HAL
open science

Concentration and compression over infinite alphabets, mixing times of random walks on random graphs

Anna Ben-Hamou

To cite this version:

Anna Ben-Hamou. Concentration and compression over infinite alphabets, mixing times of random walks on random graphs. General Mathematics [math.GM]. Université Sorbonne Paris Cité, 2016. English. NNT : 2016USPCC197 . tel-01376929v2

HAL Id: tel-01376929

https://hal.science/tel-01376929v2

Submitted on 15 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS DIDEROT (PARIS 7)
SORBONNE PARIS CITÉ

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Laboratoire de Probabilités et Modèles Aléatoires - CNRS UMR 7599

THÈSE DE DOCTORAT

Discipline : Mathématiques Appliquées

Présentée par

Anna Ben-Hamou

CONCENTRATION ET COMPRESSION SUR ALPHABETS INFINIS,

TEMPS DE MÉLANGE DE MARCHES ALÉATOIRES SUR DES GRAPHES ALÉATOIRES

Sous la direction de **Stéphane BOUCHERON** et **Justin SALEZ**

Rapporteurs :

M. **Laurent**	**MASSOULIÉ**	INRIA - Microsoft Research
M. **Roberto I.**	**OLIVEIRA**	IMPA, Brésil

Soutenue le **15 septembre 2016** devant le jury composé de :

M. Stéphane	BOUCHERON	ENS Paris - Université Paris Diderot	Directeur
Mme Elisabeth	GASSIAT	Université Paris Sud	Examinatrice
M. Jean-François	LE GALL	Université Paris Sud	Président du jury
M. Gábor	LUGOSI	Université Pompeu Fabra	Examinateur
M. Laurent	MASSOULIÉ	INRIA - Microsoft Research	Rapporteur
M. Roberto I.	OLIVEIRA	IMPA	Rapporteur
M. Justin	SALEZ	Université Paris Diderot	Directeur
Mme Perla	SOUSI	Université de Cambridge	Examinatrice

Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse, Stéphane Boucheron et Justin Salez, grâce à qui j'ai pu découvrir les domaines passionnants de la concentration et des temps de mélange. Stéphane, merci de m'avoir introduite à la recherche dès le Master, d'avoir été si encourageant, de m'avoir tant transmis, de m'avoir donné du temps pour tenter de corriger les défaillances de ma culture mathématique et d'avoir gardé le temps de plaisanter. Justin, merci de m'avoir proposé le sujet fascinant du cutoff, de m'avoir guidée, avec patience et bienveillance, sur tout le chemin de la résolution d'un problème, depuis le tâtonnement intelligent du début jusqu'à l'élaboration de la preuve la plus légère et élégante possible.

Je suis honorée que Laurent Massoulié et Roberto Oliveira aient accepté d'être rapporteurs de ma thèse, et les remercie vivement pour l'attention qu'ils ont portée au manuscrit. Je remercie également chaleureusement Elisabeth Gassiat, Jean-François Le Gall, Gábor Lugosi et Perla Sousi d'avoir bien voulu faire partie du jury, c'est pour moi un honneur de soutenir cette thèse devant eux.

J'ai eu la chance durant cette thèse d'être accueillie pendant trois mois dans le groupe théorique de Microsoft Research, et d'y travailler sous la direction de Yuval Peres. Yuval, it was a true honor to work with you and I would like to thank you for offering me this opportunity, for the innumerable things I have learned with you, for the care you take in communicating your profound insights on mixing times.

Je voudrais aussi exprimer toute ma gratitude aux personnes avec qui j'ai travaillé et discuté : cette thèse leur doit énormément. Je remercie en particulier Mesrob Ohannessian, avec qui j'ai adoré collaborer et sans qui j'aurais manqué beaucoup de la masse manquante, Elisabeth Gassiat, sans qui le codage universel me serait resté indéchiffrable, Eyal Lubetzky, who unravelled a lot of my mixing questions, and I hope to keep on exploring all those directions and more with Roberto and Gábor, who I thank a second time for the stimulating discussions we had. Estou muito feliz em trabalhar com vocês no próximo ano no Rio !

Le LPMA est un lieu privilégié pour effectuer une thèse, et je garderai un excellent souvenir de mes années dans ce laboratoire. Je remercie Francis Comets de faire du LPMA un environnement si stimulant et d'avoir toujours été soucieux de la situation des doctorants, ainsi que Valérie Juvé et Nathalie Bergame pour leur disponibilité, leur sympathie, et leur indulgence envers mes nombreuses étourderies.

Je remercie également Mihaï Brancovan, Josselin Garnier et Raphaël Lefevere : j'ai beaucoup aimé assurer les travaux dirigés de vos cours.

Je garde un très bon souvenir du petit groupe de lecture sur les temps de mélange, et remercie Mathieu Merle de l'avoir fait vivre, et de m'avoir souvent conseillée et soutenue.

Un gigantesque merci à tous les doctorants et anciens doctorants de Paris 7, vous avez rendu ces trois années extrêmement agréables et m'avez, de bien des façons, beaucoup aidée à faire cette thèse. Merci à Marc-Antoine Giuliani, Thomas Vareschi, Oriane Blondel, Christophe Poquet, Jiatu Cai, Sébastien Choukroun, Lorick Huang, Aser Cortines, Guillaume Barraquand, Shanqiu Li, Arturo Leos, Sophie Coquan, Vu-Lan Nguyen, Clément Ménassé, Adrien Genin, Thomas Galtier, Maha Khatib, Clément Walter, David Krief, Yann Chiffaudel, Elodie Odrat, Julien-Piera Vest, Amine Ismail, Roman Andreev, Mina Abdel-Sayed, Mi-Song Dupuy, Côme Hure, Ting-Ting Zhang, Remy Degenne, Alexis Bismuth, Laure Marêché, Assaf Shapira, et particulièrement Maud Thomas : Maud, merci d'avoir été une grande soeur de thèse hors-pair, une fournisseuse de première classe en bons conseils et bons chocolats, merci pour ton soutien constant et pour ton amitié.

Merci à Noufel Frikha, à ceux de Jussieu, Sarah Kaakai, Olga Lopusanschi, Alexandre Boumezoued, Yvain Bruned, Loïc de Raphelis, Minmin Wang. Thanks a lot to all the people I met at Microsoft and University of Washington, Sébastien Bubeck, Shirshendu Ganguly, Clayton Barnes, Gireeja Ranade, Janardhan Kulkarni. Je remercie aussi les personnes que j'ai eu le plaisir de rencontrer lors de conférences, Hamed Amini, Cécile Mailler, Marie Albenque, Guillaume Chapuy, Henning Sulzbach, Guillem Perarnau, Laura Eslava.

Je remercie tous ceux qui m'ont fait aimer les maths, Philippe Gallic, Serge Nicolas, et bien sûr ma mère. Et puis je remercie mon frère, qui me fera toujours rire.

Paris, le 5 septembre 2016

Résumé

Ce document rassemble les travaux effectués durant mes années de thèse. Je commence par une présentation concise des résultats principaux, puis viennent trois parties relativement indépendantes. Dans la première partie, je considère des problèmes d'inférence statistique sur un échantillon I.I.D. issu d'une loi inconnue à support dénombrable. Le premier chapitre est consacré aux propriétés de concentration du profil de l'échantillon et de la masse manquante. Il s'agit d'un travail commun avec Stéphane Boucheron et Mesrob Ohannessian. Après avoir obtenu des bornes sur les variances, nous établissons des inégalités de concentration de type Bernstein, et exhibons un vaste domaine de lois pour lesquelles le facteur de variance dans ces inégalités est tendu. Le deuxième chapitre présente un travail en cours avec Stéphane Boucheron et Elisabeth Gassiat, concernant le problème de la compression universelle adaptative d'un tel échantillon. Nous établissons des bornes sur la redondance minimax des classes enveloppes, et construisons un code quasi-adaptatif sur la collection des classes définies par une enveloppe à variation régulière. Dans la deuxième partie, je m'intéresse à des marches aléatoires sur des graphes aléatoires à degrés precrits. Je présente d'abord un résultat obtenu avec Justin Salez, établissant le phénomène de cutoff pour la marche sans rebroussement. Sous certaines hypothèses sur les degrés, nous déterminons précisément le temps de mélange, la fenêtre du cutoff, et montrons que le profil de la distance à l'équilibre converge vers la fonction de queue gaussienne. Puis je m'intéresse à la comparaison des temps de mélange de la marche simple et de la marche sans rebroussement. Enfin, la troisième partie est consacrée aux propriétés de concentration de tirages pondérés sans remise et correspond à un travail commun avec Yuval Peres et Justin Salez.

Abstract

This document presents the problems I have been interested in during my PhD thesis. I begin with a concise presentation of the main results, followed by three relatively independent parts. In the first part, I consider statistical inference problems on an I.I.D. sample from an unknown distribution over a countable alphabet. The first chapter is devoted to the concentration properties of the sample's profile and of the missing mass. This is a joint work with Stéphane Boucheron and Mesrob Ohannessian. After obtaining bounds on variances, we establish Bernstein-type concentration inequalities and exhibit a vast domain of sampling distributions for which the variance factor in these inequalities is tight. The second chapter presents a work in progress with Stéphane Boucheron and Elisabeth Gassiat, on the problem of universal adaptive compression over countable alphabets. We give bounds on the minimax redundancy of envelope classes, and construct a quasi-adaptive code on the collection of classes defined by a regularly varying envelope. In the second part, I consider random walks on random graphs with prescribed degrees. I first present a result obtained with Justin Salez, establishing the cutoff phenomenon for non-backtracking random walks. Under certain degree assumptions, we precisely determine the mixing time, the cutoff window, and show that the profile of the distance to equilibrium converges to the Gaussian tail function. Then I consider the problem of comparing the mixing times of the simple and non-backtracking random walks. The third part is devoted to the concentration properties of weighted sampling without replacement and corresponds to a joint work with Yuval Peres and Justin Salez.

Contents

Résumé détaillé **9**

I Concentration and compression on infinite alphabets **17**

Introduction **19**
1 Sampling from an unknown discrete distribution . . . 19
2 Occupancy counts and occupancy masses . . . 22
3 Karlin's formalism . . . 24

1 Concentration inequalities in infinite occupancy schemes **26**
1 Main results . . . 26
2 Distribution-free concentration . . . 29
2.1 Occupancy counts . . . 29
2.2 Missing mass . . . 32
3 Regular variation . . . 34
3.1 Definition and motivation . . . 34
3.2 Case $\alpha \in (0,1)$ . . . 35
3.3 Fast variation, $\alpha = 1$ . . . 36
3.4 Slow variation, $\alpha = 0$ . . . 37
4 Applications . . . 40
4.1 Estimating the regular variation index . . . 40
4.2 Estimating the missing mass . . . 40
4.3 Estimating the number of species . . . 43
5 Discussion . . . 44
5.1 The cost of Poissonization and negative correlation . . . 44
5.2 Extensions of regular variation . . . 44
5.3 Random measures . . . 45
6 Proofs . . . 46
6.1 Fundamental techniques . . . 46
6.2 Occupancy counts . . . 48
6.3 Missing mass . . . 50
6.4 Regular variation . . . 51
6.5 Applications . . . 54

2 Adaptive coding on countable alphabets **58**
1 Coding on infinite alphabets . . . 58
1.1 Universal source coding . . . 59
1.2 Adaptive source coding . . . 61
1.3 Regularly varying envelope classes . . . 61
1.4 The Pattern Censoring Code . . . 63
2 Main result . . . 65
3 Minimax redundancy . . . 66
3.1 Properties of minimax redundancy . . . 66
3.2 Poisson sampling . . . 67
3.3 Minimax redundancy of envelope classes . . . 68
4 Analysis of the Pattern Censoring Code . . . 69
5 Proofs . . . 72

II Cutoff for random walks on sparse random graphs **81**
1 Introduction . . . 83
1.1 Mixing times . . . 83
1.2 The cutoff phenomenon . . . 84
1.3 Random walks on random graphs with given degrees . . . 86
2 Cutoff for non-backtracking random walks . . . 88
2.1 Statement and comments . . . 89
2.2 Proof outline . . . 92
2.3 Proof details . . . 94
3 Comparing mixing times of simple and non-backtracking random walks . . . 102
3.1 The simple random walk . . . 102
3.2 Entropies on Galton-Watson trees . . . 104

III Weighted sampling without replacement **109**
1 Sampling with and without replacement . . . 111
1.1 The uniform case . . . 111
1.2 Main results . . . 112
2 A useful coupling . . . 115
3 Proofs . . . 116
3.1 Proof of Theorem 1.1 . . . 116
3.2 Proof of Theorem 1.2 . . . 117
3.3 Proof of Theorem 1.3 . . . 120

Résumé détaillé

Ce texte rassemble des travaux effectués au cours de mes trois années de thèse. Les problèmes auxquels je me suis intéressée étant assez éloignés, j'ai préféré présenter ces travaux en trois parties indépendantes. En particulier, elles peuvent être lues dans n'importe quel ordre, et chaque partie comporte son introduction propre. Ici, je tente de résumer de façon succinte les résultats obtenus.

Le profil d'un échantillon (i.i.d.)

Le cadre général de la première partie sera le suivant : on dispose d'un échantillon I.I.D. $(X_1, \dots, X_n)$ à valeurs dans un ensemble dénombrable de symboles $\mathcal{X}$ que l'on appellera souvent un *alphabet*, et qui pour nous sera l'ensemble des entiers naturels strictement positifs : $\mathcal{X} = \mathbb{N}^*$. On note alors $(p_j)_{j\geq 1}$ la loi de X_1. On utilise souvent la métaphore des urnes : l'échantillon peut être interprété comme le résultat de n lancers indépendants de boules dans une collection d'urnes, p_j correspondant à la probabilité qu'une boule tombe dans l'urne j. Que peut-on dire du nombre d'urnes occupées ? Du nombre d'urnes contenant r boules ? Si $K_{n,r}$ est définie comme le nombre d'urnes contenant r boules après n lancers, la suite $(K_{n,r})_{r\geq 1}$ est appelée le *profil* de l'échantillon. S'intéresser au profil, c'est extraire les informations concernant l'occupation des urnes, en ignorant le numéro des urnes.

Un cas particulier qui a fait l'objet de beaucoup d'attention est celui d'un nombre fini d'urnes m, où chaque urne est munie de la probabilité uniforme : $p_1 = \cdots = p_m = \frac{1}{m}$. Le profil de l'échantillon dépend alors de la manière dont on fait dépendre le nombre n de lancers du nombre m d'urnes. Par exemple, si l'on cherche à savoir combien de boules il faut lancer pour qu'au moins une urne contienne plus de deux boules (c'est-à-dire pour que $K_{n,1}$ devienne strictement plus petit que n), le *paradoxe des anniversaires* nous dit qu'il faut prendre $n \asymp \sqrt{m}$. Ou encore, si l'on lance $n = m$ boules, quel est le nombre maximal de boules contenues par une urne ? Dans ce cas, le nombre de boules contenues par une urne donnée est approximativement distribué comme une loi de Poisson de paramètre 1, et le maximum de n telles variables est approximativement $\frac{\log n}{\log\log n}$. L'étude asymptotique de la variable K_n dans différents régimes $n = n(m)$ est au coeur de l'ouvrage de Kolchin et al. [106], qui identifient cinq régimes pour lesquels K_n proprement normalisée converge en loi vers une loi normale ou vers une loi de Poisson. Par exemple, quand $n/N \to c$ avec $c \in]0, +\infty[$, l'espérance et la variance de K_n sont alors linéaires en n, et la loi limite de K_n est gaussienne.

La généralisation de ces questions à un nombre infini d'urnes est due (à notre connaisance) à Bahadur [12], qui s'intéresse essentiellement à la variable K_n. L'analyse systématique du comportement asymptotique de K_n, $K_{n,r}$, ainsi que d'autres variables d'occupation (notamment le nombre d'urnes contenant un nombre pair de boules) remonte au remarquable article de Karlin [102]. Entre beaucoup d'autres résultats dont nous parlerons plus loin, Karlin montre que, dès que le support est infini, la variable K_n vérifie une loi forte des grands nombres : $\frac{K_n}{\mathbb{E}K_n} \to 1$ presque sûrement. Un théorème central limite pour K_n a été établi par Dutko et al. [65], sous la condition $\mathrm{Var}K_n \to +\infty$. Ces résultats reposent crucialement sur la monotonie de K_n. Obtenir des convergences similaires pour les variables $K_{n,r}$, qui ne vérifient pas cette

propriété de monotonie, s'avère beaucoup plus délicat. Par exemple, si $\mathbb{E}K_{n,r}$ ne tend pas vers l'infini, la loi des grands nombres peut ne pas être vérifiée, car $K_{n,r}$ fait des sauts de taille 1. Si l'on veut pouvoir dire quelque chose du comportement asymptotique de ces variables, on doit restreindre d'une certaine façon la classe des lois considérées. Un cadre général proposé par Karlin [102] est celui des lois dites *à variation régulière*. Si l'on définit, pour tout x de $]0,1]$,

$$\vec{\nu}(x) = \left| \{j \geq 1,\, p_j \geq x\} \right|,$$

on dira que la loi $(p_j)_{j\geq 1}$ est à variation régulière d'indice $\alpha \in [0,1]$ si $\vec{\nu}(1/\cdot)$ l'est, c'est-à-dire, si $\vec{\nu}(1/n) \sim n^{\alpha}\ell(n)$ quand $n \to +\infty$, où ℓ est une fonction à variation lente. Sous cette condition, l'étude asymptotique des variables d'occupation est relativement transparente. La théorie de la variation régulière (initiée par Karamata) permet alors d'obtenir des équivalents simples pour les moments, et fournit un cadre où les variables $K_{n,r}$ vérifient la loi forte des grands nombres et un théorème central-limite.

L'étude du profil $(K_{n,r})_{r\geq 1}$ s'est révélée cruciale en statistiques, notamment dans des cas où la taille de l'échantillon est relativement petite par rapport à celle de l'alphabet, et où les méthodes d'estimation classiques telles que le maximum de vraisemblance sont mises en échec [132]. Les champs d'application sont divers. En écologie, les urnes peuvent par exemple représenter des espèces animales [73]. En linguistique, elles peuvent correspondre à des mots, ou à des suites de mots. En supposant le nombre d'urnes fini égal à m, on peut alors chercher à estimer m à partir de l'échantillon [43, 67]. En alphabet infini, la question correspondante serait de savoir combien de nouveaux symboles on observerait si l'on agrandissait la taille de l'échantillon [83]. En particulier, si l'on dispose d'un échantillon de taille n, quelle est la probabilité que le $(n+1)^{\text{e}}$ symbole soit nouveau ? Se poser cette question, c'est s'intéresser à la *masse manquante*, que l'on notera $M_{n,0}$ et qui est définie comme

$$M_{n,0} \quad = \quad \sum_{j\geq 1} p_j \mathbb{1}_{\{j\notin\{X_1,\ldots,X_n\}\}}\,,$$

la probabilité des symboles qui ne sont pas dans l'échantillon. Un estimateur de la masse manquante dont les qualités sont encore célébrées aujourd'hui a été proposé par Alan Turing, dans le contexte du décryptage du dispositif de chiffrage Enigma, utilisé par la marine allemande durant la Seconde Guerre Mondiale. L'estimateur de Good-Turing [82] consiste à approcher la masse manquante par $K_{n,1}/n$, la proportion des symboles observés une seule fois dans l'échantillon. Cet estimateur peut *a posteriori* est considéré comme un estimateur par ré-échantillonnage, de type *jackknife*.

Dans le Chapitre 1, nous présentons un travail en collaboration avec Stéphane Boucheron et Mesrob Ohannessian, qui fait l'objet d'un article à paraître dans le *Bernoulli Journal* [22]. Nous nous intéressons aux propriétés de concentration (non-asymptotique) des variables d'occupation K_n et $K_{n,r}$ et de la masse manquante $M_{n,0}$. On cherche à obtenir, pour n fixé, des bornes sur la probabilité que ces variables s'éloignent de leur espérance de plus d'un certain seuil $t \geq 0$. Ces variables sont des sommes, pondérées ou non, de variables de Bernoulli, mais celles-ci ne sont pas indépendantes, et c'est là une des difficultés principales. Il y a deux manières de contourner ce problème : la première consiste à recourir à la *Poissonnisation*. On suppose alors que la taille de l'échantillon est elle-même aléatoire, distribuée selon une loi de Poisson de paramètre n. Dans ce scenario poissonnisé, les variables de Bernoulli sont indépendantes. Une deuxième méthode consiste à utiliser *l'association négative* des variables correspondant au nombre de boules dans chaque urne. L'association négative nous permet d'obtenir, pour K_n et $K_{n,\bar{r}}$ (le nombre de symboles apparus plus de r fois), des inégalités de type Bennett, c'est-à-dire une concentration du type de la loi de Poisson, avec un facteur de variance tendu, celui obtenu par l'inégalité d'Efron-Stein-Steele. Obtenir des inégalités de concentration pour $M_{n,0}$ s'avère plus problématique et la concentration de cette variable aléatoire n'est pas garantie. Intuitivement, la masse manquante concentre

bien lorsque les symboles qui n'apparaissent pas dans un échatillon de taille n (et donc contribuent à la masse manquante) ont tous une probabilité bien plus petite que $1/n$. Le cas de la loi géométrique montre cependant que la masse manquante peut fluctuer largement autour de son espérance. Dans ce cas en effet, les symboles appartenant à la masse manquante correspondent *grosso modo* aux symboles plus grands que $j^\star$, le quantile d'ordre $1 - 1/n$. Les fluctuations concernent donc majoritairement les symboles se situant autour de ce quantile. Or la probabilité $p_{j^\star}$ est elle-même de l'ordre de $1/n$, et donc ces symboles, selon qu'ils apparaissent ou non, peuvent faire fluctuer assez fortement la masse manquante. Rechercher une inégalité de concentration dans laquelle le facteur de variance soit universellement tendu semble donc illusoire. L'étude de la transformée de Laplace de la masse manquante nous a cependant conduit à une représentation intéressante, qui est fonction de toute la suite $(\mathbb{E}K_{n,r})_{r\geq 1}$. Contrôler les déviations de la masse manquante revient alors à contrôler uniformément l'espérance des variables d'occupation. Nous obtenons finalement une inégalité de concentration pour la masse manquante valable pour toute loi $(p_j)_{j\geq 1}$, et nous montrons que, si la loi appartient au domaine des lois à variation régulière avec une queue de distribution plus lourde que celle de la loi géométrique, alors le facteur de variance dans cette inégalité a effectivement l'ordre de la vraie variance. Nous appliquons ces résultats à l'étude de l'estimateur de Good-Turing.

Codage adaptatif sur des alphabets infinis

Dans le Chapitre 2, nous présentons un travail en cours avec Stéphane Boucheron et Elisabeth Gassiat. Le cadre général est inchangé : celui d'un échantillon I.I.D. à valeurs dans $\mathbb{N}^*$, et issu d'une loi inconnue $P = (p_j)_{j\geq 1}$. On s'intéresse au problème de l'encodage du message $X_{1:n} = (X_1, \ldots, X_n)$. Les deux grandes difficultés ici sont, d'une part, le fait de ne pas connaître la loi de la source P qui génère le message, et d'autre part, l'infinité de l'alphabet $\mathcal{X} = \mathbb{N}^*$.

L'inégalité de Kraft-McMillan établit une correspondance entre les code uniquement décodable et les probabilités : à toute probabilité Q_n sur $\mathcal{X}^n$, on peut associer un code préfixe (c'est-à-dire tel qu'aucun mot de code n'est le préfixe d'un autre) dont la fonction de longueur ℓ est telle que $\ell(x) \leq \lceil -\log Q_n(x) \rceil$ pour tout $x \in \mathcal{X}^n$. Les méthodes de codage arithmétique permettent de construire un tel code. Notre problème est donc de trouver une loi Q_n qui minimise $\mathbb{E}_P\left[-\log Q_n(X_{1:n})\right]$. L'entropie de la source elle-même étant une borne inférieure pour cette quantité, on va en fait s'intéresser à la différence entre les deux (la *redondance*), qui n'est autre que la distance de Kullback-Leibler (dite aussi entropie relative) entre P^n et Q_n :

$$D(P^n, Q_n) = \mathbb{E}_P\left[\log \frac{P^n(X_{1:n})}{Q_n(X_{1:n})}\right] .$$

Si la source est connue, la probabilité de codage optimale est bien sûr donnée par P^n elle-même. Lorsque l'on sait uniquement que la source appartient à une certaine classe $\mathcal{C}$, on est ramené à un problème de codage dit *universel*. La *redondance minimax* de la classe est définie comme :

$$\overline{R}(\mathcal{C}^n) = \inf_{Q_n} \sup_{P\in\mathcal{C}} D(P^n, Q_n) ,$$

et une suite de distributions de codage (Q_n) est dite fortement universelle si $\dfrac{\sup_{P\in\mathcal{C}} D(P^n, Q_n)}{n} \to 0$ quand $n \to +\infty$. Le cas des alphabets finis a fait l'objet d'une vaste littérature [107, 140, 156]. On sait par exemple que la redondance minimax de la classe des lois I.I.D. sur un alphabet de taille k est équivalente, quand n tend vers l'infini, à $\frac{k-1}{2}\log n$. Lorsque l'alphabet est infini, le théorème de Kieffer implique qu'il n'existe pas de code universel pour la classe des lois I.I.D.. On doit donc restreindre nos ambitions, et considérer des classes plus petites. Nous nous placerons dans le cadre du *codage adaptatif* : étant donné

une collection de classes de sources, telle que l'on dispose, pour chaque classe, d'un codeur universel, peut-on construire un unique code qui atteint la redondance minimax sur chaque classe ? Formellement, une suite de distributions de codage (Q_n) est dite adaptative sur la collection de classes $(\mathcal{C}(\alpha))_{\alpha\in A}$ si, pour tout $\alpha \in A$,

$$\sup_{P\in\mathcal{C}(\alpha)} D(P^n, Q_n) = (1 + o_\alpha(1))\overline{R}(\mathcal{C}(\alpha)^n)\,.$$

Si la collection de classes est très vaste, l'adaptivité peut s'avérer très difficile, sinon impossible à obtenir. On dira qu'une suite de distributions de codage (Q_n) est adaptative à un facteur r_n près si pour tout $\alpha \in A$,

$$\sup_{P\in\mathcal{C}(\alpha)} D(P^n, Q_n) = O_\alpha(r_n)\overline{R}(\mathcal{C}(\alpha)^n)\,.$$

Lorsque la suite (r_n) croît bien plus lentement que $\overline{R}(\mathcal{C}(\alpha)^n)$, ce critère, moins contraignant, reste significatif. Nous nous intéresserons ici à des classes de loi dite *enveloppes*, introduites par [39], et définies comme l'ensemble des lois de probabilités dominées par une fonction d'enveloppe f. Si l'on note Λ_f la classe-enveloppe induite par f, un premier problème est de comprendre le comportement asymptotique de $\overline{R}(\Lambda_f^n)$. En utilisant des arguments de Poissonnisation proposés par Acharya et al. [4], nous obtenons des bornes inférieures et supérieures sur $\overline{R}(\Lambda_f^n)$, et l'introduction du formalisme de Karlin [102] nous permet d'interpréter ces bornes de façon transparente. Dans le cas particulier où la fonction f vérifie une condition de variation régulière, le théorème de Karamata et ses extensions permettent une analyse asymptotique immédiate. Notre contribution principale est la construction d'un code (Q_n), qui est adaptatif (à un facteur $\log\log n$ près) sur la famille des classes-enveloppe à variation régulière : pour tout $\alpha \in [0, 1[$, et pour toute fonction enveloppe f telle que $\vec{\nu}_f(1/n) \sim n^\alpha \ell(n)$, il existe $c_\alpha > 0$ tel que

$$\sup_{P\in\Lambda_f^n} D(P^n, Q_n) \leq (c_\alpha + o_\alpha(1)) \log\log n \overline{R}(\Lambda_f^n)\,.$$

Le code que nous proposons est largement inspiré des codes de type *auto-censurant*, tel que le code AC de Bontemps et al. [35] ou le code ETAC de Boucheron et al. [41]. Il s'agit d'un code préfixe *en ligne* dont le principe est le suivant : on code les symboles du message $X_{1:n}$ séquentiellement. Au temps $i \in \{1, \ldots, n\}$, si X_i est un nouveau symbole, c'est-à-dire s'il n'a pas déjà été observé dans $X_{1:i-1}$, alors on utilise le code d'Elias pour l'encoder et on intègre ce symbole au *dictionnaire courant*. Si X_i a déjà été observé, alors on l'encode par le mélange de Krichevski-Trofimov (KT) sur l'alphabet fini correspondant au dictionnaire courant, c'est-à-dire l'ensemble des symboles déjà observés (le code KT réalise une estimation bayésienne de la loi de la source, et le choix du prior de Jeffrey permet l'optimisation de la redondance). Bien sûr, pour que ce code soit bien décodable, il faut indiquer au décodeur lequel de ces deux codes est utilisé. Pour cela, on utilise un symbole additionnel, noté 0, et les nouveaux symboles sont en fait codés deux fois : une fois comme un 0 par KT, et une deuxième fois par Elias. De cette manière, si le décodage par KT d'un mot de code donne un 0, on sait que le prochain mot de code correspond au code Elias d'un nouveau symbole.

Ce code peut être compris comme un mélange entre d'une part le codage du dictionnaire, et d'autre part, le codage du *pattern* du message [78, 129]. Cependant, une particularité de notre analyse est de ne pas séparer les deux contributions à la redondance globale du code. Au contraire, une partie de la redondance due au codage KT nous aide à compenser une partie de l'encodage, par le code Elias, des nouveaux symboles. C'est ce qui nous permet d'obtenir l'adaptivité à un facteur $\log\log n$, au lieu d'un facteur $\log n$. Une question en suspens est de savoir si l'on peut se passer de ce facteur $\log\log n$, ou s'il y a un coût inévitable à l'adaptivité.

Marches aléatoires sur graphes aléatoires

Dans la deuxième partie de ce texte, je m'intéresse à un problème complètement différent qui est celui du temps de mélange de marches aléatoires sur des grands graphes aléatoires.

Pour motiver l'étude du temps de mélange de chaînes de Markov, on peut considérer le problème d'échantillonnage suivant. Soit $G = (V, E)$ un graphe connexe sur n sommets. Supposons que l'on veuille tirer un sommet dans ce graphe avec probabilité proportionnelle à son degré (qui peut être considéré comme une mesure de l'importance de ce sommet dans le réseau). Si n est très grand, ce problème peut s'avérer difficile, en particulier si l'on n'a qu'une connaissance locale et incomplète du réseau. Une solution approchée consiste à faire partir une marche aléatoire d'un sommet fixé $x \in V$. La marche va progressivement oublier son point de départ, se perdre dans le réseau, si bien qu'au bout d'un certain temps, elle se trouvera sur un sommet v avec probabilité très proche de $\frac{\deg(v)}{2|E|}$. On dit qu'elle a *mélangé*. Si l'on sait contrôler le *temps de mélange* de la marche aléatoire sur G, on dispose alors d'une méthode à la fois élégante théoriquement et bien souvent très efficace en pratique, pour échantillonner selon la probabilité voulue, appelée probabilité stationnaire de la chaîne. C'est le principe général des méthodes MCMC (*Monte Carlo Markov Chains*).

Pour une chaîne de Markov irréductible et apériodique sur un espace d'état fini Ω_n, on définit, pour tout ε de $]0, 1[$, le temps de mélange $t_{\text{MIX}}^{(n)}(\varepsilon)$ comme

$$t_{\text{MIX}}^{(n)}(\varepsilon) = \inf\{t \geq 0,\, \mathcal{D}_n(t) \leq \varepsilon\}\,,$$

où $\mathcal{D}_n(t)$ correspond au maximum, sur tous les points de départ $x \in \Omega_n$, de la distance en variation totale entre la loi de la chaîne au temps t partie de x et la loi stationnaire. Si l'on dispose d'une suite de chaînes de Markov sur une suite d'espaces d'état (Ω_n), on peut alors s'intéresser au comportement asymptotique de $t_{\text{MIX}}^{(n)}(\varepsilon)$ lorsque n tend vers l'infini. Au début des années 1980, Diaconis and Shahshahani [60] et Aldous [6] ont découvert un phénomène surprenant, appelé le *phénomène de cutoff* : le premier terme du développement asymptotique de $t_{\text{MIX}}^{(n)}(\varepsilon)$ peut ne pas dépendre de ε. Cela signifie qu'asymptotiquement, la distance chute abruptement de 1 à 0, en une période de temps négligeable devant le temps de mélange lui-même et que l'on appelle la fenêtre du cutoff.

Dans la Section 2, nous présentons un travail effectué avec Justin Salez, qui a fait l'objet d'un article à paraître dans les Annales de Probabilités [21]. Nous nous intéressons à la marche sans rebroussement (« non-backtracking » en anglais) sur des graphes aléatoires à degrés prescrits, générés selon le modèle de configuration. On se fixe une suite de degrés $(\deg(v))_{v\in V}$ telle que $N = \sum_{v\in V} \deg(v)$ est pair. Initialement, chaque sommet v de V est muni de $\deg(v)$ *demi-arêtes*, et l'on choisit uniformément au hasard un appariement de ces demi-arêtes, générant ainsi un graphe aléatoire (qui n'est pas nécéssairement simple) où chaque arête correspond à l'appariement de deux demi-arêtes. La marche aléatoire sans rebroussement est alors définie comme une chaîne de Markov sur l'ensemble des demi-arêtes. Sous certaines hypothèses sur la suite de degrés, nous montrons qu'avec probabilité tendant vers 1, la marche aléatoire sans rebroussement présente le phénomène de cutoff. Nous déterminons le temps de mélange $t_\star$ et la fenêtre $\omega_\star$ et nous établissons la convergence suivante :

$$\frac{t_{\text{MIX}}(\varepsilon) - t_\star}{\omega_\star} \xrightarrow{\mathbb{P}} \Phi^{-1}(\varepsilon) \quad \text{quand } n \to +\infty\,, \tag{0.1}$$

où Φ est la fonction de queue d'une variable gaussienne centrée réduite. Si l'on pose

$$\mu = \frac{1}{N} \sum_{v\in V} \deg(v) \log(\deg(v) - 1)\,,$$

et

$$\sigma^2 = \frac{1}{N} \sum_{v \in V} \deg(v) \left(\log(\deg(v) - 1) - \mu\right)^2 ,$$

alors on a : $t_\star = \frac{\log N}{\mu}$ et $\omega_\star = \sqrt{\frac{\sigma^2 \log N}{\mu^3}}$. La convergence (0.1) correspond donc à un résultat d'universalité : pour toute suite de degrés (qui vérifient certaines conditions dites de *sparsité*), le profil de la distance à l'équilibre ne dépend des degrés qu'à travers μ et σ^2. Prise en des temps de la forme $t_\star + \lambda\omega_\star$, la distance converge vers une courbe universelle, celle de la fonction de queue gaussienne $\Phi(\lambda)$. La preuve de ce résultat repose sur une analyse très fine de la variable aléatoire $P^t(x, y)$, la probabilité pour la marche sans rebroussement partie de x d'être en y au temps t. On cherche à déterminer le temps t à partir duquel cette variable concentre autour de son espérance, $1/N$. Dans le cas d-régulier analysé dans [113], le problème revient à contrôler le nombre de chemins allant de x à y. Ici, on doit compter ces chemins de façon pondérée : en effet, les degrés sont hétérogènes et différents chemins de même longueur peuvent avoir des probabilités bien différentes d'être empruntés par la marche. En utilisant une propriété de pseudo-réversibilité de la marche sans rebroussement, on représente $P^t(x, y)$ comme une somme pondérée de variables de Bernoulli faiblement dépendantes :

$$P^t(x, y) = \sum_{i,j} \mathbf{w}(i)\mathbf{w}(j) X_{i,j} , \tag{0.2}$$

où $\mathbf{w}(i)$ (resp. $\mathbf{w}(j)$) correspond à la probabilité que la marche partie de x soit en i au temps $t/2$ (resp. que la marche partie de y soit en j au temps $t/2$), et où $X_{i,j}$ correspond à l'indicatrice de l'évènement « les demi-arêtes i et j sont appariées ». Pour que cette somme concentre, il faut s'assurer qu'aucune des variables de Bernoulli n'a trop d'impact sur la somme totale, et donc, qu'aucun des poids $\mathbf{w}(i)\mathbf{w}(j)$ n'est trop grand. Intuitivement, pour t trop petit, la variable $P^t(x, y)$ n'est pas concentrée, car les chemins de longueur $t/2$ autour de x et de y ont encore un « trop grand » poids. Si l'on augmente t progressivement, les chemins de poids petit deviennent de plus en plus nombreux, jusqu'à couvrir quasiment toute la masse des chemins possibles, et cette transition a lieu très abruptement. Pour (presque) toutes les paires (x, y), et pour un seuil $\theta \approx 1/N$, la masse $\sum_{i,j} \mathbf{w}(i)\mathbf{w}(j)\mathbb{1}_{\mathbf{w}(i)\mathbf{w}(j)>\theta}$ chute abruptement de 1 à 0 au moment $t_\star$ et en une période bien plus petite $\omega_\star$. On peut décrire précisément la forme de cette chute : pour $t = t_\star + \lambda\omega_\star$, la masse $\sum_{i,j} \mathbf{w}(i)\mathbf{w}(j)\mathbb{1}_{\mathbf{w}(i)\mathbf{w}(j)>\theta}$ est très proche de $\Phi(\lambda)$.

Notons aussi qu'une des difficultés de la preuve réside dans le fait que les variables d'appariement $X_{i,j}$ dans (0.2) ne sont pas indépendantes. Grâce à la méthode des paires échangeables (une des variantes de la méthode de Stein), nous obtenons une inégalité de concentration pour des sommes indéxées par un appariement aléatoire.

L'utilisation de marches sans rebroussement semble assez naturelle. D'une certaine façon, en supprimant le « bruit » lié aux retours en arrière de la marche simple, elles sont plus en adéquation avec la structure du graphe du lui-même. D'un point de vue plus pratique, on peut supposer qu'elles permettent une exploration plus rapide et plus efficace du graphe. Les marches sans rebroussement mélangent-elles plus vite ? Dans le cas de graphes aléatoires d-réguliers, l'avantage de la marche sans rebroussement sur la marche simple a été confirmé par Lubetzky and Sly [113], qui ont montré que la marche simple mélangeait moins vite que la marche sans rebroussement. Dans ce cas, le rapport entre les deux temps de mélange est précisément donné par $\frac{d-2}{d}$, qui correspond à la vitesse d'une marche simple dans un arbre d-régulier. Cette comparaison est-elle toujours valable dans le cas non-régulier ? Remarquons tout d'abord que la réponse n'est pas immédiate. Certes la marche sans rebroussement est toujours avantagée par sa plus grande vitesse, mais l'hétérogénéité des degrés introduit un deuxième effet allant dans le sens contraire : lorsqu'elle entre dans une partie du graphe où les degrés sont relativement petits, elle y est piégée, alors que, sur un sommet de petit degré, la marche simple a relativement plus de chance de reculer. La marche

simple a donc tendance à naturellement quitter les chemins de petits degrés et à se diriger vers les sommets de plus grands degrés, auxquels la mesure stationnaire donne plus de poids. Dans la Section 3 de la deuxième partie, je montre que, même dans le cas non-régulier, la marche sans rebroussement mélange plus vite que la marche simple. La preuve repose sur une comparaison délicate de l'entropie des deux marches sur un arbre de Galton-Watson.

Tirages pondérés sans remise

Dans la troisième partie de ce texte, nous considérons un problème encore bien différent : celui de la comparaison entre des tirages avec et sans remise, lorsque les éléments sont munis de poids, et que l'on tire proportionnellement à ces poids. Il s'agit d'un travail en collaboration avec Yuval Peres et Justin Salez, soumis à un journal [23]. On considère une population finie de taille N. A chaque élément de la population sont associés une valeur $\nu(i) \in \mathbb{R}$, et un poids $\omega(i) > 0$, tel que $\sum_{i=1}^{N} \omega(i) = 1$. On s'intéresse aux propriétés de concentration de la variable

$$X \quad = \quad \nu(\mathbf{I}_1) + \cdots + \nu(\mathbf{I}_n)\,,$$

où, pour tout n-uplet $(i_1, \ldots, i_n)$ d'indices distincts de $\{1, \ldots, N\}$,

$$\mathbb{P}\left((\mathbf{I}_1, \ldots, \mathbf{I}_n) = (i_1, \ldots, i_n)\right) \quad = \quad \prod_{k=1}^{n} \frac{\omega(i_k)}{1 - \omega(i_1) - \cdots - \omega(i_{k-1})}\,.$$

On dit que $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ correspond à un échantillon pondéré sans remise. En tirant avec remise dans la population, on obtient un échantillon I.I.D. $(\mathbf{J}_1, \ldots, \mathbf{J}_n)$, où, pour tout $j \in \{1, \ldots, N\}$, $\mathbb{P}(\mathbf{J}_1 = j) = \omega(j)$. On définit alors

$$Y \quad = \quad \nu(\mathbf{J}_1) + \cdots + \nu(\mathbf{J}_n)\,.$$

La variable Y est une somme de variables aléatoires I.I.D., et la méthode de Chernoff nous permet d'obtenir des inégalités de concentration en majorant le logarithme de sa transformée de Laplace.

Dans le cas particulier de poids uniformes, $\omega(1) = \cdots = \omega(N) = \frac{1}{N}$, Hoeffding [92] a montré que, pour toute fonction convexe $f\colon \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[f(X)] \quad \leq \quad \mathbb{E}[f(Y)]\,.$$

On dit que X est plus petite que Y *en ordre convexe*. Peut-on obtenir une comparaison similaire dans le cas de poids non-uniformes ? Si les valeurs et les poids sont classés dans le même ordre, c'est-à-dire si pour tout $(i, j) \in \{1, \ldots, N\}^2$,

$$\omega(i) \geq \omega(j) \iff \nu(i) \geq \nu(j)\,,$$

alors, on montre que X est plus petite que Y en ordre convexe croissant : pour toute fonction convexe croissante $f\colon \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[f(X)] \quad \leq \quad \mathbb{E}[f(Y)]\,. \tag{0.3}$$

En particulier, pour tout $\lambda \geq 0$, $\mathbb{E}[\mathrm{e}^{\lambda X}] \leq \mathbb{E}[\mathrm{e}^{\lambda Y}]$, et toutes les bornes sur $\mathbb{P}(Y > t)$ obtenues par la méthode de Chernoff s'appliquent sans modification à X. Nous montrons comment une telle comparaison permet de contrôler simplement le nombre d'arêtes révélées lorsque l'on expose progressivement le

voisinage d'un sommet dans un graphe aléatoire distribué selon le modèle de configuration.

On peut remarquer que, dans le cas non-uniforme, $\mathbb{E}X$ et $\mathbb{E}Y$ ne sont pas nécéssairement égales. L'inégalité (0.3) ne permet donc pas d'obtenir une inégalité de concentration de X *autour de son espérance*. De plus, lorsque le nombre de tirages n se rapproche du nombre total d'éléments N, la comparaison directe avec Y ne capturera sûrement pas le bon ordre des fluctuations. En effet, la variance de Y est de l'ordre de n, alors que celle de X est plutôt de l'ordre de $n \wedge (N-n)$. Par exemple, dans le cas extrême où $n = N$, la variance de X est nulle. Lorsque les poids sont uniformes, Serfling [145] a capturé ce bon ordre de grandeur pour la variance dans l'inégalité de concentration suivante :

$$\mathbb{P}\left(X - \mathbb{E}X > t\right) \quad \leq \quad \exp\left(-\frac{2t^2}{n\left(1-\frac{n-1}{N}\right)\Delta^2}\right),$$

où $\Delta = \max\limits_{1\leq i\leq N} \nu(i) - \min\limits_{1\leq i\leq N} \nu(i)$. Nous avons tenté d'obtenir une inégalité similaire dans le cas de poids non-uniformes. Si l'on note $\alpha = \dfrac{\min_{1\leq i\leq N}\omega(i)}{\max_{1\leq i\leq N}\omega(i)}$, on montre que, pour $\alpha < 1$, et pour tout $t > 0$,

$$\max\left\{\mathbb{P}\left(X - \mathbb{E}X > t\right), \mathbb{P}\left(X - \mathbb{E}X < -t\right)\right\} \leq \exp\left(-\frac{t^2}{2v}\right),$$

où

$$v \quad = \quad \min\left(4\Delta^2 n\,,\ \frac{1+4\alpha}{\alpha(1-\alpha)}\Delta^2 N\left(\frac{N-n}{N}\right)^{\alpha}\right).$$

Si α est uniformément borné loin de 0 et 1, on obtient donc une inégalité de concentration sous-gaussienne pour X, dans laquelle le facteur de variance est soit de l'ordre de n (ce qui est le bon ordre pour des échantillons de taille $n \leq qN$, avec q fixé dans $]0,1[$), soit de l'ordre de $N\left(\frac{N-n}{N}\right)^{\alpha}$, ce qui donne une amélioration lorsque $\frac{n}{N} \to 1$. Il serait intéressant de savoir si l'on peut se passer de l'exposant α, ce qui nous permettrait d'obtenir un équivalent de l'inégalité de Serfling pour des poids non-uniformes.

Part I

Concentration and compression on infinite alphabets

In this part, we consider various problems related to sampling over infinite countable alphabets. The general setting is often referred to as an infinite urn scheme*: one may picture n balls being independently thrown onto an infinite collection of urns, each ball having probability p_j to fall in urn j. We start with a general introduction. In Chapter 1, we are interested in the concentration properties of various functions of the sample: the number of occupied urns, the number of urns containing r balls, and the missing mass. This is a work in collaboration with Stéphane Boucheron and Mesrob I. Ohannessian,* Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications *[22], to appear in the Bernoulli Journal. Chapter 2 is devoted to the different although related problem of encoding such a sample with as few bits as possible without knowing the underlying distribution. This is an ongoing work with Stéphane Boucheron and Elisabeth Gassiat.*

Introduction

1 Sampling from an unknown discrete distribution

In this part, we are concerned with inference from I.I.D. samples of an unknown discrete distribution (say a distribution over a finite or countable set which elements will be referred as symbols). Given such samples, what can we infer about the distribution? In particular, can we use the sample to learn something about the unobserved portion of the distribution? Those questions have critical importance when the sample size is small with respect to alphabet size, which is the case in many applications. In those situations, classical methods such as maximum-likelihood estimation typically fail to capture *any* information about unseen symbols. Consider the following examples: first, assume that, in a sample of size 100, we observe 40 times symbol a, and 60 times symbol b. A trivial answer, based on empirical frequencies, would be to assign probability $2/5$ to symbol a, probability $3/5$ to symbol b, and 0 to any other symbol. One can show that this is the maximum-likelihood estimator (MLE) of the distribution. In cases where the number of occurrences of each symbol is approximately linear in the sample size, this estimation procedure performs well. But assume now that we observe a sequence $(a_1, a_2, \ldots, a_n)$ where for all $i \neq j$, $a_i \neq a_j$. The maximum-likelihood estimator (empirical frequencies) is the uniform distribution over n symbols, and hence completely misses one of the essential feature of the sample: all symbols are distinct. What would seem likely is that, if we sample another time, we will discover yet another distinct symbol, an event to which the MLE assigns probability zero.

The problem of estimating an unknown discrete distribution from a small sample has a long history which can be traced back to the pioneer works of Ronan Fisher and Alan Turing. In the early 1940's, a naturalist named Corbet, coming back from a two-year expedition in Malaysia during which he could collect a sample of butterflies, approached Fisher with the following question : knowing that the sample contains n_1 butterfly species observed only once, n_2 species observed twice, and so on, is it possible to estimate the number of new species that could be discovered if another two-year expedition were to be launched? The model proposed by Fisher et al. [73] gives a general framework to address this problem. Around the same time, Alan Turing and Irving Good were working on a related problem, in an effort to break the Enigma, a encryption device used by the German Navy during World War II. In order to estimate the probability for the next cypher to be a new one (the *missing mass* of the sample), they constructed a simple estimation scheme, known as the Good-Turing estimator, corresponding to the proportion of symbols occurring only once in the sample [82].

Those works initiated a long line of research, which benefited recently from major contributions of the theoretical computer scientists. One of the central question is: how well can we estimate the distribution itself? This turns out to be a special case of density estimation [57, 58]. A common (and pessimistic) way to address the problem is through the *minimax* approach: we assume that the true distribution P lies in some class $\mathcal{C}$, and we aim at constructing a distribution Q_n, based on the observations, which would perform as well as possible (in terms of a given loss function), were the distribution P chosen adversarially as the worst distribution in the class. Even in the simplest setting where $\mathcal{C}$ is the class

of distributions over a fixed-size alphabet, this problem is not fully understood. There exist a variety of estimators (Laplace or add-one estimators, Krichevsky-Trofimov or add-one-half estimators...) and their performance intimately depends one the chosen loss function [89, 101]. One may also be interested into estimating a functional of the sampling distribution, such as its entropy, or the cardinality of its support. The specific problem of estimating the support size of a distribution arises in a wide variety of fields, such as ecology where one is interested in the number of species in a population [83], linguistics where one looks for an estimate on the number of words in a vocabulary [67], and even numismatics [71]. Bunge and Fitzpatrick [43] gives a general review of the literature on support size estimation. The estimation of symmetric functionals of the distribution (that is, functionals which do not depend on a specific labelling of symbols) has received a lot of attention in the theoretical computer science community [11, 17, 90, 96, 153]. Typically, one is interested in estimating quantities of the form $\sum f(p_i)$, one of the most studied symmetric functional being the entropy of the distribution $-\sum p_i \log p_i$. Here again, plugging in the maximum likelihood estimator can be highly sub-optimal. Let us also mention the problem of *testing* properties of the distribution: is it uniform? Is it monotone? Log-concave? Unimodular? Of particular importance in a small sample regime is the question of the *sample complexity*, that is the number of observations, necessary or sufficient, to test those properties with a small error probability [5, 18, 109].

As highlighted by Orlitsky et al. [132], in situations of under-represented data, one might considerably benefit from modelling the *profile* of the sample, instead of the number of occurrences of each individual symbol. The profile of the sample is the sequence of occupancy counts, defined as the number of symbols occurring once, the number of symbols occurring twice, three times, and so on. It is sometimes called the *histogram of histograms*, or the *fingerprint* of the sample. Let us go back to the example of the beginning. Assume that we observe the sequence $(2, 1, 3)$. The MLE of the distribution is thus the uniform distribution over $\{1, 2, 3\}$. However, let us try to take into account more specifically the profile of the sample: three symbols occur once. Under the uniform distribution over $\{1, 2, 3\}$, the probability of this event is

$$3! \times \frac{1}{3^3} = \frac{2}{9}\,.$$

However, if instead we choose the uniform on, say, $\{1, \dots, 5\}$, we would find

$$\frac{5!}{(5-3)!} \times \frac{1}{5^3} = \frac{12}{25} > \frac{2}{9}\,.$$

Orlitsky et al. [132] then propose a new estimate for the distribution, called the *high profile*, based on the maximization of the probability of the observed profile. This line of work turned out to be decisive. Although computing the high profile of a sample is often a deterring task, focusing on the profile prompted great progress related to the various questions mentioned above. For instance, the profile contains all the useful information about symmetric properties of a distribution, and the algorithm designed by Valiant and Valiant [153] to estimate symmetric functionals largely relies on the sample profile. This algorithm actually returns a distribution which generates, with high probability, a sample whose profile closely matches the one of the observed sample.

Intuitively, in a sample of size n, one would like to distinguish between "frequent" and "infrequent" symbols. Frequent symbols' probabilities could be estimated by maximum-likelihood, or by variants of the MLE, such as add-constant estimators, which give to symbol j a probability proportional to $X_{n,j} + \beta$, where $X_{n,j}$ is the number of occurrences of j and β is a constant term. Those estimators stem from Bayesian estimation with prior Dirichlet$(\beta, \dots, \beta)$. The most popular are Laplace estimators ($\beta = 1$) and Krichevsky-Trofimov ($\beta = 1/2$), which are justified by their asymptotic properties over finite alphabets. However, as observed in several places, such estimators may perform very poorly for infrequent symbols.

To cope with those rare symbols, one may rather resort to Good-Turing frequency estimation, whose principle is as follows: let $K_{n,r}$ be the number of symbols occurring r times in the samples. Good and Turing proposed to estimate the cumulative probability of symbols occurring r times by $\frac{(r+1)K_{n,r+1}}{n}$ (in particular, $K_{n,1}/n$ corresponds to the Good-Turing estimator of the missing mass). Using the fact that the mass of symbols occurring r times should be almost equally distributed among those symbols, this yields the following estimator for the probability of a symbol j occurring r times:

$$\hat{p}_{\mathrm{GT}}(j) \quad = \quad \frac{(r+1)K_{n,r+1}}{nK_{n,r}} \, .$$

This suggests the following estimation procedure: choose a threshold $r^\star$. If symbol j occurs more than $r^\star$ times, estimate its probability by its empirical frequency (possibly with an additive term β). If it occurs less than $r^\star$ times, resort to Good-Turing estimation. The problem then becomes that of choosing an optimal threshold $r^\star$ [130, 150, 153].

