\ldots,\lambda'_K)\in\mathbb{N}^{*K} \text{ s.t. } \sum_{k=1}^{K}\lambda'_k=N} \mathbf{C}_{K,\lambda'}$. The result stated below describes the generalization capacity of solutions of the minimization problem

$$\min_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \widehat{\Lambda}_n(\mathcal{C}), \tag{IV.20}$$

over the class $\mathbf{C}_{K,\lambda}$ of bucket orders $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of shape λ (*i.e.* s.t. $\lambda = (\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$), through a rate bound for their excess of distortion. Its proof is given in section 8.

Theorem 1. *Let $\widehat{C}_{K,\lambda}$ be any empirical distortion minimizer over $\mathbf{C}_{K,\lambda}$, i.e solution of* (IV.20)*. Then, for all $\delta \in (0,1)$, we have with probability at least $1-\delta$:*

$$\Lambda_P(\widehat{C}_{K,\lambda}) - \inf_{\mathcal{C}\in\mathbf{C}_K} \Lambda_P(\mathcal{C}) \leq 4\mathbb{E}\left[\mathcal{R}_n(\lambda)\right] + \kappa(\lambda)\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \left\{\inf_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C}) - \inf_{\mathcal{C}\in\mathbf{C}_K} \Lambda_P(\mathcal{C})\right\},$$

where $\kappa(\lambda) = \sum_{k=1}^{K-1} \lambda_k \times (N - \lambda_1 - \ldots - \lambda_k)$.

We point out that the Rademacher average is of order $O(1/\sqrt{n})$: $\mathcal{R}_n(\lambda) \leq \kappa(\lambda)\sqrt{2\log\left(\binom{N}{\lambda}\right)/n}$ with $\binom{N}{\lambda} = N!/(\#\mathcal{C}_1! \times \cdots \times \#\mathcal{C}_K!) = \#\mathbf{C}_{K,\lambda}$, where $\kappa(\lambda)$ is the number of terms involved in (IV.14)-(IV.19) and $\binom{N}{\lambda}$ is the multinomial coefficient, *i.e.* the number of bucket orders of shape λ. Putting aside the approximation error, the rate of decay of the distortion excess is classically of order $O_{\mathbb{P}}(1/\sqrt{n})$.

Remark 2. (Empirical Distortion Minimization over $\mathbf{C}_K$) *We point out that rate bounds describing the generalization ability of minimizers of* (IV.19) *over the whole class* $\mathbf{C}_K$ *can be obtained using a similar argument. A slight modification of Theorem 1's proof shows that, with probability larger than* $1-\delta$*, their excess of distortion is less than* $N^2(K-1)/K\sqrt{\log(N^2(K-1)\#\mathbf{C}_K/(K\delta))/(2n)}$*. Indeed, denoting by* $\lambda_{\mathcal{C}}$ *the shape of any bucket order* $\mathcal{C}$ *in* $\mathbf{C}_K$*,* $\max_{\mathcal{C}\in\mathbf{C}_K}\kappa(\lambda_{\mathcal{C}}) \leq N^2(K-1)/(2K)$*, the upper bound being attained when* K *divides* N *for* $\lambda_1 = \cdots = \lambda_K = N/K$*. In addition, we have:* $\#\mathbf{C}_K = \sum_{k=0}^{K}(-1)^{K-k}\binom{K}{k}k^N$.

Remark 3. (Alternative statistical framework) *Since the distortion* (IV.14) *involves pairwise comparisons solely, an empirical version could be computed in a statistical framework stipulating that the observations are of pairwise nature,* $(\mathbb{I}\{\Sigma_1(i_1) < \Sigma_1(j_1)\}, \ldots, \mathbb{I}\{\Sigma_n(i_n) < \Sigma_n(j_n)\})$*, where* $\{(i_s, j_s),\ s = 1, \ldots, n\}$*, are i.i.d. pairs, independent from the* Σ_s*'s, drawn from an unknown distribution* ν *on the set* $\{(i,j) : 1 \leq i < j \leq N\}$ *such that* $\nu(\{(i,j)\}) > 0$ *for all* $i < j$*. Based on these observations, more easily available in most practical applications (see e.g. [CBCTH13], [PNZ+15]), the pairwise probability* $p_{i,j}$*,* $i < j$*, can be estimated by:*

$$\frac{1}{n_{i,j}}\sum_{s=1}^{n}\mathbb{I}\{(i_s,j_s) = (i,j),\ \Sigma_s(i_s) < \Sigma_s(j_s)\},$$

with $n_{i,j} = \sum_{s=1}^{n}\mathbb{I}\{(i_s,j_s) = (i,j)\}$ *and the convention* $0/0 = 0$.

Remark 4. (Low-dimensional representations) *For any ranking agent described by its intrinsic preferences* $\Sigma \sim P$*, the challenge of dimensionality reduction consists in avoiding fully observing* Σ*. Given a solution* $\widehat{C}_{K,\lambda}$ *of* (IV.20)*, by only asking to the ranking agent to order items inside each bucket* $\widehat{C}_{K,\lambda,k}$ *for* $k \in \{1,\ldots,K\}$*, one can reconstruct the associated optimal ranking* $\Sigma_{\widehat{C}_{K,\lambda}}$ *coupled with* Σ *and verifying (see Eq.* (IV.13)*):*

$$\mathbb{E}_{\Sigma\sim P}\left[d_\tau\left(\Sigma, \Sigma_{\widehat{C}_{K,\lambda}}\right)\middle|\widehat{C}_{K,\lambda}\right] = \Lambda_P(\widehat{C}_{K,\lambda}).$$

In other words, the expected approximation error (in terms of Kendall's τ *distance) for observing* $\Sigma_{\widehat{C}_{K,\lambda}}$ *instead of* Σ *is* $\Lambda_P(\widehat{C}_{K,\lambda})$*, which is controlled by the generalization bound given in Theorem 1. This approach actually corresponds to sampling w.r.t.* $P_{\widehat{C}_{K,\lambda}}$ *instead of* P*, their Wasserstein distance being* $W_{d_\tau,1}\left(P, P_{\widehat{C}_{K,\lambda}}\right) = \Lambda_P(\widehat{C}_{K,\lambda})$.

Selecting the Shape of the Bucket Order. A crucial issue in dimensionality reduction is to determine the dimension of the simpler representation of the distribution of interest. Here we consider a complexity regularization method to select the bucket order shape λ that uses a data-driven penalty based on Rademacher averages. Suppose that a sequence $\{(K_m, \lambda_m)\}_{1 \leq m \leq M}$ of bucket order sizes/shapes is given (observe that $M \leq \sum_{K=1}^{N} \binom{N-1}{K-1} = 2^{N-1}$). In order to avoid overfitting, consider the complexity penalty given by

$$\text{PEN}(\lambda_m, n) = 2\mathcal{R}_n(\lambda_m) \tag{IV.21}$$

and the minimizer $\widehat{\mathcal{C}}_{K_{\widehat{m}},\lambda_{\widehat{m}}}$ of the penalized empirical distortion, with

$$\widehat{m} = \arg\min_{1 \leq m \leq M} \left\{ \widehat{\Lambda}_n(\widehat{\mathcal{C}}_{K_m,\lambda_m}) + \text{PEN}(\lambda_m, n) \right\} \text{ and } \widehat{\Lambda}_n(\widehat{\mathcal{C}}_{K,\lambda}) = \min_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \widehat{\Lambda}_n(\mathcal{C}). \tag{IV.22}$$

The next result shows that the bucket order thus selected nearly achieves the performance that would be obtained with the help of an oracle, revealing the value of the index m ruling the bucket order size/shape that minimizes $\mathbb{E}[\Lambda_P(\widehat{\mathcal{C}}_{K_m,\lambda_m})]$.

Theorem 2. (An oracle inequality) *Let $\widehat{\mathcal{C}}_{K_{\widehat{m}},\lambda_{\widehat{m}}}$ be any penalized empirical distortion minimizer over $\mathbf{C}_{K_{\widehat{m}},\lambda_{\widehat{m}}}$, i.e solution of* (IV.22)*. Then we have:*

$$\mathbb{E}\left[\Lambda_P(\widehat{\mathcal{C}}_{K_{\widehat{m}},\lambda_{\widehat{m}}})\right] \leq \min_{1 \leq m \leq M} \left\{ \mathbb{E}\left[\Lambda_P(\widehat{\mathcal{C}}_{K_m,\lambda_m})\right] + 2\mathbb{E}\left[\mathcal{R}_n(\lambda_m)\right] \right\} + 5M\binom{N}{2}\sqrt{\frac{\pi}{2n}}.$$

The Strong Stochastic Transitive Case. The theorem below shows that, when strong/strict stochastic transitivity properties hold for the considered distribution P, optimal buckets are those which agree with the Kemeny median.

Theorem 3. *Suppose that P is strongly/strictly stochastically transitive. Let $K \in \{1, \ldots, N\}$ and $\lambda = (\lambda_1, \ldots, \lambda_K)$ be a given bucket size and shape. Then, the minimizer of the distortion $\Lambda_P(\mathcal{C})$ over $\mathbf{C}_{K,\lambda}$ is unique and given by $\mathcal{C}^{*(K,\lambda)} = (\mathcal{C}_1^{*(K,\lambda)}, \ldots, \mathcal{C}_K^{*(K,\lambda)})$, where*

$$\mathcal{C}_k^{*(K,\lambda)} = \left\{ i \in [\![N]\!] : \sum_{l<k} \lambda_l < \sigma_P^*(i) \leq \sum_{l \leq k} \lambda_l \right\} \text{ for } k \in \{1, \ldots, K\}. \tag{IV.23}$$

In addition, for any $\mathcal{C} \in \mathbf{C}_{K,\lambda}$, we have:

$$\Lambda_P(\mathcal{C}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)}) \geq 2 \sum_{j \prec_{\mathcal{C}} i} (1/2 - p_{i,j}) \cdot \mathbb{I}\{p_{i,j} < 1/2\}. \tag{IV.24}$$

In other words, $\mathcal{C}^{*(K,\lambda)}$ is the unique bucket in $\mathbf{C}_{K,\lambda}$ that agrees with σ_P^* (*cf* Property 1). Hence, still under the hypotheses of Theorem 3, the minimizer $\mathcal{C}^{*(K)}$ of (IV.11) also agrees with σ_P^* and corresponds to one of the $\binom{N-1}{K-1}$ possible segmentations of the ordered list $(\sigma_P^{*-1}(1), \ldots, \sigma_P^{*-1}(N))$ into K segments. This property paves the way to design efficient procedures, such as the BuMeRank algorithm described in the next section, for

recovering bucket order representations with a fixed distortion rate of minimal dimension, avoiding to specify the size/shape in advance. If, in addition, condition (IV.7) is fulfilled, when $\widehat{P}_n$ is strictly stochastically transitive (which then happens with overwhelming probability, see Proposition 14 in [KCS17]), the computation of the empirical Kemeny median $\sigma^*_{\widehat{P}_n}$ is immediate from formula (IV.6) (replacing P by $\widehat{P}_n$), as well as an estimate of $\mathcal{C}^{*(K,\lambda)}$, plugging $\sigma^*_{\widehat{P}_n}$ into (IV.23) as implemented in the experiments below. When the empirical distribution $\widehat{P}_n$ is not stochastically transitive, which happens with negligible probability, the empirical median can be classically replaced by any permutation obtained from the Copeland score by breaking ties at random. The following result shows that, in the strict/strong stochastic transitive case, when the low-noise condition $\mathbf{NA}(h)$ is fulfilled, the excess of distortion of the empirical minimizers is actually of order $O_{\mathbb{P}}(1/n)$.

Theorem 4. *(Fast rates) Let λ be a given bucket order shape and $\widehat{C}_{K,\lambda}$ any empirical distortion minimizer over $\mathbf{C}_{K,\lambda}$. Suppose that P is strictly/strongly stochastically transitive and fulfills condition* (IV.7)*. Then, for any $\delta > 0$, we have with probability $1-\delta$:*

$$\Lambda_P(\widehat{\mathcal{C}}_{K,\lambda}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)}) \leq \left(\frac{2^{\binom{N}{2}+1}N^2}{h}\right) \times \frac{\log\left(\binom{N}{\lambda}/\delta\right)}{n}.$$

The proof is given in section 8.

4 The BuMeRank Algorithm: Hierarchical Recovery of a Bucket Distribution

Motivated by Theorem 3, we propose a hierarchical 'bottom-up' procedure to recover, from ranking data, a bucket order representation (agreeing with Kemeny consensus) of smallest dimension for a fixed level of distortion, that does not requires to specify in advance the bucket size K and thus avoids computing the optimum (IV.23) for all possible shape/size.

Suppose for simplicity that P is strictly/strongly stochastically transitive. One starts with the bucket order of size N defined by its Kemeny median σ^*_P:

$$\mathcal{C}(0) = (\{\sigma_P^{*-1}(1)\},\ \ldots,\ \{\sigma_P^{*-1}(N)\}).$$

The initial representation has minimum dimension, *i.e.* $d_{\mathcal{C}(0)} = 0$, and maximal distortion among all bucket order representations agreeing with σ^*_P, *i.e.* $\Lambda_P(\mathcal{C}(0)) = L^*_P$, see Corollary 1. The binary agglomeration strategy we propose, namely the BuMeRank (for '**Bu**cket **Me**rge') algorithm, consists in recursively merging two adjacent buckets $\mathcal{C}_k(j)$ and $\mathcal{C}_{k+1}(j)$ of the current bucket order $\mathcal{C}(j) = (\mathcal{C}_1(j),\ \ldots,\ C_K(j))$ into a single bucket, yielding the 'coarser' bucket order

$$\mathcal{C}(j+1) = (\mathcal{C}_1(j), \ldots, \mathcal{C}_{k-1}(j), \mathcal{C}_k(j) \cup \mathcal{C}_{k+1}(j), \mathcal{C}_{k+2}(j), \ldots, \mathcal{C}_K(j)). \tag{IV.25}$$

The pair $(\mathcal{C}_k(j), \mathcal{C}_{k+1}(j))$ chosen corresponds to that maximizing the quantity

$$\Delta_P^{(k)}(\mathcal{C}(j)) = \sum_{i \in \mathcal{C}_k(j), j \in \mathcal{C}_{k+1}(j)} p_{j,i}. \tag{IV.26}$$

The agglomerative stage $\mathcal{C}(j) \to \mathcal{C}(j+1)$ increases the dimension of the representation,

$$d_{\mathcal{C}(j+1)} = (d_{\mathcal{C}(j)} + 1) \times \binom{\#\mathcal{C}_k(j) + \#\mathcal{C}_{k+1}(j)}{\#\mathcal{C}_k(j)} - 1, \tag{IV.27}$$

while reducing the distortion by $\Lambda_P(\mathcal{C}(j)) - \Lambda_P(\mathcal{C}(j+1)) = \Delta_P^{(k)}(\mathcal{C}(j))$.

BuMeRank Algorithm

1. **Input.** Training data $\{\Sigma_i\}_{i=1}^n$, maximum dimension $d_{\max} \geq 0$, distortion tolerance $\epsilon \geq 0$.
2. **Initialization.** Compute empirical Kemeny median $\sigma^*_{\widehat{P}_n}$ and $\mathcal{C}(0) = (\{\sigma^{*-1}_{\widehat{P}_n}(1)\}, \ldots, \{\sigma^{*-1}_{\widehat{P}_n}(N)\})$. Set $K \leftarrow N$.
3. **Iterations.** While $K \geq 3$ and $\widehat{\Lambda}_n(\mathcal{C}(N-K)) > \epsilon$,
   a) Compute $k \in \arg\max_{1 \leq l \leq K-1} \Delta^{(l)}_{\widehat{P}_n}(\mathcal{C}(N-K))$ and $\mathcal{C}(N-K+1)$.
   b) If $d_{\mathcal{C}(N-K+1)} > d_{\max}$: go to 4. Else: set $K \leftarrow K-1$.
4. **Output.** Bucket order $\mathcal{C}(N-K)$.

For notational convenience, the BuMeRank algorithm is defined taking full rankings Σ_i's as input, but it remains valid in the pairwise comparisons framework (see Remark 3). This algorithm is specifically designed for finding the bucket order $\mathcal{C}$ of minimal dimension $d_{\mathcal{C}}$ (i.e. of maximal size K) such that a bucket distribution in $\mathbf{P}_{\mathcal{C}}$ approximates well the original distribution P (i.e. with small distortion $\Lambda_P(\mathcal{C})$). The next result formally supports this idea in the limit case of P being a bucket distribution.

Theorem 5. *Let P be a strongly/strictly stochastically transitive bucket distribution and denote $K^* = \max\{K \in \{2, \ldots, N\}, \exists$ bucket order $\mathcal{C}$ of size K s.t. $P \in \mathbf{P}_{\mathcal{C}}\}$.*
(i) There exists a unique K^*-shape λ^* such that $\Lambda_P(\mathcal{C}^{*(K^*,\lambda^*)}) = 0$.
(ii) For any bucket order $\mathcal{C}$ such that $P \in \mathbf{P}_{\mathcal{C}}$: $\mathcal{C} \neq \mathcal{C}^{*(K^*,\lambda^*)} \Rightarrow d_{\mathcal{C}} > d_{\mathcal{C}^{*(K^*,\lambda^*)}}$.
(iii) The BuMeRank algorithm, runned with $d_{max} = N! - 1$, $\epsilon = 0$ and theoretical quantities (σ^*_P, $\Delta_P^{(k)}$'s and Λ_P) instead of estimates, returns $\mathcal{C}^{*(K^*,\lambda^*)}$.

The proof is deferred to section 8. Hence, the BuMeRank algorithm allows to recover the bucket order $\mathcal{C}$ with minimal dimension such that $P \in \mathbf{P}_{\mathcal{C}}$.

5 Numerical Experiments

We provide numerical experiments on both real and artificial datasets.

5.1 Real-World Datasets

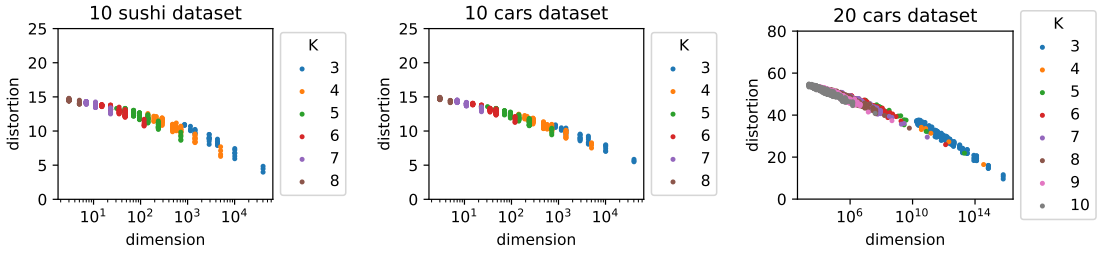


Figure IV.1: Dimension-Distortion plot for different bucket sizes on real-world preference datasets.

In this section we illustrate the relevance of our approach through real-world ranking datasets, which exhibit the type of sparsity considered in the present chapter. The first one is the well-known Sushi dataset (see [Kam03]), which consists of full rankings describing the preferences of $n = 5000$ individuals over $N = 10$ sushi dishes. We also considered the two *Cars preference datasets*[2] (see [EA13]). It consists of pairwise comparisons of users between N different cars. In the first dataset, 60 users are asked to make all the possible 45 pairwise comparisons between 10 cars (around 3000 samples). In the second one, 60 users are asked to make (randomly selected) 38 comparisons between 20 cars (around 2500 samples). For each dataset, the empirical ranking $\sigma^*_{\widehat{P}_n}$ is computed based on the empirical pairwise probabilities. In Figure IV.1, the dimension $d_{\mathcal{C}}$ (in logarithmic scale) *vs* distortion $\widehat{\Lambda}_n(\mathcal{C})$ diagram is plotted for each dataset, for several bucket sizes (K) and shapes (λ). These buckets are obtained by segmenting $\sigma^*_{\widehat{P}_n}$ with respect to λ as explained at the end of the previous section. Each color on a plot corresponds to a specific size K, and each point in a given color thus represents a bucket order of size K. As expected, on each plot the lowest distortion is attained for high-dimensional buckets (i.e., of smaller size K). These numerical results shed light on the sparse character of these empirical ranking distributions. Indeed, the dimension $d_{\mathcal{C}}$ can be drastically reduced, by choosing the size K and shape λ in an appropriate manner, while keeping a low distortion for the representation. We provide in the next subsection additional dimension/distortion plots on toy datasets for different distributions which underline the sparsity observed here: specifically, these empirical distributions show intermediate behaviors between a true bucket distribution and a uniform distribution (i.e., without exhibiting bucket sparsity).

5.2 Toy Datasets

We now provide an illustration of the notions we introduced in this chapter, in particular of a bucket distribution and of our distortion criteria. For $N = 6$ items, we fixed a bucket order $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3)$ of shape $\lambda = (2, 3, 1)$ and considered a bucket distribution $P \in \mathbf{P}_{\mathcal{C}}$.

[2] `http://users.cecs.anu.edu.au/~u4940058/CarPreferences.html`, First experiment.

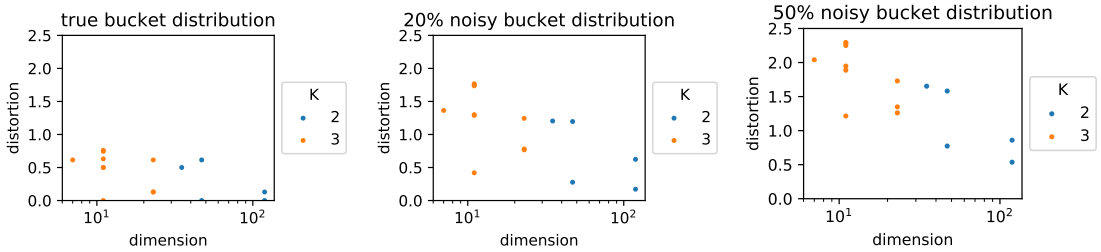


Figure IV.2: Dimension-Distortion plot for different bucket sizes on simulated datasets.

Specifically, P is the uniform distribution over all the permutations extending the bucket order $\mathcal{C}$ and has thus its pairwise marginals such that $p_{j,i} = 0$ as soon as $(i,j) \in \mathcal{C}_k \times \mathcal{C}_l$ with $k < l$. In Figure IV.2, the first plot on the left is a scatterplot of all buckets of size $K \in \{2,3\}$ where for any bucket $\mathcal{C}'$ of size K, the horizontal axis is the distortion $\Lambda_P(\mathcal{C}')$ (see (IV.14)) and the vertical axis is the dimension of $\mathbf{P}_{\mathcal{C}'}$ in log scale. On the left plot, one can see that one bucket of size $K = 3$ attains a null distortion, i.e. when $\mathcal{C}' = \mathcal{C}$, and two buckets of size $K = 2$ as well, i.e. when $\mathcal{C}' = (\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{C}_3)$ and when $\mathcal{C}' = (\mathcal{C}_1, \mathcal{C}_2 \cup \mathcal{C}_3)$. Then, a dataset of 2000 samples from P was drawn, and for a certain part of the samples, a pair of items was randomly swapped within the sample. The middle and right plot thus represent the empirical distortions $\widehat{\Lambda}_n(\mathcal{C}')$ for any $\mathcal{C}'$ computed on these datasets, where respectively 20% and 50% of the samples were contaminated. One can notice that the distortion is increasing with the noise, still, the best bucket of size 3 remains $\mathcal{C}' = \mathcal{C}$. However, the buckets $\mathcal{C}'$ attaining the minimum distortion in the noisy case are of size 2, because the distortion involves a smaller number of terms $\kappa(\lambda_{\mathcal{C}'})$ for a smaller size.

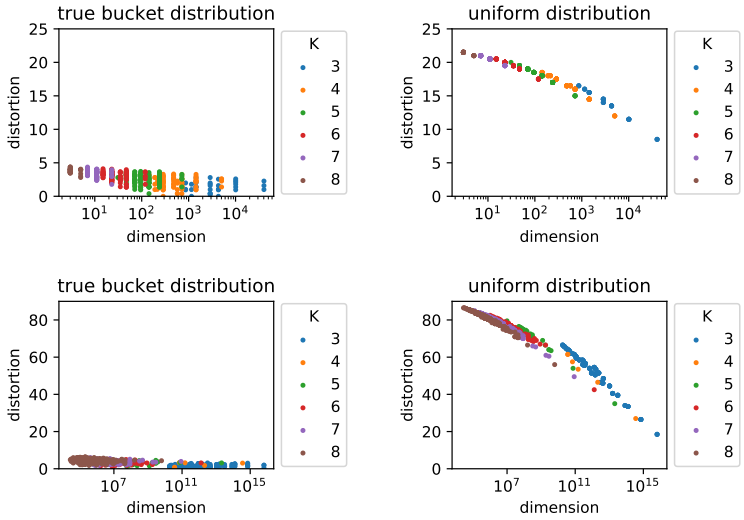


Figure IV.3: Dimension-Distortion plot for a true bucket distribution versus a uniform distribution ($N = 10$ on top and $N = 20$ below).

We now perform a second experiment. We want to compare the distortion versus dimension graph for a true bucket distribution (i.e., for a collection of pairwise marginals that respect a bucket order) and for a uniform distribution (i.e., a collection of pairwise marginals where $p_{j,i} = 0.5$ for all i, j). This corresponds to the plots on Figure IV.3. One can notice that the points are much more spread for a true bucket distribution, since some buckets will attain a very low distortion (those who agree with the true one) while some have a high distortion. In contrast, for a uniform distribution, all the buckets will perform relatively in the same way, and the scatter plot is much more compact.

6 Conclusion

In this chapter, we have developed novel theoretical concepts to represent efficiently *sparse* ranking data distributions. We have introduced a distortion measure, based on a mass transportation metric on the set of probability distributions on the set of rankings (with Kendall's τ as transportation cost) in order to evaluate the accuracy of (bucket) distribution representations and investigated the performance of empirical distortion minimizers. We have also provided empirical evidence that the notion of sparsity, which the dimensionality reduction method proposed relies on, is encountered in various real-world situations. Future research could investigate at length how to exploit such sparse representations for improving the completion of certain statistical learning tasks based on ranking data (e.g. clustering, ranking prediction), by circumventing this way the curse of dimensionality.

7 Perspective - Alternative Cost Function: The Spearman ρ Distance

As a direction of future research, we propose to extend the dimensionality reduction framework derived in this chapter from the Kendall's τ distance, to the Spearman ρ distance.

The expression of the distortion $\Lambda_P(\mathcal{C})$ obtained in Proposition 1 critically depends on the choice of the Wasserstein parameters, namely $d = d_\tau$ the Kendall's τ distance and $q = 1$. Although obtaining a closed-analytical form for the distortion is a challenging problem for general d and q, the following result shows that choosing the *Spearman ρ distance* $d = d_2$ as cost function and $q = 2$ leads to an *alternative distortion measure*:

$$\Lambda'_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_\mathcal{C}} W_{d_2,2}(P, P'),$$

that can be explicitly expressed in terms of the *triplet-wise probabilities*

$$p_{i,j,k} = \mathbb{P}_{\Sigma \sim P}\{\Sigma(i) < \Sigma(j) < \Sigma(k)\}.$$

In addition, the coupling $(\Sigma, \Sigma_\mathcal{C})$ can also be shown to be optimal in this case:

$$\Lambda'_P(\mathcal{C}) = \mathbb{E}\left[d_2^2\left(\Sigma, \Sigma_\mathcal{C}\right)\right].$$

Hence, based on the explicit formula below, the distortion can be straightforwardly estimated, just like the $p_{i,j,k}$'s, so that an analysis similar to that in section 3 in the Kendall's τ case, can be naturally carried out in order to provide statistical guarantees for the generalization capacity of empirical distortion minimization procedures.

Proposition 2. *Let $N \geq 3$ and P be any distribution on $\mathfrak{S}_N$. For any bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$, we have:*

$$\Lambda'_P(\mathcal{C}) = \frac{2}{N-2} \sum_{1 \leq k < l < m \leq K} \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_l \times \mathcal{C}_m} (N+1)p_{c,b,a} + N(p_{b,c,a} + p_{c,a,b}) + p_{b,a,c} + p_{a,c,b}$$
$$+ \frac{2}{N-2} \sum_{1 \leq k < l \leq K} \left\{ \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_l \times \mathcal{C}_l} N(p_{b,c,a} + p_{c,b,a}) + p_{b,a,c} + p_{c,a,b} \right.$$
$$\left. + \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_k \times \mathcal{C}_l} N(p_{c,a,b} + p_{c,b,a}) + p_{a,c,b} + p_{b,c,a} \right\}.$$

The proof is a straightforward consequence of the result stated below.

Lemma 2. *Let $N \geq 3$ and P be a probability distribution on $\mathfrak{S}_N$.*

(i) For any probability distribution P' on $\mathfrak{S}_N$:

$$W_{d_2,2}\left(P, P'\right) \geq \frac{2}{N-2} \sum_{a<b<c} \left\{ \sum_{(i,j,k) \in \mathfrak{s}(a,b,c)} \max(p_{i,j,k}, p'_{i,j,k}) - 1 \right\},$$

where $\mathfrak{s}(a,b,c)$ is the set of the 6 permutations of the triplet (a,b,c) and, for any $(i,j,k) \in \mathfrak{s}(a,b,c)$, $p'_{i,j,k} = \mathbb{P}_{\Sigma \sim P'}\{\Sigma(i) < \Sigma(j) < \Sigma(k)\}$.

(ii) If $P' \in \mathbf{P}_{\mathcal{C}}$ with $\mathcal{C}$ a bucket order of $[\![N]\!]$ with K buckets:

$$W_{d_2,2}\left(P, P'\right) \geq$$
$$\frac{2}{N-2} \sum_{1 \leq k < l < m \leq K} \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_l \times \mathcal{C}_m} (N+1)p_{c,b,a} + N(p_{b,c,a} + p_{c,a,b}) + p_{b,a,c} + p_{a,c,b}$$
$$+ \frac{2}{N-2} \sum_{1 \leq k < l \leq K} \left\{ \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_l \times \mathcal{C}_l} N(p_{b,c,a} + p_{c,b,a}) + p_{b,a,c} + p_{c,a,b} \right.$$
$$\left. + \sum_{(a,b,c) \in \mathcal{C}_k \times \mathcal{C}_k \times \mathcal{C}_l} N(p_{c,a,b} + p_{c,b,a}) + p_{a,c,b} + p_{b,c,a} \right\}, \tag{IV.28}$$

equality holding true when $P' = P_{\mathcal{C}}$, i.e. when P' is the distribution of $\Sigma_{\mathcal{C}}$.

Its proof is provided in section 8. This expression of the distortion $\Lambda'_P(\mathcal{C})$ in terms of the $p_{i,j,k}$'s suggests empirical methods using their estimates

$$\widehat{p}_{i,j,k} = \frac{1}{n}\sum_{s=1}^{n} \mathbb{I}\{\Sigma_s(i) < \Sigma_s(j) < \Sigma_s(k)\}.$$

As pointed in Remark 3 for pairs, observing full rankings Σ_s may be too costly in practical applications, where one would rather rely on triplet-wise comparisons $\mathbb{I}\{\Sigma_s(i_s) < \Sigma_s(j_s) < \Sigma_s(k_s)\}$:

$$\widehat{p}_{i,j,k} = \frac{1}{n_{i,j,k}}\sum_{s=1}^{n} \mathbb{I}\{(i_s,j_s,k_s) = (i,j,k),\ \Sigma_s(i_s) < \Sigma_s(j_s) < \Sigma_s(k_s)\},$$

with $n_{i,j,k} = \sum_{s=1}^{n} \mathbb{I}\{(i_s,j_s,k_s) = (i,j,k)\}$.

8 Technical Proofs

This technical section gathers the detailed proofs of this chapter's theoretical results.

Proof of Lemma 1

Consider two probability distributions P and P' on $\mathfrak{S}_N$. Fix $i \neq j$ and let (Σ, Σ') be a pair of random variables defined on a same probability space, valued in $\mathfrak{S}_N$ and such that $p_{i,j} = \mathbb{P}_{\Sigma\sim P}\{\Sigma(i) < \Sigma(j)\}$ and $p'_{i,j} = \mathbb{P}_{\Sigma'\sim P'}\{\Sigma'(i) < \Sigma'(j)\}$. Set

$$\pi_{i,j} = \mathbb{P}\left\{\Sigma'(i) < \Sigma'(j) \mid \Sigma(i) < \Sigma(j)\right\}.$$

Equipped with this notation, by the law of total probability, we have:

$$p'_{i,j} = p_{i,j}\pi_{i,j} + (1-p_{i,j})(1-\pi_{j,i}). \tag{IV.29}$$

In addition, we may write

$$\begin{aligned}\mathbb{E}\left[d_\tau(\Sigma,\Sigma')\right] &= \sum_{i<j}\mathbb{E}\left[\mathbb{I}\{(\Sigma(i)-\Sigma(j))(\Sigma'(i)-\Sigma'(j)) < 0\}\right] \\ &= \sum_{i<j}\mathbb{E}\left[\mathbb{I}\{\Sigma(i)<\Sigma(j)\}\mathbb{I}\{\Sigma'(i)>\Sigma'(j)\} + \mathbb{I}\{\Sigma(i)>\Sigma(j)\}\mathbb{I}\{\Sigma'(i)<\Sigma'(j)\}\right] \\ &= \sum_{i<j} p_{i,j}(1-\pi_{i,j}) + (1-p_{i,j})(1-\pi_{j,i}).\end{aligned}$$

Suppose that $p_{i,j} < p'_{i,j}$. Using (IV.29), we have $p_{i,j}(1-\pi_{i,j}) + (1-p_{i,j})(1-\pi_{j,i}) = p'_{i,j} + (1-2\pi_{i,j})p_{i,j}$, which quantity is minimum when $\pi_{i,j} = 1$ (and in this case $\pi_{j,i} = (1-p'_{i,j})/(1-p_{i,j})$), and then equal to $|p_{i,j} - p'_{i,j}|$. We recall that we can only set $\pi_{i,j} = 1$ if the initial assumption $p_{i,j} < p'_{i,j}$ holds. In a similar fashion, if $p_{i,j} > p'_{i,j}$, we have

$p_{i,j}(1-\pi_{i,j})+(1-p_{i,j})(1-\pi_{j,i}) = 2(1-p_{i,j})(1-\pi_{j,i})+p_{i,j}-p'_{i,j}$, which is minimum for $\pi_{j,i}=1$ (we have incidentally $\pi_{i,j}=p'_{i,j}/p_{i,j}$ in this case) and then equal to $|p_{i,j}-p'_{i,j}|$. Since we clearly have

$$W_{d_\tau,1}\left(P,P'\right) \geq$$
$$\sum_{i<j} \inf_{(\Sigma,\Sigma') \text{ s.t. } \mathbb{P}\{\Sigma(i)<\Sigma(j)\}=p_{i,j} \text{ and } \mathbb{P}\{\Sigma'(i)<\Sigma'(j)\}=p'_{i,j}} \mathbb{P}\left[(\Sigma(i)-\Sigma(j))(\Sigma'(i)-\Sigma'(j))<0\right],$$

this proves that

$$W_{d_\tau,1}\left(P,P'\right) \geq \sum_{i<j} |p'_{i,j}-p_{i,j}|.$$

As a remark, given a distribution P on $\mathfrak{S}_N$, when $P'=P_{\mathcal{C}}$ with $\mathcal{C}$ a bucket order of $[\![N]\!]$ with K buckets, the optimality conditions on the $\pi_{i,j}$'s are fulfilled by the coupling $(\Sigma,\Sigma_{\mathcal{C}})$, which implies that:

$$W_{d_\tau,1}\left(P,P_{\mathcal{C}}\right) = \sum_{i<j} |p'_{i,j}-p_{i,j}| = \sum_{1\leq k<l\leq K} \sum_{(i,j)\in\mathcal{C}_k\times\mathcal{C}_l} p_{j,i}, \tag{IV.30}$$

where $p'_{i,j}=\mathbb{P}_{\Sigma_{\mathcal{C}}\sim P_{\mathcal{C}}}\left[\Sigma_{\mathcal{C}}(i)<\Sigma_{\mathcal{C}}(j)\right] = p_{i,j}\mathbb{I}\{k=l\}+\mathbb{I}\{k<l\}$, with $(k,l)\in\{1,\ldots,K\}^2$ such that $(i,j)\in\mathcal{C}_k\times\mathcal{C}_l$.

Proof of Proposition 1

Let $\mathcal{C}$ be a bucket order of $[\![N]\!]$ with K buckets. Then, for $P'\in\mathbf{P}_{\mathcal{C}}$, Lemma 1 implies that:

$$W_{d_\tau,1}\left(P,P'\right) \geq \sum_{i<j} |p'_{i,j}-p_{i,j}| = \sum_{k=1}^{K} \sum_{i<j,(i,j)\in\mathcal{C}_k^2} |p'_{i,j}-p_{i,j}| + \sum_{1\leq k<l\leq K} \sum_{(i,j)\in\mathcal{C}_k\times\mathcal{C}_l} p_{j,i},$$

where the last equality results from the fact that $p'_{i,j}=1$ when $(i,j)\in\mathcal{C}_k\times\mathcal{C}_l$ with $k<l$. When $P'=P_{\mathcal{C}}$, the intra-bucket terms are all equal to zero. Hence, it results from (IV.30) that :

$$W_{d_\tau,1}\left(P,P_{\mathcal{C}}\right) = \sum_{1\leq k<l\leq K} \sum_{(i,j)\in\mathcal{C}_k\times\mathcal{C}_l} p_{j,i} = \Lambda_P(\mathcal{C}).$$

Proof of Theorem 1

Observe first that the excess of distortion can be bounded as follows:

$$\Lambda_P(\widehat{C}_{K,\lambda}) - \inf_{\mathcal{C}\in\mathbf{C}_K} \Lambda_P(\mathcal{C}) \leq 2 \max_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \left|\widehat{\Lambda}_n(\mathcal{C})-\Lambda_P(\mathcal{C})\right| + \left\{\inf_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C}) - \inf_{\mathcal{C}\in\mathbf{C}_K} \Lambda_P(\mathcal{C})\right\}.$$

By a classical symmetrization device (see e.g. [VdVW]), we have:

$$\mathbb{E}\left[\max_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \left|\widehat{\Lambda}_n(\mathcal{C})-\Lambda_P(\mathcal{C})\right|\right] \leq 2\mathbb{E}\left[\mathcal{R}_n(\lambda)\right]. \tag{IV.31}$$

Hence, using McDiarmid's inequality, for all $\delta \in (0,1)$ it holds with probability at least $1-\delta$:

$$\max_{\mathcal{C}\in\mathbf{C}_{K,\lambda}} \left|\widehat{\Lambda}_n(\mathcal{C}) - \Lambda_P(\mathcal{C})\right| \leq 2\mathbb{E}\left[\mathcal{R}_n(\lambda)\right] + \kappa(\lambda)\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Proof of Theorem 2

Following the proof of Theorem 8.1 in [BBL05], we have for all $m \in \{1,\ldots,M\}$,

$$\begin{aligned}\mathbb{E}\left[\Lambda_P(\widehat{\mathcal{C}}_{K_{\widehat{m}},\lambda_{\widehat{m}}})\right] &\leq \mathbb{E}\left[\Lambda_P(\widehat{\mathcal{C}}_{K_m,\lambda_m})\right] + \mathbb{E}\left[\textsc{pen}(\lambda_m,n)\right] \\ &\quad + \sum_{m'=1}^{M}\mathbb{E}\left[\left(\max_{\mathcal{C}\in\mathbf{C}_{K_{m'},\lambda_{m'}}} \Lambda_P(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C}) - \textsc{pen}(\lambda_{m'},n)\right)_+\right],\end{aligned}$$

where $x_+ = \max(x,0)$ denotes the positive part of x. In addition, for any $\delta > 0$, we have:

$$\begin{aligned}&\mathbb{P}\left\{\max_{\mathcal{C}\in\mathbf{C}_{K_m,\lambda_m}} \Lambda_P(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C}) \geq \textsc{pen}(\lambda_m,n) + \delta\right\} \\ &\quad\leq \mathbb{P}\left\{\max_{\mathcal{C}\in\mathbf{C}_{K_m,\lambda_m}} \Lambda_P(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C}) \geq \mathbb{E}\left[\max_{\mathcal{C}\in\mathbf{C}_{K_m,\lambda_m}} \Lambda_P(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C})\right] + \frac{\delta}{5}\right\} \\ &\qquad + \mathbb{P}\left\{\mathcal{R}_n(\lambda_m) \leq \mathbb{E}\left[\mathcal{R}_n(\lambda_m)\right] - \frac{2}{5}\delta\right\} \leq 2\exp\left(-\frac{2n\delta^2}{25\kappa(\lambda_m)^2}\right),\end{aligned}$$

using (IV.31) for the first inequality, and both McDiarmid's inequality and Lemma 8.2 in [BBL05] for the second inequality. Observing that $\kappa(\lambda) \leq \binom{N}{2}$ and integrating by parts conclude the proof.

Proof of Theorem 3

Consider a bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of shape λ, different from (IV.23). Hence, there exists at least a pair $\{i,j\}$ such that $j \prec_{\mathcal{C}} i$ and $\sigma_P^*(j) < \sigma_P^*(i)$ (or equivalently $p_{i,j} < 1/2$). Consider such a pair $\{i,j\}$. Hence, there exist $1 \leq k < l \leq K$ s.t. $(i,j) \in \mathcal{C}_k \times C_l$. Define the bucket order $\mathcal{C}'$ which is the same as $\mathcal{C}$ except that the buckets of i and j are swapped: $\mathcal{C}'_k = \{j\} \cup \mathcal{C}_k \setminus \{i\}$, $\mathcal{C}'_l = \{i\} \cup \mathcal{C}_l \setminus \{j\}$ and $\mathcal{C}'_m = \mathcal{C}_m$ if $m \in \{1,\ldots,K\} \setminus \{k,l\}$. Observe that

$$\begin{aligned}\Lambda_P(\mathcal{C}') - \Lambda_P(\mathcal{C}) &= p_{i,j} - p_{j,i} + \sum_{a\in\mathcal{C}_k\setminus\{i\}} p_{i,a} - p_{j,a} + \sum_{a\in\mathcal{C}_l\setminus\{j\}} p_{a,j} - p_{a,i} \\ &\quad + \sum_{m=k+1}^{l-1}\sum_{a\in\mathcal{C}_m} p_{a,j} - p_{a,i} + p_{i,a} - p_{j,a} \leq 2(p_{i,j} - 1/2) < 0.\end{aligned}$$

Considering now all the pairs $\{i,j\}$ such that $j \prec_{\mathcal{C}} i$ and $p_{i,j} < 1/2$, it follows by induction that

$$\Lambda_P(\mathcal{C}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)}) \geq 2\sum_{j\prec_{\mathcal{C}} i}(1/2 - p_{i,j}) \cdot \mathbb{I}\{p_{i,j} < 1/2\}. \tag{IV.32}$$

Proof of Theorem 4

The fast rate analysis essentially relies on the following lemma providing a control of the variance of the empirical excess of distortion

$$\widehat{\Lambda}_n(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C}^{*(K,\lambda)}) = \frac{1}{n}\sum_{s=1}^{n}\sum_{i\neq j}\mathbb{I}\{\Sigma_s(j) < \Sigma_s(i)\}\cdot\left(\mathbb{I}\left\{i \prec_{\mathcal{C}} j\right\} - \mathbb{I}\{i \prec_{\mathcal{C}^{*(K,\lambda)}} j\}\right).$$

Set $D(\mathcal{C}) = \sum_{i\neq j}\mathbb{I}\{\Sigma(j) < \Sigma(i)\}\cdot\left(\mathbb{I}\left\{i \prec_{\mathcal{C}} j\right\} - \mathbb{I}\{i \prec_{\mathcal{C}^{*(K,\lambda)}} j\}\right)$. Observe that $\mathbb{E}[D(\mathcal{C})] = \Lambda_P(\mathcal{C}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)})$.

Lemma 3. *Let λ be a given bucket order shape. We have:*

$$var\left(D(\mathcal{C})\right) \leq 2^{\binom{N}{2}}(N^2/h)\cdot\mathbb{E}[D(\mathcal{C})].$$

PROOF. As in the proof of Theorem 3, consider a bucket order $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$ of shape λ, different from (IV.23), a pair $\{i, j\}$ such that there exist $1 \leq k < l \leq K$ s.t. $(i, j) \in \mathcal{C}_k \times C_l$ and $\sigma_P^*(j) < \sigma_P^*(i)$ and the bucket order $\mathcal{C}'$ which is the same as $\mathcal{C}$ except that the buckets of i and j are swapped. We have:

$$\begin{aligned} D(\mathcal{C}')-D(\mathcal{C}) = {} & \mathbb{I}\{\Sigma(i) < \Sigma(j)\}-\mathbb{I}\{\Sigma(j) < \Sigma(i)\}+\sum_{a\in\mathcal{C}_k\setminus\{i\}}\mathbb{I}\{\Sigma(i) < \Sigma(a)\}-\mathbb{I}\{\Sigma(j) < \Sigma(a)\} \\ & + \sum_{a\in\mathcal{C}_l\setminus\{j\}}\mathbb{I}\{\Sigma(a) < \Sigma(j)\} - \mathbb{I}\{\Sigma(a) < \Sigma(i)\} \\ & + \sum_{m=k+1}^{l-1}\sum_{a\in\mathcal{C}_m}\mathbb{I}\{\Sigma(a) < \Sigma(j)\} - \mathbb{I}\{\Sigma(a) < \Sigma(i)\} + \mathbb{I}\{\Sigma(i) < \Sigma(a)\} - \mathbb{I}\{\Sigma(j) < \Sigma(a)\}. \end{aligned}$$

Hence, we have: $var(D(\mathcal{C}') - D(\mathcal{C})) \leq 4N^2$. By induction, we then obtain that:

$$\begin{aligned} var\left(D(\mathcal{C})\right) &\leq 2^{\binom{N}{2}-1}(4N^2)\#\left\{(i,j) : \ i \prec_{\mathcal{C}} j \text{ and } p_{j,i} > 1/2\right\} \\ &\leq 2^{\binom{N}{2}-1}(4N^2/h)\sum_{j\prec_{\mathcal{C}} i}(1/2 - p_{i,j})\cdot\mathbb{I}\{p_{i,j} < 1/2\} \leq 2^{\binom{N}{2}}(N^2/h)\mathbb{E}[D(\mathcal{C})], \end{aligned}$$

by combining (IV.24) with condition (IV.7).

Applying Bernstein's inequality to the i.i.d. average $(1/n)\sum_{s=1}^{n} D_s(\mathcal{C})$, where

$$D_s(\mathcal{C}) = \sum_{i\neq j}\mathbb{I}\{\Sigma_s(j) < \Sigma_s(i)\}\cdot\left(\mathbb{I}\left\{i \prec_{\mathcal{C}} j\right\} - \mathbb{I}\{i \prec_{\mathcal{C}^{*(K,\lambda)}} j\}\right),$$

for $1 \leq s \leq n$ and the union bound over the bucket orders $\mathcal{C}$ in $\mathbf{C}_{K,\lambda}$ (recall that $\#\mathbf{C}_{K,\lambda} = \binom{N}{\lambda}$), we obtain that, for all $\delta \in (0, 1)$, we have with probability larger than

$1-\delta$: $\forall \mathcal{C} \in \mathbf{C}_{K,\lambda}$,

$$\mathbb{E}[D(\mathcal{C})] = \Lambda_P(\mathcal{C}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)}) \leq \widehat{\Lambda}_n(\mathcal{C}) - \widehat{\Lambda}_n(\mathcal{C}^{*(K,\lambda)}) + \sqrt{\frac{2var(D(\mathcal{C}))\log\left(\binom{N}{\lambda}/\delta\right)}{n}} + \frac{4\kappa(\lambda)\log(\binom{N}{\lambda}/\delta)}{3n}.$$

Since $\widehat{\Lambda}_n(\widehat{C}_{K,\lambda}) - \widehat{\Lambda}_n(\mathcal{C}^{*(K,\lambda)}) \leq 0$ by assumption and using the variance control provided by Lemma 3 above, we obtain that, with probability at least $1-\delta$, we have:

$$\Lambda_P(\widehat{C}_{K,\lambda}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)}) \leq \sqrt{\frac{2^{\binom{N}{2}+1}N^2\left(\Lambda_P(\widehat{C}_{K,\lambda}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)})\right)/h \times \log(\binom{N}{\lambda}/\delta)}{n}} + \frac{4\kappa(\lambda)\log(\binom{N}{\lambda}/\delta)}{3n}.$$

Finally, solving this inequality in $\Lambda_P(\widehat{C}_{K,\lambda}) - \Lambda_P(\mathcal{C}^{*(K,\lambda)})$ yields the desired result.

Proof of Theorem 5

Straightforward if $K^* = N$: assume $K^* < N$ in the following.
(i). Existence is ensured by definition of K^* combined with Theorem 3. Assume there exist two distinct K^*-shapes λ and λ' such that $\Lambda_P(\mathcal{C}^{*(K^*,\lambda)}) = \Lambda_P(\mathcal{C}^{*(K^*,\lambda')}) = 0$. Necessarily, there exists $k \in \{1,\ldots,K-1\}$ such that, for example, $\mathcal{C}_k^{*(K^*,\lambda)} \cap \mathcal{C}_{k+1}^{*(K^*,\lambda')} \neq \emptyset$ and $\mathcal{C}_{k+1}^{*(K^*,\lambda')} \not\subseteq \mathcal{C}_k^{*(K^*,\lambda)}$. Then, define a new bucket order $\widetilde{\mathcal{C}}$ of size K^*+1 as follows:

$$\widetilde{\mathcal{C}} = \left(\mathcal{C}_1^{*(K^*,\lambda')},\ldots,\mathcal{C}_k^{*(K^*,\lambda')},\mathcal{C}_k^{*(K^*,\lambda)} \cap \mathcal{C}_{k+1}^{*(K^*,\lambda')}, \mathcal{C}_{k+1}^{*(K^*,\lambda')} \setminus \left(\mathcal{C}_k^{*(K^*,\lambda)} \cap \mathcal{C}_{k+1}^{*(K^*,\lambda')}\right), \mathcal{C}_{k+2}^{*(K^*,\lambda')},\ldots,\mathcal{C}_{K^*}^{*(K^*,\lambda')}\right).$$

Conclude observing that $\Lambda_P(\widetilde{\mathcal{C}}) = 0$ i.e. $P \in \mathbf{P}_{\widetilde{\mathcal{C}}}$, which contradicts the definition of K^*.
(ii). By Theorem 3, any bucket order $\mathcal{C}$ such that $P \in \mathbf{P}_{\mathcal{C}}$ agrees with the Kemeny median. Then, observe that such bucket order $\mathcal{C}$ of size $K < K^*$ is obtained by iteratively merging adjacent buckets of $\mathcal{C}^{*(K^*,\lambda^*)}$: otherwise, following the proof of (i), we could define a new bucket order $\widetilde{\mathcal{C}}$ of size K^*+1 such that $P \in \mathbf{P}_{\widetilde{\mathcal{C}}}$. When $K = K^*-1$, Eq. (IV.27) proves that $d_{\mathcal{C}} > d_{\mathcal{C}^{*(K^*,\lambda^*)}}$. The general result follows by induction.
(iii). By induction on $N-K^* \in \{0,\ldots,N-2\}$. Initialization is straightforward for $K^* = N$. Let $m \in \{3,\ldots,N\}$ and assume that the proposition is true for any strongly/strictly stochastically transitive bucket distribution with $K^* = m$. Let P be a strongly/strictly stochastically transitive bucket distribution with $K^* = m-1$. By definition of K^*,

the algorithm runned with distribution P cannot stop before computing $\mathcal{C}(N-m+1)$, which results from merging the adjacent buckets $\mathcal{C}_k(N-m)$ and $\mathcal{C}_{k+1}(N-m)$ (with $k \in \{1,\ldots,m-1\}$). Then consider a distribution $\widetilde{P}$ with pairwise marginals $\widetilde{p}_{i,j} = 1$ if $(i,j) \in \mathcal{C}_k(N-m) \times \mathcal{C}_{k+1}(N-m)$, $\widetilde{p}_{i,j} = 0$ if $(i,j) \in \mathcal{C}_{k+1}(N-m) \times \mathcal{C}_k(N-m)$ and $\widetilde{p}_{i,j} = p_{i,j}$ otherwise. Hence, $\widetilde{P}$ is a strongly/strictly stochastically transitive bucket distribution and $\mathcal{C}(N-m)$ is, by construction of $\widetilde{P}$, returned by the algorithm when runned with distribution $\widetilde{P}$. Hence by induction hypothesis: $\widetilde{P} \in \mathbf{P}_{\mathcal{C}(N-m)}$. Conclude observing that $\Lambda_P(\mathcal{C}(N-m)) = \Lambda_{\widetilde{P}}(\mathcal{C}(N-m)) + \sum_{i \in \mathcal{C}_k(N-m), j \in \mathcal{C}_{k+1}(N-m)} p_{j,i} = \Delta_P^{(k)}(\mathcal{C}(N-m))$, which implies that $\Lambda_P(\mathcal{C}(N-m+1)) = \Lambda_P(\mathcal{C}(N-m)) - \Delta_P^{(k)}(\mathcal{C}(N-m)) = 0$.

