Inference on dependent data : Contributions to hidden Markov and preferential attachment graph models

Ibrahim Kaddouri

université
PARIS-SACLAY

Inference on dependent data: Contributions to hidden Markov and preferential attachment graph models

Inférence sur des données dépendantes: Contributions aux modèles de Markov cachés et de graphes à attachement préférentiel

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 574 Mathématiques Hadamard (EDMH)
Spécialité de doctorat: Mathématiques appliquées
Graduate School: Mathématiques
Référent: Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire de Mathématiques d'Orsay (Université Paris-Saclay, CNRS)**, sous la direction de **Elisabeth GASSIAT**, Professeure, et la co-direction de **Zacharie NAULET**, Professeur junior.

Thèse soutenue à Paris-Saclay, le 1 décembre 2025, par

Ibrahim KADDOURI

Composition du jury

Membres du jury avec voix délibérative

Catherine MATIAS Directrice de recherche, CNRS	Présidente
Chao GAO Professeur, University of Chicago	Rapporteur & examinateur
Yohann DE CASTRO Professeur, École Centrale Lyon	Rapporteur & examinateur
Christine KERIBIN Professeure, Université Paris-Saclay	Examinatrice
Christophe GIRAUD Professeur, Université Paris-Saclay	Examinateur
Nicolas VERZELEN Directeur de recherche, INRAE Montpellier	Examinateur

THESE DE DOCTORAT
NNT: 2025UPASM035

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre: Inférence sur des données dépendantes : contributions aux modèles de Markov cachés et de graphes à attachement préférentiel

Mots clés: Données dépendantes, modèles de Markov cachés, regroupement, risque de Bayes, attachement préférentiel, détection de ruptures

Résumé: Cette thèse est consacrée à l'étude de l'inférence statistique dans des contextes où les données sont dépendantes. Une telle dépendance peut apparaître à travers une dynamique temporelle, comme dans les modèles de Markov cachés, ou à travers une évolution structurelle, comme dans les graphes à attachement préférentiel. La première partie porte sur le problème de regroupement dans les modèles i.i.d. et de Markov cachés. Nous analysons le risque de Bayes du problème de regroupement et le comparons au risque de classification. Nos résultats montrent que les minimiseurs de ces risques ne coïncident pas toujours, mais nous démontrons que cette distinction reste essentiellement théorique : en pratique, le classificateur de Bayes s'avère presque optimal pour les tâches de regroupement. Des simulations viennent compléter nos résultats théoriques et illustrer la quasi-optimalité de la stratégie de regroupement basée sur le classificateur de Bayes. La deuxième partie développe une analyse plus fine du problème de regroupement pour les modèles de Markov cachés gaussiens dans le régime de lente mélangeabilité de la chaîne cachée. Nous fournissons une caractérisation précise du risque de Bayes en fonction du rapport signal sur bruit et des propriétés de mélange de la chaîne. Nous proposons aussi procédures de regroupement quasi-optimales. De manière intéressante, notre étude révèle des comportements surprenants du risque de Bayes dans certains régimes de paramètres, montrant que la dépendance temporelle peut conduire à des phénomènes non standard absents dans le cas i.i.d. La troisième partie s'intéresse à la détection de ruptures dans des graphes à attachement préférentiel. Nous considérons le problème de la détection tardive d'une rupture qui apparait à $\tau_n = n - \Delta_n$ où n est la taille du graphe et $0 \leq \Delta_n \leq n$. Formulé comme un problème de test d'hypothèses, cela conduit à des comportements très différents selon que le graphe observé est étiqueté ou non. Pour les graphes non étiquetés, nous prouvons l'impossibilité de détecter une rupture lorsque $\Delta_n = o(n^{1/3})$, où n est la taille du graphe, ce qui constitue un progrès vers une conjecture stipulant que la détection devrait être impossible pour tout $\Delta_n = o(n^{1/2})$. En revanche, dans le cas où le graphe étiqueté est observé, nous établissons un seuil précis : la détection est possible si et seulement si $\Delta_n \to +\infty$.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Title: Inference on dependent data: Contributions to hidden Markov and preferential attachment graph models

Keywords: Dependent data, hidden Markov models, clustering, Bayes risk, preferential attachment, change-point detection

Abstract: This thesis is devoted to the study of statistical inference in settings where data are dependent. Such dependence can arise through temporal dynamics, as in Hidden Markov models, or through structural evolution, as in preferential attachment graph models. The first part focuses on clustering in both i.i.d. and Hidden Markov model (HMM) settings. We analyze the Bayes risk of clustering and compare it to the Bayes risk of classification. Our results show that the minimizers of these risks do not always coincide but we show that this distinction is largely theoretical: in practice, the Bayes classifier performs nearly optimally for clustering tasks. Simulations complement our theoretical findings and illustrate the near-optimality of classifier-based clustering strategies. The second part develops a refined analysis of clustering under Gaussian Hidden Markov models in the regime where the hidden chain mixes slowly. We provide a precise characterization of the Bayes risk in terms of the signal-to-noise ratio and the mixing properties of the chain. Based on this characterization, we propose some Bayes-optimal clustering procedures. Interestingly, our study uncovers surprising behaviors of the Bayes risk in certain parameter regimes, showing that temporal dependence can lead to non-standard phenomena that are absent in the i.i.d. case. The third part turns to change-point detection in growing networks modeled by preferential attachment with time-dependent attachment functions. We consider the late change-point problem where the change occurs at $\tau_n = n - \Delta_n$ where n is the size of the graph and $0 \leq \Delta_n \leq n$. Formulated as a hypothesis testing problem, this leads to strikingly different behaviors depending on whether the labeled or unlabeled graph is observed. For unlabeled graphs, we prove the impossibility of detecting a change when $\Delta_n = o(n^{1/3})$ where n is the size of the graph, thereby making progress toward a conjecture that detection should be impossible for all $\Delta_n = o(n^{1/2})$. By contrast, in the case where the labeled graph is observed, we establish a sharp threshold: detection is possible if and only if $\Delta_n \to +\infty$.

À ma chère famille
إلى عائلتي الغالية

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude à mes encadrants de thèse, Élisabeth et Zacharie, pour m'avoir permis de vivre cette belle aventure. Élisabeth, je te remercie sincèrement de m'avoir initié à la recherche, pour tes relectures attentives, tes conseils avisés et pour ton encadrement sans faille. Merci pour la liberté que tu m'as accordée tout au long de cette thèse et ta vigilance toujours bienveillante. J'ai également beaucoup apprécié nos échanges, parfois éloignés des mathématiques, mais toujours enrichissants et agréables. Zacharie, je te remercie du fond du cœur pour ton encadrement attentif, tes encouragements constants et tes intuitions toujours justes. Merci d'avoir supporté avec patience mes visites fréquentes, souvent à l'improviste, pour parler d'une idée, chercher un conseil ou simplement te déranger. Sache que j'ai énormément appris grâce à ces moments d'échange. Avec un peu de recul, je trouve que vous êtes le duo parfait pour encadrer une thèse, avec chacun des qualités complémentaires de l'autre. Je vous resterai, à tous deux, éternellement reconnaissant.

I would like to thank the referees, Yohann and Chao, for taking the time to read and comment the thesis. I would also like to thank Nicolas, Christine, Catherine, Rui and Christophe for accepting to make part of the jury of my thesis. It is a real pleasure and an honor to have you as members of my defense committee. Un merci tout particulier à Christophe et Nicolas pour vos orientations précieuses, vos conseils toujours pertinents et les échanges stimulants que nous avons eus. Et toi, Simo, ce fut un réel plaisir de travailler avec toi. Merci pour ton orientation, tes conseils et pour ton suivi attentif. Je t'en suis profondément reconnaissant.

Je souhaite également remercier toutes les personnes que j'ai eu la chance de côtoyer au LMO, à commencer par mes collègues de bureau, chacun doté de qualités dont j'espère avoir hérité ne serait-ce qu'un peu. Merci Rana pour ta gentillesse et pour ton aide précieuse chaque fois que je rencontrais une difficulté. Merci Leonardo pour les moments partagés au LMO et pour nos discussions autour des maths. Tu m'as toujours impressionné par ta culture générale. Sans exagérer, je pense que tu es la personne la plus organisée que j'aie jamais rencontrée. J'espère que ta rigueur m'a un peu influencé. Merci également à Victor et Hugo pour votre bonne humeur et les moments partagés au LMO.

Pendant cette thèse, j'ai eu la chance d'organiser le séminaire des étudiants, une expérience formidable. Merci Bastien et Romain pour votre collaboration. Romain, je salue ton audace d'avoir « squatté » mon bureau dès ton premier jour au labo ! Merci pour ta disponibilité chaque fois que je passais te déranger dans ton bureau, que ce soit pour réfléchir ensemble à une question, te piquer un peu de thé, te stresser pour finir ton papier ou simplement discuter. Nos échanges, sur à peu près tous les sujets, ont toujours été un véritable plaisir.

Merci Dhia pour ta gentillesse exceptionnelle, pour les délicieux gâteaux tunisiens que tu prenais le soin de m'apporter, et pour les moments conviviaux passés ensemble, au labo ou ailleurs. Je tiens aussi à remercier Éric, Charly, Cecilia, Guillermo, Vincent, Bertrand, Pierre-André, Chiara, Luca, Daan et Jean-Baptiste pour nos échanges au Cesfo, autour

d'un café ou lors de conférences, toujours agréables et stimulants.

Je tiens également à remercier l'ensemble de la communauté musulmane de CentraleSupélec et de Polytechnique, auprès de qui j'ai beaucoup grandi et mûri, tant sur le plan personnel que spirituel. Ces années passées à vos côtés ont été formidables. J'y ai découvert des personnes d'une gentillesse et d'une générosité qui resteront gravées dans ma mémoire.

Merci Souhail, Anouar, Abdelhaq et Nouamane pour votre amitié fidèle. Merci Mohamed Amine d'avoir supporté mes visites à l'improviste. Merci Nabil pour ta gentillesse hors norme, tes plats marocains toujours au-dessus des attentes et tes attentions constantes. Je me sens privilégié d'avoir un ami qui se soucie autant de mon bien-être. Merci également à Hichem, Taha, Modar et Houssem pour votre gentillesse inégalable. Houdaïfa, ton amitié qui dure depuis tant d'années m'est particulièrement chère. Merci pour ta fidélité et ta bienveillance. Désolé de ne pas encore être venu te rendre visite à Amsterdam malgré tes mille et une invitations. Anas, j'ai énormément apprécié nos discussions autour des mathématiques, de l'économie et de la politique, avec à chaque fois la certitude de me coucher un peu moins bête le soir. Merci de m'avoir initié aux GFlowNets et aux LLMs avec autant de patience. Tu ne peux pas imaginer ma joie de t'avoir enfin convaincu que les statistiques servent bel et bien à quelque chose ! Abdellah, je suis très honoré d'avoir fait ta connaissance. Merci pour ta gentillesse exceptionnelle, tes encouragements dans le sport et ton attention constante à mon bien-être. Te rencontrer a été l'un des plus grands exploits de cette thèse. Enfin, merci à Ilyes, Zakaria, Mouad, Ayoub, Abdelmoughite, Amine, Abdelmonaim et Khalifa pour les moments partagés ensemble. Que notre amitié dure à jamais.

في الختام، أودّ أن أعبّر عن عميق امتناني وعرفاني لأسرتي الكريمة على محبتها اللامحدودة وكرمها الذي لا يُضاهى. أشعر بفخرٍ وسعادةٍ غامرة لكوني محاطًا بكم، ولأنكم كنتم دائمًا السند والداعم الأوّل في حياتي.

إلى والديَّ العزيزين، أقدّم كل حبّي وامتناني على رعايتكما الدائمة، وحبكما الذي لا ينضب، ودعواتكما التي ترافقني أينما ذهبت، وتفانيكما الذي أنار دربي منذ الصغر. تبقى كلمات الشكر عاجزةً عن التعبير عمّا في قلبي من تقديرٍ واعتزازٍ بكما. أما أنتما يا أختاي العزيزتان و أنت يا أخي الغالي، فلكم كل الشكر على محبتكم الصادقة وحرصكم المستمرّ على إسعادي منذ طفولتي. شكرًا على كل لحظة استقبالٍ دافئة عند عودتي إلى طنجة، وعلى الهدايا الجميلة، والأطباق اللذيذة (من إعداد بشرى طبعا)، وعلى كل ما أغدقتموه عليّ من عنايةٍ ومحبّة منقطعة النظير. لقد كنتم ولا تزالون مصدر الفرح والسكينة في حياتي. أحمد الله تعالى على نعمة الانتماء إلى أسرةٍ طيبة القلب مثلكم، وأسأله أن يكتب لكم الخير حيثما كنتم.

وبكل امتنانٍ، أهدي هذا العمل إليكم، أنتم الذين كنتم دائمًا مصدر دعمي وإلهامي.

﴿وَمَا تَوْفِيقِي إِلَّا بِاللَّهِ عَلَيْهِ تَوَكَّلْتُ وَإِلَيْهِ أُنِيبُ﴾ (سورة هود، الآية ٨٨)

Contents

Notations

Measure theory

$(\mathbb{Y}, \mathcal{Y})$	Measurable space
$(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$	Measured space, probability space if $\mathcal{L}$ is a probability measure
$\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$	Hilbert space of measurable and square-integrable functions on $(\mathbb{Y}, \mathcal{Y})$ w.r.t. measure $\mathcal{L}$
$\sigma(X_1, \ldots, X_n)$	σ-algebra generated by the random variables $(X_1, \ldots, X_n)$
$X_n \xrightarrow{\mathbb{P}^n} c$	Convergence in probability to c : $\lim_n \mathbb{P}^n(\lvert X_n - c\rvert > \varepsilon) = 0$ for all $\varepsilon > 0$
$X_n \overset{\mathbb{P}^n}{\rightsquigarrow} X$	Convergence in distribution of $(X_n)_{n\geq 1}$ to a random variable X
$\mathrm{TV}(.,.)$	Total variation distance between two probability distributions

Algebra, functions and others

$f \circ g$	Composition of functions: $(f \circ g)(x) = f(g(x))$
$f \otimes g$	Tensor product of functions: $(f \otimes g)(x, y) = f(x)g(y)$
$\lVert \cdot \rVert_F$	Frobenius norm
$\langle ., . \rangle$	Inner product
$\Phi(\cdot)$	Cumulative distribution function (CDF) of the standard normal distribution: $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$
$\Gamma(\cdot)$	Gamma function: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$, generalizing the factorial function
$\lfloor x \rfloor$	Floor function: greatest integer $\leq x$
$a \wedge b$	Minimum of a and b
$a \vee b$	Maximum of a and b
$a \coloneqq b$, $\quad b =: a$	a is equal to b by definition
$x^+ = x \vee 0$	Positive part of x

$\arg\max$	The set of arguments that maximize a given function
$Y_{a:b}$	Tuple $(Y_a, \dots, Y_b)$; empty if $a > b$
$\binom{n}{k}$	Binomial coefficient: the number of ways to choose k elements from a set of n, defined as $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ for $0 \leq k \leq n$

Sets

$\mathbb{N} := \{0, 1, 2, \dots\}$	Set of nonnegative integers
$\mathbb{N}^* := \{1, 2, \dots\}$	Set of positive integers
$\mathbb{Z}$	Set of all integers
$\mathbb{R}$	Set of real numbers
$\mathbb{R}_+$	Set of nonnegative real numbers
$\lvert A\rvert$, $\#A$, $\mathbf{Card}(A)$	Cardinality of finite set A
$[K]$, $[\![1, K]\!]$	Shorthand for the index set $\{1, 2, \dots, K\}$
$\mathcal{P}[n]$	Set of partitions of $[n]$
$\mathcal{S}_K$	The set of permutations of $[K]$

Comparisons

$a_n \sim b_n$	$a_n/b_n \to 1$; the sequences are asymptotically equivalent
$a_n = o(b_n)$	$a_n/b_n \to 0$; a_n is negligible compared to b_n
$a_n = O(b_n)$	a_n/b_n is bounded
$a_n \lesssim b_n$	Equivalent to $a_n = O(b_n)$
$a_n \asymp b_n$	There exist $\alpha, \beta > 0$ such that $\alpha a_n \leq b_n \leq \beta a_n$; two-sided boundedness

Chapter 1

Introduction (FR)

Contents

Cette thèse aborde deux problèmes relativement indépendants d'inférence statistique, tous deux centrés sur la thématique de l'inférence à partir de données dépendantes. Elle est divisée en deux grandes parties. La première partie est consacrée au problème du regroupement (clustering) dans les modèles de mélange paramétriques et non-paramétriques, et concerne le Chapitre 3 et le Chapitre 4. La seconde partie traite du problème de détection de ruptures dans le cadre du modèle de graphe aléatoire à attachement préférentiel, et concerne le Chapitre 5.

Les modèles à variables latentes sont des modèles statistiques qui supposent la présence de variables non observées influençant les données observées. Une définition formelle est donnée ci-dessous.