Let us mention that, in a Bayesian perspective, another approach is to choose a prior more adapted to rare-event regimes than the Dirichlet priors resulting in add-constant estimators. Natural languages, for instance, are well modelled by power-laws, and the choice of a Poisson-Dirichlet prior (Pitman-Yor process [138]) in n-gram models has been advocated by [152].

The problem of estimating rare probabilities and of understanding the behaviour of the profile $(K_{n,r})_{r\geq 1}$ is also deeply related to data compression. If the sampling distribution were known, the compression problem would be solved by applying a standard method as Huffmann coding or, if we want to encode and decode sequences of symbols in an on-line way, by feeding the symbols to an arithmetic coder fitted to the sampling distribution [51]. If the sampling distribution is not known, we face a universal coding problem and data compression turns out to be intimately related to estimating the probability of the sample under the so-called logarithmic loss [44, 45, 51, 52]. Given this strong duality between compression and probability estimation, it is not surprising that the sample's profile plays a crucial role in universal coding [129, 133]. When the alphabet is infinite, Kieffer's Theorem [104] implies that there is no universal code for I.I.D. sequences. This negative result prompted several approaches, one of which consists in encoding the *pattern* of the sample. For instance the pattern of the sequence $(7, 7, 4, 2, 6, 6, 4)$ is $(1, 1, 2, 3, 4, 4, 2)$, each symbol being replaced by its rank of appearance in the sample. We then neglect the particular labelling of symbols [1, 78, 147]. Encoding a sequence may then be done in two steps: encoding of the pattern and encoding of the dictionary. In this framework, the estimation of the probability that the next symbol is new, *i.e.* of the missing mass, turns out to be pivotal, and the Good-Turing estimator has received a lot of attention in information theory [130, 131].

The asymptotic properties of the Good-Turing estimator have been intensively analysed [70, 76, 122]. However, in our small-sample perspective, it seems relevant to look for *non-asymptotic* results. The interest of the statistical learning community for the missing mass problem brought about a new perspective on the question, that of concentration inequalities [25, 26, 121, 127]. The problems addressed in [22] and presented in Chapter 1 pertain to this line of investigation. To study the Good-Turing estimator or other quantities that depend significantly on the small-count portion of the observations, we need to understand the missing mass and the profile well. Our contribution here is to give sharp concentration inequalities with explicit and reasonable constants. Those inequalities are *distribution-free*, that is, they hold for any sampling distribution. For occupancy counts (the number of symbols occurring r times), we establish Bennett's inequalities, in which the variance factor is given the Efron-Stein proxy, which is typically a tight proxy, capturing the right order of the variance. In the case of the missing mass (denoted $M_{n,0}$), we establish a sub-Gaussian inequality on the left tail, with the tight variance proxy $v_n^- = 2\mathbb{E}[K_2(n)]/n^2$, where $K_2(n)$ is the number of symbol occurring twice in a sample of random size distributed as a Poisson variable with mean n. On the right tail, we establish a Bernstein inequality,

with scale factor $1/n$ and variance proxy $v_n^+ = 2\mathbb{E}[K_{\bar{2}}(n)]/n^2$, where $K_{\bar{2}}(n)$ is the number of symbols occurring more than twice in a sample of size Poisson(n). That is to say, we show that, for all $t > 0$,

$$\mathbb{P}\left(M_{n,0} - \mathbb{E}M_{n,0} < -t\right) \quad \leq \quad \exp\left(-\frac{t^2}{2v_n^-}\right),$$

and

$$\mathbb{P}\left(M_{n,0} - \mathbb{E}M_{n,0} > t\right) \quad \leq \quad \exp\left(-\frac{t^2}{2(v_n^+ + t/n)}\right).$$

This variance proxy v_n^+ can significantly improve on the previously established variance factor $1/n$ [25, 121]. Moreover, we exhibit a vast domain of distributions (namely, distributions with a heavier tail than the Geometric) for which v_n^+ actually captures the order of the true variance. One novelty of our analysis is the consideration of the fundamentally Poissonian behaviour of the variables involved. Indeed, as sums of (weighted) Bernoulli variables, the missing mass and occupancy counts typically have Poissonian deviations and it seems relevant to abandon the pursuit of a sub-Gaussian inequality. Instead, looking for sub-Poissonian (Bennett-type) inequalities allows us to significantly improve the variance factor, to the price of a small scale-factor $1/n$. To see how one may largely benefit from Bennett's inequalities, the case of the binomial distribution with parameters n and $1/n$ is illuminating. As the log-Sobolev constant of a Bernoulli random variable with mean $1/n$ is $\frac{\log(n-1)}{1-2/n}$, the best variance factor one could hope for in a sub-Gaussian inequality is $\frac{n(1-2/n)}{2\log(n-1)} \sim \frac{n}{2\log n}$, whereas the true variance is $1 - 1/n$. The appropriate deviations are better captured by a Bennett inequality with variance proxy 1 and scale factor 1 (inducing a Bernstein inequality with variance factor 1 and scale factor 1/3). We then apply our concentration inequalities to the study of the Good-Turing estimator.

In Chapter 2, we address the problem of universal compression over infinite alphabets. As mentioned earlier, when coping with large and possibly infinite alphabets, the problem of rare counts and probabilities becomes critical to universal compression [133]. The two problems of estimating the missing mass and coding over infinite alphabets both face the same kind of challenge : universality is not achievable. The class of all discrete distributions is too large, in the sense that there is no universally consistent estimator of the missing mass [123], and there is no universal code for this class [104]. One thus has to consider smaller classes over which universality can be hoped for. Here, we will consider classes of distributions characterized by regularly varying tails, a framework related to extreme value theory [20, 55]. Following [35, 39, 41], we consider classes of sources defined by a common dominating envelope. We give new bounds on the minimax redundancy of envelope classes, and we construct a simple coding scheme which is (almost) adaptive over the collection of classes characterized by a regularly varying envelope. As suspected, the analysis of this code intimately relies on the understanding of occupancy counts and of the probability to discover a new symbol at any given time.

Let us now introduce some notation which will be used in both Chapter 1 and 2.

2 Occupancy counts and occupancy masses

Let $X_1, X_2, \cdots, X_n$ be I.I.D. observations from a fixed but unknown distribution $(p_j)_{j=1}^{\infty}$ over a discrete set of symbols (an *alphabet*) $\mathcal{X} = \mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. We assume that for all $j \geq 1$, the probability p_j is strictly positive. The terminology of *infinite urn scheme* comes from the analogy to n independent throws of balls over an infinity of urns, p_j being the probability of a ball falling into urn j, at any i-th throw. The sample size n may be fixed in advance; we call this the *binomial setting*. In the continuous time version of this setting, the sample size is itself random and distributed as a Poisson variable N with

mean n, independent of (X_i). This is the *Poisson setting*. We write all Poisson-setting quantities with functional notations, instead of subscripts used for the fixed-n scheme.

For each $j, n \in \mathbb{N}_+$, let $X_{n,j} = \sum_{i=1}^{n} \mathbb{1}_{\{X_i=j\}}$ be the number of times symbol j occurs in a sample of size n, and $X_j(n) = \sum_{i=1}^{N} \mathbb{1}_{\{X_i=j\}}$ the Poisson version. One useful property of Poisson sampling is that the variables $(X_j(n))_{j\geq 1}$ are independent. The collection $(X_{n,j})_{j\geq 1}$ is often called the *type* of the sample $(X_1, \ldots, X_n)$. In questions of under-represented data, the central objects are sets of symbols that are repeated a small number r of times. The central quantities are the *occupancy counts* $K_{n,r}$ [respectively $K_r(n)$ for the Poisson setting], defined as the number of symbols that appear exactly r times in a sample of size n:

$$K_{n,r} = \sum_{j=1}^{\infty} \mathbb{1}_{\{X_{n,j}=r\}}.$$

The collection $(K_{n,r})_{r\geq 1}$ [resp. $(K_r(n))_{r\geq 1}$] has been given many names, such as the *profile* (in information theory [133]) or the *fingerprint* (in theoretical computer science [153]) of the sample. Here we refer to them by *occupancy counts* individually, and *occupancy process* all together.

The occupancy counts then combine to yield the cumulated occupancy counts $K_{n,\overline{r}}$ [respectively $K_{\overline{r}}(n)$] and the total number of distinct symbols in the sample, or the total number of occupied urns, often called the *coverage* and denoted by K_n [respectively $K(n)$]:

$$K_{n,\overline{r}} = \sum_{j=1}^{\infty} \mathbb{1}_{\{X_{n,j}\geq r\}} = \sum_{s\geq r} K_{n,s}\,,$$

and

$$K_n = \sum_{j=1}^{\infty} \mathbb{1}_{\{X_{n,j}>0\}} = \sum_{r\geq 1} K_{n,r}\,.$$

In addition to the occupancy numbers and the number of distinct symbols, we also address the *rare* (or small-count) *probabilities* $M_{n,r}$ [respectively $M_r(t)$], defined as the probability mass corresponding to all symbols that appear exactly r times:

$$M_{n,r} = \mathbb{P}(\{j, X_{n,j} = r\}) = \sum_{j=1}^{\infty} p_j \mathbb{1}_{\{X_{n,j}=r\}}\,.$$

In particular, we focus on $M_{n,0} = \sum_{j=1}^{\infty} p_j \mathbb{1}_{\{X_{n,j}=0\}}$, which is called the *missing mass*, and which corresponds to the probability of all the unseen symbols.

Explicit formulas for the moments of the occupancy counts and masses can be derived in the binomial and Poisson settings. The occupancy counts' expectations are given by

$$\mathbb{E}K_n = \sum_{j=1}^{\infty}(1-(1-p_j)^n) \qquad \mathbb{E}K(n) = \sum_{j=1}^{\infty}(1-\mathrm{e}^{-np_j})$$

$$\mathbb{E}K_{n,r} = \sum_{j=1}^{\infty}\binom{n}{r}p_j^r(1-p_j)^{n-r} \qquad \mathbb{E}K_r(n) = \sum_{j=1}^{\infty}\mathrm{e}^{-np_j}\frac{(np_j)^r}{r!}$$

$$\mathbb{E}M_{n,r} = \sum_{j=1}^{\infty}\binom{n}{r}p_j^{r+1}(1-p_j)^{n-r} \qquad \mathbb{E}M_r(n) = \sum_{j=1}^{\infty}p_j\,\mathrm{e}^{-np_j}\frac{(np_j)^r}{r!}\,.$$

Formulas for higher moments can also be computed explicitly but their expression, especially in the binomial setting where a lot of dependencies are involved, often has an impractical form.

3 Karlin's formalism

We find it convenient to use the unifying framework proposed by Karlin [102]. Let us encode the probabilities (p_j) into the *counting measure* ν defined by

$$\nu(\,\mathrm{d}x) = \sum_{j\geq 1} \delta_{p_j}(\,\mathrm{d}x)\,, \tag{3.1}$$

where δ_p is the Dirac mass at p, and let $\vec{\nu} : (0,1] \to \mathbb{N}$ be the right tail of ν, *i.e.* for all $x \in (0,1]$,

$$\vec{\nu}(x) = \nu[x,\infty[= |\{j \geq 1,\, p_j \geq x\}|\,. \tag{3.2}$$

The function $\vec{\nu}$ is referred to as the *counting function*, $\vec{\nu}(x)$ corresponding to the number of symbols with probability larger than x. Clearly, we always have $\vec{\nu}(x) \leq 1/x$. As shown by Gnedin et al. [80], we even have

$$\vec{\nu}(x) \quad \ll \quad \frac{1}{x}\,, \tag{3.3}$$

as $x \to 0$. We also define the measure ν_1 by

$$\nu_1(\,\mathrm{d}x) = \sum_{j\geq 1} p_j \delta_{p_j}(\,\mathrm{d}x)\,. \tag{3.4}$$

Hence, for $x \in [0,1]$, $\nu_1[0,x] = \sum_{j\geq 1} p_j \mathbb{1}_{p_j \leq x}$ is the cumulated probability of symbols with probability smaller than x. Note that the expected occupancy counts and masses can be written simply as integrals against the measure ν. For instance

$$\mathbb{E}K(n) \quad = \quad \int_0^1 (1 - \mathrm{e}^{-nx})\nu(\,\mathrm{d}x)\,.$$

One may observe that we may as well integrate from 0 to ∞, and, integrating by parts, we have

$$\mathbb{E}K(n) \quad = \quad \left[-\vec{\nu}(x)(1 - \mathrm{e}^{-nx})\right]_0^\infty + n\int_0^\infty \mathrm{e}^{-nx}\vec{\nu}(x)\,\mathrm{d}x$$

Using (3.3), the first term is equal to zero and we get

$$\frac{\mathbb{E}K(n)}{n} \quad = \quad \int_0^\infty \mathrm{e}^{-nx}\vec{\nu}(x)\,\mathrm{d}x\,.$$

Hence, $n \mapsto \mathbb{E}K(n)/n$ is the Laplace transform of $\vec{\nu}$ and $n \mapsto \mathbb{E}K(n)$ characterizes the measure ν.

In both Chapter 1 and 2, we will take a particular interest in the class of *regularly varying* probability distributions, which can be seen as a generalization of the class of power laws. Even though a large part of our work is distribution-free, we prefer to state this condition in the introduction so as to fix our terminology.

Definition 1 (REGULAR VARIATION)**.** The source $P = (p_j)_{j\geq 1}$ is said to be regularly varying with index $\alpha \in [0,1]$ if the counting function $\vec{\nu}(1/\cdot)$ is regularly varying with index α (denoted $\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha$), *i.e.*

$$\vec{\nu}(1/n) \quad \underset{n\to\infty}{\sim} \quad n^\alpha \ell(n)\,, \tag{3.5}$$

where ℓ is a slowly varying function, that is, for all $x > 0$, $\frac{\ell(nx)}{\ell(n)} \underset{n\to\infty}{\to} 1$.

When $\alpha = 0$, we will sometimes require *extended regular variation*, assuming further that there exists a slowly varying function ℓ_0 such that, for all $x > 0$,

$$\frac{\ell(xn) - \ell(n)}{\ell_0(n)} \underset{n\to\infty}{\longrightarrow} \log x\,. \tag{3.6}$$

The function ℓ_0 is called the *auxiliary function* and satisfies $\ell_0(n) = o(\ell(n))$. We denote condition (3.6) by $\vec{\nu}_f \in \Pi_{\ell_0}$.

For instance, the Geometric distribution belongs to RV_0 but does not satisfies extended regular variation. However, distributions with a slightly heavier tail than the Geometric, such as frequencies proportional to $q^{j^{1-\varepsilon}}$ for $0 < q, \varepsilon < 1$, or the discretized log-Normal distribution, do satisfy it.

As noted above, $n \mapsto \mathbb{E}K(n)/n$ is the Laplace transform of $\vec{\nu}$. Hence, by Abelian-Tauberian theorems, the regular variation of $\vec{\nu}$ translates into regular variation of $\mathbb{E}K(n)$ (and other expectations).

Under the regular variation assumption, Karlin [102] and Gnedin et al. [80] investigated the asymptotic behaviour of K_n and $K_{n,r}$, and established laws of large numbers and central limit theorems for those variables.

Chapter 1

Concentration inequalities in infinite occupancy schemes

This chapter is devoted to the concentration properties of occupancy counts and of the missing mass. We summarize our results in Section 1. We give distribution-free bounds (Section 2), and then exhibit a vast domain where these results are tight, namely the domain of distributions with a heavier tail than the geometric (Section 3). In this domain, the non-asymptotic exponential concentration properties that we establish are sharp in the sense that the exponents are order-optimal, precisely capturing the scale of the variance. For this reason, we dedicate a portion of the chapter to establishing bounds on various variances. Some applications are presented in Section 4, pertaining mostly to the Good-Turing estimator of the missing mass. All the proofs are grouped in Section 6.

1 Main results

Terminology

Our concentration results mostly take the form of bounds on the logarithm of the Laplace transform. Our terminology follows closely [40]. We say that the random variable Z is *sub-Gaussian* on the right tail (resp. on the left tail) with variance factor v if, for all $\lambda \geq 0$ (resp. $\lambda \leq 0$),

$$\log \mathbb{E}\, \mathrm{e}^{\lambda(Z-\mathbb{E}Z)} \leq \frac{v\lambda^2}{2}\,. \tag{1.1}$$

We say that a random variable Z is *sub-Poisson* with variance factor v if, for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\, \mathrm{e}^{\lambda(Z-\mathbb{E}Z)} \leq v\phi(\lambda)\,, \tag{1.2}$$

with $\phi : \lambda \mapsto \mathrm{e}^{\lambda} - \lambda - 1$.

We say that a random variable Z is *sub-gamma on the right tail* with variance factor v and scale parameter c if

$$\log \mathbb{E}\, \mathrm{e}^{\lambda(X-\mathbb{E}X)} \leq \frac{\lambda^2 v}{2(1-c\lambda)} \text{ for every } \lambda \quad \text{such that} \quad 0 \leq \lambda \leq 1/c\,. \tag{1.3}$$

The random variable Z is *sub-gamma on the left tail* with variance factor v and scale parameter c, if $-Z$ is sub-gamma on the right tail with variance factor v and scale parameter c. If Z is sub-Poisson with variance factor v, then it is sub-Gaussian on the left tail with variance factor v, and sub-gamma on the right tail with variance factor v and scale parameter $1/3$.

These log-Laplace upper bounds then imply exponential tail bounds. For instance, Inequality (1.3) results in a Bernstein-type inequality for the right tail, that is, for $s > 0$ our inequalities have the form

$$\mathbb{P}\{Z > \mathbb{E}[Z] + \sqrt{2vs} + cs\} \leq \mathrm{e}^{-s},$$

while Inequality (1.1) for all $\lambda \leq 0$ entails

$$\mathbb{P}\{Z < \mathbb{E}[Z] - \sqrt{2vs}\} \leq \mathrm{e}^{-s}\,.$$

We present such results first without making distributional assumptions, beyond the structure of those quantities themselves. These concentrations then specialize in various settings, such as that of regular variation.

Main results

We proceed by giving a coarse description of our main results. In the Poisson setting, for each $r \geq 1$, $(X_j(n))_{j\geq 1}$ are independent, hence $K_r(n)$ is a sum of independent Bernoulli random variables, and it is not too surprising that it satisfies sub-Poisson, also known as Bennett, inequalities. For $\lambda \in \mathbb{R}$, we have:

$$\log \mathbb{E}\,\mathrm{e}^{\lambda(K_r(n)-\mathbb{E}K_r(n))} \leq \mathrm{Var}(K_r(t))\phi(\lambda) \leq \mathbb{E}[K_r(n)]\phi(\lambda)\,.$$

The proofs are elementary and are based on the careful application of Efron-Stein-Steele inequalities and the entropy method [40].

As for the binomial setting, the summands are not independent but we can use negative association arguments [64] (see Section 6) to obtain Bennett inequalities for the cumulated occupancy counts $K_{n,\overline{r}}$. These hold either with the Jackknife variance proxy given by the Efron-Stein inequality, $r\mathbb{E}K_{n,r}$ or with the variance proxy stemming from the negative correlation of the summands, $\mathbb{E}K_{n,\overline{r}}$. Letting $v_{n,\overline{r}} = \min(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}})$, we have, for all $\lambda \in \mathbb{R}$:

$$\log \mathbb{E}\,\mathrm{e}^{\lambda(K_{n,\overline{r}}-\mathbb{E}K_{n,\overline{r}})} \leq v_{n,\overline{r}}\phi(\lambda)\,.$$

This in turn implies a concentration inequality for $K_{n,r}$. Letting

$$v_{n,r} = 2\min\big(\max(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}), \mathbb{E}K_{n,\overline{r}}\big)\,,$$

we have, for all $s \geq 0$,

$$\mathbb{P}\Big\{|K_{n,r} - \mathbb{E}K_{n,r}| \geq \sqrt{4v_{n,r}s} + 2s/3\Big\} \leq 4\mathrm{e}^{-s}\,.$$

In the Poisson setting, we go one step further and consider the supremum over $\mathcal{J} \subset \mathbb{N}_+$

$$Z \;=\; \sum_{r\in\mathcal{J}} \frac{K_r(n) - \mathbb{E}K_r(n)}{\sqrt{\mathrm{Var}K_r(n)}}\,.$$

Noticing that Z is the supremum of an empirical process and carefully invoking Klein-Rio and Samson inequalities, we establish Bernstein inequalities for Z.

We obtain distribution-free bounds on the log-Laplace transform of $M_{n,0}$, which result in sub-Gaussian concentration on the left tail, sub-gamma concentration on the right tail with scale proxy $1/n$. More

precisely, letting $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$, we show that, for all $\lambda \leq 0$,

$$\log \mathbb{E}\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^- \frac{\lambda^2}{2}\,,$$

and, for all $\lambda \geq 0$,

$$\log \mathbb{E}\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq v_n^+ \frac{\lambda^2}{2(1-\lambda/n)}\,.$$

Those bounds follow from an interesting relation between the log-Laplace transform of the missing mass and the sequence of expected occupancy counts $(\mathbb{E}K_r(n))_{r\geq 2}$, namely, for all $\lambda \geq 0$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n)\,.$$

This inequality implies that controlling the right tail of the missing mass can be done by uniformly controlling the expected occupancy counts.

These results are distribution-free. But though the variance factor v_n^- is a sharp bound for the variance of the missing mass, v_n^+ may be much larger. This leads us to look for distribution-specific conditions ensuring that v_n^+ captures the right order for the variance, such as by using a tail asymptotic stability condition as in extreme value theory.

Karlin [102] pioneered such a condition by assuming that the function $\vec{\nu}$ satisfies a regular variation assumption, namely

$$\vec{\nu}(1/n) \sim n^{\alpha}\ell(n) \quad \text{as } n \to \infty\,, \tag{1.4}$$

with $\alpha \in [0,1]$ and ℓ slowly varying [see also 80, 127]. This condition allows us to compare the asymptotics of the various occupancy scores. In particular, when $\alpha > 0$, then $\mathbb{E}K_2(n)$ and $\mathbb{E}K_{\overline{2}}(n)$ have the same order of growth, and, divided by n^2 they both are of the same order as the variance of the missing mass. Hence, regular variation provides a framework in which our concentration inequalities are order-optimal.

To handle the case $\alpha = 0$, we move from *Karamata* to *de Haan* theory, and take $\vec{\nu}$ to have an extended regular variation property with an auxiliary function ℓ_0 that tends to $+\infty$ (see equation 3.6). This domain corresponds to light-tailed distributions which are still heavier than the geometric. In this case, $\mathbb{E}K_2(n) \ll \mathbb{E}K_{\overline{2}}(n)$, and v_n^+ does not capture the right order. However, we manage to show the sub-gamma concentration of the missing mass only for n large enough, that is, that there exists n_0 such that for all $n \geq n_0$, for $\lambda > 0$, we have

$$\log \mathbb{E}\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})} \leq (v_n\lambda^2)/2(1-\lambda/n)\,,$$

with $v_n \asymp \mathrm{Var}M_{n,0}$.

Back to our examples of applications, considerable insight may be gained from these concentration results. For instance, heavy tails lead to multiplicative concentration for $M_{n,0}$ (strong law of large numbers) and the consistency of the Good-Turing estimator: $\frac{G_{n,0}}{M_{n,0}} \xrightarrow{p} 1$. Generally, new estimators can be derived and shown to be consistent in a unified framework, once one is able to estimate α consistently. For instance, when $\vec{\nu}(1/.)$ is regularly varying with index α, $\hat{\alpha} = K_{n,1}/K_n$ is a consistent estimator of α. Then, to estimate the number of new species in a sample twice the size of the original, we immediately get that $\widehat{K}_{2n} = K_n + \frac{2^{\hat{\alpha}}-1}{\hat{\alpha}}K_{n,1}$ is a consistent estimator of K_{2n}. This methodology is very similar to extreme value theory [20]: harnessing limiting expressions and tail parameter estimation. These results strengthen and extend the contribution of [127], which is restricted to power-laws and implicit constants in the inequalities. Beyond consistency results, we also obtain confidence intervals for the Good-Turing estimator in the Poisson setting, using empirical quantities.

Historical notes and related work

There exists a vast literature on the occupancy scheme, as formulated here and in many other variations. The asymptotic behaviour of K_n and $K_{n,r}$ has been intensively studied in the finite case with m urns and uniform probabilities. The number of urns m may then depend in a certain way on the number of samples n [98, 106]. Of particular relevance to our problem is the pioneer paper of Karlin [102], who built on earlier work by Bahadur [12], credited as one of the first to study the infinite occupancy scheme. Karlin's main results were to establish central limit theorems in an infinite setting, under a condition of regular variation. He also derived strong laws of large numbers. Gnedin et al. [80] present a general review of these earlier results, as well as more contemporary work on this problem. The focus continues to be central limit theorems, or generally asymptotic results. For example the work of Hwang and Janson [94] (effectively) provides a local limit theorem for K_n provided that the variance tends to infinity. Somewhat less asymptotic results have also been proposed, in the form of variance analysis and normal approximations, such as in the work of Bogachev et al. [31] and Barbour and Gnedin [13].

Besides occupancy counts analysis, a distinct literature investigates the number of species and missing mass problems. These originated in the works of Fisher et al. [73], Good [82], and Good and Toulmin [83], and generated a long line of research to this day [43]. Here, instead of characterizing the asymptotic behavior of these quantities, one is interested in estimating $K_{\lambda n} - K_n$ for a $\lambda > 1$, that is the number of discoveries when the sample size is multiplied by λ, or estimating $M_{n,0}$: estimators are proposed, and then their statistical properties are studied. One recently studied property by McAllester and Schapire [122], McAllester and Ortiz [121], and Acharya et al. [2], is that of concentration, which sets itself apart from the CLT-type results in that it is non-asymptotic in nature. Based on this, Ohannessian and Dahleh [127] showed that in the regular variation setting of Karlin, one could show multiplicative concentration, and establish strong consistency results. Conversely, characterizing various aspects of concentration allows one to systematically design new estimators. For example, this was illustrated in Ohannessian and Dahleh [127] for the estimation of rare probabilities, to both justify and extend Good's [82] work that remains relevant in some of the aforementioned applications, especially computational linguistics.

2 Distribution-free concentration

2.1 Occupancy counts

2.1.1 Variance bounds

In order to understand the fluctuations of occupancy counts K_n, $K(n)$, $K_{n,r}$, $K_r(n)$, we start by reviewing and stating variance bounds. We start with the Poisson setting where occupancy counts are sums of independent Bernoulli random variables with possibly different success probabilities, and thus variance computations are straightforward. There are exact expressions [see for example 80, Equation (4)], but we may also derive more tractable and tight bounds. We start by stating a well-known connection between the variance of the number of occupied urns and the expected number of singletons [80][102]. In the binomial setting, similar bounds can be derived using the Efron-Stein-Steele inequalities, outlined in Section 6.1.1 [see 40, Section 3.1].

Proposition 2.1. *In the Poisson setting, we have*

$$\frac{\mathbb{E}K_1(2n)}{2} \leq \mathrm{Var}(K(n)) \leq \mathbb{E}K_1(n)\,.$$

In the binomial setting, we have

$$\mathrm{Var}(K_n) \leq \mathbb{E}\left[K_{n,1}(1 - M_{n,0})\right] \leq \mathbb{E}K_{n,1}\,.$$

Remark 2.1. *Despite its easy proof, the bound* $\mathrm{Var}(K_n) \leq \mathbb{E}K_{n,1}$ *is a much sharper bound than the obvious self-bounding upper bound presented in [40, Example 3.9],* $Var(K_n) \leq \mathbb{E}K_n$*. In some scenarii, for example when the sampling distribution is light-tailed or even geometric,* $\mathbb{E}K_n$ *tends to* ∞ *whereas* $\mathbb{E}K_{n,1}$ *and* $Var(K_n)$ *remain bounded.*

Another straightforward bound on $\mathrm{Var}(K_n)$ comes from the fact that the Bernoulli variables $(\mathbb{1}_{\{X_{n,j}>0\}})_{j\geq 1}$ are negatively correlated. Thus, ignoring the covariance terms, we get

$$\mathrm{Var}(K_n) \leq \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{1}_{\{X_{n,j}>0\}}) = \sum_{j=1}^{\infty} (1-p_j)^n (1-(1-p_j)^n) = \mathbb{E}K_{2n} - \mathbb{E}K_n\,.$$

Let us denote this bound by $\mathrm{Var}^{\mathrm{ind}}(K_n) = \mathbb{E}K_{2n} - \mathbb{E}K_n$, as it is a variance proxy obtained by considering that the summands in K_n are independent. One can observe that the expression of $\mathrm{Var}^{\mathrm{ind}}(K_n)$ is very similar to the variance in the Poissonized setting, $\mathrm{Var}(K(n)) = \mathbb{E}K(2n) - \mathbb{E}K(n)$. It is insightful to compare the true variance, the Poissonized proxy, and the negative correlation proxy, to quantify the price one pays by resorting to the latter two as an approximation for the first. We revisit this in more detail in Section 5.1.

We now investigate the fluctuations of the individual occupancy counts $K_{n,r}$ and $K_r(n)$, and that of the cumulative occupancy counts $K_{n,\overline{r}} = \sum_{s\geq r} K_{n,s}$ and $K_{\overline{r}}(n) = \sum_{s\geq r} K_s(n)$.

Proposition 2.2. *In the Poisson setting, for* $r \geq 1$*,* $n \geq 0$*,*

$$\mathrm{Var}(K_{\overline{r}}(n)) \leq \min\left(r\mathbb{E}K_r(n), \mathbb{E}K_{\overline{r}}(n)\right)\,.$$

In the binomial setting, for $r, n \geq 1$*,*

$$\mathrm{Var}(K_{n,\overline{r}}) \leq \min\left(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}}\right)\,.$$

For each setting, the first bound follows from Efron-Stein-Steele inequalities, the second from negative correlation. These techniques are presented briefly in Sections 6.1.1 and 6.1.2 respectively.

Remark 2.2. *Except for* $r = 1$*, there is no clear-cut answer as to which of these two bounds is the tightest. In the regular variation scenario with index* $\alpha \in]0,1]$ *as explored in [80], the two bounds are asymptotically of the same order, indeed,*

$$\frac{r\mathbb{E}K_{n,r}}{\mathbb{E}K_{n,\overline{r}}} \underset{n\to+\infty}{\sim} \alpha\,,$$

see Section 3 for more on this. □

Bounds on $\mathrm{Var}(K_r(n))$ can be easily derived as $K_r(n)$ is a sum of independent Bernoulli random variables. Moreover, noticing that $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$ and that $K_{n,\overline{r}}$ and $K_{n,\overline{r+1}}$ are positively correlated, the following bound is immediate.

Proposition 2.3. *In the Poisson setting, for* $r \geq 1$*,* $n \geq 0$*,*

$$\mathrm{Var}(K_r(n)) \leq \mathbb{E}K_r(n)\,.$$

In the binomial setting, for $r, n \geq 1$,

$$\begin{aligned}\mathrm{Var}(K_{n,r}) &\leq \min\left(r\mathbb{E}K_{n,r} + (r+1)\mathbb{E}K_{n,r+1}, \mathbb{E}K_{n,\overline{r}} + \mathbb{E}K_{n,\overline{r+1}}\right) \\ &\leq 2\min\left(\max(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}), \mathbb{E}K_{n,\overline{r}}\right) .\end{aligned}$$

2.1.2 Concentration inequalities

Concentration inequalities refine variance bounds. These bounds on the logarithmic moment generating functions are indeed Bennett (sub-Poisson) inequalities with the variance upper bounds stated in the preceding section. For $K_{n,\overline{r}}$, the next proposition gives a Bernstein inequality where the variance factor is the Efron-Stein upper bound on the variance.

Proposition 2.4. *Let $r \geq 1$, and let $v_{n,\overline{r}} = \min(r\mathbb{E}K_{n,r}, \mathbb{E}K_{n,\overline{r}})$. Then, for all $\lambda \in \mathbb{R}$,*

$$\log \mathbb{E}\, \mathrm{e}^{\lambda(K_{n,\overline{r}} - \mathbb{E}K_{n,\overline{r}})} \leq v_{n,\overline{r}}\phi(\lambda) ,$$

with $\phi : \lambda \mapsto \mathrm{e}^{\lambda} - \lambda - 1$.

It is worth noting that the variance bound $\mathbb{E}K_{n,\overline{r}}$ in this concentration inequality can also be obtained using a variant of Stein's method known as size-biased coupling [15, 48].

A critical element of the proof of Proposition 2.4 is to use the fact that each $K_{n,\overline{r}}$ is a sum of negatively associated random variables (Section 6.1.2). This is not the case for $K_{n,r}$, and thus negative association cannot be invoked directly. To deal with this, we simply use the observation of Ohannessian and Dahleh [127] that since $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$, the concentration of $K_{n,r}$ follows from that of those two terms. We can show the following.

Proposition 2.5. *Let*

$$v_{n,r} = 2\min\left(\max(r\mathbb{E}K_{n,r}, (r+1)\mathbb{E}K_{n,r+1}), \mathbb{E}K_{n,\overline{r}}\right) .$$

Then, for $s \geq 0$,

$$\mathbb{P}\left\{|K_{n,r} - \mathbb{E}K_{n,r}| \geq \sqrt{4v_{n,r}s} + 2s/3\right\} \leq 4\mathrm{e}^{-s} .$$

In the favorable Poisson setting, we may go one step further and consider the supremum of the normalized *occupancy process*, letting $\sigma_r^2 = \mathrm{Var}(K_r(n))$ for each r, we consider

$$Z = \sup_{r \in \mathcal{J}} \frac{K_r(n) - \mathbb{E}K_r(n)}{\sigma_r} ,$$

where $\mathcal{J} \subset \mathbb{N}^*$. As for each $r \in \mathcal{J}$, $K_r(n)$ is a sum of independent centered (but not necessarily identically distributed) random variables, this supremum may be considered as a supremum of an empirical process with uniformly bounded and centered components. The suprema considered here differ from the suprema involved in Talagrand's inequalities (See [40]) as they are build by summing a countable rather than a finite collection of independent random variables. Nevertheless, the dimension-free nature of Talagrand's inequalities suggests that the supremum as well as its normalized countrepart satisfy genuine concentration inequalities.

Theorem 2.1. *Let $\nu = 2\frac{\mathbb{E}Z}{\min_{r\in\mathcal{J}}\sigma_r} + 1$ and $\nu' = \frac{\mathbb{E}Z}{\min_{r\in\mathcal{J}}\sigma_r} + \max_{r\in\mathcal{J}}\frac{\mathbb{E}K_r(n)}{\sigma_r^2}$. Then,*

$$\mathbb{E}Z \leq \sqrt{2\log|\mathcal{J}|} + \frac{\log|\mathcal{J}|}{\min_{r\in\mathcal{J}}\sigma_r}$$

$$\mathrm{Var} Z \leq \min\left(\nu, \nu'\right) .$$

For $t \geq 0$,

$$\mathbb{P}\left\{Z \geq \mathbb{E}Z + t\right\} \leq \exp\left(-\frac{t}{4}\log\left(1 + 2\ln\left(1 + \frac{t}{\nu}\right)\right)\right)$$

$$\mathbb{P}\left\{Z \leq \mathbb{E}Z - t\right\} \leq \exp\left(-\frac{t^2}{2\left(\nu' + 2t/3\right)}\right) .$$

The bounds are mostly interesting if $\mathcal{J}$ is chosen in such a way that $\min_{r\in\mathcal{J}} \sigma_r \geq \sqrt{\log|\mathcal{J}|/2}$, then the expectation bound is not larger than $2\sqrt{2\log|\mathcal{J}|}$, while the variance bound is not larger than 9. Exponential tail bounds entail that with probability larger than $1-\delta$, uniformly over $\mathcal{J}$,

$$|K_r(n) - \mathbb{E}K_r(n)| \leq \sqrt{\mathrm{Var} K_r(n)}\left(2\sqrt{2\log|\mathcal{J}|} + \sqrt{\kappa'\log 1/\delta} + \kappa''\log 1/\delta\right)$$

where κ', κ'' are universal constants.

2.2 Missing mass

2.2.1 Variance bound

Recall that $M_{n,0} = \sum_{j=1}^{\infty} p_j \mathbb{1}_{\{X_{n,j}=0\}}$, and we can readily show that the summands are negatively associated weighted Bernoulli random variables (Section 6.1.2). This results in a handy upper bound for the variance of the missing mass.

Proposition 2.6. *In the Poisson setting,*

$$\mathrm{Var}(M_0(n)) = 2\mathbb{E}K_2(n)/n^2 - \mathbb{E}K_2(2n)/2n^2 \leq \frac{2\mathbb{E}K_2(n)}{n^2} ,$$

while in the binomial setting,

$$\mathrm{Var}(M_{n,0}) \leq \sum_{j=1}^{\infty} p_j^2 \mathrm{Var}\left(\mathbb{1}_{\{X_{n,j}=0\}}\right) \leq \frac{2\mathbb{E}K_2(n)}{n^2} .$$

Note that whereas the expected value of the missing mass is connected to the number of singletons, its variance may be upper bounded using the number of *doubletons* (in the Poisson setting). This connection was already pointed out in [82] and [69].

2.2.2 Concentration of the left tail

Moving on to the concentration properties of the missing mass, we first note that as a sum of weighted sub-Poisson random variables (following [40]), the missing mass is itself a sub-gamma random variable on both tails. It should not come as a surprise that the left tail of $M_{n,0}$ is sub-Gaussian with the variance factor derived from negative association. This had already been pointed out by [122] and McAllester and Ortiz [121].

Proposition 2.7. *[121] In the Poisson setting, the missing mass $M_0(n)$ is sub-Gaussian on the left tail with variance factor the effective variance* $\mathrm{Var}(M_0(n)) = \sum_{j=1}^{\infty} p_j^2 \,\mathrm{e}^{-np_j}(1 - \mathrm{e}^{-np_j})$.

In the binomial setting, the missing mass $M_{n,0}$ is sub-Gaussian on the left tail with variance factor $v = \sum_{j=1}^{\infty} p_j^2 \mathrm{Var}\left(\mathbb{1}_{\{X_{n,j}=0\}}\right)$ or $v_n^- = 2\mathbb{E}K_2(n)/n^2$.

For $\lambda \leq 0$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \leq \frac{v\lambda^2}{2} \leq \frac{2v_n^-\lambda^2}{2}$$

2.2.3 Concentration of the right tail

The following concentration inequalities for the right tail of the missing mass mostly rely on the following proposition, which bounds the log-Laplace transform of the missing mass in terms of the sequence of expected occupancy counts $\mathbb{E}K_r(n)$, for $r \geq 2$.

Proposition 2.8. *For all $\lambda > 0$,*

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n)\,.$$

This suggests that if we have a uniform control on the expected occupancy scores $(\mathbb{E}K_r(t))_{r\geq 2}$, then the missing mass has a sub-gamma right tail, with some more or less accurate variance proxy, and scale factor $1/n$.

The next theorem shows that the missing mass is sub-gamma on the right tail with variance proxy $2\mathbb{E}K_{\overline{2}}(n)/n^2$ and scale proxy $1/n$. Despite its simplicity and its generality, this bound exhibits an intuitively correct scale factor: if there exist symbols with probability of order $1/n$, they offer the major contribution to the fluctuations of the missing mass.

Theorem 2.2. *In the binomial as well as in the Poisson setting, the missing mass is sub-gamma on the right tail with variance factor $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$ and scale factor $1/n$. For $\lambda \geq 0$,*

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \frac{v_n^+\lambda^2}{2(1-\lambda/n)}\,.$$

If the sequence $(\mathbb{E}K_r(n))_{r\geq 2}$ is non-increasing, the missing mass is sub-gamma on the right tail with variance factor $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and scale factor $1/n$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \frac{v_n^-\lambda^2}{2(1-\lambda/n)}\,.$$

Remark 2.3. *McAllester and Ortiz [121] and Berend and Kontorovich [25] point out that for each Bernoulli random variable $Y_j = \mathbb{1}_{\{X_{n,j}=0\}}$, for all $\lambda \in \mathbb{R}$*

$$\log \mathbb{E}\,\mathrm{e}^{\lambda(Y_j-\mathbb{E}Y_j)} \leq \frac{\lambda^2}{4\mathrm{c}_{\mathrm{LS}}(\mathbb{E}Y_j)}\,,$$

where $\mathrm{c}_{\mathrm{LS}}(p) = \log((1-p)/p)/(1-2p)$ (or 2 if $p = 1/2$) is the optimal logarithmic Sobolev constant for Bernoulli random variables with success probability p (this sharp and non-trivial result has been proven independently by a number of people: Kearns and Saul [103], Berend and Kontorovich [25], Raginsky and Sason [139], Berend and Kontorovich [26]; the constant also appears early on in the exponent of one of Hoeffding's inequalities [92, Theorem 1, Equation (2.2)]). From this observation, thanks to the negative association of the $(Y_j)_{j\geq 1}$, it follows that the missing mass is sub-Gaussian with variance factor

$$w_n = \sum_{j=1}^{\infty} \frac{p_j^2}{2\mathrm{c}_{\mathrm{LS}}((1-p_j)^n)} \leq \sum_{j=1}^{\infty} \frac{p_j^2}{2\log((1-p_j)^{-n})} \leq \sum_{j=1}^{\infty} \frac{p_j^2}{2np_j} \leq \frac{1}{2n}\,. \tag{2.1}$$

An upper bound on w_n does not mean that w_n is necessarily larger than $\mathbb{E}K_{\overline{2}}(n)/n^2$. Nevertheless, it is possible to derive a simple lower bound on w_n that proves to be of order $O(1/n)$.

Assume that the sequence $(p_j)_{j\geq 1}$ is such that $p_j \leq 1/4$ for all $j \geq 1$. Then

$$\begin{aligned} w_n &\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2}{2\mathrm{C}_{\mathrm{LS}}((1-p_j)^n)} \\ &\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2(1-2(1-p_j)^n)}{2n\log(1/(1-p_j))} \\ &\geq \sum_{j:p_j\geq 1/n} \frac{p_j^2(1-2/\mathrm{e})}{2np_j/(1-p_j)} \\ &\geq \frac{3(1-2/\mathrm{e})}{8n}(1-\sum_{j:p_j<1/n} p_j) \\ &\geq \frac{3}{32n}\left(1-\sum_{j:p_j<1/n} p_j\right), \end{aligned}$$

and the statement follows from the observation that $\lim_{n\to\infty}\sum_{j:p_j<1/n} p_j = 0$.

The variance factor w_n from (2.1) is usually larger than $2\mathbb{E}K_{\overline{2}}(n)/n^2$. In the scenarios discussed in Section 3, $(2\mathbb{E}K_{\overline{2}}(n)/n^2)/w_n$ even tends to 0 as n tends to infinity. □

3 Regular variation

3.1 Definition and motivation

Are the variance bounds in the results of Section 2 tight? In some pathological situations this may not be the case.

In particular, we may conjecture that Theorem 2.2 is likely to be sharp when the first terms of the sequence $(\mathbb{E}K_r(n))_{r\geq 2}$ grow at the same rate as $\mathbb{E}K(n)$, or at least as $\mathbb{E}K_{\overline{2}}(n)$. We see in what follows that the regular variation framework introduced by Karlin [102] leads to such asymptotic equivalents. The most useful aspect of these equivalent growth rates is a simple characterization of the variance of various quantities, particularly relative to their expectation. We focus on the right tail of the missing mass, which exhibits the highest sensitivity to this asymptotic behavior, by trying to specialize Theorem 2.2 under regular variation.

Regularly varying frequencies can be seen as generalizations of power-law frequencies. One possible definition is as follows: for $\alpha \in (0,1)$, the sequence $(p_j)_{j\geq 1}$ is said to be regularly varying with index $-1/\alpha$ if , for all $\kappa \in \mathbb{N}_+$,

$$\frac{p_{\kappa j}}{p_j} \underset{j\to\infty}{\sim} \kappa^{-\frac{1}{\alpha}}.$$

It is easy to see that pure power laws do indeed satisfy this definition. However, in order to extend the regular variation hypothesis to $\alpha = 0$ and 1, we need a more flexible definition.

Henceforth, following Karlin [102], we say that the probability mass function $(p_j)_j$ is regularly varying with index $\alpha \in [0,1]$, if $\vec{\nu}(1/\cdot)$ is α-regularly varying in the neighbourhood of ∞, which reads as

$$\vec{\nu}(1/n) \underset{n\to\infty}{\sim} n^{\alpha}\ell(n),$$

where ℓ is a slowly varying function. We use the notation $\vec{\nu}(1/\cdot) \in \mathrm{RV}_{\alpha}$.

We now note that when $\alpha \in (0,1)$, the regular variation assumption on $(p_j)_{j\geq 1}$ is indeed equivalent to the regular variation assumption on the counting function $\vec{\nu}$ [see 80, Proposition 23]: if $(p_j)_{j\geq 1}$ is

regularly varying with index $-1/\alpha$ as j tends to infinity, then $\vec{\nu}(1/\cdot)$ is α-regularly varying [see also 30]. The second definition however lends itself more easily to generalization to $\alpha = 0$ and 1.

In what follows, we treat these three cases separately: the nominal regular variation case with $\alpha \in (0,1)$ strictly, the *fast variation* case with $\alpha = 1$, and the *slow variation* case with $\alpha = 0$.

In the latter case, that is if $\vec{\nu}(1/n) \sim \ell(n)$, we find that the mere regular variation hypothesis is not sufficient to obtain asymptotic formulas. For this reason, we introduce further control in the form of an *extended* regular variation hypothesis (see equation 3.6).

Remark 3.1. *Before we proceed, as further motivation, we note that the regular variation hypothesis is very close to being a necessary condition for exponential concentration. For example, considering Proposition 2.8, we see that if the sampling distribution is such that the ratio $\mathbb{E}K_{\overline{2}}(n)/\mathbb{E}K_2(n)$ remains bounded, then we are able to capture the right variance factor. Now, defining the shorthand $\Phi_{\overline{2}}(t) = \mathbb{E}K_{\overline{2}}(t)$ and $\Phi_2(t) = \mathbb{E}K_2(t)$ following the notation of [80], we have*

$$\Phi'_{\overline{2}}(t) = \frac{2\Phi_2(t)}{t}\,.$$

Hence, $\Phi_{\overline{2}}(t)/\Phi_2(t) = 2\Phi_{\overline{2}}(t)/t\Phi'_{\overline{2}}(t)$, and if instead of boundedness, we further require that this ratio converges to some finite limit, then, by the converse part of Karamata's Theorem [see 55, Theorem B.1.5], we find that $\Phi_{\overline{2}}$ (and then Φ_2) is regularly varying, which in turn implies that $\vec{\nu}(1/t)$ is regularly varying. We elaborate on this further in our discussions, in Section 5.2. □

3.2 Case $\alpha \in (0,1)$

We first consider the case $0 < \alpha < 1$. The next theorem states that when the sampling distribution is regularly varying with index $\alpha \in (0,1)$, the variance factors in the Bernstein inequalities of Proposition 2.7 and Theorem 2.2 are of the same order as the variance of the missing mass.

Theorem 3.1. *Assume that $\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha$ with $\alpha \in (0,1)$. Then the missing mass $M_{n,0}$ (or $M_0(n)$) is sub-Gaussian on the left tail with variance factor $v_n^- = 2\mathbb{E}K_2(n)/n^2$ and sub-gamma on the right tail with variance factor $v_n^+ = 2\mathbb{E}K_{\overline{2}}(n)/n^2$. The variance factors satisfy*

$$\lim_n \frac{v_n^-}{\mathrm{Var}(M_{n,0})} = \frac{1}{1-2^{\alpha-2}}\,,$$

$$\lim_n \frac{v_n^+}{\mathrm{Var}(M_{n,0})} = \frac{2}{\alpha(1-2^{\alpha-2})}\,.$$

The second ratio deteriorates when α approaches 0, implying that the variance factor for the right tail gets worse for lighter tails. We do not detail the proof of Theorem 3.1, except to note that it follows from Proposition 2.7, Theorem 2.2, and the following asymptotics [see also 80, 127]:

Theorem 3.2. *[102] If the counting function $\vec{\nu}$ is regularly varying with index $\alpha \in (0,1)$, we have*

- *Number of occupied urns*

$$K_n \underset{+\infty}{\sim} \mathbb{E}K_n \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_n \underset{+\infty}{\sim} \Gamma(1-\alpha)n^\alpha \ell(n)\,,$$

- *Number of urns with $r \geq 1$ balls*

$$K_{n,r} \underset{+\infty}{\sim} \mathbb{E}K_{n,r} \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_{n,r} \underset{+\infty}{\sim} \frac{\alpha\Gamma(r-\alpha)}{r!} n^\alpha \ell(n)\,.$$

and the same hold for the corresponding Poissonized quantities.

Note that all expected occupancy counts are of the same order, and the asymptotics for $\mathbb{E}K_{\overline{2}}(n)$ follows directly from the difference between $\mathbb{E}K(n)$ and $\mathbb{E}K_1(n)$.

Gnedin et al. [80] observe that, thanks to the general binomial formula, in the regular variation scenario with $0 < \alpha < 1$, the almost sure limits of $\frac{K_{n,r}}{K_n}$ define a probability distribution over $\mathbb{N}$. Indeed, we have

$$\frac{K_{n,r}}{K_n} \underset{n\to\infty}{\to} \frac{\alpha\Gamma(r-\alpha)}{r!\Gamma(1-\alpha)} = (-1)^{r+1}\binom{\alpha}{r} := Q_\alpha(r) \quad \text{a.s.}\,,$$

and $\sum_{r\geq 1} Q_\alpha(r) = 1$. Let us denote by Q_n the empirical occupancy measure, *i.e.* for all $r \geq 1$, $Q_n(r) = \frac{K_{n,r}}{K_n}$ The next result outlines that under the regular variation assumption, the sequence (Q_n) almost surely converges towards the scale-free occupancy probability distribution q_α with respect to the total variation distance. It is a direct consequence of Scheffé's Lemma.

Proposition 3.1. *If $\vec{\nu} \in \mathrm{RV}_\alpha, \alpha \in]0,1[$, then, almost surely,*

$$d_{\mathrm{TV}}(Q_n, Q_\alpha) = \frac{1}{2}\sum_{r=1}^{+\infty} |Q_n(r) - Q_\alpha(r)| \underset{n\to\infty}{\longrightarrow} 0\,.$$

3.3 Fast variation, $\alpha = 1$

We refer to the regular variation regime with $\alpha = 1$ as *fast variation*[1]. From the perspective of concentration, this represents a relatively "easy" scenario. In a nutshell, this is because the variance of various quantities grows much slower than their expectation.

The result of this section is to simply state that Theorem 3.1 continues to hold as is for $\alpha = 1$. The justification for this, however, is different. In particular, the asymptotics of Theorem 3.2 do not apply: the number of distinct symbols K_n and the singletons $K_{n,1}$ continue to have comparable growth order, but now their growth dominates that of $K_{n,r}$ for all $r \geq 2$. Intuitively, under fast variation almost all symbols appear only once in the observation, with only a vanishing fraction of symbols appearing more than once. We formalize this in the following Theorem.

Theorem 3.3. *[102] Assume $\vec{\nu}(1/n) = n\ell(n)$ with $\ell \in \mathrm{RV}_0$ (note that ℓ tends to 0 at ∞). Define $\ell_1 : [1,\infty) \to \mathbb{R}_+$ by*

$$\ell_1(y) = \int_y^\infty \frac{\ell(u)}{u}\,\mathrm{d}u\,.$$

Then $\ell_1 \in \mathrm{RV}_0$ and $\lim_{t\to\infty} \ell_1(t)/\ell(t) = \infty$ and the following asymptotics hold:

- *Number of occupied urns*

$$K_n \underset{+\infty}{\sim} \mathbb{E}K_n \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_n \underset{+\infty}{\sim} n\ell_1(n)\,,$$

- *Number of urns with one ball*

$$K_{n,1} \underset{+\infty}{\sim} \mathbb{E}K_{n,1} \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_{n,1} \underset{+\infty}{\sim} n\ell_1(n)\,,$$

- *Number of urns with $r \geq 2$ balls*

$$K_{n,r} \underset{+\infty}{\sim} \mathbb{E}K_{n,r} \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_{n,r} \underset{+\infty}{\sim} \frac{n\ell(n)}{r(r-1)}\,.$$

and the same hold for the corresponding Poissonized quantities.

1. Sometimes *rapid variation* is used [80], but this conflicts with [30].

As the expected missing mass scales like $\mathbb{E}K_1(n)/n$ while its variance scales like $\mathbb{E}K_2(n)/n^2$, Theorem 3.3 quantifies our claim that this is an "easy" concentration. To establish Theorem 3.1, it remains to show that $\mathbb{E}K_{\overline{2}}(n)$ is also of the same order as $\mathbb{E}K_2(n)$, with the correct limiting ratio for $\alpha = 1$. For this, we give the following proposition, which is in fact sufficient to prove Theorem 3.1 for both $0 < \alpha < 1$ and $\alpha = 1$.