Proof of Lemma 2

We start with proving the first assertion.

(i). Consider a coupling (Σ, Σ') of two probability distributions P and P' on $\mathfrak{S}_N$. Define the triplet-wise probabilities $p_{i,j,k} = \mathbb{P}_{\Sigma \sim P}\{\Sigma(i) < \Sigma(j) < \Sigma(k)\}$ and $p'_{i,j,k} = \mathbb{P}_{\Sigma' \sim P'}\{\Sigma'(i) < \Sigma'(j) < \Sigma'(k)\}$. For clarity's sake, we will assume that $\tilde{p}_{i,j,k} = \min(p_{i,j,k}, p'_{i,j,k}) > 0$ for all triplets (i,j,k), the extension to the general case being straightforward. We also denote $\bar{p}_{i,j,k} = \max(p_{i,j,k}, p'_{i,j,k})$. Given two pairs of three distinct elements of $[\![N]\!]$, (i,j,k) and (a,b,c), we define the following quantities:

$$\begin{aligned}
\pi_{a,b,c|i,j,k} &= \mathbb{P}\left\{\Sigma'(a) < \Sigma'(b) < \Sigma'(c) \mid \Sigma(i) < \Sigma(j) < \Sigma(k)\right\}, \\
\pi'_{a,b,c|i,j,k} &= \mathbb{P}\left\{\Sigma(a) < \Sigma(b) < \Sigma(c) \mid \Sigma'(i) < \Sigma'(j) < \Sigma'(k)\right\}, \\
\tilde{\pi}_{a,b,c|i,j,k} &= \pi_{a,b,c|i,j,k}\mathbb{I}\{p_{i,j,k} \leq p'_{i,j,k}\} + \pi'_{a,b,c|i,j,k}\mathbb{I}\{p_{i,j,k} > p'_{i,j,k}\}, \\
\bar{\pi}_{a,b,c|i,j,k} &= \pi_{a,b,c|i,j,k}\mathbb{I}\{p_{i,j,k} > p'_{i,j,k}\} + \pi'_{a,b,c|i,j,k}\mathbb{I}\{p_{i,j,k} \leq p'_{i,j,k}\}.
\end{aligned}$$

The motivation for defining the $\tilde{\pi}_{a,b,c|i,j,k}$'s is that the coupling condition $\tilde{\pi}_{i,j,k|i,j,k} = 1$, which implies $\bar{\pi}_{i,j,k|i,j,k} = \frac{\tilde{p}_{i,j,k}}{\bar{p}_{i,j,k}}$, is always feasible. By contrast, it necessarily holds that $\pi_{i,j,k|i,j,k} < 1$ (resp. $\pi'_{i,j,k|i,j,k} < 1$) when $p'_{i,j,k} < p_{i,j,k}$ (resp. $p_{i,j,k} < p'_{i,j,k}$). Throughout the proof, the triplets (a,b,c) always correspond to permutations of (i,j,k). Now write:

$$\mathbb{E}\left[d_2\left(\Sigma, \Sigma'\right)^2\right] = \sum_{i=1}^{N} \mathbb{E}[\Sigma(i)^2] + \mathbb{E}[\Sigma'(i)^2] - 2\mathbb{E}[\Sigma(i)\Sigma'(i)],$$

where

$$\mathbb{E}[\Sigma(i)^2] = \mathbb{E}[(1 + \sum_{j \neq i} \mathbb{I}\{\Sigma(j) < \Sigma(i)\})^2] = 1 + \sum_{j \neq i}(N+1)p_{j,i} - \sum_{k \neq i,j} p_{j,i,k}$$

and

$$\begin{aligned}
\mathbb{E}[\Sigma(i)\Sigma'(i)] = 1 + \sum_{j \neq i} p_{j,i} + p'_{j,i} + \mathbb{P}\{\Sigma(j) < \Sigma(i), \Sigma'(j) < \Sigma'(i)\} \\
+ \sum_{k \neq i,j} \mathbb{P}\{\Sigma(j) < \Sigma(i), \Sigma'(k) < \Sigma'(i)\}.
\end{aligned}$$

Hence,

$$\mathbb{E}\left[d_2\left(\Sigma,\Sigma'\right)^2\right] = \sum_{a<b<c}\sum_{(i,j,k)\in\mathfrak{s}(a,b,c)}\frac{1}{N-2}\left\{(N-1)(p_{j,i}+p'_{j,i}) - 2\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(j)<\Sigma'(i)\}\right\}$$
$$- p_{j,i,k} - p'_{j,i,k} - 2\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(k)<\Sigma'(i)\}, \tag{IV.33}$$

where $\mathfrak{s}(a,b,c)$ is the set of the 6 permutations of the triplet (a,b,c). Some terms involved in Eq. (IV.33) can be simplified when summing over $\mathfrak{s}(a,b,c)$, namely:

$$\sum_{(i,j,k)\in\mathfrak{s}(a,b,c)}\frac{N-1}{N-2}(p_{j,i}+p'_{j,i}) - p_{j,i,k} - p'_{j,i,k} = \frac{4N-2}{N-2}.$$

We now simply have:

$$\mathbb{E}\left[d_2\left(\Sigma,\Sigma'\right)^2\right] = \sum_{a<b<c}\frac{4N-2}{N-2} - 2\sum_{(i,j,k)\in\mathfrak{s}(a,b,c)}\frac{1}{N-2}\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(j)<\Sigma'(i)\}$$
$$+\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(k)<\Sigma'(i)\}. \tag{IV.34}$$

Observe that for all triplets (a,b,c) and (i,j,k):

$$\mathbb{P}(\Sigma'(a)<\Sigma'(b)<\Sigma'(c),\Sigma(i)<\Sigma(j)<\Sigma(k))$$
$$+\mathbb{P}(\Sigma'(i)<\Sigma'(j)<\Sigma'(k),\Sigma(a)<\Sigma(b)<\Sigma(c)) = \pi_{a,b,c|i,j,k}p_{i,j,k} + \pi'_{a,b,c|i,j,k}p'_{i,j,k}.$$

Then, by the law of total probability, we have for all distinct $i,\ j,\ k$,

$$\begin{aligned}
&\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(j)<\Sigma'(i)\}\\
&\quad= \frac{1}{2}\{\pi_{j,k,i|j,k,i}p_{j,k,i} + \pi'_{j,k,i|j,k,i}p'_{j,k,i}\}\\
&\quad+ \frac{1}{2}\{\pi_{k,j,i|k,j,i}p_{k,j,i} + \pi'_{k,j,i|k,j,i}p'_{k,j,i}\} + \frac{1}{2}\{\pi_{j,i,k|j,i,k}p_{j,i,k} + \pi'_{j,i,k|j,i,k}p'_{j,i,k}\}\\
&\quad+ \frac{1}{2}\{\pi_{j,i,k|j,k,i}p_{j,k,i} + \pi'_{j,i,k|j,k,i}p'_{j,k,i} + \pi_{j,k,i|j,i,k}p_{j,i,k} + \pi'_{j,k,i|j,i,k}p'_{j,i,k}\}\\
&\quad+ \frac{1}{2}\{\pi_{k,j,i|j,k,i}p_{j,k,i} + \pi'_{k,j,i|j,k,i}p'_{j,k,i} + \pi_{j,k,i|k,j,i}p_{k,j,i} + \pi'_{j,k,i|k,j,i}p'_{k,j,i}\}\\
&\quad+ \frac{1}{2}\{\pi_{j,i,k|k,j,i}p_{k,j,i} + \pi'_{j,i,k|k,j,i}p'_{k,j,i} + \pi_{k,j,i|j,i,k}p_{j,i,k} + \pi'_{k,j,i|j,i,k}p'_{j,i,k}\},
\end{aligned}$$

and

$$
\begin{aligned}
&\mathbb{P}\{\Sigma(j) < \Sigma(i), \Sigma'(k) < \Sigma'(i)\} \\
&\qquad = \frac{1}{2}\{\pi_{j,k,i|j,k,i} p_{j,k,i} + \pi'_{j,k,i|j,k,i} p'_{j,k,i}\} \\
&\qquad + \frac{1}{2}\{\pi_{k,j,i|k,j,i} p_{k,j,i} + \pi'_{k,j,i|k,j,i} p'_{k,j,i}\} \\
&\qquad + \frac{1}{2}\{\pi_{k,j,i|j,k,i} p_{j,k,i} + \pi'_{k,j,i|j,k,i} p'_{j,k,i} + \pi_{j,k,i|k,j,i} p_{k,j,i} + \pi'_{j,k,i|k,j,i} p'_{k,j,i}\} \\
&\qquad + \mathbb{P}\{\Sigma'(j) < \Sigma'(k) < \Sigma'(i), \Sigma(j) < \Sigma(i) < \Sigma(k)\} \\
&\qquad + \mathbb{P}\{\Sigma'(k) < \Sigma'(i) < \Sigma'(j), \Sigma(j) < \Sigma(k) < \Sigma(i)\} \\
&\qquad + \mathbb{P}\{\Sigma'(k) < \Sigma'(j) < \Sigma'(i), \Sigma(j) < \Sigma(i) < \Sigma(k)\} \\
&\qquad + \mathbb{P}\{\Sigma'(k) < \Sigma'(i) < \Sigma'(j), \Sigma(k) < \Sigma(j) < \Sigma(i)\} \\
&\qquad + \mathbb{P}\{\Sigma'(k) < \Sigma'(i) < \Sigma'(j), \Sigma(j) < \Sigma(i) < \Sigma(k)\},
\end{aligned}
$$

which implies:

$$
\begin{aligned}
&\mathbb{P}\{\Sigma(j) < \Sigma(i), \Sigma'(k) < \Sigma'(i)\} + \mathbb{P}\{\Sigma(k) < \Sigma(i), \Sigma'(j) < \Sigma'(i)\} \\
&\qquad = \pi_{j,k,i|j,k,i} p_{j,k,i} + \pi'_{j,k,i|j,k,i} p'_{j,k,i} + \pi_{k,j,i|k,j,i} p_{k,j,i} + \pi'_{k,j,i|k,j,i} p'_{k,j,i} \\
&\qquad + \pi_{k,j,i|j,k,i} p_{j,k,i} + \pi'_{k,j,i|j,k,i} p'_{j,k,i} + \pi_{j,k,i|k,j,i} p_{k,j,i} + \pi'_{j,k,i|k,j,i} p'_{k,j,i} \\
&\qquad + \frac{1}{2}\left\{\pi_{j,k,i|j,i,k} p_{j,i,k} + \pi'_{j,k,i|j,i,k} p'_{j,i,k} + \pi_{j,i,k|j,k,i} p_{j,k,i} + \pi'_{j,i,k|j,k,i} p'_{j,k,i}\right\} \\
&\qquad + \frac{1}{2}\left\{\pi_{k,i,j|j,k,i} p_{j,k,i} + \pi'_{k,i,j|j,k,i} p'_{j,k,i} + \pi_{j,k,i|k,i,j} p_{k,i,j} + \pi'_{j,k,i|k,i,j} p'_{k,i,j}\right\} \\
&\qquad + \frac{1}{2}\left\{\pi_{k,j,i|j,i,k} p_{j,i,k} + \pi'_{k,j,i|j,i,k} p'_{j,i,k} + \pi_{j,i,k|k,j,i} p_{k,j,i} + \pi'_{j,i,k|k,j,i} p'_{k,j,i}\right\} \\
&\qquad + \frac{1}{2}\left\{\pi_{k,i,j|k,j,i} p_{k,j,i} + \pi'_{k,i,j|k,j,i} p'_{k,j,i} + \pi_{k,j,i|k,i,j} p_{k,i,j} + \pi'_{k,j,i|k,i,j} p'_{k,i,j}\right\} \\
&\qquad + \frac{1}{2}\left\{\pi_{k,i,j|j,i,k} p_{j,i,k} + \pi'_{k,i,j|j,i,k} p'_{j,i,k} + \pi_{j,i,k|k,i,j} p_{k,i,j} + \pi'_{j,i,k|k,i,j} p'_{k,i,j}\right\},
\end{aligned}
$$

which is invariant under permutation of the indices j and k. Hence,

$$
\begin{aligned}
H(a,b,c) &= \sum_{(i,j,k)\in\mathfrak{s}(a,b,c)} \frac{1}{N-2}\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(j)<\Sigma'(i)\}+\mathbb{P}\{\Sigma(j)<\Sigma(i),\Sigma'(k)<\Sigma'(i)\} \\
&= \sum_{(i,j,k)\in\mathfrak{s}(a,b,c)} \left\{ \frac{2N-1}{2(N-2)}\tilde{\pi}_{j,k,i|j,k,i} + \frac{N-1}{N-2}(\tilde{\pi}_{k,j,i|j,k,i}+\tilde{\pi}_{j,i,k|j,k,i}) \right. \\
&\qquad \left. + \frac{N-1}{2(N-2)}(\tilde{\pi}_{k,i,j|j,k,i}+\tilde{\pi}_{i,j,k|j,k,i}) + \frac{1}{2}\tilde{\pi}_{i,k,j|j,k,i} \right\}\tilde{p}_{j,k,i} \\
&+ \left\{ \frac{2N-1}{2(N-2)}\bar{\pi}_{j,k,i|j,k,i} + \frac{N-1}{N-2}(\bar{\pi}_{k,j,i|j,k,i}+\bar{\pi}_{j,i,k|j,k,i}) \right. \\
&\qquad \left. + \frac{N-1}{2(N-2)}(\bar{\pi}_{k,i,j|j,k,i}+\bar{\pi}_{i,j,k|j,k,i}) + \frac{1}{2}\bar{\pi}_{i,k,j|j,k,i} \right\}\bar{p}_{j,k,i},
\end{aligned}
\tag{IV.35}
$$

which is maximum when $\tilde{\pi}_{j,k,i|j,k,i}=1$ (which implies $\bar{\pi}_{j,k,i|j,k,i}=\frac{\tilde{p}_{j,k,i}}{\bar{p}_{j,k,i}}$) and $\bar{\pi}_{k,j,i|j,k,i}+\bar{\pi}_{j,i,k|j,k,i}=1-\frac{\tilde{p}_{j,k,i}}{\bar{p}_{j,k,i}}$ for all $(i,j,k)\in\mathfrak{s}(a,b,c)$ and then verifies:

$$
\begin{aligned}
H(a,b,c) &\leq \sum_{(i,j,k)\in\mathfrak{s}(a,b,c)} \frac{N}{N-2}\tilde{p}_{i,j,k} + \frac{N-1}{N-2}\bar{p}_{i,j,k} \\
&= \frac{1}{N-2}\sum_{(i,j,k)\in\mathfrak{s}(a,b,c)} N(p_{i,j,k}+p'_{i,j,k}) - \bar{p}_{i,j,k} \\
&= \frac{1}{N-2}\left\{ 2N - \sum_{(i,j,k)\in\mathfrak{s}(a,b,c)} \bar{p}_{i,j,k} \right\},
\end{aligned}
\tag{IV.36}
$$

which concludes the first part of the proof.

(ii). Now consider the particular case $P'\in\mathbf{P}_{\mathcal{C}}$, with $\mathcal{C}$ a bucket order of $[\![N]\!]$ with K buckets. We propose to prove that $\min_{P'\in\mathbf{P}_{\mathcal{C}}} W_{d_2,2}(P,P') = W_{d_2,2}(P,P_{\mathcal{C}}) = \mathbb{E}[d_2^2(\Sigma,\Sigma_{\mathcal{C}})]$ and to obtain an explicit expression. Given three distinct indices $(a,b,c)\in[\![N]\!]^3$, we consider the following four possible scenarios.

Case 1: $(a,b,c)\in\mathcal{C}_q^3$ are in the same bucket. The maximizing conditions for $H(a,b,c)$ in Eq. (IV.35) are $\tilde{\pi}_{j,k,i|j,k,i}=1$ and $\bar{\pi}_{k,j,i|j,k,i}+\bar{\pi}_{j,i,k|j,k,i}=1-\frac{\tilde{p}_{j,k,i}}{\bar{p}_{j,k,i}}$ for all $(i,j,k)\in\mathfrak{s}(a,b,c)$. All are verified when $\Sigma'=\Sigma_{\mathcal{C}}$ as $\Sigma(i)<\Sigma(j)<\Sigma(k)$ iff $\Sigma_{\mathcal{C}}(i)<\Sigma_{\mathcal{C}}(j)<\Sigma_{\mathcal{C}}(k)$. Hence:

$$
H(a,b,c)\leq\frac{2N-1}{N-2},
$$

with equality when $\Sigma'=\Sigma_{\mathcal{C}}$.

Case 2: $(a,b,c) \in \mathcal{C}_q \times \mathcal{C}_r \times \mathcal{C}_s$ are in three different buckets (e.g. $q < r < s$). For all $(j,k,i) \in \mathfrak{s}(a,b,c) \setminus \{(a,b,c)\}$, $p'_{j,k,i} = \tilde{p}_{j,k,i} = 0$. Hence, $H(a,b,c)$ writes without the terms related to the five impossible events $\Sigma'(j) < \Sigma'(k) < \Sigma'(i)$. Moreover, $\bar{p}_{j,k,i} = p_{j,k,i}$ and $\bar{\pi}_{a,b,c|j,k,i} = 1$ so the sum of the corresponding contributions in $H(a,b,c)$ is:

$$\frac{N-1}{N-2}(p_{b,a,c} + p_{a,c,b}) + \frac{N-1}{2(N-2)}(p_{b,c,a} + p_{c,a,b}) + \frac{1}{2}p_{c,b,a}. \tag{IV.37}$$

We have $p_{a,b,c} \leq p'_{a,b,c} = 1$ so $\tilde{\pi}_{a,b,c|a,b,c} = 1$ and for all $(i,j,k) \in \mathfrak{s}(a,b,c)$, $\bar{\pi}_{i,j,k|a,b,c} = p_{i,j,k}$. The sum of the corresponding contributions in $H(a,b,c)$ is:

$$\frac{2N-1}{N-2}p_{a,b,c} + \frac{N-1}{N-2}(p_{b,a,c} + p_{a,c,b}) + \frac{N-1}{2(N-2)}(p_{b,c,a} + p_{c,a,b}) + \frac{1}{2}p_{c,b,a}. \tag{IV.38}$$

Finally, by summing expressions (IV.37) and (IV.38),

$$H(a,b,c) = \frac{2N-1}{N-2}p_{a,b,c} + \frac{2(N-1)}{N-2}(p_{b,a,c} + p_{a,c,b}) + \frac{N-1}{N-2}(p_{b,c,a} + p_{c,a,b}) + p_{c,b,a}.$$

Case 3: $(a,b,c) \in \mathcal{C}_q \times \mathcal{C}_r \times \mathcal{C}_r$ are in two different buckets such that one item (here a) is ranked first among the triplet (i.e. $q < r$). For all $(j,k,i) \in \mathfrak{s}(a,b,c) \setminus \{(a,b,c),(a,c,b)\}$, $p'_{j,k,i} = \tilde{p}_{j,k,i} = 0$. Hence, $H(a,b,c)$ writes without the terms related to the four impossible events $\Sigma'(j) < \Sigma'(k) < \Sigma'(i)$. For all $(j,k,i) \in \mathfrak{s}(a,b,c)$, $\pi_{a,b,c|j,k,i} + \pi_{a,c,b|j,k,i} = 1$ and the sum of the corresponding contributions in $H(a,b,c)$ is:

$$\begin{aligned}
&\left(\frac{2N-1}{2(N-2)}\pi_{a,b,c|a,b,c} + \frac{N-1}{N-2}\pi_{a,c,b|a,b,c}\right)p_{a,b,c} + \left(\frac{2N-1}{2(N-2)}\pi_{a,c,b|a,c,b} + \frac{N-1}{N-2}\pi_{a,b,c|a,c,b}\right)p_{a,c,b} \\
&+ \left(\frac{N-1}{2(N-2)}\pi_{a,b,c|b,c,a} + \frac{1}{2}\pi_{a,c,b|b,c,a}\right)p_{b,c,a} + \left(\frac{N-1}{N-2}\pi_{a,b,c|b,a,c} + \frac{N-1}{2(N-2)}\pi_{a,c,b|b,a,c}\right)p_{b,a,c} \\
&+ \left(\frac{N-1}{2(N-2)}\pi_{a,c,b|c,b,a} + \frac{1}{2}\pi_{a,b,c|c,b,a}\right)p_{c,b,a} + \left(\frac{N-1}{N-2}\pi_{a,c,b|c,a,b} + \frac{N-1}{2(N-2)}\pi_{a,b,c|c,a,b}\right)p_{c,a,b}.
\end{aligned} \tag{IV.39}$$

Observe that the expression above is maximum when $\pi_{a,b,c|a,b,c} = \pi_{a,c,b|a,c,b} = \pi_{a,b,c|b,c,a} = \pi_{a,b,c|b,a,c} = \pi_{a,c,b|c,b,a} = \pi_{a,c,b|c,a,b} = 1$, which is verified if $\Sigma' = \Sigma_{\mathcal{C}}$. In this case, (IV.39) writes:

$$\frac{2N-1}{2(N-2)}(p_{a,b,c} + p_{a,c,b}) + \frac{N-1}{N-2}(p_{b,a,c} + p_{c,a,b}) + \frac{N-1}{2(N-2)}(p_{b,c,a} + p_{c,b,a}). \tag{IV.40}$$

Now consider $(j,k,i) \in \{(a,b,c),(a,c,b)\}$: $p'_{a,b,c} + p'_{a,c,b} = 1$ and the corresponding

contributions in $H(a,b,c)$ sum as follows:

$$\begin{aligned}
&\left\{ \frac{2N-1}{2(N-2)}\pi'_{a,b,c|a,b,c} + \frac{N-1}{N-2}(\pi'_{b,a,c|a,b,c} + \pi'_{a,c,b|a,b,c}) \right. \\
&\left. + \frac{N-1}{2(N-2)}(\pi'_{b,c,a|a,b,c} + \pi'_{c,a,b|a,b,c}) + \frac{1}{2}\pi'_{c,b,a|a,b,c} \right\} p'_{a,b,c} \\
&+ \left\{ \frac{2N-1}{2(N-2)}\pi'_{a,c,b|a,c,b} + \frac{N-1}{N-2}(\pi'_{c,a,b|a,c,b} + \pi'_{a,b,c|a,c,b}) \right. \\
&\left. + \frac{N-1}{2(N-2)}(\pi'_{c,b,a|a,c,b} + \pi'_{b,a,c|a,c,b}) + \frac{1}{2}\pi'_{b,c,a|a,c,b} \right\} p'_{a,c,b},
\end{aligned}$$

which is maximum when $\pi'_{a,c,b|a,b,c} = \pi'_{c,a,b|a,b,c} = \pi'_{c,b,a|a,b,c} = 0$ and $\pi'_{a,b,c|a,c,b} = \pi'_{b,a,c|a,c,b} = \pi'_{b,c,a|a,c,b} = 0$: both conditions are verified for $\Sigma' = \Sigma_{\mathcal{C}}$. Then, the expression above is upper bounded by:

$$\frac{2N-1}{2(N-2)}(p_{a,b,c} + p_{a,c,b}) + \frac{N-1}{N-2}(p_{b,a,c} + p_{c,a,b}) + \frac{N-1}{2(N-2)}(p_{b,c,a} + p_{c,b,a}), \qquad \text{(IV.41)}$$

with equality when $\Sigma' = \Sigma_{\mathcal{C}}$. Finally, by summing (IV.40) and (IV.41),

$$H(a,b,c) \leq \frac{2N-1}{N-2}(p_{a,b,c} + p_{a,c,b}) + \frac{2(N-1)}{N-2}(p_{b,a,c} + p_{c,a,b}) + \frac{N-1}{N-2}(p_{b,c,a} + p_{c,b,a}),$$

with equality when $\Sigma' = \Sigma_{\mathcal{C}}$.

Case 4: $(a,b,c) \in \mathcal{C}_q \times \mathcal{C}_q \times \mathcal{C}_r$ are in two different buckets such that one item (here c) is ranked last among the triplet (i.e. $q < r$). By symmetry with the previous situation, we obtain:

$$H(a,b,c) \leq \frac{2N-1}{N-2}(p_{a,b,c} + p_{b,a,c}) + \frac{2(N-1)}{N-2}(p_{a,c,b} + p_{b,c,a}) + \frac{N-1}{N-2}(p_{c,a,b} + p_{c,b,a}),$$

with equality when $\Sigma' = \Sigma_{\mathcal{C}}$.

Part 2

Risk-Aware Reinforcement Learning

"Quand tu joues au go, faut être *aware*...
Si t'es pas *aware*, tes pierres sont mortes, et toi avec."
Jean-Claude Van Damme

CHAPTER V

Profitable Bandits

Abstract

Originally motivated by default risk management applications, this chapter investigates a novel problem, referred to as the *profitable bandit problem* here. At each step, an agent chooses a subset of the $K \geq 1$ possible actions. For each action chosen, she then respectively pays and receives the sum of a random number of costs and rewards. Her objective is to maximize her cumulated profit. We adapt and study three well-known strategies in this purpose, that were proved to be most efficient in other settings: kl-UCB, Bayes-UCB and Thompson Sampling. For each of them, we prove a finite time regret bound which, together with a lower bound we obtain as well, establishes asymptotic optimality in some cases. Our goal is also to *compare* these three strategies from a theoretical and empirical perspective both at the same time. We give simple, self-contained proofs that emphasize their similarities, as well as their differences. While both Bayesian strategies are automatically adapted to the geometry of information, the numerical experiments carried out show a slight advantage for Thompson Sampling in practice.

1 Introduction

Before providing a general formulation of the profitable bandits problem, we first motivate it with a credit risk management application.

1.1 Motivation

A general and well-known problem for lenders and investors is to choose which prospective clients they should grant loans to, so as to manage credit risk and maximize their profit. A classical supervised learning approach, referred to as *credit risk scoring* consists in ranking all the possible profiles of potential clients, viewed through a collection of socio-economic features Z by means of a (real valued) scoring rule $s(Z)$: ideally, the higher the score $s(Z)$, the higher the default probability. A wide variety of learning algorithms have been proposed to build, from a historical database, a scoring function optimizing ranking performance measures such as the ROC curve or its summary, the

AUC criterion, see *e.g.* [Wes00], [Tho00], [LYW+04], [Yan07] or [CF04]: the *credit risk screening* process then consists in selecting the prospects whose score is below a certain threshold. However, this approach has a serious drawback in general, insofar as new scoring rules are often constructed from truncated information only, namely historical data (the input features X and the observed debt payment behavior) corresponding to past clients, eligible prospects who have been selected by means of a previous scoring rule, jeopardizing thus the screening procedure when applied to prospects who would have been previously non eligible for credit. Hence, the credit risk problem leads to an exploration vs exploitation dilemma there is no way around for: should clients be used for improving the credit risk estimates, or should they be treated according to the level of risk estimated when they arrive? Lenders thus need sequential strategies able to solve this dilemma.

For simplicity, here we consider the very stylized situation, where each individual from a given category applies for a loan of the same amount in expectation. Extension of the general ideas developed in this chapter to more realistic situations will be the subject of further research. In this chapter, we propose a mathematical model that addresses this issue. We propose several strategies, prove their optimality (by giving a lower bound on the inefficiency of any *uniformly efficient* strategy, together with tight regret analyses) and empirically compare their performance in numerical experiments.

1.2 Model

We assume that the population (of credit applicants) is stratified according to $K \geq 1$ categories $a \in \{1, \dots, K\}$. For each category a, the credit risk is modelled by a probability distribution ν_a. We assume that at each step $t \in \{1, \dots, T\}$, where T denotes the total number of time steps (or time horizon), the agent is presented a random number $n_a(t) \geq 1$ of clients of each category a. She must choose a subset $\mathcal{A}_t \subset \{1, \dots, K\}$ of categories to which they grant the loans. We denote by $X_{a,c,t} - L_{a,c,t}$ the profit brought by the client number c of category a at step t, $L_{a,c,t}$ being the loan amount and $X_{a,c,t}$ the corresponding reimbursement. In addition, we assume that all loans $L_{a,c,t}$ for the same category a have the same known expectation τ_a. We assume that the variables $\{X_{a,c,t}\}$ are independent, and that $X_{a,c,t}$ has distribution ν_a and expectation μ_a. We further assume that for any category $a \in \{1, \dots, K\}$, the $n_a(t)$'s are bounded i.e. there exist two positive integers $(n_a^-, n_a^+) \in \mathbb{N}^{*2}$ such that: $n_a^- \leq n_a(t) \leq n_a^+$ for all $t \geq 1$.

Here and throughout, a *sequential strategy* is a set of mappings specifying for each t which categories to choose at time t given the past observations only. In other words, denoting by $I_t = \big(X_{a,c,s}, n_a(s)\big)_{1 \leq s \leq t, a \in \mathcal{A}_s, 1 \leq c \leq n_a(s)}$ the vector of variables observed up to time $t \geq 1$, a strategy specifies a sequence $(\mathcal{A}_t)_{t \geq 1}$ of random subsets such that, for each $t \geq 2$, $\mathcal{A}_t$ is $\sigma(I_{t-1})$-measurable.

It is the goal pursued in this work to define a strategy maximizing the expected

cumulated profit given by

$$S_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\mathbb{I}\{a \in \mathcal{A}_t\}\sum_{c=1}^{n_a(t)} X_{a,c,t} - L_{a,c,t}\right].$$

This is equivalent to minimizing the *expected regret*

$$\begin{aligned} R_T &= \sum_{a\in\mathcal{A}^*} \Delta_a \widetilde{N}_a(T) - S_T \\ &= \sum_{a\in\mathcal{A}^*} \Delta_a \left(\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]\right) + \sum_{a\notin\mathcal{A}^*} |\Delta_a|\mathbb{E}[N_a(T)], \end{aligned}$$

where $\widetilde{N}_a(T) = \mathbb{E}\left[\sum_{t=1}^{T} n_a(t)\right]$ is the expected total number of clients from category a over the T rounds, $N_a(t) = \sum_{s=1}^{t} n_a(s)\mathbb{I}\{a \in \mathcal{A}_s\}$ is the number of observations obtained from category a up to time $t \geq 1$, $\Delta_a = \mu_a - \tau_a$ is the (unknown) expected profit provided by a client of category a and $\mathcal{A}^* = \left\{a \in \{1,\dots,K\}, \Delta_a > 0\right\}$ is the set of profitable categories.

1.3 Illustrative Example

Let us consider the credit risk problem in which a bank wants to identify categories of the population they should accept to loan. It may be naturally formulated as a bandit problem with K arms representing the K categories of the population considered. The bank pays τ_a when loaning to any member of some category $a \in \{1,\dots,K\}$. Each client $c \in \{1,\dots,n_a(t)\}$ of category a receiving a loan from the bank at time step t is characterized by her capacity to reimburse it, namely the Bernoulli r.v. $B_{a,c,t} \sim \mathcal{B}(p_a)$ with $p_a \in [0,1]$:

- $B_{a,c,t} = 0$ in case of credit default, occurs with probability $1 - p_a$: the bank gets no refunding,
- $B_{a,c,t} = 1$ otherwise, occurs with probability p_a: the bank gets refunded $(1+\rho_a)\tau_a$ with τ_a the loan amount and ρ_a the interest rate.

All individuals from the same category are considered as independent i.e. the $B_{a,c,t}$'s are i.i.d. realizations of $\mathcal{B}(p_a)$. Hence the refunding $X_{a,c,t}$ received by the bank writes as follows: $X_{a,c,t} = (1+\rho_a)\tau_a B_{a,c,t}$. Therefore the bank should accept to loan to people belonging to all categories $a \in \{1,\dots,K\}$ such that $\mathbb{E}[X_{a,1,1}] > \tau_a$. This condition rewrites:

$$p_a > \frac{1}{1+\rho_a}. \tag{V.1}$$

Hence the role of the bank is to sequentially identify categories verifying Eq. (V.1) in order to maximize its cumulative profit over the T rounds.

1.4 State of the Art

In the multi-armed bandit (MAB) problem, a learner has to sequentially explore and exploit different sources in order to maximize the cumulative gain. In the stochastic setting, each source (or *arm*) is associated with a distribution generating random rewards. The optimal strategy in hindsight then consists in always pulling the arm with highest expectation. Many approaches have been proposed for solving this problem such as the UCB1 algorithm ([ACBF02]) for bounded rewards or the Thompson Sampling heuristic first proposed in [Tho33]. More recently many algorithms have been proven to be asymptotically optimal, particularly in the case of exponential family distributions, such as kl-UCB ([GC11]), Bayes-UCB ([Kau16]) and Thompson Sampling ([KKM12], [KKM13]). In this chapter we consider a variation of the MAB problem, where, at each time step, the learner may pull several arms simultaneously or no arm at all. To each arm is associated a known threshold and the goal is to maximize the cumulative profit which sums, for each arm pulled by the learner, the difference between the mean reward and the corresponding threshold. This threshold is typically the price to pay for observing a reward from a given arm, e.g. a coin that has to be inserted in a slot machine. Here the optimal strategy consists in always pulling the arms whose expectations are above their respective thresholds. The case where all arms share the same threshold is studied in [RSL17] with a different definition of regret, which only penalizes pulls of non-profitable arms and hence do not refer to the notion of profit. A similar problem has been tackled in [LGC16] in a best arm identification setting with fixed time horizon and for a unique threshold, where rate-optimal strategies are studied. The purpose of this chapter is however different, and we argue that the strategies proposed here are more relevant in many applications (e.g. bank loan management, see Section 1.1).

Indeed, in this chapter we mainly focus on deriving asymptotically optimal strategies in the case of one-dimensional exponential family distributions. Section 2 contains an asymptotic lower bound for the profitable bandit problem for any *uniformly efficient* policy. The three following sections (respectively 4, 5 and 6) are devoted to the adaptation of three celebrated MAB strategies (respectively kl-UCB, Bayes-UCB and Thompson Sampling) to the present problem. We provide in each case a finite-time regret analysis. Asymptotical optimality properties of these algorithms are discussed in Section 7. The final Section 8 contains an empirical comparison of the three strategies through numerical experiments.

2 Lower Bound

The goal of this section is to give an asymptotic lower bound on the expected regret of any *uniformly efficient* strategy. In this purpose, we adapt the argument of [LR85], rewritten by [GMS16], on asymptotic lower bounds for the expected regret in MAB problems. First we define the models $\mathcal{D}_1, \dots, \mathcal{D}_K$ where, for any arm $a \in \{1, \dots, K\}$, $\mathcal{D}_a$ is the set of possible distributions ν_a. Then, we introduce the class of *uniformy efficient* policies that we focus on.

Definition 1. *A strategy is uniformly efficient if, for any profitable bandit problem* $(\nu_a, \tau_a)_{1\leq a\leq K} \in \prod_{a=1}^{K} \mathcal{D}_a \times \mathbb{R}$*, it satisfies for all arms* $a \in \{1, \dots, K\}$ *and for all* $\alpha \in]0, 1]$*,* $\mathbb{E}[N_a(T)] = o(\widetilde{N}_a(T)^\alpha)$ *if* $\mu_a < \tau_a$ *or* $\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] = o(\widetilde{N}_a(T)^\alpha)$ *if* $\mu_a > \tau_a$*.*

Now we can state our lower bound which applies to these strategies.

Theorem 1. *For all models* $\mathcal{D}_1, \dots, \mathcal{D}_K$*, for all uniformly efficient strategies, for all profitable bandit problems* $(\nu_a, \tau_a)_{1\leq a\leq K} \in \prod_{a=1}^{K} \mathcal{D}_a \times \mathbb{R}$*, for all non-profitable arms* a *such that* $\mu_a < \tau_a$*,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a)},$$

where $\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a) = \inf\{KL(\nu_a, \nu'_a), \nu'_a \in \mathcal{D}_a, \mu'_a > \tau_a\}$ *with* $KL(\nu_a, \nu'_a)$ *the Kullback-Leibler divergence between distributions* ν_a *and* ν'_a *and* μ'_a *the expectation of* ν'_a*.*

In the remainder of the chapter, we mainly focus on proposing asymptotically optimal strategies inspired by classical algorithms for MAB, namely KL-UCB ([GC11] and [CGM+13]), BAYES-UCB ([Kau16]) and THOMPSON SAMPLING ([KKM12] and [KKM13]). For each policy, we prove a corresponding upper bound on its expected regret which will be hopefully tight with respect to the lower bound stated above.

3 Preliminaries

This section provides preliminary remarks and notions required by the subsequent analysis.

3.1 Comparison with the Classical Bandit Framework

We briefly compare our setting with some usual MAB conventions.

One-Armed Problems. We point out that the objective of a profitable bandit problem, characterized by K pairs of reward distributions and thresholds $\{(\nu_1, \tau_1), \dots, (\nu_K, \tau_K)\}$, can be equivalently reformulated as simultaneously solving K independent instances of one-armed subproblems: $\{(\nu_1, \tau_1)\}$, $\dots$, $\{(\nu_K, \tau_K)\}$. In other words, we could without loss of generality only consider one-armed instances of the profitable bandit problem i.e. the case $K = 1$. Nevertheless, we will still write this chapter in the general case $K \geq 1$ in order to refer to MAB notations and to our main motivating application, credit risk, which naturally formulates with several categories. As a consequence of this 'separation' property, the theoretical guarantees on the expected regret that we provide for different policies come with simpler proofs than in MAB: the proofs proposed in this chapter contain all core ideas of regret analyses of some of the most sucessful bandit strategies (THOMPSON SAMPLING, BAYES-UCB and KL-UCB) with a somewhat simpler and thus more accessible setting.

Per Round Numbers of Observations. Another difference with the classical MAB model (where at each round $t \geq 1$ the learner observes only one reward drawn from pulled arm a) is that we consider here a more general setting where a random number $n_a(t)$ of

i.i.d. rewards sampled from ν_a are observed. On the other hand, a multiplicative constant (larger than or equal to 1) appears in the upper bounds on the expected regret that we propose for different policies and some parts of their proofs become more intricate.

3.2 One-Dimensional Exponential Family

We consider arms with distributions belonging to a one-dimensional exponential family. It should be noted that the KL-UCB-4P algorithm presented next, as KL-UCB, can be shown to apply to the non-parametric setting of bounded distributions, although the resulting approach has weaker optimality properties (see Section 4.1).

Definitions and Properties. A one-dimensional canonical exponential family is a set of probability distributions $\mathcal{P}_\Theta = \{\nu_\theta, \theta \in \Theta\}$ indexed by a *natural parameter* θ living in the parameter space $\Theta =]\theta^-, \theta^+[\subseteq \mathbb{R}$ and where for all $\theta \in \Theta$, ν_θ has a density $f_\theta(x) = A(x)\exp(G(x)\theta - F(\theta))$ with respect to a reference measure ξ. $A(x)$ and the sufficient statistic $G(x)$ are functions that characterize the exponential family and $F(\theta) = \log \int A(x)\exp(G(x)\theta)d\xi(x)$ is the normalization function. For notational simplicity, we only consider families with $G(x) = x$, which includes many usual distributions (*e.g.* normal, Bernoulli, gamma among others) but not heavy-tailed distributions, commonly used in financial models, such as Pareto ($G(x) = \log(x)$) or Weibull ($G(x) = x^\ell$ with $\ell > 0$). Nevertheless generalizing all the results proved in this chapter to a general sufficient statistic $G(x)$ is straightforward and boils down to considering empirical sufficient statistics $\hat{g}(n) = (1/n)\sum_{s=1}^n G(X_s)$ instead of empirical means. We additionally assume that F is twice differentiable with a continuous second derivative (classic assumption, see e.g. [Was13]) which implies that $\mu : \theta \mapsto \mathbb{E}_{X\sim\nu_\theta}[X]$ is strictly increasing and thus one-to-one in θ. We denote $\mu^- = \mu(\theta^-)$ and $\mu^+ = \mu(\theta^+)$. The Kullback-Leibler divergence between two distributions ν_θ and $\nu_{\theta'}$ in the same exponential family admits the following closed form expression as a function of the natural parameters θ and θ':

$$K(\theta, \theta') := KL(\nu_\theta, \nu_{\theta'}) = F(\theta') - [F(\theta) + F'(\theta)(\theta' - \theta)].$$

We also introduce the KL-divergence between two distributions $\nu_{\mu^{-1}(x)}$ and $\nu_{\mu^{-1}(y)}$:

$$\begin{aligned} d(x, y) :&= K(\mu^{-1}(x), \mu^{-1}(y)) \\ &= \sup_\lambda \{\lambda x - \log \mathbb{E}_{\mu^{-1}(y)}[\exp(\lambda X)]\}, \end{aligned} \tag{V.2}$$

where the last equality comes from the proof of Lemma 3 in [KKM13]. This last expression of d allows to build a confidence interval on x based on a fixed number of i.i.d. samples from $\nu_{\mu^{-1}(x)}$ by applying the Cramér-Chernoff method (see e.g. [BLM13]).

Examples. In Table V.1 we recall some usual examples of one-dimensional exponential families. For some of these distributions that are characterized by two parameters (namely normal, gamma, Pareto and Weibull), one of the two parameters is fixed to define one-dimensional families.

We mainly investigate the profitable bandit problem in the parametric setting, where all distributions $\{\nu_{\theta_a}\}_{1\leq a\leq K}$ belong to a known one-dimensional canonical exponential family $\mathcal{P}_\Theta$ as defined above.

Distribution	Density	Parameter θ
Bernoulli $\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x}\mathbb{I}\{x \in \{0,1\}\}$	$\log\left(\frac{\lambda}{1-\lambda}\right)$
Normal $\mathcal{N}(\lambda, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$	$\frac{\lambda}{\sigma^2}$
Gamma $\Gamma(k, \lambda)$	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}\mathbb{I}\{x \geq 0\}$	$-\lambda$
Poisson $\mathcal{P}(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}\mathbb{I}\{x \in \mathbb{N}\}$	$\log(\lambda)$
Pareto(x_m, λ)	$\frac{\lambda x_m^\lambda}{x^{\lambda+1}}\mathbb{I}\{x \geq x_m\}$	$-\lambda - 1$
Weibull(ℓ, λ)	$\ell\lambda(x\lambda)^{\ell-1}e^{-(\lambda x)^\ell}\mathbb{I}\{x \geq 0\}$	$-\lambda^\ell$

Table V.1: Usual examples of one-dimensional exponential families (parameters σ^2, k, x_m and ℓ are fixed).

3.3 Index Policies

All bandit strategies considered in this chapter are *index policies*: they are fully characterized by an index $u_a(t)$ which is computed at each round $t \geq 1$ for each arm separately; only arms with an index larger than the threshold τ_a are chosen. Index policies are formally described in Algorithm 2.

4 The KL-UCB-4P Algorithm

We introduce the KL-UCB-4P algorithm, '4P' meaning 'for profit', as a variant of the UCB1 algorithm ([ACBF02]) and more precisely of its improvement KL-UCB introduced in [GC11]. It is defined by the index

$$u_a(t) \quad = \quad \sup\left\{q \quad > \quad \hat{\mu}_a(t) \quad : \quad N_a(t)d(\hat{\mu}_a(t), q) \quad \leq \quad \log t \; + \; c\log\log t\right\} \quad ,$$

where $\hat{\mu}_a(t) = (1/N_a(t))\sum_{s=1}^{t}\mathbb{I}\{a \in \mathcal{A}_s\}\sum_{c=1}^{n_a(s)}X_{a,c,s}$ is the empirical average reward at time t, d is the divergence induced by the Kullback-Leibler divergence defined in Equation (V.2) and c is a positive constant typically smaller than 3. Due to its special importance for bounded rewards, we name KL-Bernoulli-UCB-4P the case $d = d_{\text{Bern}} : (x, y) \mapsto x\log(x/y) + (1-x)\log((1-x)/(1-y))$ and KL-Gaussian-UCB-4P the choice $d = d_{\text{Gauss}} : (x, y) \mapsto 2(x-y)^2$.

Analysis for One-Dimensional Exponential Family. We show for the KL-UCB-4P algorithm a finite-time regret bound that proves its asymptotic optimality up to a multiplicative constant n_a^+/n_a^- (see Section 7 for further discussion). To this purpose, we upper-bound the expected number of times non-profitable arms are pulled and profitable ones are not. The analysis is sketched below, while detailed proofs are deferred to section 10.

Theorem 2. *The* kl-UCB-4P *algorithm satisfies the following properties.*
(i). For any non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ and all $\epsilon > 0$,

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)\frac{n_a^+(\log T + c \log\log T)}{n_a^- d(\mu_a, \tau_a)} + n_a^+ \left\{1 + \frac{H_1(\epsilon)}{T^{\beta_1(\epsilon)}}\right\},$$

where $H_1(\epsilon)$ and $\beta_1(\epsilon)$ are positive functions of ϵ depending on n_a^-, μ_a and τ_a.
(ii). For any profitable arm $a \in \mathcal{A}^*$, if $T \geq \max(3, n_a^+)$ and $c \geq 3$, we have:

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \big\{ e(2c+3)\log\log T + n_a^+ + 3 \big\}.$$

Sketch of Proof. The analysis goes as follows:

(i). For a **non-profitable arm** $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, we must upper bound $\mathbb{E}[N_a(T)]$. At first, a sub-optimal arm is drawn because its confidence bonus is large. But after some $K_T \approx \kappa \log(T)$ draws (where κ is the information constant given in the theorem), the index $u_a(t)$ can be large only when the empirical mean of the observations deviates from its expectation, which has small probability. Thus, we write

$$\mathbb{E}[N_a(T)] \leq n_a^+ \left\{ K_T + \sum_{t \geq 1} \mathbb{P}(a \in \mathcal{A}_{t+1}, N_a(t) > K_T) \right\} .$$

One obtains that K_T gives the main term in the regret. The contribution of the remaining sum is negligible: denoting $d^+(x,y) = d(x,y)\mathbb{I}\{x < y\}$, we observe that:

$$\begin{aligned}
&(a \in \mathcal{A}_{t+1}) = (u_a(t) \geq \tau_a) \\
&\subset (d^+(\hat{\mu}_a(t), \tau_a) \leq d(\hat{\mu}_a(t), u_a(t))) \\
&= \left(d^+(\hat{\mu}_a(t), \tau_a) \leq \frac{\log(t) + c\log\log(t)}{N_a(t)} \right) .
\end{aligned}$$

As a deviation from the mean, the last event proved to have small probability when $N_a(t) > K_T$. Summing over these probabilities produces a term negligible compared to K_T.

(ii). For a **profitable arm** $a \in \mathcal{A}^*$, we must upper bound $\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]$. We write

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin \mathcal{A}_{t+1}) ,$$

and we control the defavorable events by noting that

$$(a \notin \mathcal{A}_{t+1}) = (u_a(t) < \tau_a) \subset (u_a(t) < \mu_a) ,$$

where the probability of the last event can be upper bounded by means of a self-normalized deviation inequality such as in Lemma 10 in [CGM+13].

4.1 Extension to General Bounded Rewards

In this subsection, we consider rewards that are bounded in $[0,1]$ and we build confidence intervals $u_a(t)$ with Bernoulli and Gaussian KL divergence, i.e. $d = d_{\text{Bern}}$ or $d = d_{\text{Gauss}}$, which respectively define kl-Bernoulli-UCB-4P and kl-Gaussian-UCB-4P algorithms. Then, with the same proof as in the one-dimensional exponential family setting, we obtain similar guarantees as in Theorem 2 except that the divergence d is either d_{Bern} or d_{Gauss}. By Pinsker's inequality, $d_{\text{Bern}}(\mu_a, \tau_a) > d_{\text{Gauss}}(\mu_a, \tau_a)$, which implies that kl-Bernoulli-UCB-4P performs always better than kl-Gaussian-UCB-4P. However, this upper bound is not tight w.r.t. the lower bound stated in Theorem 1 obtained for general bounded distributions. Hence, none of these two approaches is asymptotically optimal. A truly non-parametric, optimal strategy might be obtained by the use of Empirical-Likelihood (EL) confidence intervals, as in [CGM+13], but this is beyond the scope of this chapter.

5 The Bayes-UCB-4P Algorithm

This section introduces the Bayes-UCB-4P algorithm.

Analysis. We now propose a Bayesian index policy which is derived from Bayes-UCB ([Kau16]). For all arms $a \in \{1, \dots, K\}$, a prior distribution is chosen for the unknown mean μ_a. At each round $t \geq 1$, we compute the posterior distribution $\pi_{a,t}$ using the $N_a(t)$ observed realizations of ν_a. We compute the quantile $q_a(t) = Q(1 - 1/(t(\log t)^c); \pi_{a,t})$, where $Q(\alpha, \pi)$ denotes the quantile of order α of the distribution π. The Bayes-UCB-4P is the index policy defined by $u_a(t) = q_a(t)$. In other words, arm a is pulled ($a \in \mathcal{A}_{t+1}$) whenever the quantile $q_a(t)$ of the posterior is larger than the threshold τ_a. The following results, proved in section 10, show that Bayes-UCB-4P is asymptotically optimal up to a multiplicative constant n_a^+/n_a^- (see Section 7).