Définition 1.0.1. *Un modèle à variables latentes est un processus stochastique bivarié $(X_t, Y_t)_{t\in\mathbb{N}}$ où seules les observations $(Y_t)_{t\in\mathbb{N}}$ sont accessibles. Les variables aléatoires $(X_t)_{t\in\mathbb{N}}$ sont appelées variables cachées (ou latentes) et ne sont pas observées.*

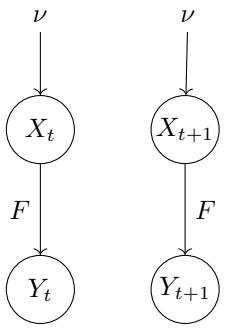
1.1 Modèles de mélange et modèles de Markov cachés

Les modèles de mélange constituent une famille particulière de modèles à variables latentes dans lesquels on suppose que chaque observation est générée à partir d'**une seule** composante latente, chacune étant associée à une distribution spécifique. Les modèles de mélange apparaissent naturellement lorsque les observations proviennent de plusieurs groupes ayant chacun leurs propres caractéristiques. La variable latente identifie alors le groupe auquel un individu appartient. Dans cette thèse, nous étudions des modèles dans lesquels les variables latentes prennent un nombre fini de valeurs. C'est le cas des modèles de mélange finis qui nous intéressent, que nous définissons comme suit.

Définition 1.1.1 (Modèle de mélange fini)**.** *Soit $\mathbb{X}$ un espace fini et $(\mathbb{Y}, \mathcal{Y})$ un espace mesurable. Soit ν une distribution de probabilité sur $\mathbb{X}$ et $F = (F_x)_{x\in\mathbb{X}}$ un vecteur de distributions de probabilité sur $\mathbb{Y}$. Un processus $(Y_t)_{t\in\mathbb{N}}$ suit un modèle de mélange avec paramètre $\theta = (K, \nu, F)$ s'il existe une suite de variables aléatoires (non observées) $(X_t)_{t\in\mathbb{N}}$ telle que :*

- *$(X_t)_{t\in\mathbb{N}}$ sont i.i.d. suivant la loi ν, à valeurs dans $\mathbb{X}$ avec $|\mathbb{X}| = K$;*
- *Conditionnellement à $(X_t)_{t\in\mathbb{N}}$, les variables $(Y_t)_{t\in\mathbb{N}}$ sont indépendantes ;*
- *Pour tout $t \in \mathbb{N}$, conditionnellement à $\{X_t = x\}$, Y_t suit la loi F_x.*

Dans ce cas, les variables $(Y_t)_{t\in\mathbb{N}}$ sont i.i.d. suivant la distribution $\mathbb{P}_\theta = \sum_{x\in\mathbb{X}} \nu_x F_x$ où $\nu_x = \nu(\{x\})$.



(b) Modèle de mélange avec paramètres (ν, F)

Figure 1.1: Graphe acyclique dirigé représentant un modèle de mélange. Les étiquettes des arêtes indiquent les noyaux de transition.

Au-delà de leur pouvoir descriptif, les modèles de mélange constituent un outil généraliste en inférence statistique, servant de base à des problèmes comme l'estimation de paramètres, le regroupement non supervisé, la classification supervisée, la détection de

ruptures et la segmentation. Leur polyvalence leur a permis d'être largement utilisés dans des domaines tels que la bioinformatique, l'économétrie ou encore le traitement du signal. Les références fondamentales incluent McLachlan and Peel [2000] pour la théorie classique, Frühwirth-Schnatter [2006] pour les approches bayésiennes, et McLachlan and Basford [1988] pour les applications en classification. Les développements récents sont présentés dans le *Handbook of Mixture Analysis* Frühwirth-Schnatter et al. [2019], qui souligne leur rôle dans l'inférence moderne.

Un autre modèle à variables latentes d'intérêt particulier dans cette thèse est le modèle de Markov caché.

Définition 1.1.2 (Modèle de Markov caché)**.** *On dit qu'un processus bivarié* $(X_t, Y_t)_{t\in\mathbb{N}}$ *suit un modèle de Markov caché si :*

- *Le processus* $(X_t)_{t\in\mathbb{N}}$ *est une chaîne de Markov ;*
- *Conditionnellement à* $(X_t)_{t\in\mathbb{N}}$*, les variables* $(Y_t)_{t\in\mathbb{N}}$ *sont indépendantes ;*
- *Conditionnellement à* $(X_t)_{t\in\mathbb{N}}$*, la loi de* Y_t *dépend uniquement de* X_t*.*

Les deux derniers points se traduisent par :

$$\mathcal{L}\left((Y_t)_{t\in\mathbb{N}} \mid (X_t)_{t\in\mathbb{N}}\right) = \bigotimes_{t\in\mathbb{N}} \mathcal{L}\left(Y_t \mid X_t\right)$$

où $\mathcal{L}(.)$ *désigne la loi.*

Dans cette thèse, nous nous intéressons aux chaînes de Markov homogènes à espace d'états fini. Voir Cappé et al. [2005] pour des modèles de Markov cachés plus généraux.Dans le cas où la chaîne de Markov cachée est homogène et prend un nombre fini de valeurs, la définition précédente s'énonce ainsi.

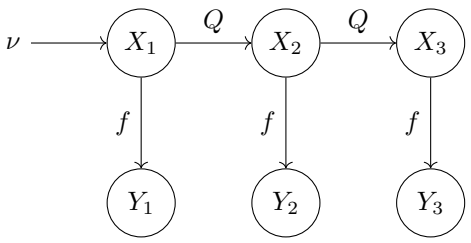
Définition 1.1.3 (Modèle de Markov caché homogène à espace d'états fini)**.** *Soit* $\mathbb{X}$ *un ensemble fini et* $(\mathbb{Y}, \mathcal{Y})$ *un espace mesurable. Soient* $(\nu_x)_{x\in\mathbb{X}}$ *une distribution de probabilité sur* $\mathbb{X}$*,* $Q = (Q_{x,x'})_{x,x'\in\mathbb{X}}$ *une matrice de transition, et* $F = (F_x)_{x\in\mathbb{X}}$ *un vecteur de lois de probabilité sur* $(\mathbb{Y}, \mathcal{Y})$*. Le processus* $(X_t, Y_t)_{t\in\mathbb{N}}$ *suit un modèle de Markov caché (HMM) avec paramètres* $\theta = (K, \nu, Q, f)$ *si :*

- *Le processus* $(X_t)_{t\in\mathbb{N}}$ *est une chaîne de Markov homogène à valeurs dans* $\mathbb{X}$*, de loi initiale* ν *et noyau de transition* Q*, avec* $|\mathbb{X}| = K$ *;*
- *Conditionnellement à* $(X_t)_{t\in\mathbb{N}}$*, les variables* $(Y_t)_{t\in\mathbb{N}}$ *sont indépendantes ;*
- *Pour tout* $t \in \mathbb{N}$*, conditionnellement à* $\{X_t = x\}$*,* Y_t *suit la loi* F_x*.*

L'ensemble $\mathbb{X}$ *est appelé* espace d'états *du HMM, et* $\mathbb{Y}$ *l'*espace des observations*. Enfin, la famille de distributions* $(F_x)_{x\in\mathbb{X}}$ *est appelée lois d'émission.*

Le modèle HMM permet de modéliser une dépendance entre les observations grâce à la structure markovienne des variables cachées $(X_t)_{t\in\mathbb{N}}$. Il est à noter que dans le cas où toutes les lignes de la matrice de transition Q sont identiques et où la distribution initiale ν est la distribution stationnaire, la structure de dépendance disparaît et le modèle se ramène au modèle de mélange défini à la Définition 1.1.1. Introduits dans les travaux pionniers de Leonard E. Baum et de ses collaborateurs dans les années 1960 Baum and Petrie [1966], comme cadre mathématique pour analyser les séquences générées par des processus stochastiques sous-jacents non observables, les HMM ont ensuite été largement

étudiés d'un point de vue théorique et utilisés dans de nombreuses applications : de la reconnaissance vocale Rabiner [1989] à la modélisation climatique Khiatani and Ghose [2017], en passant par la génomique Scharpf et al. [2008], Qin et al. [2010] et la finance Guo et al. [2023], Pakštaitė et al. [2025]. Leur succès repose sur leur simplicité, leur flexibilité, leur interprétabilité et leur facilité de mise en œuvre. Rabiner and Juang [1986] propose une introduction simple aux modèles de Markov cachés. Pour un traitement plus approfondi et moderne, le livre Cappé et al. [2005] offre une vue d'ensemble complète des aspects théoriques et pratiques.



(a) Modèle de Markov caché avec paramètres (ν, Q, f)

Figure 1.2: Graphe acyclique dirigé représentant un modèle de Markov caché. Les étiquettes des arêtes indiquent les noyaux de transition.

La principale distinction entre les modèles de mélange et les modèles de Markov cachés découle de leur structure de dépendance sous-jacente. Bien que cette différence puisse sembler simple, elle a des implications importantes pour l'analyse de divers problèmes d'inférence dans les deux contextes. Une partie de cette thèse est consacrée à l'exploration de certains problèmes d'inférence dans ces deux cadres.

1.2 Identifiabilité : Une condition préalable pour une inférence bien posée

Une question naturelle à poser avant de réaliser toute tâche d'inférence sur un modèle à variable latente est : étant donné un processus observé $(Y_t)_{t\in\mathbb{N}}$, est-il possible de définir de manière unique le modèle à variable latente qui le génère ? Autrement dit, dans quelles conditions peut-on garantir l'unicité, au sens où le paramètre qui produit ce processus est unique ? Si une telle unicité existe, on dit que le modèle est identifiable. Avant de définir l'identifiabilité, définissons la notion de modèle statistique.

Définition 1.2.1 (Modèle statistique)**.** *Un modèle statistique est une famille paramétrée de distributions de probabilités $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ définie sur un espace mesurable commun.*

Nous définissons d'abord l'identifiabilité dans le cas général.

Définition 1.2.2. *Un modèle statistique $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ est identifiable quand l'application $\theta \mapsto \mathbb{P}_\theta$ est injective.*

1.2.1 Identifiabilité dans le modèle de mélange

Tout d'abord, on observe que dans le modèle de mélange, la distribution des observations reste inchangée sous toute permutation des états cachés. Plus précisément, soit σ une permutation de l'espace latent $\mathbb{X}$, c'est-à-dire une bijection de $\mathbb{X}$ sur lui-même. Pour un tuple de paramètres $\theta = (K, \nu, F)$, définissons la version permutée $\theta' = \big(K, (\nu_{\sigma(x)})_{x\in\mathbb{X}}, (F_{\sigma(x)})_{x\in\mathbb{X}}\big)$.

Les paramètres θ et θ' génèrent le même mélange puisque $\sum_{x\in\mathbb{X}} \nu_x F_x = \sum_{x\in\mathbb{X}} \nu_{\sigma(x)} F_{\sigma(x)}$ et ainsi, les labels des composantes du mélange ne peuvent jamais être récupérées. C'est la raison pour laquelle l'identifiabilité dans le modèle de mélange est définie à une permutation près des labels. Nous définissons l'identifiabilité dans les modèles de mélange comme suit.

Définition 1.2.3 (Identifiabilité dans les modèles de mélange)**.** *Nous identifions* $\mathbb{X}$ *avec* $[K] \coloneqq \{1, \dots, K\}$ *lorsque* $|\mathbb{X}| = K$*. Définissons le modèle de mélange*

$$\mathcal{P}_\Theta \coloneqq \left\{ \sum_{x\in[K]} \nu_x F_x \;\middle|\; \theta = (K, \nu, F) \in \Theta \right\}.$$

Le modèle $\mathcal{P}_\Theta$ *est dit* identifiable *(à une permutation des labels près) si pour tous* $\theta = (K, \nu, F)$ *et* $\theta' = (K', \nu', F')$ *dans* Θ

$$\sum_{x\in[K]} \nu_x F_x = \sum_{x\in[K']} \nu'_x F'_x \quad \Longrightarrow \quad \theta = \sigma(\theta')$$

pour une permutation σ *de* $[K]$*, où*

$$\sigma(\theta') = \left(K', (\nu'_{\sigma(x)})_{x\in[K']}, (F'_{\sigma(x)})_{x\in[K']}\right).$$

Notez que l'identifiabilité des modèles de mélange est définie en utilisant uniquement la distribution marginale d'une seule observation. Cela est dû au fait que les données sont i.i.d. dans le modèle de mélange, ce qui implique que la distribution jointe de plusieurs observations ne porte aucune information supplémentaire pour l'identifiabilité. Elle correspond simplement au produit de distributions marginales identiques. Un inconvénient fondamental des modèles de mélange est que, sans hypothèses supplémentaires sur les composants du mélange, le modèle n'est pas identifiable. Par exemple, on peut dupliquer un composant en le divisant en deux composants avec des distributions identiques et la moitié du poids d'origine, ce qui donne un mélange statistiquement indiscernable de l'original, mais avec des paramètres différents et même un nombre de composants différent. En l'absence d'hypothèses supplémentaires sur les composants du mélange, l'analyse de tout problème d'inférence devient mal posée, car les paramètres du modèle ne sont pas définis de manière unique. Bien que l'identifiabilité ne puisse pas être attendue en général, des travaux récents ont établi l'identifiabilité pour certaines classes structurées de modèles. Nous passons en revue ci-dessous certains contextes où l'identifiabilité est garantie et où l'inférence statistique est donc significative.

Mélange fini de distributions paramétriques

Considérons les mélanges de la forme

$$\sum_{x\in[K]} \nu_x F_x,$$

où les distributions $(F_x)_{x\in[K]}$ appartiennent à une famille paramétrique. Teicher [1963] démontre l'identifiabilité de tous les mélanges finis de lois Gamma et gaussiennes. Dans la continuité de ce travail, Barndorff-Nielsen [1965] montre qu'un mélange de distributions appartenant à la même famille exponentielle paramétrique est identifiable sous de faibles conditions de régularité. Yakowitz and Spragins [1968] étend ces résultats d'identifiabilité à plusieurs familles, incluant la famille des produits de n lois exponentielles, la famille gaussienne multivariée, leur union, la famille des lois de Cauchy unidimensionnelles, ainsi que d'autres.

Translation d'une densité symétrique

Nous considérons ici les mélanges de la forme

$$\sum_{x\in[K]} \nu_x F(\cdot - \mu_x),$$

où F est une fonction de répartition symétrique, c'est-à-dire telle que

$$\forall x \in \mathbb{R}, \quad F(-x) + F(x) = 1.$$

Bordes et al. [2006] établissent l'identifiabilité non paramétrique du modèle dans le cas $K = 2$. Pour un nombre plus élevé de composantes, Hunter et al. [2007] donnent une condition garantissant l'identifiabilité des paramètres du modèle $(\nu_x)_{x\in[K]}$ et $(\mu_x)_{x\in[K]}$. Plus précisément, le théorème suivant est démontré.

Théorème 1.2.1 (Hunter et al. [2007]). *Les paramètres* $\left((\nu_x)_{x\in[K]}, (\mu_x)_{x\in[K]}\right)$ *peuvent être identifiés si et seulement si l'équivalence suivante est vraie :*
Pour tous $(\eta'_x)_{x\in[K]}$, $(\nu'_x)_{x\in[K]}$, *la convolution*

$$\sum_{x\in[K]} \nu_x \delta_{\mu_x} \star \sum_{x\in[K]} \nu'_x \delta_{-\mu'_x}$$

est symétrique si et seulement si

$$\sum_{x\in[K]} \nu'_x \delta_{-\mu'_x} = \sum_{x\in[K]} \nu_x \delta_{-\mu_x}.$$

Ils en déduisent que, lorsque $K = 2$, l'identifiabilité des paramètres est assurée si et seulement si $\nu_1 \notin \{0, 1/2, 1\}$. Bien que la classe des densités symétriques soit adaptée dans certains contextes, elle s'avère souvent trop restrictive pour rendre compte de la complexité de nombreux modèles pratiques. Il apparaît en revanche que l'introduction d'une forme spécifique de dépendance entre les observations permet de rétablir l'identifiabilité pour les mélanges de translations, à condition de considérer la loi jointe de deux observations consécutives. Voir la Remarque 1.2.5.

Mélanges multidimensionnels

Nous considérons à présent les modèles de mélange multidimensionnels, où les variables observées sont réparties en au moins $d \geq 3$ blocs. Dans ce cadre, la loi jointe des observations peut s'écrire

$$\sum_{x\in[K]} \nu_x \bigotimes_{j=1}^{d} F_{x,j}, \tag{1.1}$$

où, pour chaque $x \in [K]$, les distributions $(F_{x,j})_{j\in[d]}$ sont des mesures de probabilité définies sur d espaces éventuellement distincts.