Proposition 3.2. *Assume that the counting function $\vec{\nu}$ satisfies the regular variation condition with index $\alpha \in [0, 1]$, then for all $r \geq 2$,*

$$K_{\overline{r}}(n) \underset{+\infty}{\sim} \mathbb{E}K_{\overline{r}}(n) \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_{\overline{r}}(n) \underset{+\infty}{\sim} \frac{\Gamma(r-\alpha)}{(r-1)!}\, \vec{\nu}(1/n)\,.$$

Thus, when $\alpha = 1$, $\mathbb{E}K_r(n)$ and $\mathbb{E}K_{\overline{r}}(n)$ for $r \geq 2$ all grow like $n\ell(n)$, which is dominated by the $n\ell_1(n)$ growth of $\mathbb{E}K(n)$ and $\mathbb{E}K_1(n)$, as $\ell(n)/\ell_1(n) \to 0$. Specializing for $r = 2$, we do find that our proxies still capture the right order of the variance of the missing mass, and that we have the desired limit of Theorem 3.1, $\lim_n v_n^-/v_n^+ = \frac{1}{2}$.

Remark 3.2. *When $0 < \alpha < 1$, another good variance proxy would have have been $2\mathbb{E}K(n)/n^2$. For $\alpha = 1$, however, singletons should be removed to get the correct order.*

We also note that when $\alpha = 1$, the missing mass is even more stable. If we let v_n denote either $2\mathbb{E}K_2(n)/n^2$ or $2\mathbb{E}K_{\overline{2}}(n)/n^2$, then we have the following comparison between the expectation and the fluctuations of the missing mass, with the appropriate constants:

$$\frac{\sqrt{v_n}}{\mathbb{E}M_{n,0}} \sim \begin{cases} \frac{c_\alpha}{\sqrt{\vec{\nu}(1/n)}} & \text{for } 0 < \alpha < 1\,, \\ \frac{c_1}{\sqrt{\vec{\nu}(1/n)}} \cdot \frac{\ell(n)}{\sqrt{\ell_1(n)}} & \text{for } \alpha = 1\,. \end{cases} \qquad \square$$

3.4 Slow variation, $\alpha = 0$

The setting where the counting function $\vec{\nu}$ satisfies the regular variation condition with index 0 represents a challenge. Recall that this means that $\vec{\nu}(z/n)/\vec{\nu}(1/n)$ converges to 1 as n goes to infinity, yet to deal with this case we need to control the speed of this convergence, exemplified by the notion of extended regular variation that was introduced by de Haan [See 30, 55]. As we illustrate in the end of this section, one may face rather irregular behavior without such a hypothesis.

Definition 2. A measurable function $\ell : \mathbb{R}^+ \to \mathbb{R}^+$ has the extended slow variation property, if there exists a non-negative measurable function $\ell_0 : \mathbb{R}^+ \to \mathbb{R}^+$ such that for all $x > 0$

$$\lim_{n\to\infty} \frac{\ell(nx) - \ell(n)}{\ell_0(n)} \underset{n\to\infty}{\to} \log(x)\,.$$

The function $\ell_0(\cdot)$ is called an auxiliary function. When a function ℓ has the extended slow variation property with auxiliary function ℓ_0, we denote it by $\ell \in \Pi_{\ell_0}$.

Note that the auxiliary function is always slowly varying and grows slower than the original function, namely it satisfies $\lim_{n\to\infty} \ell(n)/\ell_0(n) = \infty$. Furthermore, any two possible auxiliary functions are asymptotically equivalent, that is if a_1 and a_2 are both auxiliary functions for ℓ, then $\lim_{n\to\infty} a_1(n)/a_2(n) = 1$.

The notion of extended slow variation and the auxiliary function give us the aforementioned control needed to treat the $\alpha = 0$ case on the same footing as the $0 < \alpha < 1$ case. In particular, in what follows in this section we assume that $\vec{\nu}(1/.) \in \Pi_{\ell_0}$, with the additional requirement that the auxiliary function ℓ_0 tends to $+\infty$.

Remark 3.3. *This domain corresponds to light-tailed distributions just above the geometric distribution (the upper-exponential part of Gumbel's domain). For the geometric distribution with frequencies* $p_j = (1-q)q^{k-1}$, $j = 1, 2, \dots$, *the counting function satisfies* $\vec{\nu}(1/n) \sim_\infty \log_{1/q}(n) \in \mathrm{RV}_0$, *but the auxiliary function* $\ell_0(n) = \log(1/q)$ *does not tend to infinity. Frequencies of the form* $p_j = cq^{\sqrt{j}}$ *on the other hand do fit this framework.* □

Theorem 3.4. *[80] Assume that* $\ell(t) = \vec{\nu}(1/t)$ *is in* Π_{ℓ_0}, *with* $\ell_0 \to \infty$. *The following asymptotics hold*

- *Number of occupied urns*

$$K_n \underset{+\infty}{\sim} \mathbb{E}K_n \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_n \underset{+\infty}{\sim} \ell(n)\,,$$

- *Number of urns with more than* $r \geq 1$ *balls*

$$K_{n,\bar{r}} \underset{+\infty}{\sim} \mathbb{E}K_{n,\bar{r}} \text{ a.s.} \quad \text{and} \quad \mathbb{E}K_{n,\bar{r}} \underset{+\infty}{\sim} \ell(n)\,,$$

- *Number of urns with* $r \geq 1$ *balls*

$$K_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}K_{n,r} \quad \text{and} \quad \mathbb{E}K_{n,r} \underset{+\infty}{\sim} \frac{\ell_0(n)}{r}\,,$$

- *Mass of urns with* $r \geq 0$ *balls*

$$M_{n,r} \overset{\mathbb{P}}{\sim} \mathbb{E}M_{n,r} \quad \text{and} \quad \mathbb{E}M_{n,r} \underset{+\infty}{\sim} \frac{\ell_0(n)}{n}\,.$$

The same equivalents hold for the corresponding Poissonized quantities.

Remark 3.4. *In this case, the expectations* $(\mathbb{E}K_{n,r})_{r\geq 1}$ *are of the same order but are much smaller than* $\mathbb{E}K_n$, *and the variables* K_n *and* $K_{n,\bar{r}}$ *are all almost surely equivalent to* $\ell(n)$. *It is also remarkable that all the expected masses* $(\mathbb{E}M_{n,r})_{r\geq 0}$ *are equivalent.* □

The variance of the missing mass is of order $2\mathbb{E}K_2(n)/n^2 \sim \ell_0(n)/n^2$, whereas the proxy $2\mathbb{E}K_{\overline{2}}(n)/n^2$ is of much faster order $2\ell(n)/n^2$, and is thus inadequate. By exploiting more carefully the regular variation hypothesis, we obtain uniform control over $(\mathbb{E}K_r(n))_{r\geq 1}$ for large enough n, leading to a variance proxy of the correct order.

Theorem 3.5. *Assume that* ℓ *defined by* $\ell(x) = \vec{\nu}(1/x)$ *is in* Π_{ℓ_0} *with* $\ell_0(n) \to \infty$ *as* $n \to \infty$, *and let* $v_n = 12\ell_0(n)/n^2$. *We have:*

(i) $\mathrm{Var}(M_{n,0}) \sim \frac{3\ell_0(n)}{4n^2}$, *thus* $v_n \asymp \mathrm{Var}(M_{n,0})$.

(ii) There exists $n_0 \in \mathbb{N}$ *that depends on* $\vec{\nu}$ *such that for all* $n > n_0$, *for all* $\lambda > 0$,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \leq \frac{v_n \lambda^2}{2(1-\lambda/n)}\,.$$

The same results hold for $M_0(t)$.

Remark 3.5. *By standard Chernoff bounding, Theorem 3.5 implies that there exists* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$, $s \geq 0$,

$$\mathbb{P}\left\{M_{n,0} \geq \mathbb{E}M_{n,0} + \sqrt{2v_n s} + \frac{s}{n}\right\} \quad \leq \quad \mathrm{e}^{-s}\,. \qquad \square$$

3.4.1 Too slow variation

We conclude this section by motivating why it is crucial to have a heavy-enough tail in order to obtain meaningful concentration. For example, even under regular variation when $\alpha = 0$, but $\vec{\nu}$ is not in a de Haan class Π_{ℓ_0} with $\ell_0(n) \to \infty$, the behavior of the occupancy counts and their moments may be quite irregular. In this section, we collect some observations on those light-tailed distributions. We start with the geometric distribution which represents in many respects a borderline case.

The geometric case is an example of slow variation: $\vec{\nu}(1/\cdot) \in \mathrm{RV}_0$. Indeed, with $p_k = (1-q)^{k-1}q$, $0 < q < 1$, we have

$$\begin{aligned}\vec{\nu}(x) &= \sum_{k=1}^{+\infty} \mathbb{1}_{\{p_k \geq x\}} \\ &= |k \in \mathbb{N}_+, (1-q)^{k-1}q \geq x| \\ &= 1 + \left\lfloor \frac{\log(x/q)}{\log(1-q)} \right\rfloor ,\end{aligned}$$

and thus $\vec{\nu}(x) \underset{x\to 0}{\sim} \ell(1/x)$, with ℓ slowly varying.

In this case, $\mathrm{Var}(K(n)) = \mathbb{E}K(2n) - \mathbb{E}K(n) \to \frac{\log(2)}{\log(1/1-q)}$.

Proposition 3.3. *When the sampling distribution is geometric with parameter $q \in (0,1)$, letting $M_n = \max(X_1, \ldots, X_n)$,*

$$\mathbb{E}M_n \geq \mathbb{E}K_n \geq \mathbb{E}M_n - \frac{1-q}{q^2} .$$

In the case of geometric frequencies, the missing mass can fluctuate widely with respect to its expectation, and one cannot expect to obtain sub-gamma concentration with both the correct variance proxy and scale factor $1/n$. Indeed, intuitively, the symbol which primarily contributes to the missing mass' fluctuations, is the quantile of order $1 - 1/n$. With $F(k) = \sum_{j=1}^{k} p_j$, and $F^{\leftarrow}$ the generalized inverse of F,

$$\begin{aligned}j^* = F^{\leftarrow}(1-1/n) &= \inf\{j \geq 1, F(j) \geq 1 - 1/n\} \\ &= \inf\{j \geq 1, \sum_{k>j} p_k \leq 1/n\} .\end{aligned}$$

Omitting the slowly varying functions, when $\vec{\nu}(1/\cdot) \in \mathrm{RV}_\alpha$, $0 < \alpha < 1$, j^* is of order $n^{\frac{\alpha}{1-\alpha}}$ and p_{j^*} is of order $n^{-\frac{1}{1-\alpha}}$. The closer to 1 is α, the smaller the probability of j^*. When α goes to 0, this probability becomes $1/n$. With geometric frequencies, j^* is $\frac{\log(n)}{\log(1/1-q)}$ and p_{j^*} is $\frac{q}{n(1-q)}$. Hence, around the quantile of order $1 - 1/n$, there are symbols which may contribute significantly to the missing mass' fluctuations.

Another interesting case consists of distributions which are very light-tailed, in the sense that $\frac{p_{k+1}}{p_k} \to 0$ when $k \to \infty$. An example of these is the Poisson distribution $\mathcal{P}(\lambda)$, for which $\frac{p_{k+1}}{p_k} = \frac{\lambda}{k} \underset{k\to+\infty}{\longrightarrow} 0$. The next proposition shows that for such concentrated distributions, the missing mass essentially concentrates on two points.

Proposition 3.4. *In the infinite urns scheme with probability mass function $(p_k)_{k\in\mathbb{N}}$, if $p_k > 0$ for all k and $\lim_{k\to\infty} \frac{p_{k+1}}{p_k} = 0$, then there exists a sequence of integers $(u_n)_{n\in\mathbb{N}}$ such that*

$$\lim_{n\to\infty} \mathbb{P}\left\{M_{n,0} \in \{\overline{F}(u_n), \overline{F}(u_n+1)\}\right\} = 1 ,$$

where $\overline{F}(k) = \sum_{j>k} p_j$.

4 Applications

4.1 Estimating the regular variation index

When working in the regular variation setting, the most basic estimation task is to estimate the regular variation index α. We already mentioned in Section 1 the fact that, when $\vec{\nu} \in \mathrm{RV}_\alpha, \alpha \in (0,1)$, the ratio $K_{n,1}/K_n$ provides a consistent estimate of α. This is actually only one among a family of estimators of α that one may construct. The next result shows this, and is a direct consequence of Proposition 3.2.

Proposition 4.1. *If $\vec{\nu} \in \mathrm{RV}_\alpha, \alpha \in (0,1]$, then for all $r \geq 1$*

$$\frac{rK_{n,r}}{K_{n,\bar{r}}}$$

is a strongly consistent estimator of α.

Thus, writing $k_n = \max\{r, K_{n,r} > 0\}$, at time n, we can have up to k_n non-trivial estimators of α. One would expect these estimators to offer various bias-variance trade-offs, and one could ostensibly select an "optimal" r via model selection.

4.2 Estimating the missing mass

The Good-Turing estimation problem [82] is that of estimating $M_{n,r}$ from the observation $(X_1, X_2, \cdots, X_n)$. For large scores r, designing estimators for $M_{n,r}$ is straightforward: we assume that the empirical distribution mimics the sampling distribution, and that the empirical probabilities $\frac{rK_{n,r}}{n}$ are likely to be good estimators. The question is more delicate for rare events. In particular, for $r = 0$, it may be a bad idea to assume that there is no missing mass $M_{n,0} = 0$, that is to assign a zero probability to the set of symbols that do not appear in the sample. Various "smoothing" techniques have thus developed, in order to adjust the maximum likelihood estimator and obtain more accurate probabilities.

In particular, Good-Turing estimators attempt to estimate $(M_{n,r})_r$ from $(K_{n,r})_r$ for all r. They are defined as

$$G_{n,r} = \frac{(r+1)K_{n,r+1}}{n}\,.$$

The rationale for this choice comes from the following observations.

$$\mathbb{E}G_{n,0} = \frac{\mathbb{E}\left[K_{n,1}\right]}{n} = \mathbb{E}M_{n-1,0} = \mathbb{E}M_{n,0} + \frac{\mathbb{E}M_{n,1}}{n}\,, \tag{4.1}$$

and

$$\mathbb{E}G_{n,r} = \frac{(r+1)\mathbb{E}K_{n,r+1}}{n} = \mathbb{E}M_{n-1,r}\,. \tag{4.2}$$

In the Poisson setting, there is no bias: $\mathbb{E}G_r(t) = (r+1)\frac{\mathbb{E}K_{r+1}(t)}{t} = \mathbb{E}M_r(t)$.

Here, we primarily focus on the estimation of the missing masses $M_{n,0}$ and $M_0(n)$, though most of the methodology extends also to $r > 0$, with the appropriate concentration results. From (4.1) and (4.2), Good-Turing estimators look like slightly biased estimators of the relevant masses. In particular, the bias $\mathbb{E}G_{n,0} - \mathbb{E}M_{n,0}$ is always positive but smaller than $1/n$. It is however far from obvious to determine scenarios where these estimators are consistent and where meaningful confidence regions can be constructed.

When trying to estimate the missing mass $M_{n,0}$ or $\mathbb{E}M_{n,0}$, consistency needs to be redefined since the estimand is not a fixed parameter of interest but a random quantity whose expectation further depends on n. Additive consistency, that is bounds on $\widehat{M}_{n,0} - M_{n,0}$ is not a satisfactory notion, because, as $M_{n,0}$ tends to 0, the trivial constant estimator 0 would be universally asymptotically consistent. Relative consistency,

that is control on $(\widehat{M}_{n,0} - M_{n,0})/M_{n,0}$ looks like a much more reasonable notion. It is however much harder to establish.

In order to establish relative consistency of a missing mass estimator, we have to check that $\mathbb{E}[\widehat{M}_{n,0} - M_{n,0}]$ is not too large with respect to $\mathbb{E}M_{n,0}$, and that both $\widehat{M}_{n,0}$ and $M_{n,0}$ are concentrated around their mean values.

As shown in [127], the Good-Turing estimator of the missing mass is not universally consistent in this sense. This occurs principally in very light tails, such as those described in Section 3.4.1.

Proposition 4.2. *[127] When the sampling distribution is geometric with small enough $q \in (0,1)$, there exists $\eta > 0$, and a subsequence n_i such that for i large enough, $G_{n_i,0}/M_{n_i,0} = 0$ with probability no less than η.*

On the other hand, the concentration result of Corollary 3.1 gives a law of large numbers for $M_{n,0}$ (by a direct application of the Borel-Cantelli lemma), which in turn implies the strong multiplicative consistency of the Good-Turing estimate.

Corollary 4.1. *We have the following two regimes of consistency for the Good-Turing estimator of the missing mass.*

(i) If the counting function $\vec{\nu}$ is such that $\mathbb{E}K_{n,2}/\mathbb{E}K_{n,1}$ remains bounded and $\mathbb{E}K_{n,1} \to +\infty$ (in particular, when $\vec{\nu}$ is regularly varying with index $\alpha \in (0,1]$ or $\alpha = 0$ and $\vec{\nu} \in \Pi_a$ with $a \to \infty$),

$$\frac{M_{n,0}}{\mathbb{E}M_{n,0}} \xrightarrow{\mathbb{P}} 1\,,$$

and the Good-Turing estimator of $M_{n,0}$ defined by $G_{n,0} = K_{n,1}/n$, is multiplicatively consistent in probability:

$$\frac{G_{n,0}}{M_{n,0}} \xrightarrow{\mathbb{P}} 1\,.$$

(ii) If furthermore $\mathbb{E}K_{n,\overline{2}}/\mathbb{E}K_{n,1}$ remains bounded and if, for all $c > 0$, $\sum_{n=0}^{\infty} \exp(-c\mathbb{E}K_{n,1}) < \infty$ (in particular, when $\vec{\nu}$ is regularly varying with index $\alpha \in (0,1]$), then these two convergences occur almost surely.

Remark 4.1. *One needs to make assumptions on the sampling distribution to guarantee the consistency of the Good-Turing estimator. In fact, there is no hope to find a universally consistent estimator of the missing mass without any such restrictions, as shown recently by Mossel and Ohannessian [123].* □

Consistency is a desirable property, but the concentration inequalities provide us with more power, in particular in terms of giving confidence intervals that are asymptotically tight. For brevity, we focus here on the Poisson setting to derive concentration inequalities which in turn yield confidence intervals. A similar, but somewhat more tedious, methodology yields confidence intervals in the binomial setting as well.

4.2.1 Concentration inequalities for $G_0(n) - M_0(n)$

The results of this section are given in the Poisson setting for simplicity. Let us note however that the same type of result hold in the binomial setting. In the Poisson setting, the analysis of the Good-Turing estimator is illuminating. As noted earlier, the first pleasant observation is that the Good-Turing estimator is an unbiased estimator of the missing mass. Second, the variance of $G_0(n) - M_0(n)$ is simply related to occupancy counts:

$$\mathrm{Var}(G_0(n) - M_0(n)) = \frac{1}{n^2}\left(\mathbb{E}K_1(n) + 2\mathbb{E}K_2(n)\right)\,. \tag{4.3}$$

Third, simple yet often tight concentration inequalities can be obtained for $G_0(n) - M_0(n)$.

Proposition 4.3. *The random variable $G_0(n) - M_0(n)$ is sub-gamma on the right tail with variance factor $\mathrm{Var}(G_0(n) - M_0(n))$ and scale factor $1/n$, and sub-gamma on the left tail with variance factor $3\mathbb{E}K(n)/n^2$ and scale factor $1/n$.*

For all $\lambda \geq 0$,

(i) $\log \mathbb{E}\, \mathrm{e}^{\lambda(G_0(n)-M_0(n))} \leq \mathrm{Var}(G_0(n) - M_0(n)) n^2 \phi\left(\frac{\lambda}{n}\right)$,

(ii) $\log \mathbb{E}\, \mathrm{e}^{\lambda(M_0(n)-G_0(n))} \leq \frac{3\mathbb{E}K(n)}{2n^2} \frac{\lambda^2}{1-\lambda/n}$.

We are now in a position to build confidence intervals for the missing mass.

Proposition 4.4. *With probability larger than $1 - 4\delta$, the following hold*

$$M_0(n) \leq G_0(n) + \frac{1}{n}\left(\sqrt{6K(n)\log\frac{1}{\delta}} + 5\log\frac{1}{\delta}\right) .$$

$$M_0(n) \geq G_0(n) - \frac{1}{n}\left(\sqrt{2(K_1(n) + 2K_2(n))\log\frac{1}{\delta}} + 4\log\frac{1}{\delta}\right) .$$

To see that these confidence bounds are asymptotically tight, consider the following central limit theorem. A similar results can be paralleled in the binomial setting.

Proposition 4.5. *If the counting function $\vec{\nu}$ is regularly varying with index $\alpha \in (0, 1]$, the following central limit theorem holds for the ratio $G_0(n)/M_0(n)$:*

$$\frac{\mathbb{E}K_1(n)}{\sqrt{\mathbb{E}K_1(n) + 2\mathbb{E}K_2(n)}}\left(\frac{G_0(n)}{M_0(n)} - 1\right) \rightsquigarrow \mathcal{N}(0, 1) .$$

4.2.2 Some simulations

We are interested in the behaviour of the ratio $\frac{G_{n,0}}{M_{n,0}}$ for various sampling distributions. For each sample size, we collect $1e4$ values.

We first consider the Poisson distribution with mean 1, illustrating a phenomenon of concentration on very few points, and the failure of the Good-Turing estimator to estimate the missing mass.

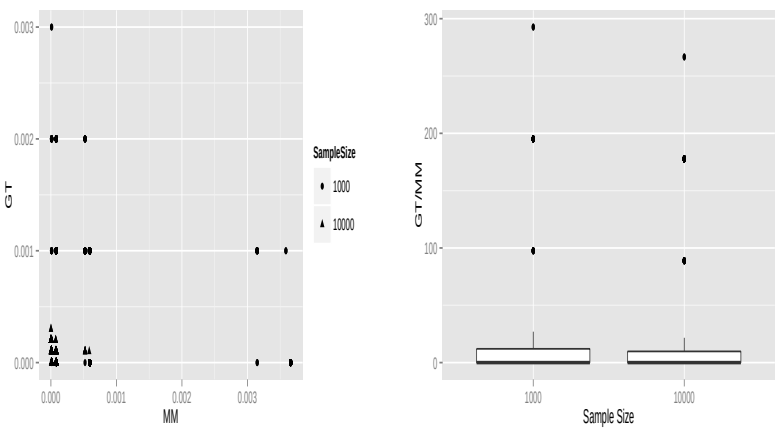


Figure 1.1 – Poisson

The case of the Geometric distribution with parameter 1/2 illustrates the irregular behaviour of the ratio $G_{n,0}/M_{n,0}$ in the lower part of Gumbel's domain. Each violin plot represents a density estimate for $G_{n,0}/M_{n,0}$.

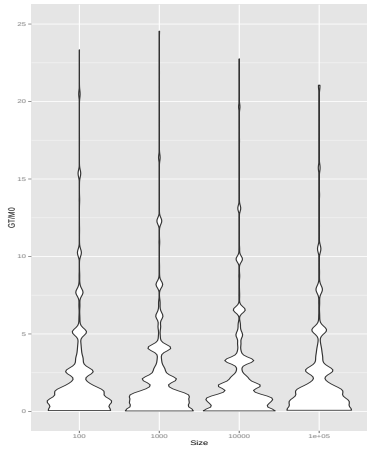


Figure 1.2 – Geometric

In contrast, the case of the discretized Pareto distribution with scale, location and shape parameter equal to 1 confirms the performances of the Good-Turing estimator on heavy-tailed distribution.

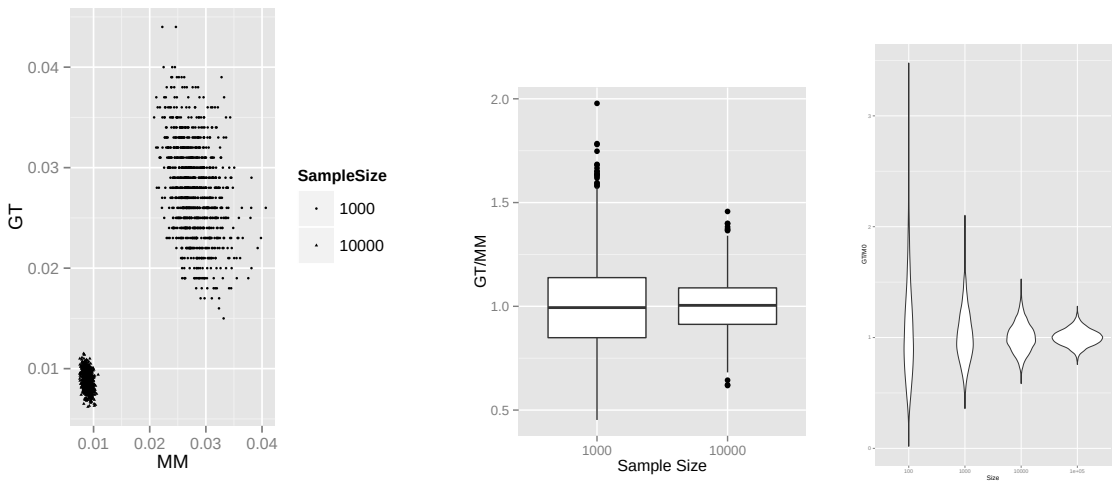


Figure 1.3 – Pareto

4.3 Estimating the number of species

Fisher's number of species problem Fisher et al. [73] consists of estimating $K_{(1+\tau)n} - K_n$ for $\tau > 0$, the number of distinct new species one would observe if the data collection runs for an additional fraction τ of time. This was posed primarily within the Poisson model in the original paper [73] and later by [67], but the same question may also be asked in the binomial model. The following estimates come from straightforward computations on the asymptotics given in Theorems 3.2, 3.3 and 3.4.

Proposition 4.6. *If the counting function $\vec{\nu}$ is regularly varying with index $\alpha \in (0, 1]$, letting $\hat{\alpha}$ be any*

of the estimates $rK_{n,r}/K_{n,\bar{r}}$ of α from Proposition 4.1, then any of the following quantities

$$(\tau^{\hat{\alpha}}-1)K_n\,,\ \frac{\tau^{\hat{\alpha}}-1}{\hat{\alpha}}K_{n,1}\,,\ \text{and}\ \left(\prod_{k=2}^{r}\frac{k}{k-1-\hat{\alpha}}\right)\frac{\tau^{\hat{\alpha}}-1}{\hat{\alpha}}K_{n,r}\,,\ r\geq 2\,,$$

is a strongly consistent estimate of $K_{\tau n}-K_n$, the number of newly discovered species when the sample size is multiplied by τ.

If the counting function $\vec{\nu}$ is in Π_{ℓ_0}, with $\ell_0(n)\to+\infty$, then, for each $r\geq 1$,

$$\log(\tau)rK_{n,r}$$

is an estimate of $K_{\tau n}-K_n$, consistent in probability.

5 Discussion

To conclude this section, we review our results in a larger context, and propose some connections, extensions, and open problems.

5.1 The cost of Poissonization and negative correlation

Resorting to Poissonization or negative correlation may have a price. It may lead to variance overestimates. [80, Lemma 1] asserts that for some constant c

$$|\mathrm{Var}(K(n))-\mathrm{Var}(K_n)|\leq\frac{c}{n}\max\left(1,\mathbb{E}K_1(n)^2\right)$$

This bound conveys a mixed message. As $\mathbb{E}K_1(n)/n$ tends to 0, it asserts that

$$|\mathrm{Var}(K(n))-\mathrm{Var}(K_n)|\,/\mathbb{E}K_1(n)$$

tends to 0. But there exist scenarios where $\mathbb{E}K_1(n)^2/n$ tends to infinity. It is shown in [80] that $\mathbb{E}K_1(n)^2/(n\mathrm{Var}(K(n)))$ tends to 0, so that, as soon as $n\mathrm{Var}(K(n))$ tends to infinity (which might not always be the case), the two variances $\mathrm{Var}(K_n)$ and $\mathrm{Var}(K(n))$ are asymptotically equivalent.

It would be interesting to find necessary and sufficient conditions under which there is equivalence. Though these aren't generally known, it is instructive to compare $\mathrm{Var}(K(n))$, $\mathrm{Var}(K_n)$ and $\mathrm{Var}^{\mathrm{ind}}(K_n)$ the variance upper bound obtained from negative correlation by bounding their differences. For instance, one can show that for any sampling distribution we have:

$$\frac{\mathbb{E}K_2(2n)}{n}\leq\mathrm{Var}(K(n))-\mathrm{Var}^{\mathrm{ind}}(K_n)\leq\frac{2\mathbb{E}K_2(n)}{n}\,,$$

and

$$0\leq\mathrm{Var}^{\mathrm{ind}}(K_n)-\mathrm{Var}(K_n)\leq\frac{(\mathbb{E}K_{n,1})^2}{n}-\frac{\mathbb{E}K_{2n,2}}{2n-1}\,.$$

These bounds are insightful but, without any further assumptions on the sampling distribution, they are not sufficient to prove asymptotic equivalence.

5.2 Extensions of regular variation

The regular variation hypothesis is an elegant framework, allowing one to derive, thanks to Karamata and Tauberian Theorems, simple and intelligible equivalents for various moments. As we have seen in

Remark 3.1, regular variation comes very close to being a necessary condition for exponential concentration. It may however seem too stringent. Without getting too specific, let us mention that other less demanding hypotheses also yield the asymptotic relative orders that work in favor of the concentration of the missing mass. For instance, referring back to Remark 3.1, one could instead ask for:

$$0 < \liminf_{t\to\infty} \frac{\Phi_{\overline{2}}(t)}{\Phi_2(t)} \leq \limsup_{t\to\infty} \frac{\Phi_{\overline{2}}(t)}{\Phi_2(t)} < \infty\,.$$

Recalling that $\Phi'_{\overline{2}}(t) = \frac{2\Phi_2(t)}{t}$, and applying Corollary 2.6.2. of [30], one obtains that Φ_2 is in the class OR of O-regularly varying functions and $\Phi_{\overline{2}}$ is in the class ER of extended regularly varying functions, that is, for all $\lambda \geq 1$

$$0 < \liminf_{t\to\infty} \frac{\Phi_2(\lambda t)}{\Phi_2(t)} \leq \limsup_{t\to\infty} \frac{\Phi_2(\lambda t)}{\Phi_2(t)} < \infty\,,$$

and

$$\lambda^d \leq \liminf_{t\to\infty} \frac{\Phi_{\overline{2}}(\lambda t)}{\Phi_{\overline{2}}(t)} \leq \limsup_{t\to\infty} \frac{\Phi_{\overline{2}}(\lambda t)}{\Phi_{\overline{2}}(t)} \leq \lambda^c\,,$$

for some constants c and d. Observe that this result, which is the equivalent of Karamata's Theorem, differs from the regular variation setting, in the sense that the control on the derivative Φ_2 is looser than the one on $\Phi_{\overline{2}}$, whereas, in the Karamata Theorem, both the function and its derivative inherit the regular variation property.

We can in turn show that $\Phi(t) = \mathbb{E}K(t)$ is in the class OR and, by Theorem 2.10.2 of [30], this is equivalent to $\vec{\nu}(1/\cdot) \in OR$, as Φ is the Laplace-Stieltjes transform of $\vec{\nu}$.

5.3 Random measures

As noted by [80], the asymptotics for the moments of the occupancy counts in the regular variation setting is still valid when the frequencies $(p_j)_{j\geq 1}$ are random, in which case the measure ν is defined by

$$\mathbb{E}\left[\sum_{j=1}^{\infty} f(p_j)\right] = \int_0^1 f(x)\nu(\mathrm{d}x)\,,$$

for all functions $f \geq 0$. We can also define the measure ν_1 by

$$\mathbb{E}\left[\sum_{j=1}^{\infty} p_j f(p_j)\right] = \int_0^1 f(x)\nu_1(\mathrm{d}x)\,,$$

for all functions $f \geq 0$. This measure corresponds to the distribution of the frequency of the first discovered symbol.

For instance, when $(p_j)_{j\geq 1}$ are Poisson-Dirichlet(α,0) with $0 < \alpha < 1$, the measure $\nu_1(\mathrm{d}x)$ is the size-biased distribution of PD$(\alpha, 0)$, that is Beta $\mathcal{B}(1-\alpha,\alpha)$ [see 138]. Thus we have:

$$\begin{aligned} \nu_1[0,x] &= \frac{1}{\mathcal{B}(1-\alpha,\alpha)} \int_0^x t^{-\alpha}(1-t)^{\alpha-1}\mathrm{d}t \\ &\underset{x\to 0}{\sim} \frac{x^{1-\alpha}}{(1-\alpha)\mathcal{B}(1-\alpha,\alpha)} \end{aligned}$$

and, by [Proposition 13 80], this is equivalent to

$$\vec{\nu}(x) \underset{x\to 0}{\sim} \frac{1}{\alpha\mathcal{B}(1-\alpha,\alpha)} x^{-\alpha}\,.$$

Thus, denoting by $N(x)$ the random number of frequencies p_j which are larger than x, the expectation $\vec{\nu}(x) = \mathbb{E}N(x)$ is regularly varying. One can also show that the mass-partition mechanism of the distribution $\mathrm{PD}(\alpha, 0)$ almost surely generates $N(x)$ to be regularly varying. To see this, refer to [Proposition 10 138] or to [Proposition 2.6 29] which assert that the limit

$$L := \lim_{n\to\infty} np_n^\alpha$$

exists almost surely. This is equivalent to

$$N(x) \underset{x\to 0}{\sim} x^{-\alpha} L, \quad \text{almost surely.}$$

The $\mathrm{PD}(\alpha, 0)$ distribution can be generated through a Poisson process with intensity measure $\nu([x, \infty]) = cx^{-\alpha}$. Without entering into further details, let us mention that similar almost sure results hold even when the intensity measure ν is not a strict power, but satisfies the property

$$\nu([x, \infty]) \underset{x\to 0}{\sim} x^{-\alpha}\ell(x)\,,$$

with ℓ slowly varying, [Section 6 81]. Working with a regular variation hypothesis thus gives us more flexibility than assuming specific Bayesian priors.

6 Proofs

6.1 Fundamental techniques

6.1.1 Efron-Stein-Steele inequalities

Our variance bounds mostly follow from the Efron-Stein-Steele Inequality [66], which states that when a random variable is expressed as a function of many independent random variables, its variance can be controlled by the sum of the local fluctuations.

Theorem 6.1. *Let $\mathcal{X}$ be some set, $(X_1, X_2, \cdots, X_n)$ be independent random variables taking values in $\mathcal{X}$, $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function of n variables, and $Z = f(X_1, X_2, \cdots, X_n)$.*

For all $i \in \{1, \cdots, n\}$, let $X^{(i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$ and $\mathbb{E}^{(i)}Z = \mathbb{E}[Z|X^{(i)}]$. Then, letting $v = \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}^{(i)}Z)^2]$,

$$\mathrm{Var}[Z] \le v\,.$$

If $X_1', \cdots, X_n'$ are independent copies of $X_1, \cdots, X_n$, then letting $Z_i' = f(X_1, \cdots, X_{i-1}, X_i', X_{i+1}, \cdots, X_n)$,

$$v = \sum_{i=1}^n \mathbb{E}[(Z - Z_i')_+^2] \le \sum_{i=1}^n \mathbb{E}[(Z - Z_i)^2]\,,$$

where the random variables Z_i are arbitrary $X^{(i)}$-measurable and square-integrable random variables.

6.1.2 Negative association

The random variables K_n, $K_{n,r}$, and $M_{n,r}$ are sums or weighted sums of Bernoulli random variables. These summands depend on the scores $(X_{n,j})_{j\ge 1}$ and therefore are not independent. Transforming the fixed-n binomial setting into a continuous time Poisson setting is one way to circumvent this problem. This is the Poissonization method. In this setting, the score variables $(X_j(n))_{j\ge 1}$ are independent Poisson

variables with respective means np_j. Results valid for the Poisson setting can then be transferred to the fixed-n setting, up to approximation costs. For instance, [80] (Lemma 1) provide bounds on the discrepancy between expectations and variances in the two settings. (See also our discussion in Section 5.1).

Another approach to deal with the dependence is to invoke the notion of negative association, which provides a systematic comparison between moments of certain monotonic functions of the occupancy scores. In our present setting, this will primarily be useful for bounding the logarithmic moment generating function, which is an expectation of products, by products of expectations, thus recovering the structure of independence. This has already been used to derive exponential concentration for occupancy counts [see 64, 122, 127, 148]. It is also useful for bounding variances. We use this notion throughout the proofs, and therefore present it here formally.

Definition 3 (NEGATIVE ASSOCIATION)**.** Real-valued random variables $Z_1, \dots, Z_K$ are said to be negatively associated if, for any two disjoint subsets A and B of $\{1, \dots, K\}$, and any two real-valued functions $f : \mathbb{R}^{|A|} \mapsto \mathbb{R}$ and $g : \mathbb{R}^{|B|} \mapsto \mathbb{R}$ that are both either coordinate-wise non-increasing or coordinate-wise non-decreasing, we have:

$$\mathbb{E}\left[f(Z_A) \cdot g(Z_B)\right] \leq \mathbb{E}\left[f(Z_A)\right] \cdot \mathbb{E}\left[g(Z_B)\right] .$$

In particular, as far as concentration properties are concerned, sums of negatively associated variables can only do better than sums of independent variables.

Theorem 6.2. *[64] For each $n \in \mathbb{N}$, the occupancy scores $(X_{n,j})_{j\geq 1}$ are negatively associated.*

As monotonic functions of negatively associated variables are also negatively associated, the variables $(\mathbb{1}_{\{X_{n,j}>0\}})_{j\geq 1}$ (respectively $(\mathbb{1}_{\{X_{n,j}=0\}})_{j\geq 1}$) are negatively associated as increasing (respectively decreasing) functions of $(X_{n,j})_{j\geq 1}$. This is of pivotal importance for our proofs of concentration results for K_n and $M_{n,0}$. For $r \geq 1$, the variables $(\mathbb{1}_{\{X_{n,j}=r\}})_{j\geq 1}$ appearing in $K_{n,r}$ are not negatively associated. However, following [127], one way to deal with this problem is to observe that

$$K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}} ,$$

recalling that $K_{n,\overline{r}} = \sum_{j=1}^{\infty} \mathbb{1}_{\{X_{n,j}\geq r\}}$ is the number of urns that contain at least r balls and that the Bernoulli variables appearing in $K_{n,\overline{r}}$ are negatively associated.

6.1.3 Regular variation theorems

We state some useful properties of regularly varying functions, which are used in this chapter, as well as in the next one [See 30, 55, for proofs and refinements]..

Theorem 6.3. (KARAMATA'S THEOREM) *Let $f \in \mathrm{RV}_\alpha$ and assume that there exists $t_0 > 0$ such that f is positive and locally bounded on $[t_0, +\infty[$.*

(i) If $\alpha \geq -1$, then

$$\int_{t_0}^{t} f(s)\,\mathrm{d}s \underset{t\to+\infty}{\sim} \frac{tf(t)}{\alpha+1} .$$

(ii) If $\alpha < -1$, or $\alpha = 1$ and $\int_0^\infty f(s)\,\mathrm{d}s < \infty$, then

$$\int_{t}^{+\infty} f(s)\,\mathrm{d}s \underset{t\to+\infty}{\sim} \frac{tf(t)}{-\alpha-1} .$$

Theorem 6.4. (POTTER-DREES INEQUALITIES.)

(i) If $f \in \mathrm{RV}_\alpha$, then for all $\delta > 0$, there exists $t_0 = t_0(\alpha)$, such that for all t, x: $\min(t, tx) > t_0$,

$$(1-\delta)x^\alpha \min\left(x^\delta, x^{-\delta}\right) \le \frac{f(tx)}{f(t)} \le (1+\delta)x^\alpha \max\left(x^\delta, x^{-\delta}\right)$$

(ii) If $\ell \in \Pi_{\ell_0}$, then for all δ_1, δ_2, there exists t_0 such that for all $t \ge t_0$, for all $x \ge 1$,

$$(1-\delta_2)\frac{1-x^{\delta_1}}{\delta_1} - \delta_2 < \frac{\ell(tx)-\ell(t)}{\ell_0(t)} < (1+\delta_2)\frac{x^{\delta_1}-1}{\delta_1} + \delta_2\,.$$

6.2 Occupancy counts

6.2.1 Variance bounds for occupancy counts

Proof of Proposition 2.1. Note that in the Poisson setting,

$$\frac{\mathrm{d}\mathbb{E}K(t)}{\mathrm{d}t} = \frac{\mathbb{E}K_1(t)}{t} = \mathbb{E}M_0(t)\,.$$

This entails

$$\begin{aligned}\mathrm{Var}(K(n)) &= \sum_{j=1}^{\infty} \mathrm{e}^{-np_j}(1-\mathrm{e}^{-np_j}) \\ &= \mathbb{E}K(2n) - \mathbb{E}K(n) \\ &= \int_n^{2n} \mathbb{E}M_0(t)\,\mathrm{d}s\,.\end{aligned}$$

Now, as $\mathbb{E}M_0(t)$ is non-increasing,

$$\frac{\mathbb{E}K_1(2n)}{2} = n\mathbb{E}M_0(2n) \le \mathrm{Var}(K(n)) \le n\mathbb{E}M_0(n) = \mathbb{E}K_1(n)\,.$$

Moving on to the binomial setting, let K_n^i denote the number of occupied urns when the ith ball is replaced by an independent copy. Then

$$\mathrm{Var}(K_n) \le \mathbb{E}\left[\sum_{i=1}^{n}(K_n - K_n^i)_+^2\right],$$

where $(K_n - K_n^i)_+$ denotes the positive part. Now, $K_n - K_n^i$ is positive if and only if ball i is moved from a singleton into in a non-empty urn. Thus $\mathrm{Var}(K_n) \le \mathbb{E}\left[K_{n,1}(1-M_{n,0})\right]$. □

Proof of Proposition 2.2. The bound $r\mathbb{E}K_{n,r}$ follows from the Efron-Stein inequality: denoting by $K_{n,\overline{r}}^{(i)}$ the number of cells with occupancy score larger than r when ball i is removed, then

$$K_{n,\overline{r}} - K_{n,\overline{r}}^{(i)} = \begin{cases} 1 & \text{if ball } i \text{ is in a } r\text{-ton} \\ 0 & \text{otherwise.}\end{cases}$$

And thus, we get $\sum_{i=1}^{n}(K_{n,\overline{r}} - K_{n,\overline{r}}^{(i)})^2 = rK_{n,r}$.

The second bound follows from the negative association of the variables $(\mathbb{1}_{\{X_{n,j}\ge r\}})_j$ (negative corre-

lation is actually sufficient):

$$\mathrm{Var}\left(\sum_{j=1}^{\infty} \mathbb{1}_{\{X_{n,j}\geq r\}}\right) \leq \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{1}_{\{X_{n,j}\geq r\}}) \leq \mathbb{E}K_{n,\overline{r}}\,. \qquad \square$$

6.2.2 Concentration inequalities for occupancy counts

Proof of Proposition 2.4. Let $X_{n,j}$ denote the occupancy score of cell $j, j \in \mathbb{N}$, then

$$K_{n,\overline{r}} = \sum_{j=1}^{\infty} \mathbb{I}_{\{X_{n,j}\geq r\}}\,.$$

As noted in Section 6.1.2, $K_{n,\overline{r}}$ is a sum of negatively associated Bernoulli random variables. Moreover, the Efron-Stein inequality implies that for each $j \in \mathbb{N}$,

$$\mathrm{Var}(\mathbb{I}_{\{X_{n,j}\geq r\}}) \leq r\mathbb{E}\mathbb{I}_{\{X_{n,j}=r\}}\,.$$

Thus we have

$$\begin{aligned}
\log \mathbb{E}\,\mathrm{e}^{\lambda(K_{n,\overline{r}}-\mathbb{E}K_{n,\overline{r}})} &\leq \sum_{j=1}^{\infty} \log \mathbb{E}\,\mathrm{e}^{\lambda(\mathbb{I}_{\{X_{n,j}\geq r\}}-\mathbb{E}\mathbb{I}_{\{X_{n,j}\geq r\}})} \\
&\leq \sum_{j=1}^{\infty} \mathrm{Var}(\mathbb{I}_{\{X_{n,j}\geq r\}})\phi(\lambda) \\
&\leq \phi(\lambda)\sum_{j=1}^{\infty} r\mathbb{E}\mathbb{1}_{\{X_{n,j}=r\}} \\
&= \phi(\lambda) r\mathbb{E}K_{n,r}\,,
\end{aligned}$$

where the first inequality comes from negative association, the second inequality is Bennett's inequality for Bernoulli random variables, and the last inequality comes from the Efron-Stein inequality. The other bound comes from the fact that $\mathrm{Var}(\mathbb{I}_{\{X_{n,j}\geq r\}}) \leq \mathbb{E}\mathbb{I}_{\{X_{n,j}\geq r\}}$. $\square$

Proof of Proposition 2.5. As $K_{n,r} = K_{n,\overline{r}} - K_{n,\overline{r+1}}$,

$$\begin{aligned}
&\{K_{n,r} \geq \mathbb{E}K_{n,r} + x\} \\
&\qquad \subseteq \left\{K_{n,\overline{r}} \geq \mathbb{E}K_{n,\overline{r}} + \frac{x}{2}\right\} \cup \left\{K_{n,\overline{r+1}} \leq \mathbb{E}K_{n,\overline{r+1}} - \frac{x}{2}\right\}\,.
\end{aligned}$$

By Proposition 2.4, Bernstein inequalities hold for both $K_{n,\overline{r}}$ and $K_{n,\overline{r+1}}$, with variance proxies $\mathbb{E}rK_{n,r}$ (or $\mathbb{E}K_{n,\overline{r}}$) and $(r+1)K_{n,r+1}$ (or $\mathbb{E}K_{n,\overline{r+1}} \leq \mathbb{E}K_{n,\overline{r}}$) respectively. Hence,

$$\begin{aligned}
&\mathbb{P}\{K_{n,r} \geq \mathbb{E}K_{n,r} + x\} \\
&\quad \leq \exp\left(-\frac{x^2/4}{2(r\mathbb{E}K_{n,r} + x/6)}\right) + \exp\left(-\frac{x^2/4}{2((r+1)\mathbb{E}K_{n,r+1})}\right) \\
&\quad \leq 2\exp\left(-\frac{x^2/4}{2(\max(r\mathbb{E}K_{n,r}, (r+1)K_{n,r+1}) + x/6)}\right)\,.
\end{aligned}$$

The same reasoning works for the alternative variance proxies and for the left tails. $\square$

Proof of Theorem 2.1. For each $r \in \mathcal{J}$, $K_r(n)$ satisfies a Bennett and thus a Bernstein Inequality with

variance proxy $\mathrm{Var}(K_r(n)) = \sigma_r^2$ and scale proxy 1. Hence, for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\, \mathrm{e}^{\frac{\lambda}{\sigma_r}(K_r(n) - \mathbb{E}K_r(n))} \leq \frac{\lambda^2}{2(1 - \frac{\lambda}{\sigma_r})}\,.$$

Then by Corollary 2.6 in [40],

$$\mathbb{E}Z = \mathbb{E} \max_{r \in \mathcal{J}} \frac{K_r(n) - \mathbb{E}K_r(n)}{\sigma_r} \leq \sqrt{2 \log |\mathcal{J}|} + \frac{\log |\mathcal{J}|}{\min_{r \in \mathcal{J}} \sigma_r}\,.$$

As summands are not identically distributed, the Bousquet-Rio bound on the variance of suprema of bounded centered empirical processes is not applicable, but the variance bound can be obtained as a corollary of the main result from [105] (Theorems 12.9 in [40]) The proof of the two tail bounds consist in carefully invoking the Klein-Rio and Samson inequalities (Theorems 12.9 and 12.11 in [40]). □

6.3 Missing mass

6.3.1 Variance bounds for the missing mass

Proof of Proposition 2.6. In the Poisson setting,

$$\mathrm{Var}(M_0(n)) = \sum_{j=1}^{\infty} p_j^2\, \mathrm{e}^{-np_j} \left(1 - \mathrm{e}^{-np_j}\right) \leq \sum_{j=1}^{\infty} p_j^2\, \mathrm{e}^{-np_j} = \frac{2}{n^2} \mathbb{E}K_2(n)\,.$$

In the binomial setting, by negative correlation,

$$\mathrm{Var}(M_{n,0}) \leq \sum_{j=1}^{\infty} p_j^2\, (1 - (1 - p_j)^n)\, (1 - p_j)^n \leq \sum_{j=1}^{\infty} p_j^2\, \mathrm{e}^{-np_j} = \frac{2}{n^2} \mathbb{E}K_2(n)\,.$$

□

6.3.2 Concentration inequalities for the missing mass

Proof of Proposition 2.7. Letting $Y_j = \mathbb{1}_{\{X_{n,j}=0\}}$, we have, for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] &= \log \mathbb{E}\left[\mathrm{e}^{\lambda \sum_{j=1}^{\infty} p_j (Y_j - \mathbb{E}Y_j)}\right] \\
&\leq \sum_{j=1}^{\infty} \log \mathbb{E}\left[\mathrm{e}^{\lambda p_j (Y_j - \mathbb{E}[Y_j])}\right] \\
&\leq \sum_{j=1}^{\infty} \mathrm{Var}(Y_j) \phi(\lambda p_j)\,,
\end{aligned}$$

where the first inequality comes from negative association, and the second is Bennett's inequality for Bernoulli random variables.

Noting that $\lim_{\lambda \to 0_-} \phi(\lambda)/\lambda^2 = \lim_{\lambda \to 0_+} \phi(\lambda)/\lambda^2 = 1/2$, the function $\lambda \mapsto \phi(\lambda)/\lambda^2$ has a continuous increasing extension on $\mathbb{R}$. Hence, for $\lambda \leq 0$, we have $\phi(\lambda) \leq \lambda^2/2$, and

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})}\right] \leq \sum_{j=1}^{\infty} p_j^2 \mathrm{Var}(Y_j) \frac{\lambda^2}{2}\,.$$

Recall that $\sum_{j=1}^{\infty} p_j^2 \mathrm{Var}(Y_j) \leq 2\mathbb{E}K_2(n)/n^2$ (Proposition 2.6). □

Proof of Proposition 2.8. From the beginning of the proof of Proposition 2.7, that is, thanks to negative association and to the fact that each Bernoulli random variable satisfies a Bennett inequality,

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \quad \leq \quad \sum_{j=1}^{\infty} \mathrm{e}^{-np_j}\,\phi(\lambda p_j)\,.$$

Now, using the power series expansion of ϕ,

$$\begin{aligned}\sum_{j=1}^{\infty} \mathrm{e}^{-np_j}\,\phi(\lambda p_j) &= \sum_{j=1}^{\infty} \mathrm{e}^{-np_j} \sum_{r=2}^{\infty} \frac{(\lambda p_j)^r}{r!} \\ &= \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \sum_{j=1}^{\infty} \mathrm{e}^{-np_j}\,\frac{(np_j)^r}{r!}\,.\end{aligned}$$

We recognize that for each $r \geq 2$, $\sum_{j=1}^{\infty} \mathrm{e}^{-np_j}\,\frac{(np_j)^r}{r!} = \mathbb{E}K_r(n)$, so that

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \quad \leq \quad \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n)\,.$$

□

Proof of Theorem 2.2. Using Proposition 2.8 and noticing that for each $r \geq 2$, $\mathbb{E}K_r(n) \leq \mathbb{E}K_{\overline{2}}(n)$, we immediately obtain that

$$\begin{aligned}\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] &\leq \mathbb{E}K_{\overline{2}}(n) \sum_{r=2}^{\infty} \left(\frac{\lambda}{n}\right)^r \\ &= \lambda^2 \frac{\mathbb{E}K_{\overline{2}}(n)/n^2}{1-\lambda/n}\,,\end{aligned}$$

which concludes the proof. □

6.4 Regular variation

Proof of Proposition 3.2. By monotonicity of $K_{n,\overline{r}}$, we have the following strong law for any sampling distribution

$$K_{n,\overline{r}} = \sum_{s=r}^{\infty} K_{n,s} \underset{+\infty}{\sim} \sum_{s=r}^{\infty} \mathbb{E}K_s(n) \quad \text{a.s.}\,,$$

[see 80, the discussion after Proposition 2]. Recall that $X_j(n) \sim \mathcal{P}(np_j)$ and that, if $Y \sim \mathcal{P}(\lambda)$, then $\mathbb{P}\left[Y \leq k\right] = \frac{\Gamma(k+1,\lambda)}{k!}$, where $\Gamma(z,x) = \int_x^{+\infty} \mathrm{e}^{-t} t^{z-1}\,\mathrm{d}t$ is the incomplete Gamma function. Hence

$$\begin{aligned}\mathbb{E}K_{\overline{r}}(n) &= \sum_{j=1}^{\infty} \mathbb{P}\left[X_j(n) \geq r\right] \\ &= \sum_{j=1}^{\infty} \frac{1}{(r-1)!} \int_0^{np_j} \mathrm{e}^{-t} t^{r-1}\,\mathrm{d}t \\ &= \frac{1}{(r-1)!} \int_0^1 \int_0^{nx} \mathrm{e}^{-t} t^{r-1}\,\mathrm{d}t\,\nu(\,\mathrm{d}x)\,.\end{aligned}$$

By Fubini's Theorem, we have

$$\mathbb{E}K_{\overline{r}}(n) \quad = \quad \frac{1}{(r-1)!}\int_0^{+\infty} \mathrm{e}^{-z} z^{r-1} \vec{\nu}(z/n)\,\mathrm{d}z\,.$$

Finally, by the Tauberian Theorem, we obtain,

$$\mathbb{E}K_{\overline{r}}(n) \quad \underset{+\infty}{\sim} \quad \frac{\vec{\nu}(1/n)}{(r-1)!}\Gamma(r-\alpha)\,.$$

In particular, when $0<\alpha<1$,

$$K_{n,\overline{r}} \underset{+\infty}{\sim} \frac{rK_{n,r}}{\alpha} \quad \text{a.s.}\,. \qquad \square$$

Proof of Theorem 3.5. Let us recall Proposition 2.8:

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0}-\mathbb{E}M_{n,0})}\right] \leq \sum_{r=2}^{\infty}\left(\frac{\lambda}{n}\right)^r \mathbb{E}K_r(n).$$

Now, bounding each $\mathbb{E}K_r(n)$ by $\mathbb{E}K_{\overline{2}}(n)$ is not sufficient to get the right order for the variance: $\mathbb{E}K_{\overline{2}}(n)$ is of order $\ell(n)$ whereas $\mathrm{Var}M_{n,0}$ is of order $\ell_0(n)/n^2$.