Theorem 3. *When running the* Bayes-UCB-4P *algorithm the following assertions hold.*
(i). For any non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ and for all $\epsilon > 0$ there exists a problem-dependent constant $N_a(\epsilon)$ such that for all $T \geq N_a(\epsilon)$,

$$\mathbb{E}[N_a(T)] \leq \left(\frac{1+\epsilon}{1-\epsilon}\right) \frac{n_a^+ (\log T + c \log\log T)}{n_a^- \, d(\mu_a, \tau_a)} + n_a^+ \left\{ 1 + H_2 + \frac{H_3(\epsilon)}{T^{\beta_2(\epsilon)}} \right\},$$

where H_2, $H_3(\epsilon)$ and $\beta_2(\epsilon)$ are respectively a constant and two positive functions of ϵ depending on n_a^-, τ_a, μ_a and a constant μ_0^- verifying $\mu^- < \mu_0^- \leq \mu_a$.
(ii). For any profitable arm $a \in \mathcal{A}^*$, if $T \geq t_a$ and $c \geq 5$,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \left\{ \frac{e(2(c-2)+4)}{A} \log\log T + t_a + 1 \right\},$$

where $t_a = \max(e/A, 3, A, n_a^+, A n_a^+)$ and A is a constant depending on the chosen prior distribution.

Sketch of Proof. We present the main steps of the proof of Theorem 3 (see section 10 for the complete version). The idea is to capitalize on the analysis of kl-UCB-4P, and to relate the quantiles of the posterior distributions to the Kullback-Leibler upper-confidence bounds.

(i). For a non-profitable arm $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, we want to upper bound $\mathbb{E}[N_a(T)]$. Again, we use the following decomposition:

$$\mathbb{E}[N_a(T)] \leq n_a^+ \left\{ K_T + \sum_{t \geq 1} \mathbb{P}(a \in \mathcal{A}_{t+1}, N_a(t) > K_T) \right\},$$

where $K_T \approx \kappa \log(T)$ of the same order of magnitude as the asymptotic lower bound derived in Theorem 1. This cut-off K_T is expected to be the dominant term in our upper bound, since the contribution of the remaining sum is negligible compared to K_T: when $N_a(t) > K_T$, we first observe that

$$(a \in \mathcal{A}_{t+1}) = (q_a(t) \geq \tau_a) = \left(\pi_{a,t}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c} \right), \tag{V.3}$$

where the $\pi_{a,t}$ is the posterior distribution on μ_a at round t and $q_a(t)$ is, under $\pi_{a,t}$, the quantile of order $1 - \frac{1}{t(\log t)^c}$. The key ingredient here is Lemma 4 from [Kau16], which relates a quantile of the posterior to an upper confidence bound on the empirical mean:

$$\pi_{a,t}([\tau_a, \mu^+[) \lesssim \sqrt{N_a(t)} e^{-N_a(t) d(\hat{\mu}_a(t), \tau_a)}.$$

This permits to conclude as for kl-UCB-4P.

(ii). For a profitable arm $a \in \mathcal{A}^*$, we must upper bound $\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]$. We write

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin \mathcal{A}_{t+1}).$$

Then we note that for all $t \geq 1$,

$$(a \notin \mathcal{A}_{t+1}) = (q_a(t) < \tau_a) = \left(\pi_{a,t}([\tau_a, \mu^+[) < \frac{1}{t(\log t)^c} \right).$$

Using again the bridge between posterior quantiles and upper-confidence bounds of Lemma 4 in [Kau16]:

$$\pi_{a,t}([\tau_a, \mu^+[) \gtrsim \frac{e^{-N_a(t) d(\hat{\mu}_a(t), \tau_a)}}{N_a(t)},$$

we can again argue as for kl-UCB-4P.

6 The TS-4P Algorithm

This section introduces the TS-4P algorithm, a variant of the Thompson Sampling approach.

Analysis. The TS-4P algorithm described in this section is inspired from the analysis of Thompson Sampling provided in [KKM13]. Although the guarantees given in Section 5 for Bayes-UCB-4P are valid for any prior distribution, the Bayesian approach proposed in this section will be analyzed only for Jeffreys priors (see [KKM13] for more details). $\pi_a(0)$ will refer to the prior distribution on θ_a and $\pi_a(t)$ to the posterior distribution updated with the $N_a(t)$ observations collected from arm a up to time t. At each time step $t \geq 1$, sample $\theta_a(t) \sim \pi_a(t)$ and define the TS-4P algorithm (see Algorithm 2) pulling arm a (i.e. $a \in \mathcal{A}_{t+1}$) if $u_a(t) = \mu(\theta_a(t))$ is larger than or equal to τ_a.

Theorem 4. *When running the* TS-4P *algorithm the following assertions hold.*
(i). For any non-profitable arm $a \in \{1, \ldots, K\} \setminus \mathcal{A}^*$ and for all $\epsilon \in]0, 1[$,

$$\mathbb{E}[N_a(T)] \leq \left(\frac{1+\epsilon}{1-\epsilon}\right) \frac{n_a^+ \log T}{n_a^- \, d(\mu_a, \tau_a)} + H_4,$$

where H_4 is a problem dependent constant.
(ii). For any profitable arm $a \in \mathcal{A}^*$,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq H_5,$$

with H_5 a problem dependent constant.

Sketch of Proof. Here we give the main steps of the proof of Theorem 4 (see section 10 for complete proof).

(i). For a non-profitable arm $a \in \{1, \ldots, K\} \setminus \mathcal{A}^*$, we must upper bound $\mathbb{E}[N_a(T)]$. We first write:

$$\mathbb{E}[N_a(T)] \lesssim n_a^+ \left\{ K_T + \sum_{t \geq 1} \mathbb{P}\left(a \in \mathcal{A}_{t+1}, E_a(t), N_a(t) > K_T\right) \right\},$$

where $K_T \approx \kappa \log(T)$ is, as in the proofs of kl-UCB-4P and Bayes-UCB-4P, a cut-off corresponding to the main term in our bound as suggested by the asymptotic lower bound in Theorem 1 and $E_a(t)$ is a high probability event ensuring that the current empirical mean at times t, namely $\hat{\mu}_a(t)$, is well concentrated around the true mean μ_a. It remains to prove that the sum of defavorable events (for $N_a(t) > K_T$ and under $E_a(t)$) is negligible compared to K_T. Observe that the following holds:

$$\mathbb{P}(a \in \mathcal{A}_{t+1}, E_a(t), N_a(t) > K_T) \leq \mathbb{P}(\mu(\theta_a(t)) \geq \tau_a, E_a(t), N_a(t) > K_T), \tag{V.4}$$

where $\theta_a(t)$ is sampled from the posterior distribution $\pi_a(t)$. Then we upper bound the right-hand side expression in Eq. (V.4) thanks to the deviation inequality stated in Theorem 4 in [KKM13] and that we recall in Lemma 2 in section 10. Summing over these probabilities produces a term negligible compared to K_T.

(ii). For a profitable arm $a \in \mathcal{A}^*$, we must upper bound $\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]$, which we decompose as follows:

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \sum_{t=1}^{T-1} \mathbb{P}(a \notin \mathcal{A}_{t+1}).$$

Then, we control the defavorable events: for all $t \geq 1$ and $b \in]0,1[$,

$$\sum_{t=1}^{T-1} \mathbb{P}(a \notin \mathcal{A}_{t+1}) \lesssim \sum_{t=1}^{+\infty} \mathbb{P}\left(\mu(\theta_a(t)) < \tau_a, E_a(t) \middle| N_a(t) > t^b\right) + \sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right),$$

where the first series is proved to converge thanks to Lemma 2 and the second too by Lemma 3 provided in section 10. We point out that our proof of Lemma 3, which is a much simplified version of the proof of Proposition 5 in [KKM13], takes advantage of the independence of arms in our objective (see Section 3.1).

7 Asymptotic Optimality

A direct consequence of theorems 2, 3 and 4 is the following asymptotic upper bound on the regret of kl-UCB-4P (with $c \geq 3$), Bayes-UCB-4P (with $c \geq 5$) and TS-4P:

$$\limsup_{T \to \infty} \frac{R_T}{\log T} \leq \sum_{a,\, \mu_a < \tau_a} \frac{n_a^+ |\Delta_a|}{n_a^- \, d(\mu_a, \tau_a)}.$$

Observe that this asymptotic upper bound on the regret is tight with the asymptotic lower bound in Section 2 when $n_a^+ = n_a^-$ for all non-profitable arms $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$, which is achieved if and only if the $n_a(t)$'s are constant. In this particular case these three algorithms are asymptotically optimal.

8 Numerical Experiments

We perform three series of numerical experiments for three different one-dimensional exponential families: Bernoulli, Poisson and exponential. In each scenario, we consider five arms ($K = 5$) with distributions belonging to the same one-dimensional exponential family. For all arms $a \in \{1, \dots, 5\}$ and time steps $t \in \{1, \dots, T\}$, $n_a(t) - 1$ is sampled from a Poisson distribution $\mathcal{P}(\lambda_a)$, where $(\lambda_1, \dots, \lambda_5) = (3, 4, 5, 6, 7)$. Moreover, the time horizon is chosen equal to $T = 10000$ and the regret is empirically averaged over 10000 independent trajectories. Our experiments also include algorithms, all index policies, whose theoretical properties have not been discussed in this chapter, namely:

- UCB-V-4P: same index as UCB-V introduced in [AMS09] and using empirical estimates of the variance of each distribution,
- KL-Emp-UCB-4P: same index as empirical KL-UCB introduced in [CGM+13] and using the empirical likelihood principle,

- KL-UCB$^+$-4P: derived from KL-UCB$^+$ introduced in [Kau16] and defined by the index $u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq \log(t(\log t)^c/N_a(t))\right\}$.

We also define KL-Bernoulli-UCB$^+$-4P by replacing the divergence d by d_{Bern} in the index of KL-UCB$^+$-4P.

Scenario 1: Bernoulli. In the first scenario, the $K = 5$ categories have Bernoulli distributions $\mathcal{B}(p_a)$ with parameters $(p_1, \dots, p_5) = (0.1, 0.3, 0.5, 0.5, 0.7)$ and thresholds $(\tau_1, \dots, \tau_5) = (0.2, 0.2, 0.4, 0.6, 0.8)$. Hence the profitable arms are the second and the third ones. Notice that although arms 3 and 4 have the same distribution, namely $\mathcal{B}(0.5)$, their thresholds are different such that arm 3 is profitable but not arm 4.

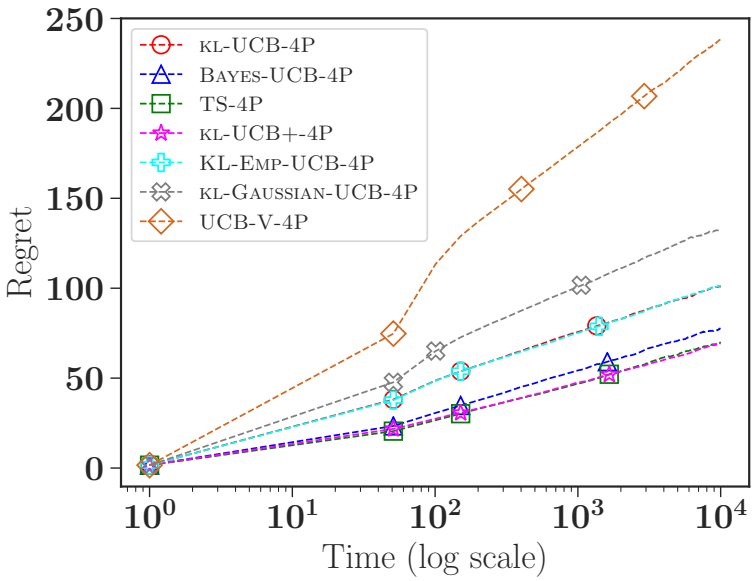


Figure V.1: Regret of various algorithms as a function of time in the Bernoulli scenario.

Observe that KL-Gaussian-UCB-4P produces large regret, which confirms the discussion in section 4.1 stating that it always performs worse than KL-Bernoulli-UCB-4P, which here coincides with KL-UCB-4P.

Scenario 2: Poisson. In the second scenario, the five categories $a \in \{1, \dots, 5\}$ have Poisson distributions $\mathcal{P}(\theta_a)$ with respective mean parameters $(\theta_1, \dots, \theta_5) = (1, 2, 3, 4, 5)$ and thresholds $(\tau_1, \dots, \tau_5) = (2, 1, 4, 3, 6)$: the profitable arms are 2 and 4. In order to run KL-Emp-UCB-4P which assumes boundedness, the rewards are truncated at a maximal value equal to 100.

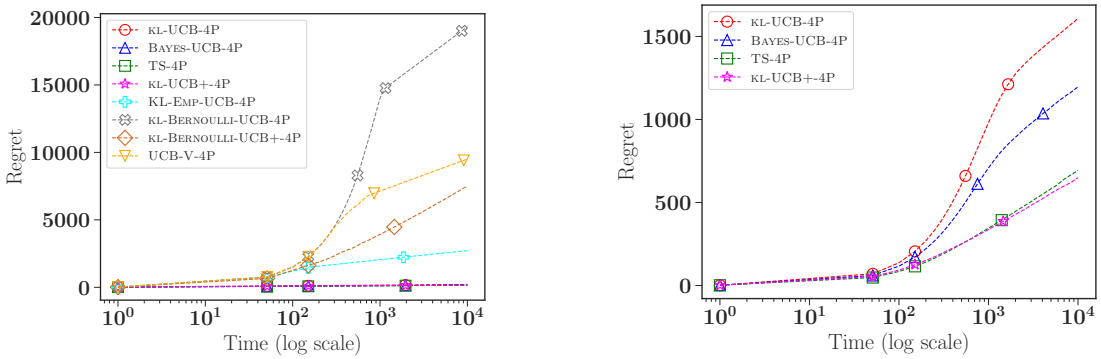


Figure V.2: Regret of various algorithms as a function of time in the Poisson scenario. The right hand-side plot only displays the best performing policies on a harder problem.

The right-hand side plot in Figure V.2 only displays the regret of the best performing strategies on a harder problem with same distributions but thresholds closer to expectations: $(\tau_1, \dots, \tau_5) = (1.1, 1.9, 3.1, 3.9, 5.1)$.

Scenario 3: Exponential. In the third scenario, we consider exponential distributions $\mathcal{E}(\lambda_a)$ with respective mean values $(\lambda_1^{-1}, \dots, \lambda_5^{-1}) = (1, 2, 3, 4, 5)$ and thresholds $(\tau_1, \dots, \tau_5) = (2, 1, 4, 3, 6)$. As in the Poisson scenario, the rewards are truncated at a maximal value of 100.

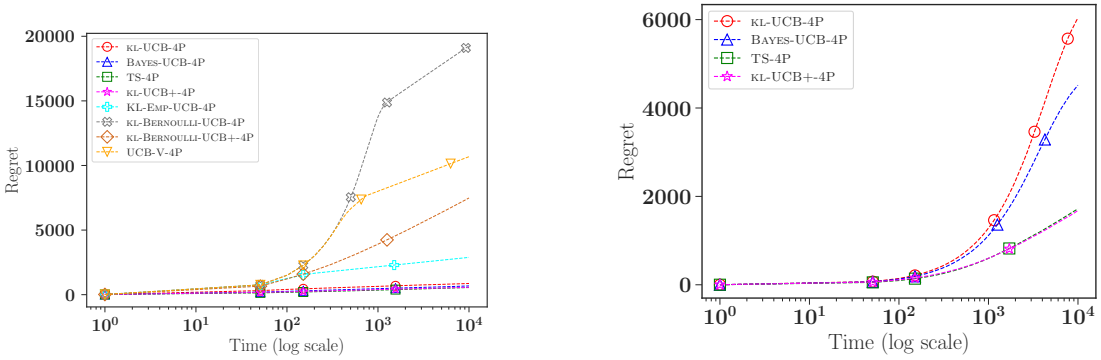


Figure V.3: Regret of various algorithms as a function of time in the exponential scenario. The right hand-side plot only displays the best performing policies on a harder problem.

The right-hand side plot in Figure V.3 only displays the best performing strategies. Here again, the distributions are kept the same but the problem is made harder with new thresholds: $(\tau_1, \dots, \tau_5) = (1.1, 1.9, 3.1, 3.9, 5.1)$.

Interpretation. In each scenario and for each algorithm, the regret curve presents a linear regime corresponding to a logarithmic growth as a function of time. We observe

that the best performing policies (i.e. with small regret) are those adapting to the parametric family of the reward distributions: through the Kullback-Leibler divergence for KL-UCB-4P and KL-UCB$^+$-4P, or through prior distributions for BAYES-UCB-4P and TS-4P. By contrast, KL-GAUSSIAN-UCB-4P always uses the Gaussian Kullback-Leibler divergence, both KL-BERNOULLI-UCB-4P and KL-BERNOULLI-UCB$^+$-4P the Bernoulli divergence and KL-EMP-UCB-4P only assumes that the rewards are bounded. Hence we see that prior knowledge on reward distributions is critical in the efficiency of these algorithms.

9 Conclusion

Motivated by credit risk evaluation of different populations in a sequential context, this chapter introduces the *profitable bandit problem*, evaluates its difficulty by giving an asymptotic lower bound on the expected regret and proposes and theoretically analyzes three algorithms, KL-UCB-4P, BAYES-UCB-4P and TS-4P, by giving finite-time upper bounds on their expected regret for reward distributions belonging to a one-dimensional exponential family. All three algorithms are proven to be asymptotically optimal in the particular setting where for each catefory, a same number of clients is presented to the loaner at each time step. An extension to general bounded distributions is proposed through two algorithms KL-BERNOULLI-UCB-4P and KL-GAUSSIAN-UCB-4P coming with finite-time analysis directly derived from the analysis of KL-UCB-4P. We finally compare all these strategies empirically and also against other policies inspired from other multi-armed bandits algorithms. BAYES-UCB-4P and TS-4P perform the best in our numerical experiments and we observe that policies having prior information on the distributions, through appropriate prior distribution for BAYES-UCB-4P and TS-4P or Kullback-Leibler divergence for KL-UCB-4P, perform much better than non-adaptive strategies like KL-BERNOULLI-UCB-4P and KL-GAUSSIAN-UCB-4P.

10 Technical Proofs

The technical proofs are collected below.

Proof of Theorem 1

We use the inequality (F) in Section 2 in [GMS16], a consequence of the contraction of entropy property, which straightforwardly extends from the classical multi-armed bandit setting to ours where several arms can be pulled at each round t and a number $n_a(t) \geq 1$ of observations are observed simultaneously for each pulled arm a. Then we have

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) \geq \mathrm{kl}(\mathbb{E}_\nu[Z], \mathbb{E}_{\nu'}[Z]), \tag{V.5}$$

where Z is any $\sigma(I_T)$-measurable random variable with values in $[0,1]$. Consider a thresholding bandit problem $(\nu_a, \tau_a)_{1\leq a\leq K} \in \prod_{a=1}^{K} \mathcal{D}_a \times \mathbb{R}$ with at least one non-profitable arm

$a \in \{1, \dots, K\}$, we define a modified problem (ν', τ) such that $\nu'_k = \nu_k$ for all $k \neq a$ and $\nu'_a \in \mathcal{D}_a$ verifies $\mu'_a > \tau_a$. Then, considering $Z = N_a(T)/\widetilde{N}_a(T)$, Eq. (V.5) rewrites as follows:

$$\begin{aligned}\mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) &\geq \mathrm{kl}(\mathbb{E}_\nu[N_a(T)]/\widetilde{N}_a(T), \mathbb{E}_{\nu'}[N_a(T)]/\widetilde{N}_a(T)) \\ &\geq \left(1 - \frac{\mathbb{E}_\nu[N_a(T)]}{\widetilde{N}_a(T)}\right) \log\left(\frac{\widetilde{N}_a(T)}{\widetilde{N}_a(T) - \mathbb{E}_{\nu'}[N_a(T)]}\right) - \log(2),\end{aligned}$$

where we used for the last inequality that for all $(p, q) \in [0, 1]^2$,

$$\mathrm{kl}(p, q) \geq (1 - p) \log\left(\frac{1}{1 - q}\right) - \log(2).$$

Then, by uniform efficiency it holds: $\mathbb{E}_\nu[N_a(T)] = o(\widetilde{N}_a(T))$ and $\widetilde{N}_a(T) - \mathbb{E}_{\nu'}[N_a(T)] = o(\widetilde{N}_a(T)^\alpha)$ for all $\alpha \in (0, 1]$. Hence for all $\alpha \in (0, 1]$,

$$\liminf_{T\to\infty} \frac{1}{\log T}\mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a, \nu'_a) \geq \liminf_{T\to\infty} \frac{1}{\log T} \log\left(\frac{\widetilde{N}_a(T)}{\widetilde{N}_a(T)^\alpha}\right) = 1 - \alpha.$$

Taking the limit $\alpha \to 0$ in the right-hand side and taking the infimum over all distributions $\nu'_a \in \mathcal{D}_a$ such that $\mu'_a > \tau_a$ in the left-hand side conclude the proof.

Proof of Theorem 2

For any arm $a \in \{1, \dots, K\}$, the average reward at time t is denoted by $\hat{\mu}_a(t) = S_a(t)/N_a(t)$ where $S_a(t) = \sum_{s=1}^t \sum_{c=1}^{n_a(s)} X_{a,c,s}\mathbb{I}\{a \in \mathcal{A}_s\}$ and $N_a(t) = \sum_{s=1}^t n_a(s)\mathbb{I}\{a \in \mathcal{A}_s\}$. For every positive integer s, we also denote by $\hat{\mu}_{a,s} = (X_{a,1} + \cdots + X_{a,s})/s$ with $X_{a,1}, \dots, X_{a,s}$ the first s samples pulled from arm a, so that $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$. The upper confidence bound for μ_a appearing in KL-UCB-4P is then given by:

$$u_a(t) = \sup\left\{q > \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t\right\}.$$

For all $(x, y) \in [\mu^-, \mu^+]^2$, define $d^+(x, y) = d(x, y)\mathbb{I}\{x < y\}$.

(i). Let $a \in \{1, \dots, K\} \setminus \mathcal{A}^*$ be a non-profitable arm i.e. such that $\mu_a < \tau_a$. Given $\epsilon \in]0, 1[$, we upper bound the expectation of $N_a(T)$ as follows,

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^T n_a(t)\mathbb{I}\{a \in \mathcal{A}_t\}\right] \leq n_a^+ \mathbb{E}\left[\sum_{t=1}^T \mathbb{I}\{a \in \mathcal{A}_t\}\right].$$

Now observe for $t \geq 1$ that $a \in \mathcal{A}_{t+1}$ implies $u_a(t) \geq \tau_a$ and hence,

$$d^+(\hat{\mu}_a(t), \tau_a) \leq d(\hat{\mu}_a(t), u_a(t)) = \frac{\log t + c \log\log t}{N_a(t)}.$$

Then,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{I}\{a \in \mathcal{A}_t\} \\
&= 1 + \sum_{t=1}^{T-1} \mathbb{I}\{a \in \mathcal{A}_{t+1}\} \sum_{s=1}^{t} \sum_{1 \le i_1 < \cdots < i_s \le t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1,\ldots,i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1,\ldots,t\} \setminus \{i_1,\ldots,i_s\}} \mathcal{A}_i \right\} \\
&\qquad \times \mathbb{I}\left\{ (n_a(i_1) + \cdots + n_a(i_s)) d^+\left(\hat{\mu}_{a, n_a(i_1)+\cdots+n_a(i_s)}, \tau_a\right) \le \log t + c \log\log t \right\}.
\end{aligned} \tag{V.6}$$

Given $\epsilon \in]0,1[$, we upper bound the last indicator function appearing in Eq. (V.6) by

$$\begin{aligned}
&\mathbb{I}\{s < K_T\} + \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \ge K_T, k d^+(\hat{\mu}_{a,k}, \tau_a) \le \log T + c \log\log T \right\} \\
&\le \mathbb{I}\{s < K_T\} + \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \ge K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \le \frac{d(\mu_a, \tau_a)}{1+\epsilon} \right\},
\end{aligned} \tag{V.7}$$

where $K_T = \left\lceil (1+\epsilon) \frac{\log T + c \log\log T}{n_a^- d(\mu_a, \tau_a)} \right\rceil$. The last expression in Eq. (V.7) is not using the indices $t, i_1, \ldots, i_s$ which allows us to exchange the sums over t and s in Eq. (V.6) and to obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{I}\{a \in \mathcal{A}_t\} \\
&\le 1 + \sum_{s=1}^{T} \left(\mathbb{I}\{s < K_T\} + \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \ge K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \le \frac{d(\mu_a, \tau_a)}{1+\epsilon} \right\} \right) \\
&\qquad \times \sum_{t=1}^{T-1} \mathbb{I}\{a \in \mathcal{A}_{t+1}\} \sum_{1 \le i_1 < \cdots < i_s \le t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1,\ldots,i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1,\ldots,t\} \setminus \{i_1,\ldots,i_s\}} \mathcal{A}_i \right\} \\
&\le K_T + \sum_{s=K_T}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ d^+(\hat{\mu}_{a,k}, \tau_a) \le \frac{d(\mu_a, \tau_a)}{1+\epsilon} \right\},
\end{aligned}$$

where the last inequality is implied by

$$\sum_{t=1}^{T-1} \mathbb{I}\{a \in \mathcal{A}_{t+1}\} \sum_{1 \le i_1 < \cdots < i_s \le t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1,\ldots,i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1,\ldots,t\} \setminus \{i_1,\ldots,i_s\}} \mathcal{A}_i \right\} \le 1. \tag{V.8}$$

Hence,

$$\mathbb{E}[N_a(T)] \leq n_a^+ \left\{ K_T + \sum_{s=K_T}^{+\infty} \sum_{k=n_a^- s}^{+\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a, \tau_a)}{1+\epsilon} \right) \right\}$$
$$\leq (1+\epsilon)\frac{n_a^+}{n_a^-}\frac{\log T + c\log\log T}{d(\mu_a,\tau_a)} + n_a^+ \left\{ 1 + \frac{H_1(\epsilon)}{T^{\beta_1(\epsilon)}} \right\},$$

comes from Lemma 1 with $H_1(\epsilon)$ and $\beta_1(\epsilon)$ positive functions of ϵ.

(ii). Now consider $a \in A^*$ i.e. verifying $\mu_a > \tau_a$. It follows,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=2}^{T} n_a(t) \mathbb{I}\{a \notin \mathcal{A}_t\} \right] \leq n_a^+ \sum_{t=1}^{T-1} \mathbb{P}\left(u_a(t) < \mu_a \right).$$

Let $t \in \{1, \ldots, T-1\}$ and observe that $(u_a(t) < \mu_a) \subset (d^+(\hat{\mu}_a(t), \mu_a) > d(\hat{\mu}_a(t), u_a(t)))$. Hence for $c \geq 3$ and $t \geq \max(3, n_a^+)$,

$$\begin{aligned}
&\mathbb{P}\left(u_a(t) < \mu_a \right) \\
&\leq \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t), \mu_a) > \delta_t \right) \leq (\delta_t \log(n_a^+ t) + 1)\exp(-\delta_t + 1) \\
&= \frac{e((\log t)^2 + c\log(t)\log\log(t) + \log(n_a^+)\log(t) + c\log(n_a^+)\log\log(t) + 1)}{t(\log t)^c} \\
&\leq \frac{e(2c+3)}{t\log t},
\end{aligned}$$

where $\delta_t = \log t + c\log\log t > 1$ and the second inequality results from the self-normalized concentration inequality stated in Lemma 10 in [CGM+13]. Then by summing over t,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \left\{ 2 + n_a^+ + e(2c+3) \sum_{t=3}^{T-1} \frac{1}{t\log t} \right\}$$
$$\leq n_a^+ \left\{ e(2c+3)\log\log T + n_a^+ + 3 \right\}.$$

Lemma 1. *Let $a \in \{1, \ldots, K\} \setminus \mathcal{A}^*$ a non-profitable arm (i.e. $\mu_a < \tau_a$), $\epsilon \in]0,1[$ and $K_T = \left\lceil f(\epsilon)\frac{\log T + c\log\log T}{n_a^- d(\mu_a,\tau_a)} \right\rceil$ with f a function such that $f(\epsilon') > 1$ for all $\epsilon' \in]0,1[$. Then there exist $H(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that*

$$\sum_{s=K_T}^{+\infty} \sum_{k=n_a^- s}^{+\infty} \mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a,\tau_a)}{f(\epsilon)} \right) \leq \frac{H(\epsilon)}{T^{\beta(\epsilon)}},$$

where $H(\epsilon)$ and $\beta(\epsilon)$ are positive functions of ϵ depeding on μ_a, τ_a and n_a^-.

proof. Observe that $d^+(\hat{\mu}_{a,k}, \tau_a) \leq d(\mu_a,\tau_a)/f(\epsilon)$ if and only if $\hat{\mu}_{a,k} \geq r(\epsilon)$ where $r(\epsilon) \in]\mu_a, \tau_a[$ verifies $d(r(\epsilon), \tau_a) = d(\mu_a,\tau_a)/f(\epsilon)$. Thus,

$$\mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{d(\mu_a,\tau_a)}{f(\epsilon)} \right) = \mathbb{P}\left(\hat{\mu}_{a,k} \geq r(\epsilon) \right) \leq e^{-k d(r(\epsilon), \mu_a)}$$

and

$$\begin{aligned}\sum_{s=K_T}^{T}\sum_{k=n_a^- s}^{n_a^+ s}\mathbb{P}\left(d^+(\hat{\mu}_{a,k},\tau_a)\leq\frac{d(\mu_a,\tau_a)}{f(\epsilon)}\right)&\leq\sum_{s=K_T}^{+\infty}\sum_{k=n_a^- s}^{+\infty}e^{-kd(r(\epsilon),\mu_a)}\\&=\frac{1}{1-e^{-d(r(\epsilon),\mu_a)}}\sum_{s=K_T}^{+\infty}e^{-n_a^- sd(r(\epsilon),\mu_a)}\\&=\frac{e^{-n_a^- d(r(\epsilon),\mu_a)K_T}}{(1-e^{-d(r(\epsilon),\mu_a)})\left(1-e^{-n_a^- d(r(\epsilon),\mu_a)}\right)}\\&\leq\frac{H(\epsilon)}{T^{\beta(\epsilon)}},\end{aligned}$$

where $H(\epsilon)=\left[\left(1-e^{-d(r(\epsilon),\mu_a)}\right)\left(1-e^{-n_a^- d(r(\epsilon),\mu_a)}\right)\right]^{-1}$ and $\beta(\epsilon)=f(\epsilon)d(r(\epsilon),\mu_a)/d(\mu_a,\tau_a)$.

Proof of Theorem 3

We first recall that the posterior distribution on the mean of a distribution belonging to an exponential family only depends on the number of observations n and the empirical mean x (see e.g. Lemma 1 in [Kau16]): for a given arm $a\in\{1,\dots,K\}$, we denote this posterior by $\pi_{a,n,x}$. Given two constants $\mu_0^-<\mu^-$ and $\mu_0^+<\mu^+$ verifying $\mu_0^-\leq\mu_a\leq\mu_0^+$ for all arms $a\in\{1,\dots,K\}$, we define the truncated empirical mean: $\bar{\mu}_a(t)=\min(\max(\hat{\mu}_a(t),\mu_0^-),\mu_0^+)$. Then, for any arm $a\in\{1,\dots,K\}$ and time step $t\geq 1$, the posterior distribution involved in Bayes-UCB-4P defines as follows:

$$\pi_{a,t}=\pi_{a,N_a(t),\bar{\mu}_a(t)}.$$

(i). Let $a\in\{1,\dots,K\}\setminus\mathcal{A}^*$ be a non-profitable arm (i.e. $\mu_a<\tau_a$). We upper bound the expectation of $N_a(T)$ as follows:

$$\begin{aligned}&\mathbb{E}[N_a(T)]=\mathbb{E}\left[\sum_{t=1}^{T}n_a(t)\mathbb{I}\{a\in\mathcal{A}_t\}\right]\leq n_a^+\mathbb{E}\left[1+\sum_{t=1}^{T-1}\mathbb{I}\{q_a(t)\geq\tau_a\}\right]\\&=n_a^+\mathbb{E}\left[1+\sum_{t=1}^{T-1}\mathbb{I}\left\{\pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a,\mu^+[)\geq\frac{1}{t(\log t)^c},a\in\mathcal{A}_{t+1}\right\}\right]\\&\leq n_a^+\mathbb{E}\Bigg[1+\sum_{t=1}^{T-1}\mathbb{I}\left\{\bar{\mu}_a(t)<\tau_a,\pi_{a,N_a(t),\bar{\mu}_a(t)}([\tau_a,\mu^+[)\geq\frac{1}{t(\log t)^c},a\in\mathcal{A}_{t+1}\right\}\qquad\text{(V.9)}\\&\qquad+\sum_{t=1}^{T-1}\mathbb{I}\{\bar{\mu}_a(t)\geq\tau_a,a\in\mathcal{A}_{t+1}\}\Bigg].\qquad\text{(V.10)}\end{aligned}$$

Using Lemma 4 in [Kau16], the first sum in (V.9) is upper bounded by

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{I}\left\{ B\sqrt{N_a(t)} e^{-N_a(t) d^+(\bar{\mu}_a(t), \tau_a)} \geq \frac{1}{t(\log t)^c}, a \in \mathcal{A}_{t+1} \right\} \\
&= \sum_{t=1}^{T-1} \mathbb{I}\left\{a \in \mathcal{A}_{t+1}\right\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1, \ldots, i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1, \ldots, t\} \setminus \{i_1, \ldots, i_s\}} \mathcal{A}_i \right\} \\
&\quad \times \mathbb{I}\left\{ B\sqrt{n_a(i_1) + \cdots + n_a(i_s)} e^{-(n_a(i_1) + \cdots + n_a(i_s)) d^+\left(\bar{\mu}_{a, n_a(i_1) + \cdots + n_a(i_s)}, \tau_a\right)} \geq \frac{1}{t(\log t)^c} \right\},
\end{aligned} \tag{V.11}$$

where B is a constant depending on μ_0^-, μ_0^+ and on prior densities. Then we upper bound the last indicator function appearing in Eq. (V.11) by

$$\begin{aligned}
&\mathbb{I}\{s < K_T\} + \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \geq K_T, k d^+(\bar{\mu}_{a,k}, \tau_a) \leq \log T + c \log\log T + \frac{1}{2}\log k + \log B \right\} \\
&\leq \mathbb{I}\{s < K_T\} + \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \geq K_T, k d^+(\hat{\mu}_{a,k}, \tau_a) \leq \log T + c \log\log T + \frac{1}{2}\log k + \log B \right\} \\
&\quad + \mathbb{I}\{\hat{\mu}_{a,k} < \mu_0^-\}.
\end{aligned} \tag{V.12}$$

We are now able to upper bound the right-hand side expression in Eq. (V.11) by injecting Eq. (V.12) and switching the sums on indices t and s, which leads to

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{I}\left\{ \bar{\mu}_a(t) < \tau_a, \pi_{a, N_a(t), \bar{\mu}_a(t)}([\tau_a, \mu^+[) \geq \frac{1}{t(\log t)^c}, a \in \mathcal{A}_{t+1} \right\} \\
&\leq K_T - 1 + \sum_{s=1}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{ s \geq K_T, k d^+(\hat{\mu}_{a,k}, \tau_a) \leq \log T + c \log\log T + \frac{1}{2}\log k + \log B \right\} \\
&\quad + \mathbb{I}\{\hat{\mu}_{a,k} < \mu_0^-\},
\end{aligned} \tag{V.13}$$

where we used the same argument as in Eq. (V.8) to get rid of the sum over t.

Given $\epsilon \in]0,1[$ we define $K_T = \left\lceil \frac{1+\epsilon}{1-\epsilon} \frac{\log T + c \log\log T}{n_a^- d(\mu_a, \tau_a)} \right\rceil$ and denote by $N_a(\epsilon)$ the constant such that $T \geq N_a(\epsilon)$ implies:

$$K_T \geq \left\lceil \frac{3}{n_a^-} \right\rceil \quad \text{and} \quad \frac{1}{n_a^- K_T} \left(\frac{1}{2} \log(n_a^- K_T) + \log(B) \right) \leq \frac{\epsilon}{1+\epsilon} d(\mu_a, \tau_a), \tag{V.14}$$

where the first inequality ensures that for all $k \geq n_a^- K_T$, the function $k \mapsto \log(x)/x$ decreases. Hence, the first indicator function appearing in the right-hand side in Eq.

(V.13) is upper bounded by

$$\mathbb{I}\left\{s \geq K_T, d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{1-\epsilon}{1+\epsilon} d(\mu_a, \tau_a)\right\}. \tag{V.15}$$

By combining equations (V.9), (V.13) and (V.15) we obtain

$$\begin{aligned}\mathbb{E}[N_a(T)] \leq n_a^+ \Bigg\{ K_T + \sum_{s=K_T}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}\left(d^+(\hat{\mu}_{a,k}, \tau_a) \leq \frac{1-\epsilon}{1+\epsilon} d(\mu_a, \tau_a)\right) \\ + \sum_{s=1}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} < \mu_0^-) + \sum_{t=1}^{T-1} \mathbb{P}(\bar{\mu}_a(t) \geq \tau_a, a \in \mathcal{A}_{t+1}) \Bigg\},\end{aligned} \tag{V.16}$$

where the first sum can be upper bounded by $H_3(\epsilon) T^{-\beta_2(\epsilon)}$ with $H_3(\epsilon) > 0$ and $\beta_2(\epsilon) > 0$ thanks to Lemma 1. We upper bound the second sum in Eq. (V.16) with Chernoff inequality:

$$\begin{aligned}\sum_{s=1}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} < \mu_0^-) &\leq \sum_{s=1}^{+\infty} \sum_{k=n_a^- s}^{+\infty} e^{-k d(\mu_0^-, \mu_a)} \\ &= \frac{e^{-n_a^- d(\mu_0^-, \mu_a)}}{\left(1 - e^{-d(\mu_0^-, \mu_a)}\right)\left(1 - e^{-n_a^- d(\mu_0^-, \mu_a)}\right)}.\end{aligned}$$

Finally, we upper bound the third sum in Eq. (V.16) by

$$\begin{aligned}&\mathbb{E}\left[\sum_{t=1}^{T-1} \mathbb{I}\{\hat{\mu}_{a,s} \geq \tau_a, a \in \mathcal{A}_{t+1}\}\right] \\ &\leq \mathbb{E}\Bigg[\sum_{t=1}^{T-1} \mathbb{I}\{a \in \mathcal{A}_{t+1}\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{a \in \bigcap_{i \in \{i_1, \ldots, i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1, \ldots, t\} \setminus \{i_1, \ldots, i_s\}} \mathcal{A}_i\right\} \\ &\qquad \times \mathbb{I}\left\{\hat{\mu}_{a, n_a(i_1) + \cdots + n_a(i_s)} \geq \tau_a\right\}\Bigg] \\ &\leq \sum_{s=1}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}(\hat{\mu}_{a,k} \geq \tau_a) \leq \frac{e^{-n_a^- d(\tau_a, \mu_a)}}{\left(1 - e^{-d(\tau_a, \mu_a)}\right)\left(1 - e^{-n_a^- d(\tau_a, \mu_a)}\right)},\end{aligned} \tag{V.17}$$

where we respectively used Eq. (V.8) and Chernoff inequality in the two last inequalities.

(ii). Now consider $a \in A^*$. We have,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=1}^{T-1} n_a(t+1)\mathbb{I}\{a \notin \mathcal{A}_{t+1}\}\right] = n_a^+ \sum_{t=1}^{T-1} \mathbb{P}(q_a(t) < \tau_a)$$
$$\leq n_a^+ \left\{ t_0 - 1 + \sum_{t=t_0}^{T-1} \mathbb{P}\left(\hat{\mu}_a(t) < \tau_a, N_a(t) \geq (\log t)^2\right) + \sum_{t=1}^{T-1} \mathbb{P}\left(q_a(t) < \tau_a, N_a(t) \leq (\log t)^2\right) \right\}, \tag{V.18}$$

where $t_0 = \max(t_1, t_2)$ with t_1 the smallest integer verifying $C^2 t_0 (\log t_0)^{2c} \geq 1$, which implies for all $t \geq t_1$ that $\bar{\mu}_a(t) \leq q_a(t)$, and $t_2 = \lceil \exp(2/d(\tau_a, \mu_a)) \rceil$ to ensure that $d(\tau_a, \mu_a)(\log t)^2 \geq 2 \log t$ for all $t \geq t_2$. To upper bound the first sum in Eq. (V.18) we write for $t \geq t_0$,

$$\mathbb{P}\left(\hat{\mu}_a(t) < \tau_a, N_a(t) \geq (\log t)^2\right) \leq \sum_{s=\lceil (\log t)^2 \rceil}^{t} \mathbb{P}(\hat{\mu}_{a,s} < \tau_a) \leq \sum_{s=\lceil (\log t)^2 \rceil}^{+\infty} e^{-s d(\tau_a, \mu_a)}$$
$$\leq e^{-d(\tau_a, \mu_a)(\log t)^2} \leq \frac{1}{t^2}.$$

To upper bound the second sum in Eq. (V.18) use again Lemma 4 in [Kau16],

$$\mathbb{P}\left(q_a(t) < \tau_a, N_a(t) \leq (\log t)^2\right) = \mathbb{P}\left(\pi_{a, N_a(t), \bar{\mu}_a(t)}([\tau_a, \mu^+[) < \frac{1}{t(\log t)^c}, N_a(t) \leq (\log t)^2\right)$$
$$\leq \mathbb{P}\left(\frac{A e^{-N_a(t) d(\bar{\mu}_a(t), \tau_a)}}{N_a(t)} < \frac{1}{t(\log t)^c}, N_a(t) \leq (\log t)^2\right)$$
$$= \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t), \tau_a) > \log\left(\frac{At(\log t)^c}{N_a(t)}\right), N_a(t) \leq (\log t)^2\right)$$
$$\leq \mathbb{P}\left(N_a(t) d^+(\hat{\mu}_a(t), \tau_a) > \log(At) + (c-2)\log\log t\right),$$

where A is a constant depending on μ_0^-, μ_0^+ and on prior densities. Then for $c \geq 5$, using the self-normalized deviation inequality stated in Lemma 10 in [CGM$^+$13], we have,

$$\mathbb{P}(N_a(t) d^+(\hat{\mu}_a(t), \tau_a) > \log(At) + (c-2)\log\log t) \leq (\delta_t \log(n_a^+ t) + 1)\exp(-\delta_t + 1)$$
$$= (At(\log(t))^{c-2})^{-1}\{e((\log(t))^2 + (c-2)\log(t)\log\log(t) + \log(An_a^+)\log(t)$$
$$+ (c-2)\log(n_a^+)\log\log(t) + \log(A)\log(n_a^+) + 1)\}$$
$$\leq \frac{e(2(c-2)+4)}{At\log(t)},$$

where we assumed $t \geq t_a = \max(e/A, 3, A, n_a^+, An_a^+)$ to ensure the last inequality and that $\delta_t = \log(At) + (c-2)\log\log(t) > 1$. Then by summing over t,

$$\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] \leq n_a^+ \left\{ t_a + \frac{e(2(c-2)+4)}{A} \sum_{t=3}^{T-1} \frac{1}{t\log t} \right\}$$
$$\leq n_a^+ \{e(2(c-2)+4)\log\log T + t_a + 1\}.$$

Proof of Theorem 4

We first introduce some notations. Denote by $(X_{a,s})_{s\geq 1}$ i.i.d. samples from distribution ν_a. Let $L(\theta) = (1/2)\min(1, \sup_x p(x|\theta))$ and for any $\delta_a > 0$,

$$E_{a,s} = \left(\exists s' \in \{1,\dots,s\}, p(X_{a,s'}|\theta_a) \geq L(\theta_a), \left| \frac{\sum_{u=1,u\neq s'}^{s} X_{a,u}}{s-1} - \mu_a \right| \leq \delta_a \right)$$

is an event where there is at least one 'likely' observation of arm a (namely $X_{a,s'}$) and such that the empirical sufficient statistic is close to its true mean. We also define $E_a(t) = E_{a,N_a(t)}$.

Remark 1. *In the definition of $E_{a,s}$, the 'likely' observation $X_{a,s'}$ is only needed for technical reasons when the Jeffreys prior $\pi_a(0)$ is improper (see Remark 8 in [KKM13] for further discussion).*

We now recall the Theorem 4 in [KKM13], an important result on the posterior concentration under the event $E_a(t)$.

Lemma 2. *There exists problem-dependent constants $C_{1,a}$ and $N_{1,a}$ and a function $\Delta \mapsto C_{2,a}(\Delta)$ such that for $\delta_a \in]0,1[$ and $\Delta > 0$ verifying $1 - \delta_a C_{2,a}(\Delta) > 0$, it holds whenever $N_a(t) \geq N_{1,a}$:*

$$\mathbb{P}\left(\mu(\theta_a(t)) \geq \mu_a + \Delta, E_a(t) | (X_{a,s})_{1\leq s\leq N_a(t)}\right) \leq C_{1,a} N_a(t) e^{-(N_a(t)-1)(1-\delta_a C_{2,a}(\Delta))d(\mu_a, \mu_a+\Delta)}$$

and

$$\mathbb{P}\left(\mu(\theta_a(t)) \leq \mu_a - \Delta, E_a(t) | (X_{a,s})_{1\leq s\leq N_a(t)}\right) \leq C_{1,a} N_a(t) e^{-(N_a(t)-1)(1-\delta_a C_{2,a}(\Delta))d(\mu_a, \mu_a-\Delta)}.$$

Thanks to these concentration inequalities we can derive bounds on the expected number of pulls of any arm.

For all arms $a \in \{1,\dots,K\}$ and $t \geq 1$, $\theta_a(t)$ is a r.v. sampled from the posterior distribution $\pi_a(t)$ on θ_a obtained after $N_a(t)$ observations. For all $s \geq 1$, we also denote by $\theta_{a,s}$ a r.v. sampled from the posterior distribution resulting from the first s observations pulled from arm a (with arbitrary choice when some of these random variables are pulled together), so that $\theta_a(t) = \theta_{a,N_a(t)}$.

We now prove Theorem 4.

(i). Let $a \in \{1,\dots,K\} \setminus \mathcal{A}^*$ be a non-profitable arm (i.e. $\mu_a < \tau_a$). We upper bound the expectation of $N_a(T)$ as follows:

$$\mathbb{E}[N_a(T)] = \mathbb{E}\left[n_a(t) \sum_{t=1}^{T} \mathbb{I}\{a \in \mathcal{A}_t\} \right] \leq n_a^+ \left\{ 1 + \sum_{t=1}^{T-1} \mathbb{P}\left(a \in \mathcal{A}_{t+1}, E_a(t)\right) + \mathbb{P}\left(a \in \mathcal{A}_{t+1}, E_a(t)^c\right) \right\}. \tag{V.19}$$

First observe that the first sum in the right-hand side in Eq. (V.19) is equal to

$$\mathbb{E}\Bigg[\sum_{t=1}^{T-1}\mathbb{I}\left\{a\in\mathcal{A}_{t+1}\right\}\sum_{s=1}^{t}\sum_{1\leq i_1<\cdots<i_s\leq t}\mathbb{I}\left\{a\in\bigcap_{i\in\{i_1,\ldots,i_s\}}\mathcal{A}_i, a\notin\bigcup_{i\in\{1,\ldots,t-1\}\setminus\{i_1,\ldots,i_s\}}\mathcal{A}_i\right\}$$
$$\times\,\mathbb{I}\left\{\mu(\theta_{a,n_a(i_1)+\cdots+n_a(i_s)})\geq\tau_a, E_{a,n_a(i_1)+\cdots+n_a(i_s)}\right\}\Bigg].$$

Then, given $\epsilon\in]0,1[$, by choosing $\delta_a\leq\epsilon/C_{2,a}(|\Delta_a|)$, defining $K_T=\left\lceil\frac{1+\epsilon}{1-\epsilon}\frac{\log T}{n_a^- d(\mu_a,\tau_a)}\right\rceil$ and observing that $\mathbb{I}\{\mu(\theta_{a,n_a(i_1)+\cdots+n_a(i_s)})\geq\tau_a, E_{a,n_a(i_1)+\cdots+n_a(i_s)}\}$ is upper bounded by $\mathbb{I}\{s<K_T\}+\sum_{k=n_a^- s}^{n_a^+ s}\mathbb{I}\{s\geq K_T,\mu(\theta_{a,k})\geq\tau_a,E_{a,k}\}$, we obtain:

$$\begin{aligned}
\sum_{t=1}^{T-1}\mathbb{P}\left(a\in\mathcal{A}_{t+1},E_a(t)\right)&\leq K_T-1+\sum_{s=K_T}^{T}\sum_{k=n_a^- s}^{n_a^+ s}\mathbb{P}\left(\mu(\theta_{a,k})\geq\tau_a,E_{a,k}\right)\\
&\leq K_T-1+\sum_{s=K_T}^{T}\sum_{k=n_a^- s}^{n_a^+ s}C_{1,a}ke^{-(k-1)(1-\epsilon)d(\mu_a,\tau_a)}\\
&\leq\frac{1+\epsilon}{1-\epsilon}\frac{\log T}{n_a^- d(\mu_a,\tau_a)}+C_{1,a}T(n_a^+K_T)^2e^{-(n_a^-K_T-1)(1-\epsilon)d(\mu_a,\tau_a)}\\
&\leq\frac{1+\epsilon}{1-\epsilon}\frac{\log T}{n_a^- d(\mu_a,\tau_a)}+C_{1,a}e^{(1-\epsilon)d(\mu_a,\tau_a)}\frac{(n_a^+K_T)^2}{T^\epsilon},
\end{aligned}$$

where we used in the first inequality Eq. (V.8). In the second and third inequalities we assumed T larger than $N_a(\epsilon)$ verifying $T\geq N_a(\epsilon)\Rightarrow K_T\geq\max(N_{1,a}/n_a^-,N_{2,a})$ with $N_{1,a}$ defined in Lemma 2 and $N_{2,a}$ such that the function $u\mapsto u^2e^{-(n_a^-u-1)(1-\epsilon)d(\mu_a,\tau_a)}$ is decreasing for $u\geq N_{2,a}$.