Les premiers résultats d'identifiabilité sont apparus dans la littérature pour des cas de faible dimension, tels que $K = 2$ et $d = 2$ ou $d = 3$ Hall and Zhou [2003]. D'autres travaux ont ensuite considéré la situation où K est inconnu et $d \geq 2$ Hall et al. [2005]. Un résultat fondateur de Kruskal [1977] établit que le modèle (1.1) est identifiable lorsque $d = 3$, sous l'hypothèse que les mesures de probabilité ont un support fini. Dans la continuité, Allman et al. [2009] ont montré que l'identifiabilité est assurée pour tout $d \geq 3$, à condition que, pour chaque $j = 1, \ldots, d$, la famille de mesures $F_{1,j}, \ldots, F_{K,j}$ soit linéairement indépendante. Enfin, certains arguments spectraux jouent un rôle central dans la démonstration de l'identifiabilité pour d'autres modèles Anandkumar et al. [2012], Gassiat et al. [2016], ainsi que dans la construction d'estimateurs Abraham et al. [2025, 2022].

1.2.2 Identifiabilité dans le modèle de Markov caché

Tout d'abord, il est important de noter que, comme dans les modèles de mélange, la loi des observations reste inchangée par une permutation des états cachés. Plus précisément, soit $\theta = (K, \nu, Q, F)$ et soit $\theta' = (K, \nu', Q', F')$ où

$$\big(\forall (x, x') \in \mathbb{X}\big) \quad \nu'_x = \nu_{\sigma(x)}, \quad Q'_{x,x'} = Q_{\sigma(x),\sigma(x')}, \quad F'_x = F_{\sigma(x)}.$$

Alors, en notant $\mathbb{P}_\theta^{(n)}$ la loi jointe de $(Y_i)_{i\in[n]}$ et en utilisant les propriétés du modèle de Markov caché, on peut montrer facilement que

$$(\forall y_{1:n} \in \mathbb{Y}^n)\ \ \mathbb{P}_\theta^{(n)}(y_{1:n}) = \sum_{x_{1:n}\in\mathbb{X}^n} \nu_{x_1} Q_{x_1,x_2} \cdots Q_{x_{n-1},x_n} \prod_{i=1}^{n} F_{x_i}(y_i). \tag{1.2}$$

D'après l'expression ci-dessus, on a directement $(\forall y_{1:n} \in \mathbb{Y}^n)\ \mathbb{P}_\theta^{(n)}(y_{1:n}) = \mathbb{P}_{\theta'}^{(n)}(y_{1:n})$. Cela signifie que, même en utilisant la loi de plusieurs observations, les labels des composantes ne peuvent pas être identifiés. Par conséquent, les résultats d'identifiabilité doivent toujours être interprétés modulo ces permutations des états cachés, sauf si un étiquetage canonique est disponible. Ainsi, dans ce qui suit, lorsque $|\mathbb{X}| = K$, nous supposerons sans perte de généralité que $\mathbb{X} = [K]$. Nous définissons l'identifiabilité dans le modèle de Markov caché comme suit.

Définition 1.2.4 (Identifiabilité dans les modèles de Markov cachés)**.** *Nous identifions* $\mathbb{X}$ *avec* $[K] \coloneqq \{1, \dots, K\}$ *lorsque* $|\mathbb{X}| = K$*. On définit le modèle de Markov caché*

$$\mathcal{P}_\Theta \coloneqq \left\{ \mathbb{P}_\theta^{(n)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}.$$

où $\mathbb{P}_\theta^{(n)}$ *est la loi jointe de* $Y_{1:n}$ *sous le HMM de paramètre* θ*. On dit que* $\mathcal{P}_\Theta$ *est* identifiable *(à une permutation des labels près) à partir de la loi de* n *observations si, pour tout* $\theta = (K, \nu, Q, F)$ *et* $\theta' = (K', \nu', Q', F')$ *dans* Θ*,*

$$\mathbb{P}_\theta^{(n)} = \mathbb{P}_{\theta'}^{(n)} \quad \Longrightarrow \quad (\exists \sigma \in \mathcal{S}_K)\ \ \theta = \sigma(\theta')$$

où

$$\sigma(\theta') \coloneqq \left(K, \left(\nu'_{\sigma(x)}\right)_{x\in[K]}, \left(Q'_{\sigma(x),\sigma(x')}\right)_{x,x'\in[K]}, \left(F'_{\sigma(x)}\right)_{x\in[K]} \right).$$

Il est important de noter que notre intérêt porte sur l'identifiabilité à partir de la loi jointe de plusieurs observations, car la structure markovienne cachée ne devient utile que dans ce contexte. Lorsqu'on considère une seule observation, aucun résultat d'identifiabilité au-delà de ceux des modèles de mélange n'est à attendre.

Les résultats fondateurs sur l'identifiabilité des modèles de Markov cachés (HMM) remontent aux travaux de Baum and Petrie [1966], Petrie [1969], qui ont établi l'identifiabilité pour les HMM discrets. Ils ont montré que, lorsque le nombre d'états cachés K est connu et que les distributions d'émission sont linéairement indépendantes, le modèle est identifiable. Cependant, ces résultats étaient limités au cas des alphabets d'observation finis.

Une avancée majeure est intervenue avec l'application des méthodes de décomposition tensorielle, en particulier le théorème de Kruskal Kruskal [1977] à l'identifiabilité des HMM. Allman et al. [2009] ont appliqué ce théorème afin de démontrer l'identifiabilité de certains HMM discrets (dans un sens légèrement différent de la définition 2.2.5), sous des conditions de rang modérées. Cette approche a introduit un cadre algébrique puissant pour l'étude de

l'identifiabilité, consistant à traiter la loi jointe sur un nombre fini d'observations comme un tenseur de faible rang, et à exploiter l'unicité des décompositions tensorielles pour identifier les paramètres.

Des travaux ultérieurs ont étendu les résultats d'identifiabilité au cadre non paramétrique. En particulier, Gassiat et al. [2016] ont démontré l'identifiabilité du HMM non paramétrique sous l'hypothèse que les distributions d'émission sont distinctes et linéairement indépendantes. Ils ont établi le théorème suivant.

Théorème 1.2.2 (Gassiat et al. [2016]). *Soit $(X_t, Y_t)_{t\in\mathbb{N}}$ un HMM de paramètre $\theta \in \Theta$. On suppose que Θ est l'ensemble des paramètres θ tels que :*

- $(\forall x \in [K]) \quad \nu_x > 0$,
- *Q est inversible,*
- *Les mesures de probabilité $(F_x)_{x\in[K]}$ sont linéairement indépendantes,*

alors, le modèle $\mathcal{P}_\Theta = \left\{ \mathbb{P}_\theta^{(3)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}$ est identifiable, où $\mathbb{P}_\theta^{(3)}$ désigne la loi de (Y_1, Y_2, Y_3).

Autrement dit, sous des hypothèses de régularité légères, les paramètres du HMM peuvent être identifiés à partir de la loi jointe de trois observations consécutives. L'hypothèse d'indépendance linéaire peut être relâchée à la condition plus simple que les lois d'émission soient distinctes entre elles, au prix de considérer la loi jointe de plus de trois observations, comme détaillé dans le théorème suivant.

Théorème 1.2.3 (Alexandrovich et al. [2016b]). *Soit $(X_t, Y_t)_{t\in\mathbb{N}}$ un HMM de paramètre $\theta \in \Theta$. On suppose que Θ est l'ensemble des paramètres θ tels que :*

- *Q est inversible, irréductible et apériodique (Voir Norris [1997] pour une définition formelle),*
- *Les mesures de probabilité $(F_x)_{x\in[K]}$ sont distinctes entre elles,*

alors, le modèle $\mathcal{P}_\Theta = \left\{ \mathbb{P}_\theta^{((2K+1)(K^2-2K+2)+1)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}$ est identifiable au sens de la définition 1.2.4.

Remark 1.2.5. Il est important de noter que l'identifiabilité est également assurée dans d'autres modèles qui ne relèvent pas du cadre des modèles de mélange ni des modèles de Markov cachés. Considérons par exemple des observations de loi marginale $\sum_{x\in[K]} \nu_x F(\cdot - m_x)$ mais qui ne sont pas indépendantes. Plus précisément, supposons que les observations soient issues du modèle

$$Y_i = m_{X_i} + \varepsilon_i, \tag{1.3}$$

où $(\varepsilon_i)_{i\in\mathbb{N}}$ est une suite de variables aléatoires i.i.d. à valeurs réelles, et $(m_j)_{j\in[K]}$ sont des réels.

Dans Gassiat and Rousseau [2016], les auteurs montrent que, lorsque les variables latentes $(X_i)_{i\in\mathbb{N}}$ ne sont pas indépendantes, le modèle (1.3) est identifiable sans aucune hypothèse sur F. Plus précisément, si le processus $(X_i)_{i\in\mathbb{N}}$ prend K valeurs distinctes et si F est une loi de probabilité quelconque, alors, à condition que les paramètres de translation soient distincts et que la matrice Q représentant la loi jointe de (X_1, X_2) soit de rang plein, il est possible de retrouver $\left(K, Q, (\nu_x)_{x\in[K]}, (m_x)_{x\in[K]}, F\right)$ à partir de la loi de (Y_1, Y_2).

Bien entendu, les centres $(m_x)_{x\in[K]}$ ne sont identifiables qu'à une translation près, sauf si l'un des centres est fixé à l'avance. Notons également qu'aucune hypothèse n'est imposée

sur F. La seule hypothèse structurelle requise est que Q soit de rang plein. Enfin, dans le cas où il n'y a que deux états latents ($K = 2$), l'hypothèse que Q soit de rang plein revient à supposer que les variables X_1 et X_2 sont dépendantes, ce qui constitue une condition très simple d'identifiabilité.

Les conditions classiques d'identifiabilité pour les modèles de mélange et les modèles de Markov cachés étant désormais établies, le problème de l'inférence des paramètres est bien posé. La section suivante présente un panorama des méthodes d'estimation et des algorithmes développés pour ces deux classes de modèles.

1.3 Estimation dans les modèles de mélange et les modèles de Markov cachés

Dans cette section, nous passons en revue les principales techniques d'estimation utilisées pour le modèle de mélange fini et le modèle de Markov caché.

1.3.1 L'algorithme d'Espérance-Maximization (EM)

Soit $\theta \in \Theta$ le paramètre du modèle et $\mathbb{P}_\theta^{(n)}$ la loi jointe des observations $Y = (Y_i)_{i\in[n]}$. Soit $X = (X_i)_{i\in[n]}$ les variables latentes associées. On note la vraisemblance des données observées sous le paramètre θ par $L_n(\theta; Y)$. Un estimateur naturel de θ dans le cadre paramétrique est l'estimateur du maximum de vraisemblance. L'algorithme d'Espérance-Maximization (EM), introduit par Dempster et al. [1977], est une procédure itérative largement utilisée pour approximer l'estimateur du maximum de vraisemblance lorsque la maximisation directe de la vraisemblance n'est pas réalisable. Cela est généralement le cas dans les modèles à variables latentes, incluant les modèles de mélange et les modèles de Markov cachés. L'estimateur du maximum de vraisemblance (MLE) est défini par :

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta, Y).$$

Comme les variables latentes X ne sont pas observées, la maximisation directe est en général difficile. L'algorithme EM contourne cette difficulté en maximisant itérativement l'espérance de la log-vraisemblance complète, définie à travers les étapes suivantes à partir d'un point initial $\theta^{(0)} \in \Theta$:

- **E-step (Espérance) :** Calculer la quantité intermédiaire :

$$R(\theta, \theta^{(j)}) = \mathbb{E}_{\theta^{(j)}}[\log L_n(\theta, (X, Y)) \mid Y],$$

  qui est l'espérance conditionnelle de la log-vraisemblance complète étant donné les données observées Y, sous l'estimation courante du paramètre $\theta^{(j)}$.

- **M-step (Maximization) :** Mettre à jour le paramètre en maximisant cette espérance :

$$\theta^{(j+1)} \in \arg\max_{\theta\in\Theta} R(\theta, \theta^{(j)}).$$

Ces étapes sont répétées jusqu'à convergence, typiquement lorsque l'augmentation relative de la vraisemblance devient inférieure à une tolérance $\varepsilon > 0$:

$$\frac{L_n(\theta^{(j)}, Y) - L_n(\theta^{(j-1)}, Y)}{L_n(\theta^{(j-1)}, Y)} < \varepsilon.$$

La sortie de l'algorithme est alors $\theta^{(j)}$. Wu [1983] démontre que, sous certaines conditions de régularité, la suite des itérés converge vers un point stationnaire de la fonction de vraisemblance. Cependant, l'algorithme EM ne garantit pas la convergence vers un maximum global. Il peut se bloquer dans des maxima locaux sous-optimaux, en particulier dans des espaces de grande dimension ou des problèmes mal conditionnés. Puisque l'algorithme est déterministe, son résultat dépend fortement du paramètre initial $\theta^{(0)}$. Une stratégie pratique courante consiste à exécuter l'algorithme EM plusieurs fois à partir d'initialisations différentes et à conserver la solution ayant la vraisemblance la plus élevée. Pour les lois paramétriques usuelles, les étapes E et M admettent souvent des expressions simplifiées, permettant une implémentation efficace via des mises à jour récursives explicites. Dans le cas des HMM, les deux étapes de l'algorithme EM admettent une formule explicite. Toutefois, ces formules récursives dépendent de la loi *a posteriori* des états cachés, qui peut être calculée à l'aide du célèbre algorithme Forward-Backward Cappé et al. [2005]. Pour des garanties théoriques rigoureuses sur la convergence de cet algorithme, voir Balakrishnan et al. [2017].

1.3.2 Estimation spectrale

Les algorithmes spectraux constituent une classe de méthodes qui estiment les paramètres des modèles à variables latentes en exploitant la structure algébrique des tenseurs ou matrices de moments d'ordre faible, tels que les moments d'ordre deux et trois. Ces méthodes évitent généralement l'optimisation non convexe en s'appuyant sur des outils d'algèbre linéaire, tels que les décompositions en valeurs propres, en valeurs singulières, et la décomposition tensorielle. Des développements récents, en particulier dans Anandkumar et al. [2012], ont utilisé cette technique pour l'estimation des paramètres de certains modèles de Markov cachés paramétriques. Cette technique a ensuite été affinée afin de couvrir également l'estimation dans le cadre des modèles de Markov cachés non-paramétriques. Plus précisément, considérons un HMM non-paramétrique de paramètres $\theta = (K, \nu, Q, F)$ et supposons que ν est la loi stationnaire de la matrice de transition Q de la chaîne de Markov cachée et que les lois d'émission admettent des densités par rapport à une mesure dominante commune $\mathcal{L}$ sur $(\mathbb{Y}, \mathcal{Y})$. Soit $F = (F_k)_{k\in[K]}$ où $F_k = f_k d\mathcal{L}$. Notons $\left(\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L}), \|.\|\right)$ l'espace de Hilbert des fonctions à carré intégrable sur $\mathbb{Y}$ par rapport à la mesure $\mathcal{L}$, muni du produit scalaire usuel $\langle ., . \rangle$ sur $\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$. Soit $(\varphi_a)_{a\in\mathbb{N}}$ une base orthonormée de $\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$. Pour $a, b, c \in \mathbb{N}$, considérons :

$$\begin{aligned}
\mathbf{L}(a) &= \mathbb{E}\left[\varphi_a(Y_1)\right] \\
\mathbf{N}(a, b) &= \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\right] \\
\mathbf{P}(a, c) &= \mathbb{E}\left[\varphi_a(Y_1)\varphi_c(Y_3)\right] \\
\mathbf{M}(a, b, c) &= \mathbb{E}\left[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)\right]
\end{aligned}$$

En utilisant l'Équation (1.2), on peut exprimer les quantités ci-dessus en termes de la loi initiale ν, de la matrice de transition Q et des coordonnées des projections des lois d'émission sur la base $(\varphi_a)_{a\in\mathbb{N}}$. À l'aide d'opérations algébriques simples telles que les diagonalisation et décomposition en valeurs singulières, et en exploitant les hypothèses d'identifiabilité, ces expressions peuvent être inversées et les paramètres du modèle exprimés en fonction de $\mathbf{L}$, $\mathbf{N}$, $\mathbf{P}$ et $\mathbf{M}$. L'estimation est alors obtenue en remplaçant les lois jointes par leurs contreparties empiriques dérivées des données. Cette approche présente l'avantage majeur d'être non itérative, évitant ainsi les optima locaux et la sensibilité à l'initialisation propres aux méthodes de type EM. Ce cadre constitue la base des procédures d'estimation développées dans Anandkumar et al. [2012], Gassiat et al. [2016], Abraham et al. [2022], De Castro et al. [2017], où la consistance de l'estimation est assurée par des

techniques spectrales et algébriques. Plus récemment, Abraham et al. [2022] ont introduit un estimateur spectral à noyau qui affine les méthodes spectrales précédentes afin de contrôler l'erreur d'estimation en norme suprême plutôt qu'en norme $\mathbf{L}^2$. Ce raffinement vise spécifiquement le cas des modèles de Markov cachés à deux états latents. Les auteurs établissent le théorème suivant.