We explore more carefully the structure of $\mathbb{E}K_r(n)$ and show that these quantities are uniformly (in r) bounded by a function of order $\ell_0(n)$ for large enough n, that is, that there exists $n^*\in\mathbb{N}$ and $C\in\mathbb{R}_+$ such that for all $n\geq n^*$, for all $r\geq 1$, $\mathbb{E}K_r(n)\leq C\ell_0(n)$.

Before going into the proof, we observe that for $r\geq n/\ell_0(n)$, the result is true. Indeed, from the identity $\sum_{r=1}^{\infty} r\mathbb{E}K_r(n)=n$, we deduce that $r\mathbb{E}K_{\overline{r}}(n)\leq n$, so that for $r\geq n/\ell_0(n)$, $\mathbb{E}K_{\overline{r}}(n)\leq \ell_0(n)$. Thus we assume that $r\leq n/\ell_0(n)$.

First, we easily deal with the contribution to $\mathbb{E}K_r(n)$ of the symbols with probability less than $1/n$. Indeed

$$I_1^r \quad := \quad \int_0^{1/n} \mathrm{e}^{-nx}\frac{(nx)^r}{r!}\nu(\mathrm{d}x) \leq \int_0^{1/n} \mathrm{e}^{-nx}\frac{(nx)^2}{2!}\nu(\mathrm{d}x) \leq \mathbb{E}K_2(n)\,.$$

As $\mathbb{E}K_2(n)\sim \ell_0(n)/2$, for all δ_0, there exists n_0 such that for all $n\geq n_0$, for all $r\geq 1$, $I_1^r\leq \frac{1+\delta_0}{2}\ell_0(n)$.

For the the contribution of the symbols with probability larger than $1/n$, integration by part and change of variable yield:

$$\begin{aligned} I_2^r \quad &:= \quad \int_{1/n}^{1} \mathrm{e}^{-nx}\frac{(nx)^r}{r!}\nu(\mathrm{d}x) \\ &= \quad \left[\mathrm{e}^{-nx}\frac{(nx)^r}{r!}(-\vec{\nu}(x))\right]_{1/n}^{1} + \int_{1/n}^{1} \mathrm{e}^{-nx}\frac{n^r}{r!}(rx^{r-1}-nx^r)\vec{\nu}(x)\mathrm{d}x \\ &= \quad \frac{\vec{\nu}(1/n)e^{-1}}{r!} + \int_1^{\infty} \mathrm{e}^{-z}\left(\frac{z^{r-1}}{(r-1)!}-\frac{z^r}{r!}\right)\vec{\nu}(z/n)\mathrm{d}z\,. \end{aligned}$$

As $\int_1^{\infty} \mathrm{e}^{-z}\left(\frac{z^{r-1}}{(r-1)!}-\frac{z^r}{r!}\right)\mathrm{d}z = -\mathbb{P}[\mathcal{P}(1)=r] = -e^{-1}/r!$, we can rearrange the previous expression:

$$I_2^r = \ell_0(n)\int_1^{\infty} \mathrm{e}^{-z}\left(\frac{z^r}{r!}-\frac{z^{r-1}}{(r-1)!}\right)\frac{\vec{\nu}(1/n)-\vec{\nu}(z/n)}{\ell_0(n)}\mathrm{d}z\,.$$

Notice that when $z\in[1,r]$, the integrand is negative, so we simply ignore this part of the integral

and restrict ourselves to

$$I_3^r := \int_r^\infty \mathrm{e}^{-z} \left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!} \right) \frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{\ell_0(n)} \mathrm{d}z \,,$$

which we try to bound by a constant term for n greater than some integer that does not depend on r.

The main ingredient of our proof is the next version of the Potter-Drees Inequality (see Theorem 6.4 in Section 6.1.3 and [55, point 4 of Corollary B.2.15]): for $\ell \in \Pi_{\ell_0}$, for arbitrary δ_1, δ_2, there exists t_0 such that for all $t \geq t_0$, and for all $x \leq 1$ with $tx \geq t_0$,

$$(1-\delta_2)\frac{1-x^{-\delta_1}}{\delta_1} - \delta_2 < \frac{\ell(t) - \ell(tx)}{\ell_0(t)} < (1+\delta_2)\frac{x^{-\delta_1}-1}{\delta_1} + \delta_2 \,.$$

Thus, for arbitrary δ_1, δ_2, there exists n_1 such that, for all $n \geq n_1$, for all $z \in [1, n/n_1]$,

$$\frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{\ell_0(n)} \leq (1+\delta_2)\frac{z^{\delta_1}-1}{\delta_1} + \delta_2 \,.$$

As $r \leq n/\ell_0(n)$, taking, if necessary, n large enough so that $\ell_0(n) \geq n_1$, we have $r \leq n/n_1$ and

$$\begin{aligned} I_3^r &\leq \int_r^{n/n_1} \mathrm{e}^{-z} \left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!} \right) \left((1+\delta_2)\frac{z^{\delta_1}-1}{\delta_1} + \delta_2 \right) \mathrm{d}z \\ &\qquad + \int_{n/n_1}^\infty \mathrm{e}^{-z} \left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!} \right) \frac{\vec{\nu}(1/n) - \vec{\nu}(z/n)}{\ell_0(n)} \mathrm{d}z \\ &=: I_4^r + I_5^r \,, \end{aligned}$$

with

$$\begin{aligned} I_4^r &\leq \delta_2 + \frac{1+\delta_2}{\delta_1} \int_r^\infty \mathrm{e}^{-z} \left(\frac{z^{r+\delta_1}}{r!} - \frac{z^r}{r!} + \frac{z^{r-1}}{(r-1)!} - \frac{z^{r-1+\delta_1}}{(r-1)!} \right) \mathrm{d}z \\ &\leq \delta_2 + \frac{1+\delta_2}{\delta_1} \int_r^\infty \mathrm{e}^{-z} \left(\frac{z^{r+\delta_1}}{r!} - \frac{z^{r-1+\delta_1}}{(r-1)!} \right) \mathrm{d}z \\ &= \delta_2 + \frac{1+\delta_2}{\delta_1} \left(\frac{\Gamma(r+1+\delta_1, r)}{\Gamma(r+1)} - \frac{\Gamma(r+\delta_1, r)}{\Gamma(r)} \right) , \end{aligned}$$

where $\Gamma(a,x) = \int_x^\infty \mathrm{e}^{-t} t^{a-1} \mathrm{d}t$ is the incomplete Gamma function. Using the fact that

$$\Gamma(a,x) = (a-1)\Gamma(a-1,x) + x^{a-1}\,\mathrm{e}^{-x} \,,$$

we have

$$\begin{aligned} I_4^r &\leq \delta_2 + \frac{1+\delta_2}{\delta_1 \Gamma(r+1)} \big(\Gamma(r+1+\delta_1, r) - (r+\delta_1)\Gamma(r+\delta_1, r) + \delta_1 \Gamma(r+\delta_1, r) \big) \\ &= \delta_2 + \frac{1+\delta_2}{\delta_1} \left(\frac{r^{r+\delta_1}\,\mathrm{e}^{-r}}{r!} + \delta_1 \frac{\Gamma(r+\delta_1, r)}{\Gamma(r+1)} \right) . \end{aligned}$$

By Stirling's inequality, for all r,

$$\frac{r^{r+\delta_1}\,\mathrm{e}^{-r}}{r!} \leq r^{\delta_1} (2\pi r)^{-1/2} \quad .$$

Thus, taking $\delta_1 = 1/4$, the right-hand term is uniformly bounded by 1. And $\frac{\Gamma(r+\delta_1,r)}{\Gamma(r+1)}$ is also bounded by

1. Thus

$$I_4^r \leq \delta_2 + \frac{1+\delta_2}{\delta_1}(1+\delta_1)\,,$$

and

$$\begin{aligned} I_5^r &\leq \frac{\vec{\nu}(1/n)}{\ell_0(n)} \int_{n/n_1}^{\infty} \mathrm{e}^{-z} \left(\frac{z^r}{r!} - \frac{z^{r-1}}{(r-1)!} \right) \mathrm{d}z \\ &= \frac{\vec{\nu}(1/n)}{\ell_0(n)} \mathrm{e}^{-n/n_1} \frac{(n/n_1)^r}{r!} \\ &\leq \frac{\vec{\nu}(1/n)}{\ell_0(n)} \mathrm{e}^{-n/n_1} \frac{(n/n_1)^{\lfloor n/n_1 \rfloor}}{\lfloor n/n_1 \rfloor!}\,. \end{aligned}$$

By Stirling's inequality, this bound is smaller than $\frac{\vec{\nu}(1/n)}{\ell_0(n)}(2\pi(n/n_1))^{-1/2}$, which tends to 0 as $n \to \infty$. Thus there exists n_2 such that for all $n \geq n_2$, and all $r \leq n/n_1$, $I_5^r \leq \delta_2$.

In the end, we get that for all $\delta_0 \geq 0$, δ_1 with $0 \leq \delta_1 \leq 1/4$, and $\delta_2 \geq 0$, there exists $n^* = \max(n_0, n_1, n_2)$ such that for all $n \geq n^*$, for all $r \geq 1$,

$$\mathbb{E}K_r(n) \leq \ell_0(n) \left(\frac{1+\delta_0}{2} + \delta_2 + \frac{1+\delta_2}{\delta_1}(1+\delta_1) + \delta_2 \right)\,.$$

Taking for instance $\delta_1 = 1/4$ and $\delta_0 = \delta_2 = 1/15$, we have that for large enough n and for all $r \geq 1$,

$$\mathbb{E}K_r(n) \leq 6\ell_0(n)\,,$$

and

$$\log \mathbb{E}\left[\mathrm{e}^{\lambda(M_{n,0} - \mathbb{E}M_{n,0})} \right] \leq \frac{12\ell_0(n)}{n^2} \cdot \frac{\lambda^2}{2(1-\lambda/n)}\,.$$

□

Proof of Proposition 3.4. Under the condition of the Proposition 3.4, from [86], with probability tending to 1, the sample is gap-free, hence the missing mass is $\overline{F}(\max(X_1, \ldots, X_n))$.

The condition of the Proposition implies the condition described in [10], i.e. $\lim_{n\to+\infty} \frac{\overline{F}(n+1)}{\overline{F}(n)} = 0$, to ensure the existence of a sequence of integers $(u_n)_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left\{ \max(X_1, \ldots, X_n) \in \{u_n, u_n+1\} \right\} = 1\,.$$

□

6.5 Applications

Proof of Corollary 4.1. Let us assume that $\mathbb{E}K_{n,1} \to \infty$. Using the fact that $0 \leq \mathbb{E}G_{n,0} - \mathbb{E}M_{n,0} \leq 1/n$, we notice that as soon as $\mathbb{E}K_{n,1} \to \infty$, $\mathbb{E}G_{n,0} \underset{+\infty}{\sim} \mathbb{E}M_{n,0}$. We also recall that, for any sampling distribution,

$$|\mathbb{E}K(n) - \mathbb{E}K_n| \to 0\,,$$

and

$$|\mathbb{E}K_r(n) - \mathbb{E}K_{n,r}| \to 0\,,$$

[see Lemma 1 80]. Now by Chebyshev's inequality,

$$\mathbb{P}\left[\left|\frac{M_{n,0}}{\mathbb{E}M_{n,0}}-1\right|>\epsilon\right] \leq \frac{\mathrm{Var}(M_{n,0})}{\epsilon^2(\mathbb{E}M_{n,0})^2} \leq \frac{2\mathbb{E}K_2(n)}{\epsilon^2 n^2(\mathbb{E}M_{n,0})^2} \sim \frac{2(\mathbb{E}K_{n,2}+o(1))}{\epsilon^2(\mathbb{E}K_{n,1})^2}\,.$$

On the other hand,

$$\mathbb{P}\left[\left|\frac{K_{n,1}}{\mathbb{E}K_{n,1}}-1\right|>\epsilon\right] \leq \frac{\mathrm{Var}(K_{n,1})}{\epsilon^2(\mathbb{E}K_{n,1})^2} \leq \frac{\mathbb{E}K_{n,1}+2\mathbb{E}K_{n,2}}{\epsilon^2(\mathbb{E}K_{n,1})^2}\,,$$

showing that if, furthermore, $\mathbb{E}K_{n,2}/\mathbb{E}K_{n,1}$ remains bounded, the ratios $M_{n,0}/\mathbb{E}M_{n,0}$, $G_{n,0}/\mathbb{E}G_{n,0}$ and thus $M_{n,0}/G_{n,0}$ converge to 1 in probability. To get almost sure convergence, we use Theorem 2.2 to get that when $\mathbb{E}K_{n,1}\to\infty$,

$$\begin{aligned}\mathbb{P}\left[\left|\frac{M_{n,0}}{\mathbb{E}M_{n,0}}-1\right|>\epsilon\right] &\leq 2\exp\left(-\frac{\epsilon^2(\mathbb{E}M_{n,0})^2}{2(2\mathbb{E}K_{\overline{2}}(n)/n^2+\mathbb{E}M_{n,0}/n)}\right)\\ &= 2\exp\left(-\frac{\epsilon^2(\mathbb{E}K_{n,1}+o(\mathbb{E}K_{n,1}))^2}{2(2\mathbb{E}K_{n,\overline{2}}+\mathbb{E}K_{n,1}+o(\mathbb{E}K_{n,1}))}\right)\,.\end{aligned}$$

If $\mathbb{E}K_{n,\overline{2}}/\mathbb{E}K_{n,1}$ remains bounded, this becomes smaller than $c_1\exp\left(-c_2\epsilon^2\mathbb{E}K_{n,1}\right)$. Hence, if $\exp(-c\mathbb{E}K_{n,1})$ is summable for all $c>0$, we can apply the Borel-Cantelli lemma and obtain the almost sure convergence of $M_{n,0}/\mathbb{E}M_{n,0}$ to 1. Moreover, by Proposition 2.5,

$$\mathbb{P}\left[\left|\frac{K_{n,1}}{\mathbb{E}K_{n,1}}-1\right|>\epsilon\right] \leq 4\exp\left(-\frac{\epsilon^2(\mathbb{E}K_{n,1})^2}{2(4\max(\mathbb{E}K_{n,1},2\mathbb{E}K_{n,2})+2/3)}\right)\,,$$

which shows that under these assumptions $K_{n,1}/\mathbb{E}K_{n,1}$ also tends to 1 almost surely. □

Proof of Proposition 4.3. The random variable $G_0(n)-M_0(n)$ is a sum of independent, centered and bounded random variables, namely

$$G_0(n)-M_0(n)=\frac{1}{n}\sum_{j=1}^{\infty}\left(\mathbb{1}_{X_j(n)=1}-np_j\mathbb{1}_{X_j(n)=0}\right)\,.$$

Bound (i) follows immediately from the observation that each $\mathbb{1}_{X_j(n)=1}-np_j\mathbb{1}_{X_j(n)=0}$ satisfies a Bennett inequality: for all $\lambda\geq 0$,

$$\begin{aligned}\log\mathbb{E}\,\mathrm{e}^{\lambda(G_0(n)-M_0(n))} &\leq \sum_{j=1}^{\infty}\mathrm{Var}(\mathbb{1}_{X_j(n)=1}-np_j\mathbb{1}_{X_j(n)=0})\phi\left(\frac{\lambda}{n}\right)\\ &= \mathrm{Var}(G_0(n)-M_0(n))n^2\phi\left(\frac{\lambda}{n}\right)\,.\end{aligned}$$

Bound (ii) follows from the observation that each $\mathbb{1}_{X_j(n)=0} - \frac{1}{tp_j}\mathbb{1}_{X_j(n)=1}$ satisfies a Bennett inequality,

$$\begin{aligned}
\log \mathbb{E}\, \mathrm{e}^{\lambda(M_0(n)-G_0(n))} &\leq \sum_{j=1}^{\infty} \mathrm{Var}\left(\mathbb{1}_{X_j(n)=0} - \frac{1}{np_j}\mathbb{1}_{X_j(n)=1}\right)\phi\left(\lambda p_j\right) \\
&= \sum_{j=1}^{\infty}\left(1+\frac{1}{np_j}\right)\mathrm{e}^{-np_j}\phi(\lambda p_j) \\
&= \sum_{r\geq 2}\left(\frac{\lambda}{n}\right)^r \sum_{j=1}^{\infty}\left(1+\frac{1}{np_j}\right)\mathrm{e}^{-np_j}\frac{(np_j)^r}{r!} \\
&= \sum_{r\geq 2}\left(\frac{\lambda}{n}\right)^r \left(\mathbb{E}K_r(n) + \frac{1}{r}\mathbb{E}K_{r-1}(n)\right) \\
&\leq \sum_{r\geq 2}\left(\frac{\lambda}{n}\right)^r \frac{3\mathbb{E}K(n)}{2}\,,
\end{aligned}$$

which concludes the proof. □

Proof of Proposition 4.4. With probability greater than $1-2\delta$, by Proposition 4.3,

$$G_0(n) - M_0(n) \leq \frac{1}{n}\sqrt{2(\mathbb{E}K_1(n) + 2\mathbb{E}K_2(n))\log\frac{1}{\delta}} + \frac{\log\frac{1}{\delta}}{3n}$$

and

$$G_0(n) - M_0(n)) \geq -\frac{1}{n}\sqrt{6\mathbb{E}K(n)\log\frac{1}{\delta}} - \frac{\log\frac{1}{\delta}}{n}\,.$$

We may now invoke concentration inequalities for $K_1(n)+2K_2(n)$ and $K(n)$. Indeed, with probability greater than $1-\delta$, $K(n) \geq \mathbb{E}K(n) - \sqrt{2\mathbb{E}K(n)\log\frac{1}{\delta}}$ which entails $\sqrt{\mathbb{E}K(n)} \leq \sqrt{K(n) + \frac{\log\frac{1}{\delta}}{2}} + \sqrt{\frac{\log\frac{1}{\delta}}{2}}$.

We have $2K_2(n) + K_1(n) \geq 2\mathbb{E}K_2(n) + \mathbb{E}K_1(n) - \sqrt{4(2\mathbb{E}K_2(n) + \mathbb{E}K_1(n))\log\frac{1}{\delta}}$ with probability greater than $1-\delta$, which entails

$$\sqrt{2\mathbb{E}K_2(n) + \mathbb{E}K_1(n)} \leq \sqrt{(2K_2(n) + K_1(n)) + \log\frac{1}{\delta}} + \sqrt{\log\frac{1}{\delta}}\,,$$

which concludes the proof. □

Proof of Proposition 4.5. The covariance matrix $\mathrm{Cov}(n)$ of $(G_0(n), M_0(n))$ can be written in terms of the expected occupancy counts as

$$\mathrm{Cov}(n) = \frac{1}{n^2}\begin{pmatrix}\mathbb{E}K_1(n) & 0 \\ 0 & 2\mathbb{E}K_2(n)\end{pmatrix} - \frac{\mathbb{E}K_2(2n)}{2n^2}\begin{pmatrix}1 & 1 \\ 1 & 1\end{pmatrix}.$$

From [102], we have

$$\mathrm{Cov}(n)^{-1/2}\begin{pmatrix}G_0(n) - \mathbb{E}G_0(n) \\ M_0(n) - \mathbb{E}M_0(n)\end{pmatrix} \rightsquigarrow \mathcal{N}(0, I_2)\,,$$

where I_2 is the identity matrix, which can be rewritten as

$$\Sigma(n)^{-1/2}\begin{pmatrix}\frac{G_0(n)}{\mathbb{E}G_0(n)} - 1 \\ \frac{M_O(n)}{\mathbb{E}M_0(n)} - 1\end{pmatrix} \rightsquigarrow \mathcal{N}(0, I_2)\,,$$

with $\Sigma(n) = (\mathbb{E}G_0(n))^{-2}\mathrm{Cov}(n)$.

The delta method applied to the function $(x_1, x_2) \mapsto x_1/x_2$ yields

$$\left((1 \ \ -1) \ \Sigma(n) \ \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)^{-1/2} \left(\frac{G_0(n)}{M_0(n)} - 1 \right) \rightsquigarrow \mathcal{N}(0,1)\,,$$

and

$$\left((1 \ \ -1) \ \Sigma(n) \ \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)^{-1/2} = \frac{\mathbb{E}K_1(n)}{\sqrt{\mathbb{E}K_1(n) + 2\mathbb{E}K_2(n)}}\,,$$

which concludes the proof. □

Remark 6.1. *The same central limit theorem holds in the binomial setting. The proof is essentially the same, the only difference being that $\mathbb{E}G_{n,0}$ and $\mathbb{E}M_{n,0}$ are no longer equal. However, the bias becomes negligible with respect to the fluctuations, that is, for v_n either $n^\alpha \ell(n)$ or $n\ell_1(n)$*

$$\sqrt{v_n} \left(\frac{\mathbb{E}G_{n,0}}{\mathbb{E}M_{n,0}} - 1 \right) \underset{n\to\infty}{\to} 0\,. \qquad \square$$

Chapter 2

Adaptive coding on countable alphabets

This chapter presents an ongoing work with Stéphane Boucheron and Elisabeth Gassiat.

1 Coding on infinite alphabets

The problem we address here is that of encoding a sequence $X_{1:n} = (X_1, \ldots, X_n)$, where each symbol takes values in a countable alphabet $\mathcal{X}$, and is generated by a source $P = (p_j)_{j \geq 1}$. The sequence $X_{1:n}$ is assumed to be I.I.D.. One says that the source is stationary and memoryless.

A *lossless binary source code* (or *code* for short) is a one-to-one map from finite sequences of symbols in the alphabet $\mathcal{X}$ (here $\mathcal{X} = \mathbb{N}_+$) to finite sequences of binary $\{0, 1\}$ symbols. Given a source P, the task of source coding is to minimize the *expected codelength*:

$$\mathbb{E}[\ell(X_{1:n})] = \sum_{x_{1:n} \in \mathcal{X}^n} P^n(x_{1:n}) \ell(x_{1:n}).$$

By the source coding theorem, the Shannon entropy of the source is a lower bound to the expected codelength of any lossless binary code (in this chapter, log denotes the base-2 logarithm). Therefore, one way to measure the performance of any particular code is by its *expected redundancy*, defined as the excess expected length $\mathbb{E}[\ell(X_{1:n})] - H(P^n)$. This is meaningful when $H(P) < \infty$, which we will assume to be the case.

A code is *uniquely decodable* if any concatenation of codewords can be parsed into codewords in a unique way. The Kraft-McMillan inequality asserts that for a uniquely decodable code over $\mathcal{X}^* = \bigcup_{n \geq 0} \mathcal{X}^n$, the codelength map $x_{1:n} \mapsto \ell(x_{1:n})$ satisfies

$$\sum_{x_{1:n} \in \mathcal{X}^n} 2^{-\ell(x_{1:n})} \leq 1,$$

and that conversely, given codelengths that satisfy such an inequality, there exists a corresponding uniquely decodable code (even a *prefix* code, *i.e.* a code such that no codeword is the prefix of an other codeword). The Kraft-McMillan inequality thus establishes a deep correspondence between codes over $\mathcal{X}^n$ and probability distributions over $\mathcal{X}^n$, and we may refer to an arbitrary probability distribution Q_n on $\mathcal{X}^n$ as a *coding distribution* [51].

When the source is P, the expected code length of a coding distribution Q_n is thus given by

$\mathbb{E}_P\left[-\log Q_n(X_{1:n})\right]$, and the expected redundancy is the Kullback-Leibler divergence (or relative entropy) between P^n and Q_n:

$$D(P^n, Q_n) = \sum_{x_{1:n}\in\mathcal{X}^n} P^n(x_{1:n}) \log \frac{P^n(x_{1:n})}{Q_n(x_{1:n})} = \mathbb{E}_P\left[\log \frac{P^n(X_{1:n})}{Q_n(X_{1:n})}\right].$$

The theoretically optimal coding probabilities are given by the source P^n itself. By using methods such as the arithmetic coding, codes corresponding to P^n can be designed to have a redundancy that remains bounded by 1 for all n. Arithmetic coding allows to encode a message sequentially (*online*) such that, for all $n \geq 1$, the length of the codeword of $x_{1:n}$ is less than $\lceil -\log P^n(x_{1:n})\rceil$. But this requires that P is known.

1.1 Universal source coding

In universal coding, one attempts to construct a coding distribution Q_n that achieves low redundancy across an entire source class $\mathcal{C}^n$, without knowing in advance which $P \in \mathcal{C}$ is actually generating the sequence. Such a construction is called coding with respect to $\mathcal{C}^n$.

To assess a code with respect to a source class, we study the *maximal expected redundancy* defined as

$$\overline{R}(Q_n, \mathcal{C}^n) = \sup_{P\in\mathcal{C}} D(P^n, Q_n)\,.$$

which is essentially as high as the redundancy could grow if P is chosen adversarially at every n.

The infimum of $\overline{R}(Q_n, \mathcal{C}^n)$ over all Q_n in $\mathcal{M}_1(\mathcal{X}^n)$ (the set of probability distributions on $\mathcal{X}^n$), is called the *minimax redundancy* of $\mathcal{C}^n$:

$$\overline{R}(\mathcal{C}^n) = \inf_{Q_n\in\mathcal{M}_1(\mathcal{X}^n)} \overline{R}(Q_n, \mathcal{C}^n).$$

The minimax redundancy is a property of the source class $\mathcal{C}$ and represents the best a code could hope for in terms of a guaranteed expected redundancy over the class $\mathcal{C}$.

A code $(Q_n)_{n\geq 1}$ is said to be *weakly universal* over $\mathcal{C}$ if, for $P \in \mathcal{C}$, $\frac{1}{n}D(P^n, Q_n) \to 0$ as $n \to \infty$. One may require further *strong universality* with respect to $\mathcal{C}$, by asking that the uniform convergence $\frac{1}{n}\overline{R}(Q_n, \mathcal{C}^n) \to 0$. A source class $\mathcal{C}$ is said to have a non-trivial minimax redundancy rate when $\overline{R}(\mathcal{C}^n) = o(n)$, that is, when there exists a strongly universal code over $\mathcal{C}$.

A more stringent redundancy measure is the *minimax regret* (or *worst-case redundancy*) defined as

$$\widehat{R}(\mathcal{C}^n) = \inf_{Q_n\in\mathcal{M}_1(\mathcal{X}^n)} \sup_{P\in\mathcal{C}} \sup_{x_{1:n}\in\mathcal{X}^n} \log \frac{P^n(x_{1:n})}{Q_n(x_{1:n})}.$$

The class $\mathcal{C}_k^n$ of stationary memoryless sources of length n over a finite alphabet A_k of size k has been deeply investigated. Xie and Barron [156] showed that, for k fixed, as $n \to \infty$,

$$\widehat{R}(\mathcal{C}_k^n) \quad = \quad \frac{k-1}{2}\log\frac{n}{2\pi} + \log\frac{\Gamma\left(\frac{1}{2}\right)^k}{\Gamma\left(\frac{k}{2}\right)} + o_k(1)\,.$$

Moreover, the minimax redundancy satisfies $\widehat{R}(\mathcal{C}_k^n) \geq \overline{R}(\mathcal{C}_k^n) \geq \widehat{R}(\mathcal{C}_k^n) - \log(\mathrm{e})$. The Krichevsky-Trofimov coder [107] can be shown to be asymptotically maximin (even though not asymptotically minimax). As we will use this particular coding procedure in the construction of our code, we recall its general principle here: the justification of the KT coding distribution comes from Bayesian statistics. The minimax redundancy

$\overline{R}(\mathcal{C}_k^n)$ indeed has the following Bayesian representation.

$$\begin{aligned}\overline{R}(\mathcal{C}_k^n) &= \sup_{\pi\in\mathcal{M}_1(\mathcal{C}^k)} \inf_{Q_n\in\mathcal{M}_1(A_k^n)} \int D(P^n, Q_n)\,\mathrm{d}\pi(P) \\ &= \sup_{\pi\in\mathcal{M}_1(\mathcal{C}^k)} \int D(P^n, P_\pi^n)\,\mathrm{d}\pi(P)\,,\end{aligned}$$

where, for $\pi \in \mathcal{M}_1(\mathcal{C}^k)$, P_π is the mixture distribution given by $P_\pi(j) = \int P(j)\,\mathrm{d}\pi(P)$. With this representation, the problem comes down to finding an *a priori* distribution on $\mathcal{C}^k$ which would result, by mixture, in a "good" coding distribution. The appropriate prior is called the Jeffrey's prior and, in the case of stationary memoryless sources over a finite alphabet of size k, is given by the Dirichlet law with parameter $(1/2, \ldots, 1/2)$. That is to say, we will choose as prior on $\mathcal{M}_1(\mathcal{C}^k) = \left\{\theta = (\theta_1, \ldots, \theta_k),\ \sum_{j=1}^k \theta_j = 1\right\}$, the distribution π given by

$$\pi(\theta) = \frac{\Gamma\left(\frac{k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k} \prod_{j=1}^k \theta_j^{-\frac{1}{2}}\,.$$

The Krichevsky-Trofimov coding distribution then corresponds to

$$\mathrm{KT}(X_{1:n}) = \int_\theta P_\theta(X_{1:n})\,\mathrm{d}\pi(\theta)\,.$$

With our usual notation $X_{n,j} = \sum_{i=1}^n \mathbb{1}_{\{X_i=j\}}$, we have

$$\mathrm{KT}(X_{1:n}) = \frac{\Gamma\left(\frac{k}{2}\right)\prod_{j=1}^k \Gamma\left(X_{n,j}+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k \Gamma\left(n+\frac{k}{2}\right)}\,,$$

and

$$\mathrm{KT}(j\big|X_{1:n}) = \frac{X_{n,j}+\frac{1}{2}}{n+\frac{k}{2}}\,.$$

When the alphabet size is fixed, the per-symbol redundancy thus decreases to 0 very fast (as $\log n/n$), and we have at our disposal coding procedures such as KT which achieve the minimax redundancy. In numerous applications however, it might seem unrealistic to assume that the alphabet size is fixed, or even much smaller than the message length. What can we say when the alphabet $\mathcal{X}$ is infinite countable? In this domain, one first faces an impossibility result: there is no weakly universal code over the class of stationary memoryless sources over an infinite alphabet, which is a consequence of Kieffer's Theorem.

Theorem 1.1 ([104])**.** *Let Λ be a class of sources over $\mathcal{X}^{\mathbb{N}}$. For $\mathbb{P} \in \Lambda$, we let $\mathbb{P}_n$ denote the marginal distribution of $\mathbb{P}$ over the first n coordinates. The two following statements are equivalent*

- *There exists a coding distribution $(Q_n)_{n\geq 1}$ on $\mathcal{X}^n$ such that*

$$\forall \mathbb{P} \in \Lambda,\ \frac{D(\mathbb{P}_n, Q_n)}{n} \underset{n\to\infty}{\to} 0\,.$$

- *There exists a probability $Q^\star$ on $\mathcal{X}$ such that*

$$\forall \mathbb{P} \in \mathcal{C},\ \mathbb{E}_{\mathbb{P}_1}\left[-\log Q^\star(X)\right] < \infty\,.$$

The class of stationary memoryless sources on $\mathbb{N}_+$ clearly do not satisfy Kieffer's condition: for all $Q^\star$ on $\mathbb{N}_+$ with infinite support, one can choose an increasing sequence $(x_n)_{n\geq 1}$ such that $-\log Q^\star(x_n) \geq 2^n$,

and then choose the probability P with support $(x_n)_{n\geq 1}$ defined by $P(x_n) = 2^{-n}$. We then have $\mathbb{E}_P\left[-\log Q^\star(X)\right] = \infty$.

This prompted several approaches to cope with infinite alphabets [129]. A first approach is to restrict to source classes satisfying Kieffer's condition [68, 87]. A second direction consists in modifying the preformance criterion, and considering the encoding of the sequence's *pattern*. For instance, the pattern of the sequence $(8, 19, 4, 4, 19)$ is $(1, 2, 3, 3, 2)$. If $\mathcal{C}_\phi^n$ denotes the class of all pattern distributions induced by *i.i.d.* sources of length n over $\mathcal{X}^n$, then

$$0.3n^{1/3} \leq \overline{R}(\mathcal{C}_\phi^n) \leq \widehat{R}(\mathcal{C}_\phi^n) \leq n^{1/3}(\log n)^4 \,.$$

(see Acharya et al. [1], Garivier [78], Shamir [147]). Regardless of the alphabet size, patterns can be compressed with a per-symbol redundancy decreasing to zero.

A third direction, coming from statistical learning theory, is to look for *adaptivity*.

1.2 Adaptive source coding

Given a possibly very large collection of sources, a universal code attempts to minimize redundancy, that is the difference between the expected codeword length and the expected codeword length that would be achieved by a code tailored to the source. Adaptive coding considers a more general setting. Assume we are facing a collection of source classes, such that for each class, a good universal coder is available (and each class has a non trivial minimax redundancy rate), is it possible to engineer a single coding method that performs well over all classes in the collection? The notion of adaptivity comes from the mathematical statistics language : an estimator is said to be adaptive over a collection of models if it achieves the minimax over all models [54]. The search for adaptive codes make sense if the the union of source classes does not have a non-trivial redundancy rate, which is the case of stationary memoryless sources over an infinite alphabet.

Let $(\mathcal{C}(\alpha))$ be a collection of source classes indexed by $\alpha \in A$. A sequence $(Q_n)_{n\geq 1}$ of coding probabilities is said to be *asymptotically adaptive* with respect to a collection $(\mathcal{C}(\alpha))_{\alpha\in A}$ of source classes if for all $\alpha \in A$

$$\overline{R}(Q_n, \mathcal{C}(\alpha)^n) = \sup_{P\in\mathcal{C}(\alpha)} D(P^n, Q_n) \leq (1 + o_\alpha(1))\overline{R}(\mathcal{C}(\alpha)) \tag{1.1}$$

as n tends to infinity. If the inequality (1.1) holds with a factor other than $(1 + o_\alpha(1))$ (that may depend on α) larger than 1 to the right, then we say that there is adaptivity *within* this factor. Note that Q_n cannot depend on α or else the problem is simply one of universality.

1.3 Regularly varying envelope classes

In our situation, the alphabet $\mathcal{X}$ is the set of positive integers $\mathbb{N}_+$, and the source P is given by $(p_j)_{j\geq 1}$. As outlined earlier, in this countable alphabet setting, the class of all stationary memoryless distributions has trivial minimax redundancy: there is no universal code (even weakly universal) for the class of I.I.D. distributions over $\mathbb{N}_+$ [87, 88, 104]. This challenge has prompted several approaches: imposing constraints on the sources classes as in [39] and subsequent papers [3, 34, 35, 41], or redefining the performance criteria by focusing on pattern coding as popularized in [72, 78, 129, 133]. Here we pursue the line of research initiated in [39], we deal with collection of so-called envelope classes, but the adaptive code we introduce and investigate will turn out to be a pattern encoder in the spirit of [129, 133].

We start by recalling the definition of an envelope source class, as introduced by [39].

Definition 4 (ENVELOPE CLASSES)**.** Let $(f_j)_{j\geq 1}$ be a probability distribution on $\mathbb{N}_+$ such that $f_1 \geq f_2 \geq \ldots$ and let $\ell_f \geq 0$. The *envelope class* Λ_f defined by the distribution $(f_j)_{j\geq 1}$ and the integer ℓ_f is

the collection of distributions which are dominated by f after ℓ_f:

$$\Lambda_f \quad = \quad \Big\{ P : \forall j \geq 1,\ p_{j+\ell_f} \leq f_j \ \Big\} .$$

Envelope classes provide a framework where the search for adaptive coding strategies is feasible. Their relevance to practical applications may be questioned : Falahatgar et al. [72] point out that in natural language processing where the (large but finite) alphabet may be considered as the set of words of some natural language, there is no special reason to privilege any ordering on the alphabet. This is implicitly done when defining classes with a decreasing envelope. But the tight redundancy bounds established in [72] reveal that when dealing with infinite alphabets, some ordering has to be assumed if sublinear redundancy is desired.

We will interested in the full range of *regularly varying* envelope classes.

Let us denote by ν_f, $\vec{\nu}_f$, $\nu_{1,f}$ the quantities defined respectively in (3.1), (3.2), (3.4) when the underlying distribution is given by the envelope frequencies $(f_j)_{j\geq 1}$, that is,

$$\nu_f(\,\mathrm{d}x) = \sum_{j\geq 1} \delta_{f_j}(\,\mathrm{d}x)\,, \quad \vec{\nu}_f(x) = \nu[x,1]\,, \quad \nu_{1,f}[0,x] = \int_0^x y\nu_f(\,\mathrm{d}y)\,.$$

We define regularly varying envelope classes as follows.

Definition 5 (REGULAR VARIATION)**.** The envelope class Λ_f is said to be regularly varying with index $\alpha \in [0,1]$ if the function $\vec{\nu}_f(1/\cdot)$ belongs to RV_α, *i.e.* if $\vec{\nu}_f(1/n) \sim n^\alpha \ell(n)$, with ℓ slowly varying.

By abuse of notation, we will sometimes write $f \in \mathrm{RV}_\alpha$ instead of $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$. Also the simple symbol $\mathbb{E}$ denotes the expectation with respect to the source P, while $\mathbb{E}_f$ will denote the expectation with respect to the envelope distribution $(f_j)_{j\geq 1}$.

As observed in the previous chapter, under regular variation, the expected occupancy counts $\mathbb{E}_f K_n$, $\mathbb{E}_f K_{n,r}$ are nicely related to the function $\vec{\nu}_f(1/n)$ (see Theorems 3.2, 3.3 and 3.4 in Chapter 1 for the asymptotics in the three regimes $0 < \alpha < 1$, $\alpha = 1$ and $\alpha = 0$ respectively).

The introduction of envelope classes combined with the search for adaptive codes allowed to gain a lot of insight about redundancies of source classes over infinite alphabets. For instance, Boucheron et al. [39] showed that, of Λ_f is an envelope class, and if $\overline{F}(u) = \sum_{j>u} f_j$, then

$$\overline{R}(\Lambda_f) \quad \leq \quad \inf_{u\leq n} \left\{ n\overline{F}(u) \log \mathrm{e} + \frac{u-1}{2} \log n \right\} + 2\,. \tag{1.2}$$

This inequality conveys an insightful message: if the envelope $(f_j)_{j\geq 1}$ is known, the following coding strategy seems natural. Choose a threshold u such that $\overline{F}(u) \approx \frac{1}{n}$. Encode symbol j with an appropriate code for I.I.D. sources over the finite-size alphabet $\{0, 1, \ldots, u\}$ (*e.g.* Krichevsky-Trofimov). Symbol 0 corresponds to an *escape* symbol, meaning that it indicates that symbol j is greater than u and that one has to encode it with another (more costly) procedure, the Elias encoding for integers. The redundancy of such a censuring procedure should attain the upper bound in (1.2). If the envelope is not known, one has to choose the threshold u based on the sample only. The principle of the AC-code (auto-censuring) designed by Bontemps [34] is to choose $u_n = \max(X_1, \ldots, X_n)$, and Bontemps et al. [35] showed that this code is adaptive over the collection of envelope source classes with finite and non-decreasing hazard rate. When the envelope has a heavy tail however, the maximum of the sample can be extremely large, and the AC-code fails to mimic the tail behaviour of the source. For such envelopes, Boucheron et al. [41] proposed another procedure to select the threshold, and designed a general code called the ETAC code (expanding threshold auto-censuring), which is shown to be adaptive, within a $\log n$ factor, over the collection of envelope source classes characterized by a regularly varying envelope with index $\alpha \in]0,1[$.

The code we propose here aims at taking into account both the case $\alpha = 0$ (the so-called Gumbel domain of light-tailed distributions) and $\alpha \in]0, 1[$ (the Fréchet domain of heavy tailed distributions). It achieves adaptivity over those source classes, to the price of a $\log \log n$ factor.

1.4 The Pattern Censoring Code

The code we construct performs an online encoding or decoding of a sequence of symbols. It pertains to the family of censoring codes described in [34, 35, 41]: first occurrences of symbols are censored, that is they are encoded using a general purpose encoder for integers (namely, the Elias code [68]) and implicitly inserted into a dictionary; symbols that have already been observed are fed to a Krichevsky-Trofimov encoder (KT) that works on the current dictionary. The Krichevsky-Trofimov encoder actually performs a variant of pattern coding. Thus, the code does two things: it progressively encodes the dictionary each time a new symbol occurs, and it encodes the sequence using the KT code on the finite-size alphabet containing the symbols seen so far.

We now describe it more formally, using the same notations for occupancy counts as in the previous chapter ($X_{n,j}$, K_n...). We start with our input sequence $X_{1:n} = (X_1, \ldots, X_n)$ of symbols from $\mathcal{X} = \mathbb{N}_+$.

Encoding:

— The dictionary is initialized with the symbol 0: $\mathcal{D}_0 = \{0\}$.

— At every index i corresponding to an input symbol, maintain a dictionary $\mathcal{D}_i$ as follows:

$$\mathcal{D}_i \quad = \quad \{0\} \cup \{j \geq 1,\, X_{i,j} > 0\} \ .$$

Note that, at time i, the size of the dictionary $\mathcal{D}_i$ is $K_i + 1$.

— Create a *censored* sequence $\widetilde{X}_{1:n}$ such that every symbol X_i that does not belong to $\mathcal{D}_{i-1}$ is replaced by the special 0 symbol:

$$\widetilde{X}_i \quad = \quad X_i \mathbb{1}_{\{X_i \in \mathcal{D}_{i-1}\}} \ .$$

— The variable K_n (the number of distinct symbols in $X_{1:n}$) then corresponds precisely to the number of *redacted* (censored-out) input symbols. Let $i_{1:K_n}$ be their indices. Extract the subsequence $X_{i_{1:K_n}}$ of all such redacted symbols.

— Perform an instantaneously decodable lossless progressive encoding (in the style of Mixture / arithmetic coding) of the censored sequence $\widetilde{X}_{1:n}$, assuming decoder side-information about past symbols. Call the resulting string C_M.

— Perform an instantaneously decodable lossless encoding (in the style of Elias / integer coding) of each redacted symbol in $X_{i_{1:K_n}}$ individually rather than as a sequence, assuming decoder side-information about past symbols. Call the resulting string C_E.

— Interleave the coded redacted symbols of C_E just after each coded 0 symbol of C_M, to form the overall code.

Decoding:

— Decode the interleaved C_M and C_E strings until exhaustion, as follows.

— Decode C_M to obtain $\widetilde{X}_{1:n}$ progressively.

— When a 0 is encountered, decode the single interleaved redacted symbol from C_E to take the place of the 0 symbol in the decoded sequence, then move back to decoding C_M.

— Note that at all times the decoder knows the entire past sequence, and therefore the decoder side past side-information hypothesis is upheld.

We now give the details of the encoding of the censored sequence $\widetilde{X}_{1:n}$ and of the dictionary $X_{i_{1:K_n}}$.

We also take additional care in guaranteeing that our code is *instantaneously decodable*. To encode $\widetilde{X}_{1:n}$, we start by appending an extra 0 at the end of the original censored sequence, to signal the termination of the input. We therefore in fact encode $\widetilde{X}_{1:n}0$ into C_M. We do this by performing a progressive arithmetic coding [141] using coding probabilities $\widetilde{Q}_{n+1}(\widetilde{X}_{1:n}0)$ given by:

$$\widetilde{Q}_{n+1}(\widetilde{X}_{1:n}0) = \widetilde{Q}_{n+1}(0 \mid \widetilde{X}_{1:n}, \mathcal{D}_n) \prod_{i=0}^{n-1} \widetilde{Q}_{i+1}(\widetilde{X}_{i+1} \mid \widetilde{X}_{1:i}, \mathcal{D}_i)\,,$$

where the predictive probabilities $\widetilde{Q}_{i+1}$ are a variant of Krichevsky-Trofimov mixtures on dictionary $\mathcal{D}_i$: for $j \in \mathcal{D}_i$,

$$\widetilde{Q}_{i+1}\left(\widetilde{X}_{i+1} = j \mid \widetilde{X}_{1:i}, \mathcal{D}_i\right) = \frac{\widetilde{X}_{i,j} + \frac{1}{2}}{i + \frac{K_i+1}{2}}\,,$$

where, for $j \in \mathcal{D}_i$, $\widetilde{X}_{i,j}$ is the number of occurrences of symbol j in $\widetilde{X}_{1:i}$. Note that

$$\widetilde{X}_{i,j} \quad = \quad \begin{cases} K_i & \text{if } j = 0, \\ X_{i,j} - 1 & \text{if } j \in \mathcal{D}_i \setminus \{0\}. \end{cases}$$

What these coding probabilities represent, in effect, is a mixture code consisting of progressively enlarging the alphabet based on the symbols seen so far, and feeding an arithmetic coder with Krichevsky-Trofimov mixtures over this growing alphabet. Thanks to K_i being determined by the data, the enlargement of the alphabet is performed online.

The subsequence $X_{i_{1:N}}$ of redacted symbols is encoded into the string C_E as follows. For each $i \in i_{1:N}$, we encode $X_i + 1$ using Elias penultimate coding [68], where the $+1$ is added to make sure these values are strictly greater than 1. Thus, if $X_i = j$ and $X_{i-1,j} = 0$, the cost of inserting this new symbol in the dictionary is $\log(j+1) + 2\log\log(j+1)$. Corresponding to the extra 0 appended to the censored sequence, we append an encoded 1 to C_E. Since no other encoded redacted symbol but this one results in a 1, it unambiguously signals to the decoder that the 0 symbol decoded from C_M is in fact the termination signal. This ensures that the overall code is instantaneously decodable, and that it therefore corresponds to an implicit coding probability Q_n.

Note that, for $i \geq 0$, conditionally on the first i symbols $X_{1:i}$, the expected instantaneous redundancy of the encoding of symbol X_{i+1} is given by

$$\begin{aligned} \mathbb{E}_P\left[\log \frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})} \Big| X_{1:i}\right] \quad = \quad & \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0} \log\left(\frac{p_j\left(i + \frac{K_i+1}{2}\right)}{X_{i,j} - \frac{1}{2}}\right) \\ & + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0} \log\left(\frac{p_j\left(i + \frac{K_i+1}{2}\right)}{K_i + \frac{1}{2}}\right) \\ & + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0} \left(\log(j+1) + 2\log\log(j+1)\right)\,. \end{aligned}$$

The first two sums correspond to the KT code on the censored sequence and the last sum to the Elias encoding of the alphabet. Let us point out right away that our analysis will have the particularity not to separate the contributions of those two different codes. Instead, we rearrange the total redundancy as

follows:

$$\mathbb{E}_P\left[\log\frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})}\Big|X_{1:i}\right] = \sum_{j\geq 1} p_j \log\left(ip_j\left(1+\frac{K_i+1}{2i}\right)\right) + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0}\log\left(\frac{1}{X_{i,j}-\frac{1}{2}}\right) \quad (1.3)$$
$$+\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\left(\frac{j+1}{K_i+\frac{1}{2}}\right) + 2\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\log(j+1)\,.$$

Note that, in the term

$$\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\left(\frac{j+1}{K_i+\frac{1}{2}}\right)$$

the $\log(j+1)$ contribution of Elias encoding is now counter-balanced by the $K_i+1/2$ coming from the encoding of the 0 symbols in the censored sequence. This will turn out to be crucial: this is what allows us to obtain adaptivity within a $\log\log n$ factor (which comes from the residual part of Elias code) instead of a $\log n$ factor. Note that the encoding of the 0 symbols in the censored sequence is then pivotal, this is one of the main differences with the AC or ETAC codes: in those previous codes, zeros were actually not encoded, and the KT code was working with the counts stemming from the sequence $X_{1:i}$ instead of $\widetilde{X}_{1:i}$. Here, the KT code considers 0 as a proper symbol in the sequence $\widetilde{X}_{1:i}$, and this allows us to gain the term $\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\left(\frac{p_j\left(i+\frac{K_i+1}{2}\right)}{K_i+\frac{1}{2}}\right)$, to the price of decreasing the other counts $X_{i,j}$ by one, which does not really affects the redundancy.

2 Main result

We analyse the performance of the Pattern Censoring Code (Q_n) on regularly varying envelope classes Λ_f, with index $\alpha\in[0,1[$. For an envelope distribution $f=(f_j)_{j\geq 1}$, let $\mathbf{R}_f(n)$ be defined as

$$\mathbf{R}_f(n) = \log(\mathrm{e})\int_1^n \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t\,.$$

The following result is a consequence of Acharya et al. [4], and allows to relate the minimax redundancy of Λ_f^n to $\mathbf{R}_f(n)$.

Theorem 2.1 ([4])**.** *Let Λ_f be an envelope source class, with $f\in \mathrm{RV}_\alpha$ and $\alpha\in[0,1[$. Then*

$$\overline{R}(\Lambda_f^n) \asymp \mathbf{R}_f(n)\,.$$

If $\alpha=0$, this is even an asymptotic equivalence. In the case $\alpha=1$, we have

$$\overline{R}(\Lambda_f^n) \asymp \mathbb{E}_f[K_n] \gg \mathbf{R}_f(n)\,.$$

Theorem 2.2. *Let (Q_n) be the coding distribution associated to the Pattern Censoring Code. For all $\alpha\in[0,1[$, for all envelope distribution f with $\vec{\nu}_f(1/\cdot)\in\mathrm{RV}_\alpha$ and all integer $\ell_f\geq 0$, there exists constants $a_f,b_f>0$ such that*

$$(a_f+o_f(1))\mathbf{R}_f(n)\leq \overline{R}(\Lambda_f^n)\leq \overline{R}(Q_n,\Lambda_f^n)\leq (b_f+o_f(1))\mathbf{R}_f(n)\log\log n\,. \quad (2.1)$$

In particular, the Pattern Censoring Code is adaptive, within a $\log\log n$ factor, with respect to the collection $\left((\Lambda_f)_{\substack{f\in\mathrm{RV}_\alpha\\ \ell_f\geq 0}}\right)_{\alpha\in[0,1[}$.

Remark 2.1. *Let us mention several possible improvements of this Theorem, that we would like to address in the future. Is it possible that a more sophisticated analysis of the Pattern Censoring Code could show that, in the case of non-decreasing hazard rate envelopes, this code is properly adaptive and thus performs as well as the AC Code of Bontemps [34]? Is it possible to take into account the case $\alpha = 1$, corresponding to extremely heavy tails? Finally and more generally, can we do without the* $\log\log n$ *factor or is there an incompressible price to adaptivity?*

Before analysing the redundancy of the code, we first need to understand the minimax redundancy of envelope classes Λ_f^n. This is the purpose of the next section, in which we give upper and lower bounds on $\overline{R}(\Lambda_f^n)$. The techniques mostly rely on Poissonization arguments as introduced in Acharya et al. [4], and Karlin's formalism then allows us to directly capture the order of those bounds and to gain insight on the quantities involved. In particular, we will derive Theorem 2.1. In section 4, we proceed with the analysis of the code, establishing the right-hand side inequality in (2.1). Most of the proofs are grouped in Section 5.