In order to upper bound the second sum in the right-hand side in Eq. (V.19) we first introduce the following events:

$$B_{a,s}=(\forall s'\in\{1,\ldots,s\},p(X_{a,s'}|\theta_a)\leq L(\theta_a))$$

and

$$D_{a,s}=\left(\exists s'\in\{1,\ldots,s\},\left|\frac{\sum_{u=1,u\neq s'}^{s}X_{a,u}}{s-1}-\mu_a\right|>\delta_a\right).$$

Then observing that $E_a(t)^c \subset B_{a,N_a(t)} \bigcup D_{a,N_a(t)}$ and it holds

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{P}(a \in \mathcal{A}_{t+1}, E_a(t)^c) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T-1} \mathbb{I}\left\{a \in \mathcal{A}_{t+1}\right\} \sum_{s=1}^{t} \sum_{1 \leq i_1 < \cdots < i_s \leq t} \mathbb{I}\left\{ a \in \bigcap_{i \in \{i_1,\ldots,i_s\}} \mathcal{A}_i, a \notin \bigcup_{i \in \{1,\ldots,t-1\} \setminus \{i_1,\ldots,i_s\}} \mathcal{A}_i \right\} \right. \\
&\qquad \left. \times \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{I}\left\{B_{a,k}\right\} + \mathbb{I}\left\{D_{a,k}\right\} \right] \\
&\leq \sum_{s=1}^{T} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}\left(B_{a,k}\right) + \mathbb{P}\left(D_{a,k}\right) \\
&\leq \sum_{s=1}^{+\infty} n_a^+ s \mathbb{P}\left(p(X_{a,1}|\theta_a) < L(\theta_a)\right)^{n_a^- s} + \left(n_a^+ s\right)^2 \left(e^{-(n_a^- s - 1) d(\mu_a - \delta_a, \mu_a)} + e^{-(n_a^- s - 1) d(\mu_a + \delta_a, \mu_a)}\right) < +\infty,
\end{aligned}$$

where we used Eq. (V.8) in the second inequality.

(ii). Now consider $a \in A^*$ i.e. verifying $\mu_a > \tau_a$. Let $b \in]0,1[$, we have:

$$\begin{aligned}
&\widetilde{N}_a(T) - \mathbb{E}[N_a(T)] = \mathbb{E}\left[\sum_{t=2}^{T} n_a(t) \mathbb{I}\{a \notin \mathcal{A}_t\}\right] \leq n_a^+ \sum_{t=1}^{T-1} \mathbb{P}(\mu(\theta_a(t)) < \tau_a) \\
&\leq n_a^+ \left\{ \sum_{t=1}^{T-1} \mathbb{P}\left(\mu(\theta_a(t)) < \tau_a, E_a(t) \middle| N_a(t) > t^b\right) + \sum_{t=1}^{T-1} \mathbb{P}\left(E_a(t)^c \middle| N_a(t) > t^b\right) + \sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \right\}.
\end{aligned} \tag{V.20}$$

By applying Lemma 2, the first sum in Eq. (V.20) is upper bounded by

$$N_{0,a}^{1/b} + \sum_{t=\left\lceil N_{0,a}^{1/b} \right\rceil}^{+\infty} C_{1,a} t^b e^{-(t^b - 1)(1 - \delta_a C_{2,a}(|\Delta_a|)) d(\mu_a, \tau_a)} < +\infty,$$

where $N_{0,a} = \max(N_{1,a}, N_{3,a})$ with $N_{3,a}$ such that the function $u \mapsto u e^{-(u-1)(1-\delta_a C_{2,a}(|\Delta_a|)) d(\mu_a, \tau_a)}$ is decreasing for $u \geq N_{3,a}$.

By applying Chernoff inequality we upper bound the second sum in Eq. (V.20) by

$$\begin{aligned}
&\sum_{t=1}^{T-1} \mathbb{P}\left(E_a(t)^c \middle| N_a(t) > t^b\right) \leq \sum_{t=1}^{T} \sum_{s=\lceil t^b / n_a^+ \rceil}^{t} \sum_{k=n_a^- s}^{n_a^+ s} \mathbb{P}(B_{a,k}) + \mathbb{P}(D_{a,k}) \\
&\leq \sum_{t=1}^{+\infty} n_a^+ t^2 \mathbb{P}(p(X_{a,1}|\theta_a) \leq L(\theta_a))^{\frac{n_a^-}{n_a^+} t^b} + 2(n_a^+)^2 t^3 \left(e^{-\left(\frac{n_a^-}{n_a^+} t^b - 1\right) d(\mu_a - \delta_a, \mu_a)} + e^{-\left(\frac{n_a^-}{n_a^+} t^b - 1\right) d(\mu_a + \delta_a, \mu_a)} \right) \\
&< +\infty.
\end{aligned}$$

Finally we upper bound the third sum in Eq. (V.20) with the following result, inspired from Proposition 5 in [KKM13]. In our case its proof is simpler as there are no dependencies between arms in the objective of the profitable bandit problem.

Lemma 3. *For any profitable arm $a \in A^*$ and any $b \in]0,1[$, there exists a problem-dependent constant $C_b < +\infty$ such that*

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \leq C_b.$$

Then, by using the Bernstein-Von-Mises theorem telling us that $\lim_{j\to+\infty} \mathbb{P}(\mu(\theta_a(\tau_j)) < \tau_a) = 0$, we deduce that there exists a constant $C \in]0,1[$ such that for all $j \geq 0$, $\mathbb{P}(\mu(\theta_a(\tau_j)) < \tau_a) \leq C$. Hence,

$$\sum_{t=1}^{+\infty} \mathbb{P}\left(N_a(t) \leq t^b\right) \leq \sum_{t=1}^{+\infty} \left(t^b + 1\right) C^{t^{1-b}-1} < +\infty.$$

Proof of Lemma 3

In all this proof we consider a fixed profitable arm $a \in A^*$. We follow the lines of the proof of Proposition 5 in [KKM13] : let t_j be the occurence of the j-th play of the arm a (with $t_0 = 0$ by convention). Let $\xi_j = t_{j+1} - t_j - 1$, it corresponds to the number of time steps between the j-th and the $(j+1)$-th play of arm a. Hence, $t - N_a(t) \leq \sum_{j=0}^{N_a(t)} \xi_j$ and we have

$$\begin{aligned}
\mathbb{P}\left(N_a(t) \leq t^b\right) &\leq \mathbb{P}\left(\exists j \in \left\{0, \dots, \lfloor t^b \rfloor, \xi_j \geq t^{1-b} - 1\right\}\right) \\
&\leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}\left(\xi_j \geq t^{1-b} - 1\right) \\
&\leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}\left(\mu(\theta_a(\tau_j)) < \tau_a\right)^{t^{1-b}-1}.
\end{aligned}$$

CHAPTER VI

Max K-Armed Bandit: On the ExtremeHunter Algorithm and Beyond

Abstract

This chapter is devoted to the study of the *max K-armed bandit problem*, which consists in sequentially allocating resources in order to detect extreme values. Our contribution is twofold. We first significantly refine the analysis of the ExtremeHunter algorithm carried out in [CV14], and next propose an alternative approach, showing that, remarkably, Extreme Bandits can be reduced to a classical version of the bandit problem to a certain extent. Beyond the formal analysis, these two approaches are compared through numerical experiments.

1 Introduction

In a classical multi-armed bandit (MAB in abbreviated form) problem, the objective is to find a strategy/policy in order to sequentially explore and exploit K sources of gain, referred to as *arms*, so as to maximize the expected cumulative gain. Each arm $a \in \{1, \ldots, K\}$ is characterized by an unknown probability distribution ν_a. At each round $t \geq 1$, a strategy π picks an arm $A_t = \pi((A_1, X_{A_1,1}), \ldots, (A_{t-1}, X_{A_{t-1},t-1}))$ and receives a random reward $X_{A_t,t}$ sampled from distribution ν_{A_t}. Whereas usual strategies aim at finding and exploiting the arm with highest expectation, the quantity of interest in many applications such as medicine, insurance or finance may not be the sum of the rewards, but rather the *extreme* observations (even if it might mean replacing loss minimization by gain maximization in the formulation of the practical problem). In such situations, classical bandit algorithms can be significantly sub-optimal: the 'best' arm should not be defined as that with highest expectation, but as that producing the maximal values. This setting, referred to as *extreme bandits* in [CV14], was originally introduced by [CS05] by the name of *max K-armed bandit problem*. In this framework, the goal pursued is to obtain the highest possible reward during the first $T \geq 1$ steps.

For a given arm a, we denote by

$$G_T^{(a)} = \max_{1 \leq t \leq T} X_{a,t}$$

the maximal value taken until round $T \geq 1$ and assume that, in expectation, there is a unique optimal arm

$$a^* = \arg\max_{1 \leq a \leq K} \mathbb{E}[G_T^{(a)}] \, .$$

The *expected extreme regret* of a strategy π is here defined as

$$R_T = \mathbb{E}[G_T^{(a^*)}] - \mathbb{E}[G_T^{(\pi)}], \tag{VI.1}$$

where $G_T^{(\pi)} = \max_{1 \leq t \leq T} X_{A_t,t}$ is the maximal value observed when implementing strategy π. When the supports of the reward distributions (*i.e.* the ν_a's) are bounded, no-regret is expected provided that every arm can be sufficiently explored, refer to [NLB16] (see also [DS16] for a PAC approach). If infinitely many arms are possibly involved in the learning strategy, the challenge is then to explore and exploit optimally the unknown reservoir of arms, see [CV15]. When the rewards are unbounded in contrast, the situation is quite different: the best arm is that for which the maximum $G_T^{(a)}$ tends to infinity faster than the others. In [NLB16], it is shown that, for unbounded distributions, no policy can achieve no-regret without restrictive assumptions on the distributions. In accordance with the literature, we focus on a classical framework in extreme value analysis. Namely, we assume that the reward distributions are *heavy-tailed*. Such Pareto-like laws are widely used to model extremes in many applications, where a conservative approach to risk assessment might be relevant (*e.g.* finance, environmental risks). Like in [CV14], rewards are assumed to be distributed as second order Pareto laws in the present chapter. For the sake of completeness, we recall that a probability law with cdf $F(x)$ belongs to the (α, β, C, C')-second order Pareto family if, for every $x \geq 0$,

$$|1 - Cx^{-\alpha} - F(x)| \leq C' x^{-\alpha(1+\beta)} \, , \tag{VI.2}$$

where α, β, C and C' are strictly positive constants, see *e.g.* [Res07]. In this context, [CV14] have proposed the ExtremeHunter algorithm to solve the *extreme bandit* problem and provided a regret analysis.

The contribution of this chapter is twofold. First, the regret analysis of the ExtremeHunter algorithm is significantly improved, in a nearly optimal fashion. This essentially relies on a new technical result of independent interest (see Theorem 1 below), which provides a bound for the difference between the expectation of the maximum among independent realizations $X_1, \ldots, X_T$ of a (α, β, C, C')-second order Pareto distribution, $\mathbb{E}[\max_{1 \leq t \leq T} X_t]$ namely, and its rough approximation $(TC)^{1/\alpha}\Gamma(1 - 1/\alpha)$. As a by-product, we propose a more simple Explore-Then-Commit strategy that offers the same theoretical guarantees as ExtremeHunter. Second, we explain how extreme bandit can be reduced to a classical bandit problem to a certain extent. We show that a regret-minimizing strategy such as Robust-UCB (see [BCL13]), applied on correctly

left-censored rewards, may also reach a very good performance. This claim is supported by theoretical guarantees on the number of pulls of the best arm a^* and by numerical experiments both at the same time. From a practical angle, the main drawback of this alternative approach consists in the fact that its implementation requires some knowledge of the complexity of the problem (*i.e.* of the gap between the first-order Pareto coefficients of the first and second arms). In regard to its theoretical analysis, efficiency is proved for large horizons only.

This chapter is organized as follows. Section 2 presents the technical result mentioned above, which next permits to carry out a refined regret analysis of the ExtremeHunter algorithm in 3. In 4, the regret bound thus obtained is proved to be nearly optimal: precisely, we establish a lower bound under the assumption that the distributions are close enough to Pareto distributions showing the regret bound is sharp in this situation. In 5, reduction of the extreme bandit problem to a classical bandit problem is explained at length, and an algorithm resulting from this original view is then described. Finally, we provide a preliminary numerical study that permits to compare the two approaches from an experimental perspective.

2 Second-Order Pareto Distributions: Approximation of the Expected Maximum Among i.i.d. Realizations

In the extreme bandit problem, the key to controlling the behavior of explore-exploit strategies is to approximate the expected payoff of a fixed arm $a \in \{1, \ldots, K\}$. The main result of this section, stated in Theorem 1, provides such control: it significantly improves upon the result originaly obtained by [CV14] (see Theorem 1 therein). As shall be next shown in Section 3, this refinement has substantial consequences on the regret bound.

In [CV14], the distance between the expected maximum of independent realizations of a (α, β, C, C')-second order Pareto and the corresponding expectation of a Fréchet distribution $(TC)^{1/\alpha}\Gamma(1-1/\alpha)$ is controlled as follows:

$$\left|\mathbb{E}\left[\max_{1\leq t\leq T} X_t\right] - (TC)^{1/\alpha}\Gamma(1-1/\alpha)\right| \leq \frac{4D_2C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T^{\beta-1/\alpha}} + (2C'T)^{\frac{1}{(1+\beta)\alpha}}.$$

Notice that the leading term of this bound is $(2C'T)^{1/((1+\beta)\alpha)}$ as $T \to +\infty$. Below, we state a sharper result where, remarkably, this (exploding) term disappears, the contribution of the related component in the approximation error decomposition being proved as (asymptotically) negligible in contrast.

Theorem 1. (Fréchet approximation bound) *If $X_1, \ldots, X_T$ are i.i.d. r.v.'s drawn from a (α, β, C, C')-second order Pareto distribution with $\alpha > 1$ and $T \geq Q_1$, where Q_1*

is the constant depending only on α, β, C and C' given in VI.3 below, then,

$$\begin{aligned}&\left|\mathbb{E}\left[\max_{1\le t\le T} X_t\right] - (TC)^{1/\alpha}\Gamma(1-1/\alpha)\right| \\ &\le \frac{4D_2C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T^{\beta-1/\alpha}} + 2(2C'T)^{\frac{1}{(1+\beta)\alpha}} e^{-HT^{\frac{\beta}{\beta+1}}} \\ &= o(T^{1/\alpha}),\end{aligned}$$

where $H = C(2C')^{1/(\alpha(1+\beta))}/2$. In particular, if $\beta \ge 1$, we have:

$$\left|\mathbb{E}\left[\max_{1\le t\le T} X_t\right] - (TC)^{1/\alpha}\Gamma(1-1/\alpha)\right| = o(1) \text{ as } T \to +\infty.$$

We emphasize that the bound above shows that the distance of $\mathbb{E}[\max_{1\le t\le T} X_t]$ to the Fréchet mean $(TC)^{1/\alpha}\Gamma(1-\frac{1}{\alpha})$ actually vanishes as $T \to \infty$ as soon as $\beta \ge 1$, a property that shall be useful in Section 3 to study the behavior of learning algorithms in the extreme bandit setting.

PROOF. Assume that $T \ge Q_1$, where

$$Q_1 = \frac{1}{2C'} \max\left\{(2C'/C)^{(1+\beta)/\beta},\ (8C)^{1+\beta}\right\}. \tag{VI.3}$$

As in the proof of Theorem 1 in [CV14], we consider the quantity $B = (2C'T)^{1/((1+\beta)\alpha)}$ that serves as a cut-off between tail and bulk behaviors. Observe that

$$\begin{aligned}\left|\mathbb{E}\left[\max_{1\le t\le T} X_t\right] - (TC)^{1/\alpha}\Gamma(1-1/\alpha)\right| \le& \\ &\left|\int_0^\infty \left\{1 - \mathbb{P}\left(\max_{1\le t\le T} X_t \le x\right) - 1 + e^{-TCx^{-\alpha}}\right\} \mathrm{d}x\right| \\ &\le \left|\int_0^B \left\{\mathbb{P}\left(\max_{1\le t\le T} X_t \le x\right) - e^{-TCx^{-\alpha}}\right\} \mathrm{d}x\right| \\ &\qquad + \left|\int_B^\infty \left\{\mathbb{P}\left(\max_{1\le t\le T} X_t \le x\right) - e^{-TCx^{-\alpha}}\right\} \mathrm{d}x\right|.\end{aligned}$$

For $p \in \{2, \beta+1\}$, we set $D_p = \Gamma(p - \frac{1}{\alpha})/\alpha$. Equipped with this notation, we may write

$$\left|\int_B^\infty \left\{\mathbb{P}\left(\max_{1\le t\le T} X_t \le x\right) - e^{-TCx^{-\alpha}}\right\} \mathrm{d}x\right| \le \frac{4D_2C^{1/\alpha}}{T^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T^{\beta-1/\alpha}}.$$

Instead of loosely bounding the bulk term by B, we write

$$\left|\int_0^B \left\{\mathbb{P}\left(\max_{1\le t\le T} X_t \le x\right) - e^{-TCx^{-\alpha}}\right\} \mathrm{d}x\right| \le B\,\mathbb{P}\left(X_1 \le B\right)^T + \int_0^B e^{-TCx^{-\alpha}} \mathrm{d}x\ . \tag{VI.4}$$

First, using (VI.2) and the inequality $C'B^{-(1+\beta)\alpha} \leq CB^{-\alpha}/2$ (a direct consequence of VI.3), we obtain

$$\begin{aligned}\mathbb{P}(X_1 \leq B)^T &\leq \left(1 - CB^{-\alpha} + C'B^{-(1+\beta)\alpha}\right)^T \\ &\leq \left(1 - \frac{1}{2}CB^{-\alpha}\right)^T \leq e^{-\frac{1}{2}TCB^{-\alpha}} = e^{-HT^{\beta/(\beta+1)}}.\end{aligned}$$

Second, the integral in VI.4 can be bounded as follows:

$$\int_0^B e^{-TCx^{-\alpha}} \mathrm{d}x \leq Be^{-TCB^{-\alpha}} = (2C'T)^{1/((1+\beta)\alpha)} e^{-2HT^{\beta/(\beta+1)}}.$$

This concludes the proof.

3 The ExtremeHunter and ExtremeETC Algorithms

In this section, the tighter control provided by Theorem 1 is used in order to refine the analysis of the ExtremeHunter algorithm (Algorithm 3) carried out in [CV14]. This theoretical analysis is also shown to be valid for ExtemeETC, a novel algorithm we next propose, that greatly improves upon ExtremeHunter, regarding computational efficiency.

3.1 Further Notations and Preliminaries

Throughout the chapter, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{I}\{\mathcal{E}\}$ and $\bar{\mathcal{E}}$ means the complementary event of $\mathcal{E}$. We assume that the reward related to each arm $a \in \{1, \ldots, K\}$ is drawn from a $(\alpha_a, \beta_a, C_a, C')$-second order Pareto distribution. Sorting the tail indices by increasing order of magnitude, we use the classical notation for order statistics: $\alpha_{(1)} \leq \cdots \leq \alpha_{(K)}$. We assume that $\alpha_{(1)} > 1$, so that the random rewards have finite expectations, and suppose that the strict inequality $\alpha_{(1)} < \alpha_{(2)}$ holds true. We also denote by $N_a(t)$ the number of times the arm a is pulled up to time t. For $1 \leq a \leq K$ and $t \geq 1$, the r.v. $\widetilde{X}_{a,t}$ is the reward obtained at the t-th draw of arm a if $t \leq N_a(T)$ or a new r.v. drawn from ν_a independent from the other r.v.'s otherwise.

We start with a preliminary lemma supporting the intuition that the tail index α fully governs the extreme bandit problem. It will allow to show next that the algorithm picks the right arm after the exploration phase, see Lemma 2.

Lemma 1. (Optimal arm) *For T larger than some constant Q_4 depending only on $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ and C', the optimal arm for the extreme bandit problem is given by:*

$$a^* = \underset{1\leq a\leq K}{arg\ min}\, \alpha_a = \underset{1\leq a\leq K}{arg\ max}\, V_a, \tag{VI.5}$$

where $V_a = (TC_a)^{1/\alpha_a}\Gamma(1 - \frac{1}{\alpha_a})$.

PROOF. We first prove the first equality. It follows from Theorem 1 that there exists a constant Q_2, depending only on $\{(\alpha_a, \beta_a, C_a)\}_{1\le a\le K}$ and C', such that for any arm $a \in \{1, \ldots, K\}$, $|\mathbb{E}[G_T^{(a)}] - V_a| \le V_a/2$. Then for $a \neq a^*$ we have, for all $T > Q_2$, $V_a/2 \le \mathbb{E}[G_T^{(a)}] \le \mathbb{E}[G_T^{(a^*)}] \le 3V_{a^*}/2$. Recalling that V_a is proportional to T^{1/α_a}, it follows that $\alpha_{a^*} = \min_{1\le a\le K} \alpha_a$. Now consider the following quantity:

$$Q_3 = \max_{a\neq a^*} \left[\frac{2C_a^{\frac{1}{\alpha_a}} \Gamma(1-\frac{1}{\alpha_a})}{C_{a^*}^{\frac{1}{\alpha_{a^*}}} \Gamma(1-\frac{1}{\alpha_{a^*}})} \right]^{\frac{1}{\frac{1}{\alpha_{a^*}} - \frac{1}{\alpha_a}}} . \tag{VI.6}$$

For $T > Q_4 = \max(Q_2, Q_3)$, we have $V_{a^*} > 2V_a$ for any suboptimal arm $a \neq a^*$, which proves the second equality.

From now on, we assume that T is large enough for Lemma 1 to apply.

3.2 The ExtremeHunter Algorithm ([CV14])

Before developing a novel analysis of the extreme bandit problem in Section 3.2 (see Theorem 2), we recall the main features of ExtremeHunter, and in particular the estimators and confidence intervals involved in the indices of this optimistic policy.

Algorithm 3 ExtremeHunter ([CV14])

Input: K: number of arms, T: time horizon, $b > 0$ such that $b \le \min_{1\le a\le K} \beta_a$, N: minimum number of pulls of each arm (Eq. (VI.9)).
Initialize: Pull each arm N times.
for $a = 1, \ldots, K$ **do**
  Compute estimators $\widehat{h}_{a,KN} = \widetilde{h}_a(N)$ (Eq. (VI.8)) and $\widehat{C}_{a,KN} = \widetilde{C}_a(N)$ (Eq. (VI.7))
  Compute index $B_{a,KN}$ (Eq. (VI.12))
end for
Pull arm $A_{KN+1} = \arg\max_{1\le a\le K} B_{a,KN}$
for $t = KN + 2, \ldots, T$ **do**
  Update estimators $\widehat{h}_{A_{t-1},t-1}$ and $\widehat{C}_{A_{t-1},t-1}$
  Update index $B_{A_{t-1},t-1}$
  Pull arm $A_t = \arg\max_{1\le a\le K} B_{a,t-1}$
end for

Theorem 1 states that for any arm $a \in \{1, \ldots, K\}$, $\mathbb{E}[G_T^{(a)}] \approx (C_a T)^{1/\alpha_a}\Gamma(1 - 1/\alpha_a)$. Consequently, the optimal strategy in hindsight always pulls the arm $a^* = \arg\max_{1\le a\le K}\{(TC_a)^{1/\alpha_a}\Gamma(1-1/\alpha_a)\}$. At each round and for each arm $a \in \{1, \ldots, K\}$, ExtremeHunter algorithm ([CV14]) estimates the coefficients α_a and C_a (but not β_a, see Remark 2 in [CV14]). The corresponding confidence intervals are detailed below. Then, following the *optimism-in-the-face-of-uncertainty* principle (see [ACBF02] and references therein), the strategy plays the arm maximizing an optimistic plug-in estimate

of $(C_aT)^{1/\alpha_a}\Gamma(1-1/\alpha_a)$. To that purpose, Theorem 3.8 in [CK14a] and Theorem 2 in [CK+14b] provide estimators $\widetilde{\alpha}_a(T')$ and $\widetilde{C}_a(T')$ for α_a and C_a respectively, after T' draws of arm a. Precisely, the estimate $\widetilde{\alpha}_a(T')$ is given by

$$\widetilde{\alpha}_a(T') = \log\left(\frac{\sum_{t=1}^{T'}\mathbb{I}\{X_t > e^r\}}{\sum_{t=1}^{T'}\mathbb{I}\{X_t > e^{r+1}\}}\right),$$

where r is chosen in an adaptive fashion based on Lepski's method, see [Lep90], while the estimator of C_a considered is

$$\widetilde{C}_a(T') = T'^{-2b/(2b+1)}\sum_{t=1}^{T'}\mathbb{I}\{\widetilde{X}_{a,t} \geq T'^{\widetilde{h}_a(T')/(2b+1)}\}, \tag{VI.7}$$

where

$$\widetilde{h}_a(T') = \min(1/\widetilde{\alpha}_a(T'), 1)\ . \tag{VI.8}$$

The authors also provide finite sample error bounds for $T' \geq N$, where

$$N = A_0(\log T)^{2(2b+1)/b}, \tag{VI.9}$$

with b a known lower bound on the β_a's ($b \leq \min_{1\leq a\leq K}\beta_a$), and A_0 a constant depending only on $(\alpha_a, \beta_a, C_a)_{1\leq k\leq K}$ and C'. These error bounds naturally define confidence intervals of respective widths Λ_1 and Λ_2 at level δ_0 defined by

$$\delta_0 = T^{-\rho}, \quad \text{where} \quad \rho = \frac{2\alpha_{a^*}}{\alpha_{a^*}-1}. \tag{VI.10}$$

More precisely, we have

$$\mathbb{P}\left(\left|\frac{1}{\alpha_a} - \widetilde{h}_a(T')\right| \leq \Lambda_1(T'),\ \left|C_a - \widetilde{C}_a(T')\right| \leq \Lambda_2(T')\right) \geq 1-2\delta_0, \tag{VI.11}$$

where

$$\Lambda_1(T') = D\sqrt{\log(1/\delta_0)}T'^{-b/(2b+1)} \text{ and } \Lambda_2(T') = E\sqrt{\log(T'/\delta_0)}\log(T')T'^{-b/(2b+1)},$$

denoting by D and E some constants depending only on $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ and C'. When $N_a(t) \geq N$, denote by $\widehat{h}_{a,t} = \widetilde{h}_a(N_a(t))$ and $\widehat{C}_{a,t} = \widetilde{C}_a(N_a(t))$ the estimators based on the $N_a(t)$ observations for simplicity. ExtremeHunter's index $B_{a,t}$ for arm a at time t, the optimistic proxy for $\mathbb{E}[G_T^{(a)}]$, can be then written as

$$B_{a,t} = \widetilde{\Gamma}\left(1-\widehat{h}_{a,t}-\Lambda_1(N_a(t))\right)\left(\left(\widehat{C}_{a,t}+\Lambda_2(N_a(t))\right)T\right)^{\widehat{h}_{a,t}+\Lambda_1(N_a(t))}, \tag{VI.12}$$

where $\widetilde{\Gamma}(x) = \Gamma(x)$ if $x > 0$ and $+\infty$ otherwise.

On Computational Complexity. Notice that after the initialization phase, at each time $t > KN$, ExtremeHunter computes estimators $\widehat{h}_{A_t,t}$ and $\widehat{C}_{A_t,t}$, each having a time complexity linear with the number of samples $N_{A_t}(t)$ pulled from arm A_t up to time t. Summing on the rounds reveals that ExtremeHunter's time complexity is quadratic with the time horizon T.

Complexity	ExtremeETC	ExtremeHunter
Time	$O\big((\log T)^{\frac{2(2b+1)}{b}}\big)$	$O(T^2)$
Memory	$O\big((\log T)^{\frac{2(2b+1)}{b}}\big)$	$O(T)$

Table VI.1: Time and memory complexities required for estimating $(\alpha_a, C_a)_{1\leq a\leq K}$ in ExtremeETC and ExtremeHunter.

3.3 ExtremeETC: A Computationally Appealing Alternative

In order to reduce the restrictive time complexity discussed previously, we now propose the ExtremeETC algorithm, an *Explore-Then-Commit* version of ExtremeHunter, which offers similar theoretical guarantees.

Algorithm 4 ExtremeETC

Input: K: number of arms, T: time horizon, $b > 0$ such that $b \leq \min_{1\leq a\leq K} \beta_a$, N: minimum number of pulls of each arm (Eq. (VI.9)).
Initialize: Pull each arm N times.
for $a = 1, \ldots, K$ **do**
  Compute estimators $\widehat{h}_{a,KN} = \widetilde{h}_a(N)$ (Eq. (VI.8)) and $\widehat{C}_{a,KN} = \widetilde{C}_a(N)$ (Eq. (VI.7))
  Compute index $B_{a,KN}$ (Eq. (VI.12))
end for
Set $a_{\text{winner}} = \arg\max_{1\leq a\leq K} B_{a,KN}$
for $t = KN + 1, \ldots, T$ **do**
  Pull arm a_{winner}
end for

After the initialization phase, the *winner arm*, which has maximal index $B_{a,KN}$, is fixed and is pulled in all remaining rounds. Then ExtremeETC's time complexity, due to the computation of $\widehat{h}_{a,KN}$ and $\widehat{C}_{a,KN}$ only, is $O(KN) = O\left((\log T)^{2(2b+1)/b}\right)$, which is considerably faster than quadratic time achieved by ExtremeHunter. For clarity, Table VI.1 summarizes time and memory complexities of both algorithms.

Due to the significant gain of computational time, we used the ExtremeETC algorithm in our simulation study (Section 6) rather than ExtremeHunter.

Controlling the Number of Suboptimal Rounds. We introduce a high probability event that corresponds to the favorable situation where, at each round, all coefficients $(1/\alpha_a, C_a)_{1\leq k\leq K}$ simultaneously belong to the confidence intervals recalled in the previous subsection.

Definition 1. *The event ξ_1 is the event on which the bounds*

$$\left|\frac{1}{\alpha_a} - \widetilde{h}_a(T')\right| \leq \Lambda_1(T') \quad \textit{and} \quad \left|C_a - \widetilde{C}_a(T')\right| \leq \Lambda_2(T')$$

hold true for any $1 \leq a \leq K$ *and* $N \leq T' \leq T$.

The union bound combined with (VI.11) yields

$$\mathbb{P}(\xi_1) \geq 1 - 2KT\delta_0. \tag{VI.13}$$

Lemma 2. *For* $T > Q_5$*, where* Q_5 *is the constant defined in* (VI.15)*,* ExtremeETC *and* ExtremeHunter *always pull the optimal arm after the initialization phase on the event* ξ_1*. Hence, for any suboptimal arm* $a \neq a^*$*, we have on* ξ_1*:*

$$N_a(T) = N \quad \textit{and thus} \quad N_{a^*}(T) = T - (K-1)N.$$

proof. Here we place ourselves on the event ξ_1. For any arm $1 \leq a \leq K$, Lemma 1 in [CV14] provides lower and upper bounds for $B_{a,t}$ when $N_a(t) \geq N$

$$V_a \leq B_{a,t} \leq V_a \left(1 + F \log T \sqrt{\log(T/\delta_0)} N_a(t)^{-b/(2b+1)}\right), \tag{VI.14}$$

where F is a constant which depends only on $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ and C'. Introduce the horizon Q_5, which depends on $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ and C'

$$Q_5 = \max\left(e^{\left(F\sqrt{1+\rho}A_0^{-b/(2b+1)}\right)^2}, Q_4\right). \tag{VI.15}$$

Then the following Lemma 3, proved in section 8, tells us that for T large enough, the exploration made during the initialization phase is enough to find the optimal arm, with high probability.

Lemma 3. *If* $T > Q_5$*, we have under the event* ξ_1 *that for any suboptimal arm* $a \neq a^*$ *and any time* $t > KN$ *that* $B_{a,t} < B_{a^*,t}$ *.*

Hence the optimal arm is pulled at any time $t > KN$.

The following result immediately follows from Lemma 2.

Corollary 1. *For* T *larger than some constant depending only on* $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ *and* C' *we have under* ξ_1

$$N_{a^*}(T) \geq T/2.$$

Upper Bounding the Expected Extreme Regret. The upper bound on the expected extreme regret stated in the theorem below improves upon that given in [CV14] for ExtremeHunter. It is also valid for ExtremeETC.

Theorem 2. *For* ExtremeETC *and* ExtremeHunter*, the expected extreme regret is upper bounded as follows*

$$R_T = O\left((\log T)^{2(2b+1)/b} T^{-(1-1/\alpha_{a^*})} + T^{-(b-1/\alpha_{a^*})}\right),$$

as $T \to +\infty$*. If* $b \geq 1$*, we have in particular* $R_T = o(1)$ *as* $T \to +\infty$*.*

The proof of Theorem 2 is deferred to section 8. It closely follows that of Theorem 2 in [CV14], the main difference being that their concentration bound (Theorem 1 therein) can be replaced by our tighter bound (see Theorem 1 in the present chapter). Recall that in Theorem 2 in [CV14], the upper bound on the expected extreme regret for ExtremeHunter goes to infinity when $T \to +\infty$:

$$R_T = O\left(T^{\frac{1}{(1+b)\alpha_{a^*}}}\right). \tag{VI.16}$$

In contrast, in Theorem 2 when $b \geq 1$, the upper bound obtained vanishes when $T \to +\infty$. In the case $b < 1$, the upper bound still improves upon Eq. (VI.16) by a polynomial factor $T^{(\alpha_{a^*} b(b+1)-b)/((b+1)\alpha_{a^*})} > T^{b^2/(2\alpha_{a^*})}$.

4 Lower Bound on the Expected Extreme Regret

In this section we prove a lower bound on the expected extreme regret for ExtremeETC and ExtremeHunter in specific cases. We assume now that $\alpha_{(2)} > 2\alpha_{a^*}^2/(\alpha_{a^*} - 1)$ and we start with a preliminary result on second order Pareto distributions, proved in 8.

Lemma 4. *If X is a r.v. drawn from a (α, β, C, C')-second order Pareto distribution and r is a strictly positive constant, the distribution of the r.v. X^r is a $(\alpha/r, \beta, C, C')$-second order Pareto.*

In order to prove the lower bound on the expected extreme regret, we first establish that the event corresponding to the situation where the highest reward obtained by ExtremeETC and ExtremeHunter comes from the optimal arm a^* occurs with overwhelming probability. Precisely, we denote by ξ_2 the event such that the bound

$$\max_{a \neq a^*} \max_{1 \leq t \leq N} \widetilde{X}_{a,t} \leq \max_{1 \leq t \leq T-(K-1)N} \widetilde{X}_{a^*,t}.$$

holds true. The following lemma, proved in 8, provides a control of its probability of occurence.

Lemma 5. *For T larger than some constant depending only on $(\alpha_a, \beta_a, C_a)_{1 \leq a \leq K}$ and C', the following assertions hold true.*

(i) We have:

$$\mathbb{P}(\xi_2) \geq 1 - K\delta_0,$$

where δ_0 is given in VI.10.

(ii) Under the event $\xi_0 = \xi_1 \cap \xi_2$, the maximum reward obtained by ExtremeETC *and* ExtremeHunter *comes from the optimal arm:*

$$\max_{1 \leq t \leq T} X_{A_t,t} = \max_{1 \leq t \leq T-(K-1)N} \widetilde{X}_{a^*,t}.$$

The following lower bound shows that the upper bound (2) is actually tight in the case $b \geq 1$.

Theorem 3. *If $b \geq 1$ and $\alpha_{(2)} > 2\alpha_{a^*}^2/(\alpha_{a^*} - 1)$, the expected extreme regret of* ExtremeETC *and* ExtremeHunter *are lower bounded as follows*

$$R_T = \Omega\left((\log T)^{2(2b+1)/b} T^{-(1-1/\alpha_{a^*})}\right) .$$

proof. Here, π refers to either ExtremeETC or else ExtremeHunter. In order to bound from below $R_T = \mathbb{E}[G_T^{(a^*)}] - \mathbb{E}[G_T^{(\pi)}]$, we start with bounding $\mathbb{E}[G_T^{(\pi)}]$ as follows

$$\begin{aligned}\mathbb{E}\left[G_T^{(\pi)}\right] &= \mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t}\right] = \mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t}\mathbb{I}\{\xi_0\}\right] + \mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t}\mathbb{I}\{\bar{\xi}_0\}\right] \\ &\leq \mathbb{P}(\xi_0)\mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t} \;\middle|\; \xi_0\right] + \sum_{a=1}^{K}\mathbb{E}\left[\max_{1\leq t\leq N_a(T)} \widetilde{X}_{a,t}\mathbb{I}\{\bar{\xi}_0\}\right] , \qquad \text{(VI.17)}\end{aligned}$$

where $\widetilde{X}_{a,t}$ has been defined in 3.1. From (ii) in 5, we have

$$\mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t} \;\middle|\; \xi_0\right] = \mathbb{E}\left[\max_{1\leq t\leq T-(K-1)N} \widetilde{X}_{a^*,t} \;\middle|\; \xi_0\right] . \qquad \text{(VI.18)}$$

In addition, in the sum of expectations on the right-hand-side of VI.17, $N_a(T)$ may be roughly bounded from above by T. A straightforward application of Hölder inequality yields

$$\sum_{a=1}^{K}\mathbb{E}\left[\max_{1\leq t\leq N_a(T)} \widetilde{X}_{a,t}\mathbb{I}\{\bar{\xi}_0\}\right] \leq \sum_{a=1}^{K}\left(\mathbb{E}\left[\max_{1\leq t\leq T} \widetilde{X}_{a,t}^{\frac{\alpha_{a^*}+1}{2}}\right]\right)^{\frac{2}{\alpha_{a^*}+1}} \mathbb{P}\left(\bar{\xi}_0\right)^{\frac{\alpha_{a^*}-1}{\alpha_{a^*}+1}} . \qquad \text{(VI.19)}$$

From (i) in 5 and VI.13, we have $\mathbb{P}(\bar{\xi}_0) \leq K(2T+1)\delta_0$. By virtue of 4, the r.v. $\widetilde{X}_{a,t}^{(\alpha_{a^*}+1)/2}$ follows a $(2\alpha_a/(\alpha_{a^*}+1), \beta_a, C_a, C')$-second order Pareto distribution. Then, applying 1 to the right-hand side of (VI.19) and using the identity (VI.18), the upper bound (VI.17) becomes

$$\begin{aligned}&\mathbb{E}\left[G_T^{(\pi)}\right] \leq \mathbb{E}\left[\max_{1\leq t\leq T-(K-1)N} \widetilde{X}_{a^*,t}\mathbb{I}\{\xi_0\}\right] \\ &+ \sum_{a=1}^{K}\left((TC_a)^{\frac{\alpha_{a^*}+1}{2\alpha_a}}\Gamma\left(1-\frac{\alpha_{a^*}+1}{2\alpha_a}\right) + o\left(T^{\frac{\alpha_{a^*}+1}{2\alpha_a}}\right)\right)^{\frac{2}{\alpha_{a^*}+1}} \left(K(2T+1)\delta_0\right)^{\frac{\alpha_{a^*}-1}{\alpha_{a^*}+1}} \\ &\leq \mathbb{E}\left[\max_{1\leq t\leq T-(K-1)N} \widetilde{X}_{a^*,t}\right] + O\left(T^{-(1-1/\alpha_{a^*})}\right), \qquad \text{(VI.20)}\end{aligned}$$

where the last inequality comes from the definition of δ_0. Combining 1 and (VI.20) we finally obtain the desired lower bound

$$\begin{aligned} R_T &= \mathbb{E}\left[G_T^{(a^*)}\right] - \mathbb{E}\left[G_T^{(\pi)}\right] \\ &\geq \Gamma(1-1/\alpha_{a^*})C_{a^*}^{1/\alpha_{a^*}}\left(T^{1/\alpha_{a^*}} - (T-(K-1)N)^{1/\alpha_{a^*}}\right) + O\left(T^{-(1-1/\alpha_{a^*})}\right) \\ &= \frac{\Gamma(1-1/\alpha_{a^*})C_{a^*}^{1/\alpha_{a^*}}}{\alpha_{a^*}}(K-1)NT^{-(1-1/\alpha_{a^*})} + O\left(T^{-(1-1/\alpha_{a^*})}\right), \end{aligned}$$

where we used a Taylor expansion of $x \mapsto (1+x)^{1/\alpha_{a^*}}$ at zero for the last equality.

5 A Reduction to Classical Bandits

The goal of this section is to render explicit the connections between the max K-armed bandit considered in the present chapter and a particular instance of the classical Multi-Armed Bandit (MAB) problem.

5.1 MAB Setting for Extreme Rewards

In a situation where only the large rewards matter, an alternative to the max K-armed problem would be to consider the expected cumulative sum of the most 'extreme' rewards, that is, those which exceeds a given high threshold u. For $a \in \{1, \ldots, K\}$ and $t \in \{1, \ldots, T\}$, we denote by $Y_{a,t}$ these new rewards

$$Y_{a,t} = X_{a,t}\mathbb{I}\{X_{a,t} > u\} \, .$$

In this context, the classical MAB problem consists in maximizing the expected cumulative gain

$$\mathbb{E}\left[G^{\text{MAB}}\right] = \mathbb{E}\left[\sum_{t=1}^{T} Y_{A_t,t}\right].$$

It turns out that for a high enough threshold u, the unique optimal arm for this MAB problem, $\arg\max_{1\leq a\leq K}\mathbb{E}[Y_{a,1}]$, is also the optimal arm a^* for the max K-armed problem. We still assume second order Pareto distributions for the random variables $X_{a,t}$ and that all the hypothesis listed in Section 3.1 hold true. The rewards $\{Y_{a,t}\}_{1\leq a\leq K, 1\leq t\leq T}$ are also heavy-tailed so that it is legitimate to attack this MAB problem with the Robust UCB algorithm ([BCL13]), which assumes that the rewards have finite moments of order $1+\epsilon$

$$\max_{1\leq a\leq K}\mathbb{E}\left[|Y_{a,1}|^{1+\epsilon}\right] \leq v \, , \tag{VI.21}$$

where $\epsilon \in (0,1]$ and $v > 0$ are known constants. Given our second order Pareto assumptions, it follows that Eq. (VI.21) holds with $1+\epsilon < \alpha_{(1)}$. Even if the knowledge of such constants ϵ and v is a strong assumption, it is still fair to compare Robust

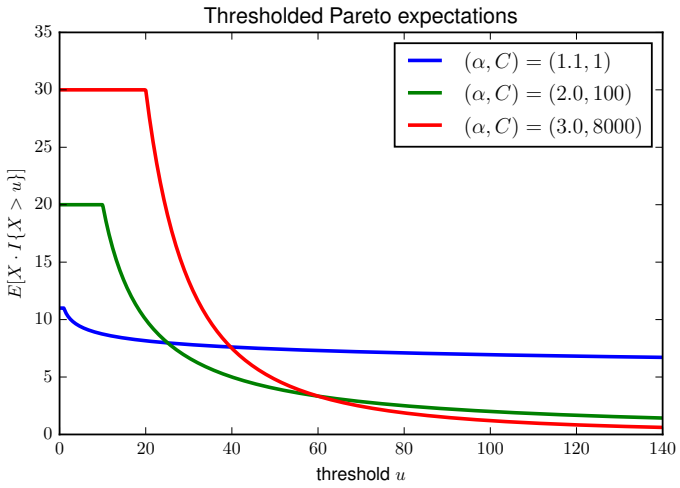


Figure VI.1: Expected values $\mathbb{E}[Y] = \frac{\alpha}{\alpha-1}\frac{C}{u^{\alpha-1}}$ of the thresholded rewards $Y = X\mathbb{I}\{X > u\}$ (with X an (α, C)-Pareto r.v.) as a function of the thresholding level u.

UCB to ExtremeETC/Hunter, which also has strong requirements. Indeed, ExtremeETC/Hunter assumes that b and T are known and verify conditions depending on unknown problem parameters (e.g. $T \geq Q_1$, see Eq. (VI.3)).

The following Lemma, whose proof is postponed to section 8, ensures that the two bandit problems are equivalent for high thresholds.

Lemma 6.

$$\textit{If} \quad u > \max\left(1, \left(\frac{2C'}{\min_{1\leq a\leq K} C_a}\right)^{\frac{1}{\min_{1\leq a\leq K}\beta_a}}, \left(\frac{3\max_{1\leq a\leq K} C_a}{\min_{1\leq a\leq K} C_a}\right)^{\frac{1}{\alpha_{(2)}-\alpha_{(1)}}}\right), \quad \text{(VI.22)}$$

then the unique best arm for the MAB problem is arg $min_{1\leq a\leq K}\alpha_a = a^*$.

Remark 1. *Tuning the threshold u based on the data is a difficult question, outside our scope. A standard practice is to monitor a relevant output (e.g. estimate of α) as a function of the threshold u and to pick the latter as low as possible in the stability region of the output. This is related to the Lepski's method, see e.g. [BT15], [CK14a], [HW85].*

5.2 Robust UCB Algorithm ([BCL13])

For the sake of completeness, we recall below the main feature of Robust UCB and make explicit its theoretical guarantees in our setting. The bound stated in the following proposition is a direct consequence of the regret analysis conducted by [BCL13].

Proposition 1. *Applying the* Robust UCB *algorithm of [BCL13] to our MAB problem, the expected number of times we pull any suboptimal arm $a \neq a^*$ is upper bounded as follows*

$$\mathbb{E}[N_a(T)] = O\left(\log T\right) \ .$$

proof. See proof of Proposition 1 in [BCL13].

Hence, in expectation, Robust UCB pulls fewer times suboptimal arms than ExtremeETC/Hunter. Indeed with ExtremeETC/Hunter,

$$N_a(T) \geq N = \Theta((\log T)^{2(2b+1)/b}).$$

Remark 2. *Proposition 1 may be an indication that the Robust UCB approach performs better than* ExtremeETC/Hunter. *Nevertheless, guarantees on its expected extreme regret require sharp concentration bounds on $N_a(T)$ ($a \neq a^*$), which is out of the scope of this chapter.*

Algorithm 5 Robust UCB with truncated mean estimator ([BCL13])

Input: $u > 0$ s.t. Eq. (VI.22) holds, $\epsilon \in (0, 1]$ and $v > 0$ s.t. Eq. (VI.21) holds.
Initialize: Pull each arm once.
for $t \geq K + 1$ **do**
  for $a = 1, \ldots, K$ **do**
    Update truncated mean estimator
    $\widehat{\mu_a} \leftarrow \frac{1}{N_a(t-1)} \sum_{s=1}^{t-1} Y_{a,s} \mathbb{I}\left\{A_s = a, Y_{a,s} \leq \left(\frac{v N_a(s)}{\log(t^2)}\right)^{\frac{1}{1+\epsilon}}\right\}$
    Update index
    $B_a \leftarrow \widehat{\mu_a} + 4v^{1/(1+\epsilon)} \left(\frac{\log t^2}{N_a(t-1)}\right)^{\epsilon/(1+\epsilon)}$
  end for
  Play arm $A_t = \arg\ \max_{1 \leq a \leq K} B_a$
end for

6 Numerical Experiments

In order to illustrate some aspects of the theoretical results presented previously, we consider a time horizon $T = 10^5$ with $K = 3$ arms and exact Pareto distributions with parameters given in VI.2. Here, the optimal arm is the second one (incidentally, the distribution with highest mean is the first one).

We have implemented Robust UCB with parameters $\epsilon = 0.4$, which satisfies $1 + \epsilon < \alpha_2 = 1.5$, v achieving the equality in VI.21 (ideal case) and a threshold u equal to the lower bound in VI.22 plus 1 to respect the strict inequality. ExtremeETC is runned with $b = 1 < +\infty = \min_{1 \leq a \leq K} \beta_a$. In this setting, the most restrictive condition on the time horizon, $T > KN \approx 7000$ (given by VI.9), is checked, which places us in the validity framework of ExtremeETC. The resulting strategies are compared to each other and

	Arm 1	Arm $a^* = 2$	Arm 3
α_a	15	1.5	10
C_a	10^8	1	10^5
$\mathbb{E}[X_{a,1}]$	3.7	3	3.5
$\mathbb{E}[\max_{1\leq t\leq T} X_{a,t}]$	7.7	$\mathbf{5.8 \cdot 10^3}$	11

Table VI.2: Pareto distributions used in the experiments.

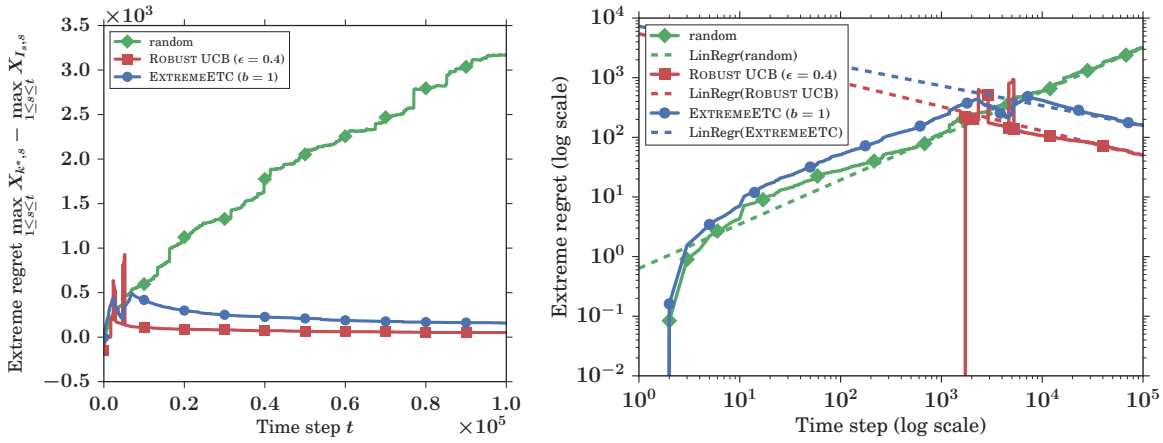


Figure VI.2: Averaged extreme regret (over 1000 independent simulations) for ExtremeETC, Robust UCB and a uniformly random strategy. VI.2b is the log-log scaled counterpart of VI.2a with linear regressions computed over $t = 5 \cdot 10^4, \ldots, 10^5$.

to the random strategy pulling each arm uniformly at random, but not to Threshold Ascent algorithm [SS06] which is designed only for bounded rewards. Precisely, 1000 simulations have been run and Figure VI.2 depicts the extreme regret (VI.1) in each setting averaged over these 1000 trajectories. These experiments empirically support the theoretical bounds in 2: the expected extreme regret of ExtremeETC converges to zero for large horizons. On the log-log scale (VI.2b), ExtremeETC's extreme regret starts linearly decreasing after the initialization phase, at $T > KN \approx 7000$, which is consistent with 2. The corresponding linear regression reveals a slope ≈ -0.333 (with a coefficient of determination $R^2 \approx 0.97$), which confirms 2 and 3 yielding the theoretical slope $-(1 - 1/\alpha_{a^*}) = -1/3$.

7 Conclusion

This chapter brings two main contributions. It first provides a refined regret bound analysis of the performance of the ExtremeHunter algorithm in the context of the

max K-armed bandit problem that significantly improves upon the results obtained in the seminal contribution [CV14], also proved to be valid for ExtremeETC, a computationally appealing alternative we introduce. In particular, the obtained upper bound on the regret converges to zero for large horizons and is shown to be tight when the tail of the rewards is sufficiently close to a Pareto tail (second order parameter $b \geq 1$). On the other hand, this chapter offers a novel view of this approach, interpreted here as a specific version of a classical solution (*Robust UCB*) of the MAB problem, in the situation when only very large rewards matter. Based on these encouraging results, several lines of further research can be sketched. In particular, future work could investigate to which extent the lower bound established for ExtremeETC/Hunter holds true for any strategy with exploration stage of the same duration, and whether improved performance is achievable with alternative stopping criteria for the exploration stage.

8 Technical Proofs

This section contains the proofs of some results stated in the present chapter.