Theorem 1.3.1. *Sous des hypothèses standards garantissant l'identifiabilité du HMM et en supposant que les densités d'émission soient s-Hölder régulières, il existe un estimateur* $\left(\hat{f}_j\right)_{j\in[K]}$ *et une permutation* τ *tels que, pour une constante* C *suffisamment grande,*

$$\mathbb{P}_\theta\left(\|\hat{f}_j - f_{\tau(j)}\|_\infty \geq C\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right) \xrightarrow[n\to+\infty]{} 0.$$

De plus, la convergence en espérance est également vérifiée : pour une constante $C' > 0$*,*

$$\mathbb{E}\left[\|\hat{f}_j - f_{\tau(j)}\|_\infty\right] \leq C'\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}.$$

Une version raffinée de cet estimateur sera utilisée au Chapitre 3 pour la construction d'une procédure de regroupement par substitution (plug-in). Dans le même cadre des HMM non paramétriques à deux états latents, Abraham et al. [2025] ont proposé un estimateur par seuillage en blocs d'ondelettes, montré comme étant adaptatif à la régularité de chaque densité.

1.3.3 Estimation des moindres carrés pénalisés

Le paradigme des moindres carrés permet la construction de procédures d'estimation minimax adaptatives dans un cadre non-paramétrique. Dans cette section, nous décrivons une méthode d'estimation des paramètres du Modèle de Markov Caché (HMM) en utilisant l'approche des moindres carrés pénalisés. Définissons :

$$(\forall k \in [K])\,(\forall M \in \mathbb{N}^\star) \quad f_{M,k} = \sum_{m=1}^{M} \langle f_k, \varphi_m\rangle \varphi_m$$

Soit $\theta = (K, \nu, Q, (f_k)_{k\in[K]})$ l'ensemble des paramètres d'un HMM. Notons g_θ la densité jointe d'un triplet d'observations (Y_1, Y_2, Y_3) :

$$g_\theta(y_1, y_2, y_3) = \sum_{k_1,k_2,k_3=1}^{K} \pi_{k_1} Q_{k_1,k_2} Q_{k_2,k_3} f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3)$$

Pour toute fonction de carré intégrable $t \in \mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$, on définit la fonctionnelle de contraste :

$$C(t) = \|t - g_\theta\|_2^2 = \|t\|_2^2 - 2\langle t, g_\theta\rangle + \|g_\theta\|_2^2.$$

qui est minimale pour $t = g_\theta$. Comme $\|g_\theta\|_2^2$ est constante en t, minimiser $C(t)$ revient à minimiser $\|t\|_2^2 - 2\langle t, g_\theta\rangle$. Étant donné un échantillon $(Y_s)_{s\in[N+2]}$ d'observations d'un HMM, on considère donc la fonction de contraste empirique :

$$\gamma_N(t) = \|t\|_2^2 - \frac{2}{N}\sum_{s=1}^{N} t(Y_s, Y_{s+1}, Y_{s+2}),$$

Supposons que l'on dispose d'un estimateur $\hat{Q}$ de Q. Par exemple, on peut utiliser l'estimateur spectral de la section précédente. En se basant sur $\hat{Q}$, on considère $\mathcal{S}(\hat{Q}, M)$,

la collection des fonctions g_θ telles que $\theta = \left(K, \hat{\nu}, \hat{Q}, f_M\right)$, où $\hat{\nu}$ est la loi stationnaire de $\hat{Q}$ et $f_M = (f_{M,k})_{k\in[K]}$. On définit alors

$$\hat{g}_M = \underset{t\in\mathcal{S}(\hat{Q},M)}{\arg\min}\, \gamma_N(t)$$

Pour choisir la valeur optimale de M, on cherche à minimiser l'erreur quadratique $\|\hat{g}_M - g^\star\|_2^2$. Comme $\|g^\star\|_2^2$ est fixée, cela revient approximativement à minimiser $\gamma_N(\hat{g}_M)$. Toutefois, pour prendre en compte les fluctuations stochastiques du processus empirique γ_N, on introduit une fonction de pénalisation $\mathrm{pen}(N, M)$. On sélectionne alors :

$$\hat{M} = \underset{M=1,\dots,N}{\arg\min} \left\{\gamma_N(\hat{g}_M) + \mathrm{pen}(N, M)\right\}.$$

Avec ce choix, l'estimateur final des moindres carrés pénalisés est défini par :

$$\hat{g} := \hat{g}_{\hat{M}}.$$

Il est important de noter que cette procédure suppose la connaissance préalable de l'ordre K du HMM. Dans Lehéricy [2019], l'auteur améliore cette approche en introduisant un estimateur consistant de l'ordre K. Outre l'établissement de la consistance pour l'estimation de l'ordre, la méthode proposée fournit également des estimateurs des paramètres du HMM qui sont minimax adaptatifs, à un facteur logarithmique près, par rapport à la régularité globale du modèle. Une approche plus fine, permettant une estimation minimax adaptative état par état, est présentée dans Lehéricy [2018].

1.3.4 Estimateur du maximum de vraisemblance pénalisé

Dans le cadre des modèles de Markov cachés (HMM), l'estimateur du maximum de vraisemblance (MLE) est défini comme la valeur du paramètre qui maximise la vraisemblance des observations, à savoir la fonction

$$L_n : (\theta, y_{1:n}) \longmapsto \mathbb{P}_\theta^{(n)}(y_{1:n}),$$

où $\theta = (K, \nu, Q, F)$ désigne le paramètre du modèle et $\mathbb{P}_\theta^{(n)}$ est la loi jointe de $Y_{1:n}$ sous θ. La construction du MLE dans ce cadre suit les étapes suivantes. On considère tout d'abord une famille de modèles paramétriques $\left(\mathcal{M}_M\right)_{M\in I}$, où chaque $\mathcal{M}_M$ est constitué de produits de n mesures de probabilité sur l'espace d'observation $(\mathbb{Y}, \mathcal{Y})$. Pour chaque $M \in I$, on définit

$$\Theta_{K,M} = \{\theta = (K, \nu, Q, F) : F \in \mathcal{M}_M\}.$$

Soit $y_{1:n}$ une réalisation fixée du vecteur aléatoire $Y_{1:n}$. Pour des valeurs fixées de K et M, le maximiseur de la vraisemblance est donné par

$$\hat{\theta}_{n,K,M} \in \underset{\theta\in\Theta_{K,M}}{\arg\max}\, L_n(\theta, y_{1:n}).$$

Afin de sélectionner un modèle, on introduit un terme de pénalisation $\mathrm{pen}_n(K, M)$ et on choisit

$$(\hat{K}_n, \hat{M}_n) \in \underset{(K,M)\in\mathbb{N}^*\times I}{\arg\max} \left\{\frac{1}{n}\log L_n\left(\hat{\theta}_{n,K,M}, y_{1:n}\right) - \mathrm{pen}_n(K, M)\right\}.$$

L'estimateur final est alors

$$\hat{\theta} = \hat{\theta}_{n,\hat{K}_n,\hat{M}_n}.$$

Les premières garanties théoriques pour cet estimateur ont été établies dans Vernet [2015], où la consistance postérieure et les vitesses de concentration d'un MLE non paramétrique bayésien sont démontrées. Alexandrovich et al. [2016a] ont prouvé la consistance d'un MLE non paramétrique construit à partir de HMM à espace d'états fini et de mélanges non paramétriques de densités paramétriques. Dans le cas non spécifié, Lehéricy [2021] ont étudié les HMM non paramétriques et montré que le MLE retrouve la meilleure approximation de la loi véritable.

1.4 Quelques problèmes d'inférence dans les HMMs et les modèles de mélange

1.4.1 Regroupement (Clustering)

Le *regroupement* est formellement défini comme la tâche consistant à retrouver la partition aléatoire $\Pi_n = \{A_1, \dots, A_m\}$ de l'ensemble d'indices $\{1, \dots, n\}$, où la partition est induite par les états latents $X_{1:n} = (X_1, \dots, X_n)$, de sorte que pour $k \in [\![1, m]\!]$, $i, j \in A_k$ si et seulement si $X_i = X_j$. L'objectif du problème de regroupement est d'inférer cette partition uniquement à partir des données observées $Y_{1:n} = (Y_1, \dots, Y_n)$.

Définition 1.4.1. *Un n-regroupeur g est une application mesurable de l'espace des données observées $\mathbb{Y}^n$ vers l'ensemble des partitions de $[n]$, noté $\mathcal{P}[n]$:*

$$g : \mathbb{Y}^n \to \mathcal{P}[n].$$

L'ensemble de tous les n-regroupeurs est noté $\mathcal{G}_n$.

Nous présentons ci-dessous un aperçu de plusieurs algorithmes de regroupement largement utilisés. Pour une discussion approfondie et des garanties théoriques sur le risque de regroupement de ces procédures, nous renvoyons le lecteur au Chapitre 12 de Giraud [2021].

Regroupement des K-moyennes

Définition 1.4.2 (Regroupement des K-moyennes)**.** *Soient $Y_1, \dots, Y_n \in \mathbb{R}^d$ et soit $K \geq 1$ un entier. Le* regroupement des K-moyennes *cherche une partition $\{C_1, \dots, C_K\}$ de $[\![1, n]\!]$ et des centroïdes $\{\mu_1, \dots, \mu_K\} \subset \mathbb{R}^d$ qui minimisent la variance intra-groupe totale :*

$$\sum_{k=1}^{K} \sum_{j \in C_k} \|Y_j - \mu_k\|^2, \quad \text{où } \mu_k = \frac{1}{|C_k|} \sum_{j \in C_k} Y_j.$$

L'algorithme des K-moyennes optimise un objectif non convexe et est NP-difficile en général. En pratique, on utilise l'algorithme de Lloyd. Celui-ci assigne itérativement les points à leurs centroïdes les plus proches et met à jour les centroïdes.

Définition 1.4.3 (Algorithme de Lloyd)**.** *Soient $Y_1, \dots, Y_n \in \mathbb{R}^d$ n observations et soit $K \in \mathbb{N}$ un nombre prédéfini de groupes. L'algorithme de Lloyd met à jour itérativement un ensemble de centres de groupes $\{\mu_1, \dots, \mu_K\}$ et une fonction d'affectation $\varphi : \{1, \dots, n\} \to \{1, \dots, K\}$ comme suit :*

Algorithm 1: Algorithme de Lloyd

Input: *Observations* $Y_1, \dots, Y_n \in \mathbb{R}^d$*, nombre de groupes* K
Output: *Affectations de groupes* $\varphi^{(t)}$ *et centres* $\mu_1^{(t)}, \dots, \mu_K^{(t)}$

Initialisation : *Choisir des centres initiaux* $\mu_1^{(0)}, \dots, \mu_K^{(0)}$ *;*
repeat
  Étape d'affectation :
  for $i = 1$ **to** n **do**
    $\varphi^{(t)}(i) \leftarrow \arg\min_{k \in \{1,\dots,K\}} \|Y_i - \mu_k^{(t)}\|^2$
  Étape de mise à jour :
  for $k = 1$ **to** K **do**
    $\mu_k^{(t+1)} \leftarrow \frac{1}{|\{i : \varphi^{(t)}(i) = k\}|} \sum_{i : \varphi^{(t)}(i) = k} Y_i$
until *Convergence atteinte;*

L'algorithme s'arrête lorsque les affectations $\varphi^{(t)}$ ne changent plus ou lorsque la diminution de la somme des carrés intra-groupes devient inférieure à un seuil fixé. Lu and H. Zhou [2016] établissent des garanties théoriques pour les performances de cet algorithme sur des observations sous-gaussiennes. Des garanties théoriques pour d'autres variantes de l'algorithme des K-moyennes ont également été établies. Nous renvoyons à Ndaoud [2022] pour une variante de l'algorithme de Lloyd initialisée par l'algorithme de regroupement spectral, et à Giraud and Verzelen [2018] pour une version relâchée de l'algorithme des K-moyennes.

Regroupement spectral

Le regroupement spectral est un algorithme largement utilisé pour découvrir des structures latentes de groupes dans les données. Dans le contexte des modèles de mélange gaussiens (GMMs), il offre une procédure simple mais puissante pour estimer les groupes uniquement à partir des données observées. L'algorithme spectral sert généralement d'étape initiale de regroupement, produisant une affectation grossière des groupes, laquelle peut ensuite être affinée à l'aide d'une méthode plus spécialisée.

Soient $Y_1, \dots, Y_n \in \mathbb{R}^d$ des observations i.i.d. issues d'un GMM avec K composantes. On note $\mathbf{Y} = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times d}$ la matrice de données. L'idée centrale de l'algorithme spectral est de calculer les K plus grands vecteurs propres de la matrice de Gram $\mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{n \times n}$, puis d'utiliser ces vecteurs propres pour définir une nouvelle représentation des données qui révèle la structure de regroupement. La justification théorique provient du fait que, sous un GMM bien séparé, l'espérance de la matrice de Gram admet la décomposition suivante :

$$\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = A\Theta\Theta^\top A^\top + \Gamma,$$

où :

- $A \in \mathbb{R}^{n \times K}$ est la matrice d'appartenance, définie par $A_{ik} = \mathbf{1}_{\{i \in G_k\}}$ avec G_k l'ensemble des indices appartenant au groupe k,
- $\Theta \in \mathbb{R}^{K \times d}$ contient les moyennes des composantes gaussiennes,
- $\Gamma \in \mathbb{R}^{n \times n}$ est une matrice diagonale représentant la covariance intra-groupe.

Ainsi, $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$ se compose d'une matrice signal de rang faible (rang K) structurée par l'affectation aux groupes et d'un terme de bruit additif. Cela motive l'utilisation de la meilleure approximation de rang-K de $\mathbf{Y}\mathbf{Y}^\top$ comme approximation de la structure de regroupement.

Soit $\mathbf{Y}\mathbf{Y}^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ la décomposition spectrale de la matrice de Gram, avec $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. On définit l'approximation de rang-K par :

$$\left(\mathbf{Y}\mathbf{Y}^\top\right)_{(K)} = \sum_{k=1}^{K} \lambda_k v_k v_k^\top.$$

L'embedding spectral des données consiste à extraire les K premiers vecteurs propres $v_1, \ldots, v_K \in \mathbb{R}^n$. Ces vecteurs propres peuvent être empilés en colonnes pour former une matrice $V \in \mathbb{R}^{n \times K}$, où chaque ligne correspond à une représentation de faible dimension d'une observation. Un algorithme de regroupement, tel que les K-moyennes, est ensuite appliqué aux lignes de V pour retrouver la structure de groupe. L'algorithme suivant résume cette procédure de regroupement :

Algorithm 2: Regroupement spectral sous un modèle de mélange gaussien

Input: Matrice de données $\mathbf{Y} \in \mathbb{R}^{n \times d}$, nombre de groupes K
Output: Affectations des groupes pour les n observations
Étape 1 : Calcul de la matrice de Gram
$\mathbf{G} \leftarrow \mathbf{Y}\mathbf{Y}^\top$
Étape 2 : Décomposition spectrale de la matrice de Gram
Calcul des valeurs et vecteurs propres :
$\mathbf{G} = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ où $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
Étape 3 : Construction de l'approximation de rang-K
$\mathbf{G}_{(K)} \leftarrow \sum_{k=1}^{K} \lambda_k v_k v_k^\top$
Étape 4 : Construction de l'embedding spectral
$V \leftarrow [v_1 \; v_2 \; \cdots \; v_K] \in \mathbb{R}^{n \times K}$
Étape 5 : Regroupement des points projetés
Appliquer l'algorithme des K-moyennes aux lignes de V pour obtenir les étiquettes de groupes
return *Affectation de chaque observation à son groupe*

Löffler et al. [2021] établissent des garanties théoriques sur la performance de l'algorithme de clustering spectral dans les modèles de mélange gaussiens à covariance isotrope. Il est montré que l'algorithme atteint des performances minimax optimales, à condition que le nombre de classes soit fixé et que le rapport signal-bruit soit suffisamment élevé.