3 Minimax redundancy

In this section, we give upper and lower bounds for the minimax redundancy of envelope source classes. The techniques are very much inspired by [4]. We show that the Poissonization arguments used in this paper to bound the minimax regret can be extended to minimax redundancy. When particularizing to regularly varying envelopes, we show that the resulting bounds are tight up to constant factors. When restricting further to slowly varying envelopes, they are properly tight.

3.1 Properties of minimax redundancy

Recall that the minimax redundancy of a class $\mathcal{C}^n$ of stationary memoryless sources is defined as

$$\overline{R}(\mathcal{C}^n) \quad = \quad \inf_{Q_n \in \mathcal{M}_1(\mathcal{X}^n)} \sup_{P \in \mathcal{C}} \mathbb{E}_P \left[\log \frac{P^n(X_{1:n})}{Q_n(X_{1:n})} \right] .$$

The sequence $(\overline{R}(\mathcal{C}^n))_{n\geq 1}$ has the following properties (we refer to Gassiat [79]).

Proposition 3.1.

- $(\overline{R}(\mathcal{C}^n))_{n\geq 1}$ *is increasing.*
- $(\overline{R}(\mathcal{C}^n))_{n\geq 1}$ *is sub-additive: for all* $n, m \geq 1$,

$$\overline{R}(\mathcal{C}^{n+m}) \quad \leq \quad \overline{R}(\mathcal{C}^n) + \overline{R}(\mathcal{C}^m) .$$

- *The minimax redundancy is equal to the Bayesian redundancy:*

$$\begin{aligned} \overline{R}(\mathcal{C}^n) \quad &= \quad \sup_{\pi \in \mathcal{M}_1(\mathcal{C})} \inf_{Q_n \in \mathcal{M}_1(\mathcal{X}^n)} \int D(P^n, Q_n) \, \mathrm{d}\pi(P) \\ &= \quad \sup_{\pi \in \mathcal{M}_1(\mathcal{C})} \int D(P^n, P_\pi^n) \, \mathrm{d}\pi(P) , \end{aligned}$$

  where P_π^n *is the mixture distribution given by* $P_\pi(j) = \int P(j) \, \mathrm{d}\pi(P)$.

Moreover, as noted by Acharya et al. [4], the minimax redundancy of a class of I.I.D. sources is equal to the minimax redundancy of the induced class on types. More precisely, for $P \in \mathcal{C}$, let us denote by $\tau(P^n)$ the distribution of the *type* of a sequence $(X_1, \ldots, X_n) \sim P^n$, *i.e.* $\tau(P^n)$ is the probability

distribution of the sequence $(X_{n,j})_{j\geq 1}$. For a class $\mathcal{C}$ of sources over $\mathcal{X}$, we define the class $\tau(\mathcal{C}^n)$ as

$$\tau(\mathcal{C}^n) \quad = \quad \{\tau(P^n),\, P \in \mathcal{C}\}\ .$$

Then we have

$$\overline{R}(\mathcal{C}^n) = \overline{R}(\tau(\mathcal{C}^n)) \quad \text{and} \quad \widehat{R}(\mathcal{C}^n) = \widehat{R}(\tau(\mathcal{C}^n))\,.$$

3.2 Poisson sampling

In Poisson sampling, we assume that the number of symbols is distributed as a Poisson random variable N with mean n, denoted $N \sim \mathcal{P}(n)$. Let $\mathcal{C}^{\mathcal{P}(n)}$ be the Poissonized version of $\mathcal{C}^n$:

$$\mathcal{C}^{\mathcal{P}(n)} = \left\{P^{\mathcal{P}(n)},\, P \in \mathcal{C}\right\},$$

where, for all $P \in \mathcal{C}$ and $x_{1:k} \in \mathcal{X}^*$,

$$P^{\mathcal{P}(n)}(x_{1:k}) = \mathbb{P}(N = k).P^k(x_{1:k}).$$

By convention, the empty sequence has probability zero. For $j \geq 1$, let us denote by $X_j(n)$ the number of occurrences of symbol j in a Poisson sample with size $\mathcal{P}(n)$. Then $X_j(n)$ is distributed as $\mathcal{P}(nP(j))$, and a very useful property of Poisson sampling is that the symbol counts $(X_j(n))_{j\geq 1}$ are independent. Let us also note that, as in the fixed-n setting, the redundancy of $\mathcal{C}^{\mathcal{P}(n)}$ is equal to the type-redundancy:

$$\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) = \overline{R}\left(\tau(\mathcal{C}^{\mathcal{P}(n)})\right) \quad \text{and} \quad \widehat{R}(\mathcal{C}^{\mathcal{P}(n)}) = \widehat{R}\left(\tau(\mathcal{C}^{\mathcal{P}(n)})\right)\ . \tag{3.1}$$

(see Acharya et al. [4]).

We will often resort to the concentration properties of the Poisson distribution.

Lemma 3.1. *Let $N \sim \mathcal{P}(n)$. Then, for all $t > 0$,*

$$\begin{aligned}
\mathbb{P}\left(N \geq n + t\right) &\leq \exp\left(-\frac{t^2}{2(n+t)}\right)\ ,\\
\mathbb{P}\left(N \leq n - t\right) &\leq \exp\left(-\frac{t^2}{2n}\right)\ .
\end{aligned}$$

Proposition 3.2. *For all class $\mathcal{C}$, the Poissonized minimax redundancy satisfies*

$$\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) \quad = \quad \inf_{(Q_k)} \sup_{P\in\mathcal{C}} \sum_{k\geq 0} \mathbb{P}(N = k) D(P^k, Q_k)\,,$$

where the infimum is taken over sequences $(Q_k)_{k\geq 0}$, such that, for all $k \geq 0$, Q_k is a probability distribution over $\mathcal{X}^k$.

To benefit from the independence property of Poisson sampling, we would like to relate the redundancy of the fixed-n setting with that of the Poisson-length sample. This was done in Theorem 2 of [4] for the minimax regret, and in Falahatgar et al. [72], the authors show that $\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) \leq 2\overline{R}(\mathcal{C}^n)$. Proposition 3.3 gives similar bounds in the two directions.

Proposition 3.3. *For any class $\mathcal{C}$ with $\overline{R}(\mathcal{C}) < \infty$, we have*

$$\overline{R}\left(\mathcal{C}^{\mathcal{P}(n-n^{2/3})}\right) + o_{\mathcal{C}}(1) \leq \overline{R}(\mathcal{C}^n) \leq (1 - o(1))\overline{R}\left(\mathcal{C}^{\mathcal{P}(n+n^{2/3})}\right)$$

3.3 Minimax redundancy of envelope classes

In the case of an envelope source class $\mathcal{C} = \Lambda_f$, the class $\tau\left(\Lambda_f^{\mathcal{P}(n)}\right)$ can be written as

$$\tau\left(\Lambda_f^{\mathcal{P}(n)}\right) = \left\{\prod_{j=1}^{\infty} \mathcal{P}(np_j) : (p_j)_{j\geq 1} \in \mathcal{M}_1(\mathcal{X}),\, \forall j \geq 1,\, p_{j+\ell_f} \leq f_j\right\}. \tag{3.2}$$

Let us define, for $\lambda \geq 0$,

$$\mathcal{P}^\star(\lambda) \quad = \quad \{\mathcal{P}(\mu) : \mu \leq \lambda\}\,,$$

the class of Poisson distributions with mean smaller than λ, we have the following bounds.

Lemma 3.2. *Let Λ_f be an envelope class. Then for all $n \geq 0$,*

$$\sum_{j=1}^{\infty} \overline{R}(\mathcal{P}^\star(nf_j)) \leq \overline{R}\left(\Lambda_f^{\mathcal{P}(n)}\right) \leq \ell_f \overline{R}(\mathcal{P}^\star(n)) + \sum_{j=1}^{\infty} \overline{R}(\mathcal{P}^*(nf_j))\,.$$

We now have almost all the ingredients to establish our bounds on $\overline{R}\left(\Lambda_f^n\right)$. We first prove the upper bound, which is simply obtained by using the fact that $\overline{R}\left(\Lambda_f^n\right) \leq \widehat{R}\left(\Lambda_f^n\right)$, and applying the bounds of [4] on minimax regrets. We then interpret those bounds using Karlin's formalism, which allows us to gain insight on the quantities involved.

Proposition 3.4. *For any envelope function f, the minimax redundancy of Λ_f^n satisfies*

$$\overline{R}\left(\Lambda_f^n\right) \quad \leq \quad \log(\mathrm{e})\left(\int_{1/n}^{1} \frac{\vec{\nu}_f(x)}{2x}\,\mathrm{d}x + \vec{\nu}_f(1/n) + n\nu_{1,f}[0,1/n]\right) + O(\ell_f \log n)\,. \tag{3.3}$$

Remark 3.1. *Note that, as soon as the support of $(f_j)_{j\geq 1}$ is infinite, we have $\overline{R}(\Lambda_f^n) \gg \log n$. Hence, one may neglect the term $O(\ell_f \log n)$. Now, Karlin's formalism combined with the regular variation assumption on the envelope allows us to evaluate the terms of (3.3). In the case $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_0$, Karamata's Theorem implies that the dominant term is the first one, which can be written as*

$$\mathbf{R}_f(n) \quad = \quad \log(\mathrm{e})\int_1^n \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t\,.$$

When $\alpha = 1$, the leading term is now $n\vec{\nu}_{1,f}[0,1/n]$, which is of order $\mathbb{E}_f[K_n] \gg \vec{\nu}_f(1/n)$. And when $0 < \alpha < 1$, the first three terms are of the same order, that of $\vec{\nu}_f(1/n) \asymp \mathbb{E}_f[K_n] \asymp \mathbf{R}_f(n)$. Hence, this establishes the upper bound in Theorem 2.1.

In the other direction, we need a lower bound on $\overline{R}\left(\mathcal{P}^\star(nf_j)\right)$, which is given in the following lemma.

Lemma 3.3. *For $\lambda \geq 1$, the redundancy of $\mathcal{P}^\star(\lambda)$ satisfies*

$$\overline{R}(\mathcal{P}^\star(\lambda)) \quad \geq \quad \frac{\log\lambda}{2} - 5\,.$$

Combining Proposition 3.3, the lower bound in Lemma 3.2, and Lemma 3.3, we obtain the following.

Proposition 3.5. *Let $m = n - n^{2/3}$. For any envelope function f and for large enough n,*

$$\overline{R}\left(\Lambda_f^n\right) \quad \geq \quad \log(\mathrm{e})\int_1^m \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t - 5\vec{\nu}_f(1/m) - 1\,.$$

Remark 3.2. *Assume $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_0$. Then the function $\mathbf{R}_f(n)$ also is of slow variation. Thus, $\mathbf{R}_f(m) \sim \mathbf{R}_f(n)$. Moreover, as noted earlier, $\vec{\nu}_f(1/n) = o(\mathbf{R}_f(n))$. Hence, combining this with Proposition 3.4 and Remark 3.1, we obtain that, if $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_0$, then*

$$\overline{R}\left(\Lambda_f^n\right) \underset{n\to\infty}{\sim} \log(\mathrm{e}) \int_1^n \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t\,.$$

Note that the lower bound in Proposition 3.5 might be irrelevant. For instance, when $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$ with $\alpha > \log(\mathrm{e})/10$, the right-hand side becomes negative. However, an order-optimal lower bound in the case of heavy-tailed envelopes (that is, when $\alpha > 0$) was established by Boucheron et al. [41]: the expected redundancy of a class Λ_f^n is lower-bounded by the expected number of distinct symbols in a sample of size n drawn according to the envelope distribution $(f_j)_{j\geq 1}$. Combining the two bounds, we have the following proposition.

Proposition 3.6. *Let $m = n - n^{2/3}$. For any envelope function f, there exists a constant $c_f > 0$ such, for large enough n,*

$$\overline{R}\left(\Lambda_f^n\right) \geq (\mathbf{R}_f(m) - 5\vec{\nu}_f(1/m) - 1) \vee (\mathbb{E}_f[K_n] - c_f)\,.$$

Recall that, when $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$ for $0 < \alpha < 1$, we have: $\mathbb{E}_f[K_n] \asymp \vec{\nu}_f(1/n) \asymp \mathbf{R}_f(n)$. Hence, combining Propositions 3.4 and 3.6 establishes Theorem 2.1.

4 Analysis of the Pattern Censoring Code

Before proceeding with the analysis of the code, we first state some useful comparisons between the expected occupancy counts, the missing mass and the measure ν_1, under the source and under the envelope, as well as some asymptotics which are valid under the regular variation hypothesis on the envelope. As our code is fundamentally related to occupancy counts and to the occurrences of new symbols, those will be very helpful to evaluate the contribution of each term to the redundancy.

Lemma 4.1. *We always have*

$$\mathbb{E}M_{n,0} \leq \frac{\mathbb{E}K_n}{n}\,.$$

Moreover, if $P \in \Lambda_f$, then $\mathbb{E}K_n \leq \ell_f + \mathbb{E}_f[K_n]$, and, if $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$ with $\alpha \in [0,1[$, then

$$\mathbb{E}_f[K_n] \underset{+\infty}{\sim} \Gamma(1-\alpha)\vec{\nu}_f(1/n)\,,$$

and, for all $\varepsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$,

$$\nu_{1,f}[0,1/n] \leq \frac{(\alpha+\varepsilon)\vec{\nu}_f(1/n)}{(1-\alpha)n}\,.$$

We now proceed with the analysis of the redundancy of the PCC Code. As announced at (1.3), we rearrange the instantaneous redundancy of symbol $i+1$, given $X_{1:i}$, as follows

$$\begin{aligned}
\mathbb{E}_P\left[\log\frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})}\Big|X_{1:i}\right] &= \sum_{j\geq 1} p_j \log\left(ip_j\left(1+\frac{K_i+1}{2i}\right)\right) + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0}\log\left(\frac{1}{X_{i,j}-\frac{1}{2}}\right) \\
&\quad + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\left(\frac{j+1}{K_i+\frac{1}{2}}\right) + 2\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0}\log\log(j+1) \\
&:= A+B+C+D\,.
\end{aligned}$$

We now average over the first i symbols. Starting with term A, we use the fact that for all $x \geq 0$, $\log(1+x) \leq \log(\mathrm{e})x$ and obtain

$$\mathbb{E}[A] \quad \leq \quad \sum_{j\geq 1} p_j \log(ip_j) + \log(\mathrm{e})\frac{\mathbb{E}[K_i]+1}{2i}\,.$$

Moving on to term B, thanks to Jensen's inequality for conditional expectation, we have

$$\begin{aligned}\mathbb{E}[B] \quad &= \quad \mathbb{E}\left[\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0}\mathbb{E}\left[\log\frac{1}{X_{i,j}-\frac{1}{2}} \mid X_{i,j}>0\right]\right]\\ &\leq \quad \mathbb{E}\left[\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0}\log\mathbb{E}\left[\frac{1}{X_{i,j}-\frac{1}{2}} \mid X_{i,j}>0\right]\right]\end{aligned}$$

Lemma 4.2. *Let $X \sim \mathcal{B}(n,p)$. Then*

$$\mathbb{E}\left[\frac{1}{X-\frac{1}{2}} \mid X>0\right] \quad \leq \quad \frac{1}{np}+\frac{9}{(np)^2}$$

Resorting to Lemma 4.2, we have

$$\begin{aligned}\mathbb{E}[B] \quad &\leq \quad \mathbb{E}\left[\sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}>0}\log\left(\frac{1}{ip_j}\left(1+\frac{9}{ip_j}\right)\right)\right]\\ &\leq \quad -\sum_{j\geq 1} p_j\mathbb{P}(X_{i,j}>0)\log(ip_j)+\log(\mathrm{e})\frac{9\mathbb{E}[K_i]}{i}\,.\end{aligned}$$

Hence,

$$\mathbb{E}[A]+\mathbb{E}[B] \quad \leq \quad \sum_{j\geq 1} p_j\mathbb{P}(X_{i,j}=0)\log(ip_j)+\log(\mathrm{e})\frac{10\mathbb{E}[K_i]}{i}\,.$$

Notice that, as soon as $p_j \leq 1/i$, $\log(ip_j)$ is less than zero. We may thus only retain the terms with $p_j \geq 1/i$. Using $\log(ip_j) \leq ip_j$ and that, for all $x \in [0,1]$, $x^2(1-x)^i \leq \left(\frac{2}{i+2}\right)^2\left(1-\frac{2}{i+2}\right)^i \leq 1/i^2$, we have

$$\begin{aligned}\sum_{j\geq 1} p_j\mathbb{P}(X_{i,j}=0)\log(ip_j) \quad &\leq \quad \sum_{j,\,p_j\geq 1/i} ip_j^2(1-p_j)^i\\ &\leq \quad \frac{\vec{\nu}(1/i)}{i} \leq \frac{\vec{\nu}_f(1/i)+\ell_f}{i}\,.\end{aligned}$$

As $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$ with $\alpha \in [0,1[$, we have, thanks to Lemma 4.1, that there exists $i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$,

$$\mathbb{E}[A]+\mathbb{E}[B] \quad \leq \quad 16\Gamma(1-\alpha)\frac{\vec{\nu}_f(1/i)}{i}\,. \tag{4.1}$$

It now remains to bound the terms C and D. Using that $K_i + 1/2 \geq \frac{1}{2}(K_i+1)$, we have

$$\begin{aligned} C &\leq M_{i,0} + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0} \log\left(\frac{j+1}{K_i+1}\right) \\ &= M_{i,0} + \sum_{j\geq 1} p_j \mathbb{1}_{X_{i,j}=0} \log\left(\frac{j+1}{\sum_{\ell\neq j}\mathbb{1}_{X_{i,\ell}>0}+1}\right). \end{aligned}$$

Recall that the variables $(X_{i,j})_{j\geq 1}$ are negatively associated (see Chapter 1, Section 6.1.2). Hence, for all $j \geq 1$,

$$\mathbb{E}\left[\mathbb{1}_{X_{i,j}=0}\log\left(\frac{j+1}{\sum_{\ell\neq j}\mathbb{1}_{X_{i,\ell}>0}+1}\right)\right] \leq \mathbb{P}(X_{i,j}=0)\mathbb{E}\left[\log\left(\frac{j+1}{\sum_{\ell\neq j}\mathbb{1}_{X_{i,\ell}>0}+1}\right)\right].$$

Resorting to Jensen's inequality and noticing that $\sum_{\ell\neq j}\mathbb{1}_{X_{i,\ell}>0}+1 \geq K_i$, we get

$$\mathbb{E}[C] \leq \mathbb{E}[M_{i,0}] + \sum_{j\geq 1} p_j \mathbb{P}(X_{i,j}=0)\log\mathbb{E}\left[\frac{j+1}{K_i}\right].$$

Lemma 4.3. *The variable K_i satisfies*

$$\mathbb{E}\left[\frac{1}{K_i}\right] \leq \frac{5}{\mathbb{E}[K_i]}.$$

Using Lemma 4.3, we obtain

$$\mathbb{E}[C] \leq (1+\log(5))\mathbb{E}[M_{i,0}] + \sum_{j\geq 1} p_j \mathbb{P}(X_{i,j}=0)\log\left(\frac{j+1}{\mathbb{E}[K_i]}\right). \tag{4.2}$$

Bounding the sum in the right-hand side of (4.2) is the difficult part. To control it, we need to use the regular decay of the envelope. However, bounding this term by the corresponding quantity under the envelope distribution is not straightforward, because $\mathbb{E}[K_i]$, which is always smaller than $\mathbb{E}_f[K_i]+\ell_f$, now appears in the denominator. Lemma 4.4 gives a way to relate it to the envelope, and Lemma 4.5 gives an evaluation when the envelope is regularly varying.

Lemma 4.4. *If $P \in \Lambda_f$,*

$$\sum_{j\geq 1} p_j \mathbb{P}(X_{i,j}=0)\log\left(\frac{j+1}{\mathbb{E}[K_i]}\right) \leq \sum_{j\geq \vec{\nu}_f(1/i)} f_j \log\left(\frac{j}{\vec{\nu}_f(1/i)}\right) + \frac{6(\mathbb{E}_f[K_i]+\ell_f)}{5i}.$$

Lemma 4.5. *Assume that $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$ with $\alpha \in [0,1[$ or $\vec{\nu}_f(1/\cdot) \in \Pi_{\ell_0}$. Then there exists $i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$,*

$$\sum_{j\geq \vec{\nu}_f(1/i)} f_j \log\left(\frac{j}{\vec{\nu}_f(1/i)}\right) \leq \frac{4-\alpha}{(1-\alpha)^2}\cdot\frac{\vec{\nu}_f(1/i)}{i}.$$

Putting together equations (4.1), (4.2) and Lemma 4.4 and 4.5, we obtain

$$\mathbb{E}[A+B+C] \leq \left(21\Gamma(1-\alpha) + \frac{4-\alpha}{(1-\alpha)^2}\right)\frac{\vec{\nu}_f(1/i)}{i}, \tag{4.3}$$

for i large enough.

As for term D, we have the following bound.

Lemma 4.6. *There exists $i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$*

$$\mathbb{E}[D] \quad \leq \quad \frac{5}{2}\left(\Gamma(1-\alpha) + \frac{1}{1-\alpha}\right)\frac{\log\log(i)\vec{\nu}_f(1/i)}{i}\,.$$

As the bound on $\mathbb{E}D$ in Lemma 4.6 is much than the bound on $\mathbb{E}[A+B+C]$ in (4.3), we finally obtain that there exists $i_0 \in \mathbb{N}$ such that for all $i \geq i_0$,

$$\mathbb{E}_P\left[\log\frac{P(X_{i+1})}{Q(X_{i+1}|X_{1:i})}\Big|X_{1:i}\right] \quad \leq \quad c_\alpha\frac{\log\log(i)\vec{\nu}_f(1/i)}{i}\,,$$

with $c_\alpha = 3\left(\Gamma(1-\alpha) + \frac{1}{1-\alpha}\right)$.

Hence, for $n \geq i_0$, we obtain

$$\begin{aligned}\overline{R}(Q_n, \Lambda_f^n) \quad &\leq \quad C(i_0, f) + \sum_{i=i_0}^{n} c_\alpha\frac{\log\log(i)\vec{\nu}_f(1/i)}{i} \\ &\leq \quad C(i_0, f) + \log\log(n)\sum_{i=1}^{n} c_\alpha\frac{\vec{\nu}_f(1/i)}{i}\,,\end{aligned}$$

which establishes the upper bound in Theorem 2.2.

5 Proofs

Proof of Proposition 3.2. We have

$$\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) = \inf_{Q\in\mathcal{M}_1(\mathcal{X}^*)}\sup_{P\in\mathcal{C}} D(P^{\mathcal{P}(n)}, Q).$$

Let $P \in \mathcal{C}$ and $Q \in \mathcal{M}_1(\mathcal{X}^*)$. The distribution Q can be written as $Q = \sum_{k\geq 0} q(k)Q_k$, where $(q(k))_{k\geq 0}$ is a probability distribution over $\mathbb{N}$, and, for each $k \geq 0$, Q_k is a distribution over $\mathcal{C}^k$. Hence

$$\begin{aligned}D(P^{\mathcal{P}(n)}, Q) \quad &= \quad \sum_{k\geq 0}\mathbb{P}(N=k)\sum_{x\in\mathcal{C}^k} P^k(x)\log\frac{\mathbb{P}(N=k)P^k(x)}{q(k)Q_k} \\ &= \quad D(\mathcal{P}(n), q) + \sum_{k\geq 0}\mathbb{P}(N=k)D(P^k, Q_k)\,. \qquad (5.1)\end{aligned}$$

Maximizing in P and minimizing in $(q(k))_{k\geq 0}$ and $(Q_k)_{k\geq 0}$, we get

$$\begin{aligned}\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) \quad &= \quad \inf_{(Q_k)}\inf_{(q(k))}\left(D(\mathcal{P}(n), q) + \sup_{P\in\mathcal{C}}\sum_{k\geq 0}\mathbb{P}(N=k)D(P^k, Q_k)\right) \\ &= \quad \inf_{(q(k))} D(\mathcal{P}(n), q) + \inf_{(Q_k)}\sup_{P\in\mathcal{C}}\sum_{k\geq 0}\mathbb{P}(N=k)D(P^k, Q_k)\,.\end{aligned}$$

The first term is equal to zero for $q = \mathcal{P}(n)$, implying that the distribution Q which achieves the minimax redundancy is also a Poisson mixture. Hence

$$\overline{R}(\mathcal{C}^{\mathcal{P}(n)}) \quad = \quad \inf_{(Q_k)}\sup_{P\in\mathcal{C}}\sum_{k\geq 0}\mathbb{P}(N=k)D(P^k, Q_k)\,.$$

□

Proof of Proposition 3.3. We start with the lower bound on $\overline{R}(\mathcal{C}^n)$. Let $m = n - n^{2/3}$, and let M be a Poisson random variable with mean m. By Proposition 3.2,

$$\begin{aligned}
\overline{R}(\mathcal{C}^{\mathcal{P}(m)}) &= \inf_{(Q_k)} \sup_{P\in\mathcal{C}} \sum_{k\geq 0} \mathbb{P}(M=k) D(P^k, Q_k) \\
&\leq \inf_{(Q_k)} \sum_{k\geq 0} \mathbb{P}(M=k) \sup_{P\in\mathcal{C}} D(P^k, Q_k) \\
&= \sum_{k\geq 0} \mathbb{P}(M=k) \inf_{Q_k} \sup_{P\in\mathcal{C}} D(P^k, Q_k) \\
&= \sum_{k\geq 0} \mathbb{P}(M=k) \overline{R}(\mathcal{C}^k)\,.
\end{aligned}$$

Using the fact that the sequence $(\overline{R}(\mathcal{C}^k))_{k\geq 0}$ is increasing and sub-additive (see Proposition 3.1), we have

$$\begin{aligned}
\overline{R}(\mathcal{C}^{\mathcal{P}(m)}) &\leq \overline{R}(\mathcal{C}^n) + \sum_{k>n} \mathbb{P}(M=k) \left(\overline{R}(\mathcal{C}^k) - \overline{R}(\mathcal{C}^n)\right) \\
&\leq \overline{R}(\mathcal{C}^n) + \sum_{k>n} \mathbb{P}(M=k) \overline{R}(\mathcal{C}^{k-n}) \\
&\leq \overline{R}(\mathcal{C}^n) + \overline{R}(\mathcal{C}) \sum_{k>n} \mathbb{P}(M=k)(k-n)\,.
\end{aligned}$$

Resorting to Lemma 3.1, we have

$$\begin{aligned}
\sum_{k>n} \mathbb{P}(M=k)(k-n) &= \mathbb{E}\left[(M-n)\mathbb{1}_{\{M>n\}}\right] = \int_0^\infty \mathbb{P}\left(M - m > t + n^{2/3}\right) \mathrm{d}t \\
&\leq \int_0^n \mathrm{e}^{-\frac{n^{4/3}}{2(m+n^{2/3})}} \,\mathrm{d}t + \int_n^\infty \mathrm{e}^{-\frac{t^2}{6t}} \,\mathrm{d}t \\
&\leq n\,\mathrm{e}^{-n^{1/3}/2} + 6\,\mathrm{e}^{-n/6} \to 0\,.
\end{aligned}$$

This establishes the lower bound on $\overline{R}(\mathcal{C}^n)$ in Proposition 3.3. Let us now proceed with the other direction. Let now $m = n + n^{2/3}$ and M be a Poisson random variable with mean m. Using the Bayesian representation of the minimax redundancy (see Proposition 3.1), we have

$$\overline{R}(\mathcal{C}^{\mathcal{P}(m)}) = \sup_{\pi\in\mathcal{M}_1(\mathcal{C})} \inf_{Q\in\mathcal{M}_1(\mathcal{X}^*)} \int D(P^{\mathcal{P}(m)}, Q)\,\mathrm{d}\pi(P)\,.$$

Fix $\pi \in \mathcal{M}_1(\mathcal{C})$. Resorting to equation (5.1) in the proof of Proposition 3.2, we have

$$\begin{aligned}
&\inf_{Q\in\mathcal{M}_1(\mathcal{X}^*)} \int D(P^{\mathcal{P}(m)}, Q)\,\mathrm{d}\pi(P) \\
&\quad= \inf_{(Q_k),(q(k))} \int \left(D(\mathcal{P}(m), q) + \sum_{k\geq 0} \mathbb{P}(M=k) D(P^k, Q_k) \right) \mathrm{d}\pi(P) \\
&\quad= \inf_{(Q_k)} \int \sum_{k\geq 0} \mathbb{P}(M=k) D(P^k, Q_k)\,\mathrm{d}\pi(P) \\
&\quad= \sum_{k\geq 0} \mathbb{P}(M=k) \inf_{Q_k} \int D(P^k, Q_k)\,\mathrm{d}\pi(P)\,.
\end{aligned}$$

We claim that the sequence $\left(\inf_{Q_k}\int D(P^k,Q_k)\,\mathrm{d}\pi(P)\right)_{k\geq 0}$ is increasing. Indeed, let $k\geq 0$ and let $Q_{k+1}\in\mathcal{M}_1(\mathcal{X}^k)$. Denote by $Q_{k+1}^{(k)}$ its restriction to the first k symbols. Then, for all $P\in\mathcal{M}_1(\mathcal{X})$, $D(Q_{k+1},P^{k+1})\geq D(P^k,Q_{k+1}^{(k)})$. Hence for all Q_{k+1} there exist $Q_k\in\mathcal{M}_1(\mathcal{X}^k)$ such that $\int D(P^k,Q_k)\,\mathrm{d}\pi(P)\leq\int D(P^{k+1},Q_{k+1})\,\mathrm{d}\pi(P)$, which gives the desired result. We get

$$\begin{aligned}\overline{R}(\mathcal{C}^{\mathcal{P}(m)}) &\geq \sup_\pi\sum_{k\geq n}\mathbb{P}(M=k)\inf_{Q_k}\int D(P^k,Q_k)\,\mathrm{d}\pi(P)\\ &\geq \mathbb{P}(M\geq n)\sup_\pi\inf_{Q_n}\int D(P^n,Q_n)\,\mathrm{d}\pi(P)\\ &\geq \mathbb{P}(M\geq n)\overline{R}(\mathcal{C}^n)\,.\end{aligned}$$

Now, using again Lemma 3.1, we have

$$\mathbb{P}(M\geq n)\quad\geq\quad 1-\exp\left(-\frac{n^{4/3}}{2m}\right)\ \to 1\,,$$

which concludes the proof. □

Proof of Proposition 3.4. As announced, we start by the obvious fact that $\overline{R}\left(\Lambda_f^n\right)\leq\widehat{R}\left(\Lambda_f^n\right)$. Now Theorem 14 of [4] states that

$$\widehat{R}\left(\Lambda_f^n\right)\quad\leq\quad 1+\widehat{R}\left(\Lambda_f^{\mathcal{P}(n)}\right)\,,$$

and, by (3.1), $\widehat{R}\left(\Lambda_f^{\mathcal{P}(n)}\right)=\widehat{R}\left(\tau(\Lambda_f^{\mathcal{P}(n)})\right)$. Now, using (3.2) and the fact that for $j\leq\ell_f$, we still have the crude bound $p_j\leq 1$,

$$\tau(\Lambda_f^{\mathcal{P}(n)})\quad\subset\quad\left(\prod_{j=1}^{\ell_f}\mathcal{P}^\star(n)\right)\prod_{j=1}^{\infty}\mathcal{P}^\star(nf_j)\,.$$

As increasing the class can not reduce the regret, and as, for two classes $\mathcal{C}_1,\mathcal{C}_2$, $\widehat{R}(\mathcal{C}_1\times\mathcal{C}_2)=\widehat{R}(\mathcal{C}_1)+\widehat{R}(\mathcal{C}_2)$, we have

$$\widehat{R}\left(\tau(\Lambda_f^{\mathcal{P}(n)})\right)\quad\leq\quad\ell_f\widehat{R}\left(\mathcal{P}^\star(n)\right)+\sum_{j\geq 1}\widehat{R}\left(\mathcal{P}^\star(nf_j)\right)\,.$$

Now Lemma 17 of [4] provide us with the following bounds: if $\lambda\leq 1$,

$$\widehat{R}\left(\mathcal{P}^\star(\lambda)\right)=\log\left(2-\mathrm{e}^{-\lambda}\right)\leq\log(\mathrm{e})\lambda\,,$$

and, if $\lambda>1$,

$$\widehat{R}\left(\mathcal{P}^\star(\lambda)\right)\leq\log\left(\sqrt{\frac{2\lambda}{\pi}}+2\right)\,.$$

For aesthetic purposes (getting the common $\log(\mathrm{e})$ constant in front of each terms), we find it worth to notice that the bound above can be very slightly improved to

$$\widehat{R}\left(\mathcal{P}^\star(\lambda)\right)\leq\log\left(\sqrt{\frac{2\lambda}{\pi}}+\frac{3}{2}\right)\,.$$

Hence, as $\frac{3}{2} + \sqrt{\frac{2}{\pi}} \leq \mathrm{e}$, we obtain

$$\widehat{R}\left(\Lambda_f^n\right) \quad \leq \quad O(\ell_f \log n) + \sum_{j, f_j \geq 1/n} \log\left(\mathrm{e}\sqrt{nf_j}\right) + \log(\mathrm{e}) \sum_{j, f_j < 1/n} nf_j \,.$$

Now, using the integral representation and integrating by parts gives

$$\begin{aligned}\sum_{f_j \geq 1/n} \log\left(\mathrm{e}\sqrt{nf_j}\right) &= \log(\mathrm{e})\vec{\nu}_f(1/n) + \int_{1/n}^1 \frac{\log(nx)}{2} \nu_f(\,\mathrm{d}x) \\ &= \log(\mathrm{e})\left(\vec{\nu}_f(1/n) + \int_{1/n}^1 \frac{\vec{\nu}_f(x)}{2x}\,\mathrm{d}x\right).\end{aligned}$$

Also, we may write $\sum_{j, f_j < 1/n} f_j = \nu_{1,f}[0, 1/n]$, which gives the desired result. □

Proof of Lemma 3.3. Using the Bayesian representation of the minimax redundancy, we have

$$\overline{R}(\mathcal{P}^\star(\lambda)) \quad = \quad \sup_{\pi \in \mathcal{M}_1([0,\lambda])} \int D(\mathcal{P}(\mu), \mathcal{P}_\pi)\,\mathrm{d}\pi(\mu)\,,$$

where $\mathcal{P}_\pi = \int \mathcal{P}(\mu)\,\mathrm{d}\pi(\mu)$. In particular, taking π equal to the uniform distribution over $[0, \lambda]$, we get

$$\overline{R}(\mathcal{P}^\star(\lambda)) \quad \geq \quad \int_0^\lambda \frac{1}{\lambda} \sum_{k \geq 0} \mathbb{P}(\mathcal{P}(\mu) = k) \log \frac{\mathbb{P}(\mathcal{P}(\mu) = k)}{\mathbb{P}(\mathcal{P}_\pi = k)}\,\mathrm{d}\mu\,.$$

We have

$$\mathbb{P}(\mathcal{P}_\pi = k) \quad = \quad \frac{1}{\lambda}\int_0^\lambda \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\,\mathrm{d}\mu \;=\; \frac{\mathbb{P}(\mathcal{P}(\lambda) > k)}{\lambda} \;\leq\; \frac{1}{\lambda}\,.$$

Hence

$$\overline{R}(\mathcal{P}^\star(\lambda)) \quad \geq \quad \log\lambda - \frac{1}{\lambda}\int_0^\lambda H(\mathcal{P}(\mu))\,\mathrm{d}\mu\,.$$

Using Stirling's bound $k! \leq \mathrm{e}^{1/12k}\left(\frac{k}{\mathrm{e}}\right)^k \sqrt{2\pi k}$, we have, for all $\mu \in [0, \lambda]$,

$$\begin{aligned}\frac{H(\mathcal{P}(\mu))}{\log(\mathrm{e})} &= \mu - \mu\ln\mu + \sum_{k \geq 0} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\ln(k!) \\ &\leq \mu - \mu\ln\mu + \sum_{k \geq 1} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\left(k\ln k - k + \frac{\ln(2\pi k)}{2} + \frac{1}{12k}\right) \\ &\leq \sum_{k \geq 1} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\left(k\ln k + \frac{\ln(2\pi k)}{2}\right) - \mu\ln\mu + \frac{1}{12}\,.\end{aligned}$$

We use Jensen's inequality to obtain

$$\sum_{k \geq 1} \frac{\mathrm{e}^{-\mu}\mu^k}{k!} k\ln k \quad = \quad \mu \sum_{k \geq 0} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\ln(k+1) \;\leq\; \mu\ln(1+\mu)\,,$$

and

$$\sum_{k\geq 1} \frac{\mathrm{e}^{-\mu}\mu^k}{k!} \ln k \quad \leq \quad (1-\mathrm{e}^{-\mu}) \ln\left(\frac{\mu}{1-\mathrm{e}^{-\mu}}\right) \leq \ln\mu + \frac{1}{\mathrm{e}}\,,$$

where the last inequality is due to the fact that the function $x \mapsto x\ln x$ is larger than $-1/\mathrm{e}$ for all $x \geq 0$. We get

$$\begin{aligned} \frac{H(\mathcal{P}(\mu))}{\log(\mathrm{e})} &\leq \frac{\ln\mu}{2} + \mu\ln\left(1+\frac{1}{\mu}\right) + \frac{\ln(2\pi)}{2} + \frac{1}{2\,\mathrm{e}} + \frac{1}{12} \\ &\leq \frac{\ln\mu}{2} + 3\,, \end{aligned}$$

which is smaller than 3 for $\mu \leq 1$. Hence

$$\begin{aligned} \overline{R}(\mathcal{P}^\star(\lambda)) &\geq \log\lambda - 3\log(\mathrm{e}) - \frac{1}{\lambda}\int_1^\lambda \frac{\log\mu}{2}\,\mathrm{d}\mu \\ &\geq \frac{\log\lambda}{2} - 5\,. \end{aligned}$$

□

Proof of Proposition 3.5. Let $m = n - n^{2/3}$. Thanks to Proposition 3.3, for n large enough

$$\overline{R}\left(\Lambda_f^n\right) \quad \geq \quad \overline{R}\left(\Lambda_f^{\mathcal{P}(m)}\right) - 1\,.$$

Now, by Lemmas 3.2 and 3.3,

$$\overline{R}\left(\Lambda_f^{\mathcal{P}(m)}\right) \quad \geq \quad \sum_{j=1}^{\infty} \overline{R}\left(\mathcal{P}^\star(mf_j)\right) \geq \sum_{j,\, f_j\geq 1/m} \overline{R}\left(\mathcal{P}^\star(mf_j)\right) \geq \sum_{j,\, f_j\geq 1/m} \left(\frac{\log(mf_j)}{2} - 5\right)\,.$$

Using the integral representation and integrating by parts,

$$\begin{aligned} \sum_{j,\, f_j\geq 1/m} \frac{\log(mf_j)}{2} &= \frac{\log(\mathrm{e})}{2}\int_{1/m}^1 \ln(mx)\nu_f(\,\mathrm{d}x) \\ &= \frac{\log(\mathrm{e})}{2}\left([-\vec{\nu}_f(x)\ln(mx)]_{1/m}^1 + \int_{1/m}^1 \frac{\vec{\nu}_f(x)}{x}\,\mathrm{d}x\right) \\ &= \log(\mathrm{e})\int_1^m \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t\,. \end{aligned}$$

We obtain

$$\overline{R}\left(\Lambda_f^n\right) \quad \geq \quad \log(\mathrm{e})\int_1^m \frac{\vec{\nu}_f(1/t)}{2t}\,\mathrm{d}t - 5\vec{\nu}_f(1/m) - 1\,.$$

□

Proof of Lemma 4.1. The first inequality is due to the fact that, for all $x \in [0,1]$,

$$nx(1-x)^n \quad \leq 1-(1-x)^n\,.$$

Hence

$$\begin{aligned}\mathbb{E}M_{n,0} &= \int_0^1 x(1-x)^n \nu(\,\mathrm{d}x) \\ &\leq \frac{1}{n}\int_0^1 \left(1-(1-x)^n\right)\nu(\,\mathrm{d}x) \;=\; \frac{\mathbb{E}K_n}{n}\,.\end{aligned}$$

The relation between $\mathbb{E}K_n$ and $\mathbb{E}_f K_n$ is easily obtained by noticing that the function $x \mapsto 1-(1-x)^n$ is increasing on $[0,1]$, which gives

$$\begin{aligned}\mathbb{E}K_n &= \sum_{j\geq 1}\left(1-(1-p_j)^n\right) \\ &\leq \ell_f + \sum_{j\geq 1}\left(1-(1-p_{j+\ell_f})^n\right) \\ &\leq \ell_f + \sum_{j\geq 1}\left(1-(1-f_j)^n\right) \;=\; \ell_f + \mathbb{E}_f K_n\,.\end{aligned}$$

Assume now that $\vec{\nu}_f(1/\cdot) \in \mathrm{RV}_\alpha$, for $\alpha \in [0,1[$. As pointed out in the previous chapter, we always have $\mathbb{E}_f K_n = \mathbb{E}_f K(n) + o(1)$, so we may move to the simpler Poisson setting without affecting the asymptotic behaviour. We have

$$\mathbb{E}_f K(n) \;=\; \int_0^1 \left(1-\mathrm{e}^{-nx}\right)\nu_f(\,\mathrm{d}x) \;=\; \int_0^1 n\,\mathrm{e}^{-nx}\vec{\nu}_f(x)\,\mathrm{d}x\,,$$

and the equivalence $\mathbb{E}_f K(n) \sim \Gamma(1-\alpha)\vec{\nu}_f(1/n)$ follows from the Tauberian Theorem for monotone densities (see for instance Bingham et al. [30]). The last statement is obtained by Karamata's Theorem (see Section 6.1.3 of Chapter 1), which gives:

$$\begin{aligned}\nu_{1,f}[0,1/n] &= \int_0^{1/n} x\nu_f(\,\mathrm{d}x) \\ &= \left[-\vec{\nu}_f(x)x\right]_0^{1/n} + \int_0^{1/n}\vec{\nu}_f(x)\,\mathrm{d}x \\ &= \int_n^\infty \frac{\vec{\nu}_f(1/t)}{t^2}\,\mathrm{d}t - \frac{\vec{\nu}_f(1/n)}{n}\,.\end{aligned}$$

By Karamata's Theorem, when $0 \leq \alpha < 1$,

$$\int_n^\infty \frac{\vec{\nu}_f(1/t)}{t^2}\,\mathrm{d}t \underset{n\to\infty}{\sim} \tfrac{\vec{\nu}_f(1/n)}{(1-\alpha)n}\,.$$

Thus, when $0 < \alpha < 1$,

$$\nu_{1,f}[0,1/n] \;\sim\; \frac{\alpha\vec{\nu}_f(1/n)}{(1-\alpha)n}\,,$$

and when $\alpha = 0$, $\nu_{1,f}[0,1/n] \ll \vec{\nu}_f(1/n)/n$. In both cases, we have the desired result. □

Proof of Lemma 4.2. Using the fact that, for $k \geq 1$,

$$\frac{1}{k-\frac{1}{2}} = \frac{1}{k+1}\left(1+\frac{3}{2k-1}\right) \leq \frac{1}{k+1} + \frac{9}{(k+1)(k+2)}\,,$$

we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{X-\frac{1}{2}} \mid X>0\right] &\leq \frac{1}{1-(1-p)^n}\sum_{k=1}^{n}\binom{n}{k}p^k(1-p)^{n-k}\left(\frac{1}{k+1}+\frac{9}{(k+1)(k+2)}\right) \\
&= \frac{1}{1-(1-p)^n}\left(\frac{\mathbb{P}\left(\mathcal{B}(n+1,p)\geq 2\right)}{(n+1)p}+\frac{9\mathbb{P}\left(\mathcal{B}(n+2,p)\geq 3\right)}{p^2(n+1)(n+2)}\right) \\
&\leq \frac{1}{np}+\frac{9}{(np)^2}\,.
\end{aligned}$$

□

Proof of Lemma 4.3. The variable K_i is known to be a self-bounding function [38], which implies in particular that it satisfies a sub-Poisson concentration inequality with variance factor $\mathbb{E}K_i$, and thus a sub-Gaussian inequality on the left tail. Hence

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K_i}\right] &= \int_0^\infty \mathbb{P}\left(\frac{1}{K_i}>t\right)\mathrm{d}t = \int_1^\infty \frac{\mathbb{P}(K_i<t)}{t^2}\,\mathrm{d}t \\
&\leq \int_{\mathbb{E}[K_i]/2}^\infty \frac{1}{t^2}\,\mathrm{d}t + \int_1^{\mathbb{E}[K_i]/2}\frac{1}{t^2}\mathbb{P}\left(K_i-\mathbb{E}[K_i]<-\frac{\mathbb{E}[K_i]}{2}\right)\mathrm{d}t \\
&\leq \frac{2}{\mathbb{E}[K_i]}+\mathrm{e}^{-\mathbb{E}[K_i]/8} \leq \frac{5}{\mathbb{E}[K_i]}\,,
\end{aligned}$$

where we used the fact that, for all $x\geq 0,\ \mathrm{e}^{-x/8}\leq 3/x$. □

Proof of Lemma 4.4. We have

$$\sum_{j\geq 1}p_j\mathbb{P}(X_{i,j}=0)\log\left(\frac{j+1}{\mathbb{E}[K_i]}\right) \leq \sum_{j\geq 1}p_j\mathbb{P}(X_{i,j}=0)\log\left(\frac{j+1}{\mathbb{E}_f[K_i]+\ell_f+1}\right)+\mathbb{E}[M_{i,0}]\log\frac{\mathbb{E}_f[K_i]+\ell_f+1}{\mathbb{E}[K_i]}\,.$$

Let us first deal with the second term. Using that $\mathbb{E}[M_{i,0}]\leq\mathbb{E}[K_i]/i$ (see Lemma 4.1), we get that this term is smaller than $g(\mathbb{E}K_i)$ with $g: x\mapsto \frac{x}{i}\log\left(\frac{\mathbb{E}_f K_i+\ell_f+1}{x}\right)$. Maximizing the function g, we see that the worst possible source would be one which satisfies $\mathbb{E}[K_i]=\frac{\mathbb{E}_f[K_i]+\ell_f+1}{2}$. Hence, this second term is always smaller than $\frac{\mathbb{E}_f[K_i]+\ell_f+1}{2i}$. Moving on to first term, we use the following bound:

$$\mathbb{E}_f[K_i] \geq \sum_{j\leq\vec{\nu}_f(1/i)}\left(1-\left(1-f_j\right)^i\right) \geq \left(1-\frac{1}{\mathrm{e}}\right)\vec{\nu}_f(1/i)\,.$$

Hence

$$\sum_{j\geq 1}p_j\mathbb{P}(X_{i,j}=0)\log\left(\frac{j+1}{\mathbb{E}_f[K_i]+\ell_f+1}\right) \leq \log\left(\frac{1}{1-\mathrm{e}^{-1}}\right)\mathbb{E}[M_{i,0}]+\sum_{j\geq 1}p_j\log\left(\frac{j+1}{\vec{\nu}_f(1/i)+\ell_f+1}\right)\,.$$

Now, for $j<\vec{\nu}_f(1/i)+\ell_f$, the summands are negative and we simply omit them to get

$$\begin{aligned}
\sum_{j\geq\vec{\nu}_f(1/i)+\ell_f}p_j\log\left(\frac{j+1}{\vec{\nu}_f(1/i)+\ell_f+1}\right) &= \sum_{j\geq\vec{\nu}_f(1/i)}p_{j+\ell_f}\log\left(\frac{j+\ell_f+1}{\vec{\nu}_f(1/i)+\ell_f+1}\right) \\
&\leq \sum_{j\geq\vec{\nu}_f(1/i)}f_j\log\left(\frac{j}{\vec{\nu}_f(1/i)}\right)\,,
\end{aligned}$$

where we used $p_{j+\ell_f}\leq f_j$ and the fact that for all $a,x,y\geq 0$ with $x\leq y$, we have $\frac{x+a}{y+a}\leq\frac{x}{y}$. □

Proof of Lemma 4.5. Using the integral representation, we have

$$\sum_{j\geq \vec{\nu}_f(1/i)} f_j \log\left(\frac{j}{\vec{\nu}_f(1/i)}\right) \quad = \quad \int_0^{1/i} x \log \frac{\vec{\nu}_f(x)}{\vec{\nu}_f(1/i)} \nu_f(\,\mathrm{d}x)\,.$$

By the regular variation assumption on $\vec{\nu}_f(1/\cdot)$, Potter-Drees Inequality (see Section 6.1.3 of Chapter 1)) implies that: for all $\varepsilon, \delta > 0$, $\exists i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$, for all $x \in (0, 1/i]$,

$$\frac{\vec{\nu}_f(x)}{\vec{\nu}_f(1/i)} \quad \leq \quad \left(\frac{1}{xi}\right)^{\alpha} + \varepsilon \left(\frac{1}{xi}\right)^{\alpha+\delta}\,.$$

Taking crudely $\varepsilon = \delta = 1$ and bounding α by 1, we obtain that for i large enough and $x \in (0, 1/i]$,

$$\frac{\vec{\nu}_f(x)}{\vec{\nu}_f(1/i)} \quad \leq \quad 2\left(\frac{1}{xi}\right)^{2}\,.$$

Hence

$$\int_0^{1/i} x \log \frac{\vec{\nu}_f(x)}{\vec{\nu}_f(1/i)} \nu_f(\,\mathrm{d}x) \quad \leq \quad \nu_{1,f}[0, 1/i] + 2\log(\,\mathrm{e}) \int_0^{1/i} x \ln\left(\frac{1}{xi}\right) \nu_f(\,\mathrm{d}x)\,.$$

By Fubini's Theorem,

$$\begin{aligned}\int_0^{1/i} x \ln\left(\frac{1}{xi}\right) \nu_f(\,\mathrm{d}x) \quad &= \quad \int_0^{1/i} x \int_1^{\frac{1}{xi}} \frac{1}{t}\,\mathrm{d}t\, \nu_f(\,\mathrm{d}x) \\ &= \quad \int_1^{\infty} \frac{1}{t} \int_0^{\frac{1}{ti}} x \nu_f(\,\mathrm{d}x)\,\mathrm{d}t \;=\; \int_i^{\infty} \frac{\nu_{1,f}[0, 1/t]}{t}\,\mathrm{d}t\,.\end{aligned}$$

Now, the last statement in Lemma 4.1 implies that, for all $\varepsilon > 0$, for i large enough and for all $t \geq i$, $\nu_{1,f}[0, 1/t] \leq \frac{(\alpha+\varepsilon)\vec{\nu}_f(1/t)}{(1-\alpha)t}$. Taking $\varepsilon = 1 - \alpha$ and resorting to Karamata's Theorem, we have

$$\int_i^{\infty} \frac{\nu_{1,f}[0, 1/t]}{t}\,\mathrm{d}t \quad \leq \quad \frac{1}{1-\alpha} \int_i^{\infty} \frac{\vec{\nu}_f(1/t)}{t^2}\,\mathrm{d}t \;\sim\; \frac{\vec{\nu}_f(1/i)}{(1-\alpha)^2 i}\,.$$

In the end, we obtain that for i large enough,

$$\sum_{j\geq \vec{\nu}_f(1/i)} f_j \log\left(\frac{j}{\vec{\nu}_f(1/i)}\right) \quad \leq \quad \frac{4-\alpha}{(1-\alpha)^2} \cdot \frac{\vec{\nu}_f(1/i)}{i}\,.$$

□

Proof of Lemma 4.6. Decomposing the sum, we have

$$\sum_{j\geq 1} p_j (1-p_j)^i \log\log(j+1) \quad \leq \quad \sum_{j\geq \vec{\nu}_f(1/i)+\ell_f} p_j \log\log(j+1) + \mathbb{E}[M_{i,0}] \log\log\left(\vec{\nu}_f(1/i) + \ell_f + 1\right)\,,$$

and

$$\sum_{j\geq \vec{\nu}_f(1/i)+\ell_f} p_j \log\log(j+1) \quad \leq \quad \sum_{j\geq \vec{\nu}_f(1/i)} f_j \log\log(j+\ell_f+1)\,.$$

Now, we resort to the integral representation. We notice that, as $\vec{\nu}_f(x) \ll 1/x$ when $x \to 0$ (which is true as soon as the support is infinite, see [80]), then, for large enough i we have $\vec{\nu}_f(x) + \ell_f + 1 \leq 1/x$ for all

$x \in]0, 1/i]$. We then integrate by parts to get

$$\begin{aligned}
\sum_{j \geq \vec{\nu}_f(1/i)} f_j \log\log(j + \ell_f + 1) &= \int_0^{1/i} x \log\log\left(\vec{\nu}_f(x) + \ell_f + 1\right) \nu_f(\,\mathrm{d}x) \\
&\leq \int_0^{1/i} x \log\log\left(\frac{1}{x}\right) \nu_f(\,\mathrm{d}x) \\
&= \left[-\vec{\nu}_f(x) x \log\log\left(\frac{1}{x}\right)\right]_0^{1/i} + \int_0^{1/i} \left(\log\log\left(\frac{1}{x}\right) + \frac{\log(\mathrm{e})}{\ln x}\right) \vec{\nu}_f(x)\,\mathrm{d}x \\
&\leq \int_0^{1/i} \log\log\left(\frac{1}{x}\right) \vec{\nu}_f(x)\,\mathrm{d}x \\
&= \int_i^{\infty} \frac{\log\log(t)\vec{\nu}_f(1/t)}{t^2}\,\mathrm{d}t\,.
\end{aligned}$$

We used that, as $\alpha < 1$, the limit of $\vec{\nu}_f(x) x \log\log(1/x)$ as $x \to 0$ is equal to 0. Now, the function $t \mapsto \frac{\log\log(t)\vec{\nu}_f(1/t)}{t^2}$ belongs to $\mathrm{RV}_{\alpha-2}$. Hence, by Karamata's Theorem (see Section 6.1.3 of Chapter 1),

$$\int_i^{\infty} \frac{\log\log(t)\vec{\nu}_f(1/t)}{t^2}\,\mathrm{d}t \underset{i\to\infty}{\sim} \frac{\vec{\nu}_f(1/i)}{(1-\alpha)i} \log\log(i)\,.$$

We thus obtain that, for all $\varepsilon > 0$, there exists $i_0 \in \mathbb{N}$ such that, for all $i \geq i_0$

$$\mathbb{E}[D] \leq 2(1+\varepsilon)\left(\Gamma(1-\alpha) + \frac{1}{1-\alpha}\right) \frac{\log\log(i)\vec{\nu}_f(1/i)}{i}\,.$$

□

Part II

Cutoff for random walks on sparse random graphs

This part is devoted to the question of mixing times of random walks on sparse random graphs. We introduce the problem in Section 1. Section 2 presents an article with Justin Salez, Cutoff for non-backtracking random walks on sparse random graphs *[21], to appear in the* Annals of Probability. *Section 3 presents an ongoing work with Eyal Lubetzky and Yuval Peres on the comparison of the mixing times of simple and non-backtracking random walks on sparse random graphs.*

1 Introduction

1.1 Mixing times

Let $(X_t)_{t\geq 0}$ be a Markov chain a finite state-space Ω, with transition matrix P, which we assume irreducible and aperiodic. Then, there exists a unique stationary distribution π defined by the equation

$$\pi P \quad = \quad \pi\,,$$

and the chain converges to π, in the sense that, for all $(x,y)\in\Omega^2$,

$$P^t(x,y) \quad \underset{t\to\infty}{\to} \quad \pi(y)\,.$$

One natural question is that of quantifying the speed at which this convergence occurs: after a given number of steps, how close is the chain from equilibrium? To answer such a question, we first need to select an appropriate metric to measure the distance between two probability distributions. The *total-variation distance* between the law of the chain at time t (when started from x) and the stationary distribution is defined as

$$\begin{aligned}\mathcal{D}_x(t) &= \max_{A\subset\Omega}\big(P^t(x,A)-\pi(A)\big)\\ &= \frac{1}{2}\sum_{y\in\Omega}\Big|P^t(x,y)-\pi(y)\Big|\,.\end{aligned}$$

As we are looking for a uniform control over all possible starting points, we will consider the *worst-case* distance

$$\mathcal{D}(t) \quad = \quad \max_{x\in\Omega}\mathcal{D}_x(t)\,.$$

The function $\mathcal{D}(\cdot)$ decreases from almost 1 at time 0 and tends to 0 at $+\infty$. By the Convergence Theorem, we know that the decay is exponential in time: there exist $C>0$ and $\alpha\in(0,1)$ such that $\mathcal{D}(t)\leq C\alpha^t$. In the reversible case, *i.e.* if for all $(x,y)\in\Omega^2$, the detailed balance equation

$$\pi(x)P(x,y)=\pi(y)P(y,x)$$

holds, one can precisely describe the asymptotic rate of this exponential decay. Letting

$$\lambda_\star \quad = \quad \max\{|\lambda|,\ \lambda \text{ eigenvalue of } P,\ \lambda\neq 1\}\,,$$

we have

$$\mathcal{D}(t)^{1/t} \quad \underset{t\to\infty}{\to} \quad \lambda_\star\,. \tag{1.1}$$

The quantity $1-\lambda_\star$ corresponds to the difference between the two largest eigenvalues (in absolute value) of the matrix P, and is called the *absolute spectral gap* of the chain. The inverse of the spectral gap, $t_{\text{REL}} = (1-\lambda_\star)^{-1}$, is called the *relaxation time*. One may observe that, in this setting, the transition matrix is fixed, and we are interested in the asymptotics for the distance as the time t tends to infinity. In particular, by (1.1), $\lambda_\star$ captures the mixing properties of the chain as $t \to \infty$. In the past three decades, with the growing interest for large-size networks, the perspective changed quite radically and the focus moved from fixed chains studied as $t \to \infty$, to chains studied on growing state-spaces. This led to a different asymptotic analysis: some target distance to equilibrium ε is fixed, and we want to understand the time it takes for the chain to reach this distance, as the *size* of the state-space tends to ∞. More precisely, we now consider a sequence of state-spaces $\Omega^{(n)}$ (generally we have $n = |\Omega^{(n)}|$), a sequence of transition matrices $P^{(n)}$, associated with a sequence of distances $\mathcal{D}^{(n)}(t)$, and we define, for $0 < \varepsilon < 1$, the *mixing time* as

$$t_{\text{MIX}}^{(n)}(\varepsilon) \quad = \quad \min\left\{t \geq 0,\, \mathcal{D}^{(n)}(t) \leq \varepsilon\right\}.$$

For ease of notation, we will often omit the dependence in n.