Proof of Lemma 3

For $T > Q_3$ (defined in Eq. (VI.6)), one has $V_{a^*} > 2 \max_{a \neq a^*} V_a$, which implies that $\max_{a \neq a^*} V_a/(V_{a^*} - V_a) < 1$. Hence

$$\max_{a \neq a^*} e^{\left(F\sqrt{1+\rho} A_0^{-b/(2b+1)} \frac{V_a}{V_{a^*}-V_a}\right)^2} < e^{\left(F\sqrt{1+\rho} A_0^{-b/(2b+1)}\right)^2} \leq Q_5 \, .$$

Then, as $N_a(t) \geq N$ and by definitions of N (Eq. (VI.9)) and δ_0 (Eq. (VI.10)), we have for $T > Q_5$ that for any suboptimal arm $a \neq a^*$

$$\begin{aligned} &(C_a T)^{1/\alpha_a} \Gamma(1 - 1/\alpha_a) \left(1 + F \log T \sqrt{\log(T/\delta_0)} N_a(t)^{-b/(2b+1)}\right) \\ &< (C_{a^*} T)^{1/\alpha_{a^*}} \Gamma(1 - 1/\alpha_{a^*}) \, , \end{aligned}$$

which implies, using Eq. (VI.14), that under ξ_1: $B_{a,t} < B_{a^*,t}$ for $t > KN$.

Proof of Theorem 2

We want to upper bound $R_T = \mathbb{E}[G_T^{(a^*)}] - \mathbb{E}[G_T^{(\pi)}]$. To do so, we lower bound $\mathbb{E}[G_T^{(\pi)}]$ as follows

$$\mathbb{E}\left[G_T^{(\pi)}\right] = \mathbb{E}\left[\max_{t \leq T} X_{A_t,t}\right] \geq \mathbb{E}\left[\max_{\{t \leq T, A_t = a^*\}} X_{A_t,t}\right] = \mathbb{E}\left[\max_{\{t \leq N_{a^*}(T)\}} \widetilde{X}_{a^*,t}\right] .$$

Thus

$$\mathbb{E}\left[G_T^{(\pi)}\right] \geq \mathbb{E}\left[\max_{t \leq T-(K-1)N} \widetilde{X}_{a^*,t} \mathbb{I}\{\xi_1\}\right] ,$$

where we used that under ξ_1, $N_{a^*}(T) = T - (K-1)N$. Now we call the following result (Lemma 7, proved in section 8), giving a lower bound on the expected maximum of i.i.d. second order Pareto r.v. given some event.

Lemma 7. *Let $X_1, ..., X_{T'}$ be i.i.d. samples from an (α, β, C, C')-second order Pareto distribution. Let ξ be an event of probability larger than $1-\delta$. If $\delta < 1/2$ and $T' \geq \max\left(4c, (4c)^{1/\beta}\log(2)C(2C')^{1/\beta}, 8\log^2(2)\right)$ for a given constant c depending only on β, C and C', we have*

$$\mathbb{E}\left[\max_{1\leq t\leq T'} X_t \mathbb{I}\{\xi\}\right] \geq (T'C)^{1/\alpha}\Gamma\left(1-\frac{1}{\alpha}\right) - \left(4+\frac{8}{\alpha-1}\right)(T'C)^{1/\alpha}\delta'^{1-1/\alpha}$$
$$- 2\left(\frac{4D_2C^{1/\alpha}}{T'^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T'^{\beta-1/\alpha}} + 2(2C'T')^{1/(\alpha(1+\beta))}e^{-HT'^{\beta/(\beta+1)}}\right).$$

Then, applying Lemma 7 with $\xi = \xi_1$ and $\delta = \delta_0$ we obtain after simplification

$$R_T \leq H'T^{1/\alpha_{a^*}}\Big\{\frac{1}{T} + \frac{1}{T^b} + \frac{K}{T}(\log T)^{2(2b+1)/b} + \delta_0^{1-1/\alpha_{a^*}}$$
$$+ T^{1/(\alpha_{a^*}(1+\beta_{a^*}))}e^{-H_{a^*}(T/2)^{\beta/(\beta+1)}}\Big\},$$

where $H_{a^*} = \frac{1}{2}C_{a^*}(2C')^{1/(\alpha_{a^*}(1+\beta_{a^*}))}$ and H' is a constant depending only on $(\alpha_a, \beta_a, C_a)_{1\leq a\leq K}$ and C'. The definition of δ_0 concludes the proof.

Proof of Lemma 7

We follow the proof of Lemma 2 in [CV14] except that we use Theorem 1 instead of their Theorem 1. Let x_δ be such that $\mathbb{P}(\max_{1\leq t\leq T'} X_t \leq x_\delta) = 1-\delta$. Then we have

$$\mathbb{E}\left[\max_{1\leq t\leq T'} X_t \mathbb{I}\{\xi\}\right] = \mathbb{E}\left[\max_{1\leq t\leq T'} X_t\right] - \mathbb{E}\left[\max_{1\leq t\leq T'} X_t \mathbb{I}\{\bar{\xi}\}\right]$$
$$= \mathbb{E}\left[\max_{1\leq t\leq T'} X_t\right] - \int_0^{x_\delta} \mathbb{P}\left(\max_{1\leq t\leq T'} X_t\mathbb{I}\{\bar{\xi}\} > x\right) \mathrm{d}x$$
$$- \int_{x_\delta}^{\infty} \mathbb{P}\left(\max_{1\leq t\leq T'} X_t\mathbb{I}\{\bar{\xi}\} > x\right) \mathrm{d}x$$
$$\geq \mathbb{E}\left[\max_{1\leq t\leq T'} X_t\right] - \delta x_\delta - \int_{x_\delta}^{\infty} \mathbb{P}\left(\max_{1\leq t\leq T'} X_t\mathbb{I}\{\bar{\xi}\} > x\right) \mathrm{d}x\,,$$

where the inequality comes from $\mathbb{P}\left(\max_{1\leq t\leq T'} X_t\mathbb{I}\{\bar{\xi}\} > x\right) \leq \mathbb{P}\left(\bar{\xi}\right) \leq \delta$. Since $T' \geq \log(2)\max\left(C(2C')^{1/\beta}, 8\log(2)\right)$ and $\delta < 1/2$, we have from Lemma 3 in [CV14]

$$\left|\mathbb{P}\left(\max_{1\leq t\leq T'} X_t \leq (T'C/\log(1/(1-\delta)))^{1/\alpha}\right) - (1-\delta)\right|$$
$$\leq (1-\delta)\left(\frac{4}{T'}\left(\log\frac{1}{1-\delta}\right)^2 + \frac{2C'}{C^{1+\beta}}\left(\log\frac{1}{1-\delta}\right)^{1+\beta}\right)$$
$$\leq \frac{4}{T'}(2\delta)^2 + \frac{2C'}{C^{1+\beta}}(2\delta)^{1+\beta} \leq c\delta\max\left(\frac{\delta}{T'}, \frac{\delta^\beta}{T'^\beta}\right) \leq c\delta\max\left(\frac{1}{T'}, \frac{1}{T'^\beta}\right),$$

where c is a constant that depends only on C, C' and β. As we have $c\max(T'^{-1}, T'^{-\beta}) \leq 1/4$, this implies

$$x_- = (T'C/\log(1/(1-2\delta)))^{1/\alpha} \leq x_\delta \leq (T'C/\log(1/(1-\delta/2)))^{1/\alpha} = x_+ \ .$$

It follows

$$\mathbb{E}\left[\max_{1\leq t\leq T'} X_t \mathbb{I}\{\xi\}\right] \geq \mathbb{E}\left[\max_{1\leq t\leq T'} X_t\right] - \delta x_+ - \int_{x_-}^{\infty} \mathbb{P}\left(\max_{1\leq t\leq T'} X_t > x\right) \mathrm{d}x \ .$$

From Theorem 1 we deduce

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq t\leq T'} X_t \mathbb{I}\{\xi\}\right] &\geq \mathbb{E}\left[\max_{1\leq t\leq T'} X_t\right] - \delta x_+ - \int_{x_-}^{\infty} (1 - e^{-T'Cx^{-\alpha}}) \,\mathrm{d}x \\ &- \left(\frac{4D_2 C^{1/\alpha}}{T'^{1-1/\alpha}} + \frac{2C'D_{\beta+1}}{C^{\beta+1-1/\alpha}T'^{\beta-1/\alpha}} + 2(2C'T')^{1/(\alpha(1+\beta))} e^{-HT'^{\beta/(\beta+1)}}\right).\end{aligned}$$

From the proof of Lemma 2 in [CV14] we have for δ small enough

$$\int_{x_-}^{\infty} (1 - e^{-T'Cx^{-\alpha}}) \,\mathrm{d}x \leq \frac{8}{\alpha - 1} (T'C)^{1/\alpha} \delta^{1-1/\alpha}$$

and

$$\delta x_+ \leq 4(T'C)^{1/\alpha} \delta^{1-1/\alpha} \ .$$

Theorem 1 concludes the proof.

Proof of Lemma 4

Let F and F_r be respectively the cumulative distribution functions of X and X^r. For $x \geq 0$,

$$F_r(x) = \mathbb{P}(X^r \leq x) = \mathbb{P}(X \leq x^{1/r}) = F(x^{1/r}) \ .$$

As X follows an (α, β, C, C')-second order Pareto distribution we have

$$|1 - Cx^{-\alpha/r} - F_r(x)| = |1 - Cx^{-\alpha/r} - F(x^{1/r})| \leq C' x^{-(\alpha/r)(1+\beta)} \,,$$

which concludes the proof.

Proof of Lemma 5

We first state the following result (Lemma 8, proved in section 8), yielding high probability lower and upper bounds for the maximum of i.i.d. second order Pareto r.v.

Lemma 8. *For $X_1, ..., X_{T'}$ i.i.d samples drawn from an (α, β, C, C')-second-order Pareto distribution we define high probability lower and upper bound*

$$\ell(T', \delta) = \left(\frac{T'C}{2\log\frac{1}{\delta}}\right)^{1/\alpha} \quad \textit{and} \quad L(T', \delta) = \left(\frac{4T'C}{\log\frac{1}{1-\delta}}\right)^{1/\alpha} \,,$$

where $\delta \in (0,1)$ *can depend on* T' *and is such that* $\lim_{T'\to\infty} \ell(T',\delta) = \infty$ *and* $\lim_{T'\to\infty} L(T',\delta) = \infty$. *For* T' *large enough such that* $C\ell(T',\delta)^{-\alpha} \geq 2C'\ell(T',\delta)^{-\alpha(1+\beta)}$, $CL(T',\delta)^{-\alpha} \geq C'L(T',\delta)^{-\alpha(1+\beta)}$ *and* $L(T',\delta)^{-\alpha} \leq \frac{1}{4C}$ *we have*

$$\mathbb{P}\left(\max_{1\leq t\leq T'} X_t \leq \ell(T',\delta)\right) \leq \delta \quad \textit{and} \quad \mathbb{P}\left(\max_{1\leq t\leq T'} X_t \geq L(T',\delta)\right) \leq \delta. \tag{VI.23}$$

With the notations of Lemma 8, we respectively denote by ℓ_a and L_a the high probability lower and upper bounds for any arm a. Using Eq. (VI.23) we have by a union bound that with probability higher than $1 - K\delta_0$

$$\max_{1\leq t\leq T-(K-1)N} \widetilde{X}_{a^*,t} \geq \ell_{a^*}(T-(K-1)N,\ \delta_0)\,,$$

and for any suboptimal arm $a \neq a^*$

$$\max_{1\leq t\leq N} \widetilde{X}_{a,t} \leq L_a(N,\ \delta_0)\ .$$

Under this event, using the definition of the confidence level δ_0 we observe for T larger than some constant that for any suboptimal arm $a \neq a^*$, $L_a(N,\delta_0) \leq \ell_{a^*}(T-(K-1)N,\delta_0)$, which concludes the proof.

Proof of Lemma 8

For the high probability lower bound we write:

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq t\leq T'} X_t \leq \ell(T',\delta)\right) &= \mathbb{P}(X_1 \leq \ell(T',\delta))^{T'} \\
&\leq \left(1 - C\ell(T',\delta)^{-\alpha} + C'\ell(T',\delta)^{-\alpha(1+\beta)}\right)^{T'} \\
&\leq \left(1 - \frac{1}{2}T'C\ell(T',\delta)^{-\alpha}\right)^{T'} \leq e^{-\frac{1}{2}T'C\ell(T',\delta)^{-\alpha}} = \delta.
\end{aligned}$$

And for the high probability upper bound:

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq t\leq T'} X_t \leq L(T',\delta)\right) &= \mathbb{P}(X_1 \leq L(T',\delta))^{T'} \\
&\geq \left(1 - CL(T',\delta)^{-\alpha} - C'L(T',\delta)^{-\alpha(1+\beta)}\right)^{T'} \\
&\geq (1 - 2CL(T',\delta)^{-\alpha})^{T'} \geq e^{-4T'CL(T',\delta)^{-\alpha}} = 1-\delta\ .
\end{aligned}$$

Proof of Lemma 6

From Theorem 1, we have for any arm $a \in \{1,\ \ldots,\ K\}$,

$$\begin{aligned}
\mathbb{E}\left[X_{a,1}\mathbb{I}\{X_{a,1} > u\}\right] &\leq \int_0^\infty \mathbb{P}(X_{a,1}\mathbb{I}\{X_{a,1} > u\} \geq x)\,\mathrm{d}x \\
&= u(1-F_a(u)) + \int_u^\infty (1-F_a(x))\,\mathrm{d}x \leq M_a + \Delta_a\,,
\end{aligned}$$

where $M_a = (C_a\alpha_a/(\alpha_a-1))u^{-\alpha_a+1}$ and $\Delta_a = (C'\alpha_a(1+\beta_a)/(\alpha_a(1+\beta_a)-1))u^{-\alpha_a(1+\beta_a)+1}$. Similarly, we have $\mathbb{E}[X_{a,1}\mathbb{I}\{X_{a,1} > u\}] \geq M_a - \Delta_a$.
For u large enough, we want to prove that $M_{a^*} - \Delta_{a^*} > M_a + \Delta_a$ for any arm $a \neq a^*$, which would prove that $\arg\max_{1\leq a\leq K} \mathbb{E}[Y_{a,1}] = a^*$. First, we observe for $u > \max(1, (2C'/\min_{1\leq a\leq K} C_a)^{1/\min_{1\leq a\leq K}\beta_a})$ that $\Delta_a < \frac{1}{2}M_a$. Then, for

$$u > (3\max_{1\leq a\leq K} C_a/\min_{1\leq a\leq K} C_a)^{1/(\alpha_{(2)}-\alpha_{(1)})},$$

we have that $\frac{1}{2}M_{a^*} > \frac{3}{2}M_a$ for any arm $a \neq a^*$, which concludes the proof.

CHAPTER VII

Atomic Distributional Reinforcement Learning

The content of this final chapter was mainly designed during an internship at Google DeepMind (London) with the support of Mark Rowland, Will Dabney, Bilal Piot and Rémi Munos.

Abstract

In reinforcement learning (RL), an agent is interacting with its environment by taking actions and receiving rewards in order to find an optimal policy, i.e. a strategy maximizing the expected total return in each state. This chapter is motivated by the more challenging problem of distributional reinforcement learning (DRL) where one is interested about the whole distribution of the return, not only its expected value. We build on recent work where the distributional returns are modelled by atomic distributions and the approximation errors are measured with p-Wasserstein metrics. We first introduce two new distributional operators, for *policy evaluation* and *control* respectively, that are both contraction mappings. Then, we show that the projected atomic operators obtained by minimizing the 2-Wasserstein distance lead to a natural extension of non-distributional RL. In particular, we derive the *atomic Bellman equations* describing the dynamics of the optimal atoms. Numerical experiments in a simple two states MDP setting are provided as illustrations.

1 Introduction

In reinforcement learning (RL), an agent seeks to maximize the expected sum of discounted future rewards by sequentially interacting with his environment. This total return defines policy-dependent value functions of the environment's state and the agent's action. The objective is then to find an optimal policy maximizing these value functions in each state. In contrast, distributional reinforcement learning (DRL) consists in considering the whole distribution of the sum of rewards and not only its expected value. Such distributional approaches have shown to be very effective in practice while ensuring strong theoretical guarantees, see e.g. single-actor algorithms [BDM17, DRBM18,

DOSM18], distributed training algorithms [GDA$^+$17, BMHB$^+$18], and theoretical analysis [RDK$^+$19, RBD$^+$18, QMX18]. In particular, and as shall be recalled in this chapter, the usual RL tools such as Bellman's equations and operators, originally designed for expected values, turn out to generalize well to their distributional versions. Nevertheless, learning a whole distribution is more challenging than learning only its mean value, and different DRL approaches have been developed based on different distributional approximation schemes. For instance, [DRBM18] and [RBD$^+$18] both approximate distributional returns with atomic distributions but consider different metrics for evaluating approximation errors: respectively the 1-Wasserstein distance W_1 and the Cramér distance. The approach developed in this chapter relies on the following *two design biases*.

(a) Probability distributions D on $\mathbb{R}$ are approximated by atomic distributions $D_{\omega,\theta} = \sum_{i=1}^{N} \omega_i \delta_{\theta_i}$.

(b) Approximation errors are measured with the 2-Wasserstein distance W_2: the smaller $W_2(D, D_{\omega,\theta})$, the better $D_{\omega,\theta}$ approximates D.

Our contribution is threefold.

(i) First, we introduce two new distributional operators, one for policy evaluation and the other for the control task, that are both contraction mappings.

(ii) Then, we describe the projected operators resulting from choices (a) and (b) and prove that they are also contractions. In addition, these atomic operators provide the *atomic Bellman equations* generalizing the usual non-distributional Bellman equations.

(iii) Finally, we propose new DRL algorithms as multiatomic extensions of the TD learning and Q-learning methods.

The chapter is organized as follows. In section 2, a few concepts and results pertaining to distributional reinforcement learning are briefly recalled as well as the Wasserstein metrics that we consider to quantify how well a distribution is approximated by another one. The atomic distributions that will serve as proxies for distributional returns are defined in section 4; among them, we identify the optimal atomic approximations for different Wasserstein distances. We define and analyse the resulting projected Bellman operators in section 5. The section 6 focuses on the 2-Wasserstein case, mainly by introducing the atomic Bellman equation. Numerical experiments are presented in section 8. Finally, some concluding remarks are collected in section 9.

2 Preliminaries

Let us first recall some notations from the introductory chapter. The set of probability distributions on a set $\mathcal{E}$ (either countable or $\mathbb{R}$ throughout the chapter) is denoted by $\mathcal{P}(\mathcal{E})$ and the Lebesgue measure on $\mathbb{R}$ by λ. For any probability distribution D, $Y \sim D$

means that Y is a random variable sampled from D. For a random variable Y valued in a countable set $\mathcal{Y}$ and a mapping $\nu : \mathcal{Y} \to \mathcal{P}(\mathcal{E})$, we denote by $\nu(Y) \in \mathcal{P}(\mathcal{E})$ the *mixture distribution* of the following random variable:

$$\sum_{y \in \mathcal{Y}} \mathbb{I}\{Y = y\} U_y,$$

where $U_y \sim \nu(y)$ and Y are independent for any $y \in \mathcal{Y}$.

2.1 Markov Decision Process

We consider a Markov decision process (MDP) described by the tuple $(\mathcal{X}, \mathcal{A}, P, R)$ with countable state space $\mathcal{X}$, countable action space $\mathcal{A}$, transition kernel $P : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$ and distributional reward function $R : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$. In particular, if an agent is in state $x \in \mathcal{X}$ and takes an action $a \in \mathcal{A}$, then he receives a reward $R_0 \sim R(x, a)$ and the next state X_1 is sampled from the distribution $P(\cdot|x, a) \in \mathcal{P}(\mathcal{X})$. A policy $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ maps any state $x \in \mathcal{X}$ to a distribution over the actions $\pi(\cdot|x) \in \mathcal{P}(\mathcal{A})$. Given a discount factor $\gamma \in [0, 1)$, we define the distributional return $Z^\pi(x, a)$ of a policy π after taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$ as the probability distribution of the random variable

$$\sum_{t=0}^{\infty} \gamma^t R_t, \tag{VII.1}$$

given $X_0 = x$, $A_0 = a$ and for all $t \in \mathbb{N}$, $R_t \sim R(X_t, A_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$ and $A_{t+1} \sim \pi(\cdot|X_{t+1})$. We recall that classical (non-distributional) RL mainly focuses on expected returns, through the state-action value function $Q^\pi(x, a) = \mathbb{E}_{Z_0 \sim Z^\pi(x,a)}[Z_0]$ and the value function $V^\pi(x) = \mathbb{E}_{A_0 \sim \pi(\cdot|x)}[Q^\pi(x, A_0)]$ verifying Bellman's equation ([Bel66]):

$$\forall (x, a), \quad Q^\pi(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q^\pi(X_1, A_1)], \tag{VII.2}$$

where $R_0 \sim R(x, a)$, $X_1 \sim P(\cdot|x, a)$ and $A_1 \sim \pi(\cdot|X_1)$. The optimal policies can be characterized by means of the optimal state-action value function $Q^*(x, a)$, which verify Bellman's optimality equation:

$$\forall (x, a), \quad Q^*(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q^*(X_1, a')]. \tag{VII.3}$$

Then, denoting by $V^*(x) = \max_a Q^*(x, a)$ the optimal value function, a policy π^* is optimal if for all x,

$$\mathbb{E}[Q^*(x, A_0)] = V^*(x), \quad \text{with } A_0 \sim \pi^*(\cdot|x).$$

Bellman Operators. In the policy evaluation task, one wants to compute Q^π for a given policy π, while in the control task, the goal it to approach Q^*. The usual dynamic programming way for solving these two tasks is based on two operators. First, the Bellman operator T^π ([Bel66]) defined by: for all $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$T^\pi Q(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q(X_1, A_1)], \quad \text{with } X_1 \sim P(\cdot|x, a), A_1 \sim \pi(\cdot|X_1).$$

Second, the Bellman optimality operator T defined by:

$$TQ(x,a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q(X_1, a')], \quad \text{with } X_1 \sim P(\cdot|x,a).$$

In particular, the Bellman operator T^π (resp. Bellman optimality operator T) is known to be a γ-contraction[1] for the sup norm and its repeated application to an initial Q-function to converge exponentially fast to its unique fixed point Q^π (resp. Q^*) ([BT96]).

2.2 Wasserstein Distance

In the case one is interested in the whole distribution of the discounted sum of future rewards in Eq. (VII.1), more general distributional operators introduced in [BDM17] are required. Before defining these distributional operators, we first recall the Wasserstein metrics on which our analysis will rely.

Definition 1. *Let D_1 and D_2 be two distributions on $\mathbb{R}$ with finite moments and respective cumulative distribution functions (c.d.f.'s) F_1 and F_2.*

(i) For $p \in [1, +\infty)$, the p-Wasserstein distance between D_1 and D_2 is

$$W_p(D_1, D_2) = \left(\int_{\tau=0}^{1} \left| F_1^{-1}(\tau) - F_2^{-1}(\tau) \right|^p d\tau \right)^{\frac{1}{p}},$$

where $F^{-1} : \tau \mapsto \inf\{z \in \mathbb{R}, F(z) \geq \tau\}$ is the generalized inverse distribution function of any c.d.f. F.

(ii) For $p = +\infty$, the ∞-Wasserstein distance is the essential supremum of $|F_1^{-1} - F_2^{-1}|$ over the interval $(0, 1]$:

$$W_\infty(D_1, D_2) = \operatorname*{ess\ sup}_{\tau \in (0,1]} |F_1^{-1}(\tau) - F_2^{-1}(\tau)|.$$

We denote by $\mathcal{Z}$ the set of state-action distribution functions with finite moments:

$$\mathcal{Z} = \left\{ Z : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathbb{R}) \text{ s.t. } \forall p \geq 1, \forall (x,a) \in \mathcal{X} \times \mathcal{A}, \mathbb{E}_{Z_0 \sim Z(x,a)}[|Z_0|^p] < \infty \right\}.$$

Then, we recall the maximal form of the Wasserstein distance introduced in [BDM17]: for all $(Z, Z') \in \mathcal{Z}^2$,

$$\widetilde{W}_p(Z, Z') = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} W_p(Z(x,a), Z'(x,a)).$$

2.3 Distributional Bellman Operators

Here, we recall the two *distributional Bellman operators* introduced in [BDM17], the first for policy evaluation and the second for control.

The Distributional Bellman Operator (DBO). The distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ ([BDM17]) is defined as follows[2]: for all $Z \in \mathcal{Z}$ and $(x,a) \in \mathcal{X} \times \mathcal{A}$, $\mathcal{T}^\pi Z(x,a)$

[1] A function mapping a metric space to itself is called a γ-contraction (resp. a non-expansion) if it is Lipschitz continuous with Lipschitz constant $\gamma < 1$ (resp. $\kappa \leq 1$).

[2] We refer to [RBD+18] for a more rigorous measure theoretic definition of $\mathcal{T}^\pi$.

is the distribution of the random variable

$$R_0 + \gamma Z_1,$$

where $R_0 \sim R(x,a)$ and $Z_1 \sim Z(X_1, A_1)$ with $X_1 \sim P(\cdot|x,a)$ and $A_1 \sim \pi(\cdot|X_1)$. We know from Lemma 3 in [BDM17] that for any Wasserstein order $p \in [1, +\infty]$, the distributional Bellman operator $\mathcal{T}^\pi$ is a γ-contraction in the metric $\widetilde{W}_p$. Moreover, the unique fixed point of $\mathcal{T}^\pi$ is Z^π, which leads to the distributional version of Bellman's equation:

$$Z^\pi = \mathcal{T}^\pi Z^\pi.$$

The Distributional Bellman Optimality Operator (DBOO). In [BDM17], a distributional Bellman optimality operator $\mathcal{T} : \mathcal{Z} \to \mathcal{Z}$ is defined as any operator following greedy policies. Formally, for any $Z \in \mathcal{Z}$, there exists a greedy policy π_Z for Z, i.e. such that for all $x \in \mathcal{X}$:

$$\sum_{a \in \mathcal{A}} \pi_Z(a|x) \mathbb{E}_{Z_0 \sim Z(x,a)}[Z_0] = \max_{a \in \mathcal{A}} \mathbb{E}_{Z_0 \sim Z(x,a)}[Z_0],$$

verifying $\mathcal{T}Z = \mathcal{T}^{\pi_Z} Z$. Nevertheless, these optimality operators are not as well-behaved as DBO's: indeed, the DBOO's are not contractive mappings as shown by Proposition 1 in [BDM17].

3 1-Step Distributional Bellman Operators

We introduce new distributional Bellman operators, namely the '1-SDBO' (for evaluation) and the '1-SDBOO' (for control), taking only into account the randomness induced by the first step/transition. In a certain sense, they can be seen as less ambitious variants of the DBO and the DBOO. Most noteworthy is the control setting, where we will show that the 1-SDBOO is a contraction mapping, contrary to the DBOO $\mathcal{T}$.

The 1-Step Distributional Bellman Operator (1-SDBO). Given a policy π, we define the 1-SDBO $\mathbb{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ by: for all $Z \in \mathcal{Z}$, for all (x,a), $\mathbb{T}^\pi Z(x,a)$ is the probability distribution of

$$R_0 + \gamma \mathbb{E}[Z_1|X_1, A_1],$$

where $R_0 \sim R(x,a)$, $X_1 \sim P(\cdot|x,a)$, $A_1 \sim \pi(\cdot|X_1)$ and $Z_1 \sim Z(X_1, A_1)$.

Lemma 1. *For all $p \in [1, +\infty]$, the 1-SDBO $\mathbb{T}^\pi$ is a γ-contraction in $\widetilde{W}_p$.*

The proof is deferred to section 11. From Lemma 1 we deduce that the 1-SDBO $\mathbb{T}^\pi$ has a unique fixed point, namely a state-action distribution function Z whose expected value is equal to $Q^\pi(x,a)$ in each (x,a). Indeed, taking the expectation on both sides of the fixed point equation $Z = \mathbb{T}^\pi Z$ shows that the mean of $Z(x,a)$ solves the non-distributional Bellman equation (VII.2). Moreover, one easily proves that if the rewards

are deterministic i.e. $R(x,a) = \delta_{r(x,a)}$ with $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the fixed-point of $\mathbb{T}^\pi$ is simply:

$$(x,a) \mapsto \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} P(x'|x,a)\pi(a'|x')\delta_{r(x,a)+\gamma Q^\pi(x',a')}.$$

The 1-Step Distributional Bellman Optimality Operator (1-SDBOO). We define the 1-SDBOO $\mathbb{T} : \mathcal{Z} \to \mathcal{Z}$ by: for all $Z \in \mathcal{Z}$, for all (x,a), $\mathbb{T}Z(x,a)$ is the probability distribution of

$$R_0 + \gamma \max_{a'} \mathbb{E}[Z_{1,a'}|X_1],$$

where $R_0 \sim R(x,a)$, $X_1 \sim P(\cdot|x,a)$ and $Z_{1,a'} \sim Z(X_1,a')$ for any action a'. As the 1-SDBO, the 1-SDBOO is a contraction as stated in the following result proved in section 11.

Lemma 2. *For all $p \in [1,+\infty]$, the 1-SDBOO $\mathbb{T}$ is a γ-contraction in $\widetilde{W}_p$.*

The 1-SDBOO being a contraction, it thus has a unique fixed point. Here again, by observing that the expectation of the fixed point equation $Z = \mathbb{T}Z$ reduces to the non-distributional Bellman optimality equation (VII.3), we conclude that the expected value of this fixed point is $Q^*(x,a)$ in each (x,a). In addition, for deterministic rewards $R = \delta_r$, the fixed-point of $\mathbb{T}$ is:

$$(x,a) \mapsto \sum_{x' \in \mathcal{X}} P(x'|x,a)\delta_{r(x,a)+\gamma V^*(x')}.$$

In the next section, we introduce our approximation procedure for approaching general distributions on $\mathbb{R}$ by simpler 'atomic' distributions, described by finite numbers of particles.

4 Atomic Approximation

In practice, computing the image of some Z by the DBO/1-SDBO/1-SDBOO is hardly possible as it lives in a large subspace of $\mathcal{Z}$. On the other hand, parameterized distributions are simpler to deal with in practical implementations. Here, we focus on the parametric class of atomic distributions and our approach will consist in (optimally) projecting $\mathcal{T}^\pi Z$, $\mathbb{T}^\pi Z$ or $\mathbb{T}Z$ into this subspace.

4.1 Atomic Distributions

In [BDM17], distributional returns are approximated by categorical distributions, where the probability weights over the atoms are learned but their locations are predefined. In [DRBM18] and more recently in [RDK+19], the state-action distributions are approximated by uniform averages of Dirac distributions, whose locations are learned. In this chapter, we promote a combination of these two approaches: we first characterize the optimal atoms' locations for given probability masses, before analysing optimal mass allocation. Formally, for a given number of atoms $N \geq 1$ we define the following parametric classes of probability measures on $\mathbb{R}$.

Definition 2. *Let* $\Delta_N = \{\omega = (\omega_1, \dots, \omega_N) \in [0,1]^N \text{ s.t. } \omega_1 + \cdots + \omega_N = 1\}$ *be the probability simplex and* $\mathbb{S}_N$ *the subset of* $\mathbb{R}^N$ *of sorted vectors:*

$$\mathbb{S}_N = \{\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^N \text{ s.t. } \theta_1 \leq \cdots \leq \theta_N\}.$$

(i) The set of atomic distributions is

$$\mathcal{D}_N = \Big\{ D_{\omega,\theta} = \sum_{i=1}^{N} \omega_i \delta_{\theta_i} \text{ s.t. } \theta = (\theta_1, \dots, \theta_N) \in \mathbb{S}_N, \omega = (\omega_1, \dots, \omega_N) \in \Delta_N \Big\},$$

where ω_i *is the probability mass allocated to the* i*-th smallest atom* θ_i*.*

(ii) The set of state-action atomic distribution functions is

$$\mathcal{Z}_N = \mathcal{D}_N^{\mathcal{X}\times\mathcal{A}} = \Big\{ Z_{\Omega,\Theta} : \mathcal{X}\times\mathcal{A} \to \mathcal{D}_N \text{ s.t. } \Theta : \mathcal{X}\times\mathcal{A} \to \mathbb{S}_N, \Omega : \mathcal{X}\times\mathcal{A} \to \Delta_N \Big\},$$

where $Z_{\Omega,\Theta}(x,a) = D_{\Omega(x,a),\Theta(x,a)}$ *for all* (x,a)*.*

In other words, an element $Z_{\Omega,\Theta}$ of $\mathcal{Z}_N$ associates to each state-action pair (x,a) an atomic distribution

$$Z_{\Omega,\Theta}(x,a) = \sum_{i=1}^{N} \Omega_i(x,a) \delta_{\Theta_i(x,a)},$$

where $\Theta(x,a) = (\Theta_1(x,a), \dots, \Theta_N(x,a)) \in \mathbb{S}_N$ and $\Omega(x,a) = (\Omega_1(x,a), \dots, \Omega_N(x,a)) \in \Delta_N$. Moreover, we define for any ω,Ω,

$$\mathcal{D}_\omega = \{D_{\omega',\theta} \in \mathcal{D}_N \text{ s.t. } \omega' = \omega\} \text{ and } \mathcal{Z}_\Omega = \{Z_{\Omega',\Theta} \in \mathcal{Z}_N \text{ s.t. } \Omega' = \Omega\}.$$

We measure the approximation error between a distribution D on $\mathbb{R}$ and some proxy $D_{\omega,\theta} \in \mathcal{D}_N$ (with respective c.d.f.'s F and $F_{\omega,\theta}$) in terms of their Wasserstein distance, which in this case rewrites:

$$W_p(D, D_{\omega,\theta}) = \left(\sum_{i=1}^{N} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} \left| F^{-1}(\tau) - \theta_i \right|^p d\tau \right)^{\frac{1}{p}},$$

where $\overline{\omega} = (\overline{\omega}_1, \dots, \overline{\omega}_N)$ is the vector of the cumulative sums of ω's components:

$$\overline{\omega}_i = \sum_{j \leq i} \omega_j, \ \forall i \in \{1, \dots, N\}.$$

This expression of $W_p(D, D_{\omega,\theta})$ results from the fact that the generalized inverse distribution function of $D_{\omega,\theta}$ is piecewise constant: for all $i \in \{1, \dots, N\}$,

$$F^{-1}_{\omega,\theta}(\tau) = \theta_i, \ \forall \tau \in (\overline{\omega}_{i-1}, \overline{\omega}_i].$$

In the same way, we define for any Ω the cumulative probability function $\overline{\Omega} : (x, a) \mapsto (\overline{\Omega}_1(x, a), \dots, \overline{\Omega}_N(x, a))$ such that

$$\overline{\Omega}_i(x, a) = \sum_{j \leq i} \Omega_j(x, a).$$

For fixed probabilities, we characterize in the next subsection the best proxy among atomic distributions i.e. the optimal spatial distribution of atoms minimizing the Wasserstein error.

4.2 2-Wasserstein Error Minimizers

Our strategy consists in minimizing the 2-Wasserstein approximation error to characterize optimal atomic distributions. Although this choice may seem arbitrary, we show next that the usual non-distributional Bellman equations and operators pertain to this setting in the particular monoatomic case $N = 1$ (see Remark 1). We have,

$$W_2(D, D_{\omega,\theta}) = \left(\sum_{i=1}^{N} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} \left(F^{-1}(\tau) - \theta_i \right)^2 d\tau \right)^{\frac{1}{2}},$$

which is, for fixed D and ω, minimal if and only if for all $i \in \{1, \dots, N\}$ such that $\omega_i \neq 0$, θ_i is equal to the following trimmed mean:

$$\theta^*_{\omega,i} = \frac{1}{\omega_i} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} F^{-1}(\tau) d\tau. \tag{VII.4}$$

Moreover, notice that if for instance F is continuous, then this trimmed mean also interprets as the conditional expectation $\mathbb{E}_{Y \sim D}[Y | F^{-1}(\overline{\omega}_{i-1}) \leq Y \leq F^{-1}(\overline{\omega}_i)]$. Similarly, we approximate state-action distribution functions $Z \in \mathcal{Z}$ by atomic distribution functions $Z_{\Omega,\Theta} \in \mathcal{Z}_N$. For a given mass allocation function Ω, the unique minimizer $Z_{\Omega,\Theta^*_\Omega}$ in $\mathcal{Z}_\Omega$ of the 2-Wasserstein approximation error $W_2(Z(x, a), Z_{\Omega,\Theta^*_\Omega}(x, a))$ for all (x, a) is given by: for all $i \in \{1, \dots, N\}$ such that $\Omega_i(x, a) \neq 0$,

$$\Theta^*_{\Omega,i}(x, a) = \frac{1}{\Omega_i(x, a)} \int_{\tau=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} F^{-1}_{x,a}(\tau) d\tau, \tag{VII.5}$$

where $F_{x,a}$ denotes the c.d.f. of $Z(x, a)$.

In the monoatomic case $N = 1$ where the whole distribution $Z(x, a)$ is summarized by a single scalar, the minimum W_2-error is attained at the (global) mean. More generally for any number of atoms $N \geq 1$, this mean simply expresses as the average of the N trimmed means:

$$\sum_{i=1}^{N} \Omega_i(x, a) \Theta^*_{\Omega,i}(x, a) = \mathbb{E}_{Z_0 \sim Z(x,a)}[Z_0]. \tag{VII.6}$$

This property turns out to be useful in Q-learning methods where Q-functions, which are expected values, must be computed even in distributional Q-learning variants such as C51 ([BDM17]) and QR-DQN ([DRBM18]).

5 Atomic Bellman Operators

Here, we introduce the *atomic Bellman operators*: they are compositions of some distributional Bellman operator followed by a W_2-projection on the space of atomic distributions. These atomic Bellman operators allow to design DRL methods restricted to atomic distributions.

5.1 Atomic Projections

Motivated by the atomic approximation scheme discussed in the previous section, we define for any probability weights $\omega \in \Delta_N$, the $(2,\omega)$-atomic projection $\Pi_{2,\omega}$ as follows: for any distribution $D \in \mathcal{P}(\mathbb{R})$ with finite moments and c.d.f. F,

$$\Pi_{2,\omega} D = D_{\omega,\theta^*_\omega} \in \mathcal{D}_\omega \text{ with } \theta^*_\omega = \left(\theta^*_{\omega,1}, \ldots, \theta^*_{\omega,N}\right),$$

where $\theta^*_{\omega,i}$ is given by Eq. (VII.4). Similarly for any mass allocation function $\Omega : \mathcal{X}\times\mathcal{A} \to \Delta_N$, we define the Ω-atomic projection $\Pi_{2,\Omega}$ by: for all distribution functions $Z \in \mathcal{Z}$,

$$\Pi_{2,\Omega} Z : (x,a) \mapsto \Pi_{2,\Omega(x,a)} Z(x,a) \text{ for } (x,a) \in \mathcal{X}\times\mathcal{A}.$$

By denoting $\Theta^*_\Omega(x,a) = \theta^*_{\Omega(x,a)}$, we have: $\Pi_{2,\Omega} Z = Z_{\Omega,\Theta^*_\Omega} \in \mathcal{Z}_\Omega$.

Lemma 3. *Let $N \geq 1$ and $\Omega : \mathcal{X}\times\mathcal{A} \to \Delta_N$. The atomic projection $\Pi_{2,\Omega}$ is a non-expansion in $\widetilde{W}_\infty$.*

By combining this atomic projection with DRL operators, we define below the atomic Bellman operators.

5.2 The Atomic Bellman Operators

Given a policy π, we define atomic Bellman operators as compositions of the DBO/1-SDBO/1-SDBOO followed by the $(2,\Omega)$-atomic projection, i.e. respectively:

$$\Pi_{2,\Omega}\mathcal{T}^\pi,\ \Pi_{2,\Omega}\mathbb{T}^\pi \text{ and } \Pi_{2,\Omega}\mathbb{T}.$$

By analogy with the usual names and notations for non-distributional operators, we will refer to $\mathcal{T}^\pi_\Omega = \Pi_{2,\Omega}\mathcal{T}^\pi$ as the *atomic Bellman operator* (ABO), to $\mathbb{T}^\pi_\Omega = \Pi_{2,\Omega}\mathbb{T}^\pi$ as the *1-step atomic Bellman operator* (1-SABO) and to $\mathbb{T}_\Omega = \Pi_{2,\Omega}\mathbb{T}$ as the *1-step atomic Bellman optimality operator* (1-SABOO).

Remark 1. *In the monoatomic case $N = 1$, where necessarily $\Omega(x,a) = \Omega_1(x,a) \equiv 1$, we point out that $\mathcal{T}^\pi_\Omega = \mathbb{T}^\pi_\Omega$ is simply the usual non-distributional Bellman operator T^π and $\mathbb{T}_\Omega$ ($= \Pi_{2,\Omega}\mathcal{T}$) the non-distributional Bellman optimality operator T.*

Then, by combining Lemma 3 with respectively Lemma 3 in [BDM17], Lemma 1 and Lemma 2, we obtain that the ABO, 1-SABO and 1-SABOO are all three contractive maps.

Corollary 1. *For any number of atoms $N \geq 1$, policy π and state-action mass allocation function $\Omega : \mathcal{X} \times \mathcal{A} \to \Delta_N$, the ABO $\mathcal{T}^\pi_\Omega$, the 1-SABO $\mathbb{T}^\pi_\Omega$ and 1-SABOO $\mathbb{T}_\Omega$ are all γ-contractions for the metric $\widetilde{W}_\infty$.*

The next subsection evaluates the loss of accuracy induced by our atomic projections: we focus on the distortion of the distributional fixed points.

5.3 $\mathbf{W}_\infty$-Approximation Error of the Atomic Model

For the ∞-Wasserstein distance, we provide an upper bound on the approximation error resulting from approaching the distributional returns by atomic distributions. For all distributional operators $T \in \{\mathcal{T}^\pi, \mathbb{T}^\pi, \mathbb{T}\}$, denoting by Z (resp. Z_Ω) the fixed point of T (resp. $T_\Omega = \Pi_{2,\Omega} T$), we prove that $\widetilde{W}_\infty(Z_\Omega, Z)$ is essentially of order $O(1/N)$.

Proposition 1. *Let π be a policy. For any operator $T \in \{\mathcal{T}^\pi, \mathbb{T}^\pi, \mathbb{T}\}$, let Z and Z_Ω be the respective fixed points of T and $T_\Omega = \Pi_{2,\Omega} T$. Then, the $\widetilde{W}_\infty$ distance between Z and Z_Ω is upper bounded as follows:*

$$\widetilde{W}_\infty(Z_\Omega, Z) \leq \frac{1}{1-\gamma} \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \epsilon_\Omega(x,a),$$

where

$$\epsilon_\Omega(x,a) = \max_{1 \leq i \leq N, \Omega_i(x,a) \neq 0} F^{-1}_{x,a}(\overline{\Omega}_i(x,a)) - F^{-1}_{x,a}(\overline{\Omega}_{i-1}(x,a)+),$$

with $F_{x,a}$ the c.d.f. of $Z(x,a)$ and $F^{-1}_{x,a}(\tau+)$ the right limit[3] of the quantile function $F^{-1}_{x,a}$ at any point $0 \leq \tau < 1$.

In particular for $\Omega_i(x,a) \equiv \frac{1}{N}$, if $Z(x,a)$ has a convex support and a density lower bounded by $M > 0$ for all (x,a), then $\epsilon_\Omega(x,a) \leq \frac{1}{M \cdot N}$ and the approximation error $\widetilde{W}_\infty(Z_\Omega, Z)$ is thus of order $O(1/N)$. This rate is empirically verified in Figure VII.1.

6 Atomic Bellman Equations

The purpose of this section is to provide explicit expressions for the ABO, 1-SABO and 1-SABOO and to deduce generalizations of Bellman's equations. By considering atomic distribution functions Z, as in DRL algorithms such as QR-DQN in [DRBM18], and by assuming that the rewards are deterministic ($R = \delta_r$), the trimmed means of $\mathcal{T}^\pi Z(x,a)$, $\mathbb{T}^\pi Z(x,a)$ and $\mathbb{T} Z(x,a)$ can all be written in closed-form. These formulas will lead us to atomic Bellman equations.

[3] We recall that c.d.f.'s are right-continuous with left limits while quantile functions are left-continuous with right limits, see e.g. [EH13].

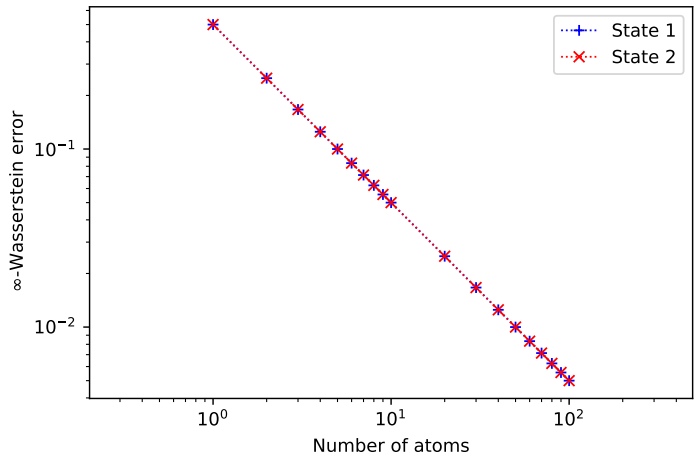


Figure VII.1: Log-log plot of the approximation error $W_\infty(Z^\pi_\Omega(x,a), Z^\pi(x,a))$ (with $\Omega_i(x,a) \equiv 1/N$ and Z^π_Ω the fixed point of $\mathcal{T}^\pi_\Omega$) in function of the number of atoms N: it follows the rate $O(1/N)$ of Proposition 1. Same two states MDP setting as in section 8.

6.1 The Atomic Bellman Equation

For deterministic rewards and $Z = Z_{\Omega,\Theta} \in \mathcal{Z}_N$ an atomic distribution function, we have an explicit expression for the projected distribution $\mathcal{T}^\pi_{\Omega'} Z_{\Omega,\Theta}(x,a)$ (with Ω' potentially different from Ω) as shown in the next result.

Lemma 4. *Let π be a policy and $Z_{\Omega,\Theta} \in \mathcal{Z}_N$. Let $\Omega' : \mathcal{X}\times\mathcal{A} \to \Delta_N$ and $\Theta' : \mathcal{X}\times\mathcal{A} \to \mathbb{S}_N$ such that $Z_{\Omega',\Theta'} = \mathcal{T}^\pi_{\Omega'} Z_{\Omega,\Theta}$. Then, assuming deterministic rewards $R = \delta_r$, we have for all $(x,a) \in \mathcal{X}\times\mathcal{A}$ and $i \in \{1,\dots,N\}$ such that $\Omega'_i(x,a) \neq 0$,*

$$\Theta'_i(x,a) = r(x,a) + \frac{\gamma}{\Omega'_i(x,a)} \sum_{\theta \in \widetilde{\Theta}} \mu^\pi_i(\Theta, x, a, \theta) \cdot \theta \,,$$

with the set $\widetilde{\Theta} = \{\Theta_j(x',a') : (x',a',j) \in \mathcal{X}\times\mathcal{A}\times\{1,\dots,N\}\}$ and the weight functions μ^π_i given by:

$$\begin{aligned}\mu^\pi_i(\Theta, x, a, \theta) &= \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[H^\pi_{x,a}(\theta),\, G^\pi_{x,a}(\theta)\big]\big), \\ &= \big(\min\{\overline{\Omega'}_i(x,a),\, G^\pi_{x,a}(\theta)\} - \max\{\overline{\Omega'}_{i-1}(x,a),\, H^\pi_{x,a}(\theta)\}\big)_+ \,,\end{aligned}$$

where λ is the Lebesgue measure on $\mathbb{R}$, $(z)_+ = \max\{0,z\}$ denotes the positive part of any $z \in \mathbb{R}$, $G^\pi_{x,a}$ is the c.d.f. of $Z_{\Omega,\Theta}(X_1,A_1)$ given $X_1 \sim P(\cdot|x,a)$ and $A_1 \sim \pi(\cdot|X_1)$:

$$G^\pi_{x,a}(\theta) = \sum_{(x',a') \in \mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x') \sum_{j=1}^N \Omega_j(x',a') \mathbb{I}\{\Theta_j(x',a') \leq \theta\},$$

and $H^\pi_{x,a} : \theta \mapsto G^\pi_{x,a}(\theta-)$ *is the left limit function of* $G^\pi_{x,a}$ *also equal to*

$$H^\pi_{x,a}(\theta) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x') \sum_{j=1}^{N} \Omega_j(x',a')\mathbb{I}\{\Theta_j(x',a') < \theta\}.$$

The proof of Lemma 4 is deferred to section 11. We emphasize that $\widetilde{\Theta}$ is a set of unique elements: if for instance $\mathcal{X}$ and $\mathcal{A}$ are both finite, then the cardinality of $\widetilde{\Theta}$ may be strictly smaller than $|\mathcal{X}|\cdot|\mathcal{A}|\cdot N$ in the case of repeated valued of the atoms $\Theta_j(x',a')$.

The fixed point Z_{Ω,Θ^π} of the ABO $\mathcal{T}^\pi_\Omega$ is the unique solution of the *atomic Bellman equation*: $\mathcal{T}^\pi_\Omega Z_{\Omega,\Theta^\pi} = Z_{\Omega,\Theta^\pi}$.

Proposition 2. (Atomic Bellman Equation) *For deterministic rewards* $R = \delta_r$*, the fixed point* Z_{Ω,Θ^π} *of the ABO* $\mathcal{T}^\pi_\Omega$ *is given by the unique solution* Θ^π *of the* atomic Bellman equation*: for all* $(x,a) \in \mathcal{X}\times\mathcal{A}$*, for all* $i \in \{1,\dots,N\}$ *such that* $\Omega_i(x,a) \neq 0$*,*

$$\Theta^\pi_i(x,a) = r(x,a) + \frac{\gamma}{\Omega_i(x,a)} \sum_{\theta\in\widetilde{\Theta^\pi}} \mu^\pi_i(\Theta^\pi, x, a, \theta)\cdot\theta\,, \tag{VII.7}$$

where $\widetilde{\Theta^\pi} = \{\Theta^\pi_j(x',a') : (x',a',j) \in \mathcal{X}\times\mathcal{A}\times\{1,\dots,N\}\}$ *and the weight functions* μ^π_i *are given by Lemma 4 with* $\Omega' = \Omega$*.*

For clarity purpose, we point out that in the particular case of uniform probability weights $\Omega_i(x,a) \equiv 1/N$, Eq. (VII.7) writes as:

$$\Theta^\pi_i(x,a) = r(x,a) + \gamma N \sum_{\theta\in\widetilde{\Theta^\pi}} \mu^\pi_i(\Theta^\pi, x, a, \theta)\cdot\theta\,,$$

with

$$\mu^\pi_i(\Theta^\pi, x, a, \theta) = \Big(\min\Big\{\frac{i}{N},\, G^\pi_{x,a}(\theta)\Big\} - \max\Big\{\frac{i-1}{N},\, H^\pi_{x,a}(\theta)\Big\}\Big)_+.$$

An immediate consequence of Proposition 2 is that the average of the N atoms $\Theta^\pi_i(x,a)$ is equal to the state-action value function $Q^\pi(x,a)$.