Regroupement basé sur les modèles

Le regroupement basé sur les modèles McNicholas [2016] est une approche statistique qui repose sur les modèles de mélange finis pour effectuer le regroupement. Soit $\mathbf{Y} = [Y_1, \ldots, Y_n]^\top \in \mathbb{R}^{n \times d}$ un ensemble de n observations, supposées indépendamment issues d'une loi de mélange fini de densité

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y, \alpha_k),$$

où $\forall k \in [K]$, $\pi_k > 0$ et $\sum_{k=1}^{K} \pi_k = 1$. Dans les applications typiques de regroupement, les densités des composantes $f_k(y, \alpha_k)$ partagent la même forme fonctionnelle, de sorte

que $f_k(y, \alpha_k) = f(y, \alpha_k)$ pour tout $k \in [K]$, où f est une famille paramétrique connue. L'ensemble complet des paramètres du modèle est noté $\theta = \big(\pi, (\alpha_k)_{k\in[K]}\big)$. Sous cette formulation, la vraisemblance des données observées s'écrit :

$$L_n(\theta; \mathbf{Y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(Y_i, \alpha_k).$$

Une fois les paramètres estimés — généralement via l'algorithme Espérance-Maximisation (EM) — chaque observation est assignée à la composante maximisant la probabilité a posteriori :

$$\forall i \in [n], \quad \hat{z}_i := \arg\max_{k\in[K]} \frac{\hat{\pi}_k f(Y_i, \hat{\alpha}_k)}{\sum_{j=1}^{K} \hat{\pi}_j f(Y_i, \hat{\alpha}_j)}.$$

Le regroupement est alors effectué sur la base de la partition définie par les étiquettes estimées $\hat{z}_i$.
Cette méthodologie s'étend naturellement aux modèles de Markov cachés (HMM). Une fois les paramètres du modèle estimés, le calcul des lois postérieures des variables latentes devient plus complexe en raison des dépendances temporelles entre les états cachés, qui compliquent la structure de la vraisemblance. Néanmoins, l'algorithme Forward–Backward Cappé et al. [2005] fournit un moyen efficace d'évaluer ces postérieures. Cela permet alors d'adopter une approche de regroupement analogue à celle des modèles de mélange paramétriques, en utilisant la partition induite par les états cachés estimés. Concrètement, les variables latentes sont estimées par

$$\forall i \in [n], \quad \hat{z}_i := \arg\max_{k\in[K]} \phi_{\hat{\theta},i|1:n}(k),$$

où $\hat{\theta}$ désigne l'estimateur des paramètres du HMM et $\phi_{\hat{\theta},i|1:n}$ est la loi postérieure de l'état caché associé à la i-ème observation, étant donné la séquence observée $Y_{1:n}$ sous le paramètre $\hat{\theta}$. Puisqu'un algorithme de regroupement (clusterer) peut être obtenu à partir d'un classifieur en ne conservant que la partition induite, il est naturel d'utiliser pour le regroupement la partition générée par $(\hat{z}_i)_{i\in[n]}$. Cependant, cette pratique courante parmi les utilisateurs de HMM ne repose pas sur une justification rigoureuse en termes de théorie de la décision : elle imite en réalité le classifieur bayésien (minimiseur du risque en classification) plutôt que le regroupement bayésien (minimiseur du risque en clustering). Des définitions formelles sont données au Chapitre 3. La question est d'importance pratique, dans la mesure où le classifieur bayésien possède une expression en forme fermée et est en général plus facile à calculer. Chaque fois que le regroupement bayésien peut être dérivé du classifieur bayésien, cela représente un avantage computationnel évident. Or, la relation entre le regroupement bayésien et le classifieur bayésien n'avait encore jamais été étudiée. Une partie essentielle du Chapitre 3 est consacrée à l'examen et à l'éclaircissement de ce lien.

1.4.2 Autres problèmes d'inférence

En plus des tâches classiques telles que l'estimation, la classification et le regroupement, plusieurs autres problèmes d'inférence apparaissent dans le cadre des modèles de mélanges et des modèles de Markov cachés (HMM). Une tâche centrale d'inférence dans les HMM est de retrouver la séquence des états cachés $X_1, \ldots, X_n$ à partir de la séquence observée $Y_1, \ldots, Y_n$. Contrairement au classifieur de Bayes, qui assigne à chaque observation l'état caché le plus probable via le classifieur marginal *Maximum a posteriori*, la récupération de la séquence entière des états cachés requiert une approche différente.

En effet, en raison de la structure de dépendance imposée par la chaîne de Markov cachée, la loi a posteriori jointe $\mathcal{L}(X_{1:n} \mid Y_{1:n})$ n'est pas simplement le produit des lois postérieures individuelles $\mathcal{L}(X_t \mid Y_{1:n})$. En conséquence, une séquence formée en choisissant à chaque instant l'état le plus probable peut ne pas correspondre à un chemin valide du modèle, en particulier lorsque certaines transitions d'état sont interdites. Pour pallier ce problème, on utilise l'algorithme de *Viterbi*, qui permet d'identifier, via une programmation dynamique, la séquence qui maximise la distribution a posteriori jointe des états cachés donnés les observations. Voir Cappé et al. [2005].

Cet algorithme diffère du Maximum a posteriori dans le type d'optimalité recherchée. Le classifieur de Bayes maximise la loi postérieure marginale de chaque état caché indépendamment, tandis que l'algorithme de Viterbi identifie la séquence d'états la plus probable *dans son ensemble* en maximisant la loi postérieure jointe. Ce dernier aspect est particulièrement important dans les applications où la cohérence temporelle et les contraintes du modèle doivent être respectées, car il garantit que le chemin inféré est valide au regard de la dynamique de transition du HMM. Voir Rabiner [1989], Forney [1973].

1.5 Contributions au problème de regroupement

Dans les Chapitres 3 et 4, nos contributions concernent principalement le problème du regroupement. Nous aborderons plusieurs questions fondamentales, qui seront traitées dans les cadres i.i.d. et HMM. Bien que notre étude se concentre surtout sur le cadre non-paramétrique, nous examinerons également le cadre gaussien afin d'identifier la dépendance exacte du risque de Bayes de regroupement par rapport aux paramètres du modèle, ainsi que le gain permis par la structure de dépendance en termes de performance de regroupement (voir Chapitre 4).

Dans le Chapitre 3, nous supposerons qu'il existe une constante absolue δ telle que $\min_{x,x'} Q_{x,x'} \geq \delta$ et $\min_x \nu_x \geq \delta$, où Q désigne la matrice de transition de la chaîne cachée et ν la loi initiale, ce qui garantit que le processus caché explore suffisamment tous les états. Cette hypothèse sera relâchée dans le Chapitre 4. Nos contributions au problème du regroupement se résument comme suit :

1. **Lien entre classifieur de Bayes et regroupeur de Bayes :** Le classifieur de Bayes et le regroupeur de Bayes sont définis comme les minimiseurs des risques de classification et de regroupement, respectivement. Une question centrale est de savoir s'il existe un lien clair entre ces deux objets et, si oui, sous quelles hypothèses structurelles sur les données ce lien peut être précisé. Cette question est d'un grand intérêt pratique : le classifieur de Bayes admet souvent une expression simple, et est donc beaucoup plus facile à calculer, tandis que le regroupeur de Bayes est souvent plus abstrait et moins directement accessible. Ainsi, chaque fois que le regroupeur de Bayes peut être déduit du classifieur de Bayes, cela apporte non seulement un éclairage conceptuel mais aussi un avantage algorithmique concret. Notre analyse révèle un résultat frappant : le regroupement induit par le classifieur de Bayes coïncide avec celui du regroupeur de Bayes si et seulement si les observations sont i.i.d. issues d'un mélange à exactement deux composantes. Dans tous les autres cas—mélanges à plus de deux composantes ou données dépendantes générées par un HMM—il existe toujours une configuration de paramètres pour laquelle le regroupement produit par le classifieur de Bayes diffère de celui du regroupeur de Bayes avec probabilité positive. Ces résultats sont établis rigoureusement dans les Théorèmes 3.3.1, 3.3.4, 3.3.6 et 3.3.8 du Chapitre 3.

2. **Ordre de grandeur des risques de Bayes :** Sous quelles conditions le risque de

Bayes du regroupement est-il comparable à celui de la classification ? Cette question est particulièrement importante car le risque de Bayes de classification admet une expression en forme simple, ce qui le rend analysable de manière fine. Cette tractabilité suggère que, chaque fois que les deux risques sont comparables, toute borne théorique établie pour la classification peut être transférée au regroupement. Nos résultats montrent que cette comparabilité n'a lieu que dans deux régimes : soit lorsque le nombre de classes est limité à deux, soit lorsque le risque de Bayes de classification ne décroît pas à une vitesse exponentielle en fonction de la taille de l'échantillon n. En dehors de ces scénarios, les deux risques se comportent de manière fondamentalement différente. Voir les Théorèmes 3.3.2, 3.3.5, 3.3.7, 3.3.9 et le Corollaire 1.

3. **Dépendance aux paramètres du modèle :** Un aspect central de notre analyse réside dans la compréhension de la dépendance des risques de Bayes aux paramètres du modèle, et en particulier aux densités de population qui caractérisent les distributions des observations. Nous identifions la quantité clé qui gouverne la difficulté intrinsèque des problèmes de classification et de regroupement :

$$\Lambda := \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left(\sum_{x \neq x_0} f_x(y) \right) d\mathcal{L}(y),$$

où les f_x désignent les densités par rapport à une mesure dominante $\mathcal{L}$. Dans le cas particulier de deux populations ($J = 2$), on obtient

$$\Lambda = 1 - \|F_1 - F_2\|_{\mathrm{TV}},$$

où $\|\cdot\|_{\mathrm{TV}}$ désigne la distance en variation totale.

4. **Apprentissage du regroupement :** Est-il possible de construire un regroupeur, sans connaissance des paramètres réels, dont l'excès de risque tend vers zéro quand n croît ? Nous apportons une réponse positive à cette question en construisant une procédure de regroupement (clustering) qui atteint des performances quasi optimales. Plus précisément, nous montrons que, lorsque les densités d'émission sont s-Hölder régulières et sous des hypothèses de régularité supplémentaires, il existe une procédure de regroupement $\tilde{g}$ telle que

$$\mathcal{R}_n^{\mathrm{clust}}(\theta, \tilde{g}) - \inf_g \mathcal{R}_n^{\mathrm{clust}}(\theta, g) = \mathcal{O}\left(\left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}} \right).$$

Ce résultat établit une justification théorique de l'utilisation du classifieur de Bayes par plug-in dans la pratique. Pour les praticiens travaillant avec des modèles de Markov cachés (HMM), le message clé est que, malgré la différence conceptuelle entre le classifieur de Bayes et le regroupeur de Bayes, le classifieur de Bayes peut être utilisé en toute confiance comme substitut pour le regroupement. Pour renforcer cette conclusion, nous présentons des expériences numériques dans un cadre non paramétrique, qui illustrent clairement que les procédures de regroupement basées sur les HMM sont capables de révéler des structures latentes même dans des situations où les méthodes conventionnelles échouent.

5. **Caractérisation précise du risque de Bayes :** Comment la structure de dépendance du modèle de Markov caché améliore-t-elle les performances du clustering? Nous illustrerons la valeur ajoutée de la dépendance dans le régime où la chaîne

de Markov cachée mélange lentement et où les densités d'émission sont gaussiennes. Cela repose sur une caractérisation précise de la dépendance du risque de Bayes en clustering par rapport aux paramètres du modèle. Nous construirons également des procédures de clustering optimales. C'est le sujet du chapitre 4.

À la suite de cette étude sur le regroupement, nous avons été naturellement amenés à explorer d'autres problèmes d'inférence dans les modèles de mélanges et HMM, en particulier la détection de ruptures et la segmentation. Au cours de cette exploration, nous avons rencontré une conjecture formulée dans Bet et al. [2025] concernant la détection de rupture dans un cadre totalement différent : les graphes aléatoires à attachement préférentiel. Cette rencontre inattendue a suscité notre intérêt et nous a conduits à étudier ce problème.

1.6 Le modèle de graphe aléatoire à attachement préférentiel

Le modèle d'attachement préférentiel (PA), avec ses nombreuses variantes, constitue l'un des cadres les plus utilisés pour modéliser des graphes aléatoires en science des réseaux. Ces modèles permettent d'étudier des réseaux réels tels que les réseaux sociaux, les réseaux de citations ou le Web, où les nouveaux sommets ont tendance à se connecter à ceux déjà bien connectés. Depuis deux décennies, des travaux approfondis ont permis d'établir des résultats solides sur leurs propriétés structurelles, notamment la distribution des degrés, la structure locale, etc. Voir van der Hofstad [2016, 2024]. Introduit par Barabási and Albert [1999a], le modèle d'attachement préférentiel a suscité une attention considérable en raison de son mécanisme de croissance local simple et de sa capacité à générer naturellement des distributions de degrés en loi de puissance, comme souvent observé dans les réseaux réels. Le mécanisme d'attachement préférentiel définit une suite de multigraphes aléatoires $(G_t)_{t\geq 1}$ sur les ensembles de sommets $\{1,\dots,t\}$. Il n'y a aucune perte de généralité à supposer que ces graphes sont orientés, en adoptant la convention selon laquelle les flèches vont des sommets portant les plus grands labels vers ceux portant les plus petits labels. Pour être un peu plus précis, dans ce qui suit, un *graphe étiqueté* renvoie à la définition suivante.

Définition 1.6.1 (Graphe étiqueté)**.** *Un (multi)graphe étiqueté $\mathfrak{g}$ est un couple $(\mathcal{V},\mathcal{E})$ où $\mathcal{V}$ est l'ensemble des sommets et $\mathcal{E}\subset\mathcal{V}^2$ est le multiensemble des arêtes orientées, sans boucle. Pour une arête $(u,v)\in\mathcal{E}$, on adopte la convention que la flèche va de u vers v, et on note pour simplifier $u\to_{\mathfrak{g}} v$.*

Rappelons que dans un multigraphe, deux sommets peuvent être reliés par plusieurs arêtes. Il existe de nombreuses variantes du mécanisme d'attachement préférentiel. Le modèle le plus simple est le modèle de Barabási–Albert Barabási and Albert [1999a], qui est défini ci-dessous.

Définition 1.6.2 (Modèle de Barabási–Albert)**.** *On construit la suite de graphes $(G_t)_{t\geq 1}$ de la façon suivante :*

- *G_1 est le graphe réduit à un sommet 1.*
- *Le sommet 2 se connecte à 1 pour former G_2.*
- *Pour $n\geq 3$, on construit G_n à partir de G_{n-1} en reliant le nouveau sommet n à un sommet $v\in\{1,\dots,n-1\}$ choisi avec probabilité proportionnelle à son degré :*

$$\mathbb{P}\left(n\to_{G_n} v \mid G_{n-1}\right)=\frac{d_v(G_{n-1})}{\sum_{j=1}^{n-1} d_j(G_{n-1})}.$$

où $d_v(G_{n-1})$ est le degré du sommet v dans le graphe G_{n-1}.

Alors que le modèle canonique est le modèle de Barabási–Albert (BA), plusieurs variantes ont été développées afin de capturer des propriétés plus nuancées observées dans les réseaux réels. Par exemple, le modèle de Barabási–Albert peut naturellement être adapté pour générer des multigraphes, où plusieurs arêtes entre deux sommets sont autorisées. Le mécanisme central reste inchangé ; la différence principale réside dans le fait qu'à chaque étape, le sommet nouvellement ajouté forme m arêtes au lieu d'une seule. Ces m connexions sont établies en sélectionnant les sommets existants — soit simultanément, soit successivement — selon la même règle d'attachement préférentiel. Cela permet des arêtes multiples et reflète des schémas de connectivité plus complexes observés dans les réseaux réels. Le modèle suivant est couramment utilisé à cette fin.

Définition 1.6.3 (Modèle d'attachement préférentiel avec degré sortant m)**.** *Le modèle d'attachement préférentiel (AP) avec degré sortant m et fonction d'attachement f génère une suite de graphes $(G_t)_{t\in\mathbb{N}^\star}$ sur les ensembles de sommets $V_t = \{1, \dots, t\}$, définie inductivement comme suit :*

- *Initialiser G_1 comme étant le graphe constitué d'un unique sommet 1 sans arêtes.*
- *Soit G_2 le graphe formé de deux sommets 1 et 2, reliés par m arêtes.*
- *Pour $t \geq 3$ et étant donné G_{t-1}, définir une suite de graphes intermédiaires :*

$$G_{t,0}, G_{t,1}, \dots, G_{t,m}$$

où :

  - *$G_{t,0}$ est le graphe G_{t-1} auquel on ajoute un nouveau sommet isolé noté t.*
  - *Pour chaque $i \in [m]$, construire $G_{t,i}$ à partir de $G_{t,i-1}$ en ajoutant une arête entre v_t et un sommet existant $v \in V_{t-1}$, tiré selon la loi :*

$$\mathbb{P}(v_{t,i} = v \mid G_{t,i-1}) = \frac{f(d_v(G_{t,i-1}))}{\sum_{j=0}^{t-1} f(d_j(G_{t,i-1}))}$$

où $d_v(G)$ désigne le degré du sommet v dans le graphe G.

- *Définir $G_t := G_{t,m}$.*

Lorsque $f(x) = x$, on parle d'attachement préférentiel linéaire avec degré sortant m.
Lorsque $f(x) = x + \delta$, on parle d'attachement préférentiel affine avec degré sortant m.
Lorsque f est constante, on parle d'attachement uniforme (UA).

Notons que lorsque f n'est pas linéaire, le modèle est plus difficile à analyser car, contrairement aux modèles linéaires où le dénominateur dans la probabilité d'attachement est déterministe et peut être calculé explicitement, le dénominateur n'admet ici pas d'expression fermée simple. C'est la raison pour laquelle une grande partie de la littérature sur le sujet se concentre sur le modèle d'attachement préférentiel linéaire. Le modèle d'attachement préférentiel affine généralise le modèle de Barabási–Albert en introduisant un degré de liberté supplémentaire qui permet d'ajuster le biais en faveur des sommets de haut degré grâce à un paramètre ajustable δ. Plus précisément, la probabilité d'attachement devient :

$$\mathbb{P}\left(n \to_{G_n} v \mid G_{n-1}\right) = \frac{d_v(G_{n-1}) + \delta}{\sum_{j=1}^{n-1} \left(d_j(G_{n-1}) + \delta\right)}$$

Par exemple, des valeurs élevées de δ atténuent l'*effet préférentiel* car elles garantissent que même les sommets de faible degré ont une chance non négligeable de recevoir de nouvelles arêtes. Lorsque δ devient très grand, la propriété préférentielle disparaît, ce qui revient à un modèle d'attachement uniforme où chaque sommet est sélectionné avec la même probabilité à chaque étape. D'autres variantes du modèle permettent que le paramètre δ varie au cours du temps, c'est-à-dire que sa valeur peut changer pendant la construction du graphe.