1.2 The cutoff phenomenon

A remarkable phenomenon was discovered in the early eighties by Diaconis and Shahshahani [60] and Aldous [6]: there are situations where, as $n \to \infty$, the mixing time $t_{\text{MIX}}(\varepsilon)$ does not depend on ε, at least to first order. The distance remains close to 1 for a certain amount of time and then abruptly drops to 0. This surprising phenomenon was first discovered in the context of card shuffling: given a certain procedure for shuffling a deck of n cards, how many shuffles do we need to do for the deck to be mixed? It turns out that, for certain procedures, there exists a quite precise number of shuffles slightly below which the deck is far from being mixed, and slightly above which it is almost completely mixed. In 1981, [60] first singled out this phenomenon for random uniform transpositions, and in 1983, [6] established it for the random walk on the hypercube $\{0,1\}^n$. The term *cutoff* and the general formalization appeared shortly after, in the seminal paper by Aldous and Diaconis [7].

Formally, a chain is said to exhibit cutoff if, for all $0 < \varepsilon < 1$,

$$\frac{t_{\text{MIX}}^{(n)}(\varepsilon)}{t_{\text{MIX}}^{(n)}(1-\varepsilon)} \quad \underset{n\to+\infty}{\longrightarrow} \quad 1\,. \tag{1.2}$$

For instance, in the case of the random walk on the hypercube analysed by Aldous [6], the state-space is $\{0,1\}^n$ and a transition consists in choosing uniformly at random a vector among $\{\mathbf{0}\} \cup \{e_i\}_{i=1}^n$ where $\mathbf{0}$ is the zero vector and where, for $1 \leq i \leq n$, e_i is the vector with 0 at all coordinates except a 1 in position i, and adding it, modulo 2, to the current state. Then, for all $0 < \varepsilon < 1$, we have

$$\frac{4t_{\text{MIX}}^{(n)}(\varepsilon)}{n\log n} \quad \underset{n\to+\infty}{\longrightarrow} \quad 1\,.$$

This means that if we rescale time by $\frac{1}{4}n\log n$, then the distance to stationarity approaches the step function:

$$\lim_{n\to\infty} \mathcal{D}\left(\frac{c}{4}n\log n\right) \quad = \quad \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases}$$

A classical example of a random walk on a group which does not have a cut-off is the random walk on

the cycle C_n driven by the uniform distribution on $\{-1,0,1\}$. It takes $\Theta(n^2)$ steps for $\mathcal{D}(t)$ to reach $1/2$, then it takes another $\Theta(n^2)$ steps to go from $1/2$ to $1/4$, and so on. Here, the mixing time is of order n^2, but there is no sharp transition. When the cutoff occurs however, one may properly speak of the mixing time of the chain, without referring to the target ε anymore, and it is common to take $t_{\text{MIX}}^{(n)} = t_{\text{MIX}}^{(n)}(1/4)$. The rate of convergence in (1.2) is addressed by the notion of *cutoff window*: a sequence of Markov chains has cutoff at $t_{\text{MIX}}^{(n)}$ with window ω_n if $\omega_n = o(t_{\text{MIX}}^{(n)})$ and

$$\begin{aligned} \lim_{\alpha\to+\infty} \liminf_{n\to+\infty} \mathcal{D}\left(t_{\text{MIX}}^{(n)} - \alpha\omega_n\right) &= 1\,, \\ \lim_{\alpha\to+\infty} \limsup_{n\to+\infty} \mathcal{D}\left(t_{\text{MIX}}^{(n)} + \alpha\omega_n\right) &= 0\,. \end{aligned}$$

It may even be possible to describe very precisely the *shape* of the distance inside the cutoff window. In the case of the random walk on the hypercube, Diaconis et al. [61] showed that, for all $\alpha \in \mathbb{R}$,

$$\mathcal{D}\left(\frac{1}{4}n\log n + \alpha n\right) = \phi\left(\mathrm{e}^{-2\alpha}/\sqrt{8}\right) + o(1)\,,$$

where $\phi(x) = \frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2/2}\,\mathrm{d}t$.

Since the pioneered examples of the eighties, cutoff was identified in a variety of contexts. One celebrated example is the so-called *riffle shuffle*, as modelled by the Gilbert-Shannon-Reeds transition matrix, and proved to have cutoff at time $\frac{3\log_2 n}{2}$ with a constant-order window by Bayer and Diaconis [19]. For more on card shuffling and cutoff for random walks on finite groups, we refer to see Diaconis [59], Chen and Saloff-Coste [47], and the survey by Saloff-Coste [144].

Establishing this phenomenon rigorously requires a very detailed understanding of the underlying chain, and often constitutes a challenging task even in situations with a high degree of symmetry. It is now believed to be a widespread phenomenon and many natural families of Markov chains are conjectured to exhibit cutoff. However there are very few general results about it. For instance, the historical case of random walks on the symmetric group $\mathfrak{S}_n$ is still far from being completely understood (see [144]). To cite one of the many open problems in this domain, it is an open question to determine whether every random walk on $\mathfrak{S}_n$ driven by the uniform distribution on a minimal generating set exhibits cut-off. In this direction, Berestycki and Sengul [27] recently established cutoff when the walk is induced by a given conjugacy class (that is, by the uniform distribution over the permutations with a cycle structure).

Interacting particle systems in statistical mechanics provide a rich class of dynamics displaying cutoff. One emblematic example is the *stochastic Ising model* at high enough temperature, for which the cutoff phenomenon has been established successively on the complete graph (Levin et al. [111]), on lattices (Ding et al. [62], Lubetzky and Sly [115]), and finally on any sequence of graphs (Lubetzky and Sly [116]). Other examples include the *Potts model* (Cuff et al. [53]), the *East process* (Ganguly et al. [77]), or the *Simple Exclusion process* on the cycle (Lacoin [108]).

The problem of singling out abstract conditions under which the cutoff phenomenon occurs, without necessarily pinpointing its precise location, has drawn considerable attention. In 2004, Peres [134] proposed a simple spectral criterion for reversible chains, known as the *product condition*:

$$t_{\text{MIX}}(1-\lambda_\star) \underset{n\to\infty}{\to} +\infty\,, \tag{1.3}$$

that is, the mixing time is much larger than the relaxation time. From the inequality

$$(t_{\text{REL}} - 1)\log\left(\frac{1}{2\varepsilon}\right) \le t_{\text{MIX}}(\varepsilon) \le \log\left(\frac{1}{\varepsilon\min_{x\in\Omega}\pi(x)}\right) t_{\text{REL}}\,, \tag{1.4}$$

valid for any reversible, irreducible and aperiodic Markov chain (see [110, Theorems 12.3 and 12.4]), one easily sees that, on such chains, condition (1.3) is necessary for cutoff. It can be shown to be sufficient for ℓ_2-cutoff, that is, when measuring the distance between $P^t(x,.)$ and π by

$$\left\| \frac{P^t(x,.)}{\pi(\cdot)} - 1 \right\|^2_{\ell_2(\pi)} \quad = \quad \sum_{y\in\Omega} \left(\frac{P^t(x,y)}{\pi(y)} - 1 \right)^2 \pi(y)\,.$$

However, for the total-variation distance, counter-examples of reversible chains satisfying (1.3) without cutoff have quickly been constructed (see Levin et al. [110, Chapter 18] and Chen and Saloff-Coste [47, Section 6]). Still, the product condition is widely believed to be sufficient for "most" chains. This has already been verified for birth-and-death chains (Ding et al. [63]) and, more generally, for random walks on trees (Basu et al. [16]). The latter result relies on a promising characterization of cutoff in terms of the concentration of hitting times of "worst" (in some sense) sets. See also Oliveira [128], Peres and Sousi [135], Griffiths et al. [85] and Hermon [91].

Now, an important family of chains is the family of random walks on *expander graphs* with degree d. Let $(G_n)_{n\geq 1}$ be a sequence of d-regular graphs with $|V(G_n)| \to \infty$ and let $\lambda_2^{(n)}$ be the second largest eigenvalue of the transition matrix of the simple random walk on G_n. One says that the family (G_n) is an expander family if there exists $\alpha > 0$, such that, for all $n \geq 1$,

$$1 - \lambda_2^{(n)} \quad \geq \quad \alpha\,.$$

Note that one can always make the walk lazy (that is, stay on the current state with probability 1/2, and move according to P with probability 1/2) to ensure that all the eigenvalues are positive and that $\frac{1-\lambda_2^{(n)}}{2}$ corresponds to the absolute spectral gap of the chain. Then, by the right-hand side of (1.4) and noticing that, as the graph is regular, the stationary distribution is uniform on V_n, one sees that the mixing time of the lazy random walk on expanders satisfies $t_{\text{MIX}} = O\left(\log |V_n|\right)$. On the other hand, on any d-regular graph with vertex set V, we have

$$t_{\text{MIX}}(\varepsilon) \quad \geq \quad \frac{\log\left(|V|(1-\varepsilon)\right)}{\log d}\,.$$

(This can be seen easily by noticing that the number of states on which the chain can be after t steps is smaller than d^t). Hence, expander graphs achieve the fastest mixing time (up to constant factors) among regular graphs with bounded degree. Moreover, according to the product condition, they should exhibit cutoff. However, up to 2010, no explicit example of expander graphs with cutoff was known. Then, Lubetzky and Sly [114] constructed both a 3-regular expander with cutoff and without cutoff. A conjecture of Peres (2004) is that, on every *transitive* expander with bounded degree, the simple random walk (SRW) has cutoff. Recently, Lubetzky and Peres [112] showed that cutoff occurs for NBRW and SRW on all Ramanujan graphs. Ramanujan graphs were introduced by Lubotzky et al. [117], and are optimal expanders in the spectral sense: any eigenvalue λ of the adjacency matrix is either $\pm d$ or satisfies $|\lambda| \leq 2\sqrt{d-1}$. In light of the Alon-Boppana Theorem [126], such graphs achieve the largest possible spectral gap and [112] confirmed their remarkable mixing properties.

1.3 Random walks on random graphs with given degrees

A decisive change of perspective was made in 2010, when Lubetzky and Sly [113] considered the SRW on a graph which, instead of being fixed, is itself random. A classical result of Pinsker [136] (see also [42]) states that random $d-$regular graphs with d fixed are *expanders* with high probability (*i.e.* with probability tending to 1 as $n \to \infty$). The celebrated result of [75] shows that, with high probability,

they are even optimal expanders, the second eigenvalue of their adjacency matrix being $2\sqrt{d-1}+o(1)$ (we say that they are weakly Ramanujan). In particular, the SRW on such graphs satisfies the product condition, and should therefore exhibit cutoff. This long-standing conjecture was confirmed only recently in the impressive work of Lubetzky and Sly [113], who also determined the precise cutoff window and typical profile of the distance inside the window. Their result relies on refined path counting arguments and is actually derived from the analysis of the non-backtracking random walk (NBRW) itself, via a clever transfer argument. Their main theorem is the following. Let Φ be the tail function of the standard normal: for all $\lambda \in \mathbb{R}$,

$$\Phi(\lambda) \quad = \quad \frac{1}{2\pi}\int_{\lambda}^{\infty} \mathrm{e}^{-\frac{u^2}{2}}\,\mathrm{d}u\,.$$

Theorem 1.1 (Lubetzky and Sly [113])**.** *Let $G \sim \mathcal{G}(n,d)$ be a random regular graph with degree $d \geq 3$ fixed. Then, for all $0<\varepsilon<1$,*

(i) the mixing time of the NBRW *satisfies*

$$t_{\mathrm{MIX}}^{(n)}(\varepsilon) = \log_{d-1}(dn) + O_{\mathbb{P}}(1).$$

(ii) the mixing time of the SRW *satisfies*

$$t_{\mathrm{MIX}}^{(n)}(\varepsilon) = \frac{d}{(d-2)}\log_{d-1} n + (\Lambda + o_{\mathbb{P}}(1))\,\Phi^{-1}(\varepsilon)\sqrt{\log_{d-1} n}\,,$$

where $\Lambda = \frac{2\sqrt{d(d-1)}}{(d-2)^{3/2}}$.

The intuition behind this result is as follows: as random regular graphs are locally tree-like, the probability that the NBRW started from x is at a given vertex y after t steps is about $(d-1)^{-t}$ if y is at distance t from x. This becomes close to $1/n$ when t is equal to $\log_{d-1} n$. In particular, the walk mixes when its graph-distance from the origin reaches $\log_{d-1} n$, which corresponds both to the diameter and the typical graph-distance between two vertices in $\mathcal{G}(n,d)$. For the NBRW which, on a tree, can only go forward, this happens in precisely $\log_{d-1} n$ steps, and the cutoff window is extremely sharp. Now, at time t, the height of the SRW on a d-regular tree is about $\frac{d-2}{d}t$. Hence, it reaches distance $\log_{d-1} n$ at time $\frac{d}{d-2}\log_{d-1} n$, with Gaussian fluctuations of order $\sqrt{\log n}$. The mixing time of the SRW is thus $d/(d-2)$ times larger than that of the NBRW. This confirms the practical advantage of NBRW over SRW for efficient network sampling and exploration, and complements a well-known spectral comparison for regular expanders due to Alon et al. [9], as well as a recent result by Cooper and Frieze [50] on the cover time of random regular graphs. For other ways of speeding up random walks, see Cooper [49].

A natural question is whether the NBRW and the SRW still have cutoff on random graphs which are not regular. Fix a graphic degree sequence $d_1,\ldots,d_n$ and let G be a random uniform graph on this degree sequence. Can we describe the mixing time of the random walk on G ? In [21], we focused on non-backtracking random walks in the more general setting of the configuration model (described below). We show that, for any degree sequence satisfying some sparsity conditions, the NBRW exhibits the cutoff phenomenon, and that the cutoff location and window only depends on the degrees through two very simple statistics: the empirical mean and variance of the logarithmic degrees [1]. We also show that the profile of the distance inside the window approaches a universal shape, namely, that of the Gaussian tail function. Section 2 is devoted to this result.

1. A weaker version of our result on the NBRW was independently proved by Berestycki et al. [28], namely that $t_{\mathrm{MIX}}(\varepsilon) = t_\star + O_{\mathbb{P}}(\sqrt{\log N})$ (under a more restrictive assumption on degrees). The main result of this paper is to establish the cutoff for the SRW on G and we will go into it more deeply in Section 3.

Most real-world networks, such as the web graph, are directed: there might be an edge from vertex u to vertex v but not from v to u. However, studying random walks on random directed graphs is very challenging: those chains are indeed non-reversible. Moreover, the equilibrium measure itself is hard to understand. The stationary measure of a vertex may not depend only on its in-degree, but also on its in-coming neighbourhood, and actually on the whole graph. In a remarkable work, Bordenave et al. [36] showed that random walks on random directed graphs with given (in- and out-) degrees exhibit the cutoff phenomenon with high probability. They give the precise location and window, and show that the rescaled distance inside the cutoff window converges to the Gaussian tail function. They also obtain a precise description of the stationary measure.

In many practical situations, the use of non-backtracking random walks seems natural. By cancelling the noise created by backtracks, non-backtracking random walks seem to be more tightly related to the structure of the graph itself. For instance, the analysis of the non-backtracking matrix turns out to be decisive in problems related to community detection in the Stochastic Block Model [37, 56, 120, 124].

Berestycki et al. [28] established the cutoff (starting from a typical vertex) for the SRW on random graphs with given degrees. A natural question is to determine whether, as in the regular case, the NBRW still mixes faster than the SRW. Indeed, as noted above, in the d-regular case, all the paths of length t have the same probability, and mixing occurs when the walk has reached a typical distance from its origin. In this case, the SRW is clearly slowed down, and the delay factor is precisely its speed, $(d-2)/d$. In the non-regular case, different paths can have very different weights, and mixing occurs when t is large enough to see paths with a "reasonable" weight. In addition to the speed effect, heterogeneous degrees create another opposite effect: the disadvantage of the NBRW would be to be "trapped" in some low-degree paths, whereas the backtracking possibility of the SRW would prevent it from this kind of pitfall. The SRW is in a way smarter: when it enters a low-degree part of the graph, it has relatively more chance to backtrack and go explore higher-degree parts of the graph, which are given more weight by the stationary distribution. It turns out that this second effect is not strong enough to compensate for the slowdown of the SRW, and that, even in the non-regular case, non-backtracking random walks mix faster. This confirms the practical advantage of non-backtracking random walks: not only do they discover the graph faster (at least on random d-regular graphs, Cooper and Frieze [50] showed that the *cover time* of NBRW is asymptotic to $n \log n$, which is a general lower bound for the cover time, whereas the one of SRW is larger by a factor $\frac{d-1}{d-2}$), but they also get faster to stationarity. This problem is addressed in Section 3.

2 Cutoff for non-backtracking random walks

Given a finite set V and a function $\deg\colon V \to \{2, 3, \ldots\}$ such that

$$N \quad := \quad \sum_{v \in V} \deg(v) \tag{2.1}$$

is even, we construct a graph G with vertex set V and degrees $(\deg(v))_{v \in V}$ as follows. We form a set $\mathcal{X}$ by "attaching" $\deg(v)$ *half-edges* to each vertex $v \in V$:

$$\mathcal{X} := \{(v, i)\colon v \in V, 1 \leq i \leq \deg(v)\}.$$

We then simply choose a *pairing* π on $\mathcal{X}$ (i.e., an involution without fixed points), and interpret every pair of matched half-edges $\{x, \pi(x)\}$ as an edge between the corresponding vertices. Loops and multiple edges are allowed.

The *non-backtracking random walk* (NBRW) on the graph $G = G(\pi)$ is a discrete-time Markov chain

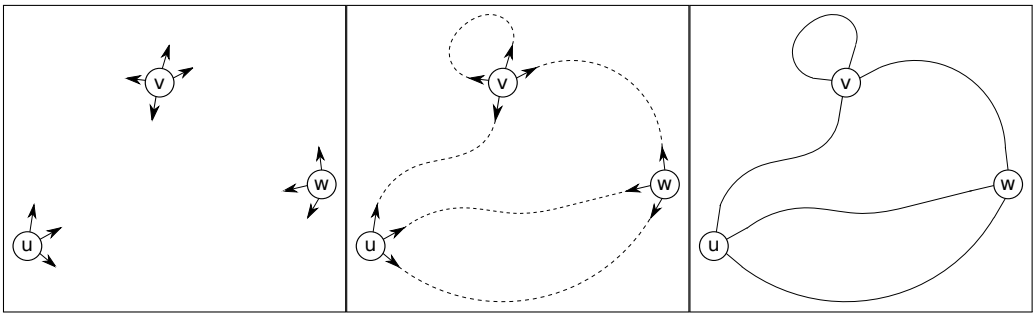


Figure 2.1 – A set of half-edges $\mathcal{X}$, a pairing π and the resulting graph G

with state space $\mathcal{X}$ and transition matrix

$$P(x,y) = \begin{cases} \frac{1}{\deg(\pi(x))} & \text{if y is a neighbour of } \pi(x) \\ 0 & \text{otherwise.} \end{cases}$$

In this definition and throughout the paper, two half-edges $x = (u,i)$ and $y = (v,j)$ are called *neighbours* if $u = v$ and $i \neq j$, and we let $\deg(x) := \deg(u) - 1$ denote the number of neighbours of the half-edge $x = (u,i)$. In words, the chain moves at every step from the current state x to a uniformly chosen neighbour of $\pi(x)$.

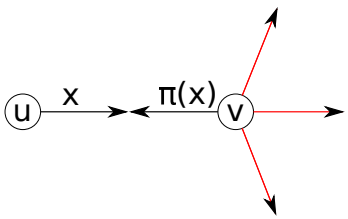


Figure 2.2 – The non-backtracking moves from x (in red)

Note that the matrix P is symmetric with respect to π: for all $x, y \in \mathcal{X}$,

$$P(\pi(y), \pi(x)) = P(x,y). \tag{2.2}$$

In particular, P is doubly stochastic: the uniform law on $\mathcal{X}$ is invariant for the chain. The *worst-case total-variation distance to equilibrium* at time $t \in \mathbb{N}$ is

$$\mathcal{D}(t) \quad := \quad \max_{x \in \mathcal{X}} \mathcal{D}_x(t), \qquad \text{where} \qquad \mathcal{D}_x(t) := \frac{1}{2} \sum_{y \in \mathcal{X}} \left| P^t(x,y) - \frac{1}{N} \right|. \tag{2.3}$$

This quantity is non-increasing in t, and the number of transitions that have to be made before it falls below a given threshold $0 < \varepsilon < 1$ is the *mixing time*:

$$t_{\text{MIX}}(\varepsilon) \quad := \quad \inf \{ t \in \mathbb{N} \colon \mathcal{D}(t) < \varepsilon \}.$$

2.1 Statement and comments

We are concerned with the typical profile of the function $t \mapsto \mathcal{D}(t)$ under the so-called *configuration model* (see e.g., [154]), i.e. when the pairing π is chosen uniformly at random among the $(N-1)!!$

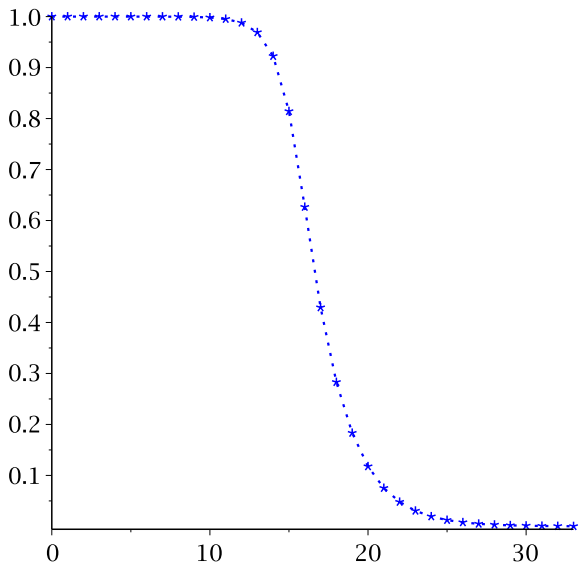


Figure 2.3 – Distance to stationarity along time for the NBRW on a random graph with 10^6 degree $3-$vertices and 10^6 degree $4-$vertices

possible pairings on $\mathcal{X}$. In order to study large-size asymptotics, we let the vertex set V and degree function $\deg\colon V \to \mathbb{N}$ depend on an implicit parameter $n \in \mathbb{N}$, which we omit from the notation for convenience. The same convention applies to all related quantities, such as N or $\mathcal{X}$. All asymptotic statements are understood as $n \to \infty$. Our interest is in the *sparse regime*, where the number N of half-edges diverges at a faster rate than the maximum degree. Specifically, we assume that

$$\Delta := \max_{v \in V} \deg(v) \quad = \quad N^{o(1)}, \tag{2.4}$$

As the behaviour of the NBRW at degree-2 vertices is deterministic, we assume that

$$\min_{v \in V} \deg(v) \quad \geq \quad 3. \tag{2.5}$$

Remarkably enough, the asymptotics in this regime depends on the degrees through two simple statistics: the mean logarithmic degree of an half-edge

$$\mu \quad := \quad \frac{1}{N} \sum_{v \in V} \deg(v) \log\left(\deg(v) - 1\right), \tag{2.6}$$

and the corresponding variance

$$\sigma^2 \quad := \quad \frac{1}{N} \sum_{v \in V} \deg(v) \left\{\log\left(\deg(v) - 1\right) - \mu\right\}^2. \tag{2.7}$$

We will also need some control on the third absolute moment:

$$\varrho \quad := \quad \frac{1}{N} \sum_{v \in V} \deg(v) \left|\log\left(\deg(v) - 1\right) - \mu\right|^3. \tag{2.8}$$

It might help the reader to think of μ, σ and ϱ as being fixed, or bounded away from 0 and ∞. However, we only impose the following (much weaker) condition:

$$\frac{\sigma^2}{\mu^3} >> \frac{(\log\log N)^2}{\log N} \quad \text{and} \quad \frac{\sigma^3}{\varrho\sqrt{\mu}} >> \frac{1}{\sqrt{\log N}}. \tag{2.9}$$

Our main result states that on most graphs with degrees $(\deg(v))_{v\in V}$, the NBRW exhibits a remarkable behaviour, visible on Figure 2.3 and known as a *cutoff*: the distance to equilibrium remains close to 1 for a rather long time, roughly

$$t_\star := \frac{\log N}{\mu}, \tag{2.10}$$

and then abruptly drops to nearly 0 over a much shorter time scale [2], of order

$$\omega_\star := \sqrt{\frac{\sigma^2 \log N}{\mu^3}}. \tag{2.11}$$

Moreover, the cutoff shape inside this window approaches a surprisingly simple function $\Phi\colon \mathbb{R} \to [0,1]$, namely the tail distribution of the standard normal:

$$\Phi(\lambda) := \frac{1}{2\pi}\int_\lambda^\infty e^{-\frac{u^2}{2}}\,\mathrm{d}u.$$

It is remarkable that this limit shape does not depend at all on the precise degrees.

Theorem 2.1 (Cutoff for the NBRW on sparse graphs)**.** *For every* $0 < \varepsilon < 1$*,*

$$\frac{t_{\text{MIX}}(\varepsilon) - t_\star}{\omega_\star} \xrightarrow{\mathbb{P}} \Phi^{-1}(\varepsilon).$$

Equivalently, for $t = t_\star + \lambda\omega_\star + o(w_\star)$ *with* $\lambda \in \mathbb{R}$ *fixed, we have* $\mathcal{D}(t) \xrightarrow{\mathbb{P}} \Phi(\lambda)$*.*

It is interesting to compare this with the $d-$regular case (i.e., $\deg\colon V \to \mathbb{N}$ constant equal to d) studied by Lubetzky and Sly [113]: by a remarkably precise path counting argument, they establish cutoff within constantly many steps around $t_\star = \log N/\log(d-1)$. To appreciate the effect of heterogeneous degrees, recall that μ and σ are the mean and variance of $\log Z$, where Z is the degree of a uniformly sampled half-edge. Now, by Jensen's Inequality,

$$t_\star \quad \geq \quad \frac{\log N}{\log \mathbb{E}[Z]}\,,$$

and the less concentrated Z, the larger the gap. The right-hand side is a well-known characteristic length in G, namely the typical inter-point distance (see e.g., [155]). One notable effect of heterogeneous degrees is thus that the mixing time becomes significantly larger than the natural graph distance. A heuristic explanation is as follows: in the regular case, all paths of length t between two points are equally likely for the NBRW, and mixing occurs as soon as t is large enough for many such paths to exist. In the non-regular case however, different paths have very different weights, and most of them actually have a negligible chance of being seen by the walk. Consequently, one has to make t larger in order to see paths with a "reasonable" weight. Even more remarkable is the impact of heterogeneous degrees on the cutoff width $\omega_\star$, which satisfies $\omega_\star >> \log\log N$ against $\omega_\star = \Theta(1)$ in the regular case. Finally, the gaussian

2. The fact that $\omega_\star << t_\star$ follows from condition (2.4).

limit shape Φ itself is specific to the non-regular case and is directly related to the fluctuations of degrees along a typical trajectory of the NBRW.

Remark 2.1 (Simple graphs)**.** *A classical result by Janson [95] asserts that the graph produced by the configuration model is simple (no loops or multiple edges) with probability asymptotically bounded away from* 0*, as long as*

$$\sum_{v\in V} \deg(v)^2 = O(N)\,. \tag{2.12}$$

Moreover, conditionally on being simple, it is uniformly distributed over all simple graphs with degrees $(\deg(v))_{v\in V}$*. Thus, every property which holds* **whp** *under the configuration model also holds* **whp** *for the uniform simple graph model. In particular, under (2.12), the conclusion of Theorem 2.1 extends to simple graphs.*

Remark 2.2 (IID degrees)**.** *A common setting consists in generating an infinite* IID *degree sequence* $(\deg(v))_{v\in\mathbb{N}}$ *from some fixed degree distribution* Q *and then restricting it to the index set* $V = \{1,\dots,n\}$ *for each* $n \geq 1$*. Let* D *denote a random integer with distribution* Q*. Assuming that*

$$\mathbb{P}\left(D \leq 2\right) = 0, \quad \mathrm{Var}(D) > 0, \text{ and } \quad \mathbb{E}\left[e^{\theta D}\right] < \infty \text{ for some } \theta > 0,$$

ensures that the conditions (2.4), (2.5) and (2.9) hold almost surely. Thus, Theorem 2.1 applies with the parameters μ, σ *and* N *now being random. But the latter clearly concentrate around their deterministic counterparts, in the following sense:*

$$\begin{aligned} N &= n\mathbb{E}[D] + O_{\mathbb{P}}\left(n^{\frac{1}{2}}\right) \\ \mu &= \mu_\star + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \quad \text{with } \mu_\star = \mathbb{E}[D\log(D-1)]/\mathbb{E}[D] \\ \sigma &= \sigma_\star + O_{\mathbb{P}}\left(n^{-\frac{1}{2}}\right) \quad \text{with } \sigma_\star^2 = \mathbb{E}\left[D\left\{\log(D-1) - \mu_\star\right\}^2\right]/\mathbb{E}[D]. \end{aligned}$$

Those error terms are small enough to allow one to substitute $n, \mu_\star, \sigma_\star$ *for* N, μ, σ *without affecting the convergence stated in Theorem 2.1.*

2.2 Proof outline

The proof of Theorem 2.1 is divided into two (unequal) halves: for

$$t = t_\star + \lambda w_\star + o(w_\star), \tag{2.13}$$

with $\lambda \in \mathbb{R}$ fixed, we show that

$$\min_{x\in\mathcal{X}} \mathbb{E}\left[\mathcal{D}_x\left(t\right)\right] \geq \Phi(\lambda) - o(1)\,, \tag{2.14}$$

$$\max_{x\in\mathcal{X}} \mathcal{D}_x\left(t\right) \leq \Phi(\lambda) + o_{\mathbb{P}}(1). \tag{2.15}$$

Note that this actually shows that the maximization over all possible states in (2.3) is irrelevant. The lower bound (2.14) is proved in Section 2.3.1. The difficult part is the upper bound (2.15), due to the worst-case maximization: our approximations for a given initial state $x \in \mathcal{X}$ need to be valid with probability $1 - o(1/N)$, so that we may then take union bound. Our starting point is the key identity

$$P^t(x, \pi(y)) = \sum_{(u,v)\in\mathcal{X}\times\mathcal{X}} P^{t/2}(x,u)P^{t/2}(y,v)\mathbb{1}_{\{\pi(u)=v\}}\,, \tag{2.16}$$

which follows from the symmetry (2.2). As a first approximation, let us assume that the balls of radius $t/2$ around x and y consist of disjoint trees, as in Figure 2.4.

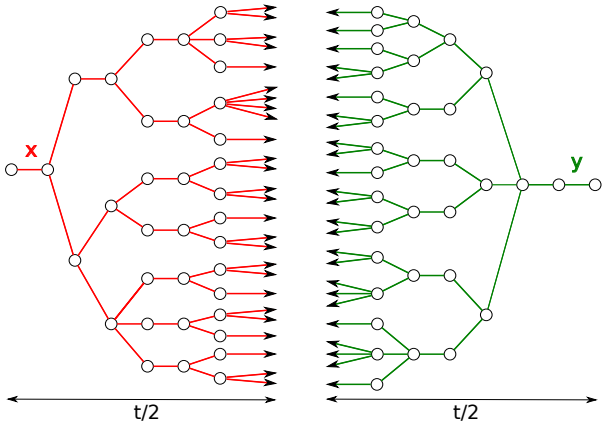


Figure 2.4 – The tree-approximation

This is made rigorous by a particular exposure process described in Section 2.3.2. Then the *weight* $\mathbf{w}(u) := P^{t/2}(x,u)$ (resp. $\mathbf{w}(v) := P^{t/2}(y,v)$) can be unambiguously written as the inverse product of degrees along the unique path from x to u (resp. y to v). A second approximation consists in eliminating those paths whose weight exceeds some threshold $\theta > 0$ (the correct choice turns out to be $\theta \approx \frac{1}{N}$):

$$P^t(x,\pi(y)) \quad \approx \quad \sum_{u,v} \mathbf{w}(u)\mathbf{w}(v)\mathbb{1}_{(\mathbf{w}(u)\mathbf{w}(v)\leq\theta)}\mathbb{1}_{(\pi(u)=v)}\,.$$

Conditionally on the two trees of height $t/2$, this is a weighted sum of weakly dependent Bernoulli variables, and the large-weight truncation should prevent it from deviating largely from its expectation. We make this argument rigorous in Lemma 2.5, using Stein's method of exchangeable pairs. Provided the exposure process did not reveal too many pairs of matched half-edges, the conditional expectation of $\mathbb{1}_{(\pi(u)=v)}$ remains close to $1/N$, and we obtain the new approximation

$$NP^t(x,\pi(y)) \quad \approx \quad \sum_{u,v} \mathbf{w}(u)\mathbf{w}(v)\mathbb{1}_{(\mathbf{w}(u)\mathbf{w}(v)\leq\theta)}\,.$$

Now, the right-hand side corresponds to the quenched probability that the product of weights seen by two independent NBRWs of length $t/2$, one starting from x and the other from y, does not exceed θ. The last step consists in approximating those trajectories by independent uniform samples $X_1^\star,\dots,X_t^\star$ from $\mathcal{X}$:

$$\begin{aligned}
\sum_{u,v} \mathbf{w}(u)\mathbf{w}(v)\mathbb{1}_{\mathbf{w}(u)\mathbf{w}(v)\leq\theta} \quad &\approx \quad \mathbb{P}\left[\frac{1}{\deg(X_1^\star)}\cdots\frac{1}{\deg(X_t^\star)} \leq \theta\right] \\
&\approx \quad \mathbb{P}\left[\frac{\sum_{k=1}^t(\mu - \log\deg(X_k^\star))}{\sigma\sqrt{t}} \leq \frac{\mu t + \log\theta}{\sigma\sqrt{t}}\right] \\
&\approx \quad 1-\Phi(\lambda)\,,
\end{aligned}$$

by the central limit theorem (recall that $\theta \approx 1/N$ and $t \approx t_\star + \lambda\omega_\star$). Consequently,

$$\mathcal{D}_x(t) = \sum_y \left(\frac{1}{N} - P^t(x,\pi(y))\right)_+ \quad \approx \quad \Phi(\lambda)\,,$$

as desired. This argument is made rigorous in Lemmas 2.6, 2.7 and 2.8.

2.3 Proof details

2.3.1 The lower bound

Fix $t \geq 1$, and a parameter $\theta \in (0,1)$. Choose two distinct states $x, y \in \mathcal{X}$ uniformly at random. Let $P^t_\theta(x,y)$ denote the contribution to $P^t(x,y)$ from paths having weight less than θ. Note that $P^t_\theta(x,y) < P^t(x,y)$ if and only if some path of length t from x to y has weight larger than θ, implying in particular that $P^t(x,y) > \theta$. Thus,

$$\frac{1}{N} - P^t_\theta(x,y) \quad \leq \quad \left(\frac{1}{N} - P^t(x,y)\right)_+ + \frac{1}{N}\mathbf{1}_{P^t(x,y)>\theta}.$$

Summing over all $y \in \mathcal{X}$ and observing that there can not be more than $1/\theta$ half-edges $y \in \mathcal{X}$ satisfying $P^t(x,y) > \theta$, we obtain

$$1 - \sum_{y \in \mathcal{X}} P^t_\theta(x,y) \quad \leq \quad \mathcal{D}_x(t) + \frac{1}{\theta N}.$$

Now, the left-hand side is the quenched probability (i.e., conditional on the pairing π) that a NBRW $\{X_k\}_{0 \leq k \leq t}$ on $G(\pi)$ starting at x satisfies $\prod_{k=1}^t \frac{1}{\deg(X_k)} > \theta$. Taking expectation w.r.t. the pairing, we arrive at

$$\mathbb{P}\left(\prod_{k=1}^t \frac{1}{\deg(X_k)} > \theta\right) \quad \leq \quad \mathbb{E}[\mathcal{D}_x(t)] + \frac{1}{\theta N}, \tag{2.17}$$

where the average is now taken over both the NBRW and the pairing (annealed law). A useful property of the uniform pairing is that it can be constructed sequentially, the pairs being revealed along the way, as we need them. We exploit this degree of freedom to generate the walk $\{X_k\}_{k \geq 0}$ and the pairing simultaneously, as follows. Initially, all half-edges are unpaired and $X_0 = x$; then at each time $k \geq 1$,

(i) if X_{k-1} is unpaired, we pair it with a uniformly chosen other unpaired half-edge; otherwise, $\pi(X_{k-1})$ is already defined and no new pair is formed.

(ii) in both cases, we let X_k be a uniformly chosen neighbour of $\pi(X_{k-1})$.

The sequence $\{X_k\}_{k \geq 0}$ is then exactly distributed according to the annealed law. Now, if we sample uniformly from $\mathcal{X}$ instead of restricting the random choice made at (i) to unpaired half-edges, then the uniform neighbour chosen at step (ii) also has the uniform law on $\mathcal{X}$. This creates a coupling between the process $\{X_k\}_{k \geq 1}$ and a sequence $\{X^\star_k\}_{k \geq 1}$ of IID samples from $\mathcal{X}$, valid until the first time T where the uniformly chosen half-edge or its uniformly chosen neighbour is already paired. As there are less than $2k$ paired half-edges by step k, a crude union-bound yields

$$\mathbb{P}(T \leq t) \leq \frac{2t^2}{N}.$$

Consequently,

$$\left|\mathbb{P}\left(\prod_{k=1}^t \frac{1}{\deg(X_k)} > \theta\right) - \mathbb{P}\left(\prod_{k=1}^t \frac{1}{\deg(X^\star_k)} > \theta\right)\right| \quad \leq \quad \frac{2t^2}{N}. \tag{2.18}$$

On the other hand, since $\{X_1^\star, \dots, X_t^\star\}$ are IID, Berry-Esseen's inequality implies

$$\left| \mathbb{P}\left(\prod_{k=1}^{t} \frac{1}{\deg(X_k^\star)} > \theta \right) - \Phi\left(\frac{\mu t + \log\theta}{\sigma\sqrt{t}} \right) \right| \quad \leq \quad \frac{\varrho}{\sigma^3\sqrt{t}}. \tag{2.19}$$

We may now combine (2.17), (2.18), (2.19) to obtain

$$\mathbb{E}[\mathcal{D}_x(t)] \quad \geq \quad \Phi\left(\frac{\mu t + \log\theta}{\sigma\sqrt{t}} \right) - \frac{1}{\theta N} - \frac{2t^2}{N} - \frac{\varrho}{\sigma^3\sqrt{t}}.$$

With t as in (2.13) and $\theta = (\log N)/N$, the right-hand side is $\Phi(\lambda) + o(1)$, thanks to our assumptions on μ, σ, ϱ. This establishes the lower bound (2.14).

2.3.2 The upper-bound

Following Lubetzky and Sly [113], we call $x \in \mathcal{X}$ a *root* (written $x \in \mathcal{R}$) if the (directed) ball of radius r centered at x (denoted by $\mathcal{B}_x$) is a tree, where

$$r \quad := \quad \left\lfloor \frac{\log N}{10 \log \Delta} \wedge \log\log N \right\rfloor. \tag{2.20}$$

Note that $1 << r << \omega_\star$ by assumptions (2.4) and (2.9). The first proposition below shows that we may restrict our attention to paths between roots. The second proposition provides a good control on such paths.

Proposition 2.3 (Roots are quickly reached)**.**

$$\max_{x \in \mathcal{X}} P^r(x, \mathcal{X} \setminus \mathcal{R}) \quad \xrightarrow{\mathbb{P}} \quad 0.$$

Proposition 2.4 (Roots are well inter-connected)**.** *For t as in (2.13),*

$$\min_{x \in \mathcal{R}} \min_{y \in \mathcal{R} \setminus \mathcal{B}_x} P^t(x, \pi(y)) \geq \frac{1 - \Phi(\lambda) - o_{\mathbb{P}}(1)}{N}.$$

Let us first see how those results imply the upper-bound (2.15). Observe that

$$\mathcal{D}(t+r) \quad \leq \quad \max_{x \in \mathcal{X}} P^r\left(x, \mathcal{X} \setminus \mathcal{R}\right) + \max_{x \in \mathcal{R}} \mathcal{D}_x(t)\,.$$

The first term is $o_{\mathbb{P}}(1)$ by Proposition 2.3. For the second one, we write

$$\mathcal{D}_x(t) \quad = \quad \sum_{y \in \mathcal{R} \setminus \mathcal{B}_x} \left(\frac{1}{N} - P^t(x, \pi(y)) \right)_+ + \sum_{y \in \mathcal{B}_x \cup (\mathcal{X} \setminus \mathcal{R})} \left(\frac{1}{N} - P^t(x, \pi(y)) \right)_+.$$

Proposition 2.4 ensures that the first sum is bounded by $\Phi(\lambda) + o_{\mathbb{P}}(1)$ uniformly in $x \in \mathcal{R}$. To see that the second sum is $o_{\mathbb{P}}(1)$ uniformly in $x \in \mathcal{R}$, it suffices to bound its summands by $1/N$ and observe that $|\mathcal{B}_x| \leq \Delta^r = o(N)$ by (2.20), while

$$|\mathcal{X} \setminus \mathcal{R}| \quad = \quad \sum_{x \in \mathcal{X}} P^r(x, \mathcal{X} \setminus \mathcal{R}) \quad = \quad o_{\mathbb{P}}(N).$$

(The first equality because P is doubly stochastic, the second by Proposition 2.3.)

Proof of Proposition 2.3. Define $R := \left\lfloor \frac{\log N}{5 \log \Delta} \right\rfloor$ and fix $x \in \mathcal{X}$. The ball of radius R around x can

be generated sequentially, its half-edges being paired one after the other with uniformly chosen other unpaired half-edges, until the whole ball has been paired. Observe that at most $k = \frac{\Delta\left((\Delta-1)^R-1\right)}{\Delta-2}$ pairs are formed. Moreover, for each of them, the number of unpaired half-edges having an already paired neighbour is at most $\Delta(\Delta-1)^R$ and hence the conditional chance of hitting such a half-edge (thereby creating a cycle) is at most $p = \frac{\Delta(\Delta-1)^R-1}{N-2k-1}$. Thus, the probability that more than one cycle is found is at most

$$(kp)^2 \quad = \quad O\left(\frac{(\Delta-1)^{4R}}{N^2}\right) \; = \; o\left(\frac{1}{N}\right).$$

Summing over all $x \in \mathcal{X}$ (union bound), we obtain that with high probability, no ball of radius R in $G(\pi)$ contains more than one cycle.

To conclude the proof, we now fix a pairing π with the above property, and we prove that the NBRW on $G(\pi)$ starting from any $x \in \mathcal{X}$ satisfies

$$\mathbb{P}\left(X_t \text{ is not a root}\right) \quad \leq \quad 2^{1-t}, \tag{2.21}$$

for all $t \leq R - r$. The claim is trivial if the ball of radius R around x is acyclic. Otherwise, it contains a single cycle $\mathcal{C}$, by assumption. Write $d(x, \mathcal{C})$ for the minimum length of a non-backtracking path from x to some $z \in \mathcal{C}$. The non-backtracking property ensures that if $d(X_t, \mathcal{C}) < d(X_{t+1}, \mathcal{C})$ for some $t < R - r$, then $X_{t+1}, X_{t+2}, \ldots, X_{R-r}$ are all roots. Indeed, as soon as the NBRW makes a step away from $\mathcal{C}$ on one of the disjoint trees rooted to $\mathcal{C}$, it can only go further away from it. By (2.5), the conditional chance that $d(X_{t+1}, \mathcal{C}) = d(X_t, \mathcal{C}) + 1$ given the past is at least $1/2$ (unless $d(X_t, \mathcal{C}) = 1$, which can only happen once). This shows (2.21). We then specialize to $t = r$. □

The remainder of the section is devoted to the proof of Proposition 2.4. By union bound, it is enough to fix two distinct half-edges $x, y \in \mathcal{X}$ and establish that, for every $\varepsilon > 0$,

$$\mathbb{P}\left(x \in \mathcal{R}, y \in \mathcal{R} \setminus \mathcal{B}_x, P^t(x, \pi(y)) \leq \frac{1 - \Phi(\lambda) - \varepsilon}{N}\right) \quad = \quad o\left(\frac{1}{N^2}\right). \tag{2.22}$$

To do so, we shall analyse a special procedure that generates a uniform pairing on $\mathcal{X}$ together with a two-tree forest $\mathfrak{F}$ keeping track of certain paths from x and from y. Initially, all half-edges are unpaired and $\mathfrak{F}$ is reduced to its two ancestors, x and y. We then iterate the following three steps:

1. An unpaired half-edge $z \in \mathfrak{F}$ is selected according to some rule (see below).
2. z is paired with a uniformly chosen other unpaired half-edge z'.
3. If neither z' nor any of its neighbours was already in $\mathfrak{F}$, then all neighbours of z' become children of z in the forest $\mathfrak{F}$.

The *exploration stage* stops when no unpaired half-edge is compatible with the selection rule. We then complete the pairing π by matching all the remaining unpaired half-edges uniformly at random: this is the *completion stage*.

The condition in step 3 ensures that $\mathfrak{F}$ remains a forest: any $z \in \mathfrak{F}$ determines a unique sequence $(z_0, \ldots, z_h)$ in $\mathfrak{F}$ such that $z_0 \in \{x, y\}$, z_i is a child of z_{i-1} for each $1 \leq i \leq h$, and $z_h = z$. We shall naturally refer to h and z_0 as the *height* and *ancestor* of z, respectively. We also define the *weight* of z as

$$\mathbf{w}(z) := \prod_{i=1}^{h} \frac{1}{\deg(z_i)}.$$

Note that this quantity is the quenched probability that the sequence $(z_0, \ldots, z_h)$ is realized by a NBRW

on G starting from z_0. In particular,

$$\mathbf{w}(z) \quad \leq \quad P^h(z_0, z). \tag{2.23}$$

Our rule for step 1 consists in selecting the smallest half-edge (according to the lexicographic order on $\mathcal{X}$) among all unpaired $z \in \mathfrak{F}$ with height $\mathbf{h}(z) < t/2$ and weight $\mathbf{w}(z) > w_{\text{MIN}}$, where

$$w_{\text{MIN}} \quad := \quad N^{-\frac{2}{3}}. \tag{2.24}$$

The role of this parameter is to limit the number of pairs formed at the exploration stage, see (2.28) below. As outlined in Section 2.2, we shall be interested in

$$\mathfrak{W} \quad := \sum_{(u,v)\in\mathcal{H}_x\times\mathcal{H}_y} \mathbf{w}(u)\mathbf{w}(v)\mathbf{1}_{\mathbf{w}(u)\mathbf{w}(v)\leq\theta},$$

where $\mathcal{H}_x$ (resp. $\mathcal{H}_y$) denotes the set of unpaired half-edges with height $\frac{t}{2}$ and ancestor x (resp. y) in $\mathfrak{F}$ at the end of the exploration stage, and where

$$\theta \quad := \quad \frac{1}{N(\log N)^2}. \tag{2.25}$$

Write $\overline{\mathfrak{W}}$ for the quantity obtained by replacing $\leq$ with $>$ in $\mathfrak{W}$, so that

$$\mathfrak{W} + \overline{\mathfrak{W}} \quad = \sum_{(u,v)\in\mathcal{H}_x\times\mathcal{H}_y} \mathbf{w}(u)\mathbf{w}(v) \quad \geq \sum_{z\in\mathcal{H}_x\cup\mathcal{H}_y} \mathbf{w}(z) - 1,$$

thanks to the inequality $ab \geq a + b - 1$ for $a, b \in [0, 1]$. Now, let $\mathfrak{U}$ denote the set of unpaired half-edges in $\mathfrak{F}$. By construction, at the end of the exploration stage, each $z \in \mathfrak{U}$ must have height $t/2$ or weight less than w_{MIN}, so that

$$\sum_{z\in\mathcal{H}_x\cup\mathcal{H}_y} \mathbf{w}(z) \quad \geq \quad \sum_{z\in\mathfrak{U}} \mathbf{w}(z) - \sum_{z\in\mathfrak{F}} \mathbf{w}(z)\mathbf{1}_{(\mathbf{w}(z)<w_{\text{MIN}})}.$$

Therefore, (2.22) is a consequence of the following four technical lemmas.