Property 1. *Let* Z_{Ω,Θ^π} *be the fixed point of the ABO* $\mathcal{T}^\pi_\Omega$*. Then for all* (x,a)*,*

$$\sum_{i=1}^{N} \Omega_i(x,a)\Theta^\pi_i(x,a) = Q^\pi(x,a).$$

The proof is straightforward by averaging the atomic Bellman equations (VII.7) and observing that $\sum_i \Omega_i(x,a)\Theta^\pi_i(x,a)$ is the solution of the (non-distributional) Bellman equation. Property 1 is also verified empirically: see Figure VII.2 in section 8. Notice that for any (x,a), the two distributions $Z_{\Omega,\Theta^\pi}(x,a)$ and $Z^\pi(x,a)$ have the same expectation, namely $Q^\pi(x,a)$. Nevertheless, these two distributions are not equal in general (they are if $Z^\pi \in \mathcal{Z}_\Omega$).

6.2 The 1-Step Atomic Bellman Equation

Similarly for the 1-SABO applied on some atomic distribution function, we have the following formula in the case of deterministic rewards.

Lemma 5. *Let $N \geq 1$, $\Omega' : \mathcal{X} \times \mathcal{A} \to \Delta_N$ be a mass allocation function, and assume deterministic rewards $R = \delta_r$. The image of any $Z_{\Omega,\Theta} \in \mathcal{Z}_N$ by the 1-SABO $\mathbb{T}^\pi_{\Omega'}$ is $Z_{\Omega',\Theta'} = \mathbb{T}^\pi_{\Omega'} Z_{\Omega,\Theta}$, where for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, $i \in \{1,\dots,N\}$ such that $\Omega'_i(x,a) \neq 0$,*

$$\Theta'_i(x,a) = r(x,a) + \frac{\gamma}{\Omega'_i(x,a)} \sum_{q \in \mathcal{Q}(\Theta)} \mu^\pi_i(\Theta, x, a, q) \cdot q \,,$$

with $\mathcal{Q}(\Theta)$ the set of Q-values $Q(x',a') = \sum_{j=1}^N \Omega_j(x',a')\Theta_j(x',a')$:

$$\mathcal{Q}(\Theta) = \left\{ \sum_{j=1}^N \Omega_j(x',a')\Theta_j(x',a') : (x',a') \in \mathcal{X} \times \mathcal{A} \right\},$$

and the weight functions μ^π_i given by:

$$\mu^\pi_i(\Theta, x, a, q) = \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[H^\pi_{x,a}(q),\, G^\pi_{x,a}(q)\big]\big),$$

where $G^\pi_{x,a}$ is the c.d.f. of $Q(X_1, A_1)$ given $X_1 \sim P(\cdot|x,a)$, $A_1 \sim \pi(\cdot|X_1)$:

$$G^\pi_{x,a}(q) = \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} P(x'|x,a)\pi(a'|x')\mathbb{I}\{Q(x',a') \leq q\},$$

and

$$H^\pi_{x,a}(q) = G^\pi_{x,a}(q-) = \sum_{(x',a') \in \mathcal{X} \times \mathcal{A}} P(x'|x,a)\pi(a'|x')\mathbb{I}\{Q(x',a') < q\}.$$

The fixed point Z_{Ω,Θ^π} of the 1-SABO $\mathbb{T}^\pi_\Omega$ provides the *1-step atomic Bellman equation.*

Proposition 3. (1-Step Atomic Bellman Equation) *For deterministic rewards $R = \delta_r$, the fixed point Z_{Ω,Θ^π} of the 1-SABO $\mathbb{T}^\pi_\Omega$ is given by the unique solution Θ^π of the* 1-step atomic Bellman equation*: for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, for all $i \in \{1,\dots,N\}$ such that $\Omega_i(x,a) \neq 0$,*

$$\Theta^\pi_i(x,a) = r(x,a) + \frac{\gamma}{\Omega_i(x,a)} \sum_{q \in \mathcal{Q}(\Theta^\pi)} \mu^\pi_i(\Theta^\pi, x, a, q) \cdot q \,, \tag{VII.8}$$

where the weight functions μ^π_i are given by Eq. (5) with $\Omega' = \Omega$.

As in Property 1 for the solution of the ABO, the solution Θ^π described in Proposition 3 satisfies the following averaging property.

Property 2. *Let Z_{Ω,Θ^π} be the fixed point of the 1-SABO $\mathbb{T}^\pi_\Omega$. Then for all (x,a),*

$$\sum_{i=1}^N \Omega_i(x,a)\Theta^\pi_i(x,a) = Q^\pi(x,a).$$

In the next subsection, we give similar results concerning the 1-SABOO.

6.3 The 1-Step Atomic Bellman Optimality Equation

Similarly to Lemmas 4 and 5, the following result holds when applying the 1-SABOO to an atomic distribution function.

Lemma 6. *Let $N \geq 1$, $\Omega' : \mathcal{X} \times \mathcal{A} \to \Delta_N$ be a mass allocation function, and assume deterministic rewards $R = \delta_r$. The image of any $Z_{\Omega,\Theta} \in \mathcal{Z}_N$ by the 1-SABOO $\mathbb{T}_{\Omega'}$ is $Z_{\Omega',\Theta'} = \mathcal{T}_{\Omega'} Z_{\Omega,\Theta}$, where for all (x, a), i such that $\Omega'_i(x, a) \neq 0$,*

$$\Theta'_i(x,a) = r(x,a) + \frac{\gamma}{\Omega'_i(x,a)} \sum_{q \in \mathcal{Q}_{\max}(\Theta)} \mu_i^*(\Theta, x, a, q) \cdot q \,,$$

with the set of maximal Q-values $Q(x', a') = \sum_{j=1}^N \Omega_j(x', a')\Theta_j(x', a')$

$$\mathcal{Q}_{\max}(\Theta) = \left\{ \max_{a' \in \mathcal{A}} Q(x', a') : x' \in \mathcal{X} \right\},$$

and the weight functions μ_i^* given by:

$$\mu_i^*(\Theta, x, a, q) = \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[H^*_{x,a}(q),\, G^*_{x,a}(q)\big]\big),$$

where $G^*_{x,a}$ is the c.d.f. of $\max_{a'} Q(X_1, a')$ given $X_1 \sim P(\cdot|x, a)$:

$$G^*_{x,a}(q) = \sum_{x' \in \mathcal{X}} P(x'|x,a) \mathbb{I}\Big\{ \max_{a' \in \mathcal{A}} Q(x', a') \leq q \Big\},$$

and

$$H^*_{x,a}(q) = G^*_{x,a}(q-) = \sum_{x' \in \mathcal{X}} P(x'|x,a) \mathbb{I}\Big\{ \max_{a' \in \mathcal{A}} Q(x', a') < q \Big\}.$$

Then, we derive *1-step atomic Bellman optimality equation.*

Proposition 4. (1-Step Atomic Bellman Optimality Equation) *For deterministic rewards $R = \delta_r$, the fixed point Z_{Ω,Θ^*} of the 1-SABOO $\mathbb{T}_\Omega$ is given by the unique solution Θ^* of the* 1-step atomic Bellman optimality equation*: for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, for all $i \in \{1, \ldots, N\}$ such that $\Omega_i(x, a) \neq 0$,*

$$\Theta^*_i(x,a) = r(x,a) + \frac{\gamma}{\Omega_i(x,a)} \sum_{q \in \mathcal{Q}_{\max}(\Theta^*)} \mu_i^*(\Theta^*, x, a, q) \cdot q, \tag{VII.9}$$

where the weight functions μ_i^* are given by Eq. (6) with $\Omega' = \Omega$.

Here also, the averaging property holds: it allows to recover the optimal state-action value function Q^* from the atoms Θ^*.

Property 3. *Let Z_{Ω,Θ^*} be the fixed point of the 1-SABOO $\mathbb{T}_\Omega$. Then for all (x, a),*

$$\sum_{i=1}^N \Omega_i(x,a)\Theta^*_i(x,a) = Q^*(x,a).$$

From Atomic Bellman Operators to DRL. In practice, the model P is unknown to the learner. Hence, the update rules from lemmas 4, 5, 6 cannot be exactly applied. In the next section, we promote DRL approaches approximating the weight functions μ_i^π, μ_i^* by plug-in estimators obtained by sequentially learning the c.d.f.'s $G_{x,a}^\pi, G_{x,a}^*$ and their left limits $H_{x,a}^\pi, H_{x,a}^*$ in a temporal difference fashion as in [MSK$^+$10], [MSK$^+$12].

7 Atomic DRL Algorithms

From our atomic operators/equations, we derive two DRL algorithms for policy evaluation and one for the control task. For each of the three algorithms, we describe one iteration based on a single transition. We denote by $N \geq 1$ the number of atoms, $\Omega_i(x,a)$ the probability weights, $\Theta_i(x,a)$ the atoms, $G_{x,a}(\theta)$ the c.d.f.'s, $H_{x,a}(\theta)$ the 'left limit' functions. All our methods rely on two nested learning procedures, both based on the computation of exponentially weighted moving averages:

(1) the first for learning the functions $G_{x,a}, H_{x,a}$ with a learning rate $0 < \beta \leq 1$,

(2) the second for learning the atoms $\Theta_i(x,a)$ with a learning rate $0 < \alpha \leq 1$.

7.1 Atomic Temporal-Difference Learning

Based on the (1-step) atomic Bellman equations derived in the previous section, we adapt the temporal-difference (TD) learning algorithm ([Sut88]), suited to non-distributional RL, to our multiatomic framework. We present two algorithms for the policy evaluation task. Consider a policy π and a single transition $x, a, r(x,a), X_1, A_1$ such that $X_1 \sim P(\cdot|x,a), A_1 \sim \pi(\cdot|X_1)$.

Atomic Temporal-Difference (ATD) - Stochastic Approximation of the ABO. Initialize $\widetilde{\Theta} = \emptyset$, then update for all $x' \in \mathcal{X}$, $a' \in \mathcal{A}$, $j \in \{1, \ldots, N\}$,

(a) $\theta \leftarrow \Theta_j(x', a')$,

(b) $\widetilde{\Theta} \leftarrow \widetilde{\Theta} \cup \{\theta\}$,

(c) $G_{x,a}(\theta) \leftarrow (1-\beta) G_{x,a}(\theta) + \beta \sum_{k=1}^N \Omega_k(X_1, A_1) \mathbb{I}\{\Theta_k(X_1, A_1) \leq \theta\}$,

(d) $H_{x,a}(\theta) \leftarrow (1-\beta) H_{x,a}(\theta) + \beta \sum_{k=1}^N \Omega_k(X_1, A_1) \mathbb{I}\{\Theta_k(X_1, A_1) < \theta\}$,

(e) $\forall 1 \leq i \leq N$, $\mu_i(\Theta, x, a, \theta) \leftarrow \max\big\{0, \min\{\overline{\Omega}_i(x,a), G_{x,a}(\theta)\} - \max\{\overline{\Omega}_{i-1}(x,a), H_{x,a}(\theta)\}\big\}$.

Then, return the updated atoms in state-action (x,a): for $1 \leq i \leq N$,

$$\Theta_i(x,a) \leftarrow (1-\alpha)\Theta_i(x,a) + \alpha\Big(r(x,a) + \frac{\gamma}{\Omega_i(x,a)} \sum_{\theta \in \widetilde{\Theta}} \mu_i(\Theta, x, a, \theta) \cdot \theta\Big).$$

1-Step Atomic Temporal-Difference (1-SATD) - Stochastic Approximation of the 1-SABO. Initialize $\mathcal{Q}(\Theta) = \emptyset$, then update for all $x' \in \mathcal{X}$, $a' \in \mathcal{A}$,

(a) $q \leftarrow \sum_{j=1}^{N} \Omega_j(x', a')\Theta_j(x', a')$,

(b) $\mathcal{Q}(\Theta) \leftarrow \mathcal{Q}(\Theta) \cup \{q\}$,

(c) $G_{x,a}(q) \leftarrow (1-\beta)G_{x,a}(q) + \beta\mathbb{I}\{\sum_{k=1}^{N} \Omega_k(X_1, A_1)\Theta_k(X_1, A_1) \leq q\}$,

(d) $H_{x,a}(q) \leftarrow (1-\beta)H_{x,a}(q) + \beta\mathbb{I}\{\sum_{k=1}^{N} \Omega_k(X_1, A_1)\Theta_k(X_1, A_1) < q\}$,

(e) $\forall 1 \leq i \leq N$, $\mu_i(\Theta, x, a, q) \leftarrow \max\big\{0, \min\{\overline{\Omega}_i(x, a), G_{x,a}(q)\} - \max\{\overline{\Omega}_{i-1}(x, a), H_{x,a}(q)\}\big\}$.

Then, return the updated atoms in state-action (x, a): for $1 \leq i \leq N$,

$$\Theta_i(x, a) \leftarrow (1-\alpha)\Theta_i(x, a) + \alpha\Big(r(x, a) + \frac{\gamma}{\Omega_i(x, a)} \sum_{q \in \mathcal{Q}(\Theta)} \mu_i(\Theta, x, a, q) \cdot q\Big).$$

We point out that both the ATD and the 1-SATD algorithms coincide with TD(0) (for Q-functions) if $N = 1$ and $\beta = 1$.

7.2 Atomic Q-Learning

Now, we rely on the 1-step atomic Bellman optimality equation to define our DRL algorithm for the control task. Consider a single transition $x, a, r(x, a), X_1$ with $X_1 \sim P(\cdot|x, a)$.

Atomic Q-Learning - Stochastic Approximation of the 1-SABOO. Initialize $\mathcal{Q}_{\max}(\Theta) = \emptyset$, then update for all $x' \in \mathcal{X}$,

(a) $q \leftarrow \max_{a' \in \mathcal{A}} \sum_{j=1}^{N} \Omega_j(x', a')\Theta_j(x', a')$,

(b) $\mathcal{Q}_{\max}(\Theta) \leftarrow \mathcal{Q}_{\max}(\Theta) \cup \{q\}$,

(c) $G_{x,a}(q) \leftarrow (1-\beta)G_{x,a}(q) + \beta\mathbb{I}\{\max_{a' \in \mathcal{A}} \sum_{k=1}^{N} \Omega_k(X_1, a')\Theta_k(X_1, a') \leq q\}$,

(d) $H_{x,a}(q) \leftarrow (1-\beta)H_{x,a}(q) + \beta\mathbb{I}\{\max_{a' \in \mathcal{A}} \sum_{k=1}^{N} \Omega_k(X_1, a')\Theta_k(X_1, a') < q\}$,

(e) $\forall 1 \leq i \leq N$, $\mu_i(\Theta, x, a, q) \leftarrow \max\big\{0, \min\{\overline{\Omega}_i(x, a), G_{x,a}(q)\} - \max\{\overline{\Omega}_{i-1}(x, a), H_{x,a}(q)\}\big\}$.

Then, return the updated atoms in state-action (x, a): for $1 \leq i \leq N$,

$$\Theta_i(x, a) \leftarrow (1-\alpha)\Theta_i(x, a) + \alpha\Big(r(x, a) + \frac{\gamma}{\Omega_i(x, a)} \sum_{q \in \mathcal{Q}_{\max}(\Theta)} \mu_i(\Theta, x, a, q) \cdot q\Big).$$

We point out that this approach can be seen as an extension of the Q-learning algorithm ([Wat89]). Indeed, Atomic Q-Learning boils down to Q-learning in the monoatomic case $N = 1$ combined with c.d.f. learning rate $\beta = 1$.

8 Distributional Policy Evaluation in a Two States MDP

We consider a policy evaluation setting in a 2 states MDP characterized by:

- state space $\mathcal{X} = \{x_1, x_2\}$,
- action space $\mathcal{A} = \{a\}$,
- transition probabilities: for all $(x, x') \in \{x_1, x_2\}^2$, $P(x'|x, a) = 1/2$,
- deterministic rewards $R(x, a) = \delta_{r(x,a)}$ with $r(x_1, a) = 0$, $r(x_2, a) = 1$,
- discount rate $\gamma = 1/2$.

In fact, it is the same MDP as in Figure I.9 except that the action space is restricted to the singleton $\{a_1\}$. Then, we denote by π the unique policy: it verifies $\pi(a|x_1) = \pi(a|x_2) = 1$. Basic computation shows that $Z^\pi(x, a)$ is a uniform distribution on the interval $[0, 1]$ if $x = x_1$ or on $[1, 2]$ if $x = x_2$.

The Figure VII.2 displays four dotted lines corresponding to the trajectories of the $N = 4$ atoms of an atomic distribution $Z_{\Omega,\Theta}$ ($\Omega_i(x, a) \equiv 1/4$, $\Theta_i(x, a) \equiv (i-1)/4$ initially) on which we recursively apply the atomic Bellman operator $\mathcal{T}^\pi_\Omega$, by implementing the formula from Lemma 4. In less than ten iterations, the atoms have already converged to the values of the atoms $\Theta^\pi_i(x, a)$ of the fixed point distribution $Z_{\Omega,\Theta^\pi}(x, a)$ for the two states $x \in \{x_1, x_2\}$. Moreover, the dark dashed line representing the average of the four atoms experimentally confirms the Property 1 as $Q^\pi(x_1, a) = 1/2$ and $Q^\pi(x_2, a) = 3/2$. On Figure VII.3, obtained by running the tabular ATD algorithm introduced in the previous section, the atoms are converging to the same values as with the exact method; the probabilities $\Omega_i(x, a) \equiv 1/4$ are uniform and, in each of the 1000 independent instances, the atoms $\Theta_i(x, a)$ are randomly initialized from the uniform distribution $\mathcal{U}([0, 1])$.

9 Conclusion

In this chapter, we extended existing DRL approaches based on atomic distributional approximations in terms of p-Wasserstein metrics. We discussed different approximation schemes for particular Wasserstein metrics (W_1, W_2, W_∞) and gave contraction guarantees for the corresponding projected Bellman operators. A careful study of the 2-Wasserstein case led to our main contribution, namely a generalization of the Bellman equations. Given the transition probabilities, these new atomic Bellman equations allow to approximate the policy distribution by a fixed point distribution with same expectation. The empirical study of a simple two states MDP is presented as an illustration of the theory.

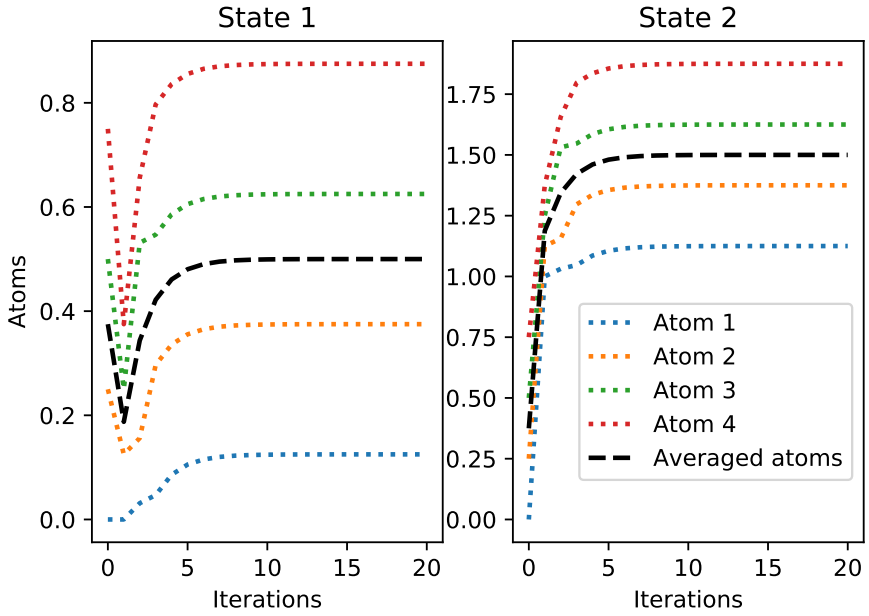


Figure VII.2: Exact dynamic programming approach ('model P is known'). The $N = 4$ atoms (dotted lines) converge to the atoms of the fixed point distribution $Z_{\Omega,\Theta^\pi}(x, a)$ by iteratively applying the atomic Bellman operator $\mathcal{T}_\Omega^\pi$ (state $x = x_1$ on the left, state $x = x_2$ on the right). As expected from Property 1, the average of the 4 atoms (dashed line) converges to the theoretical Q-value in each state: $Q^\pi(x_1, a) = 1/2$ and $Q^\pi(x_2, a) = 3/2$.

10 Perspective - Optimal Mass Allocation

So far we have been projecting distributions for given probabilities over the atoms. Here, we define a notion of optimality for the probability functions Ω and we show its link with the fixed point of the atomic Bellman operator.

Definition 3. *Let $Z \in \mathcal{Z}$. A mass allocation function $\Omega^* : \mathcal{X} \times \mathcal{A} \to \Delta_N$ is Z-optimal if it minimizes for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ the 2-Wasserstein error*

$$W_2(Z(x, a), \Pi_{2,\Omega^*} Z(x, a)) = \min_{\omega \in \Delta_N} W_2(Z(x, a), \Pi_{2,\omega} Z(x, a)).$$

Notice that if $Z = Z_{\Omega,\Theta} \in \mathcal{Z}_N$, then $\Omega^* = \Omega$ is Z-optimal. Moreover, $\Pi_{2,\Omega^*} Z = Z$ and the Wasserstein error is equal to zero everywhere on $\mathcal{X} \times \mathcal{A}$: $\widetilde{W}_2(Z, \Pi_{2,\Omega^*} Z(x, a)) = 0$.

W_2-Error & Trimmed Variances. The 2-Wasserstein optimal mass functions are simply maximizing dot products with the squared trimmed means.

Proposition 5. *Let $Z \in \mathcal{Z}$. A mass allocation function Ω^* is Z-optimal if and only if for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\Omega^*(x, a)$ maximizes over Δ_N the dot product*

$$\Omega^*(x, a) \in \underset{\omega \in \Delta_N}{arg\ max}\ \langle \omega, \theta^{*2}_{\omega,x,a} \rangle,$$

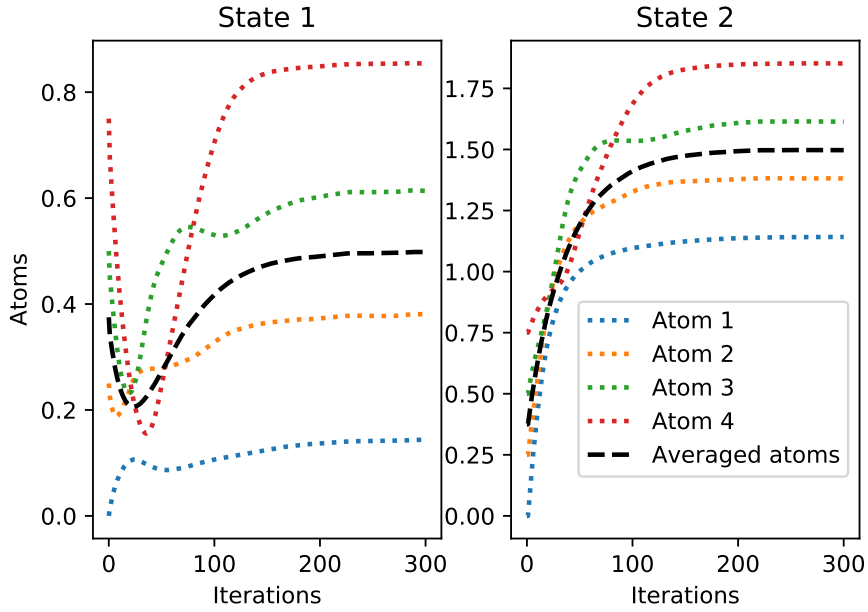


Figure VII.3: Stochastic DRL approach ('model P is unknown'). The curves are averaged over 1000 instances of the ATD algorithm run on 300 iterations with learning rates $\alpha = \beta = 0.1$.

with trimmed means vector $\theta^*_{\omega,x,a}$ *given by Eq. (VII.4) with* $F = F_{x,a}$ *the c.d.f. of* $Z(x,a)$*:*

$$\forall i \text{ s.t. } \omega_i \neq 0, \quad \theta^*_{\omega,x,a,i} = \frac{1}{\omega_i} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} F^{-1}_{x,a}(\tau) d\tau,$$

and componentwise squared operation: $\theta^{*2}_{\omega,x,a} = (\theta^{*2}_{\omega,x,a,1}, \ldots, \theta^{*2}_{\omega,x,a,N})$.

The proof relies on decomposing the squared Wasserstein error as a weighted sum of trimmed variances. Proposition 5 reduces the optimal mass allocation problem to maximizing dot products over the simplex, which may be easily implemented in practical algorithms.

11 Technical Proofs

For notational convenience, we will sometimes denote the expectation of a distribution D by $\mathbb{E}[D] = \mathbb{E}_{Y\sim D}[Y]$, and the p-Wasserstein distance between the respective distributions D, D' of two random variables Y, Y' by $W_p(Y,Y') = W_p(D,D')$.

Partition Lemma

We state a stronger version of the 'partition lemma' given in [BDM17]: it will be useful to prove lemmas 1 and 2.

Lemma 7 (Partition Lemma)**.** *Let $(C_i)_{i\in\mathbb{N}}$ be a sequence of binary random variables valued in $\{0,1\}$ such that $\sum_{i\geq 0} C_i = 1$ almost surely. Then for all $p \in [1,+\infty]$ and real-valued random variables Y, Y',*

$$W_p^p(Y,Y') \leq \sum_{i\geq 0} W_p^p(C_iY, C_iY').$$

We point out that Lemma 7 implies Lemma 1 in [BDM17], which states that

$$W_p(Y,Y') \leq \sum_{i\geq 0} W_p(C_iY, C_iY').$$

In fact, the proof of Lemma 1 in [BDM17] also proves Lemma 7.

Proof of Lemma 1

Let $p \in [1,+\infty]$, Z, Z' in $\mathcal{Z}$. We have (with some abuse of notations):

$$\begin{aligned}
W_p^p(\mathbb{T}^\pi Z(x,a), \mathbb{T}^\pi Z'(x,a)) &= W_p^p(R_0 + \gamma\mathbb{E}[Z_1|X_1,A_1], R_0 + \gamma\mathbb{E}[Z_1'|X_1,A_1]) \\
&\leq \gamma^p W_p^p(\mathbb{E}[Z_1|X_1,A_1], \mathbb{E}[Z_1'|X_1,A_1]) \\
&\leq \gamma^p \sum_{x',a'} W_p^p(\mathbb{I}\{X_1=x',A_1=a'\}\mathbb{E}[Z(x',a')], \mathbb{I}\{X_1=x',A_1=a'\}\mathbb{E}[Z'(x',a')]) \\
&\leq \gamma^p \sum_{x',a'} \mathbb{P}(x'|x,a)\pi(a'|x')|\mathbb{E}[Z(x',a')] - \mathbb{E}[Z'(x',a')]|^p \\
&\leq \gamma^p \sup_{x',a'} W_p^p(Z(x',a'), Z'(x',a')) = \gamma^p \widetilde{W}_p^p(Z,Z'),
\end{aligned}$$

where we used Lemma 7 in the second inequality and Hölder's inequality in the last one.

Proof of Lemma 2

Let $p \in [1,+\infty]$, Z, Z' in $\mathcal{Z}$. We have:

$$\begin{aligned}
W_p^p(\mathbb{T}Z(x,a), \mathbb{T}Z'(x,a)) &= W_p^p(R_0 + \gamma \max_{a'} \mathbb{E}[Z_{1,a'}|X_1], R_0 + \gamma \max_{a'} \mathbb{E}[Z_{1,a'}'|X_1]) \\
&\leq \gamma^p W_p^p(\max_{a'} \mathbb{E}[Z_{1,a'}|X_1], \max_{a'} \mathbb{E}[Z_{1,a'}'|X_1]) \\
&\leq \gamma^p \sum_{x'} W_p^p(\mathbb{I}\{X_1=x'\} \max_{a'} \mathbb{E}[Z(x',a')], \mathbb{I}\{X_1=x'\} \max_{a'} \mathbb{E}[Z'(x',a')]) \\
&\leq \gamma^p \sum_{x'} \mathbb{P}(x'|x,a)| \max_{a'} \mathbb{E}[Z(x',a')] - \max_{a''} \mathbb{E}[Z'(x',a'')]|^p \\
&\leq \gamma^p \sum_{x'} \mathbb{P}(x'|x,a) \max_{a'} |\mathbb{E}[Z(x',a')] - \mathbb{E}[Z'(x',a')]|^p \\
&\leq \gamma^p \sup_{x',a'} W_p^p(Z(x',a'), Z'(x',a')) = \gamma^p \widetilde{W}_p^p(Z,Z'),
\end{aligned}$$

where we used Lemma 7 in the second inequality and Hölder's inequality in the last one.

Proof of Lemma 3

Let us prove that the projection $\Pi_{2,\Omega}$ is a non-expansion in $\widetilde{W}_\infty$. Let $(Z, Z') \in \mathcal{Z}^2$ such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, $Z(x,a)$ and $Z'(x,a)$ have respective c.d.f.'s $F_{x,a}$ and $G_{x,a}$. We have:

$$W_\infty\big(\Pi_{2,\Omega}Z(x,a), \Pi_{2,\Omega}Z'(x,a)\big) = \max_{1\le i\le N} \frac{1}{\Omega_i(x,a)} \left| \int_{(\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} F^{-1}_{x,a}(\tau)d\tau - \int_{(\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} G^{-1}_{x,a}(\tau)d\tau \right|.$$

Then,

$$W_\infty\big(\Pi_{2,\Omega}Z(x,a), \Pi_{2,\Omega}Z'(x,a)\big) \le \max_{1\le i\le N} \operatorname*{ess\,sup}_{\tau\in(\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} \big|F^{-1}_{x,a}(\tau) - G^{-1}_{x,a}(\tau)\big| = W_\infty(Z(x,a), Z'(x,a)),$$

where the inequality holds because quantile functions are left continuous with right limits. Taking the supremum over $\mathcal{X} \times \mathcal{A}$ on both sides of the previous inequality concludes the proof.

Proof of Lemma 4

First notice that for $Z = Z_{\Omega,\Theta} \in \mathcal{Z}_\Omega$ and deterministic rewards $R(x,a) = \delta_{r(x,a)}$,

$$\mathcal{T}^\pi Z_{\Omega,\Theta}(x,a) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x') \sum_{j=1}^N \Omega_j(x',a')\delta_{r(x,a)+\gamma\Theta_j(x',a')},$$

and, for $X_1 \sim P(\cdot|x,a), A_1 \sim \pi(\cdot|X_1)$, the mixture of Dirac distributions

$$Z_{\Omega,\Theta}(X_1, A_1) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x') \sum_{j=1}^N \Omega_j(x',a')\delta_{\Theta_j(x',a')}$$

has the following c.d.f.:

$$\forall z \in \mathbb{R}, \quad G^\pi_{x,a}(z) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x') \sum_{j=1}^N \Omega_j(x',a')\mathbb{I}\{\Theta_j(x',a') \le z\}.$$

We conclude the proof by observing that

$$\forall \tau \in (0,1], \quad G^{\pi -1}_{x,a}(\tau) = \sum_{\theta\in\widetilde{\Theta}} \theta \cdot \mathbb{I}\Big\{G^\pi_{x,a}(\theta-) < \tau \le G^\pi_{x,a}(\theta)\Big\},$$

which implies:

$$\int_{\tau=\overline{\Omega'}_{i-1}(x,a)}^{\overline{\Omega'}_i(x,a)} G^{\pi -1}_{x,a}(\tau)d\tau = \sum_{\theta\in\widetilde{\Theta}} \mu^\pi_i(\Theta, x, a, \theta) \cdot \theta\,,$$

where

$$\mu^\pi_i(\Theta, x, a, \theta) = \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[G^\pi_{x,a}(\theta-),\, G^\pi_{x,a}(\theta)\big]\big).$$

Proof of Lemma 5

Similarly to the proof of Lemma 4, we first write:

$$\mathbb{T}^\pi Z_{\Omega,\Theta}(x,a) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x')\delta_{r(x,a)+\gamma Q(x',a')},$$

with Q-values

$$Q(x',a') = \sum_{j=1}^{N} \Omega_j(x',a')\Theta_j(x',a'), \quad \forall (x',a') \in \mathcal{X}\times\mathcal{A}.$$

In addition, the mixture of Dirac distributions

$$\sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x')\delta_{Q(x',a')}$$

has the following c.d.f.:

$$\forall z \in \mathbb{R}, \quad G^\pi_{x,a}(z) = \sum_{(x',a')\in\mathcal{X}\times\mathcal{A}} P(x'|x,a)\pi(a'|x')\mathbb{I}\{Q(x',a') \le z\}.$$

We conclude the proof by observing that

$$\forall \tau \in (0,1], \quad G^{\pi -1}_{x,a}(\tau) = \sum_{q\in\mathcal{Q}(\Theta)} q \cdot \mathbb{I}\Big\{G^\pi_{x,a}(q-) < \tau \le G^\pi_{x,a}(q)\Big\},$$

which implies:

$$\int_{\tau=\overline{\Omega'}_{i-1}(x,a)}^{\overline{\Omega'}_i(x,a)} G^{\pi -1}_{x,a}(\tau)d\tau = \sum_{q\in\mathcal{Q}(\Theta)} \mu^\pi_i(\Theta,x,a,q)\cdot q,$$

where

$$\mu^\pi_i(\Theta,x,a,q) = \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[G^\pi_{x,a}(q-),\, G^\pi_{x,a}(q)\big]\big).$$

Proof of Lemma 6

We have:

$$\mathbb{T}Z_{\Omega,\Theta}(x,a) = \sum_{x'\in\mathcal{X}} P(x'|x,a)\delta_{r(x,a)+\gamma \max_{a'\in\mathcal{A}} Q(x',a')},$$

with Q-values defined as in Lemma 5:

$$Q(x',a') = \sum_{j=1}^{N} \Omega_j(x',a')\Theta_j(x',a'), \quad \forall (x',a') \in \mathcal{X}\times\mathcal{A}.$$

In addition, the mixture of Dirac distributions

$$\sum_{x'\in\mathcal{X}} P(x'|x,a)\delta_{\max_{a'\in\mathcal{A}} Q(x',a')}$$

has the following c.d.f.:

$$\forall z \in \mathbb{R}, \quad G^*_{x,a}(z) = \sum_{x' \in \mathcal{X}} P(x'|x,a) \mathbb{I}\left\{ \max_{a' \in \mathcal{A}} Q(x',a') \leq z \right\}.$$

Finally we have,

$$\forall \tau \in (0,1], \quad G^{*-1}_{x,a}(\tau) = \sum_{q \in \mathcal{Q}_{\max}(\Theta)} q \cdot \mathbb{I}\Big\{ G^*_{x,a}(q-) < \tau \leq G^*_{x,a}(q) \Big\},$$

which implies:

$$\int_{\tau=\overline{\Omega'}_{i-1}(x,a)}^{\overline{\Omega'}_i(x,a)} G^{*-1}_{x,a}(\tau) d\tau = \sum_{q \in \mathcal{Q}_{\max}(\Theta)} \mu^*_i(\Theta, x, a, q) \cdot q \,,$$

where

$$\mu^*_i(\Theta, x, a, q) = \lambda\big(\big[\overline{\Omega'}_{i-1}(x,a),\, \overline{\Omega'}_i(x,a)\big] \cap \big[G^*_{x,a}(q-),\, G^*_{x,a}(q)\big] \big).$$

Proof of Proposition 1

We have,

$$\begin{aligned} \widetilde{W}_\infty(Z_\Omega, Z) \leq \widetilde{W}_\infty(Z_\Omega, \Pi_{2,\Omega} Z) + \widetilde{W}_\infty(\Pi_{2,\Omega} Z, Z) &= \widetilde{W}_\infty(T_\Omega Z_\Omega, \Pi_{2,\Omega} T Z) + \widetilde{W}_\infty(\Pi_{2,\Omega} Z, Z) \\ &\leq \gamma \widetilde{W}_\infty(Z_\Omega, Z) + \widetilde{W}_\infty(\Pi_{2,\Omega} Z, Z), \end{aligned}$$

where the first inequality is a triangular inequality and the second inequality comes from Corollary 1 stating that $T_\Omega = \Pi_{2,\Omega} T$ is a γ-contraction for $\widetilde{W}_\infty$. Hence,

$$\widetilde{W}_\infty(Z_\Omega, Z) \leq \frac{1}{1-\gamma} \widetilde{W}_\infty(\Pi_{2,\Omega} Z, Z).$$

Then, by denoting $I(x,a) = \{i \in \{1, \dots, N\} \text{ s.t. } \Omega_i(x,a) \neq 0\}$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, we have:

$$\begin{aligned} &W_\infty(\Pi_{2,\Omega} Z(x,a), Z(x,a)) = \\ &\qquad \max_{i \in I(x,a)} \operatorname*{ess\ sup}_{\tau \in (\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} \left| F^{-1}_{x,a}(\tau) - \frac{1}{\Omega_i(x,a)} \int_{\tau'=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} F^{-1}_{x,a}(\tau') d\tau' \right| \\ &\qquad \leq \max_{i \in I(x,a)} \operatorname*{ess\ sup}_{\tau \in (\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} \frac{1}{\Omega_i(x,a)} \int_{\tau'=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} \left| F^{-1}_{x,a}(\tau) - F^{-1}_{x,a}(\tau') \right| d\tau' \\ &\leq \max_{i \in I(x,a)} \operatorname*{ess\ sup}_{\tau \in (\overline{\Omega}_{i-1}(x,a), \overline{\Omega}_i(x,a)]} \max\left\{ \left| F^{-1}_{x,a}(\tau) - F^{-1}_{x,a}(\overline{\Omega}_{i-1}(x,a)+) \right|, \left| F^{-1}_{x,a}(\tau) - F^{-1}_{x,a}(\overline{\Omega}_i(x,a)) \right| \right\} \\ &\qquad \leq \max_{i \in I(x,a)} F^{-1}_{x,a}(\overline{\Omega}_i(x,a)) - F^{-1}_{x,a}(\overline{\Omega}_{i-1}(x,a)+), \end{aligned}$$

which concludes the proof.

Proof of Proposition 5

We have,

$$
\begin{aligned}
W_2^2(Z(x,a), Z_{\Omega,\Theta^*_\Omega}(x,a)) = \\
&\sum_{i=1}^{N} \int_{\tau=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} \left(F^{-1}_{x,a}(\tau) - \frac{1}{\Omega_i(x,a)} \int_{\tau'=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} F^{-1}_{x,a}(\tau') d\tau' \right)^2 d\tau \\
&= \int_{\tau=0}^{1} F^{-1}_{x,a}(\tau)^2 d\tau - \sum_{i=1}^{N} \frac{1}{\Omega_i(x,a)} \left(\int_{\tau=\overline{\Omega}_{i-1}(x,a)}^{\overline{\Omega}_i(x,a)} F^{-1}_{x,a}(\tau) d\tau \right)^2 \\
&= \mathbb{E}_{Z_0 \sim Z(x,a)} \left[Z_0^2 \right] - \langle \Omega(x,a), \Theta^*_\Omega(x,a)^2 \rangle,
\end{aligned}
$$

where $\langle \cdot, \cdot \rangle$ is the dot product on $\mathbb{R}^N$ and the squared operation is applied on each component of the trimmed means vector: $\Theta^*_\Omega(x,a)^2 = (\Theta^*_{\Omega,1}(x,a)^2, \ldots, \Theta^*_{\Omega,N}(x,a)^2)$.

Bibliography

[ABGK12] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[ACG+17] M. Achab, S. Clémençon, A. Garivier, A. Sabourin, and C. Vernade. Max k-armed bandit: On the extremehunter algorithm and beyond. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 389–404. Springer, 2017.

[ACG18] M. Achab, S. Clémençon, and A. Garivier. Profitable bandits. *arXiv preprint arXiv:1805.02908*, 2018.

[ACP19] G. Ausset, S. Clémençon, and F. Portier. Empirical risk minimization under random censorship: Theory and practice. *arXiv preprint arXiv:1906.01908*, 2019.

[ADEH99] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[AGH+05] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.

[AGR17] J. Aledo, J. Gámez, and A. Rosete. Utopia in the solution of the bucket order problem. *Decision Support Systems*, 97:69–80, 2017.

[AGR18] J. Aledo, J. Gámez, and A. Rosete. Approaching rank aggregation problems by using evolution strategies: the case of the optimal bucket order problem. *European Journal of Operational Research*, 2018.

[AKC18] M. Achab, A. Korba, and S. Clémençon. Dimensionality reduction and (bucket) ranking: a mass transportation approach, 2018.

[AM12] A. Ali and M. Meilă. Experiments with kemeny ranking: What works when? *Mathematical Social Sciences*, 64(1):28–40, 2012.

[AMS09] J. Audibert, R. Munos, and C. Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[Arr12] K. J. Arrow. *Social choice and individual values*, volume 12. Yale university press, 2012.

[BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification : a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[BCB+12] S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[BCL13] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

[BCZ+16] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *NIPS*, page 4349–4357, 2016.

[BD18] J. Bekker and J. Davis. Beyond the selected completely at random assumption for learning from positive and unlabeled data. *CoRR*, abs/1809.03207, 2018.

[BDBC+10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1), 2010.

[BDM17] M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 449–458. JMLR. org, 2017.

[Bel66] R. Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[BGST06] J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

[BHH06] F. Bach, D. Heckerman, and E. Horvitz. Considering cost asymmetry in learning classifiers, 2006.

[BHS+19] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. 2019.

[Bis06] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[BK96] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[BK07] R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[BMHB+18] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, A. Muldal, N. Heess, and T. Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

[Bor84] J. d. Borda. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*, 1784.

[BT96] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

[BT15] S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. *Electron. J. Statist.*, 9(2):2751–2792, 2015.

[C+15] F. Chollet et al. Keras. `https://keras.io`, 2015.

[CA17] S. Clémençon and M. Achab. Ranking data with continuous labels through oriented recursive partitions. In *Advances in Neural Information Processing Systems*, pages 4600–4608, 2017.

[CBCTH13] X. Chen, P. Bennett, K. Collins-Thompson, and E. Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202. ACM, 2013.

[CDV13a] G. Clémençon, M. Depecker, and N. Vayatis. Ranking Forests. *J. Mach. Learn. Res.*, 14:39–73, 2013.

[CDV13b] S. Clémençon, M. Depecker, and N. Vayatis. An empirical comparison of learning algorithms for nonparametric scoring: the treerank algorithm and other methods. *Pattern Analysis and Applications*, 16(4):475–496, 2013.

[CF04] G. Creamer and Y. Freund. Predicting performance and quantifying corporate governance risk for latin american adrs and banks. *FINANCIAL ENGINEERING AND APPLICATIONS, MIT, Cambridge*, 2004.

[CGM+13] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, Jun. 2013.

[CJ10] S. Clémençon and J. Jakubowicz. Kantorovich distances between rankings with applications to rank aggregation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 248–263. Springer, 2010.

[CK14a] A. Carpentier and A. K. Kim. Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, 25:1133–1144, 2014.

[CK+14b] A. Carpentier, A. K. Kim, et al. Honest and adaptive confidence interval for the tail coefficient in the pareto model. *Electronic Journal of Statistics*, 8(2):2066–2110, 2014.

[CK15] W. Cowan and M. N. Katehakis. Asymptotically optimal sequential experimentation under generalized ranking. *arXiv preprint arXiv:1510.02041*, 2015.

[CLV05] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *International Conference on Computational Learning Theory*, pages 1–15. Springer, 2005.

[CLV08] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[CM04] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In *Advances in neural information processing systems*, pages 313–320, 2004.

[CMM10] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pages 442–450, 2010.

[CMRR08] C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *International conference on algorithmic learning theory*, pages 38–53. Springer, 2008.

[CMZ18] A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. *arXiv preprint arXiv:1806.01380*, 2018.

[CR14] S. Clémençon and S. Robbiano. The TreeRank Tournament algorithm for multipartite ranking. *Journal of Nonparametric Statistics*, 25(1):107–126, 2014.

[CS05] V. A. Cicirello and S. F. Smith. The max k-armed bandit: A new model of exploration applied to search heuristic selection. In *The Proceedings of the Twentieth National Conference on Artificial Intelligence*, volume 3, pages 1355–1361. AAAI Press, July 2005.

[CV09a] S. Clémençon and N. Vayatis. Empirical performance maximization based on linear rank statistics. In *NIPS*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2009.

[CV09b] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[CV10] S. Clémençon and N. Vayatis. The RankOver algorithm: overlaid classification rules for optimal ranking. *Constructive Approximation*, 32:619–648, 2010.

[CV14] A. Carpentier and M. Valko. Extreme bandits. In *Advances in Neural Information Processing Systems 27*, pages 1089–1097. Curran Associates, Inc., 2014.

[CV15] A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1133–1141, 2015.

[Day92] P. Dayan. The convergence of td (λ) for general λ. *Machine learning*, 8(3-4):341–362, 1992.

[DC$^+$14] N. De Condorcet et al. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Cambridge University Press, 2014.

[DD09] M. Deza and E. Deza. *Encyclopedia of Distances*. Springer, 2009.

[DDB18] A. Das, A. Dantcheva, and F. Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[Dev08] L. Devroye. Non-uniform random variate generation (1986), 2008.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

[DM59] D. Davidson and J. Marschak. Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, 17:274, 1959.

[DOSM18] W. Dabney, G. Ostrovski, D. Silver, and R. Munos. Implicit quantile networks for distributional reinforcement learning. *arXiv preprint arXiv:1806.06923*, 2018.

[dPNS14] M. C. du Plessis, G. Niu, and M. Sugiyama. Analysis of learning from positive and unlabeled data. In *NIPS*, pages 703–711, 2014.

[dPNS15] M. C. du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *ICML*, ICML'15, pages 1386–1394. JMLR.org, 2015.

[dPS14] M. C. du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE TRANSACTIONS on Information and Systems*, 97(5):1358–1362, 2014.

[DRBM18] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos. Distributional reinforcement learning with quantile regression. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[DS16] Y. David and N. Shimkin. Pac lower bounds and efficient algorithms for the max k-armed bandit problem. In *Proceedings of The 33nd International Conference on Machine Learning*, 2016.

[EA13] E. V. B. P. P. E. Abbasnejad, S. Sanner. Learning community-based preferences via dirichlet process mixtures of gaussian processes. In *In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[EH13] P. Embrechts and M. Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432, 2013.

[FFN08] J. Feng, Q. Fang, and W. Ng. Discovering bucket orders from full rankings. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 55–66. ACM, 2008.

[FH11] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

[Fis73] P. C. Fishburn. Binary choice probabilities: on the varieties of stochastic transitivity. *Journal of Mathematical psychology*, 10(4):327–352, 1973.

[FISS03] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

[GC11] A. Garivier and O. Cappé. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. *ArXiv e-prints*, February 2011.

[GDA+17] A. Gruslys, W. Dabney, M. G. Azar, B. Piot, M. Bellemare, and R. Munos. The reactor: A fast and sample-efficient actor-critic agent for reinforcement learning. *arXiv preprint arXiv:1704.04651*, 2017.

[GF15] J. García and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[GMPU06] A. Gionis, H. Mannila, K. Puolamäki, and A. Ukkonen. Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566. ACM, 2006.

[GMS16] A. Garivier, P. Ménard, and G. Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. *ArXiv e-prints*, February 2016.

[GMS19] A. Garivier, P. Ménard, and G. Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[GST13] N. Galichet, M. Sebag, and O. Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

[GV14] J. Garcke and T. Vanck. Importance weighted inductive transfer learning for regression. In *ECML PKDD*, pages 466–481. Springer, 2014.

[HGB+07] J. Huang, A. Gretton, K. M. Borgwardt, B. Schölkopf, and A. J. Smola. Correcting sample selection bias by unlabeled data. In *NIPS*, pages 601–608, 2007.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[Hud08] O. Hudry. NP-hardness results for the aggregation of linear orders into median orders. *Ann. Oper. Res.*, 163:63–88, 2008.

[HW85] P. Hall and A. H. Welsh. Adaptive estimates of parameters of regular variation. *Ann. Statist.*, 13(1):331–341, 03 1985.

[HZRS15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[JGP+17] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. Acm New York, NY, USA, 2017.

[JKS16] Y. Jiao, A. Korba, and E. Sibony. Controlling the distance to a kemeny consensus without computing it. In *Proceeding of ICML 2016*, 2016.

[JKSO16] M. Jang, S. Kim, C. Suh, and S. Oh. Top-k ranking from pairwise comparisons: When spectral ranking is optimal. *arXiv preprint*, 2016.

[Kam03] T. Kamishima. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 583–588. ACM, 2003.

[Kau16] E. Kaufmann. On Bayesian index policies for sequential resource allocation. *ArXiv e-prints*, January 2016.

[KB05] V. Koltchinskii and O. Beznosova. Exponential convergence rates in classification. In *Proceedings of COLT 2005*, 2005.

[KCG12] E. Kaufmann, O. Cappé, and A. Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600, 2012.

[KCS17] A. Korba, S. Clémençon, and E. Sibony. A learning theory of ranking aggregation. In *Proceeding of AISTATS 2017*, 2017.

[Kem] J. G. Kemeny. Mathematics without numbers. *Daedalus*, (88):571–591.

[Kem59] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88:571–591, 1959.

[KJ+19] R. K. Kolla, K. Jagannathan, et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

[KKM12] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *ALT*, volume 12, pages 199–213. Springer, 2012.

[KKM13] N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, pages 1448–1456, 2013.

[KM58] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[KNdPS17] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *NIPS*, pages 1674–1684, 2017.

[Lep90] O. V. Lepski. A problem of adaptive estimation in gaussian white noise. *Teoriya Veroyatnostei i ee Primeneniya*, 35(3):459–470, 1990.

[LGC16] A. Locatelli, M. Gutzeit, and A. Carpentier. An optimal algorithm for the Thresholding Bandit Problem. *ArXiv e-prints*, May 2016.

[LMC+17] S. W. Linderman, G. E. Mena, H. Cooper, L. Paninski, and J. P. Cunningham. Reparameterizing the Birkhoff polytope for variational permutation inference. *arXiv preprint arXiv:1710.09508*, 2017.

[LR85] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[LYLW16] Z. Liu, J.-a. Yang, H. Liu, and W. Wang. Transfer learning by sample selection bias correction and its application in communication specific emitter identification. *JCM*, 11:417–427, 2016.

[LYW+04] X. Li, Ying, W., J. Tuo, W. Li, and W. Liu. Applications of classification trees to consumer credit scoring methods in commercial banks. *Systems, Man and Cybernetics SMC*, 5:4112–4117, 2004.

[Mai13] O.-A. Maillard. Robust risk-averse stochastic multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 218–233. Springer, 2013.

[Mal57] C. L. Mallows. Non-null ranking models. *Biometrika*, 44(1-2):114–130, 1957.

[MSK+10] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

[MSK+12] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[MW16] A. K. Menon and R. C. Williamson. Bipartite ranking: a risk-theoretic perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

[NB16] G. Neu and G. Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *The Journal of Machine Learning Research*, 17(1):5355–5375, 2016.