Une des propriétés caractéristiques des graphes à attachement préférentiel est la propriété d'*invariance d'échelle* (*scale-free*). Cela signifie que la distribution asymptotique des degrés suit une loi de puissance : la probabilité qu'un sommet ait un degré k décroît comme $k^{-\gamma}$ pour $\gamma > 1$. Ce comportement est une caractéristique distinctive de nombreux réseaux réels, et constitue l'une des raisons principales pour lesquelles le modèle d'attachement préférentiel est devenu central en science des réseaux. Des preuves empiriques montrent que des structures d'invariance d'échelle apparaissent dans une large gamme de systèmes complexes. Par exemple : le Web présente une loi de puissance sur le nombre d'hyperliens par page ; les réseaux de citation sont dominés par quelques articles très cités ; les réseaux sociaux comprennent des individus exceptionnellement connectés ; et dans les systèmes biologiques, comme les réseaux d'interaction protéine–protéine, un petit nombre de protéines participent à de nombreuses interactions. Le modèle d'attachement préferentiel capture cette large hétérogénéité grâce à un mécanisme simple mais puissant de type *rich-get-richer* : les sommets ayant un degré plus élevé ont plus de chances d'attirer de nouvelles arêtes. Cela est illustré à la Figure 1.3 avec différents choix de la fonction d'attachement. Pour des exemples supplémentaires et des discussions, voir le Chapitre 1 de van der Hofstad [2016].

1.7 Problèmes d'inférence sous le modèle de graphes aléatoires à attachement préférentiel

Dans cette section, nous passons en revue les principaux problèmes d'inférence qui sont habituellement étudiés dans le cadre du modèle de graphe aléatoire par attachement préférentiel. Nous supposons que l'on n'observe que le graphe aléatoire par attachement préférentiel non étiqueté. Voir la Définition 5.2.3 pour une définition formelle du graphe non étiqueté.

1.7.1 Distribution asymptotique des degrés

Une des propriétés clés du modèle d'attachement préférentiel est son comportement asymptotique : la distribution des degrés des sommets converge vers une distribution à queue épaisse, souvent de type loi de puissance. Cette section passe en revue les bases mathématiques de ce phénomène et met en évidence comment différents choix de la règle d'attachement influencent la distribution asymptotique.

Il convient de noter que la distribution des degrés de type loi de puissance n'apparaît pas dans d'autres modèles de graphes aléatoires standards tels que l'attachement uniforme et les graphes d'Erdős–Rényi (voir Figure 1.4), modèle que nous définissons ci-dessous.

Définition 1.7.1 (Modèle de graphe aléatoire d'Erdős–Rényi)**.** Un graphe aléatoire d'Erdős–Rényi (ER) $G(n, p)$ *est un graphe non orienté aléatoire à n sommets où, pour chaque paire de sommets distincts $i, j \in \{1, \dots, n\}$, une arête $\{i, j\}$ est présente avec probabilité p, indépendamment de toutes les autres arêtes.*

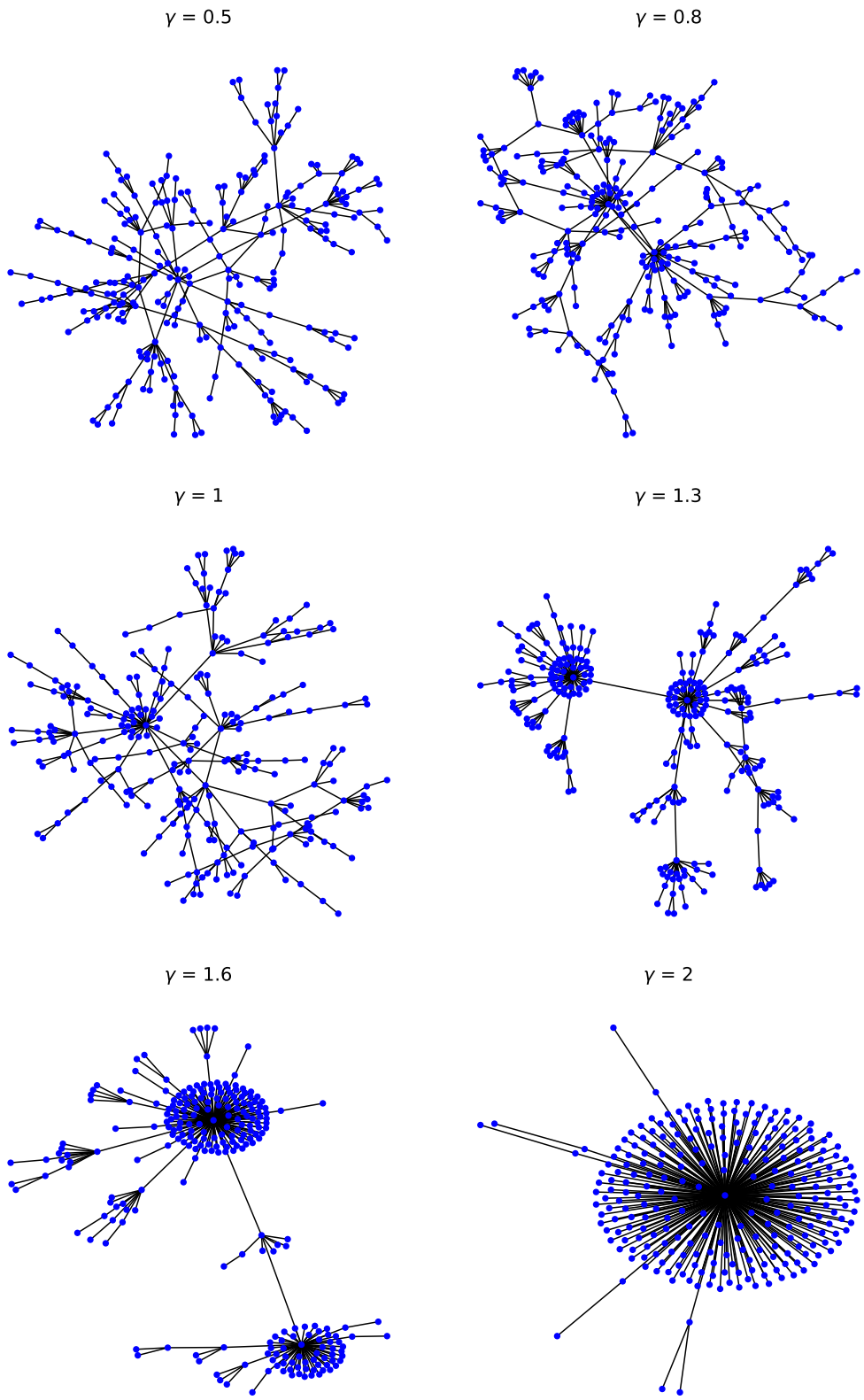


Figure 1.3: Graphes à attachement préférentiel avec fonction d'attachement $f(x) = x^\gamma$. Chaque graphe contient 250 sommets.

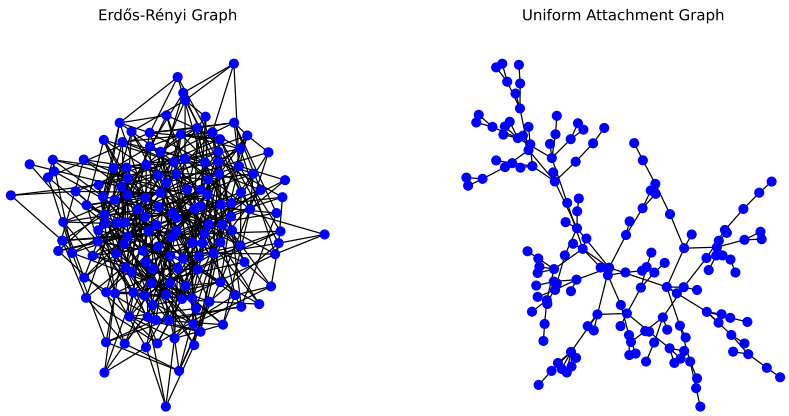


Figure 1.4: Graphe Erdős–Rényi avec paramètre $p = 0.05$ et graphe à attachement uniforme. Chaque graphe contient $n = 150$ sommets.

Soit $P_k(n)$ la proportion de sommets dans G_n ayant un degré k. La question centrale est de savoir si cette quantité se stabilise lorsque $n \to \infty$, et si oui, sous quelle forme. Le théorème suivant identifie cette distribution limite pour le modèle d'attachement préférentiel linéaire avec degré sortant m (correspondant à la fonction d'attachement $f(x) = x$, voir Définition 1.6.3) et est dû à Bollobás et al. [2001]. La distribution limite présente clairement une queue de type loi de puissance avec un exposant $\gamma = 3$, une évidence empirique initialement rapportée dans Barabási and Albert [1999b].

Théorème 1.7.1 (Bollobás et al. [2001])**.** *Pour $k \geq m$ et lorsque $n \to +\infty$, $P_k(n) \to p_k$ en probabilité, où*

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}.$$

Pour un attachement sous-linéaire (c'est-à-dire $f(x) \ll x$), le mécanisme du *rich-get-richer* est plus faible. Dans ce cas, la distribution des degrés décroît plus rapidement qu'une loi de puissance, typiquement selon une *loi exponentielle étirée* Krapivsky et al. [2000], Dereich and Mörters [2009] :

$$p_k \asymp \exp(-c \cdot k^{\alpha}),$$

avec $\alpha \in (0, 1)$. Ce régime produit des graphes plus homogènes, où les hubs sont rares et la plupart des sommets ont des degrés comparables. Dans le régime sur-linéaire (c'est-à-dire $f(x) \gg x$), un phénomène de condensation apparaît : un sommet dominant émerge, se connectant à presque tous les autres sommets, tandis que tous les autres sommets ont des degrés finis. Ce phénomène est formalisé et démontré rigoureusement dans Oliveira and Spencer [2005]. La dérivation de ces résultats utilise plusieurs techniques clés : relations récursives pour suivre le nombre attendu de sommets de degré k, méthodes de décomposition en martingales, et immersion du processus discret d'attachement préférentiel dans des processus de branchement ou de naissance en temps continu.

1.7.2 Diamètre

Définition 1.7.2 (Diamètre d'un graphe)**.** *Soit $G = (\mathcal{V}, \mathcal{E})$ un graphe connexe. Le* diamètre *de G, noté* $\mathrm{diam}(G)$*, est défini comme la distance maximale sur les plus courts*

chemins entre toutes paires de sommets :

$$\mathrm{diam}(G) = \max_{u,v\in\mathcal{V}} d_G(u,v),$$

où $d_G(u,v)$ est la distance du plus court chemin (c'est-à-dire le nombre minimal d'arêtes) entre les sommets u et v. Lorsque G n'est pas connexe, son diamètre est celui de sa plus grande composante connexe.

1.7.3 Degré maximal

Le degré maximal d'un graphe aléatoire correspond au plus haut degré parmi tous les degrés des sommets du graphe. Il reflète la présence de hubs et l'hétérogénéité du réseau. Son étude permet de comprendre à quel point un réseau est centralisé ou déséquilibré. Dans le Tableau 1.1 ci-dessous, nous résumons les propriétés clés de certains modèles de graphes usuels. Soit

$$p_k(m,\delta) = (2+\delta/m)\,\frac{\Gamma(k+\delta)\Gamma(m+2+\delta+\delta/m)}{\Gamma(m+\delta)\Gamma(k+3+\delta+\delta/m)}.$$

Remarquons que pour k grand, $p_k(m,\delta) \approx C(m,\delta)k^{-(3+\delta/m)}$.

Modèle	**Dist. des degrés asymp.**	**Degré maximal**	**Diamètre**
ER$(n,\lambda/n)$ $(\lambda>1)$	Poisson(λ)	$\Theta\left(\frac{\log n}{\log\log n}\right)$ Móri [2005]	$\Theta(\log n)$ Durrett [2006]
UA	$\left(2^{-k}\right)_{k\geq 1}$ Janson [2005]	$\Theta(\log n)$ Devroye and Lu [1995]	$\Theta(\log n)$ Pittel [1994]
BA	$\left(\frac{4}{k(k+1)(k+2)}\right)_{k\geq 1}$	$\Theta(n^{1/2})$	$\Theta(\log n)$ Bollobás and Riordan [2004]
PA linéaire $(m>1)$	$\left(\frac{2m(m+1)}{k(k+1)(k+2)}\right)_{k\geq m}$	$\Theta(n^{1/2})$	$\Theta\left(\frac{\log n}{\log\log n}\right)$ Bollobás and Riordan [2004]
PA affine $(m,\delta>0)$	$(p_k(m,\delta))_{k\geq m}$	$\Theta(n^{\frac{1}{2+\delta/m}})$	$\Theta(\log n)$ Dommers et al. [2010]

Table 1.1: Propriétés asymptotiques de certains modèles de graphes aléatoires.

Les résultats sans références dans le tableau ci-dessus, ainsi que leurs démonstrations correspondantes, se trouvent dans van der Hofstad [2016].

1.7.4 Archéologie des réseaux dans les graphes aléatoires récursifs

Dans les graphes aléatoires récursifs, tels que les graphes d'attachement préférentiel et d'attachement uniforme, l'archéologie du graphe désigne le problème consistant à identifier l'ordre d'arrivée des sommets (ou d'une partie des sommets) du graphe final observé. Bien entendu, cela suppose que le graphe final observé n'est pas étiqueté, car sinon l'ordre peut être déduit des étiquettes. Ce problème comprend :

- **Recherche de la racine :** Cela correspond à identifier le premier sommet (souvent noté sommet 1) ajouté au graphe lors de sa construction. Comme l'identification exacte de la racine est généralement impossible, l'idée est d'identifier le plus petit ensemble de confiance contenant la racine avec grande probabilité. Ce problème a été

étudié dans Bubeck et al. [2017] pour les cas d'attachement préférentiel et uniforme. Pour les graphes d'attachement préférentiel, une taille minimale de l'ensemble de confiance pour la recherche de la racine a été identifiée dans Bubeck et al. [2017] et un algorithme optimal a été proposé dans Contat et al. [2024]. Khim and Loh [2016] ont étudié la recherche de la racine dans le cas où l'arbre est obtenu par diffusion sur un arbre régulier infini, et Brandenberger et al. [2022] dans le cas d'un arbre de Galton-Watson de taille conditionnée.

- **Estimation des temps d'arrivée :** Identifier le premier sommet fournit seulement une vision partielle de l'évolution du graphe. Reconstituer l'ordre complet d'arrivée des sommets peut être beaucoup plus informatif. Cette information est particulièrement utile pour suivre la propagation de désinformation, de rumeurs ou même de virus à travers un réseau. Dans Crane and Xu [2021], un cadre général pour l'archéologie des réseaux est développé et peut être appliqué au problème d'inférence des temps d'arrivée. Plus récemment, ce problème a été étudié dans Briend et al. [2025] dans le cas de l'attachement uniforme et de l'attachement préférentiel linéaire, où un estimateur d'ordre a été introduit et montré optimal par rapport à une famille de mesures de risque.

1.7.5 Détection et localisation des ruptures

Tout d'abord, notez que ce problème ne peut pas être formalisé sous l'attachement uniforme ou linéaire, car le mécanisme d'attachement reste le même durant tout le processus de construction du graphe. Un modèle simple où ce problème peut être étudié est le modèle d'attachement préférentiel affine (voir Définition 1.6.3). Dans ce modèle, à chaque étape, un nouveau sommet rejoint le graphe et forme m arêtes vers des sommets existants, avec une probabilité d'attachement influencée par une fonction $\delta(t)$ qui modifie la préférence linéaire en fonction du degré du sommet. Plus formellement, la probabilité qu'un nouveau sommet au temps t se connecte à un sommet existant v est proportionnelle à $d_{G_{t-1}}(v)+\delta(t)$, où $d_{G_{t-1}}(v)$ est le degré de v dans G_{t-1}.

Pour analyser ce modèle rigoureusement, on introduit un processus $((G_{t,i})_{i=1}^{m})_{t\geq 1}$, où chaque $G_{t,i}$ correspond à l'état intermédiaire du graphe après que i des m nouvelles arêtes du sommet t ont été ajoutées. Cette formulation permet de modéliser précisément la règle probabiliste d'attachement pour chaque ajout d'arête. Le graphe final G_t est alors obtenu en posant $G_t = G_{t,m}$. Une des questions statistique qui peut-être étudiée dans ce contexte est de savoir si le mécanisme d'attachement, codé par $\delta(t)$, reste constant au fil du temps ou subit un changement structurel. Le problème de détection des ruptures peut être défini comme un problème de test d'hypothèse, qu'on rappelle ci-dessous.