Lemma 2.5. *For every $\varepsilon > 0$,*

$$\mathbb{P}\left(P^t(x, \pi(y)) \leq \frac{\mathfrak{W} - \varepsilon}{N}\right) \quad = \quad o\left(\frac{1}{N^2}\right).$$

Lemma 2.6. *For every $\varepsilon > 0$,*

$$\mathbb{P}\left(\sum_{z\in\mathfrak{F}} \mathbf{w}(z)\mathbf{1}_{(\mathbf{w}(z)<w_{\text{MIN}})} > \varepsilon\right) \quad = \quad o\left(\frac{1}{N^2}\right).$$

Lemma 2.7. *For every $\varepsilon > 0$,*

$$\mathbb{P}\left(\overline{\mathfrak{W}} > \Phi(\lambda) + \varepsilon\right) \quad = \quad o\left(\frac{1}{N^2}\right).$$

Lemma 2.8. *For every $\varepsilon > 0$,*

$$\mathbb{P}\left(x \in \mathcal{R}, y \in \mathcal{R} \setminus \mathcal{B}_x, \sum_{z \in \mathfrak{U}} \mathbf{w}(z) < 2 - \varepsilon\right) \quad = \quad o\left(\frac{1}{N^2}\right).$$

Proof of Lemma 2.5. Combining the representation (2.16) with the observation (2.23) yields

$$P^t(x, \pi(y)) \quad \geq \quad \sum_{(u,v) \in \mathcal{H}_x \times \mathcal{H}_y} \mathbf{w}(u)\mathbf{w}(v)\mathbf{1}_{\mathbf{w}(u)\mathbf{w}(v) \leq \theta}\mathbf{1}_{\pi(u)=v}.$$

The right-hand side can be interpreted as the weight of the uniform pairing chosen at the completion stage, provided we define the weight of a pair (u, v) as

$$\mathbf{w}(u)\mathbf{w}(v)\mathbf{1}_{u \in \mathcal{H}_x}\mathbf{1}_{v \in \mathcal{H}_y}\mathbf{1}_{\mathbf{w}(u)\mathbf{w}(v) \leq \theta}. \tag{2.26}$$

With this interpretation, Lemma 2.5 becomes a special case of the following concentration inequality (which we apply conditionally on the exploration stage, with $\mathcal{I}$ being the set of half-edges that did not get paired, and weights given by (2.26)).

Lemma 2.9. *Let $\mathcal{I}$ be an even set, $\{w_{i,j}\}_{(i,j) \in \mathcal{I} \times \mathcal{I}}$ an array of non-negative weights, and π a uniform random pairing on $\mathcal{I}$. Then for all $a > 0$,*

$$\mathbb{P}\left(\sum_{i \in \mathcal{I}} w_{i,\pi(i)} \leq m - a\right) \quad \leq \quad \exp\left\{-\frac{a^2}{4\theta m}\right\},$$

where $m = \frac{1}{|\mathcal{I}|-1}\sum_{i \in \mathcal{I}}\sum_{j \neq i} w_{i,j}$ and $\theta = \max_{i \neq j}(w_{i,j} + w_{j,i})$.

Note that in our case, $m = \frac{\mathfrak{W}}{|\mathcal{I}|-1}$. Lemma 2.5 follows easily by taking $a = \frac{\varepsilon}{|\mathcal{I}|-1}$ and observing that $|\mathcal{I}| - 1 \leq N$ and $\mathfrak{W} \leq 1$. □

Proof of Lemma 2.9. We exploit the following concentration result for Stein pairs due to Chatterjee [46] (see also Ross [143, Theorem 7.4]): let Y, Y' be bounded variables satisfying

(i) $(Y, Y') \overset{d}{=} (Y', Y)$;

(ii) $\mathbb{E}[Y' - Y|Y] = -\lambda Y$;

(iii) $\mathbb{E}[(Y' - Y)^2|Y] \leq \lambda(bY + c)$,

for some constants $\lambda \in (0, 1)$ and $b, c \geq 0$. Then for all $a > 0$,

$$\mathbb{P}(Y \leq -a) \leq \exp\left\{-\frac{a^2}{c}\right\} \quad \text{and} \quad \mathbb{P}(Y \geq a) \leq \exp\left\{-\frac{a^2}{ab + c}\right\}.$$

We shall only use the first inequality. Consider the centered variable

$$Y := \sum_{i \in \mathcal{I}} w_{i,\pi(i)} - m,$$

and let Y' be the corresponding quantity for the pairing π' obtained from π by performing a random switch: two indices i, j are sampled uniformly at random from $\mathcal{I}$ without replacement, and the pairs $\{i, \pi(i)\}$, $\{j, \pi(j)\}$ are replaced with the pairs $\{i, j\}$, $\{\pi(i), \pi(j)\}$. This changes the weight by exactly

$$\Delta_{i,j} := w_{i,j} + w_{j,i} + w_{\pi(i),\pi(j)} + w_{\pi(j),\pi(i)} - w_{i,\pi(i)} - w_{\pi(i),i} - w_{j,\pi(j)} - w_{\pi(j),j}. \tag{2.27}$$

It is not hard to see that $(\pi, \pi') \stackrel{d}{=} (\pi', \pi)$, so that (i) holds. Moreover,

$$\begin{aligned}
\mathbb{E}[Y' - Y|\pi] &= \frac{1}{|\mathcal{I}|(|\mathcal{I}|-1)} \sum_{i\in\mathcal{I}} \sum_{j\neq i} \Delta_{i,j} \\
&= \frac{4}{|\mathcal{I}|(|\mathcal{I}|-1)} \sum_{i\in\mathcal{I}} \sum_{j\neq i} w_{i,j} - \frac{4}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} w_{i,\pi(i)} \\
&= -\frac{4}{|\mathcal{I}|} Y.
\end{aligned}$$

Regarding the square $\Delta_{i,j}^2 = |\Delta_{i,j}||\Delta_{i,j}|$, we may bound the first copy of $|\Delta_{i,j}|$ by 2θ and the second by changing all minus signs to plus signs in (2.27), yielding

$$\begin{aligned}
\mathbb{E}\left[(Y' - Y)^2|\pi\right] &= \frac{1}{|\mathcal{I}|(|\mathcal{I}|-1)} \sum_{i\in\mathcal{I}} \sum_{j\neq i} \Delta_{i,j}^2 \\
&\leq \frac{8\theta}{|\mathcal{I}|(|\mathcal{I}|-1)} \sum_{i\in\mathcal{I}} \sum_{j\neq i} w_{i,j} + \frac{8\theta}{|\mathcal{I}|} \sum_{i\in\mathcal{I}} w_{i,\pi(i)} \\
&= \frac{8\theta}{|\mathcal{I}|} (2m + Y).
\end{aligned}$$

Note that taking conditional expectation with respect to Y does not affect the right-hand side. Thus, (ii) and (iii) hold with $\lambda = \frac{4}{|\mathcal{I}|}$, $b = 2\theta$ and $c = 4m\theta$. □

Proof of Lemma 2.6. We may fix $z_0 \in \{x, y\}$ and restrict our attention to the halved sum

$$Z := \sum_{z\in\mathfrak{F}} \mathbf{w}(z) \mathbf{1}_{(\mathbf{w}(z) < w_{\text{MIN}})} \mathbf{1}_{(z \text{ has ancestor } z_0)}.$$

Consider $m = \lfloor \log N \rfloor$ independent NBRWs on $G(\pi)$ starting at z_0, each being killed as soon as its weight falls below w_{MIN}, and write A for the event that their trajectories form a tree of height less than $t/2$. Clearly, $\mathbb{P}(A|\pi) \geq Z^m$. Taking expectation and using Markov inequality, we deduce that

$$\mathbb{P}(Z > \varepsilon) \quad \leq \quad \frac{\mathbb{P}(A)}{\varepsilon^m},$$

where the average is now taken over both the walks and the pairing. Recalling that $m = \lceil \log N \rceil$, it is more than enough to establish that $\mathbb{P}(A) = (o(1))^m$. To do so, we generate the m killed NBRWs one after the other, revealing the underlying pairs along the way, as described in Section 2.3.1. Given that the first $\ell - 1$ walks form a tree of height less than $t/2$, the conditional chance that the ℓth walk also fulfils the requirement is $o(1)$, uniformly in $1 \leq \ell \leq m$. Indeed,

— either its weight falls below $\eta = (1/\log N)^2$ before it ever leaves the graph spanned by the first $\ell - 1$ trajectories and reaches an unpaired half-edge: thanks to the tree structure, there are at most $\ell - 1 < m$ possible trajectories to follow, each having weight at most η, so the chance is less than

$$m\eta = o(1).$$

— or the remainder of its trajectory after the first unpaired half-edge has weight less than $\Delta w_{\text{MIN}}/\eta$: this part consists of at most $t/2$ half-edges which can be coupled with uniform samples from $\mathcal{X}$

for a total-variation cost of mt^2/N, as in Section 2.3.1. Thus, the conditional chance is at most

$$\frac{mt^2}{N} + \mathbb{P}\left(\prod_{k=1}^{t/2} \deg(X_k^\star) \geq \frac{\eta}{\Delta w_{\text{MIN}}}\right) \quad = \quad o(1),$$

by Chebychev's inequality, since $\log\left(\frac{\eta}{\Delta w_{\text{MIN}}}\right) - \frac{\mu t}{2} >> \sigma\sqrt{\frac{t}{2}}$.

□

Proof of Lemma 2.7. Set $m = \lceil (\log N)^2 \rceil$. On $G(\pi)$, let $X^{(1)}, \ldots, X^{(m)}$ and $Y^{(1)}, \ldots, Y^{(m)}$ be $2m$ independent NBRWs of length $t/2$ starting at x and y respectively. Let B denote the event that their trajectories form two disjoint trees and that for all $1 \leq k \leq m$,

$$\prod_{\ell=1}^{t/2} \frac{1}{\deg(X_\ell^{(k)})} \prod_{\ell=1}^{t/2} \frac{1}{\deg(Y_\ell^{(k)})} > \theta.$$

Then clearly, $\mathbb{P}(B|\pi) \geq \overline{\mathfrak{W}}^m$. Averaging w.r.t. the pairing π, we see that

$$\mathbb{P}\left(\overline{\mathfrak{W}} > \Phi(\lambda) + \varepsilon\right) \quad \leq \quad \frac{\mathbb{P}(B)}{(\Phi(\lambda) + \varepsilon)^m}.$$

Thus, it is enough to establish that $\mathbb{P}(B) \leq (\Phi(\lambda) + o(1))^m$. We do so by generating the $2m$ walks $X^{(1)}, Y^{(1)}, \ldots, X^{(m)}, Y^{(m)}$ one after the other along with the underlying pairing, as above. Given that $X^{(1)}, Y^{(1)}, \ldots, X^{(\ell-1)}, Y^{(\ell-1)}$ already satisfy the desired property, the conditional chance that $X^{(\ell)}, Y^{(\ell)}$ also does is at most $\Phi(\lambda) + o(1)$, uniformly in $1 \leq \ell \leq m$. Indeed,

— either one of the two walks attains length $s = \lceil 4 \log\log N \rceil$ before leaving the graph spanned by the first $2(\ell - 1)$ trajectories and reaching an unpaired half-edge: thanks to the tree structure, there are at most $\ell - 1 < m$ possible trajectories to follow for each walk, each having weight at most 2^{-s} by (2.5), so the conditional chance is at most

$$2m2^{-s} = o(1).$$

— or at least $t - 2s$ unpaired half-edges are encountered, and the product of their degrees falls below $\frac{1}{\theta}$ with conditional probability at most

$$\frac{4mt^2}{N} + \mathbb{P}\left(\prod_{k=1}^{t-2s} \deg(X_k^\star) < \frac{1}{\theta}\right) \quad = \quad \Phi(\lambda) + o(1),$$

by the same coupling as above and Berry-Essen's inequality (2.19).

□

Proof of Lemma 2.8. Let τ denote the (random) number of pairs formed during the exploration stage. For $k \geq 0$, we let $\mathfrak{U}_k$ denote the set of unpaired half-edges in the forest after $k \wedge \tau$ pairs have been formed, and we consider the random variable

$$W_k := \sum_{z \in \mathfrak{U}_k} \mathbf{w}(z).$$

Initially $W_0 = 2$, and this quantity either stays constant or decreases at each stage, depending on whether the condition appearing in step 3 is satisfied or not. More precisely, denoting by z_k (resp. z_k') the half-edge

selected at step 1 (resp. chosen at step 2) of the kth pair, we have for all $k \geq 1$,

$$W_k = W_{k-1} - \mathbf{1}_{(k \leq \tau)} \left(\mathbf{w}(z_k) \mathbf{1}_{(z'_k \in \mathfrak{U}^+_{k-1})} + \mathbf{w}(z'_k) \mathbf{1}_{(z'_k \in \mathfrak{U}_{k-1})} \right),$$

where $\mathfrak{U}^+_{k-1}$ is $\mathfrak{U}_{k-1}$ together with the unpaired neighbours of x and y. Now, let $\{\mathcal{G}_k\}_{k \geq 0}$ be the natural filtration associated with the exploration stage. Note that τ is a stopping time, that $\mathbf{w}(z_k)$ is $\mathcal{G}_{k-1}-$measurable, and that the conditional law of z'_k given $\mathcal{G}_{k-1}$ is uniform on $\mathcal{X} \backslash \{z_1, \ldots, z_k, z'_1 \ldots, z'_{k-1}\}$. Thus,

$$\begin{aligned}
\mathbb{E}[W_k - W_{k-1} | \mathcal{G}_{k-1}] &= -\mathbf{1}_{(k \leq \tau)} \frac{\mathbf{w}(z_k)(|\mathfrak{U}^+_{k-1}| - 2) + W_{k-1}}{N - 2k + 1}. \\
\mathbb{E}\left[(W_k - W_{k-1})^2 | \mathcal{G}_{k-1}\right] &= \mathbf{1}_{(k \leq \tau)} \frac{\mathbf{w}(z_k)^2(|\mathfrak{U}^+_{k-1}| - 4) + 2\mathbf{w}(z_k) W_{k-1} + \sum_{z \in \mathfrak{U}_{k-1}} \mathbf{w}(z)^2}{N - 2k + 1}.
\end{aligned}$$

To bound those quantities, observe that each half-edge in $\mathfrak{U}_{k-1}$ has weight at least $\frac{\mathbf{w}(z_k)}{\Delta}$ because its parent has been selected at an earlier iteration and our selection rule ensures that the quantity $\mathbf{w}(z_k)$ is non-increasing with k. Consequently,

$$|\mathfrak{U}_{k-1}| \frac{\mathbf{w}(z_k)}{\Delta} \leq \sum_{z \in \mathfrak{U}_{k-1}} \mathbf{w}(z) \leq 2.$$

Combining this with the bound $|\mathfrak{U}^+_{k-1}| \leq |\mathfrak{U}_{k-1}| + 2\Delta$, we arrive at

$$\begin{aligned}
\mathbb{E}[W_k - W_{k-1} | \mathcal{G}_{k-1}] &\geq -\mathbf{1}_{(k \leq \tau)} \frac{4\Delta}{N - 2k + 1} \\
\mathbb{E}\left[(W_k - W_{k-1})^2 | \mathcal{G}_{k-1}\right] &\leq \mathbf{1}_{(k \leq \tau)} \frac{4\Delta \mathbf{w}(z_k) + 2}{N - 2k + 1}.
\end{aligned}$$

Now recall that $\mathbf{w}(z_k) \geq w_{\text{MIN}}$ and $\mathbf{h}(z_k) < \frac{t}{2}$ as per our selection rule, implying

$$w_{\text{MIN}} \tau \leq \sum_{k \geq 1} \mathbf{w}(z_k) \mathbf{1}_{(\tau \geq k)} \leq \sum_{z \in \mathfrak{F}} \mathbf{w}(z) \mathbf{1}_{\left(\mathbf{h}(z) < \frac{t}{2}\right)} \leq t. \tag{2.28}$$

The right-most inequality follows from the fact that the total weight at a given height in $\mathfrak{F}$ is at most 2 (the total weight being preserved from a parent to its children, if any). We conclude that

$$\begin{aligned}
\sum_{k=1}^{\tau} \mathbb{E}[W_k - W_{k-1} | \mathcal{G}_{k-1}] &\geq -\frac{4\Delta t}{w_{\text{MIN}} N - 2t} := -m \\
\sum_{k=1}^{\tau} \mathbb{E}\left[(W_k - W_{k-1})^2 | \mathcal{G}_{k-1}\right] &\leq \frac{4\Delta t w_{\text{MIN}} + 2t}{N w_{\text{MIN}} - 2t} := v.
\end{aligned}$$

Now, fix $\varepsilon > 0$ and consider the martingale $\{M_k\}_{k \geq 0}$ defined by $M_0 = 0$ and

$$M_k := \sum_{i=1}^{k} \left\{ (W_{i-1} - W_i) \wedge \varepsilon - \mathbb{E}\left[(W_{i-1} - W_i) \wedge \varepsilon \middle| \mathcal{G}_{i-1} \right] \right\}.$$

Then the increments of $\{M_k\}_{k \geq 0}$ are bounded by ε by construction, and the above computation guarantees

that almost-surely,

$$\sum_{k=1}^{\tau} \mathbb{E}\left[(M_k - M_{k-1})^2 \,|\mathcal{G}_{k-1}\right] \quad \leq \quad v = N^{-\frac{1}{3}+o(1)}.$$

Thus, the martingale version of Bennett's inequality due to Freedman [74] yields

$$\mathbb{P}\left(M_\tau > 7\varepsilon\right) \quad \leq \quad \left(\frac{ev}{v+7\varepsilon^2}\right)^7 = N^{-\frac{7}{3}+o(1)}. \tag{2.29}$$

But on the event $\{x \in \mathcal{R}, y \in \mathcal{R} \setminus \mathcal{B}_x\}$, all paths from the set $\{x, y\}$ to itself must have length at least r, and since $r \to \infty$, we must have asymptotically

$$\begin{aligned} \{x \in \mathcal{R}, y \in \mathcal{R} \setminus \mathcal{B}_x\} \quad &\subseteq \quad \left\{\max_k (W_{k-1} - W_k) \leq \varepsilon\right\} \\ &\subseteq \quad \{W_0 - W_\tau \leq M_\tau + m\} \end{aligned}$$

With (2.29), this proves Lemma 2.8 since $W_0 - W_\tau = 2 - \sum_{z \in \mathfrak{U}} \mathbf{w}(z)$ and $m = o(1)$. □

3 Comparing mixing times of simple and non-backtracking random walks

In this section, we are interested in comparing the mixing times of the NBRW and SRW on random graphs with given degrees. We first present the result of Berestycki et al. [28], which establishes cutoff for the SRW starting from a typical point. Then, we relate the mixing time of each walk to its entropy on a Galton-Watson tree whose offspring distribution is given by the law of $\deg(X)$, where X is a uniformly chosen half-edge. This representation allows us to show that, from any given starting point, with high probability, the NBRW mixes faster than the SRW.

3.1 The simple random walk

Let $G = (V, E)$ be a random uniform graph on a degree sequence $(\deg(v))_{v \in V}$. We let $n = |V|$ and $N = \sum_{v \in V} \deg(v)$. It will be convenient to define a random variable Z distributed as the out-degree of a uniformly chosen half-edge. Equivalently, for all $k \geq 1$,

$$\mathbb{P}\left(Z = k\right) \quad = \quad \frac{(k+1)|\{v \in V, \deg(v) = k+1\}|}{N}.$$

We assume that $Z \geq 2$ (that is, the degrees are larger than 3), and, as before, we let $\mu = \mathbb{E} \log Z$.

Loosely speaking, the random walk on G can be coupled with a walk on a rooted augmented Galton-Watson tree (T, ρ) with offspring distribution Z, *i.e.* a Galton-Watson tree where the root has stochastically one more child. As the quantities involved are much easier to describe in the Galton-Watson setting, this section will mostly consider SRW and NBRW on (T, ρ), even if we do not really explain why this coupling is relevant.

Consider a SRW (X_t) on T started at $X_0 = \rho$. As $Z \geq 2$, (X_t) is transient and its loop-erased trajectory defines a unique infinite ray ξ, whose distribution is called the *harmonic measure* of the walk: this is the random location at which the walk escapes to ∞. Let ν be the speed of (X_t):

$$\nu \quad \overset{\text{a.s.}}{=} \quad \lim_{t \to \infty} \frac{|X_t|}{t},$$

where, for $x \in T$, $|x|$ denotes the graph-distance between x and ρ. Thanks to Lyons et al. [119], we have

$$\nu \quad = \quad \mathbb{E}\left[\frac{Z-1}{Z+1}\right] .$$

If ∂T denotes the boundary of T, *i.e.* the set of infinite rays of T starting from ρ, one can endow ∂T with the following metric d: for all $\beta, \eta \in \partial T$, $d(\eta, \eta) = 0$, and, if $\beta \neq \eta$,

$$d(\beta, \eta) \quad = \quad \mathrm{e}^{-|\beta \wedge \eta|} ,$$

where $\beta \wedge \eta$ is the youngest common ancestor of β and η, *i.e.* the vertex common to β and η which is furthest from ρ. With this metric, the Hölder exponent of the harmonic measure at ξ, which is also its Hausdorff dimension, is

$$\mathbf{d} \quad \overset{\text{a.s.}}{:=} \quad \lim_{t \to \infty} -\frac{1}{t} \log \mathbb{P}(\xi_t \in \xi) . \tag{3.1}$$

With this notation, Berestycki et al. [28] showed that the SRW on G starting from a fixed vertex has cutoff at time $\frac{\log n}{\nu \mathbf{d}}$ with window $\sqrt{\log n}$.

Theorem 3.1 (Berestycki et al. [28])**.** *Assume that the variable Z satisfies*

$$\mathbb{E}Z \leq K, \quad \textit{and} \quad 2 \leq Z \leq \exp\left((\log n)^{1/2-\delta}\right) , \tag{3.2}$$

for some absolute constants $K, \delta > 0$. For $v \in V$, let $\mathcal{D}_v(t)$ denote the total-variation distance to equilibrium at time t for the SRW *started at v. Then, for all $\varepsilon > 0$, there exists $\gamma > 0$, such that, with high probability, the random walk started from a uniformly chosen vertex v satisfies*

$$\mathcal{D}_v\left(\frac{\log n}{\nu \mathbf{d}} - \gamma\sqrt{\log n}\right) \quad > \quad 1 - \varepsilon ,$$

and

$$\mathcal{D}_v\left(\frac{\log n}{\nu \mathbf{d}} + \gamma\sqrt{\log n}\right) \quad < \quad \varepsilon .$$

To give a (very) imprecise intuition of the mixing time location of SRW, one may interpret the definition of $\mathbf{d}$ at (3.1) as follows: the probability that the random walk is where it is, given that it is at distance k from the root, is approximately $\mathrm{e}^{-\mathbf{d}k}$. Now, the distance at time t is about νt. Hence, the probability that SRW is where it is at time t is approximately $\mathrm{e}^{-\mathbf{d}\nu t}$. For this probability to be close to $1/n$, one has to take $t \approx \frac{\log n}{\nu \mathbf{d}}$.

Theorem 3.1 implies in particular that for all $\varepsilon > 0$, there exists $\gamma > 0$ such that, with high probability

$$\max_{v \in V} \mathcal{D}_v\left(\frac{\log n}{\nu \mathbf{d}} - \gamma\sqrt{\log n}\right) \quad > \quad 1 - \varepsilon .$$

Hence, for all $\varepsilon > 0$, the mixing time $t_{\text{MIX}}^{\text{SRW}}(\varepsilon)$ (from the worst starting point) is larger than $(1+o_{\mathbb{P}}(1))\frac{\log n}{\nu \mathbf{d}}$.[3]

Combining Theorems 2.1 and 3.1, we have,

$$t_{\text{MIX}}^{\text{NBRW}}(\varepsilon) = (1 + o_{\mathbb{P}}(1))\frac{\log n}{\mu} \quad \text{and} \quad t_{\text{MIX}}^{\text{SRW}}(\varepsilon) \geq (1 + o_{\mathbb{P}}(1))\frac{\log n}{\nu \mathbf{d}} .$$

3. In an ongoing work with Eyal Lubetzky and Yuval Peres, we are trying to obtain the cutoff phenomenon for SRW from the worst starting point, which should be the same as from a typical point.

One may first observe that, as for the NBRW, allowing heterogeneous degrees also slows down the mixing of the SRW, compared to the regular case. Indeed, Lyons et al. [119] showed that, as soon as Z is not constant,

$$\mathbf{d} \quad < \quad \log \mathbb{E}Z \,. \tag{3.3}$$

This remarkable inequality is referred to as the *dimension drop* of the harmonic measure: $\log \mathbb{E}Z$ is the Hausdorff dimension of ∂T, the boundary of T. Hence inequality (3.3) means that, as $n \to \infty$, the harmonic measure at level n is supported on an exponentially small fraction of the vertices. In particular, it implies that

$$t_{\text{MIX}}^{\text{SRW}} \quad \geq \quad \frac{\log n}{\nu \log \mathbb{E}Z} \,.$$

The right-hand side corresponds to the time when the SRW reaches distance $\frac{\log n}{\log \mathbb{E}Z}$ from the origin. Hence, as in the non-backtracking case, mixing occurs later than the time it takes to the walk to reach the typical graph-distance from its starting point.

A natural question is to compare $t_{\text{MIX}}^{\text{SRW}}$ and $t_{\text{MIX}}^{\text{NBRW}}$. Does the NBRW mix faster? In the regular case, this was answered by Lubetzky and Sly [113]: the SRW is clearly slowed down, and the delay factor is precisely the inverse of the speed on a d-regular tree, $d/(d-2)$. In the non-regular case, the answer is much less clear. There is still the speed effect which gives an advantage to the NBRW. But now, another effect comes into play: when the NBRW enters a low-degree path, it is trapped in it, whereas, on such paths, the SRW has relatively more chance to backtrack and escape those low-degree paths in favour of higher-degree paths. The SRW would naturally go to high-degree vertices, which have a higher stationary distribution.

It turns out that the speed effect still prevails and that, on random graphs with given degrees satisfying (3.2), with high probability, the NBRW mixes faster (actually, even started from the worst vertex, NBRW mixes faster than SRW started from a typical point).

Theorem 3.2. *Assume $Z \geq 2$ and $\mathbb{E}[(\log Z)^2] < \infty$. We have*

$$\nu \mathbf{d} \quad < \quad \mu \,.$$

3.2 Entropies on Galton-Watson trees

The proof of Theorem 3.2 relies on the interpretation of μ (resp. $\nu\mathbf{d}$) as the limit entropy of the NBRW (resp. SRW) on Galton-Watson tree with offspring distribution Z.

As above, let (T, ρ) be a rooted augmented Galton-Watson tree with offspring distribution Z (except the root ρ which has offspring distribution $Z+1$). Let (X_t) and (Y_t) be respectively a SRW and a NBRW on T started at ρ. Conditionally on (T, ρ), let $\mathbf{H}(X_t) = \mathbf{H}^{(T,\rho)}(X_t)$ be the entropy of the SRW on T at time t, *i.e.*

$$\mathbf{H}(X_t) \quad = \quad \sum_{x \in T} \mathbb{P}_\rho[X_t = x] \log \frac{1}{\mathbb{P}_\rho[X_t = x]} \,,$$

and let $h_t^X = \mathbb{E}[\mathbf{H}(X_t)]$. Similarly, let $\mathbf{H}(Y_t) = \mathbf{H}^{(T,\rho)}(Y_t)$ be the entropy of the NBRW on T at time t, *i.e.*

$$\mathbf{H}(Y_t) \quad = \quad \sum_{x \in T} \mathbb{P}_\rho[Y_t = x] \log \frac{1}{\mathbb{P}_\rho[Y_t = x]} \,,$$

and let $h_t^Y = \mathbb{E}[\mathbf{H}(Y_t)]$.

One has

$$\mu \overset{\text{a.s.}}{=} \lim_{t\to\infty} \frac{\mathbf{H}(Y_t)}{t} \quad \text{and} \quad \nu\mathbf{d} \overset{\text{a.s.}}{=} \lim_{t\to\infty} \frac{\mathbf{H}(X_t)}{t}\,, \tag{3.4}$$

and

$$\mu = \lim_{t\to\infty} \frac{h_t^Y}{t} \quad \text{and} \quad \nu\mathbf{d} = \lim_{t\to\infty} \frac{h_t^X}{t}\,. \tag{3.5}$$

(see Lyons et al. [119, Theorem 9.7] for the identity for $\nu\mathbf{d}$). The fact that $h_t^Y/t \to \mu$ is easily obtained: denote by T_k the set of vertices of T at distance k from the root, and let $\mathcal{F}_k$ be the σ-field generated by the restriction of T up to level k. Also, if y is a child of x in T, we write $y \prec x$, and, if $x \in T$, we write Z_x the degree -1 of x (if $x \neq \rho$, then Z_x is the number of children of x and the number of children of the root ρ is $Z_\rho + 1$). We have, for all $k \geq 1$,

$$\begin{aligned}
\mathbf{H}(Y_{k+1}) &= \sum_{y\in T_{k+1}} \mathbb{P}_\rho(Y_{k+1} = y) \log \frac{1}{\mathbb{P}_\rho(Y_{k+1} = y)} \\
&= \sum_{x\in T_k} \sum_{y\prec x} \frac{\mathbb{P}_\rho(Y_k = x)}{Z_x} \log \frac{Z_x}{\mathbb{P}_\rho(Y_k = x)} \\
&= \mathbf{H}(Y_k) + \sum_{x\in T_k} \mathbb{P}_\rho(Y_k = x) \log Z_x\,.
\end{aligned}$$

Hence, as, for all $x \in T_k$, the variables $\mathbb{P}_\rho(Y_k = x)$ and $\mathbf{H}(Y_k)$ are $\mathcal{F}_k$-measurable, and as Z_x is independent of $\mathcal{F}_k$, we have

$$\mathbb{E}\left[\mathbf{H}(Y_{k+1}) \middle| \mathcal{F}_k\right] = \mathbf{H}(Y_k) + \mu\,.$$

Together with $\mathbb{E}[\mathbf{H}(Y_1)] = \mathbb{E}[\log(Z+1)]$, this yields

$$h_t^Y = \mathbb{E}[\log(Z+1)] + (t-1)\mu\,,$$

and $\frac{h_t^Y}{t} \to \mu$.

Remark 3.1. *The representation of μ and $\nu\mathbf{d}$ given in* (3.4) *entails an interesting interpretation of mixing times for random walks on random graphs: for both walks, mixing occurs when the entropy of the walk reaches* $\log n$*, which, under our assumptions on degrees, is asymptotic to the entropy of the stationary distribution. Let us try to give a very informal intuition. As the stationary entropy is asymptotic to* $\log n$*, let us, for simplicity, move to a setting where the stationary distribution is uniform. Consider an irreducible Markov chain on a finite state-space Ω of size n with transition matrix P and uniform stationary distribution. Then, if $H(P^t(x,.))$ denotes the entropy of law of the chain after t steps when started at x, the* entropy mixing time *can be defined, for all $0 < \varepsilon < 1$, as*

$$t_{\text{ENT}}(\varepsilon) = \inf\left\{t \geq 0,\, \forall x \in \Omega,\, \frac{H(P^t(x,.))}{\log n} \geq 1 - \varepsilon\right\}\,.$$

One has: $t_{\text{ENT}}(\varepsilon) \leq t_{\text{MIX}}(\varepsilon)$ [6]. Hence, entropy mixing is a lower bound for total-variation mixing, and the asymptotic in (3.4) *entails that, in the case of random walks on random graphs with given degrees, this lower bound is achieved.*

Now, the proof of Theorem 3.2 combines (3.5) and the following result due to Benjamini et al. [24]: the sequence $(h_t^X - h_{t-1}^X)_{t\geq 1}$ is non-increasing. This was first observed in the case of random walks on groups by Kaimanovich and Vershik [100] and the analysis of entropy of random walks on *random stationary environments* was pioneered by Kaimanovich [99]. Note that the environment (T, ρ, SRW) is

stationary, in the sense that the law of (T,ρ) is equal to the law of (T,X_1), where X_1 is the position of SRW started at ρ after one step (this is the reason why we consider an *augmented* Galton-Watson tree).

Lemma 3.2 (Benjamini et al. [24], Lemma 9)**.** *The sequence* $(h_t^X - h_{t-1}^X)_{t\geq 1}$ *is non-increasing.*

Proof of Lemma 3.2. Conditionally on (T,ρ), let $\mathbf{H}(X_1,X_t) = \mathbf{H}^{(T,\rho)}(X_1,X_t)$ be the entropy of (X_1,X_t), i.e.

$$\mathbf{H}(X_1,X_t) = \sum_{x,y\in T} \mathbb{P}_\rho(X_1 = x, X_t = y) \log \frac{1}{\mathbb{P}_\rho(X_1 = x, X_t = y)}\,,$$

and $h_{1,t}^X = \mathbb{E}[\mathbf{H}(X_1,X_t)]$. One has

$$\mathbf{H}(X_1,X_t) \quad = \quad \mathbf{H}(X_1) + \sum_{x\in T} \mathbb{P}_\rho(X_1 = x) \sum_{y\in T} \mathbb{P}_x(X_{t-1} = y) \log \frac{1}{\mathbb{P}_x(X_{t-1} = y)}\,.$$

Taking expectation, we obtain

$$h_{1,t}^X = h_1^X + \mathbb{E}[\mathbf{H}^{(T,X_1)}(X_{t-1})] = h_1^X + h_{t-1}^X\,,$$

where the last equality is due to the stationarity of the environment, implying that $\mathbf{H}^{(T,X_1)}(X_{t-1})$ has the same law as $\mathbf{H}^{(T,\rho)}(X_{t-1})$. Thus

$$h_t^X - h_{t-1}^X = h_t^X - h_{1,t}^X + h_1^X = \mathbb{E}[\mathbf{H}(X_t) - \mathbf{H}(X_1,X_t)] + h_1^X\,.$$

Now $\mathbf{H}(X_1,X_t) - \mathbf{H}(X_t)$ can be seen as the conditional entropy of X_1 given X_t, written $\mathbf{H}(X_1|X_t)$. We have

$$\mathbf{H}(X_1|X_t) = \mathbf{H}(X_1|X_t,X_{t+1}) \leq \mathbf{H}(X_1|X_{t+1})\,,$$

because, conditionally on X_t, the knowledge of X_{t+1} does not provide more information about X_1, and conditioning on more information can not increase the entropy. Hence $\mathbf{H}(X_t) - \mathbf{H}(X_1,X_t)$ is non-increasing, and so is its expectation. □

To sum up, we have $h_1^X = h_1^Y = \mathbb{E}[\log(Z+1)]$; for all $t\geq 2$, $h_t^Y - h_{t-1}^Y$ is constant equal to μ, and the sequence $(h_t^X - h_{t-1}^X)$ is non-increasing. Proving that $h_2^X - h_1^X < h_2^Y - h_1^Y$ would be enough but it does not hold for $Z=2$. We will show that, as soon as $Z\geq 2$ and $\mathbb{E}[(\log Z)^2] < \infty$, we have $h_3^X - h_1^X < h_3^Y - h_1^Y$. Combined with Lemma 3.2, this will establish that, for all $t\geq 2$,

$$h_t^X - h_2^X \leq \left\lceil \frac{t-2}{2} \right\rceil (h_3^X - h_1^X) < 2\left\lceil \frac{t-2}{2} \right\rceil \mu\,,$$

and that $\lim_{t\to\infty} \frac{h_t^X}{t} < \mu$, proving Theorem 3.2.

Proof of $h_3^X - h_1^X < h_3^Y - h_1^Y$. First, let us notice that Theorem 3.2 is true when Z is constant (in this case, $\mathbf{d} = \nu = \log(d-1)$ and $\nu = \frac{d-2}{d} < 1$). We may thus assume that $\mathrm{Var}Z > 0$. Let us consider the entropy of SRW after three steps in the tree T:

$$\mathbf{H}(X_3) = \sum_{z\in T} \mathbb{P}_\rho(X_3 = z) \log\left(\frac{1}{\mathbb{P}_\rho(X_3 = z)}\right)\,.$$

Recall that $T_k = \{z\in T : |z| = k\}$. As the SRW at time 3 can be either at distance 3, or at distance 1

from the root, we have $\mathbf{H}(X_3) = A + B$, with

$$A = \sum_{z \in T_3} \mathbb{P}_\rho(X_3 = z) \log\left(\frac{1}{\mathbb{P}_\rho(X_3 = z)}\right),$$

and

$$B = \sum_{z \in T_1} \mathbb{P}_\rho(X_3 = z) \log\left(\frac{1}{\mathbb{P}_\rho(X_3 = z)}\right).$$

Note that

$$\begin{aligned} A &= \sum_{x \prec \rho} \sum_{y \prec x} \sum_{z \prec y} \frac{1}{(Z_\rho + 1)} \frac{1}{(Z_x + 1)} \frac{1}{(Z_y + 1)} \log\left((Z_\rho + 1)(Z_x + 1)(Z_y + 1)\right) \\ &= \sum_{x \prec \rho} \sum_{y \prec x} Z_y \frac{\log(Z_\rho + 1) + \log(Z_x + 1) + \log(Z_y + 1)}{(Z_\rho + 1)(Z_x + 1)(Z_y + 1)}. \end{aligned}$$

Averaging over T_3, we have

$$\begin{aligned} \mathbb{E}[A|T_1, T_2] &= \frac{1}{Z_\rho + 1} \sum_{y \prec x \prec \rho} \frac{1}{Z_x + 1} \left(\mathbb{E}\left[\frac{Z}{Z+1}\right] \big(\log(Z_\rho + 1) + \log(Z_x + 1)\big) + \mathbb{E}\left[\frac{Z \log(Z+1)}{Z+1}\right] \right) \\ &= \frac{1}{Z_\rho + 1} \sum_{x \prec \rho} \frac{Z_x}{Z_x + 1} \left(\mathbb{E}\left[\frac{Z}{Z+1}\right] \big(\log(Z_\rho + 1) + \log(Z_x + 1)\big) + \mathbb{E}\left[\frac{Z \log(Z+1)}{Z+1}\right] \right). \end{aligned}$$

Continuing in the same manner, one easily gets

$$\mathbb{E}[A] = 2\mathbb{E}\left[\frac{Z}{Z+1}\right] \mathbb{E}\left[\frac{Z}{Z+1} \log(Z+1)\right] + \mathbb{E}\left[\frac{Z}{Z+1}\right]^2 \mathbb{E}\left[\log(Z+1)\right]. \tag{3.6}$$

Turning our attention to B, using convexity of $x \mapsto x \log x$ and Jensen's Inequality for conditional expectation, we obtain

$$\mathbb{E}[B|T_1] \leq -\sum_{z \in T_1} \mathbb{E}\left[\mathbb{P}_\rho(X_3 = z)|T_1\right] \log \mathbb{E}\left[\mathbb{P}_\rho(X_3 = z)|T_1\right].$$

Now, for $z \in T_1$, accounting for whether $X_2 = \rho$ or $X_2 = y$ for some $y \prec z$, one has

$$\mathbb{P}_\rho(X_3 = z) = \sum_{x \prec \rho} \frac{1}{(Z_\rho + 1)^2} \frac{1}{Z_x + 1} + \sum_{y \prec z} \frac{1}{(Z_\rho + 1)} \frac{1}{(Z_z + 1)} \frac{1}{(Z_y + 1)}.$$

Hence

$$\begin{aligned} \mathbb{E}\left[\mathbb{P}_\rho(X_3 = z)|T_1\right] &= \frac{1}{Z_\rho + 1} \mathbb{E}\left[\frac{1}{Z+1}\right] + \frac{1}{Z_\rho + 1} \mathbb{E}\left[\frac{1}{Z+1}\right] \mathbb{E}\left[\frac{Z}{Z+1}\right] \\ &= \frac{1}{Z_\rho + 1} \mathbb{E}\left[\frac{1}{Z+1}\right] \mathbb{E}\left[\frac{2Z+1}{Z+1}\right] \\ &= \frac{1}{Z_\rho + 1} \left(1 - \mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right). \end{aligned}$$

Taking average over T_1, this yields

$$\mathbb{E}[B] \leq \left(1 - \mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right) \left(\mathbb{E}\left[\log(Z+1)\right] - \log\left(1 - \mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right)\right). \tag{3.7}$$

Observe that $\mathbb{E}\left[\frac{Z}{Z+1}\right]^2 < \mathbb{E}\left[\left(\frac{Z}{Z+1}\right)^2\right]$ (because Z is not constant). Combining this observation with an other application of Jensen's inequality to $x \mapsto \log x$ yields

$$\begin{aligned}
-\log\left(1-\mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right) &< -\log\mathbb{E}\left[1-\left(\frac{Z}{Z+1}\right)^2\right] \\
&< -\mathbb{E}\left[\log\left(1-\left(\frac{Z}{Z+1}\right)^2\right)\right] \\
&= \mathbb{E}\left[\log\left(\frac{(Z+1)^2}{2Z+1}\right)\right],
\end{aligned}$$

Plugging this into (3.7), we obtain

$$\begin{aligned}
\mathbb{E}[B] &< \left(1-\mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right)\mathbb{E}\left[\log(Z+1)\right]+\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[\frac{2Z+1}{Z+1}\right]\mathbb{E}\left[\log\left(\frac{(Z+1)^2}{2Z+1}\right)\right] \\
&< \left(1-\mathbb{E}\left[\frac{Z}{Z+1}\right]^2\right)\mathbb{E}\left[\log(Z+1)\right]+\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[\frac{2Z+1}{Z+1}\log\left(\frac{(Z+1)^2}{2Z+1}\right)\right],
\end{aligned}$$

where the second inequality comes from the fact that $\mathrm{Cov}\left(\frac{2Z+1}{Z+1}, \log\left(\frac{(Z+1)^2}{2Z+1}\right)\right) > 0$ (the two functions are increasing, and, under our assumptions $Z \geq 2$ and $\mathbb{E}(\log Z)^2 < \infty$, both have finite second moment). Hence, recalling the formula for term A at (3.6) and that $h_1^X = \mathbb{E}[\log(Z+1)]$,

$$\begin{aligned}
h_3^X - h_1^X &< 2\mathbb{E}\left[\frac{Z}{Z+1}\right]\mathbb{E}\left[\frac{Z}{Z+1}\log(Z+1)\right]+\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[\frac{2Z+1}{Z+1}\log\left(\frac{(Z+1)^2}{2Z+1}\right)\right] \\
&= 2\mathbb{E}\left[\frac{Z}{Z+1}\log(Z+1)\right]+\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[\frac{2Z+1}{Z+1}\log\left(\frac{(Z+1)^2}{2Z+1}\right)-\frac{2Z}{Z+1}\log(Z+1)\right] \\
&:= 2\mathbb{E}\left[\frac{Z}{Z+1}\log(Z+1)\right]+\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[g(Z)\right].
\end{aligned}$$

It is easy to verify that $g'(z) = -\frac{\log(2z+1)}{(z+1)^2} < 0$ for all $z > 0$. The function g is therefore decreasing on $\mathbb{R}_+$. Hence, as $\mathbb{E}[g(Z)^2] < \infty$ thanks to $Z \geq 2$ and $\mathbb{E}(\log Z)^2 < \infty$, we obtain that $\mathbb{E}\left[\frac{1}{Z+1}\right]\mathbb{E}\left[g(Z)\right] < \mathbb{E}\left[\frac{g(Z)}{Z+1}\right]$. We finally get

$$h_3^X - h_1^X < \mathbb{E}\left[\frac{2Z}{Z+1}\log(Z+1)+\frac{g(Z)}{Z+1}\right],$$

and conclude by noticing that, for all $x \geq 2$, $\frac{2x}{x+1}\log(x+1)+\frac{g(x)}{x+1} < 2\log(x)$. □

Part III

Weighted sampling without replacement

This part is devoted to the problem of sampling without replacement in a finite population, when the items are allowed to have different weights. This is a work in collaboration with Justin Salez and Yuval Peres, Weighted sampling without replacement *[23].*

1 Sampling with and without replacement

1.1 The uniform case

The analysis of the concentration properties of sampling without replacement has a long history which can be traced back to the pioneer work of Hoeffding [92].

Consider a population of N items. Each item is equipped with a *value* $\nu(i) \in \mathbb{R}$, and a weight $\omega(i) > 0$ such that

$$\sum_{i=1}^{N} \omega(i) \quad = \quad 1\,.$$

We are interested in the case where different items can have different weights, but let us first consider the uniform case where

$$\omega(1) = \cdots = \omega(N) = \frac{1}{N}\,.$$

For $n \leq N$, let $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ be a sample drawn without replacement, *i.e.*, for each n-tuple $(i_1, \ldots, i_n)$ of distinct indices in $\{1, \ldots, N\}$,

$$\mathbb{P}\left((\mathbf{I}_1, \ldots, \mathbf{I}_n) = (i_1, \ldots, i_n)\right) \quad = \quad \frac{(N-n)!}{N!}\,,$$

and let $(\mathbf{J}_1, \ldots, \mathbf{J}_n)$ be drawn with replacement, *i.e.*, for each $(j_1, \ldots, j_N) \in \{1, \ldots, N\}^n$,

$$\mathbb{P}\left((\mathbf{J}_1, \ldots, \mathbf{J}_n) = (j_1, \ldots, j_n)\right) \quad = \quad \frac{1}{N^n}\,.$$

Now let

$$X \quad = \quad \nu(\mathbf{I}_1) + \cdots + \nu(\mathbf{I}_n)\,,$$

and

$$Y \quad = \quad \nu(\mathbf{J}_1) + \cdots + \nu(\mathbf{J}_n)\,.$$

Hoeffding [92] showed that, for all continuous convex functions $f\colon \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\left[f(X)\right] \quad \leq \quad \mathbb{E}\left[f(Y)\right]\,. \tag{1.1}$$

One says that X is less than Y *in the convex order*. In particular, the Laplace transform of X is upper-bounded by the Laplace transform of Y, and all the Chernoff-type tail estimates known for Y as a sum of I.I.D. random variables directly transfer to Y. This includes the celebrated Hoeffding and Bernstein inequalities, see the book [40]. Note that, as, in the uniform case, $\mathbb{E}X = \mathbb{E}Y$, inequality (1.1) implies proper concentration results for X around its mean. For instance, thanks to Hoeffding's inequality, one has, for all $t > 0$,

$$\mathbb{P}\left(X - \mathbb{E}X > t\right) \quad \leq \quad \exp\left(-\frac{2t^2}{n\Delta^2}\right)\,,$$

where $\Delta = \max_{1\le i\le N} \nu(i) - \min_{1\le i\le N} \nu(i)$. If $\Delta = O(1)$, this yields a variance factor of order n. However, one may notice that the variance of X can be much smaller when the sample size approaches the total number of items. In the extreme case where $n = N$, the variable X is deterministic and thus its variance is zero. More precisely, if σ^2 denotes the variance of $\nu(\mathbf{J}_1)$, then

$$\mathrm{Var}Y \quad = \quad n\sigma^2 \,,$$

whereas

$$\mathrm{Var}X \quad = \quad n\left(1-\frac{n}{N}\right)\sigma^2 \,.$$

This sharpening effect in the variance of X was incorporated in the following remarkable concentration inequality due to Serfling [145]:

$$\mathbb{P}\left(X - \mathbb{E}X > t\right) \quad \le \quad \exp\left(-\frac{2t^2}{n\left(1-\frac{n-1}{N}\right)\Delta^2}\right) \,.$$

Bernstein-type concentration inequalities for X were also obtained by Bardenet and Maillard [14].

Another remarkable feature of uniform sampling without replacement is the *negative association* of the sequence $(\nu(\mathbf{I}_1), \ldots, \nu(\mathbf{I}_n))$, which was established by Joag-Dev and Proschan [97], and also implies that the Laplace transform of X is upper-bounded by that of Y.

A natural question is to determine whether similar comparisons between sampling with and without replacement also hold when the sampling procedure is no longer uniform but when different items have different weights.

1.2 Main results

In the more general setting where the weights are heterogeneous, we have, for each $n-$tuple $(i_1, \ldots, i_n)$ of distinct indices in $\{1, \ldots, N\}$,

$$\mathbb{P}\left((\mathbf{I}_1, \ldots, \mathbf{I}_n) = (i_1, \ldots, i_n)\right) \quad = \quad \prod_{k=1}^{n} \frac{\omega(i_k)}{1-\omega(i_1)-\cdots-\omega(i_{k-1})} \,,$$

and, for each $n-$tuple $(j_1, \ldots, j_n) \in \{1, \ldots, N\}^n$,

$$\mathbb{P}\left((\mathbf{J}_1, \ldots, \mathbf{J}_n) = (j_1, \ldots, j_n)\right) \quad = \quad \prod_{k=1}^{n} \omega(j_k) \,.$$

As above, we define

$$X \quad = \quad \nu(\mathbf{I}_1) + \cdots + \nu(\mathbf{I}_n) \,,$$

and

$$Y \quad = \quad \nu(\mathbf{J}_1) + \cdots + \nu(\mathbf{J}_n) \,.$$

Weighted sampling without replacement, also known as *successive sampling*, appears in a variety of contexts (see [84, 93, 142, 157]). When $n << N$, it is natural to expect Y to be a good approximation of X. For instance, the total-variation distance between $\mathbb{P}\left(\mathbf{I}_{n+1} \in \cdot \,\middle|\, (\mathbf{I}_k)_{k=1}^n\right)$ and $\mathbb{P}\left(\mathbf{J}_1 \in \cdot\,\right)$ is given by

$\sum_{k=1}^{n} \omega(\mathbf{I}_k)$, which is $O(n/N)$ provided all the weights are $O(1/N)$.

One particular case is when weights and values are arranged in the same order, *i.e.*

$$\omega(i) > \omega(j) \quad \Longrightarrow \quad \nu(i) \geq \nu(j)\,. \tag{1.2}$$

Under condition (1.2), we have the following stochastic order relation.

Theorem 1.1. *Assume that condition* (1.2) *holds. Then X is less than Y in the* increasing convex order*, i.e. for every non-decreasing, convex function $f\colon \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\left[f\left(X\right)\right] \quad \leq \quad \mathbb{E}\left[f\left(Y\right)\right]\,. \tag{1.3}$$

Under the monotonicity assumption (1.2), Theorem 1.1 establishes an exact strong stochastic ordering between X and Y. In particular, for all $\lambda \geq 0$,

$$\mathbb{E}\,\mathrm{e}^{\lambda X} \quad \leq \mathbb{E}\,\mathrm{e}^{\lambda Y} \;\leq\; \mathbb{E}\left[\mathrm{e}^{\lambda \nu(\mathbf{J}_1)}\right]^n\,.$$

The Chernoff's bound

$$\mathbb{P}\left(Y \geq a\right) \quad \leq \quad \exp\left(n \log \mathbb{E}\left[\mathrm{e}^{\lambda \nu(\mathbf{J}_1)}\right] - \theta a\right)\,, \tag{1.4}$$

yields a variety of sharp concentration results based on efficient controls on the log-Laplace transform, and Theorem 1.1 implies in particular that all upper-tail estimates derived from Chernoff's bound (1.4) apply to X without modification.

The condition (1.2) describes a sampling procedure which is sometimes referred to as size-biased sampling without replacement. It arises in many situations, including ecology, oil discovery models, in the construction of the Poisson-Dirichlet distribution [137, 138], or in the configuration model of random graphs [32, 33]. In the next example, we show how Theorem 1.1 can be used to obtain upper-tail bounds on the number of edges revealed when we progressively expose the neighbourhood of a vertex in the configuration model.