[NLB16] R. Nishihara, D. Lopez-Paz, and L. Bottou. No regret bound for extreme bandits. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[NP87] W. K. Newey and J. L. Powell. Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, pages 819–847, 1987.

[NSS+19] S. Nagpal, M. Singh, R. Singh, M. Vatsa, and N. Ratha. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*, 2019.

[PC+19] G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[Pla75] R. L. Plackett. The analysis of permutations. *Applied Statistics*, 2(24):193–202, 1975.

[PMM+17] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade. Worst-case vs average-case design for estimation from fixed pairwise comparisons. *arXiv preprint arXiv:1707.06217*, 2017.

[PNZ+15] D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015.

[PR+13] V. Perchet, P. Rigollet, et al. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693–721, 2013.

[PS16] A. Procaccia and N. Shah. Optimal aggregation of uncertain preferences. In *AAAI*, pages 608–614, 2016.

[PY10] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[QMX18] C. Qu, S. Mannor, and H. Xu. Nonlinear distributional gradient temporal-difference learning. *arXiv preprint arXiv:1805.07732*, 2018.

[Qui86] J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):1–81, 1986.

[RA05] S. Rajaram and S. Agarwal. Generalization bounds for k-partite ranking. In *NIPS 2005 Workshop on Learn to rank*, 2005.

[Rac91] S. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.

[Rak04] A. Rakotomamonjy. Optimizing Area Under Roc Curve with SVMs. In *Proceedings of the First Workshop on ROC Analysis in AI*, 2004.

[RBD+18] M. Rowland, M. G. Bellemare, W. Dabney, R. Munos, and Y. W. Teh. An analysis of categorical distributional reinforcement learning. *arXiv preprint arXiv:1802.08163*, 2018.

[RDK+19] M. Rowland, R. Dadashi, S. Kumar, R. Munos, M. G. Bellemare, and W. Dabney. Statistics and samples in distributional reinforcement learning. *arXiv preprint arXiv:1902.08102*, 2019.

[RDS+14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[Res07] S. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Number vol. 10 in Heavy-tail Phenomena: Probabilistic and Statistical Modeling. Springer, 2007.

[RFGST12] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

[Ris05] M. Risse. Why the count de borda cannot beat the marquis de condorcet. *Social Choice and Welfare*, 25(1):95, 2005.

[RJ05] F. Radlinski and T. Joachims. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248, 2005.

[RN94] G. A. Rummery and M. Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.

[RSL17] P. Reverdy, V. Srivastava, and N. Leonard. Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control*, 62(8):3788–3803, 2017.

[RU+00] R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[Rud16] S. Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[SBFWH15] B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. 2015.

[SBGW15] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *arXiv preprint arXiv:1510.05610*, 2015.

[SBW16] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1153–1157. IEEE, 2016.

[SCV13] S. R. S. Clémençon and N. Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104, 2013.

[SDP12] Y. Song, S. Dixon, and M. Pearce. A survey of music recommendation systems and future perspectives. In *9th International Symposium on Computer Music Modeling and Retrieval*, volume 4, pages 395–410, 2012.

[SJLS00] S. Singh, T. Jaakkola, M. L. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308, 2000.

[Sli14] A. Slivkins. Contextual bandits with similarity information. *The Journal of Machine Learning Research*, 15(1):2533–2568, 2014.

[SLM12] A. Sani, A. Lazaric, and R. Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

[SNK+08] M. Sugiyama, S. Nakajima, H. Kashima, P. v. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, pages 1433–1440, 2008.

[SS06] M. J. Streeter and S. F. Smith. A simple distribution-free approach to the max k-armed bandit problem. In *International Conference on Principles and Practice of Constraint Programming*, pages 560–574. Springer, 2006.

[Sto09] A. Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

[Sut88] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[SW15] N. Shah and M. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *arXiv preprint arXiv:1512.08949*, 2015.

[TGP19] L. Torossian, A. Garivier, and V. Picheny. X-armed bandits: Optimizing quantiles, cvar and other risks. In *Asian Conference on Machine Learning*, pages 252–267, 2019.

[Tho33] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[Tho00] L. Thomas. A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16:149–172, 2000.

[UPGM09] A. Ukkonen, K. Puolamäki, A. Gionis, and H. Mannila. A randomized approximation algorithm for computing bucket orders. *Information Processing Letters*, 109(7):356–359, 2009.

[VACT20] R. Vogel, M. Achab, S. Clémençon, and C. Tillier. Weighted empirical risk minimization: Sample selection bias correction based on importance sampling, 2020.

[Vap00] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Lecture Notes in Statistics. Springer, 2000.

[VdV00] A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[VdVW] A. W. Van der Vaart and J. A. Wellner. Weak convergence and empirical processes. 1996.

[Vil08] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[VZ16] S. Vakili and Q. Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

[Was13] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[Wat89] C. J. C. H. Watkins. Learning from delayed rewards. 1989.

[WD92] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[Wes00] D. West. Neural network credit scoring models. *Computers and Operations Research*, 27:1131–1152, 2000.

[Woo79] M. Woodroofe. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806, 1979.

[Yan07] Y. Yang. Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research*, 183(3):1521–1536, 2007.

[ZIJC14] A. Zimin, R. Ibsen-Jensen, and K. Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

[ZWY+17] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

APPENDIX A

Résumé des contributions

1 Ordonnancement par minimisation du risque empirique

Nous commençons par rappeler brièvement le cadre de la *minimisation du risque empirique* où l'on entend minimiser une fonction de perte en espérance par rapport à une certaine *distribution de test* P, sur la base de l'observation de réalisations indépendantes de P. Ensuite, nous présentons nos contributions au problème de correction du biais de sélection d'échantillonnage, où les observations sont échantillonnées à partir d'une *distribution d'entraînement* P' différente de P.

Tout au long de la thèse, la notation $X \sim P$ signifie que P est la distribution de probabilité de la variable aléatoire X.

1.1 Minimisation du risque empirique

Le principal paradigme de l'apprentissage prédictif est la *minimisation du risque empirique* (MRE en abrégé), voir *e.g.* [DGL96]. Dans la configuration standard, Z est une variable aléatoire (v.a. en abrégé) qui prend ses valeurs dans un espace $\mathcal{Z}$ et de distribution P, Θ est un espace de paramètres et $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ est une fonction de perte (mesurable). Le *risque* est alors défini par : $\forall \theta \in \Theta$,

$$\mathcal{R}_P(\theta) = \mathbb{E}_P\left[\ell(\theta, Z)\right], \tag{A.1}$$

et plus généralement pour toute mesure Q sur $\mathcal{Z}$: $\mathcal{R}_Q(\theta) = \int_{\mathcal{Z}} \ell(\theta, z) dQ(z)$. Dans la plupart des situations pratiques, la distribution P impliquée dans la définition du risque est inconnue et l'apprentissage est basé sur la seule observation d'un échantillon indépendant et identiquement distribué (i.i.d.) $Z_1, \ldots, Z_n$ tiré de P et le risque (A.1) doit être remplacé par une contrepartie empirique, généralement :

$$\widehat{\mathcal{R}}_P(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i) = \mathcal{R}_{\widehat{P}_n}(\theta), \tag{A.2}$$

où $\widehat{P}_n = (1/n)\sum_{i=1}^n \delta_{Z_i}$ est la mesure empirique de P et δ_z désigne la mesure de Dirac en un point z. Tout minimiseur empirique $\widehat{\theta}_n \in \arg\min_{\theta\in\Theta} \widehat{\mathcal{R}}_P(\theta)$ est alors accessible à

l'apprenant — en fait, cela n'est pas nécessairement vrai en pratique car la minimisation du risque empirique peut parfois s'avérer être difficilement réalisable — et peut être utilisé comme substitut aux paramètres optimaux (inconnus) $\theta^* \in \arg\min_{\theta\in\Theta} \mathcal{R}_P(\theta)$ vivant dans la classe d'hypothèse Θ.

Example 1. (Classification binaire) *Dans la* classification binaire*, le problème phare de l'apprentissage machine (voir par exemple [DGL96]), le but est de trouver un* classifieur $g : \mathcal{X} \to \{-1, +1\}$ *avec un risque de classification* $\mathcal{R}_P(g) = \mathbb{P}\{g(X) \neq Y\}$ *aussi faible que possible. La paire aléatoire* $Z = (X, Y)$*, avec distribution* P*, est à valeur dans* $\mathcal{Z} = \mathcal{X} \times \{-1, +1\}$*, et l'*espace des caractéristiques $\mathcal{X}$ *est typiquement un sous-ensemble de* $\mathbb{R}^d$ *(*$d \geq 1$*) ;* X *est appelé le* vecteur de caractéristique *et* Y *est l'*étiquette*. Désignant la* probabilité postérieure *par* $\eta(x) = \mathbb{P}\{Y = +1|X = x\}$ *pour tout* $x \in \mathcal{X}$*, le classifieur de Bayes* $g^*(x) = 2\mathbb{I}\{\eta(x) > \frac{1}{2}\} - 1$ *est la règle de classification optimale car elle minimise* $\mathcal{R}_P$ *: pour tout classifieur* g*,* $\mathcal{R}_P(g) \geq \mathcal{R}_P(g^*)$ *(voir le Théorème 2.1 dans [DGL96]). Empiriquement, l'apprenant reçoit un* ensemble de données d'entraînement *composé de* n *copies i.i.d.* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *de* (X, Y)*. En utilisant les notations introduites ci-dessus,*

- *l'espace des paramètres* Θ *est un ensemble* $\mathcal{G}$ *de classifieurs* g*,*
- *la fonction de perte* ℓ *est la fonction de perte zéro-un* $\ell_{0/1}$ *:*

$$\forall (g, x, y) \in \mathcal{G} \times \mathcal{X} \times \{-1, +1\}, \quad \ell_{0/1}(g, (x, y)) = \mathbb{I}\{g(x) \neq y\}.$$

La classification binaire appartient à la famille des problèmes d'*apprentissage supervisé* car elle tente d'apprendre comment étiqueter toute nouvelle observation non étiquetée X, sur la base d'exemples étiquetés (X_i, Y_i) fournis par un "professeur". En revanche, la tâche de *partitionnement de données* est un problème *d'apprentissage non supervisé*. En effet, elle consiste à trouver des groupes de similitudes dans l'espace des caractéristiques sans aucune information de type étiquette.

Example 2. (Partitionnement en k-moyennes) *Étant donné un certain nombre de groupes* $k \geq 1$*, la méthode de* partitionnement en k-moyennes *(voir par exemple [Bis06] ou [HTF09]) résout le problème de minimisation suivant :*

$$\min_{(m_1,\ldots,m_k)\in\mathcal{Z}^k} \sum_{i=1}^{n} \min_{1\leq j\leq k} ||Z_i - m_j||_2^2, \tag{A.3}$$

avec $\mathcal{Z} \subseteq \mathbb{R}^d$ *et* $||\cdot||_2$ *la norme euclidienne. Cette quantité mesure la distance euclidienne totale au carré de chaque observation* Z_i *par rapport au centre de son groupe, à savoir le point le plus proche* $m_{j(Z_i)}$*, avec* $j(Z_i) \in arg\ min_{1\leq j\leq k} ||Z_i - m_j||_2$*, parmi les* k *centroïdes* $m_1, \ldots, m_k$ *(en fait ici,* $m_{j(Z_i)}$ *n'est pas nécessairement unique). Avec nos notations de MRE, nous avons que :*

- *l'espace des paramètres est l'ensemble des* k*-uplets* $(m_1, \ldots, m_k)$ *:* $\Theta = \mathcal{Z}^k$*,*

- *la fonction de perte du partitionnement en k-moyennes est :*

$$\forall((m_1,\dots,m_k),z)\in\mathcal{Z}^k\times\mathcal{Z},\quad \ell((m_1,\dots,m_k),z)=\min_{1\le j\le k}||z-m_j||_2^2.$$

En particulier, le critère du partitionnement en k-moyennes dans l'Eq. (I.3) est, à normalisation près, égal au risque empirique $\widehat{\mathcal{R}}_P$:

$$\widehat{\mathcal{R}}_P((m_1,\dots,m_k))=\frac{1}{n}\sum_{i=1}^{n}\min_{1\le j\le k}||Z_i-m_j||_2^2.$$

La performance des minimiseurs $\widehat{\theta}_n$ de (A.2) peut être étudiée en contrôlant l'*excédent de risque* $\mathcal{R}(\widehat{\theta}_n)-\min_{\theta\in\Theta}\mathcal{R}(\theta)$, qui satisfait à l'inégalité élémentaire (voir par exemple [BBL05])

$$\begin{aligned}\mathcal{R}_P(\widehat{\theta}_n)-\min_{\theta\in\Theta}\mathcal{R}_P(\theta)&=\mathcal{R}_P(\widehat{\theta}_n)-\widehat{\mathcal{R}}_P(\widehat{\theta}_n)+\widehat{\mathcal{R}}_P(\widehat{\theta}_n)-\mathcal{R}_P(\theta^*)\\&\le\mathcal{R}_P(\widehat{\theta}_n)-\widehat{\mathcal{R}}_P(\widehat{\theta}_n)+\widehat{\mathcal{R}}_P(\theta^*)-\mathcal{R}_P(\theta^*)\le 2\sup_{\theta\in\Theta}|\widehat{\mathcal{R}}_P(\theta)-\mathcal{R}_P(\theta)|.\end{aligned}\quad(A.4)$$

Les fluctuations des écarts maximaux $\sup_{\theta\in\Theta}|\widehat{\mathcal{R}}_P(\theta)-\mathcal{R}_P(\theta)|$ dans l'Eq. (A.4) peuvent ensuite être quantifiées au moyen d'*inégalités de concentration*, sous diverses hypothèses de complexité pour la classe fonctionnelle $\mathcal{F}=\{\ell(\theta,\,\cdot):\ \theta\in\Theta\}$ (*e.g.* dimension VC, entropies métriques, moyennes de Rademacher), voir [BLM13] par exemple.

Parfois, l'échantillon d'entraînement est tiré d'une distribution d'entraînement P' différente de la distribution cible P d'intérêt : c'est ce qu'on appelle un *biais de sélection d'échantillonnage*. Notre approche pour faire face à cette situation est appelée MREP pour "minimisation du risque empirique pondéré", elle repose sur une étape de repondération par l'estimation de poids d'échantillonnage préférentiel pour chaque observation de l'ensemble de données d'entraînement.

1.2 Correction du biais de sélection d'échantillonnage

Les problèmes de sélection biaisée dans l'apprentissage machine, qui résultent souvent d'erreurs lors du processus d'acquisition des données, font désormais l'objet d'une grande attention dans la littérature, voir [BCZ+16], [ZWY+17], [BHS+19], [LYLW16] ou [HGB+07]. Nous considérons le cas où l'échantillon i.i.d. $Z'_1,\ \dots,\ Z'_n$ disponible pour la phase d'entraînement n'est pas tiré de P mais d'une autre distribution P', par rapport à laquelle P est absolument continue. Le but poursuivi est de poser les bases théoriques de l'application des idées qui sous-tendent la méthodologie de l'échantillonnage préférentiel (EP en bref) pour étendre l'approche MRE à ce dispositif d'apprentissage. Les méthodes d'EP sont largement utilisées dans l'apprentissage machine, y compris dans des contextes d'apprentissage en ligne tels que les problèmes de bandits (voir [NB16]), qui sont présentés dans la partie 2. Nous soulignons que le problème à l'étude est un cas très particulier d'*apprentissage par transfert* (voir *e.g.* [PY10], [BDBC+10] et [Sto09]), un domaine de recherche qui fait actuellement l'objet d'une grande attention dans la littérature.

La figure I.1 illustre un exemple de ce type de biais de sélection d'échantillonnage dans un contexte de classification : l'ensemble de données d'entraînement est composé d'images de quatre types d'animaux (chien, loup, tigre et singe), alors que la population cible est simplement un mélange de chiens et de loups. En d'autres termes, les étiquettes d'entraînement Y' sont à valeurs dans $\mathcal{Y} = \{\text{chien,loup,tigre,singe}\}$, alors que, pour la distribution test/cible, Y prend ses valeurs seulement dans le sous-ensemble $\{\text{chien}, \text{loup}\} \subset \mathcal{Y}$. Les niveaux de l'histogramme représentent les probabilités de classe : $\mathbb{P}\{Y' = y\}$ en bleu pour l'entraînement, et $\mathbb{P}\{Y = y\}$ en vert pour le test, pour chaque animal $y \in \mathcal{Y}$. Nous formulons ci-dessous la méthode MREP pour traiter ces questions de biais de sélection d'échantillonnage.

MRE Pondéré (MREP). La *minimisation du risque empirique pondéré* (MREP) que nous proposons au chapitre II consiste à minimiser une version pondérée du risque empirique. Nous étudions les conditions garantissant que les valeurs du paramètre θ qui minimisent presque (A.1) peuvent être obtenues par la minimisation d'une version pondérée du risque empirique basé sur les Z'_i, à savoir

$$\widetilde{\mathcal{R}}_{w,n}(\theta) = \mathcal{R}_{\widetilde{P}_{w,n}}(\theta), \tag{A.5}$$

où $\widetilde{P}_{w,n} = (1/n)\sum_{i=1}^{n} w_i \delta_{Z'_i}$ et $w = (w_1, \ldots, w_n) \in \mathbb{R}_+^n$ est un certain vecteur de poids. Les poids idéaux w^* sont donnés par la fonction de vraisemblance $\Phi(z) = (dP/dP')(z)$: $w_i^* = \Phi(Z'_i)$ pour $i \in \{1, \ldots, n\}$. Dans ce cas, la quantité (A.5) est un estimateur non biaisé du risque réel (A.1) :

$$\mathbb{E}_{P'}\left[\mathcal{R}_{\widetilde{P}_{w^*,n}}(\theta)\right] = \mathcal{R}_P(\theta), \tag{A.6}$$

et des bornes de généralisation pour l'excès de risque $\mathcal{R}_P$ des minimiseurs de $\widetilde{\mathcal{R}}_{w^*,n}$ peuvent être directement établies en étudiant les propriétés de concentration du processus empirique lié aux Z'_i et à la classe de fonctions $\{\Phi(\cdot)\ell(\theta, \cdot) : \theta \in \Theta\}$. Cependant, la *fonction d'importance* Φ est inconnue en général, tout comme la distribution P.

Dans la figure I.1, qui correspond à un problème de classification où l'ensemble de données d'entraînement $Z'_1 = (X'_1, Y'_1), \ldots, Z'_n = (X'_n, Y'_n)$, avec X'_i une image (vecteur de pixels dans $[0,1]^d$ par exemple) d'un animal de type $Y'_i \in \mathcal{Y} = \{\text{chien}, \text{loup}, \text{tigre}, \text{singe}\}$, la fonction de vraisemblance est donnée par :

$$\forall (x,y) \in [0,1]^d \times \mathcal{Y}, \quad \Phi((x,y)) = \frac{80\%}{10\%}\mathbb{I}\{y = \text{chien}\} + \frac{20\%}{40\%}\mathbb{I}\{y = \text{loup}\},$$

qui ne dépend que de y, si l'on suppose que la distribution conditionnelle de $Z = (X, Y) \sim P$ sachant $Y = y$ est la même que celle de $Z' = (X', Y') \sim P'$ sachant $Y' = y$, pour tout $y \in \mathcal{Y}$ (i.e. P et P' sont deux mélanges des mêmes 4 composantes animales mais avec des poids différents).

Contributions. Notre principale contribution à ce problème est de montrer que, dans des situations loin d'être rares en pratique, les poids (idéaux) w_i^* peuvent être estimés à partir des Z'_i combinés à des informations auxiliaires sur la population cible P. Ces cas favorables comprennent notamment :

- les problèmes de classification lorsque les probabilités de classe dans l'étape de test diffèrent de celles de l'étape d'entraînement (comme dans la figure I.1),

- la minimisation de risque dans des populations stratifiées (voir [BD18]), les strates étant représentées statistiquement de manière différente dans les populations de test et d'entraînement,

- l'apprentissage "positive-unlabeled" (voir *e.g.* [dPNS14]), qui consiste à résoudre un problème de classification binaire basé uniquement sur des données positives et non étiquetées.

Dans chacun de ces cas, nous montrons que le processus stochastique obtenu en utilisant les estimateurs des poids dans la fonction de risque empirique pondéré (A.5) est beaucoup plus complexe qu'un simple processus empirique (*i.e.* une collection de moyennes i.i.d.) mais peut cependant être étudié au moyen de *techniques de linéarisation*, dans l'esprit des extensions de la MRE établies dans [CLV08] ou [CV09a]. Des bornes de convergence pour les minimiseurs de l'estimateur du risque correspondant sont prouvées et, au-delà de ces garanties théoriques, la performance de l'approche MREP est soutenue par des résultats numériques convaincants.

1.3 Ordonnancement à partir de données étiquetées de façon binaire

Nous introduisons maintenant un autre problème de MRE, de nature *globale* contrairement à la classification binaire : le problème d'*ordonnancement à partir de données étiquetées de façon binaire*, que nous appelerons aussi *ordonnancement bipartite*, (voir [AGH$^+$05],[FISS03]), où l'on veut ordonner, au moyen de méthodes de scoring, tous les éléments de l'espace caractéristique $\mathcal{X}$, à partir de l'observation d'un ensemble de données d'entraînement composé de copies i.i.d. d'une paire aléatoire (X, Y) à valeur dans $\mathcal{X} \times \{-1, +1\}$. Intuitivement, les bonnes règles de scoring sont les fonctions $s : \mathcal{X} \to \mathbb{R}$ attribuant des scores importants $s(x)$ aux éléments $x \in \mathcal{X}$ avec une grande probabilité postérieure $\mathbb{P}\{Y = +1 | X = x\}$. L'ordonnancement trouve de nombreuses applications pratiques (voir par exemple [CDV13b]), allant d'études médicales, où les patients sont ordonnés en fonction de leur probabilité d'être malades, aux systèmes de recommandation qui classent un catalogue de produits en fonction des préférences de certains utilisateurs. Voir par exemple [BK07] à propos de certaines méthodes de *recommandation de films* utilisées lors de la compétition du "Netflix Prize". Une autre application de l'ordonnancement est la gestion du *risque de crédit*, qui servira également de motivation à notre problème de *bandits manchots rentables* exposé au chapitre V dans un contexte d'apprentissage en ligne. Nous rappelons ci-dessous quelques notions importantes pour l'ordonnancement avant d'introduire nos contributions au problème d'*ordonnancement avec étiquettes continues*, une généralisation au cas où les étiquettes Y prennent des valeurs continues.

Cadre Formel. Le cadre probabiliste de l'ordonnancement bipartite est le même que celui de la classification binaire (voir exemple 1). En effet, nous considérons une variable

aléatoire $(X, Y) \sim P$ à valeur dans $\mathcal{X} \times \{-1, +1\}$, avec un espace de caractéristiques $\mathcal{X} \subseteq \mathbb{R}^d$ $(d \geq 1)$, et avec la distribution P caractérisée par la paire (μ, η), où

- la *distribution marginale* de X est μ,
- la *probabilité postérieure* pour tout $x \in \mathcal{X}$ est

$$\eta(x) = \mathbb{P}\{Y = +1 | X = x\} = \frac{1}{2}(\mathbb{E}[Y|X = x] + 1).$$

De même, P peut être décrite par le triplet (p, G, H), où

- $p = \mathbb{P}\{Y = +1\}$ est la probabilité d'occurrence d'un cas positif,
- G et H sont respectivement les distributions conditionnelles de X sachant $Y = +1$ et de X sachant $Y = -1$.

Le problème empirique qui nous intéresse est le suivant : étant donné un échantillon $(X_1, Y_1), \ldots, (X_n, Y_n)$ de $n \geq 1$ copies i.i.d. de (X, Y), un agent apprenant souhaite sélectionner une règle de notation, c'est-à-dire une fonction mesurable $s : \mathcal{X} \to \mathbb{R}$, capable d'ordonner tout nouvel échantillon non étiqueté $X'_1, \ldots, X'_{n'}$ (de distribution commune μ) tel que, avec grande probabilité, les observations X'_t avec des scores élevés $s(X'_t)$ ont des étiquettes positives $Y'_t = +1$ plus souvent que les observations avec de plus petits scores. Ce problème se retrouve dans de nombreuses applications, parmi lesquelles les systèmes de recommandation de musique (voir [SDP12]).

Example 3. (Recommandation Musicale) *L'ordonnancement bipartite peut être utilisé pour construire un système de recommandation musicale. Considérons une collection $\mathcal{X}$ de chansons, chaque chanson $x \in \mathcal{X}$ étant modélisée par les d coordonnées $x = (x_1, \ldots, x_d)$: x_1 =“titre”, x_2 =“artiste”, x_3 =“durée”, etc. Un utilisateur d'une plateforme musicale produit un ensemble de données d'entraînement composé de n paires chanson-évaluation $(X_1, Y_1), \ldots, (X_n, Y_n)$ avec Y_i une note binaire donnée par l'utilisateur et égale à $+1$ s'il a apprécié la chanson $X_i \in \mathcal{X}$, ou bien $Y_i = -1$ s'il ne l'a pas aimée. Sur la base de ces informations partielles, la plateforme musicale veut prédire les préférences de l'utilisateur sur l'ensemble du catalogue de chansons $\mathcal{X}$, en donnant à chaque chanson $x \in \mathcal{X}$ un score $s_{user}(x)$. Dès lors, une bonne fonction de score s_{user} donne des scores élevés aux chansons que l'utilisateur est susceptible d'apprécier.*

Formellement, pour deux variables aléatoires à valeurs réelles U et U', nous rappelons que U est *stochastiquement supérieure* à U' si $\mathbb{P}\{U \geq t\} \geq \mathbb{P}\{U' \geq t\}$ pour tout $t \in \mathbb{R}$. Dès lors, le but est d'apprendre une fonction de score s telle que la v.a. conditionnelle $s(X)$ sachant $Y = +1$ soit la plus stochastiquement supérieure à $s(X)$ sachant $Y = -1$ que possible. En d'autres termes, nous voulons que s maximise la différence entre $1 - G_s(t) = \bar{G}_s(t) = \mathbb{P}\{s(X) \geq t | Y = +1\}$ et $1 - H_s(t) = \bar{H}_s(t) = \mathbb{P}\{s(X) \geq t | Y = -1\}$ pour tous les niveaux $t \in \mathbb{R}$. Ce critère fonctionnel peut également être exprimé au

moyen de la courbe ROC de toute règle de notation s, c'est-à-dire la courbe paramétrée $t \in \mathbb{R} \mapsto (\bar{H}_s(t), \bar{G}_s(t))$, ou de façon équivalente le graphique de la fonction

$$\alpha \in (0,1) \mapsto \mathrm{ROC}_s(\alpha) = \bar{G}_s \circ (1 - H_s^{-1})(1-\alpha),$$

où les éventuels points de discontinuité sont reliés par des segments. En effet, les éléments optimaux s^* sont ceux dont la courbe ROC $\mathrm{ROC}_{s^*} = \mathrm{ROC}^*$ domine toute autre courbe ROC ROC_s partout :

$$\forall \alpha \in (0,1), \quad \mathrm{ROC}^*(\alpha) \geq \mathrm{ROC}_s(\alpha).$$

Voir la figure I.2 pour un exemple. Il est bien connu que les fonctions de score optimales s^* sont les transformées strictement croissantes de la fonction de probabilité postérieure η (voir par exemple [CLV05]). Compte tenu de sa nature fonctionnelle, la courbe ROC ROC_s est souvent résumée par une quantité scalaire plus facile à manier, à savoir son aire appelée *Aire Sous la Courbe ROC* (AUC en bref) :

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X) < s(X') | Y = -1, Y' = +1\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X') | Y = -1, Y = +1\},$$

où (X', Y') est une copie i.i.d. de (X, Y). En effet, lorsque les courbes ROC de deux fonctions de score s_1 et s_2 se croisent comme dans la figure I.2, aucune des deux ne peut être considérée comme meilleure que l'autre, ou même égale, du point de vue du critère ROC. Au contraire, un critère scalaire global tel que l'AUC, permet toujours de comparer deux règles de score. Il est intéressant de noter que l'AUC s'accompagne d'une interprétation probabiliste : c'est le taux théorique de paires concordantes. L'approche MRE habituelle pour l'ordonnancement bipartite consiste à maximiser la version empirique de l'AUC, étant donné un échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$:

$$\widehat{\mathrm{AUC}}_n(s) = \frac{1}{n_+ \cdot n_-} \sum_{i:Y_i=-1} \sum_{j:Y_j=+1} \mathbb{I}\{s(X_i) < s(X_j)\} + \frac{1}{2}\mathbb{I}\{s(X_i) = s(X_j)\}, \qquad \text{(A.7)}$$

avec $n_+ = \sum_{i=1}^n \mathbb{I}\{Y_i = +1\} = n - n_-$. Remarquez que l'AUC empirique dans l'Eq. (A.7) est une somme de variables dépendantes : plus précisément, c'est une U-statistique de degré 2 (voir [CLV08]). Plusieurs algorithmes basés sur la maximisation de $\widehat{\mathrm{AUC}}_n$ ont été proposés et étudiés dans la littérature, tels que l'approche TreeRank ([CV09b]). Une extension au cas où l'étiquette Y prend au moins trois valeurs ordinales, appelée *ordonnancement multi-partite*, a également été étudiée ([RA05], [SCV13]) : nous présentons dans ce qui suit notre contribution au problème plus général de l'*ordonnancement continu*, où Y prend ses valeurs dans tout l'intervalle $[0, 1]$.

1.4 Ordonnancement à partir de données étiquetées de façon continue

Dans le chapitre III, nous considérons une tâche d'ordonnancement similaire à l'ordonnancement bipartite, la différence résidant dans la nature de l'étiquette Y, dont le support s'étend sur un continuum de valeurs scalaires : nous appelons ce problème *ordonnancement*

continu. Selon le contexte, Y peut représenter une taille, une mesure biologique, ou le cash-flow des entreprises en finance quantitative. Nous décrivons ci-dessous une application potentielle de l'ordonnancement continu pour la recommandation musicale, adaptée de l'exemple 3 (ordonnancement bipartite).

Example 4. (Recommandation Musicale Avancée)
Comme dans l'exemple 3, une plateforme musicale veut recommander intelligemment les chansons de la playlist $\mathcal{X}$ à un utilisateur donné, au moyen d'une règle de notation $s_{user} : \mathcal{X} \to \mathbb{R}$ propre à cet utilisateur. Ici aussi, l'utilisateur génère un ensemble de données d'entraînement $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ après avoir écouté n chansons $X_i \in \mathcal{X}$. Néanmoins, chaque étiquette Y_i correspond maintenant à la quantité de dopamine (alias la "molécule du plaisir") libérée par le cerveau de l'utilisateur — et mesurée par un capteur — pendant l'écoute de la i-ème chanson. Par conséquent, ces étiquettes Y_i ne sont pas des évaluations binaires comme dans l'ordonnancement bipartite, mais prennent plutôt des valeurs continues. Néanmoins, l'objectif du système de recommandation reste le même : donner des scores importantes $s_{user}(x)$ aux chansons $x \in \mathcal{X}$ qui sont susceptibles de libérer beaucoup de dopamine dans le cerveau de l'utilisateur.
• *Un exemple plus réaliste s'appuie sur les* évaluations implicites*, en particulier l'action de l'utilisateur "sauter la chanson en cours", qui ont reçu beaucoup d'attention dans la littérature récemment ([RFGST12],[RJ05],[JGP+17]). Dans ce cas, une étiquette continue Y_i est définie par :*

$$Y_i = \frac{\textit{temps d'écoute de la chanson } X_i \textit{ avant saut}}{\textit{durée totale de la chanson } X_i} \in [0, 1],$$

qui s'interprète implicitement comme une évaluation négative lorsqu'elle est proche de zéro.

Formellement, nous supposons que la paire aléatoire (X, Y) admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^{d+1}$, et que le support de Y est compact, égal à $[0, 1]$ pour simplifier. La *fonction de régression* est désignée par

$$m : x \in \mathcal{X} \mapsto \mathbb{E}[Y|X = x].$$

Nous formulons le problème de l'ordonnancement continu comme un continuum de problèmes d'ordonnancement bipartite imbriqués. En effet, pour toute valeur seuil $y \in (0, 1)$, le sous-problème d'ordonnancement bipartite lié à la paire (X, Z_y) avec $Z_y = 2\mathbb{I}\{Y > y\} - 1$ peut être considéré comme une approximation discrète du problème complet : nous désignons respectivement par $\mathrm{ROC}_{s,y}$ et $\mathrm{AUC}_{s,y}$ la courbe ROC correspondante et l'AUC de toute fonction de score mesurable $s : \mathcal{X} \to \mathbb{R}$. En d'autres termes, nous voulons résoudre simultanément tous ces sous-problèmes, c'est-à-dire identifier une règle de score s maximisant $\mathrm{ROC}_{s,y}$ et $\mathrm{AUC}_{s,y}$ pour tout $y \in (0, 1)$.
Contributions. Dans ce but, nous introduisons de nouvelles mesures de performance obtenues en intégrant $\mathrm{ROC}_{s,y}$ et $\mathrm{AUC}_{s,y}$ par rapport à la distribution marginale F_Y de Y :

$$\forall \alpha \in (0, 1), \quad \mathrm{IROC}_s(\alpha) = \int_{y=0}^{1} \mathrm{ROC}_{s,y}(\alpha) F_Y(dy) \quad \text{et} \quad \mathrm{IAUC}(s) = \int_{\alpha=0}^{1} \mathrm{IROC}_s(\alpha) d\alpha.$$

Notre analyse théorique est double. Nous montrons que :

(i) dans certaines conditions, les règles de notation optimales sont des transformées strictement croissantes de la fonction de régression m,

(ii) une règle de notation s^* est optimale si et seulement si sa courbe IROC domine uniformément n'importe quelle autre courbe IROC IROC_s :

$$\forall \alpha \in (0,1), \quad \mathrm{IROC}_{s^*}(\alpha) = \mathbb{E}[\mathrm{ROC}^*_Y(\alpha)] \geq \mathrm{IROC}_s(\alpha),$$

et son IAUC est maximale :

$$\mathrm{IAUC}(s^*) = \mathbb{E}[\mathrm{AUC}^*_Y] \geq \mathrm{IAUC}(s) \quad \text{pour tout } s.$$

En outre, nous fournissons une expression probabiliste de l'IAUC :

$$\mathrm{IAUC}(s) = \mathbb{P}\{s(X) < s(X') | Y < Y'' < Y'\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X') | Y < Y'' < Y'\},$$

où (X', Y') est une copie i.i.d. de (X, Y) et Y'' est échantillonné indépendamment de la distribution marginale F_Y de Y. A partir de cette formule, nous estimons empiriquement l'IAUC(s) à partir d'un échantillon i.i.d. $(X_1, Y_1), \ldots, (X_n, Y_n)$ comme suit :

$$\widehat{\mathrm{IAUC}}_n(s) = \frac{6}{n(n-1)(n-2)} \sum_{1 \leq i,j,k \leq n} \mathbb{I}\left\{s(X_i) < s(X_k), Y_i < Y_j < Y_k\right\} + \frac{3}{n(n-1)(n-2)} \sum_{1 \leq i,j,k \leq n} \mathbb{I}\left\{s(X_i) = s(X_k), Y_i < Y_j < Y_k\right\},$$

qui est une U-statistique de degré 3. Enfin, nous fournissons un algorithme hiérarchique, CRank, visant à maximiser $\widehat{\mathrm{IAUC}}_n$: il renvoie une règle de notation constante par morceaux obtenue en divisant récursivement l'espace des caractéristiques.

Dans la section suivante, nous nous concentrons sur une autre tâche de classement, à savoir *l'agrégation de classements*, visant à ordonner un nombre fini d'éléments, à partir de classements (complets ou incomplets) constituant les données d'entraînement.

1.5 Agrégation de classements par minimisation du risque empirique

Le problème d'ordonnancement bipartite (resp. continu) présenté précédemment consistait à produire une fonction de notation, et par conséquent un ordre sur un espace de caractéristiques $\mathcal{X}$, à partir d'observations (vectorielles) de la forme $(X_1, Y_1), \ldots, (X_n, Y_n)$ à valeurs dans $\mathcal{X} \times \{-1, +1\}$ (resp. dans $\mathcal{X} \times [0, 1]$). Dans le problème d'*agrégation de classements*, bien que l'objectif soit toujours d'établir un ordre, il existe deux différences principales avec les problèmes précédents :

- l'ensemble des éléments à classer est de cardinalité finie N, contrairement aux espaces de caractéristiques infinis $\mathcal{X} \subseteq \mathbb{R}^d$ souvent considérés dans l'ordonnancement bipartite/continu,

- les données d'entrée elles-mêmes sont des classements/comparaisons, c'est-à-dire des informations *relatives* contrairement aux étiquettes représentant des évaluations *absolues*.

Repères Historiques. L'analyse des données de classement date du 18-ème siècle avec la conception d'un système d'élection pour l'Académie des Sciences française. Différents systèmes de vote ont été proposés, chacun satisfaisant à certaines propriétés souhaitables : en particulier, la *méthode de Borda* en 1781 ([Bor84]) et son concurrent, la *méthode de Condorcet* en 1785 ([DC$^+$14]), ont créé le fameux *débat Borda-Condorcet*. Plus tard, en 1951, Arrow a prouvé un "théorème d'impossibilité" ([Arr12]) selon lequel aucune règle électorale ne peut satisfaire simultanément un ensemble de propriétés raisonnables ; les systèmes de vote sont également étudiés dans la théorie du choix social (voir [Ris05]). Nous nous concentrons ci-dessous sur un problème spécifique qui se pose dans l'analyse des données de classement, à savoir celui de la synthèse d'un ensemble de classements par une seule permutation.

Agrégation de Classements. Étant donné une liste de $N \geq 2$ éléments indexés par $[\![N]\!] = \{1, \ldots, N\}$ et $n \geq 1$ permutations $\sigma_1, \ldots, \sigma_n$ dans le *groupe symétrique* $\mathfrak{S}_N$ de l'ensemble $[\![N]\!]$, le problème d'agrégation de classements consiste à identifier une seule permutation "consensus" $\widehat{\sigma}_n$ qui résume au mieux les σ_t. Dans de nombreuses applications utilisant des systèmes de vote (par exemple les systèmes de recommandation), chaque classement σ_t est obtenu en demandant à un agent d'ordonner les N éléments par préférence. Ainsi, le consensus $\widehat{\sigma}_n$ peut être considéré comme la permutation maximisant l'accord simultané des n agents ou, de manière équivalente, minimisant leur désaccord. Étant donné une permutation $\sigma \in \mathfrak{S}_N$ et deux éléments distincts $(i, j) \in [\![N]\!]^2$, nous utiliserons la notation $i \prec j$ signifiant que i est préféré à j, c'est-à-dire que i est classé avant j dans le classement σ : $\sigma(i) < \sigma(j)$. Un jeu de données composé de $n = 4$ classements de $N = 6$ éléments est représenté dans la figure I.4. Si plusieurs méthodes ont été développées pour résoudre ce problème, nous nous concentrons ici sur l'approche du consensus de Kemeny.

Consensus de Kemeny. La méthode du *consensus de Kemeny* ([Kem]) définit le classement consensus $\widehat{\sigma}_n$ comme un minimiseur de la somme des distances $d(\widehat{\sigma}_n, \sigma_t)$ aux n permuations σ_t :

$$\widehat{\sigma}_n \in \operatorname*{arg\,min}_{\sigma \in \mathfrak{S}_N} \sum_{t=1}^{n} d(\sigma, \sigma_t),$$

où d est une métrique sur l'ensemble des permutations $\mathfrak{S}_N$. Plus précisément, la règle de Kemeny est basée sur le choix $d = d_\tau$ avec la distance d_τ de Kendall définie par : pour tous $(\sigma, \sigma') \in \mathfrak{S}_N^2$,

$$d_\tau(\sigma, \sigma') = \sum_{1 \leq i < j \leq N} \mathbb{I}\{(\sigma(i) - \sigma(j))(\sigma'(i) - \sigma'(j)) < 0\},$$

qui est le nombre de désaccords entre paires entre σ et σ'. Nous soulignons que calculer le consensus de Kemeny $\widehat{\sigma}_n$ est difficile en pratique ([DKNS01]) : nous nous

référons à [AM12] pour une discussion sur les algorithmes tractables capables d'approcher raisonnablement $\widehat{\sigma}_n$.

Cadre d'Apprentissage Statistique. Dans [KCS17], l'agrégation de classements est formulée comme un problème d'apprentissage statistique : les permutations déterministes σ_t sont remplacées par des variables aléatoires i.i.d. Σ_t avec distribution P sur $\mathfrak{S}_N$. Dans ce cadre probabiliste, le but ultime est d'identifier un véritable classement médian σ^* de P caractérisé par

$$\sigma^* \in \arg\min_{\sigma \in \mathfrak{S}_N} L_P(\sigma),$$

où $L_P(\sigma) = \mathbb{E}_{\Sigma \sim P}[d(\sigma, \Sigma)]$. Néanmoins, la distribution P étant inconnue de l'apprenant n'ayant accès qu'à un échantillon $\Sigma_1, \ldots, \Sigma_n$, le risque L_P et donc le classement médian σ^* ne peuvent être directement calculés. Cependant, en suivant le paradigme de MRE introduit précédemment, le consensus empirique

$$\widehat{\sigma}_n \in \arg\min_{\sigma \in \mathfrak{S}_N} L_{\widehat{P}_n}(\sigma) = \frac{1}{n} \sum_{t=1}^{n} d(\sigma, \Sigma_t),$$

où $\widehat{P}_n = (1/n) \sum_{t=1}^{n} \delta_{\Sigma_t}$ indique la distribution empirique, apparaît comme l'alternative naturelle à σ^*. En particulier, [KCS17] a établi des bornes minimax d'ordre $O_{\mathbb{P}}(1/\sqrt{n})$ pour l'*excédent de risque*

$$L_P(\widehat{\sigma}_n) - L_P(\sigma^*).$$

Ainsi, lorsque le nombre d'observations n augmente à l'infini, la performance de la solution empirique $\widehat{\sigma}_n$ converge vers le risque minimal $L_P(\sigma^*) = \min_{\sigma \in \mathfrak{S}_N} L_P(\sigma)$. En outre, en indiquant par

$$p_{i,j} = \mathbb{P}_{\Sigma \sim P}\{\Sigma(i) < \Sigma(j)\}$$

la probabilité que l'élément $i \in [\![N]\!]$ soit préféré à $j \in [\![N]\!] \setminus \{i\}$, les auteurs ont montré dans le cas du τ de Kendall $d = d_\tau$ que si la distribution P satisfait l'hypothèse suivante de *stricte transitivité stochastique faible* : pour tous $1 \leq i \neq j \leq N$, $p_{i,j} \neq \frac{1}{2}$ et

$$\forall k \in [\![N]\!] \setminus \{i, j\}, \quad \min(p_{i,j}, p_{j,k}) > \frac{1}{2} \Rightarrow p_{i,k} > \frac{1}{2},$$

alors la médiane de Kemeny σ^* est unique et égale au classement de Copeland σ_{Cop} (voir Théorème 5 dans [KCS17]) :

$$\sigma_{\mathrm{Cop}}(i) = 1 + \sum_{j \neq i} \mathbb{I}\Big\{p_{i,j} < \frac{1}{2}\Big\}, \quad \forall i \in [\![N]\!].$$

Empiriquement, un résultat similaire est également valable : avec une probabilité écrasante, le consensus $\widehat{\sigma}_n$ est égal au classement plug-in de Copeland $\widehat{\sigma}_{\mathrm{Cop}}$ défini pour chaque élément $i \in [\![N]\!]$ par

$$\widehat{\sigma}_{\mathrm{Cop}}(i) = 1 + \sum_{j \neq i} \mathbb{I}\Big\{\widehat{p}_{i,j} < \frac{1}{2}\Big\},$$

avec des probabilités par paire empiriques

$$\widehat{p}_{i,j} = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}\{\Sigma_t(i) < \Sigma_t(j)\}, \quad \forall 1 \leq i \neq j \leq N.$$

Il s'ensuit que ce cas spécifique du problème d'agrégation de classements peut être résolu efficacement en se basant uniquement sur les *comparaisons par paires* $\mathbb{I}\{\Sigma_t(i) < \Sigma_t(j)\}$, qui sont un cas particulier de *classements incomplets*. En d'autres termes, on peut éviter d'observer des classements complets Σ_t qui peuvent être coûteux à obtenir en pratique, surtout lorsque le nombre d'éléments N est grand.

1.6 Réduction de la dimensionnalité sur $\mathfrak{S}_N$

Dans le chapitre IV, nous proposons une généralisation de l'agrégation de classements et de l'approche de Kemeny au problème plus général de la réduction de la dimensionnalité sur le groupe symétrique $\mathfrak{S}_N$. Nous rappelons d'abord que l'espace des distributions sur $\mathfrak{S}_N$ est de dimensionnalité explosive $N!-1$ et nous soulignons que l'agrégation de classements peut être considérée comme une forme extrême de réduction de la dimensionnalité : en effet, elle résume une distribution entière P sur $\mathfrak{S}_N$ par une seule permutation médiane σ^*. Néanmoins, cette approche présente dans sa formulation même l'inconvénient de masquer la complexité de la distribution P, qui peut par exemple être multimodale et ne peut donc pas être fidèlement représentée par une seule permutation. Pour remédier à cela, nous proposons une relaxation de la méthode du consensus de Kemeny en approximant la distribution originale P par une distribution plus simple P', en suivant une approche de *transport optimal* (voir [Vil08] ou [PC$^+$19]). Plus précisément, notre méthode consiste à choisir des distributions P' dans un ensemble $\mathbf{P}_{\mathcal{C}}$ de *distributions de "bucket"*, c'est-à-dire telles que les probabilités par paire

$$p'_{i,j} = \mathbb{P}_{\Sigma' \sim P'}\{\Sigma'(i) < \Sigma'(j)\}$$

sont égales à zéro ou à un aussitôt que les deux éléments i et j appartiennent à deux cellules distinctes $\mathcal{C}_k$ et $\mathcal{C}_l$ de l'*ordre de "bucket"* $\mathcal{C} = (\mathcal{C}_1, \ldots, \mathcal{C}_K)$, qui est une partition ordonnée de $[\![N]\!]$. Formellement, $\bigcup_{k=1}^{K} \mathcal{C}_k = [\![N]\!]$, $\mathcal{C}_k \neq \emptyset$ pour tout $1 \leq k \leq K$, et si $1 \leq k < l \leq K$, alors $\mathcal{C}_k \cap \mathcal{C}_l = \emptyset$ et

$$(i,j) \in \mathcal{C}_k \times \mathcal{C}_l \Rightarrow p'_{j,i} = 1 - p'_{i,j} = 0.$$

Intuitivement, une distribution $P' \in \mathbf{P}_{\mathcal{C}}$ est contrainte de telle manière qu'elle ne peut pas générer de permutations hésitantes sur le rang relatif de deux éléments dans différentes cellules. Par exemple avec les quatre classements $\sigma_1, \ldots, \sigma_4$ de la figure I.4 et les trois ordres de bucket $\mathcal{C}, \mathcal{C}', \mathcal{C}''$ de la figure I.5,

- les deux classements σ_1 et σ_2 satisfont tous deux la structure de l'ordre de bucket $\mathcal{C}$: $\{\delta_{\sigma_1}, \delta_{\sigma_2}\} \subset \mathbf{P}_{\mathcal{C}}$,
- le classement σ_3 satisfait $\mathcal{C}'$: $\delta_{\sigma_3} \in \mathbf{P}_{\mathcal{C}'}$,

- le dernier classement σ_4 ne satisfait les contraintes d'aucun des trois ordres de bucket : $\delta_{\sigma_4} \notin \mathbf{P}_{\mathcal{C}} \cup \mathbf{P}_{\mathcal{C}'} \cup \mathbf{P}_{\mathcal{C}''}$.

Étant donné un ordre de bucket $\mathcal{C}$, nous dénotons par $\lambda = (\#\mathcal{C}_1, \ldots, \#\mathcal{C}_K)$ sa *forme* : elle décrit le nombre d'éléments contenus dans chacune des K cellules et détermine la dimensionnalité de $\mathbf{P}_{\mathcal{C}}$, à savoir $d_{\mathcal{C}} = \Pi_{1 \leq k \leq K} \#\mathcal{C}_k! - 1$.

Ensuite, le proxy optimal $P^*_{\mathcal{C}} \in \mathbf{P}_{\mathcal{C}}$ d'une distribution générale P est choisi en minimisant la *distance de Wasserstein* $W_{d_\tau,1} = \inf_{\Sigma \sim P, \Sigma' \sim P'} \mathbb{E}[d_\tau(\Sigma, \Sigma')]$:

$$P^*_{\mathcal{C}} \in \arg\min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_\tau,1}(P, P').$$

Le cas sans contrainte $K = 1$, ou de façon équivalente $\lambda = N$ et $d_{\mathcal{C}} = N! - 1$, correspond à aucune réduction de la dimensionnalité avec $\mathbf{P}_{\mathcal{C}}$ égal à tout l'espace des distributions sur $\mathfrak{S}_N$ et donc $P^*_{\mathcal{C}} = P$. Dans le cas extrême opposé $K = N$, équivalent à $\lambda = (1, \ldots, 1)$ et $d_{\mathcal{C}} = 0$, l'ensemble $\mathbf{P}_{\mathcal{C}} = \{\delta_{\sigma_{\mathcal{C}}}\}$ est réduit à un singleton : comme dans l'agrégation de classements, la distribution P est simplement approchée par un seul classement, ici la permutation unique $\sigma_{\mathcal{C}}$ telle que $i \in \mathcal{C}_{\sigma_{\mathcal{C}}(i)}$ pour tous les $i \in [\![N]\!]$.

Contributions. Notre analyse de ce problème repose sur les résultats suivants.

(i) Nous montrons que la *distorsion* $\Lambda_P(\mathcal{C}) = \min_{P' \in \mathbf{P}_{\mathcal{C}}} W_{d_\tau,1}(P, P')$ de tout ordre de bucket $\mathcal{C}$ s'écrit simplement en termes de probabilités par paire :

$$\Lambda_P(\mathcal{C}) = \sum_{1 \leq k < l \leq K} \sum_{(i,j) \in \mathcal{C}_k \times \mathcal{C}_l} p_{j,i},$$

qui peut être estimée empiriquement à partir de comparaisons par paire, de manière similaire au risque de la méthode du consensus de Kemeny.

(ii) Nous formulons et analysons la version MRE du problème d'optimisation de l'ordre de bucket :

$$\min_{\mathcal{C} \in \mathbf{C}_{K,\lambda}} \Lambda_P(\mathcal{C}),$$

où $\mathbf{C}_{K,\lambda}$ désigne l'ensemble de tous les ordres de bucket avec le même nombre de cellules K et la même forme λ.