Définition 1.7.3 (Test d'hypothèse)**.** *Soit $(\mathbb{Y}, \mathcal{Y})$ un espace mesurable et soit $\mathbb{P}_0$ et $\mathbb{P}_1$ deux mesures de probabilité sur $\mathbb{Y}$ correspondant respectivement à l'hypothèse nulle H_0 et à l'hypothèse alternative H_1. Un* test statistique *est une fonction mesurable*

$$\phi : \mathbb{Y} \to \{0, 1\},$$

où $\phi(y) = 1$ indique le rejet de l'hypothèse nulle H_0 basé sur l'observation $y \in \mathbb{Y}$.

- *L'*erreur de type I *(faux positif) est la probabilité de rejeter H_0 par le test ϕ lorsque celle-ci est vraie :*
$$\alpha(\phi) := \mathbb{P}_0(\phi = 1).$$
- *L'*erreur de type II *(faux négatif) est la probabilité d'accepter H_0 par le test ϕ lorsque H_1 est vraie :*
$$\beta(\phi) := \mathbb{P}_1(\phi = 0).$$

Le problème de détection de ruptures peut ainsi être formulé comme suit :

$$(H_0) : \delta(t) = \delta_0, \quad \text{pour tout } t, \qquad (H_1) : \delta(t) = \delta_0 \mathbf{1}_{\{t \leq \tau_n\}} + \delta_1 \mathbf{1}_{\{t > \tau_n\}}, \quad \text{pour tout } t,$$

où τ_n désigne le point de changement inconnu, et δ_0, δ_1 sont des paramètres dans $(-m, +\infty)$. L'objectif est de déterminer, **à partir du graphe final non étiqueté uniquement**, si un changement a eu lieu dans le mécanisme d'attachement et, le cas échéant, d'identifier le moment τ_n où la transition de δ_0 à δ_1 s'est produite. Ce problème a des implications importantes pour comprendre l'évolution temporelle des réseaux, notamment pour identifier des moments de changements abrupts de comportement, tels que des variations dans la dynamique de croissance ou les préférences d'attachement.

Dans Banerjee et al. [2023], les auteurs ont étudié la détection des ruptures précoces. Cela correspond à la situation où $\tau_n = \lfloor cn^\gamma \rfloor$ avec $\gamma \in (0, 1)$ et $c > 0$, où n est le nombre de nœuds du graphe. Il est montré que les propriétés structurelles du graphe aléatoire obtenu, telles que la queue de la distribution des degrés, sont déterminées uniquement par les paramètres du modèle avant le point de rupture. Le test utilisé dans ce régime se base plutôt sur le degré maximal du graphe, puisque sa distribution dépend encore asymptotiquement de γ.

La détection tardive du point de changement a été étudiée dans Bet et al. [2025], où les auteurs ont construit un test basé sur les sommets de faible degré. Il a été montré que le test détecte le changement uniquement lorsque

$$\frac{n - \tau_n}{n^{1/2}} \xrightarrow[n\to\infty]{} +\infty.$$

Conjecture 1. Soit $\tau_n = n - cn^\gamma$ avec $c > 0$ et $\gamma < 1/2$. Alors :

- Aucun test basé sur la séquence des degrés n'est puissant.
- Aucun test basé sur le graphe final non étiqueté n'est puissant.

Contrairement à la démonstration de la *possibilité de détection*, qui repose sur la construction d'un test unique dont les erreurs de Type I et II peuvent être rendues arbitrairement petites, la preuve de la conjecture de *l'impossibilité de détection* nécessite une affirmation beaucoup plus forte : il faut montrer que pour tout test basé sur le graphe observé, il est impossible de rendre simultanément petites les erreurs de Type I et II. Dans ce contexte, le concept de *contiguïté* est particulièrement important, car il fournit un critère selon lequel aucun test statistique ne peut distinguer de manière fiable entre deux suites de distributions.

Définition 1.7.4 (Contiguïté des mesures de probabilité)**.** *Soit $(\Omega_n, \mathcal{F}_n)$ une suite d'espaces mesurables, et soient $(\mathbb{P}_n)$ et $(\mathbb{Q}_n)$ des suites de mesures de probabilité définies sur ces espaces. On dit que $(\mathbb{Q}_n)$ est* contiguë *par rapport à $(\mathbb{P}_n)$, noté $\mathbb{Q}_n \triangleleft \mathbb{P}_n$, si pour toute suite d'ensembles mesurables $A_n \in \mathcal{F}_n$,*

$$\mathbb{P}_n(A_n) \to 0 \quad \Longrightarrow \quad \mathbb{Q}_n(A_n) \to 0.$$

Appliqué au contexte de la détection de ruptures dans les graphes aléatoires à attachement préférentiel, si $\mathbb{P}_n$ et $\mathbb{Q}_n$ correspondent respectivement à la distribution du graphe final non étiqueté sous l'hypothèse nulle et sous l'alternative, alors la contiguïté implique (par le premier lemme de Le Cam [Vaart, 1998, Section 6.2]) qu'aucun test (éventuellement randomisé) basé sur le graphe final non étiqueté ne peut contrôler simultanément les erreurs de Type I et II : si $(\phi_n)_{n\geq 1}$ est une suite de tests basée sur le graphe observé telle

que $\mathbb{E}_{\mathbb{P}_n}(\phi_n) \to 0$, alors $\mathbb{E}_{\mathbb{Q}_n}(\phi_n) \to 0$ également.

Dans de nombreux problèmes concrets, une technique classique peut être utilisée pour prouver la contiguïté. Cette méthode repose sur le calcul du second moment du rapport de vraisemblance sous l'hypothèse nulle. Supposons que l'on teste entre une hypothèse nulle $\mathbb{P}_n$ et une hypothèse alternative $\mathbb{Q}_n$, toutes deux définies sur le même espace mesurable $(\Omega_n, \mathcal{F}_n)$. Si $\mathbb{Q}_n$ est absolument continue par rapport à $\mathbb{P}_n$, alors pour tout $A_n, B_n \in \mathcal{F}_n$:

$$\mathbb{Q}_n(A_n) \leq \mathbb{Q}_n(B_n^c) + \mathbb{P}_n(A_n)^{1/2} \mathbb{E}_{\mathbb{P}_n}\left[\left(\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right)^2 \mathbf{1}_{B_n}\right]^{1/2}.$$

Ainsi, si l'on construit une suite d'événements $(B_n)_{n\geq 1}$ dans $\mathcal{F}_n$ telle que

$$\mathbb{Q}_n(B_n^c) \to 0, \qquad \text{et} \qquad \limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}\left[\left(\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right)^2 \mathbf{1}_{B_n}\right] < +\infty,$$

alors la contiguïté de $\mathbb{Q}_n$ par rapport à $\mathbb{P}_n$ est vérifiée et aucun test puissant n'existe. L'un des principaux avantages de cette technique est qu'elle évite la nécessité de vérifier directement la définition de la contiguïté, qui pourrait impliquer de raisonner sur tous les ensembles mesurables. À la place, le problème se réduit à un calcul de moment gérable sous l'hypothèse nulle.

1.8 Contribution au problème de détection de ruptures

Dans le Chapitre 5, nous tentons de résoudre la conjecture soulevée dans Bet et al. [2025] tout en répondant à d'autres questions liées au problème de détection et de localisation des ruptures. Plus précisément, à la lumière du cadre proposé dans Bet et al. [2025], le Chapitre 5 poursuit deux objectifs : (i) Prouver que la conjecture est vérifiée au moins pour $n - \tau_n = o(n^{1/3})$, où τ_n désigne le moment où la rupture a lieu, et (ii) Étudier le problème de détection de ruptures dans le cas où le graphe étiqueté est observé. Ci-dessous, nous présentons une énonciation informelle de nos principaux résultats.

Théorème 1.8.1 (Informel)**.** *En utilisant le graphe aléatoire à attachement préférentiel non étiqueté, la détection de la rupture n'est pas possible lorsque* $n - \tau_n = o(n^{1/3})$.

Théorème 1.8.2 (Informel)**.** *En utilisant le graphe aléatoire à attachement préférentiel étiqueté, la détection de la rupture est possible si et seulement si* $n - \tau_n \to \infty$.

Les énoncés formels de ces théorèmes sont fournis au Chapitre 5. Lorsque seul le graphe non étiqueté est observé, l'application directe de la méthode du second moment est difficile en raison de l'intractabilité du rapport de vraisemblance : son calcul nécessite de marginaliser sur tous les étiquetages possibles du graphe observé, ce qui est computationnellement irréalisable. Pour surmonter cette difficulté, nous adoptons une stratégie de réduction consistant à supposer qu'un sous-ensemble des étiquettes est révélé. Cet étiquetage partiel simplifie l'analyse et permet d'obtenir des bornes tractables sur le rapport de vraisemblance. Dans le cas où le graphe étiqueté est observé, nous avons également étudié les problèmes d'estimation des paramètres et de localisation de la rupture. Plus précisément, nous avons montré que l'estimateur du maximum de vraisemblance des paramètres du modèle est consistant et asymptotiquement normal (voir Théorème 5.3.7). Nous avons également montré que la rupture peut être localisé avec une erreur d'ordre polylogarithmique (voir Proposition 5.3.9).

Chapter 2

Introduction (EN)

Contents

This thesis addresses two relatively independent statistical inference problems, both centered around the theme of inference on dependent data. This thesis is divided into two main parts. The first part focuses on the problem of clustering in parametric and non-parametric mixture models and includes Chapter 3 and Chapter 4. The second part deals with the problem of change-point detection under the preferential attachment random graph model and includes Chapter 5.

Latent variable models are statistical models that assume the presence of unobserved variables that influence the observed data. A formal definition is given below.

Definition 2.0.1. *A latent variable model is a bivariate stochastic process $(X_t, Y_t)_{t\in\mathbb{N}}$ where only the observations $(Y_t)_{t\in\mathbb{N}}$ are accessible. The random variables $(X_t)_{t\in\mathbb{N}}$, are called hidden (or latent) variables and are not observed.*

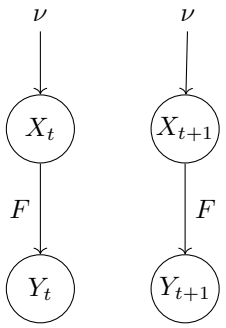
2.1 Mixture Models and Hidden Markov Models

Mixture models are a particular family of latent variable models where it is assumed that each observation is generated from **one** of several latent components, each associated with a specific distribution. Mixture models naturally arise when observations come from several groups, each with its own specific features. The latent variable then identifies the group to which an individual belongs. In this thesis, we study models in which the latent variables take a finite number of values. This is the case for finite mixture models we are interested in, which we define as follows.

Definition 2.1.1 (Finite Mixture Model)**.** *Let $\mathbb{X}$ be a finite space and $(\mathbb{Y}, \mathcal{Y})$ a measurable space. Let ν be a probability distribution over $\mathbb{X}$ and $F = (F_x)_{x\in\mathbb{X}}$ a vector of probability distributions over $\mathbb{Y}$. A process $(Y_t)_{t\in\mathbb{N}}$ follows a mixture model with parameter $\theta = (K, \nu, F)$ if there exists a sequence of (non-observed) random variables $(X_t)_{t\in\mathbb{N}}$ such that:*

- *$(X_t)_{t\in\mathbb{N}}$ are i.i.d. observations following ν, with values in $\mathbb{X}$ such that $|\mathbb{X}| = K$.*
- *Conditionally on $(X_t)_{t\in\mathbb{N}}$, the variables $(Y_t)_{t\in\mathbb{N}}$ are independent;*
- *For all $t \in \mathbb{N}$, conditionally on $\{X_t = x\}$, Y_t follows the law F_x.*

In this case, the random variables $(Y_t)_{t\in\mathbb{N}}$ are i.i.d. following the distribution $\mathbb{P}_\theta = \sum_{x\in\mathbb{X}} \nu_x F_x$ where $\nu_x = \nu(\{x\})$.



(b) Mixture model with parameter (ν, F)

Figure 2.1: Directed acyclic graph representation of a mixture model. Edge labels indicate the transition kernels.

Beyond their descriptive power, mixture models form a general-purpose tool in statistical inference, underpinning problems such as parameter estimation, unsupervised clustering, supervised classification, change-point detection, and segmentation. Their versatility has led to widespread applications in domains such as bioinformatics, econometrics, and signal processing. Foundational references include McLachlan and Peel [2000] for classical theory, Frühwirth-Schnatter [2006] for Bayesian approaches, and McLachlan and Basford [1988] for clustering applications. Recent developments are synthesized in the *Handbook of Mixture Analysis* Frühwirth-Schnatter et al. [2019], which highlights their role in modern

inference.

Another latent variable model of particular importance in this thesis is the Hidden Markov Model.

Definition 2.1.2 (Hidden Markov Model)**.** *We say that a bivariate process $(X_t, Y_t)_{t\in\mathbb{N}}$ follows a Hidden Markov Model if:*

- *The process $(X_t)_{t\in\mathbb{N}}$ is a Markov chain,*
- *Conditionally on $(X_t)_{t\in\mathbb{N}}$, the random variables $(Y_t)_{t\in\mathbb{N}}$ are independent,*
- *Conditionally on $(X_t)_{t\in\mathbb{N}}$, the distribution of Y_t depends only on X_t.*

The last two points can be summarized as

$$\mathcal{L}\left((Y_t)_{t\in\mathbb{N}} \mid (X_t)_{t\in\mathbb{N}}\right) = \bigotimes_{t\in\mathbb{N}} \mathcal{L}\left(Y_t \mid X_t\right)$$

where $\mathcal{L}(.)$ denotes the law.

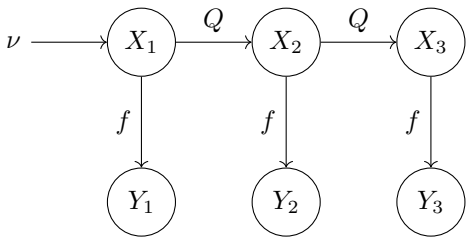
In this thesis, we will be interested mainly in homogeneous Markov chains with finite state space. See Cappé et al. [2005] for more general Hidden Markov Models. When the hidden Markov chain is homogeneous and takes only a finite number of values, the definition above translates into the following.

Definition 2.1.3 (Homogeneous finite state space HMM)**.** *Let $\mathbb{X}$ be a finite set and $(\mathbb{Y}, \mathcal{Y})$ a measurable space. Let ν be a probability distribution on $\mathbb{X}$, $Q = (Q_{x,x'})_{x,x'\in\mathbb{X}}$ a transition matrix and $F = (F_x)_{x\in\mathbb{X}}$ a vector of probability measures on $(\mathbb{Y}, \mathcal{Y})$. The process $(X_t, Y_t)_{t\in\mathbb{N}}$ follows a Hidden Markov Model (HMM) with parameter $\theta = (K, \nu, Q, f)$ if*

- *The process $(X_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain with values in $\mathbb{X}$, initial distribution ν, and transition kernel Q, such that $|\mathbb{X}| = K$,*
- *Conditionally on $(X_t)_{t\in\mathbb{N}}$, the variables $(Y_t)_{t\in\mathbb{N}}$ are independent,*
- *For all $t \in \mathbb{N}$, conditionally on $\{X_t = x\}$, Y_t follows the law F_x.*

The set $\mathbb{X}$ is called the state space *of the HMM, and $\mathbb{Y}$ is the* observation space*. Finally, the family of probability distributions $(F_x)_{x\in\mathbb{X}}$ is referred to as the emission laws.*

The HMM model allows for the modeling of dependence between the observations thanks to the Markovian structure between the hidden variables $(X_t)_{t\in\mathbb{N}}$. Note that in the case where all the lines of the transition matrix Q are equal and the initial distribution ν is the stationary distribution, the dependence structure is lost and the model boils down to the mixture model we defined in Definition 2.1.1. First introduced in the seminal work of Leonard E. Baum and his collaborators in the 1960s Baum and Petrie [1966], as a mathematical framework for analyzing sequences generated by underlying, unobservable stochastic processes, HMMs have subsequently been extensively studied from a theoretical point of view and have been used in numerous applications, from speech recognition Rabiner [1989] to climate modeling Khiatani and Ghose [2017], genomics Scharpf et al. [2008], Qin et al. [2010] and finance Guo et al. [2023], Pakštaitė et al. [2025]. They owe their success to their simplicity, flexibility, interpretability, and ease of implementation. Rabiner and Juang [1986] provide a simple introduction to Hidden Markov Models. For a more in-depth and modern treatment, the book Cappé et al. [2005] offers a comprehensive overview of both theoretical and practical aspects.



(a) Hidden Markov model with parameter (ν, Q, f)

Figure 2.2: Directed acyclic graph representation of a hidden Markov model. Edge labels indicate the transition kernels.