Example 1. Consider a random graph G on a vertex set V and with degree sequence $(\deg(v))_{v\in V}$, generated according to the configuration model, as described in Section 2 of Part II. As before, we denote by $\mathcal{X}$ the set of half-edges. Consider the process of progressively revealing the neighbourhood of a randomly chosen vertex as follows: initially all half-edges are unpaired. Choose uniformly at random a half-edge z_1 in $\mathcal{X}$ and let $\mathfrak{E}_1$ be the set formed by z_1 and its neighbours (as in Part II, we say that two half-edges are neighbours if they are distinct and share the same vertex). Then, at step $k \geq 2$, choose an unpaired half-edge z_k of $\mathfrak{E}_{k-1}$ (according to some arbitrary rule), and pair it to an other uniformly chosen unpaired half-edge z'_k. Let $\mathfrak{E}_k$ be the union of $\mathfrak{E}_{k-1}$ and z'_k and the neighbours of z'_k (note that it might be the case that $\mathfrak{E}_k = \mathfrak{E}_{k-1}$). If at some time k, there is no unpaired half-edge satisfying the rule in $\mathfrak{E}_k$, then we let $\mathfrak{E}_\ell = \mathfrak{E}_k$ for all $\ell \geq k$. Let $\mathbf{V}$ be distributed as a size-biased pick in the set V, *i.e.*

$$\mathbb{P}\left(\mathbf{V} = v\right) \quad = \quad \frac{\deg(v)}{\sum_{u\in V} \deg(u)}\,.$$

Then, for all $1 \leq k \leq |V|$ and $\varepsilon > 0$,

$$\mathbb{P}\left(|\mathfrak{E}_k| > k\mathbb{E}[\deg(\mathbf{V})] + \varepsilon\right) \quad \leq \quad \exp\left(-\frac{2\varepsilon^2}{k\Delta^2}\right)\,,$$

where $\Delta = \max_{v \in V} \deg(v)$. To see this, note that the variable $|\mathfrak{E}_k|$ is stochastically dominated by

$$X \quad = \quad \sum_{i=1}^{k} \deg(\widetilde{V}_i)\,,$$

where $(\widetilde{V}_1, \ldots, \widetilde{V}_k)$ is a *size-biased sample without replacement* in V, *i.e.* for all k-tuple $(v_1, \ldots, v_k)$ of *distinct* vertices in V,

$$\mathbb{P}\Big((\widetilde{V}_1, \ldots, \widetilde{V}_k) = (v_1, \ldots, v_k)\Big) \quad = \quad \prod_{i=1}^{k} \frac{\deg(v_i)}{N - \deg(v_1) - \cdots - \deg(v_{i-1})}\,.$$

Indeed, one may couple $|\mathfrak{E}_k|$ and X in such a way that $|\mathfrak{E}_k| \leq X$: if $(\widetilde{V}_1, \ldots, \widetilde{V}_k)$ be a size-biased sample without replacement in V, one can generate $\mathfrak{E}_k$ as follows. Let $|\mathfrak{E}_1| = \deg(\widetilde{V}_1)$. Then, at each step $\ell \geq 2$, either the pair is chosen in $\mathfrak{E}_{\ell-1}$ and $|\mathfrak{E}_\ell| = |\mathfrak{E}_{\ell-1}|$, or it is chosen outside, and, if $\tau_{\ell-1}$ denotes the number of vertices in the graph induced by $\mathfrak{E}_{\ell-1}$, then we let $|\mathfrak{E}_\ell| = |\mathfrak{E}_{\ell-1}| + \deg(\widetilde{V}_{\tau_{\ell-1}+1})$. As $\tau_k \leq k$, we have $|\mathfrak{E}_k| \leq \sum_{\ell=1}^{k} \deg(\widetilde{V}_\ell)$.

Now, by Theorem 1.1, X is less, in the increasing convex order than $Y = \sum_{i=1}^{k} \deg(\mathbf{V}_i)$ where $(\mathbf{V}_i)$ is an I.I.D. sequence with the law of $\mathbf{V}$. In particular, all the Chernoff bounds that apply to Y can be transferred to X. Hoeffding's inequality then yields the desired result.

□

Stochastic orders provide powerful tools to compare distributions of random variables and processes, and they have been used in various applications [125, 146, 151]. As other stochastic relations, the increasing convex order is only concerned with marginal distributions. One way of establishing (1.3) is thus to carefully construct two random variables X and Y with the correct marginals on a common probability space, in such a way that

$$X \quad \leq \quad \mathbb{E}[Y|X] \tag{1.5}$$

holds almost-surely. The existence of such a *submartingale coupling* clearly implies (1.3), thanks to Jensen's inequality. Quite remarkably, the converse is also true, as proved by Strassen [149].

Remark 1.1 (The uniform case)**.** *When ω is constant, the sequence $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ is exchangeable. In particular, $\mathbb{E}[X] = \mathbb{E}[Y]$, forcing equality in (1.5). Thus, (1.3) automatically extends to arbitrary convex functions, as established by Hoeffding [92]. We pointed out in Section 1.1 that another remarkable feature uniform sampling without replacement was negative association of the sequence $(\nu(\mathbf{I}_1), \ldots, \nu(\mathbf{I}_n))$ [97]. However, this result also seems to make crucial use of the exchangeability of $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$, and it is not clear whether it can be extended to more general weights,* e.g. *to monotone weights satisfying* (1.2)*. Non-uniform sampling without replacement can be more delicate and induce counter-intuitive correlations, as highlighted by Alexander [8], who showed that for two fixed items, the indicators that each is in the sample can be positively correlated.*

In the non-uniform case, $\mathbb{E}X$ and $\mathbb{E}Y$ need not be equal. Hence, Theorem 1.1 does not entail proper concentration inequalities for X around its mean. The second result of this chapter is to establish a sub-Gaussian concentration inequality for X, which holds for arbitrary weights $(\omega(i))_{i=1}^{N}$. Define

$$\Delta \quad = \quad \max_{1 \leq i \leq N} \nu(i) - \min_{1 \leq i \leq N} \nu(i)\,,$$

and

$$\alpha \quad = \quad \frac{\min_{1\le i\le N}\omega(i)}{\max_{1\le i\le N}\omega(i)}\,.$$

The case $\alpha = 1$ (uniform sampling) was analysed by Serfling [145]. We thus consider $\alpha < 1$ and show the following.

Theorem 1.2. *Assume $\alpha < 1$. For all $t > 0$,*

$$\max\left\{\mathbb{P}\left(X - \mathbb{E}X > t\right), \mathbb{P}\left(X - \mathbb{E}X < -t\right)\right\} \le \exp\left(-\frac{t^2}{2v}\right),$$

with

$$v \quad = \quad \min\left(4\Delta^2 n\,,\ \frac{1+4\alpha}{\alpha(1-\alpha)}\Delta^2 N\left(\frac{N-n}{N}\right)^{\alpha}\right). \tag{1.6}$$

Theorem 1.2 holds under the only assumption that $\alpha < 1$, but the domain of application that we have in mind is when α is bounded away from 0 and 1. In this domain, when $n/N \le q$, for some fixed $0 < q < 1$, equation (1.6) gives $v = O(\Delta^2 n)$, which corresponds to the order of the variance factor in the classical Hoeffding inequality. When $n/N \underset{N\to\infty}{\to} 1$, it can be improved up to $v = O\left(\Delta^2 n\left(\frac{N-n}{N}\right)^{\alpha}\right)$, hence displaying a sharpening effect in the variance, as identified by Serfling [145] in the uniform case. It would be interesting to know whether the α appearing in the exponent can be removed.

Finally, in this chapter, we also answer a question raised by [118]. The problem is to compare linear statistics induced by sampling in Polya urns with replacement number d versus D, for positive integers d, D with $D > d \ge 1$.

Let $\mathbf{C}$ be a population of N items, labelled from 1 to N, each item i being equipped with some value $\nu(i)$. Let $d < D$ be two positive integers. For $n \ge 1$, let $(K_1, \dots, K_n)$ and $(L_1, \dots, L_n)$ be samples generated by sampling in Polya urns with initial composition $\mathbf{C}$ and replacement numbers d and D respectively, *i.e.* each time an item is picked, it is replaced along with $d-1$ (resp. $D-1$) copies. We say that $(K_1, \dots, K_n)$ (resp. $(L_1, \dots, L_n)$) is a d-Polya (resp. D-Polya) sample. Let

$$\begin{aligned} W &= \nu(K_1) + \cdots + \nu(K_n)\,, \\ Z &= \nu(L_1) + \cdots + \nu(L_n)\,. \end{aligned}$$

Theorem 1.3. *The variable W is less than Z in the* convex order*, i.e. for every convex function $f\colon \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\left[f\left(W\right)\right] \quad \le \quad \mathbb{E}\left[f\left(Z\right)\right].$$

Remark 1.2. *[118] proved a similar result in the case where the first sample is drawn without replacement in $\mathbf{C}$ and the second is a D-Polya sample, for $D \ge 1$.*

2 A useful coupling

The proofs of Theorem 1.1 and 1.2 rely on a particular coupling of samples drawn with and without replacement. This coupling is inspired by the one described in [118] for the uniform case.

First generate an infinite sequence $(\mathbf{J}_k)_{k\ge1}$ by sampling with replacement and with probability pro-

portional to $(\omega(i))_{i=1}^N$. Now, "screen" this sequence, starting at $\mathbf{J}_1$ as follows: for $1 \leq k \leq N$, set

$$\mathbf{I}_k \quad = \quad \mathbf{J}_{T_k}\,,$$

where T_k is the random time when the kth distinct item appears in $(\mathbf{J}_i)_{i\geq 1}$.

The sequence $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ is then distributed as a sample without replacement. As above, we define $X = \sum_{k=1}^n \nu(\mathbf{I}_k)$ and $Y = \sum_{k=1}^n \nu(\mathbf{J}_k)$.

3 Proofs

3.1 Proof of Theorem 1.1

Consider the coupling of X and Y described above (Section 2). Under the monotonicity assumption (1.2), we show that (X, Y) is a submartingale coupling in the sense of (1.5). As the sequence $(\mathbf{J}_1, \ldots, \mathbf{J}_n)$ is exchangeable and as permuting $\mathbf{J}_i$ and $\mathbf{J}_j$ in this sequence does not affect X, it is sufficient to show that $\mathbb{E}\left[\nu(\mathbf{J}_1)\middle|X\right] \geq X/n$.

Let $\{i_1, \ldots, i_n\} \subset \{1, \ldots, N\}$ be a set of cardinality n, and let $\mathbf{A}$ be the event $\{\mathbf{I}_1, \ldots, \mathbf{I}_n\} = \{i_1, \ldots, i_n\}$.

$$\mathbb{E}\left[\nu(\mathbf{J}_1)\middle|\mathbf{A}\right] = \sum_{j=1}^n \mathbb{P}\left(\mathbf{J}_1 = i_j\middle|\mathbf{A}\right)\nu(i_j)\,.$$

Let us now show that, for all $1 \leq k \neq \ell \leq n$, if $\nu(i_k) \geq \nu(i_\ell)$, then $\mathbb{P}\left(\mathbf{J}_1 = i_k\middle|\mathbf{A}\right)$ is not smaller than $\mathbb{P}\left(\mathbf{J}_1 = i_\ell\middle|\mathbf{A}\right)$. First, by (1.2), one has $\omega(i_k) \geq \omega(i_\ell)$. Letting $\mathfrak{S}_n$ be the set of permutations of n elements, one has

$$\mathbb{P}\left(\{\mathbf{J}_1 = i_k\} \cap \mathbf{A}\right) \quad = \quad \sum_{\pi \in \mathfrak{S}_n, \pi(1)=k} p(\pi)\,,$$

where

$$p(\pi) \quad := \quad \omega(i_{\pi(1)})\frac{\omega(i_{\pi(2)})}{1-\omega(i_{\pi(1)})}\cdots\frac{\omega(i_{\pi(n)})}{1-\omega(i_{\pi(1)})-\omega(i_{\pi(2)})-\omega(i_{\pi(n-1)})}$$

Now, each permutation π with $\pi(1) = k$ can be uniquely associated with a permutation $\pi^\star$ such that $\pi^\star(1) = \ell$, by performing the switch: $\pi^\star(\pi^{-1}(\ell)) = k$, and letting $\pi(j) = \pi^\star(j)$, for all $j \notin \{1, \pi^{-1}(\ell)\}$. Observe that $p(\pi) \geq p(\pi^\star)$. Thus

$$\mathbb{P}\left(\mathbf{J}_1 = i_k\middle|\mathbf{A}\right) - \mathbb{P}\left(\mathbf{J}_1 = i_\ell\middle|\mathbf{A}\right) \quad = \quad \frac{1}{\mathbb{P}\left(\mathbf{A}\right)}\sum_{\pi \in \mathfrak{S}_n, \pi(1)=k}\left(p(\pi) - p(\pi^\star)\right) \geq 0\,.$$

Consequently, by Chebyshev's sum inequality,

$$\begin{aligned}
\mathbb{E}\left[\nu(\mathbf{J}_1)\middle|\mathbf{A}\right] &= n\frac{1}{n}\sum_{j=1}^{n}\mathbb{P}\left(\mathbf{J}_1 = i_j\middle|\mathbf{A}\right)\nu(i_j) \\
&\geq n\left(\frac{1}{n}\sum_{j=1}^{n}\mathbb{P}\left(\mathbf{J}_1 = i_j\middle|\mathbf{A}\right)\right)\left(\frac{1}{n}\sum_{j=1}^{n}\nu(i_j)\right) \\
&= \frac{\sum_{j=1}^{n}\nu(i_j)}{n},
\end{aligned}$$

and $\mathbb{E}\left[Y\middle|X\right] \geq X$.

□

3.2 Proof of Theorem 1.2

We only need to show that the bound in Theorem 1.2 holds for $\mathbb{P}[X - \mathbb{E}X > t]$. Indeed, replacing X by $-X$ (i.e. changing all the values to their opposite) does not affect the proof. Hence, the bound on $\mathbb{P}[X - \mathbb{E}X < -t]$ will follow directly.

Theorem 1.2 is proved using the same coupling between sampling with and without replacement as described in Section 2.

Note that, in this coupling, X is a function of the I.I.D. variables $(\mathbf{J}_i)_{i\geq 1}$:

$$X = \sum_{i=1}^{+\infty}\nu(\mathbf{J}_i)\mathbb{1}_{\{\mathbf{J}_i\notin\{\mathbf{J}_1,\ldots,\mathbf{J}_{i-1}\}\}}\mathbb{1}_{\{T_n\geq i\}}\,. \tag{3.1}$$

As such, one may obtain concentration results for X by resorting to the various methods designed for functions of independent variables.

The proof relies on the *entropy method* as described in Chapter 6 of [40]. We will show that X is such that, for all $\lambda > 0$,

$$\lambda\mathbb{E}\left[X\,\mathrm{e}^{\lambda X}\right] - \mathbb{E}\left[\mathrm{e}^{\lambda X}\right]\log\mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \leq \frac{\lambda^2 v}{2}\mathbb{E}\left[\mathrm{e}^{\lambda X}\right], \tag{3.2}$$

for v as in (1.6). Then, a classical argument due to Herbst (see [40], Proposition 6.1) ensures that, for all $\lambda > 0$,

$$\log\mathbb{E}\left[\mathrm{e}^{\lambda(X-\mathbb{E}X)}\right] \leq \frac{\lambda^2 v}{2},$$

and thus, for all $t > 0$,

$$\mathbb{P}\left(X - \mathbb{E}X > t\right) \leq \exp\left(-\frac{t^2}{2v}\right),$$

that is, the upper-tail of X is sub-Gaussian with variance factor v. Let us establish inequality (3.2). For $t \geq 1$, consider the truncated variable X_t defined by summing only from 1 to t in (3.1), i.e.

$$\begin{aligned}
X_t &= \sum_{i=1}^{t}\nu(\mathbf{J}_i)\mathbb{1}_{\{\mathbf{J}_i\notin\{\mathbf{J}_1,\ldots,\mathbf{J}_{i-1}\}\}}\mathbb{1}_{\{T_n\geq i\}} \\
&:= f(\mathbf{J}_1,\ldots,\mathbf{J}_t)\,.
\end{aligned}$$

Note that X_t converges to X almost surely as $t \to +\infty$. Then, for all $1 \leq i \leq t$, consider the perturbed variable X_t^i which is obtained by replacing $\mathbf{J}_i$ by an independent copy $\mathbf{J}_i'$, i.e.

$$X_t^i = f(\mathbf{J}_1,\ldots,\mathbf{J}_{i-1},\mathbf{J}_i',\mathbf{J}_{i+1},\ldots,\mathbf{J}_t)\,,$$

and let X^i be the almost sure limit of X_t^i, as $t \to +\infty$. Theorem 6.15 of [40] implies that, for all $\lambda > 0$,

$$\lambda \mathbb{E}\left[X_t \, \mathrm{e}^{\lambda X_t}\right] - \mathbb{E}\left[\mathrm{e}^{\lambda X_t}\right] \log \mathbb{E}\left[\mathrm{e}^{\lambda X_t}\right] \quad \leq \quad \sum_{i=1}^{t} \mathbb{E}\left[\lambda^2 \, \mathrm{e}^{\lambda X_t} (X_t - X_t^i)_+^2\right] . \tag{3.3}$$

We now show that this inequality still holds when we let t tend to $+\infty$. Let $\nu_{\max} = \max_{1 \leq j \leq N} \nu(j)$. For all $t \geq 1$, the variable X_t is almost surely bounded by $n\nu_{\max}$. Hence, the left-hand side of (3.3) tends to the left-hand side of (3.2). As for the right-hand side, we have that, for all $1 \leq i \leq t$,

$$\mathbb{E}\left[\lambda^2 \, \mathrm{e}^{\lambda X_t} (X_t - X_t^i)_+^2\right] \leq \lambda^2 \, \mathrm{e}^{\lambda n \nu_{\max}} \Delta^2 \mathbb{P}(i \leq T_n) \, ,$$

and $\sum_{i=1}^{+\infty} \mathbb{P}[i \leq T_n] = \mathbb{E}[T_n] < +\infty$. Hence, by dominated convergence, the right-hand side also converges, and we obtain

$$\lambda \mathbb{E}\left[X \, \mathrm{e}^{\lambda X}\right] - \mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \log \mathbb{E}\left[\mathrm{e}^{\lambda X}\right] \quad \leq \quad \sum_{i=1}^{+\infty} \mathbb{E}\left[\lambda^2 \, \mathrm{e}^{\lambda X} (X - X^i)_+^2\right] .$$

Recall that $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$ is the sequence of the first n distinct items in $(\mathbf{J}_i)_{i \geq 1}$ and that X is measurable with respect to $\sigma(\mathbf{I}_1, \ldots, \mathbf{I}_n)$, so that

$$\sum_{i=1}^{+\infty} \mathbb{E}\left[\lambda^2 \, \mathrm{e}^{\lambda X} (X - X^i)_+^2\right] = \mathbb{E}\left[\lambda^2 \, \mathrm{e}^{\lambda X} \mathbb{E}\left[\sum_{i=1}^{+\infty} (X - X^i)_+^2 \middle| \mathbf{I}_1, \ldots, \mathbf{I}_n\right]\right] .$$

Thus, letting

$$V := \mathbb{E}\left[\sum_{i=1}^{+\infty} (X - X^i)_+^2 \middle| \mathbf{I}_1, \ldots, \mathbf{I}_n\right] ,$$

our task comes down to showing that

$$V \leq \frac{v}{2} \quad \text{a.s.} .$$

Observe that for all $i \geq 1$, we have $(X - X^i)_+^2 \leq \Delta^2$ and that $X = X^i$ unless $i \leq T_n$ and one of the following two events occurs:

— $\mathbf{J}_i' \notin \{\mathbf{I}_1, \ldots, \mathbf{I}_n\}$;

— the item $\mathbf{J}_i$ occurs only once before T_{n+1}.

Let us define

$$A \quad = \quad \sum_{i=1}^{+\infty} \mathbb{E}\left[\mathbb{1}_{\{\mathbf{J}_i' \notin \{\mathbf{I}_1, \ldots, \mathbf{I}_n\}\}} \mathbb{1}_{i \leq T_n} \middle| \mathbf{I}_1, \ldots, \mathbf{I}_n\right] ,$$

and

$$B \quad = \quad \sum_{k=1}^{n} \mathbb{E}\left[\mathbb{1}_{\{\exists!\, i < T_{n+1},\, \mathbf{J}_i = \mathbf{I}_k\}} \middle| \mathbf{I}_1, \ldots, \mathbf{I}_n\right] ,$$

so that $V \leq \Delta^2 (A + B)$. Since $\mathbf{J}_i'$ is independent of everything else and since $\sigma_n := \omega(\mathbf{I}_1) + \ldots \omega(\mathbf{I}_n)$ is a measurable function of $(\mathbf{I}_1, \ldots, \mathbf{I}_n)$, we have

$$A \quad = \quad (1 - \sigma_n) \mathbb{E}\left[T_n \middle| \mathbf{I}_1, \ldots, \mathbf{I}_n\right] .$$

We use the following fact.

Lemma 3.1. *For $1 \le k \le n$, let $\tau_k = T_k - T_{k-1}$. Conditionally on $(\mathbf{I}_1, \dots, \mathbf{I}_n)$, the variables $(\tau_k)_{k=1}^n$ are independent and for all $1 \le k \le n$, τ_k is distributed as a Geometric random variables with parameters $1 - \sigma_{k-1}$.*

Proof. Let $(i_1, \dots, i_n)$ be an n-tuple of distinct elements of $\{1, \dots, N\}$ and let $t_1, \dots, t_n \ge 1$. Let also $(G_k)_{k=1}^n$ be independent Geometric random variables with parameter $(1 - \omega(i_1) - \cdots - \omega(i_{k-1}))$. We have

$$\begin{aligned}
&\mathbb{P}\left((\tau_1, \dots, \tau_n) = (t_1, \dots, t_n), (\mathbf{I}_1, \dots, \mathbf{I}_n) = (i_1, \dots, i_n)\right) \\
&\quad = \mathbb{1}_{\{t_1 = 1\}} \omega(i_1) \prod_{k=2}^n \left(\omega(i_1) + \cdots + \omega(i_{k-1})\right)^{t_k - 1} \omega(i_k) \\
&\quad = \prod_{k=1}^n \frac{\omega(i_k)}{1 - \omega(i_1) - \cdots - \omega(i_{k-1})} \prod_{k=1}^n \mathbb{P}\left(G_k = t_k\right) \\
&\quad = \mathbb{P}\left((\mathbf{I}_1, \dots, \mathbf{I}_n) = (i_1, \dots, i_n)\right) \prod_{k=1}^n \mathbb{P}\left(G_k = t_k\right) ,
\end{aligned}$$

and we obtain the desired result. □

Lemma 3.1 implies that

$$\mathbb{E}\left[T_n \middle| \mathbf{I}_1, \dots, \mathbf{I}_n\right] = \sum_{k=1}^n \frac{1}{1 - \sigma_{k-1}} .$$

In particular, $A \le n$. We also have

$$A \le \frac{1}{\alpha} \sum_{k=1}^n \frac{N - n}{N - k + 1} \le \frac{1}{\alpha}(N - n) \log\left(\frac{N}{N - n}\right) . \tag{3.4}$$

It remains to control B. Clearly $B \le n$, which shows that $V \le 2\Delta^2 n$. Moreover, for $1 \le k \le n$, we have

$$\mathbb{P}\left(\exists!\, i < T_{n+1},\, \mathbf{J}_i = \mathbf{I}_k \middle| \mathbf{I}_1, \dots, \mathbf{I}_n\right) = \mathbb{E}\left[\prod_{j=k}^n \left(1 - \frac{\omega(\mathbf{I}_k)}{\sigma_j}\right)^{\tau_{j+1} - 1} \middle| \mathbf{I}_1, \dots, \mathbf{I}_n\right] .$$

Using Lemma 3.1 and the fact that the generating function of a geometric variable G with parameter p is given by $\mathbb{E}\left[x^G\right] = \frac{px}{1-(1-p)x}$, we obtain

$$B = \sum_{k=1}^n \prod_{j=k}^n \frac{1}{1 + \frac{\omega(\mathbf{I}_k)}{1 - \sigma_j}} .$$

Thanks to the inequality the inequality $\log(1 + x) \ge x - x^2/2$ for $x \ge 0$,

$$B \le \sum_{k=1}^n \prod_{j=k}^n \frac{1}{1 + \frac{\alpha}{N - j}} \le \sum_{k=1}^n \exp\left(-\alpha \sum_{j=k}^n \frac{1}{N - j} + \frac{1}{2} \sum_{j=k}^n \frac{1}{(N - j)^2}\right) .$$

The second term in the exponent is always smaller than 1/2. Using Riemann sums, we get

$$\begin{aligned} B &\leq 2\sum_{k=1}^{n} \exp\left(-\alpha \log\left(\frac{N-k+1}{N-n}\right)\right) = 2\sum_{k=1}^{n}\left(\frac{N-n}{N-k+1}\right)^{\alpha} \\ &\leq \frac{2}{1-\alpha} N \left(\frac{N-n}{N}\right)^{\alpha}, \end{aligned}$$

Combined with (3.4), this yields

$$\begin{aligned} V &\leq \left(\frac{1}{\alpha}\left(\frac{N-n}{N}\right)^{1-\alpha} \log\left(\frac{N}{N-n}\right) + \frac{2}{1-\alpha}\right) \Delta^2 N \left(\frac{N-n}{N}\right)^{\alpha} \\ &\leq \left(\frac{\mathrm{e}^{-1}}{\alpha(1-\alpha)} + \frac{2}{1-\alpha}\right) \Delta^2 N \left(\frac{N-n}{N}\right)^{\alpha} \\ &\leq \frac{1/2+2\alpha}{\alpha(1-\alpha)} \Delta^2 N \left(\frac{N-n}{N}\right)^{\alpha}, \end{aligned}$$

where the second inequality is due to the fact that $\log(x)/x^{1-\alpha} \leq \mathrm{e}^{-1}/(1-\alpha)$ for all $x > 0$.

□

3.3 Proof of Theorem 1.3

The proof of Theorem 1.3 relies on the construction of a martingale coupling (W, Z), *i.e.* of a coupling of W and Z such that $\mathbb{E}\left[Z \middle| W\right] = W$.

Consider two urns, $\mathbf{U}_d$ and $\mathbf{U}_D$, each of them initially containing N balls, labelled from 1 to N. In each urn, arrange the balls from left to right by increasing order of their label. Then arrange $\mathbf{U}_D$ and $\mathbf{U}_d$ on top of one another. Each time we will pick a ball in $\mathbf{U}_D$, we will pick the ball just below it in $\mathbf{U}_d$. More precisely, we perform an infinite sequence of steps as follows: at step 1, we pick a ball B_1 uniformly at random in $\mathbf{U}_D$ and pick the ball just below it in $\mathbf{U}_d$. They necessarily have the same label, say j. We let $K_1 = L_1 = j$, and add, on the right part of $\mathbf{U}_D$, $D-1$ balls with label j, and, on the right part of $\mathbf{U}_d$, $d-1$ balls with label j and $D-d$ unlabelled balls. Note that, at the end of this step, the two urns still have the same number of balls, $N+D-1$. The first step is depicted in Figure 2.5. Then, at each step t, we pick a ball B_t at random among the $N+(t-1)(D-1)$ balls of $\mathbf{U}_D$ and choose the ball just below it in $\mathbf{U}_d$. There are two different possibilities:

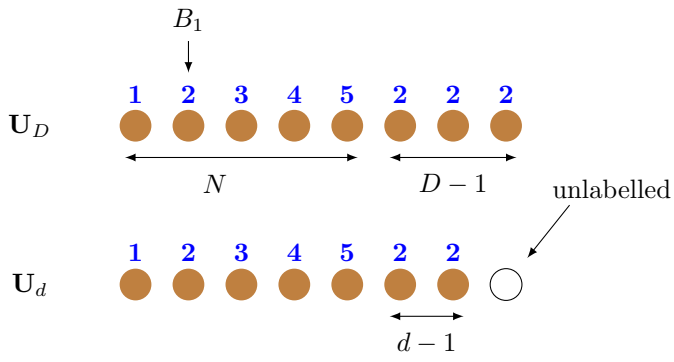
- if the ball drawn in $\mathbf{U}_d$ is unlabelled and the one drawn in $\mathbf{U}_D$ has label j, we let $L_t = j$ and add $D-1$ balls with label j on the right part of $\mathbf{U}_D$, and $D-1$ unlabelled balls on the right part of $\mathbf{U}_d$, .
- if both balls have label j, and if t corresponds to the ith time a labelled ball is drawn in $\mathbf{U}_d$, we let $L_t = K_i = j$ and add $D-1$ balls with label j on the right part of $\mathbf{U}_D$, and $d-1$ balls with label j and $D-d$ unlabelled balls on the right part of $\mathbf{U}_d$;

The sequence $(K_1, \ldots, K_n)$ records the labels of the first n labelled balls picked in $\mathbf{U}_d$, and $(L_1, \ldots, L_n)$ the labels of the first n balls picked in $\mathbf{U}_D$. Observe that $(K_1, \ldots, K_n)$ (resp. $(L_1, \ldots, L_n)$) is distributed as a d-Polya (resp. D-Polya) sample. Define

$$\begin{aligned} W &= \nu(K_1) + \cdots + \nu(K_n), \\ Z &= \nu(L_1) + \cdots + \nu(L_n). \end{aligned}$$

Let us show that $1 \leq i \leq n-1$, $\mathbb{E}\left[\nu(L_{i+1}) \middle| W\right] = \mathbb{E}\left[\nu(L_i) \middle| W\right]$. Let $\{k_1, \ldots, k_n\}$ be a multiset of cardinality n of elements of $\{1, \ldots, N\}$, and let $\mathbf{A}$ be the event $\{K_1, \ldots, K_n\} = \{k_1, \ldots, k_n\}$ (accounting

Figure 2.5 – The ball B_1 has label 2 ($N = 5$, $d = 3$, $D = 4$).



for the multiplicity of each label). Denote by $\mathcal{C}_i$ the set of $D-1$ balls added at step i. Observe that, if $B_{i+1} \in \mathcal{C}_i$, then $L_{i+1} = L_i$. Hence

$$\mathbb{E}\left[\nu(L_{i+1})\middle|\mathbf{A}\right] \quad = \quad \mathbb{E}\left[\nu(L_i)\mathbb{1}_{\{B_{i+1}\in\mathcal{C}_i\}}\middle|\mathbf{A}\right] + \mathbb{E}\left[\nu(L_{i+1})\mathbb{1}_{\{B_{i+1}\notin\mathcal{C}_i\}}\middle|\mathbf{A}\right] .$$

We have

$$\mathbb{E}\left[\nu(L_{i+1})\mathbb{1}_{\{B_{i+1}\notin\mathcal{C}_i\}}\middle|\mathbf{A}\right] \quad = \quad \frac{1}{\mathbb{P}(\mathbf{A})}\sum_{k=1}^{N}\nu(k)\sum_{\ell=1}^{N}\mathbb{P}\left(L_i = \ell, L_{i+1} = k, B_{i+1} \notin \mathcal{C}_i, \mathbf{A}\right) .$$

Notice that, on the event $B_{i+1} \notin \mathcal{C}_i$, the balls B_i and B_{i+1} are exchangeable. Hence $\mathbb{P}\left(L_i = \ell, L_{i+1} = k, B_{i+1} \notin \mathcal{C}_i\right) = \mathbb{P}\left(L_i = k, L_{i+1} = \ell, B_{i+1} \notin \mathcal{C}_i\right)$. Moreover, permuting B_i and B_{i+1} can not affect the multiset $\{K_1, \ldots, K_n\}$. Hence

$$\mathbb{E}\left[\nu(L_{i+1})\mathbb{1}_{\{B_{i+1}\notin\mathcal{C}_i\}}\middle|\mathbf{A}\right] \quad = \quad \mathbb{E}\left[\nu(L_i)\mathbb{1}_{\{B_{i+1}\notin\mathcal{C}_i\}}\middle|\mathbf{A}\right] ,$$

and $\mathbb{E}\left[\nu(L_{i+1})\middle|W\right] = \mathbb{E}\left[\nu(L_i)\middle|W\right]$. We get that, for all $1 \le i \le n$,

$$\mathbb{E}\left[\nu(L_i)\middle|W\right] \quad = \quad \mathbb{E}\left[\nu(L_1)\middle|W\right] \; = \; \mathbb{E}\left[\nu(K_1)\middle|W\right] \; = \; W/n ,$$

where the last equality comes from the exchangeability of $(K_1, \ldots, K_n)$. $\square$

Bibliography

[1] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Tight bounds for universal compression of large alphabets. In *ISIT*, pages 2875–2879, 2013.

[2] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pages 764–796, 2013.

[3] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Poissonization and universal compression of envelope classes, February 2014. Personal communication.

[4] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Universal compression of envelope classes: Tight characterization via poisson sampling. *CoRR*, abs/1405.7460, 2014. URL `http://arxiv.org/abs/1405.7460`.

[5] J. Acharya, C. Daskalakis, and G. C. Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems*, pages 3577–3598, 2015.

[6] D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Seminar on probability, XVII*, volume 986 of *Lecture Notes in Math.*, pages 243–297. Springer, Berlin, 1983.

[7] D. Aldous and P. Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, pages 333–348, 1986.

[8] K. S. Alexander et al. A counterexample to a correlation inequality in finite sampling. *The Annals of Statistics*, 17(1):436–439, 1989.

[9] N. Alon, I. Benjamini, E. Lubetzky, and S. Sodin. Non-backtracking random walks mix faster. *Communications in Contemporary Mathematics*, 9(04):585–603, 2007.

[10] C. Anderson. Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *Journal of Applied Probability*, pages 99–113, 1970.

[11] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[12] R. Bahadur. On the number of distinct values in a large sample from an infinite discrete distribution. 26A(Supp. II):67–75, 1960.

[13] A. D. Barbour and A. V. Gnedin. Small counts in the infinite occupancy scheme. *Electron. J. Probab.*, 14:no. 13, 365–384, 2009. ISSN 1083-6489. doi: 10.1214/EJP.v14-608. URL `http://dx.doi.org/10.1214/EJP.v14-608`.

[14] R. Bardenet and O.-A. Maillard. Concentration inequalities for sampling without replacement. *ArXiv e-prints*, Sept. 2013.

[15] J. Bartroff, L. Goldstein, and Ü. Işlak. Bounded size biased couplings, log concave distributions and concentration of measure for occupancy models. *arXiv preprint arXiv:1402.6769*, 2014.

[16] R. Basu, J. Hermon, and Y. Peres. Characterization of cutoff for reversible Markov chains. 2015.

[17] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[18] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete

distributions. *J. ACM*, 60(1):Art. 4, 25, 2013. ISSN 0004-5411. doi: 10.1145/2432622.2432626. URL `http://dx.doi.org/10.1145/2432622.2432626`.

[19] D. Bayer and P. Diaconis. Trailing the dovetail shuffle to its lair. *The Annals of Applied Probability*, pages 294–313, 1992.

[20] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006.

[21] A. Ben-Hamou and J. Salez. Cutoff for non-backtracking random walks on sparse random graphs. *arXiv preprint arXiv:1504.02429*, 2015.

[22] A. Ben-Hamou, S. Boucheron, and M. I. Ohannessian. Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *arXiv preprint arXiv:1412.8652*, 2014.

[23] A. Ben-Hamou, Y. Peres, and J. Salez. Weighted sampling without replacement. *arXiv preprint arXiv:1603.06556*, 2016.

[24] I. Benjamini, H. Duminil-Copin, G. Kozma, and A. Yadin. Disorder, entropy and harmonic functions. *arXiv preprint arXiv:1111.4853*, 2011.

[25] D. Berend and A. Kontorovich. On the concentration of the missing mass. *Electron. Commun. Probab.*, 18:no. 3, 7, 2013.

[26] D. Berend and A. Kontorovich. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.

[27] N. Berestycki and B. Sengul. Cutoff for conjugacy-invariant random walks on the permutation group. *arXiv preprint arXiv:1410.4800*, 2014.

[28] N. Berestycki, E. Lubetzky, Y. Peres, and A. Sly. Random walks on the random graph. *arXiv preprint arXiv:1504.01999*, 2015.

[29] J. Bertoin. *Random fragmentation and coagulation processes*. Cambridge University Press Cambridge, 2006.

[30] N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27. Cambridge university press, 1989.

[31] L. Bogachev, A. Gnedin, and Y. Yakubovich. On the variance of the number of occupied boxes. *Adv. in Appl. Math.*, 40(4):401–432, 2008.

[32] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[33] B. Bollobás. *Random graphs*. Springer, 1998.

[34] D. Bontemps. Universal coding on infinite alphabets: exponentially decreasing envelopes. *IEEE Trans. Inform. Theory*, 57(3):1466–1478, 2011.

[35] D. Bontemps, S. Boucheron, and E. Gassiat. About adaptive coding on countable alphabets. *IEEE Trans. Inform. Theory*, 60(2):808–821, 2014. URL `http://arxiv.org/abs/1202.0258`.

[36] C. Bordenave, P. Caputo, and J. Salez. Random walk on sparse random digraphs. *arXiv preprint arXiv:1508.06600*, 2015.

[37] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[38] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

[39] S. Boucheron, A. Garivier, and E. Gassiat. Coding over Infinite Alphabets. *IEEE Trans. Inform. Theory*, 55:358–373, 2009.

[40] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL `http://dx.doi.org/10.1093/acprof:oso/9780199535255.001.0001`.

[41] S. Boucheron, E. Gassiat, and M. I. Ohannessian. About adaptive coding on countable alphabets: Max-stable envelope classes. *IEEE Trans. Inform. Theory*, 61(9):4948–4967, 2015.

[42] A. Broder and E. Shamir. On the second eigenvalue of random regular graphs. In *Foundations of Computer Science, 1987., 28th Annual Symposium on*, pages 286–294, Oct 1987.

[43] J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[44] O. Catoni. *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004. Ecole d'Ete de Probabilites de Saint-Flour XXXI.

[45] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[46] S. Chatterjee. Stein's method for concentration inequalities. *Probability theory and related fields*, 138(1):305–321, 2007.

[47] G.-Y. Chen and L. Saloff-Coste. The cutoff phenomenon for ergodic Markov processes. *Electronic Journal of Probability*, 13(3):26–78, 2008.

[48] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal approximation by Stein's method*. Springer, 2010.

[49] C. Cooper. Random walks, interacting particles, dynamic networks: Randomness can be helpful. In *Structural Information and Communication Complexity*, pages 1–14. 2011.

[50] C. Cooper and A. Frieze. Vacant sets and vacant nets: Component structures induced by a random walk. *arXiv preprint arXiv:1404.4403*, 2014.

[51] T. Cover and J. Thomas. *Elements of information theory*. John Wiley & sons, 1991.

[52] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Channels*. Academic Press, 1981.

[53] P. Cuff, J. Ding, O. Louidor, E. Lubetzky, Y. Peres, and A. Sly. Glauber dynamics for the mean-field potts model. *Journal of Statistical Physics*, 149(3):432–477, 2012.

[54] A. B. Cybakov. *Introduction à l'estimation non paramétrique*, volume 41. Springer Science & Business Media, 2003.

[55] L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007.

[56] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[57] L. Devroye and L. Gyorfi. *Nonparametric density estimation: the L1 view*, volume 119. John Wiley & Sons Incorporated, 1985.

[58] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[59] P. Diaconis. The cutoff phenomenon in finite Markov chains. *Proc. Nat. Acad. Sci. U.S.A.*, 93(4): 1659–1664, 1996.

[60] P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Probability Theory and Related Fields*, 57(2):159–179, 1981.

[61] P. Diaconis, R. L. Graham, and J. A. Morrison. Asymptotic analysis of a random walk on a hypercube with many dimensions. *Random structures and algorithms*, 1(1):51–72, 1990.

[62] J. Ding, E. Lubetzky, and Y. Peres. The mixing time evolution of Glauber dynamics for the

mean-field Ising model. *Communications in Mathematical Physics*, 289(2):725–764, 2009.

[63] J. Ding, E. Lubetzky, and Y. Peres. Total variation cutoff in birth-and-death chains. *Probability theory and related fields*, 146(1-2):61–85, 2010.

[64] D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

[65] M. Dutko et al. Central limit theorems for infinite urn models. *The Annals of Probability*, 17(3): 1255–1263, 1989.

[66] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

[67] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[68] P. Elias. Universal codeword sets and representations of the integers. *IEEE Trans. Information Theory*, IT-21:194–203, 1975.

[69] W. Esty. Confidence intervals for the coverage of low coverage samples. *Ann. Statist.*, 10(1):190–196, 1982.

[70] W. Esty. The efficiency of Good's nonparametric coverage estimator. *Ann. Statist.*, 14(3):1257–1260, 1986.

[71] W. W. Esty. Confidence intervals for an occupancy problem estimator used by numismatists. *Math. Sci.*, 9(2):111–115, 1984. ISSN 0312-3685.

[72] M. Falahatgar, A. Jafarpour, A. Orlitsky, V. Pichapati, and A. T. Suresh. Universal compression of power-law distributions. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2001–2005. IEEE, 2015.

[73] R. A. Fisher, A. S. Corbet, and C. B. Williams. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *Journal of Animal Ecology*, 12(1):42–58, May 1943. ISSN 00218790. URL `http://www.jstor.org/stable/1411`.

[74] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[75] J. Friedman. *A proof of Alon's second eigenvalue conjecture and related problems.* American Mathematical Soc., 2008.

[76] W. A. Gale and G. Sampson. Good-turing frequency estimation without tears*. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[77] S. Ganguly, E. Lubetzky, and F. Martinelli. Cutoff for the east process. *arXiv preprint arXiv:1312.7863*, 2013.

[78] A. Garivier. A lower bound for the maximin redundancy in pattern coding. *Entropy*, 11:634–642, 2009.

[79] E. Gassiat. *Codage universel et identification d'ordre par sélection de modèles.* Société Mathématique de France, 2014.

[80] A. Gnedin, B. Hansen, J. Pitman, et al. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv*, 4:146–171, 2007.

[81] A. V. Gnedin. Regeneration in random combinatorial structures. *Probab. Surv.*, 7:105–156, 2010. ISSN 1549-5787. doi: 10.1214/10-PS163. URL `http://dx.doi.org/10.1214/10-PS163`.

[82] I. J. Good. The Population Frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264, 1953. URL `http://www.bibsonomy.org/bibtex/224e62d66370d0b400e268c0113deb188/nlp`.

[83] I. J. Good and G. H. Toulmin. The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased. *Biometrika*, 43(1/2):45–63, June 1956. ISSN 00063444.

URL http://www.jstor.org/stable/2333577.

[84] L. Gordon. Successive sampling in large finite populations. *Ann. Statist.*, 11(2):702–706, 1983. ISSN 0090-5364.

[85] S. Griffiths, R. Kang, R. Oliveira, and V. Patel. Tight inequalities among set hitting times in Markov chains. *Proceedings of the American Mathematical Society*, 142(9):3285–3298, 2014.

[86] R. Grübel and P. Hitczenko. Gaps in discrete random samples. *J. Appl. Probab.*, 46(4):1038–1051, 2009. ISSN 0021-9002. doi: 10.1239/jap. URL http://dx.doi.org/10.1239/jap/1261670687.

[87] L. Gyorfi, I. Pali, and E. van der Meulen. On universal noiseless source coding for infinite source alphabets. *Eur. Trans. Telecommun. & Relat. Technol.*, 4(2):125–132, 1993.

[88] L. Györfi, I. Pali, and E. C. Van Der Meulen. There is no universal source code for an infinite source alphabet. *IEEE Trans. Inform. Theory*, 40(1):267–271, 1994.

[89] Y. Han, J. Jiao, and T. Weissman. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Trans. Inform. Theory*, 61(11):6343–6354, 2015. ISSN 0018-9448. doi: 10.1109/TIT.2015.2478816. URL http://dx.doi.org/10.1109/TIT.2015.2478816.

[90] Y. Han, J. Jiao, and T. Weissman. Adaptive estimation of shannon entropy. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1372–1376. IEEE, 2015.

[91] J. Hermon. A technical report on hitting times, mixing and cutoff. *arXiv preprint arXiv:1501.01869*, 2015.

[92] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963. ISSN 0162-1459.

[93] L. Holst. Some limit theorems with applications in sampling theory. *Ann. Statist.*, 1:644–658, 1973. ISSN 0090-5364.

[94] H.-K. Hwang and S. Janson. Local limit theorems for finite and infinite urn models. *Ann. Probab.*, 36(3):992–1022, 2008. ISSN 0091-1798. doi: 10.1214/07-AOP350. URL http://dx.doi.org/10.1214/07-AOP350.

[95] S. Janson. The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, 18(1-2):205–225, 2009.

[96] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inform. Theory*, 61(5):2835–2885, 2015. ISSN 0018-9448. doi: 10.1109/TIT.2015.2412945. URL http://dx.doi.org/10.1109/TIT.2015.2412945.

[97] K. Joag-Dev and F. Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295, 1983.

[98] N. L. Johnson and S. Kotz. Urn models and their application; an approach to modern discrete probability theory. 1977.

[99] V. Kaimanovich. Boundary and entropy of random walks in random environment. *Prob. Theory and Math. Stat*, 1:573–579, 1990.

[100] V. A. Kaimanovich and A. M. Vershik. Random walks on discrete groups: boundary and entropy. *The annals of probability*, pages 457–490, 1983.

[101] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, pages 1066–1100, 2015.

[102] S. Karlin. Central limit theorems for certain infinite urn schemes. *J. Math. Mech.*, 17:373–401, 1967.

[103] M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 311–319. Morgan Kaufmann Publishers Inc., 1998.

[104] J. C. Kieffer. A unified approach to weak universal source coding. *IEEE Trans. Inform. Theory*,

24(6):674–682, 1978.

[105] T. Klein, E. Rio, et al. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[106] V. Kolchin, B. Sevast′yanov, and V. Chistyakov. *Random allocations*. V. H. Winston & Sons, Washington, D.C.; distributed by Halsted Press [John Wiley & Sons], New York-Toronto, Ont.-London, 1978. Translated from the Russian, Translation edited by A. V. Balakrishnan, Scripta Series in Mathematics.

[107] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *Information Theory, IEEE Transactions on*, 27(2):199–207, 1981.

[108] H. Lacoin. The Cutoff profile for the Simple-Exclusion process on the cycle. *arXiv preprint arXiv:1502.00952*, 2015.

[109] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. *Theory Comput.*, 9:295–347, 2013. ISSN 1557-2862. doi: 10.4086/toc.2013.v009a008. URL `http://dx.doi.org/10.4086/toc.2013.v009a008`.

[110] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[111] D. A. Levin, M. J. Luczak, and Y. Peres. Glauber dynamics for the mean-field Ising model: cut-off, critical power law, and metastability. *Probability Theory and Related Fields*, 146(1-2):223–265, 2010.

[112] E. Lubetzky and Y. Peres. Cutoff on all ramanujan graphs. *arXiv preprint arXiv:1507.04725*, 2015.

[113] E. Lubetzky and A. Sly. Cutoff phenomena for random walks on random regular graphs. *Duke Mathematical Journal*, 153(3):475–510, 2010.

[114] E. Lubetzky and A. Sly. Explicit expanders with cutoff phenomena. *Electronic Journal of Probability*, 16:419–435, 2011.

[115] E. Lubetzky and A. Sly. Cutoff for general spin systems with arbitrary boundary conditions. *Communications on Pure and Applied Mathematics*, 67(6):982–1027, 2014.

[116] E. Lubetzky and A. Sly. Universality of cutoff for the Ising model. *arXiv preprint arXiv:1407.1761*, 2014.

[117] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[118] K. Luh and N. Pippenger. Large-deviation bounds for sampling without replacement. *American Mathematical Monthly*, 121(5):449–454, 2014.

[119] R. Lyons, R. Pemantle, and Y. Peres. Ergodic theory on galton-watson trees: Speed of random walk and dimension of harmonic measure. *Ergodic Theory and Dynamical Systems*, 15(03):593–619, 1995.

[120] L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

[121] D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *The Journal of Machine Learning Research*, 4:895–911, 2003.

[122] D. A. McAllester and R. E. Schapire. On the convergence rate of Good-Turing estimators. In *COLT 2000*, pages 1–6, 2000.

[123] E. Mossel and M. I. Ohannessian. On the impossibility of learning the missing mass. *arXiv preprint arXiv:1503.03613*, 2015.

[124] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[125] A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2002. ISBN 0-471-49446-1.

[126] A. Nilli. On the second eigenvalue of a graph. *Discrete Mathematics*, 91(2):207–210, 1991.

[127] M. I. Ohannessian and M. A. Dahleh. Rare Probability Estimation under Regularly Varying Heavy Tails. *Journal of Machine Learning Research-Proceedings Track*, 23:21–1, 2012.

[128] R. I. Oliveira. Mixing and hitting times for finite Markov chains. *Electron. J. Probab*, 17(70):1–12, 2012.

[129] A. Orlitsky and N. P. Santhanam. Speaking of infinity. *IEEE Trans. Inform. Theory*, 50(10): 2215–2230, 2004. ISSN 0018-9448.

[130] A. Orlitsky and A. T. Suresh. Competitive distribution estimation: Why is good-turing good. In *Advances in Neural Information Processing Systems*, pages 2134–2142, 2015.

[131] A. Orlitsky, N. Santhanam, and J. Zhang. Always good turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

[132] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 426–435. AUAI Press, 2004.

[133] A. Orlitsky, N. P. Santhanam, and J. Zhang. Universal compression of memoryless sources over unknown alphabets. *Information Theory, IEEE Transactions on*, 50(7):1469–1481, 2004.

[134] Y. Peres. American institute of mathematics (AIM) research workshop "sharp thresholds for mixing times" (Palo Alto, December 2004). *Summary available at http://www.aimath.org/WWN/mixingtimes*, 2004.

[135] Y. Peres and P. Sousi. Mixing times are hitting times of large sets. *Journal of Theoretical Probability*, pages 1–32, 2013.

[136] M. S. Pinsker. On the complexity of a concentrator. In *7th International Telegraffic Conference*, volume 4, pages 1–318. Citeseer, 1973.

[137] J. Pitman and N. M. Tran. Size biased permutation of a finite sequence with independent and identically distributed terms. *ArXiv e-prints*, Oct. 2012.

[138] J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.

[139] M. Raginsky and I. Sason. *Concentration of Measure Inequalities in Information Theory, Communications and Coding*, volume 10. NOW, 2013. URL `http://arxiv.org/abs/1212.4663`.

[140] J. Rissanen. Stochastic complexity and modeling. *The annals of statistics*, pages 1080–1100, 1986.

[141] J. Rissanen and J. G. G. Langdon. Arithmetic coding. *IBM J. Res. Develop.*, 23(2):149–162, 1979. ISSN 0018-8646.

[142] B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement. I, II. *Ann. Math. Statist.*, 43:373–397; ibid. 43 (1972), 748–776, 1972. ISSN 0003-4851.

[143] N. Ross. Fundamentals of Stein's method. *Probab. Surv*, 8:210–293, 2011.

[144] L. Saloff-Coste. Random walks on finite groups. In *Probability on discrete structures*, pages 263–346. Springer, 2004.

[145] R. J. Serfling. Probability inequalities for the sum in sampling without replacement. *Ann. Statist.*, 2:39–48, 1974. ISSN 0090-5364.

[146] M. Shaked and J. G. Shanthikumar. *Stochastic orders*. Springer Series in Statistics. Springer, New York, 2007. ISBN 978-0-387-32915-4; 0-387-32915-3. doi: 10.1007/978-0-387-34675-5. URL `http://dx.doi.org/10.1007/978-0-387-34675-5`.

[147] G. I. Shamir. Universal lossless compression with unknown alphabets&# 8212; the average case. *Information Theory, IEEE Transactions on*, 52(11):4915–4944, 2006.

[148] Q.-M. Shao. A comparison theorem on moment inequalities between negatively associated and

independent random variables. *Journal of Theoretical Probability*, 13(2):343–356, 2000.

[149] V. Strassen. The existence of probability measures with given marginals. *Ann. Math. Statist.*, 36: 423–439, 1965. ISSN 0003-4851.

[150] A. T. Suresh. Statistical inference over large domains. 2016.

[151] R. Szekli. *Stochastic ordering and dependence in applied probability*, volume 97 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1995. ISBN 0-387-94450-8. doi: 10.1007/978-1-4612-2528-7. URL `http://dx.doi.org/10.1007/978-1-4612-2528-7`.

[152] Y. W. Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

[153] G. Valiant and P. Valiant. Estimating the Unseen: An N/Log(N)-sample Estimator for Entropy and Support Size, Shown Optimal via New CLTs. STOC '11, pages 685–694, 2011.

[154] R. van der Hofstad. *Random Graphs and Complex Networks*. 2013. Course notes available at `http://www.win.tue.nl/~rhofstad/NotesRGCN.html`.

[155] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. Distances in random graphs with finite variance degrees. *Random Structures Algorithms*, 27(1):76–123, 2005. ISSN 1042-9832.

[156] Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *Information Theory, IEEE Transactions on*, 46(2):431–445, 2000.

[157] Y. Yu. On the inclusion probabilities in some unequal probability sampling plans without replacement. *Bernoulli*, 18(1):279–289, 2012. ISSN 1350-7265. doi: 10.3150/10-BEJ337. URL `http://dx.doi.org/10.3150/10-BEJ337`.