Nous soulignons que pour la forme $\lambda = (1, \ldots, 1)$, le problème (ii) coïncide avec la méthode du consensus de Kemeny : en effet, nous avons $\Lambda_P(\mathcal{C}) = L_P(\sigma_{\mathcal{C}})$ et $\{\sigma_{\mathcal{C}} : \mathcal{C} \in \mathbf{C}_{N,\lambda}\} = \mathfrak{S}_N$ dans ce cas. Par conséquent, notre approche de réduction de la dimensionnalité prolonge naturellement l'agrégation de classements. Nous fournissons également un algorithme hiérarchique, appelé BuMeRank, qui fusionne récursivement les cellules adjacentes dans des ordres de bucket plus grossiers.

2 Apprentissage par renforcement avec aversion au risque

Cette section présente les contributions de cette thèse dans deux cadres (imbriqués) d'*apprentissage en ligne* : les bandits manchots et l'apprentissage par renforcement, le premier étant un cas particulier du second. Désormais, contrairement aux problèmes de MRE hors ligne exposés dans la section précédente, les données d'entraînement ne sont pas initialement mises à la disposition de l'apprenant. En effet, l'apprenant/décideur doit interagir avec un *environnement* pour simultanément recueillir des observations et concevoir sa propre stratégie.

2.1 Bandits manchots stochastiques

Le problème du *bandit manchot* (BM) (voir par exemple [BCB+12]) est un problème de prise de décision séquentiel rencontré par un joueur dans un casino face à $K \geq 1$ machines à sous : à chaque itération $t \in \{1, \dots, T\}$ (avec $T \geq 1$ l'horizon temporel), il choisit une machine à sous (alias "bandit à un bras", ou simplement "bras") $A_t \in \{1, \dots, K\}$ puis reçoit une récompense aléatoire $X_{A_t,t}$. Au début, le joueur/apprenant/décideur n'a aucune connaissance préalable des machines et son objectif est de maximiser sa récompense totale moyenne à travers toutes les itérations, à savoir $\mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right]$.

Dans le cadre *stochastique*, nous supposons que les récompenses $X_{a,1}, \dots, X_{a,T}$ générées par chaque machine $a \in \{1, \dots, K\}$ sont échantillonnées de façon i.i.d. depuis une distribution de probabilité ν_a sur $\mathbb{R}$ d'espérance μ_a. Dès lors, la quantité à maximiser s'écrit :

$$\mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right] = \sum_{a=1}^{K} \mu_a \mathbb{E}[N_a(T)], \tag{A.8}$$

où $N_a(t') = \sum_{t=1}^{t'} \mathbb{I}\{A_t = a\}$ indique le nombre de fois où le bras a a été tiré jusqu'à un instant $t' \geq 1$. Un *modèle* $\mathcal{D}$ de BM est un ensemble de distributions possibles ν_a avec espérances finies : chaque K-tuple $(\nu_1, \dots, \nu_K) \in \mathcal{D}^K$ caractérise une instance de problème de BM. La *stratégie optimale en rétrospective* consiste donc à toujours tirer le bras optimal a^*, supposé unique :

$$a^* = \arg\max_{1 \leq a \leq K} \mu_a,$$

où $\mu^* = \mu_{a^*} = \max_{1 \leq a \leq K} \mu_a > \max_{a \neq a^*} \mu_a$. Formellement, une stratégie de BM est une fonction $h_t \mapsto (\mathbb{P}\{A_{t+1} = 1 | h_t\}, \dots, \mathbb{P}\{A_{t+1} = K | h_t\})$, où l'*historique* des bras tirés et des récompenses obtenues jusqu'à l'itération courante t est dénotée par

$$h_t = (A_1, X_{A_1,1}, \dots, A_t, X_{A_t,t}).$$

La recherche d'une stratégie maximisant l'éq. (A.8) peut être reformulée de manière équivalente par la minimisation du *regret espéré*

$$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_{A_t,t}\right] = \sum_{a=1}^{K} \Delta_a \mathbb{E}[N_a(T)],$$

avec $\Delta_a = \mu^* - \mu_a$. Ce regret s'interprète comme le déficit global de récompenses espérées généré par une stratégie par rapport à la meilleure stratégie en rétrospective recevant la récompense $X_{a^*,t}$ à chaque étape $1 \leq t \leq T$.

Exploration contre Exploitation. Le problème du BM a été introduit à l'origine par Thompson en 1933 ([Tho33]), motivé par des essais cliniques comparant l'efficacité de plusieurs traitements par des tests sur une série de patients. Dans ce contexte médical, chaque récompense correspond à l'effet observé d'un traitement sur un patient : les traitements sous-optimaux doivent donc être rapidement identifiés, puis écartés pour sauver le plus grand nombre de patients possible. D'une part, l'éventail de tous les traitements possibles doit être suffisamment *exploré* pour permettre de repérer, avec un degré de confiance élevé, le meilleur d'entre eux ; et d'autre part, le meilleur traitement devrait être *exploité* aussi fréquemment que possible, c'est-à-dire fourni au plus grand nombre de patients, en évitant toute exploration superflue. Cet exemple met en évidence le compromis *exploration-exploitation* qui se pose dans les problèmes de bandits, y compris dans les applications modernes telles que le placement d'annonces publicitaires (voir [BCB+12]).

Bornes inférieures asymptotiques. [LR85], [BK96], [CK15] et [GMS19] ont prouvé que, asymptotiquement, le regret de toute stratégie *uniformément efficace* est borné inférieurement par une fonction logarithmique de l'horizon temporel T multipliée par une constante dépendant de la distribution et impliquant des divergences de Kullback-Leibler.

Definition 1. *Une stratégie de BM est uniformément efficace pour un modèle $\mathcal{D}$ si pour tous les problèmes de BM $(\nu_a)_{1\leq a\leq K} \in \mathcal{D}^K$ et pour tous les bras sous-optimaux $a \neq a^*$, elle vérifie :*

$$\mathbb{E}[N_a(T)] = o(T^\alpha), \quad \forall \alpha \in (0, 1].$$

Theorem 1 (Théorème 1 dans [GMS19])**.** *Pour tout modèle $\mathcal{D}$, toute stratégie de BM uniformément efficace sur $\mathcal{D}$, tout problème de BM $(\nu_1, \ldots, \nu_K) \in \mathcal{D}^K$ et tout bras sous-optimal a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_a(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}(\nu_a, \mu^*, \mathcal{D})},$$

où

$$\mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}) = \inf\{KL(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D} \text{ et } \mathbb{E}_{X'\sim\nu'_a}[X'] > x\},$$

avec KL la divergence de Kullback-Leibler entre deux distributions de probabilité.

En particulier, si le modèle $\mathcal{D}$ est une *famille exponentielle unidimensionnelle* (par exemple, les distributions de Bernoulli ou de Poisson), la divergence de Kullback-Leibler $KL(\nu, \nu')$ entre deux distributions ν, ν' dans $\mathcal{D}$ est simplement une fonction de leurs moyennes respectives $\mu = \mathbb{E}_{X\sim\nu}[X]$ et $\mu' = \mathbb{E}_{X'\sim\nu'}[X']$:

$$KL(\nu, \nu') = d(\mu, \mu').$$

Intuitivement, tout algorithme de BM "raisonnable" devrait au moins produire un regret logarithmique : nous rappelons ensuite des stratégies garantissant un regret borné

supérieurement avec une asymptotique correspondant à la limite inférieure du théorème 1.

Algorithmes asymptotiquement optimaux. Plusieurs algorithmes se sont avérés être *asymptotiquement optimaux*, en particulier dans le cas de distributions $\nu_1, \ldots, \nu_K$ appartenant à la même familiale exponentielle, tels que kl-UCB ([GC11]), Bayes-UCB ([Kau16]) et Thompson Sampling ([Tho33], [KKM12], [KKM13]). Ces stratégies sont toutes des *politiques d'indice* c'est-à-dire qu'elles s'appuient sur un certain indice $u_a(t)$ calculé à chaque tour $t \geq 1$ pour chaque bras $a \in \{1, \ldots, K\}$: une politique d'indice générique est décrite dans l'Algorithme 1.

• **L'algorithme kl-UCB.** Partant du même principe d'*optimisme face à l'incertitude* utilisé dans l'algorithme UCB1 ([ACBF02]) à travers le calcul d'intervalles de confiance autour des estimateurs empiriques des moyennes μ_a, l'algorithme kl-UCB a été introduit dans [GC11]. Il s'agit d'une politique d'indice caractérisée par l'indice suivant :

$$u_a(t) = \sup \Big\{ q > \hat{\mu}_a(t) : N_a(t) d(\hat{\mu}_a(t), q) \leq \log t + c \log\log t \Big\}, \tag{A.9}$$

où $\hat{\mu}_a(t) = (1/N_a(t)) \sum_{s=1}^{t} \mathbb{I}\{A_s = a\} X_{a,s}$ est la récompense moyenne empirique au moment t, et c est une constante positive généralement inférieure à 3. Cette stratégie est dite "optimiste" car elle tire le bras avec la borne supérieure de confiance (en anglais: "upper confidence bound", abrégé en "UCB") $u_a(t)$ la plus élevée, c'est-à-dire qu'elle considère (de façon optimiste) que la véritable moyenne μ_a est aussi grande que sa borne supérieure.

• **L'algorithme Bayes-UCB.** L'algorithme Bayes-UCB ([KCG12]) est une politique d'indice bayésienne. Il repose sur la même intuition que kl-UCB mais remplace l'UCB par un quantile supérieur :

$$u_a(t) = Q(1 - 1/(t(\log t)^c), \pi_{a,t}), \tag{A.10}$$

où $Q(\alpha, \pi_{a,t})$ désigne le quantile d'ordre α de la distribution postérieure $\pi_{a,t}$ pour le bras a au moment t.

• **L'algorithme Thompson Sampling.** La stratégie Thompson Sampling (initialement proposée dans [Tho33], et analysée dans [KKM12], [KKM13]) est une approche bayésienne consistant à échantillonner un paramètre naturel (d'une distribution exponentielle unidimensionnelle) $\theta_a(t) \sim \pi_a(t)$ de la distribution postérieure $\pi_a(t)$ mise à jour avec les $N_a(t)$ observations obtenues du bras a jusqu'au temps t. Ensuite, l'indice est donné par :

$$u_a(t) = \mu(\theta_a(t)), \tag{A.11}$$

avec $\mu(\theta)$ la valeur moyenne de la distribution exponentielle unidimensionnelle $\nu \in \mathcal{D}$ de paramètre naturel θ.

Nous présentons dans la sous-section suivante notre étude d'une variante du problème de BM adaptée aux applications de gestion du risque de crédit.

2.2 Bandits pour la gestion du risque de défaut

Dans le problème de la gestion du risque de défaut, un prêteur (généralement une banque) reçoit des demandes de crédit — qu'il peut soit accepter soit rejeter — de la part d'individus appartenant à des populations différentes. Chacune des $K \geq 1$ populations est une catégorie (alias un bras), désignée par $a \in \{1, \dots, K\}$, prédéfinie par la banque sur la base de caractéristiques telles que l'âge, le sexe, le salaire ou l'appartenance ethnique par exemple, combiné avec le montant moyen du prêt τ_a. En supposant que la banque dispose d'un budget suffisant, elle souhaite maximiser son profit total en prêtant de l'argent à toutes les catégories de clients rentables, et non pas uniquement à la catégorie la plus rentable. Ainsi, du point de vue bandit manchot, la notion d'unique bras optimal n'est plus pertinente. Plus formellement, nous considérons au chapitre V une variation du problème de BM, que nous appelons *bandits rentables*, où, à chaque itération $t \in \{1, \dots, T\}$, l'apprenant peut tirer un sous-ensemble $\mathcal{A}_t \subseteq \{1, \dots, K\}$ des bras, ou potentiellement aucun bras (c'est-à-dire $\mathcal{A}_t = \emptyset$). À chaque population $a \in \{1, \dots, K\}$ est associée une distribution inconnue ν_a et un seuil connu τ_a. Le seuil τ_a correspond au montant moyen d'argent emprunté à la banque par chaque individu de la population a. En outre, nous supposons qu'à chaque étape t, un nombre aléatoire (borné) $n_a(t)$ de personnes de la catégorie a demandent un crédit. Ensuite, le but est de maximiser le profit cumulé espéré qui s'écrit, pour chaque emprunteur $c \in \{1, \dots, n_a(t)\}$ de toutes les catégories choisies $a \in \mathcal{A}_t$, comme la différence entre le remboursement moyen $\mu_a = \mathbb{E}_{X \sim \nu_a}[X]$ et le montant moyen du prêt τ_a :

$$S_T = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K} \mathbb{I}\{a \in \mathcal{A}_t\} \sum_{c=1}^{n_a(t)} X_{a,c,t} - L_{a,c,t}\right],$$

où les variables aléatoires $X_{a,c,t}$ sont échantillonnées de façon i.i.d. à partir de ν_a et les montants aléatoires $L_{a,c,t}$ des prêts ont une espérance égale à τ_a. Ici aussi, nous reformulons l'objectif au moyen du regret espéré suivant :

$$R_T = \sum_{a \in \mathcal{A}^*} \Delta_a \widetilde{N}_a(T) - S_T = \sum_{a \in \mathcal{A}^*} \Delta_a \left(\widetilde{N}_a(T) - \mathbb{E}[N_a(T)]\right) + \sum_{a \notin \mathcal{A}^*} |\Delta_a| \mathbb{E}[N_a(T)],$$

où $\widetilde{N}_a(T) = \mathbb{E}\left[\sum_{t=1}^{T} n_a(t)\right]$ est le nombre total espéré de clients de la catégorie a au cours des T tours, $N_a(t) = \sum_{s=1}^{t} n_a(s)\mathbb{I}\{a \in \mathcal{A}_s\}$ est le nombre d'observations obtenues de la catégorie a jusqu'au temps $t \geq 1$, $\Delta_a = \mu_a - \tau_a$ est le profit espéré (inconnu) produit par un client de la catégorie a et $\mathcal{A}^* = \big\{a \in \{1, \dots, K\}, \Delta_a > 0\big\}$ est l'ensemble des bras rentables.

Politiques d'indice. Motivés par le succès des politiques d'indice de BM rappelées plus haut, nous les adaptons à ce nouveau problème de bandits rentables : à chaque itération t et pour chaque catégorie a, un indice $u_a(t)$ est calculé, et le bras a est tiré si $u_a(t)$ est supérieur au seuil connu τ_a (voir Algorithme 2). Ainsi, nous proposons trois politiques d'indice :

- l'algorithme KL-UCB-4P ("4P" signifiant "for profit") avec le même indice que KL-UCB, à savoir $u_a(t)$ défini à l'éq. (A.9),
- l'algorithme Bayes-UCB-4P avec le même indice que Bayes-UCB (voir l'éq. (A.10)),
- l'algorithme TS-4P avec le même indice que Thompson Sampling (voir l'éq. (A.11)).

Notre analyse montre que ces trois stratégies sont toutes *asymptotiquement optimales* pour le problème des bandits rentables.

Contributions. Nous étendons l'analyse de BM à notre cadre de bandits rentables par le biais de deux résultats principaux : respectivement, les bornes inférieure et supérieure sur le regret.

(i) Premièrement, nous montrons que toute stratégie *uniformément efficace* de bandits rentables produit un regret R_T asymptotiquement borné inférieurement comme suit :

$$\liminf_{T\to\infty} \frac{R_T}{\log T} \geq \sum_{a\notin\mathcal{A}^*} \frac{|\Delta_a|}{\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a)},$$

où

$$\mathcal{K}_{\inf}(\nu_a, x, \mathcal{D}_a) = \inf\{\mathrm{KL}(\nu_a, \nu'_a) : \nu'_a \in \mathcal{D}_a \text{ et } \mathbb{E}_{X'\sim\nu'_a}[X'] > x\}.$$

(ii) Si $n_a(t) = n_a$ presque sûrement pour tous les $1 \leq t \leq T$, avec la constante $n_a \geq 1$, alors les trois algorithmes que nous proposons, à savoir KL-UCB-4P, Bayes-UCB-4P et TS-4P sont tous asymptotiquement optimaux c'est-à-dire que leur regret correspond asymptotiquement à la borne inférieure. Sinon, il existe un écart multiplicatif entre nos bornes inférieure et supérieure. En effet, nous fournissons des bornes supérieures sur le regret avec l'asymptotique suivante :

$$\limsup_{T\to\infty} \frac{R_T}{\log T} \leq \sum_{a\notin\mathcal{A}^*} \left(\frac{n_a^+}{n_a^-}\right) \frac{|\Delta_a|}{\mathcal{K}_{\inf}(\nu_a, \tau_a, \mathcal{D}_a)},$$

pour les constantes $n_a^+, n_a^- \geq 1$ telles que $n_a^- \leq n_a(t) \leq n_a^+$ presque sûrement pour tous les t.

2.3 Bandits et valeurs extrêmes

Dans divers contextes d'*aversion au risque*, les quantités d'intérêt ne sont pas nécessairement des moyennes. Dans certaines applications environnementales ou financières par exemple, un décideur peut être averse au risque en s'assurant qu'il est suffisamment à l'abri d'événements désastreux tels que des inondations ou une crise financière (voir [BGST06], [Res07]). En d'autres termes, les stratégies efficaces sont conçues en accordant plus d'importance aux scénarios les plus défavorables qu'aux observations "normales".

Mathématiquement, ces scénarios pessimistes sont souvent modélisés comme des événements rares et extrêmes. Ainsi, dans de nombreux problèmes liés à la prise de risque, l'apprenant doit optimiser un critère basé sur la queue d'une certaine distribution, qui caractérise ses valeurs extrêmes.

Mesures du risque. Rappelons d'abord que l'espérance d'une variable aléatoire à valeur réelle X de fonction de répartition ("f.d.r." en abrégé) $F : x \mapsto \mathbb{P}\{X \leq x\}$ est égale à l'intégrale de la *fonction quantile* (ou *fonction de répartition inverse généralisée*) $F^{-1} : \tau \mapsto \inf\{x : F(x) \geq \tau\}$ sur l'intervalle $[0, 1]$ (voir [Dev08]) :

$$\mathbb{E}[X] = \mathbb{E}_{U \sim \mathcal{U}([0,1])}[F^{-1}(U)] = \int_{\tau=0}^{1} F^{-1}(\tau) d\tau,$$

où $\mathcal{U}([0, 1])$ est la distribution uniforme sur $[0, 1]$. Plusieurs mesures de risque ont été proposées dans la littérature pour remplacer l'espérance :

- le *quantile* $F^{-1}(\tau)$ à un certain niveau $\tau \in (0, 1]$, parfois appelé "value at risk" (VaR) (voir par exemple [ADEH99]), s'écrit aussi comme une solution du problème de minimisation asymétrique L_1 suivant :

$$F^{-1}(\tau) \in \arg\min_{\theta} \mathbb{E}[\ell_\tau^q(X - \theta)], \tag{A.12}$$

  avec la fonction de perte de la régression quantile (alias "perte pinball") $\ell_\tau^q(x) = x(\tau - \mathbb{I}\{x < 0\})$,

- la "*conditional value-at-risk*" (CVaR) $\mathrm{CVaR}_\alpha(X)$, aussi appelée "expected shortfall" ou encore "superquantile" ([RU+00]) :

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{\alpha} \int_{\tau=1-\alpha}^{1} F^{-1}(\tau) d\tau, \tag{A.13}$$

  décrit mieux la queue (droite) de la distribution que ne le fait l'espérance, car la fonction quantile F^{-1} n'est intégrée que sur une partie supérieure de l'intervalle $(0, 1)$,

- l'"*expectile*" $e_\tau(X)$, partageant des propriétés communes avec les quantiles, résout quant à lui un problème de minimisation asymétrique L_2 ([NP87]) :

$$e_\tau(X) = \arg\min_{\theta} \mathbb{E}[\ell_\tau^e(X - \theta)], \tag{A.14}$$

  avec la fonction de perte expectile $\ell_\tau^e(x) = x^2|\tau - \mathbb{I}\{x < 0\}|$.

Nous signalons qu'au niveau $\tau = 1/2$, le quantile $F^{-1}(1/2)$ est une médiane de la distribution de X et l'expectile son espérance : $e_{1/2}(X) = \mathbb{E}[X]$. De plus, si $\alpha = 1$, la CVaR coïncide avec l'espérance : $\mathrm{CVaR}_1(X) = \mathbb{E}[X]$. Empiriquement, la CVaR_α d'une certaine distribution ν peut être estimée à partir d'un échantillon i.i.d. $X_t \sim \nu$ $(1 \leq t \leq T)$ par :

$$\widehat{\mathrm{CVaR}}_\alpha = \frac{1}{\lceil \alpha T \rceil} \sum_{t=\lfloor (1-\alpha)T \rfloor}^{T} X_{\sigma(t)}, \tag{A.15}$$

où la permutation $\sigma \in \mathfrak{S}_T$ numérote les statistiques d'ordre $X_{\sigma(1)} \leq \cdots \leq X_{\sigma(T)}$. Nous rappelons que $\widehat{\mathrm{CVaR}}_\alpha$ est une L-statistique (voir [VdV00]), c'est-à-dire une combinaison linéaire des statistiques d'ordre.

Bandits Prudents. De nombreuses variantes du problème classique de BM ont été proposées pour des applications sensibles au risque. Elles consistent essentiellement à remplacer l'espérance par différentes mesures de risque. Alors que [SBFWH15] se concentre sur les quantiles, [GST13] et [KJ+19] proposent tous deux des stratégies reposant sur l'estimation de la CVaR. Des cadres généraux de bandits englobant de larges classes de critères de risque (y compris les quantiles et la CVaR) sont étudiés dans [TGP19] et [CMZ18]. Dans [SLM12], [VZ16] et [ZIJC14], la qualité d'un bras est évaluée par une combinaison de sa moyenne et de sa variance : pour deux bras ayant la même moyenne, celui ayant la plus faible variance est jugé plus sûr que l'autre. Dans [Mai13], l'aversion au risque est mesurée au moyen des fonctions génératrices des cumulants des distributions des bras ; l'approche moyenne-variance apparaît alors comme un cas particulier de cette méthode dans le scénario gaussien (c'est-à-dire lorsque ν_a est une distribution normale pour chaque bras a). Nous présentons ensuite le problème de "max K-armed bandit", également appelé "bandits extrêmes", comme une forme extrême des problèmes averses au risque évoqués précédemment. En effet, si la CVaR_α (avec $\alpha << 1$) d'une certaine distribution ν permet d'étudier sa queue droite — empiriquement, en sélectionnant la fraction α des plus grandes statistiques d'ordre $X_{\sigma(\lfloor(1-\alpha)T\rfloor)}, \ldots, X_{\sigma(T)}$ parmi un échantillon i.i.d. $(X_t)_{1\leq t\leq T}$ (voir Eq. (A.15)) —, le problème de bandits extrêmes défini ci-dessous se concentre uniquement sur l'observation maximale de cet échantillon, à savoir $\max_{1\leq t\leq T} X_t$.

Bandits Extrêmes. Dans certaines applications en médecine, assurance ou finance, la quantité d'intérêt n'est pas le rendement moyen, mais plutôt les observations *extrêmes* ([BGST06]). Du point de vue des bandits manchots, le "meilleur" bras n'est alors pas forcément celui qui a la récompense moyenne la plus élevée, mais plutôt celui produisant les valeurs maximales. Ce cadre, appelé *bandits extrêmes* dans [CV14], a été introduit à l'origine par [CS05] sous le nom de problème de "max K-armed bandit". Dans ce problème, l'objectif poursuivi est d'obtenir la plus haute récompense au cours des $T \geq 1$ étapes. Pour un bras donné $a \in \{1, \ldots, K\}$, nous dénotons par

$$G_T^{(a)} = \max_{1\leq t\leq T} X_{a,t}$$

la réalisation maximale jusqu'à l'étape $T \geq 1$ et supposons que, en moyenne, il y a un unique bras optimal

$$a^* = \arg\max_{1\leq a\leq K} \mathbb{E}\left[G_T^{(a)}\right].$$

Ensuite, le *regret extrême espéré* d'une stratégie, tirant le bras $A_t \in \{1, \ldots, K\}$ au temps t, est défini par

$$R_T = \mathbb{E}\left[G_T^{(a^*)}\right] - \mathbb{E}\left[\max_{1\leq t\leq T} X_{A_t,t}\right], \tag{A.16}$$

où $\max_{1\leq t\leq T} X_{A_t,t}$ est la valeur maximale observée par l'apprenant jusqu'à l'horizon temporel T. Lorsque les supports des K distributions des récompenses $\nu_1, \ldots, \nu_K$ sont

bornés, aucun regret n'est attendu à condition que chaque bras puisse être suffisamment exploré, comme le montrent [NLB16] et [DS16]. Si le nombre de bras est infini, le défi consiste alors à explorer et à exploiter de manière optimale le réservoir inconnu de bras, voir [CV15]. Lorsqu'au contraire les récompenses ne sont pas bornées, la situation est tout à fait différente : le meilleur bras est celui pour lequel le maximum $G_T^{(a)}$ tend vers l'infini plus vite que les autres. Dans [NLB16], il est montré que, pour des distributions non bornées, aucune politique ne peut parvenir à un regret nul sans hypothèses restrictives sur les distributions. Conformément à la littérature, nous nous concentrons sur un cadre classique en analyse des valeurs extrêmes. Plus précisément, nous supposons que chaque distribution de récompense est à *queue lourde*.

Les distributions à queue lourde sont très largement utilisées pour modéliser les extrêmes dans de nombreuses applications, lorsqu'une approche prudente par évaluation des risques est requise (par exemple en finance, ou pour gérer les risques environnementaux). Comme dans [CV14], nous supposons que les récompenses sont distribuées selon des lois de Pareto du second ordre, qui sont semblables aux distributions de Pareto classiques. Formellement, une loi de probabilité avec f.d.r. $F(x)$ appartient à la famille (α, β, C, C')-Pareto du second ordre si, pour chaque $x \geq 0$,

$$|1 - Cx^{-\alpha} - F(x)| \leq C' x^{-\alpha(1+\beta)}\,, \tag{A.17}$$

où α, β, C et C' sont des constantes strictement positives, voir *e.g.* [Res07]. Naturellement, la *distribution de Pareto* avec *indice de queue* α et paramètre d'échelle C, dont la f.d.r. est :

$$\forall x \geq C^{\frac{1}{\alpha}}, \quad F(x) = 1 - Cx^{-\alpha},$$

appartient à cette famille car elle vérifie trivialement l'équation (A.17). Ces distributions sont en effet à "queue lourde", voir la figure I.7 pour une comparaison avec une distribution normale repliée à "queue légère". Pour chaque bras $a \in \{1, \ldots, K\}$, la distribution ν_a appartient par hypothèse à la famille de Pareto du second ordre $(\alpha_a, \beta_a, C_a, C')$ avec $\alpha_a > 1$, de sorte que l'espérance de la variable aléatoire $X_{a,t} \sim \nu_a$ est finie. Dans ce contexte, [CV14] ont proposé l'algorithme ExtremeHunter pour résoudre le problème de *bandits extrêmes* et ont fourni une analyse du regret extrême avec la borne supérieure suivante :

$$R_T = O\left(T^{\frac{1}{(1+b)\alpha_{a^*}}}\right),$$

où $b > 0$ est une borne inférieure connue sur les coefficients (inconnus) β_a : $b \leq \min_a \beta_a$.

Contributions. Notre contribution à ce problème est présentée dans le chapitre VI : elle est double.

(i) Premièrement, nous améliorons significativement l'analyse du regret de ExtremeHunter par un facteur polynomial en l'horizon temporel T, en prouvant que

$$R_T = O\left((\log T)^{2(2b+1)/b} T^{-(1-1/\alpha_{a^*})} + T^{-(b-1/\alpha_{a^*})}\right),$$

et nous fournissons une borne inférieure correspondante dans un cas spécifique. Cela repose essentiellement sur une majoration plus fine de la différence entre l'espérance

du maximum parmi les réalisations indépendantes $X_1, \ldots, X_T$ d'une distribution (α, β, C, C')-Pareto du second ordre, à savoir $\mathbb{E}[\max_{1\leq i\leq T} X_i]$, et son approximation $(TC)^{1/\alpha}\Gamma(1-1/\alpha)$ avec Γ la fonction Gamma. Comme conséquence, nous proposons une stratégie plus simple du type "Explore-Then-Commit" offrant les mêmes garanties théoriques qu'ExtremeHunter.

(ii) Dans un second temps, nous expliquons comment les bandits extrêmes peuvent, dans une certaine mesure, être réduits à un problème de bandits classique. Nous montrons qu'une stratégie de BM telle que Robust-UCB (voir [BCL13]), appliquée sur des récompenses correctement tronquées à gauche $X_{a,t}\mathbb{I}\{X_{a,t} > u\}$ avec un seuil u suffisamment élevé, peut aussi être performante. Cette affirmation est soutenue par des garanties théoriques sur le nombre de tirages du meilleur bras a^* ainsi que par des expériences numériques.

Ensuite, nous considérons le cadre général de l'apprentissage par renforcement, qui inclut le problème de bandit manchot (*statique*) précédemment discuté. Nous soulignons que, à mi-chemin entre ces deux problèmes, des variantes plus *dynamiques* du BM ont également été étudiées dans la littérature : en particulier, les *bandits contextuels* (voir [Woo79], [Sli14], [PR$^+$13]), où les récompenses dépendent de covariables aléatoires observables.

2.4 Apprentissage par Renforcement

Le problème de bandit manchot évoqué plus haut peut être considéré comme une instance très spécifique du cadre plus général de l'*apprentissage par renforcement* (AR). Dans l'apprentissage par renforcement, un agent cherche à maximiser la somme espérée des récompenses futures (actualisées) en interagissant de manière séquentielle avec son environnement. Cette récompense totale définit des fonctions de valeur dépendant de l'état de l'environnement et de l'action prise par l'agent. L'objectif est alors de trouver une politique optimale maximisant ces fonctions de valeur dans chaque état. Si l'environnement est toujours dans le même état, alors l'AR est un problème de bandit où les armes sont les différentes actions. Nous présentons formellement le cadre de l'AR ci-dessous.

Mélanges. Ici ainsi qu'au chapitre VII, nous désignons par $\mathcal{P}(\mathcal{E})$ l'ensemble des distributions de probabilité sur un ensemble $\mathcal{E}$ (soit dénombrable soit $\mathbb{R}$). En outre, étant donné une variable aléatoire Y à valeur dans un ensemble dénombrable $\mathcal{Y}$ et une fonction $\nu : \mathcal{Y} \to \mathcal{P}(\mathcal{E})$, nous dénotons par $\nu(Y) \in \mathcal{P}(\mathcal{E})$ la *distribution de mélange* de la variable aléatoire suivante :

$$\sum_{y\in\mathcal{Y}} \mathbb{I}\{Y = y\}U_y,$$

où $U_y \sim \nu(y)$ et Y sont indépendantes pour tout $y \in \mathcal{Y}$.

Processus de décision markovien. Un *processus de décision markovien (PDM)* est décrit par un quadruplet $(\mathcal{X}, \mathcal{A}, P, R)$ avec

- l'ensemble d'états $\mathcal{X}$,

- l'ensemble d'actions $\mathcal{A}$,
- le noyau de transition $P : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathcal{X})$,
- la fonction de récompense $R : \mathcal{X} \times \mathcal{A} \to \mathcal{P}(\mathbb{R})$.

Pour simplifier, nous supposons que $\mathcal{X}$ et $\mathcal{A}$ sont tous deux dénombrables. Si l'environnement est dans l'état $x \in \mathcal{X}$ et si l'agent prend l'action $a \in \mathcal{A}$, alors il reçoit une récompense $R_0 \sim R(x, a)$ et l'état suivant X_1 est échantillonné depuis la distribution $P(\cdot|x, a) \in \mathcal{P}(\mathcal{X})$ de telle sorte que R_0, X_1 soient indépendants. Voir la figure I.9 pour un exemple basique de PDM à deux états, deux actions, et des *récompenses déterministes* (c'est-à-dire qu'il existe une fonction $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ telle que $R(x, a) = \delta_{r(x,a)}$ pour tout $(x, a) \in \mathcal{X} \times \mathcal{A}$).

Une politique $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{A})$ associe à tout état $x \in \mathcal{X}$ une distribution sur les actions $\pi(\cdot|x) \in \mathcal{P}(\mathcal{A})$. Avec un facteur d'actualisation $\gamma \in [0, 1[$, nous définissons la distribution $Z^\pi(x, a)$ du revenu total d'une politique π après avoir pris l'action $a \in \mathcal{A}$ dans l'état $x \in \mathcal{X}$ comme étant la *distribution de probabilité* de la variable aléatoire suivante :

$$\sum_{t=0}^{\infty} \gamma^t R_t \quad \text{étant donné que } X_0 = x, A_0 = a,$$
$$\text{et pour tout } t \in \mathbb{N}, R_t \sim R(X_t, A_t), X_{t+1} \sim P(\cdot|X_t, A_t), A_{t+1} \sim \pi(\cdot|X_{t+1}). \quad \text{(A.18)}$$

Le taux d'actualisation γ sert à la fois à assurer la convergence du revenu total, et de paramètre déterminant la valeur actuelle des récompenses futures : une petite valeur de γ donne peu d'importance aux récompenses futures. Une alternative à l'Eq. (A.18), que nous ne considérerons pas ici, est la somme des récompenses $\sum_{t=0}^{T} R_t$, qui n'a de sens que lorsqu'il existe une notion naturelle d'horizon temporel T (voir [SB18]). En général, l'AR se concentre sur les revenus espérés par le biais de la *fonction de valeur état-action*

$$Q^\pi(x, a) = \mathbb{E}_{Z_0 \sim Z^\pi(x,a)}[Z_0],$$

et la *fonction de valeur*

$$V^\pi(x) = \mathbb{E}_{A_0 \sim \pi(\cdot|x)}[Q^\pi(x, A_0)],$$

vérifiant l'*équation de Bellman* ([Bel66]) :

$$\forall (x, a), \quad Q^\pi(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q^\pi(X_1, A_1)],$$

où $R_0 \sim R(x, a)$, $X_1 \sim P(\cdot|x, a)$ et $A_1 \sim \pi(\cdot|X_1)$. Les *politiques optimales* peuvent être caractérisées au moyen de la *fonction de valeur optimale état-action* $Q^*(x, a)$, qui vérifie l'*équation d'optimalité de Bellman* :

$$\forall (x, a), \quad Q^*(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q^*(X_1, a')].$$

Ensuite, en désignant par $V^*(x) = \max_a Q^*(x, a)$ la *fonction de valeur optimale*, une politique π^* est *optimale* si et seulement si pour tout état x,

$$\mathbb{E}[Q^*(x, A_0)] = V^*(x), \quad \text{avec } A_0 \sim \pi^*(\cdot|x).$$

Opérateurs de Bellman. Dans la tâche d'*evaluation de politique*, on souhaite calculer Q^π pour une politique donnée π, alors que dans la tâche de *contrôle*, le but est d'approcher Q^*. La méthode classique de programmation dynamique pour résoudre ces deux tâches repose sur deux opérateurs. D'une part, l'*opérateur de Bellman* T^π ([Bel66]) défini par : pour tout $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ et $(x, a) \in \mathcal{X} \times \mathcal{A}$,

$$T^\pi Q(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[Q(X_1, A_1)], \quad \text{avec } X_1 \sim P(\cdot|x, a), A_1 \sim \pi(\cdot|X_1).$$

D'autre part, l'*opérateur d'optimalité de Bellman* T défini par :

$$TQ(x, a) = \mathbb{E}[R_0] + \gamma \mathbb{E}[\max_{a'} Q(X_1, a')], \quad \text{avec } X_1 \sim P(\cdot|x, a).$$

En particulier, l'opérateur de Bellman T^π (resp. opérateur d'optimalité de Bellman T) est une γ-contraction[1] pour la norme sup et son application répétée à une Q-fonction initiale converge exponentiellement rapidement vers son unique point fixe Q^π (resp. Q^*) ([BT96]).

Algorithmes d'AR. En AR, le noyau de transition P est inconnu et les opérateurs de Bellman ne peuvent donc pas être calculés exactement. D'où les méthodes pratiques d'AR telles que le "temporal-difference learning" (TD) ([Sut88]), SARSA ([RN94]), ou encore Q-learning ([Wat89]), consistant à calculer des approximations stochastiques de ces opérateurs à partir de trajectoires composées de séquences "état-action-récompense" observées. Avec une seule *transition*

$$(X_t, A_t, R_t, X_{t+1}, A_{t+1}),$$

où $R_t \sim R(X_t, A_t)$, $X_{t+1} \sim P(\cdot|X_t, A_t)$, $A_{t+1} \sim \pi(\cdot|X_{t+1})$, et le taux d'apprentissage $0 < \alpha \leq 1$, leurs règles de mise à jour sont les suivantes :

- TD(0) :
$$V(X_t) \leftarrow (1 - \alpha)V(X_t) + \alpha(R_t + \gamma V(X_{t+1})),$$
- SARSA(0) (utilisant l'action suivante A_{t+1}) :
$$Q(X_t, A_t) \leftarrow (1 - \alpha)Q(X_t, A_t) + \alpha(R_t + \gamma Q(X_{t+1}, A_{t+1})),$$
- Q-learning :
$$Q(X_t, A_t) \leftarrow (1 - \alpha)Q(X_t, A_t) + \alpha(R_t + \gamma \max_{a' \in \mathcal{A}} Q(X_{t+1}, a')).$$

Sous certaines hypothèses techniques (cadre tabulaire, états et actions visités un nombre infini de fois, taux d'apprentissage constant ou décroissant, etc.), il a été prouvé que les méthodes TD convergent vers la fonction de valeur V^π ([Sut88], [Day92]), tandis que l'algorithme SARSA(0) (combiné avec des politiques gloutonnes, voir [SJLS00]) ainsi que

[1] Une fonction d'un espace métrique vers lui-même est appelée une γ-contraction (resp. une non-expansion) si elle est Lipschitz avec constante de Lipschitz $\gamma < 1$ (resp. $\kappa \leq 1$).

Q-LEARNING (voir [WD92]) convergent tous deux vers la fonction de valeur état-action optimale Q^*.

Comme dans le cas du bandit manchot, de nombreuses variantes averses au risque de l'AR furent proposées en remplaçant les revenus espérés par des critères sensibles au risque, nous renvoyons à [GF15] pour un aperçu de telles méthodes. Notre approche pour ce problème s'appuie sur le cadre plus général de l'apprentissage par renforcement distributionnel (ARD), où l'accent n'est pas seulement mis sur les fonctions de valeur (c'est-à-dire les revenus moyens) comme dans l'AR, mais sur les distributions entières de ces mêmes revenus, ce qui permet potentiellement des applications sensibles au risque basées sur des mesures de risques comme la CVaR par exemple.

2.5 Au-delà des fonctions de valeur : les équations de Bellman atomiques

Dans l'*apprentissage par renforcement distributionnel (ARD)*, l'accent est mis sur la distribution, désignée par $Z^\pi(x,a)$, de la variable aléatoire $\sum_{t\geq 0}\gamma^t R_t$ dans l'Eq. (A.18), et non pas uniquement sur son espérance comme c'est le cas dans l'AR (non distributionnel). Comme le montre [BDM17], les outils d'AR habituels tels que les équations de Bellman (pour les revenus espérés) peuvent être généralisés aux distributions. De même, les auteurs ont proposé deux opérateurs de Bellman distributionnels : alors que le premier, désigné par $\mathcal{T}^\pi$, pour l'évaluation distributionnelle d'une politique donnée π, est une contraction, le second, pour la tâche de contrôle, ne l'est pas (voir respectivement le Lemme 3 et la Proposition 1 dans [BDM17]). Formellement, l'*opérateur distributionnel de Bellman* $\mathcal{T}^\pi$ est défini par : pour toute *fonction de distribution état-action*

$$Z : (x,a) \in \mathcal{X}\times\mathcal{A} \mapsto Z(x,a) \in \mathcal{P}(\mathbb{R}),$$

l'image de Z par $\mathcal{T}^\pi$ est la fonction de distribution état-action $\mathcal{T}^\pi Z$ donnée par :

$$\mathcal{T}^\pi Z : (x,a) \mapsto \text{ distribution de la v.a. } R_0 + \gamma Z_1, \text{ avec } R_0 \sim R(x,a), Z_1 \sim Z(X_1,A_1),$$

où $X_1 \sim P(\cdot|x,a)$ et $A_1 \sim \pi(\cdot|X_1)$. L'*équation de Bellman distributionnelle* s'écrit ensuite :

$$Z^\pi = \mathcal{T}^\pi Z^\pi.$$

En pratique, il peut s'avérer compliqué de calculer des distributions générales. Par conséquent, les approches d'ARD existantes ont été développées en projetant les distributions dans un espace paramétrique plus simple de mesures de probabilité, facilitant ainsi les calculs. Par exemple, [DRBM18] et [RBD+18] approximent tous deux les revenus distributionnels par des distributions atomiques mais considèrent des métriques différentes pour évaluer les erreurs d'approximation : respectivement la distance 1-Wasserstein [2] W_1 et la distance de Cramér.

[2] Pour $p \in [1,+\infty)$, la distance p-Wasserstein entre deux distributions D_1 et D_2 sur $\mathbb{R}$ (de f.d.r. F_1 et F_2) est $W_p(D_1,D_2) = \left(\int_{\tau=0}^1 \left|F_1^{-1}(\tau) - F_2^{-1}(\tau)\right|^p d\tau\right)^{\frac{1}{p}}$.

Projection Atomique. Notre approche au chapitre VII repose sur les deux choix suivants.

(a) Nous approximons les distributions de probabilité D sur $\mathbb{R}$ par des distributions atomiques $D_{\omega,\theta} = \sum_{i=1}^{N} \omega_i \delta_{\theta_i}$ avec $\omega_i \geq 0$ et $\omega_1 + \cdots + \omega_N = 1$, et $\theta_1 \leq \cdots \leq \theta_N$.

(b) Comme dans notre problème de réduction de la dimensionnalité sur le groupe symétrique (partie 1, chapitre IV), nous utilisons une métrique de transport optimal pour mesurer les erreurs d'approximation, à savoir la distance 2-Wasserstein W_2 :

$$W_2(D, D_{\omega,\theta}) = \left(\sum_{i=1}^{N} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} \left(F^{-1}(\tau) - \theta_i \right)^2 d\tau \right)^{\frac{1}{2}}, \tag{A.19}$$

avec les probabilités cumulées $\overline{\omega}_i = \sum_{j \leq i} \omega_j$.

Il est important de noter que pour des probabilités ω_i fixées, l'erreur d'approximation dans l'Eq. (A.19) est minimisée par rapport aux atomes θ_i si et seulement si pour tout $1 \leq i \leq N$ tel que $\omega_i \neq 0$, θ_i est égal à la *moyenne tronquée* suivante de la distribution D :

$$\theta^*_{\omega,i} = \frac{1}{\omega_i} \int_{\tau=\overline{\omega}_{i-1}}^{\overline{\omega}_i} F^{-1}(\tau) d\tau.$$

Nous soulignons que dans le cas monoatomique $N = 1$, l'unique "moyenne tronquée" est simplement l'espérance. En effet, elle s'écrit aussi comme l'intégrale de la fonction quantile sur tout l'intervalle $(0, 1)$: $\mathbb{E}_{Y \sim D}[Y] = \int_{\tau=0}^{1} F^{-1}(\tau) d\tau$, ce qui nous ramène à l'AR classique. En outre, ces moyennes tronquées peuvent être utilisées dans un contexte d'aversion au risque pour calculer des mesures de risque telles que la CVaR :

$$\mathrm{CVaR}_{1-\overline{\omega}_{i-1}}(Y) = \frac{\theta^*_{\omega,i} + \cdots + \theta^*_{\omega,N}}{N - i + 1}.$$

Nous présentons ci-dessous une étude de cas avec deux politiques ayant les mêmes performances moyennes mais des niveaux de risque différents.

Etude de cas - Politique prudente contre politique risquée. Pour le PDM décrit dans la figure I.9, combiné avec un facteur d'actualisation $\gamma = \frac{1}{2}$, les deux politiques π, π' données par $\pi(a_1|\cdot) \equiv 1$ ("toujours choisir l'action a_1") et $\pi'(a_2|\cdot) \equiv 1$ ("toujours choisir l'action a_2") partagent les mêmes fonctions de valeur :

$$\begin{aligned} V^{\pi}(x_1) = Q^{\pi}(x_1, a_1) = \frac{1}{2} = V^{\pi'}(x_1) = Q^{\pi'}(x_1, a_2) \\ \text{et} \quad V^{\pi}(x_2) = Q^{\pi}(x_2, a_1) = \frac{3}{2} = V^{\pi'}(x_2) = Q^{\pi'}(x_2, a_2). \end{aligned} \tag{A.20}$$

Cependant, π' produit des revenus déterministes, contrairement à π, et est donc la plus sûre des deux politiques. Plus précisément, les distributions des revenus dont les moyennes sont données dans l'Eq. (A.20) sont concentrées en des masses de Dirac dans le

cas de la politique "sûre" π', alors qu'elles sont uniformément réparties sur des intervalles dans le cas "risqué" :

$$Z^\pi(x_1, a_1) = \mathcal{U}([0,1]) \neq \delta_{\frac{1}{2}} = Z^{\pi'}(x_1, a_2),$$
$$\text{et} \quad Z^\pi(x_2, a_1) = \mathcal{U}([1,2]) \neq \delta_{\frac{3}{2}} = Z^{\pi'}(x_2, a_2), \quad \text{(A.21)}$$

où $\mathcal{U}([\alpha, \beta])$ désigne la distribution uniforme sur tout intervalle $[\alpha, \beta]$.
Contributions. Notre contribution est triple.

(i) Tout d'abord, nous introduisons deux nouveaux opérateurs "à 1 étape" d'ARD, qui ne traitent que l'aléa induit par la première étape. Le premier, pour l'évaluation de politique, est désigné par $\mathbb{T}^\pi$ et défini par : pour toute fonction de distribution état-action Z et $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\mathbb{T}^\pi Z(x, a)$ est la distribution de la v.a.

$$R_0 + \gamma \mathbb{E}[Z_1 | X_1, A_1], \text{ avec } R_0 \sim R(x, a), Z_1 \sim Z(X_1, A_1), X_1 \sim P(\cdot|x, a), A_1 \sim \pi(\cdot|X_1),$$

tandis que notre second opérateur d'ARD $\mathbb{T}$ (pour la tâche de contrôle) est défini de telle sorte que $\mathbb{T}Z(x, a)$ est la distribution de

$$R_0 + \gamma \max_{a'} \mathbb{E}[Z_{1,a'} | X_1], \text{ avec } R_0 \sim R(x, a), X_1 \sim P(\cdot|x, a), Z_{1,a'} \sim Z(X_1, a') \forall a' \in \mathcal{A}.$$

Il est intéressant de noter que $\mathbb{T}^\pi$ et $\mathbb{T}$ sont tous deux des contractions.

(ii) Ensuite, nous décrivons les *opérateurs projetés* résultant des choix (a) et (b) et prouvons qu'ils sont aussi des contractions. En outre, nous aboutissons aux *équations de Bellman atomiques*, qui sont les équations de point fixe des opérateurs projetés : elles généralisent les équations de Bellman habituelles (non distributionnelles) au cas multiatomique de plusieurs moyennes tronquées.

(iii) Enfin, nous proposons de nouveaux algorithmes d'ARD prolongeant au cas multiatomique les méthodes TD et Q-LEARNING.

En bref, le dernier chapitre VII fournit de nouveaux outils théoriques d'ARD, à savoir les opérateurs à 1 étape et les équations de Bellman atomiques, qui sont voués à être utilisés dans des situations d'aversion au risque.

Titre : Contribution à des problèmes statistiques d'ordonnancement et d'apprentissage par renforcement avec aversion au risque

Mots clés : minimisation du risque empirique, ordonnancement, bandit manchot, apprentissage par renforcement

Résumé : Les travaux de cette thèse se situent à l'interface de deux thématiques de l'apprentissage automatique : l'apprentissage de préférences d'une part, et l'apprentissage par renforcement de l'autre. La première consiste à percoler différents classements d'un même ensemble d'objets afin d'en extraire un ordre général, la seconde à identifier séquentiellement une stratégie optimale en observant des récompenses sanctionnant chaque action essayée. La structure de la thèse suit ce découpage thématique. En première partie, le paradigme de minimisation du risque empirique est utilisé à des fins d'ordonnancement. Partant du problème d'apprentissage supervisé de règles d'ordonnancement à partir de données étiquetées de façon binaire, une extension est proposée au cas où les étiquettes prennent des valeurs continues. Les critères de performance usuels dans le cas binaire, à savoir la courbe caractéristique de l'opérateur de réception (COR) et l'aire sous la courbe COR (ASC), sont étendus au cas continu : les métriques COR intégrée (CORI) et ASC intégrée (ASCI) sont introduites à cet effet. Le second problème d'ordonnancement étudié est celui de l'agrégation de classements à travers l'identification du consensus de Kemeny. En particulier, une relaxation au problème plus général de la réduction de la dimensionnalité dans l'espace des distributions sur le groupe symétrique est formulée à l'aide d'outils mathématiques empruntés à la théorie du transport optimal. La seconde partie de cette thèse s'intéresse à l'apprentissage par renforcement. Des problèmes de bandit manchot sont analysés dans des contextes où la performance moyenne n'est pas pertinente et où la gestion du risque prévaut. Enfin, le problème plus général de l'apprentissage par renforcement distributionnel, dans lequel le décideur cherche à connaître l'entière distribution de sa performance et non pas uniquement sa valeur moyenne, est considéré. De nouveaux opérateurs de programmation dynamique ainsi que leurs pendants atomiques mènent à de nouveaux algorithmes stochastiques distributionnels.

Title : Ranking and Risk-Aware Reinforcement Learning

Keywords : empirical risk minimization, ranking, multi-armed bandit, distributional reinforcement learning

Abstract : This thesis divides into two parts: the first part is on ranking and the second on risk-aware reinforcement learning. While binary classification is the flagship application of empirical risk minimization (ERM), the main paradigm of machine learning, more challenging problems such as bipartite ranking can also be expressed through that setup. In bipartite ranking, the goal is to order, by means of scoring methods, all the elements of some feature space based on a training dataset composed of feature vectors with their binary labels. This thesis extends this setting to the continuous ranking problem, a variant where the labels are taking continuous values instead of being simply binary. The analysis of ranking data, initiated in the 18th century in the context of elections, has led to another ranking problem using ERM, namely ranking aggregation and more precisely the Kemeny's consensus approach. From a training dataset made of ranking data, such as permutations or pairwise comparisons, the goal is to find the single 'median permutation' that best corresponds to a consensus order. We present a less drastic dimensionality reduction approach where a distribution on rankings is approximated by a simpler distribution, which is not necessarily reduced to a Dirac mass as in ranking aggregation. For that purpose, we rely on mathematical tools from the theory of optimal transport such as Wasserstein metrics. The second part of this thesis focuses on risk-aware versions of the stochastic multi-armed bandit problem and of reinforcement learning (RL), where an agent is interacting with a dynamic environment by taking actions and receiving rewards, the objective being to maximize the total payoff. In particular, a novel atomic distributional RL approach is provided: the distribution of the total payoff is approximated by particles that correspond to trimmed means.

Institut Polytechnique de Paris
91120 Palaiseau, France