The primary distinction between mixture models and hidden Markov models stems from their underlying dependence structure. Although this difference may seem simple, it has significant implications for the analysis of various inference problems in both settings. Part of this thesis is dedicated to exploring some inference problems under both settings.

2.2 Identifiability: A precondition for well-posed inference

A natural question to ask before performing any inference task on any latent variable model is: given an observed process $(Y_t)_{t\in\mathbb{N}}$, is it possible to define uniquely the latent variable model that generates it? In other words, under what conditions can we guarantee the uniqueness, in some sense, of the parameter that produces this process? If such uniqueness holds, the model is said to be identifiable. Before defining identifiability, we define the notion of statististical model.

Definition 2.2.1 (Statistical model)**.** *A statistical model is a parameterized family of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ defined on a common measurable space.*

We first define identifiability in the general case.

Definition 2.2.2. *A statistical model $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ is identifiable when the map $\theta \mapsto \mathbb{P}_\theta$ is one-to-one.*

2.2.1 Identifiability under the mixture model

First, observe that in the mixture model, the distribution of the observations remains unchanged under any permutation of the hidden states. More precisely, let σ be a permutation of the latent space $\mathbb{X}$, that is, a bijection from $\mathbb{X}$ to itself. For a parameter tuple $\theta = (K, \nu, F)$, define the permuted version $\theta' = \big(K, (\nu_{\sigma(x)})_{x\in\mathbb{X}}, (F_{\sigma(x)})_{x\in\mathbb{X}}\big)$. Then, the parameters θ and θ' generate the very same mixture since $\sum_{x\in\mathbb{X}} \nu_x F_x = \sum_{x\in\mathbb{X}} \nu_{\sigma(x)} F_{\sigma(x)}$ and thus, the components labels can not be recovered. This is the reason why identifiability under the mixture model is defined up to label permutation. We define identifiability under the mixture model in what follows.

Definition 2.2.3 (Identifiability in mixture models)**.** *We identify $\mathbb{X}$ with $[K] \coloneqq \{1, \dots, K\}$ when $|\mathbb{X}| = K$. Define the mixture model*

$$\mathcal{P}_\Theta \coloneqq \left\{ \sum_{x\in[K]} \nu_x F_x \;\middle|\; \theta = (K, \nu, F) \in \Theta \right\}.$$

We say that $\mathcal{P}_\Theta$ *is* identifiable *(up to a permutation of the labels) if for any* $\theta = (K, \nu, F)$ *and* $\theta' = (K', \nu', F')$ *in* Θ

$$\sum_{x\in[K]} \nu_x F_x = \sum_{x\in[K']} \nu'_x F'_x \quad \Longrightarrow \quad \theta = \sigma(\theta')$$

for some permutation σ *of* $[K]$*, where*

$$\sigma(\theta') = \left(K', (\nu'_{\sigma(x)})_{x\in[K']}, (F'_{\sigma(x)})_{x\in[K']}\right).$$

Observe that the identifiability of mixture models is defined using only the marginal distribution of a single observation. This is because the data are i.i.d. under the mixture model; hence, the joint distribution of multiple observations carries no additional information for identifiability. It is merely the product of identical marginals. A basic drawback of mixture models is that, without additional assumptions on the mixture components, the model fails to be identifiable. For instance, one can duplicate a component by splitting it into two components with identical distributions and half the original weight, yielding a mixture that is statistically indistinguishable from the original, yet has different parameters and even a different number of components. In the absence of additional assumptions on the mixture components, the analysis of any inference problem becomes ill-posed, as the model parameters are not uniquely defined. While identifiability cannot be expected in full generality, recent work has established identifiability for certain structured classes of models. We survey below some settings where identifiability holds and where statistical inference is thus meaningful.

Finite mixture of parametric distributions

Consider the mixtures of the form $\sum_{x\in[K]} \nu_x F_x$ where the distributions $(F_x)_{x\in[K]}$ belong to a parametric family. Teicher [1963] proves the identifiability of all finite mixtures of Gamma and normal distributions. Following this work, Barndorff-Nielsen [1965] shows that a mixture of distributions that belong to the same parametric exponential family is identifiable under mild regularity conditions. Yakowitz and Spragins [1968] extend these identifiability results to several families, including the family of n products of exponential distributions, the multivariate Gaussian family, their union, the family of one-dimensional Cauchy distributions, and others.

Translation of a symmetric density

We consider here the mixtures of the form $\sum_{x\in[K]} \nu_x F(. - \mu_x)$ where F is a symmetric distribution function, in the sense that $(\forall x \in \mathbb{R})\, F(-x) + F(x) = 1$. Bordes et al. [2006] prove the nonparametric identifiability of the model for $K = 2$. For a higher number of components, Hunter et al. [2007] provide a condition that ensures the identifiability of the model parameters $(\nu_x)_{x\in[K]}$ and $(\mu_x)_{x\in[K]}$. More precisely, the following theorem is proved.

Theorem 2.2.4 (Hunter et al. [2007])**.** *The parameters* $\left((\nu_x)_{x\in[K]}, (\mu_x)_{x\in[K]}\right)$ *can be identified if and only if the following equivalence holds:*
For all $(\mu'_x)_{1\le x\le K}$*,* $(\nu'_x)_{1\le x\le K}$*, the convolution*

$$\sum_{x\in[K]} \nu_x \delta_{\mu_x} \star \sum_{x\in[K]} \nu'_x \delta_{-\mu'_x}$$

is symmetric if and only if

$$\sum_{x\in[K]} \nu'_x \delta_{-\mu'_x} = \sum_{x\in[K]} \nu_x \delta_{-\mu_x}.$$

They deduce that when $K = 2$, identifiability of the parameters holds if and only if $\nu_1 \notin \{0, 1/2, 1\}$. Although the class of symmetric densities is suitable in certain contexts, it is often too restrictive to capture the complexity of many practical models. It turns out that introducing a specific form of dependence between observations can restore identifiability for translation mixtures, provided one considers the joint distribution of two consecutive observations. See Remark 2.2.8.

Multidimensional Mixtures

We consider now multidimensional mixture models where the observed variables are divided into at least $d \geq 3$ blocks. In this context, the joint distribution of the observations can be written as:

$$\sum_{x \in [K]} \nu_x \bigotimes_{j=1}^{d} F_{x,j}, \tag{2.1}$$

where, for each $x \in [K]$, the distributions $(F_{x,j})_{j \in [d]}$ are probability measures over d possibly distinct spaces. Initial identifiability results appeared in the literature for low-dimensional cases such as $K = 2$ and $d = 2$ or $d = 3$ (see Hall and Zhou [2003]). Further results have considered the case where K is unknown and $d \geq 2$ (see Hall et al. [2005]). A foundational result from Kruskal [1977] shows that model (2.1) is identifiable when $d = 3$, assuming the probability measures are supported on finite sets. Building on this, Allman et al. [2009] showed that identifiability holds for all $d \geq 3$ provided that for each $j = 1, \ldots, d$, the family of measures $F_{1,j}, \ldots, F_{K,,j}$ are linearly independent. Some spectral arguments will be at the heart of the proof of identifiability for other models (see Anandkumar et al. [2012], Gassiat et al. [2016]) and the construction of estimators (see Abraham et al. [2025, 2022]).

2.2.2 Identifiability under the Hidden Markov Model

To begin with, note that the distribution of a single observation in a HMM can be represented as a mixture model. Consequently, any conditions that guarantee identifiability in mixture models also ensure identifiability of at least the emission distributions and mixing proportions in the HMM. However, recovering the transition matrix requires information from the joint distribution of more than one single observation, which is the focus of this section. Importantly, the HMM structure makes it possible to relax the identifiability conditions that are necessary under the mixture model setting. It is also important to note that as in mixture models, the distribution of the observations remains unchanged under a relabeling of the hidden states. More precisely, let $\theta = (K, \nu, Q, F)$ and let $\theta' = (K, \nu', Q', F')$ where

$$\big(\forall (x, x') \in \mathbb{X}\big) \quad \nu'_x = \nu_{\sigma(x)}, \quad Q'_{x,x'} = Q_{\sigma(x),\sigma(x')}, \quad F'_x = F_{\sigma(x)}.$$

Then, denoting by $\mathbb{P}_\theta^{(n)}$ the joint distribution of $(Y_i)_{i \in [n]}$ and using the properties of the HMM, one can easily show that

$$(\forall y_{1:n} \in \mathbb{Y}^n) \;\; \mathbb{P}_\theta^{(n)}(y_{1:n}) = \sum_{x_{1:n} \in \mathbb{X}^n} \nu_{x_1} Q_{x_1,x_2} \ldots Q_{x_{n-1},x_n} \prod_{i=1}^{n} F_{x_i}(y_i). \tag{2.2}$$

From the expression above, it is straightforward that $(\forall y_{1:n} \in \mathbb{Y}^n) \;\; \mathbb{P}_\theta^{(n)}(y_{1:n}) = \mathbb{P}_{\theta'}^{(n)}(y_{1:n})$. This means that even if we make use of the joint distribution of more than one observation, the labels of the components can not be identified. As a result, the identifiability results one can hope to obtain must always be interpreted modulo a permutation of the labels,

unless a canonical labeling is available. Therefore, in what follows, when $|\mathbb{X}| = K$, we will assume without loss of generality that $\mathbb{X} = [K]$. We define identifiability under the Hidden Markov Model as follows.

Definition 2.2.5 (Identifiability in Hidden Markov Models)**.** *We identify* $\mathbb{X}$ *with* $[K] := \{1, \dots, K\}$ *when* $|\mathbb{X}| = K$*. Define the Hidden Markov Model*

$$\mathcal{P}_\Theta := \left\{ \mathbb{P}_\theta^{(n)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}.$$

where $\mathbb{P}_\theta^{(n)}$ *is the joint distribution of* $Y_{1:n}$ *under the HMM with parameter* θ*. We say that* $\mathcal{P}_\Theta$ *is* identifiable *(up to a permutation of the labels) from the distribution of* n *observations if for any* $\theta = (K, \nu, Q, F)$ *and* $\theta' = (K', \nu', Q', F')$ *in* Θ

$$\mathbb{P}_\theta^{(n)} = \mathbb{P}_{\theta'}^{(n)} \quad \Longrightarrow \quad (\exists \sigma \in \mathcal{S}_K) \;\; \theta = \sigma(\theta')$$

where

$$\sigma(\theta') := \left(K, \left(\nu'_{\sigma(x)} \right)_{x \in [K]}, \left(Q'_{\sigma(x), \sigma(x')} \right)_{x, x' \in [K]}, \left(F'_{\sigma(x)} \right)_{x \in [K]} \right).$$

It is important to note that our focus is on identifiability from the joint distribution of multiple observations, as the hidden Markov structure becomes useful only in this context. When considering a single observation, no identifiability results beyond those of mixture models can be expected.

The foundational results for HMM identifiability trace back to the works Baum and Petrie [1966], Petrie [1969], who established identifiability for discrete HMMs. They showed that if the number of hidden states K is known and the emission distributions are linearly independent, then the model is identifiable. However, these results were restricted to finite observation alphabets. A major breakthrough came with the application of tensor decomposition methods, particularly Kruskal's theorem in Kruskal [1977] to HMM identifiability. Allman et al. [2009] applied this theorem to prove the identifiability of some discrete HMMs (in a sense that is slightly different from definition 2.2.5) under mild rank conditions. This approach introduced a powerful algebraic framework for identifiability, treating the joint distribution over a finite number of observations as a low-rank tensor and using uniqueness of tensor decompositions to identify parameters. Subsequent work extended identifiability results beyond the parametric setting. Notably, Gassiat et al. [2016] proved identifiability for nonparametric HMMs under the assumption that the emission distributions are distinct and linearly independent. They have proved the following theorem.

Theorem 2.2.6 (Gassiat et al. [2016])**.** *Let* $(X_t, Y_t)_{t \in \mathbb{N}}$ *be a HMM with parameter* $\theta \in \Theta$*. Assume* Θ *is the set of parameters* θ *such that:*

- $(\forall x \in [K]) \quad \nu_x > 0$*,*
- Q *is invertible,*
- *The probability measures* $(F_x)_{x \in [K]}$ *are linearly independent,*

then, the model $\mathcal{P}_\Theta = \left\{ \mathbb{P}_\theta^{(3)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}$ *is identifiable, where* $\mathbb{P}_\theta^{(3)}$ *stands for the distribution of* (Y_1, Y_2, Y_3)*.*

In other words, under mild regularity assumptions, the HMM parameters can be identified from the joint distribution of three consecutive observations. The assumption of linear independence can be relaxed to the simpler condition that the emission distributions are distinct from one another, at the price of considering the joint distribution of more than three observations, as detailed in the next theorem.

Theorem 2.2.7 (Alexandrovich et al. [2016b]). *Let $(X_t, Y_t)_{t\in\mathbb{N}}$ be a HMM with parameter $\theta \in \Theta$. Assume Θ is the set of parameters θ such that:*

- *Q is invertible, irreducible and aperiodic (see Norris [1997] for a formal definition),*
- *The probability measures $(F_x)_{x\in[K]}$ are distinct from one another,*

then, the model $\mathcal{P}_\Theta = \left\{ \mathbb{P}_\theta^{((2K+1)(K^2-2K+2)+1)} \;\middle|\; \theta = (K, \nu, Q, F) \in \Theta \right\}$ is identifiable in the sense of definition 2.2.5.

Remark 2.2.8. It is important to note that identifiability is also ensured in other models that do not belong to the framework of mixture models and HMMs. Consider for example observations with marginal distribution $\sum_{x\in[K]} \nu_x F(. - m_x)$ but that are not independent. More precisely, assume the observations are derived of the model

$$Y_i = m_{X_i} + \varepsilon_i \tag{2.3}$$

where $(\varepsilon_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. $\mathbb{R}-$valued random variables and $(m_j)_{j\in[K]}$ are real numbers. In Gassiat and Rousseau [2016], the authors show that in case the latent variables $(X_i)_{i\in\mathbb{N}}$ are not independent, model (2.3) is identifiable without any assumption on F. More precisely, if the process $(X_i)_{i\in\mathbb{N}}$ takes K distinct values and if F is any probability distribution, then, provided that the translation parameters are distinct and the matrix Q representing the joint distribution of (X_1, X_2) has full rank, one may recover $\left(K, Q, (\nu_x)_{x\in[K]}, (m_x)_{x\in[K]}, F\right)$ from the distribution of (Y_1, Y_2). Of course, the centers $(m_x)_{x\in[K]}$ are identifiable up to a translation, unless one of the centers is specified in advance. Note also that no assumption is put on F. The only structural assumption is that Q has full rank. Notice that, with only two latent states (that is, $K = 2$), assuming that Q has full rank is the same as assuming that the variables X_1 and X_2 are dependent which is a very simple condition for identifiability.

With the standard conditions for the identifiability of mixture models and Hidden Markov Models now established, the problem of parameter inference is well-posed. The following section presents a review of estimation methods and algorithms developed for these two classes of models.

2.3 Estimation in Mixture Models and Hidden Markov Models

In this section, we survey the principal estimation techniques used for the finite mixture model and the Hidden Markov Model.

2.3.1 Expectation-Maximization (EM) Algorithm

Let $\theta \in \Theta$ be the model parameter and $\mathbb{P}_\theta^{(n)}$ the joint distribution of the observations $Y = (Y_i)_{i\in[n]}$. Let $X = (X_i)_{i\in[n]}$ the associated latent variables. Denote the likelihood of the observed data under parameter θ by $L_n(\theta, Y)$. A natural estimator of θ in the parametric setting is the Maximum Likelihood estimator. The Expectation-Maximization (EM) algorithm, introduced by Dempster et al. [1977], is a widely used iterative procedure for approximating the maximum likelihood estimator when a direct maximization of the likelihood is not feasible. This is done usually in latent variable models including mixture and Hidden Markov Models. The maximum likelihood estimator (MLE) is defined as:

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta, Y).$$

Because the latent variables X are unobserved, direct maximization is usually challenging. The EM algorithm addresses this by iteratively maximizing the expected complete-data log-likelihood, defined through the following steps starting from an initial point $\theta^{(0)} \in \Theta$:

- **E-step (Expectation):** Compute the intermediate quantity:

$$R(\theta, \theta^{(j)}) = \mathbb{E}_{\theta^{(j)}}[\log L_n(\theta, (X, Y)) \mid Y],$$

  which is the conditional expectation of the complete-data log-likelihood given the observed data Y, under the current parameter estimate $\theta^{(j)}$.

- **M-step (Maximization):** Update the parameter by maximizing this expectation:

$$\theta^{(j+1)} \in \arg\max_{\theta \in \Theta} R(\theta, \theta^{(j)}).$$

These steps are repeated until convergence, typically when the relative increase in the likelihood is smaller than a tolerance $\varepsilon > 0$:

$$\frac{L_n(\theta^{(j)}, Y) - L_n(\theta^{(j-1)}; Y)}{L_n(\theta^{(j-1)}, Y)} < \varepsilon.$$

The output of the algorithm is then $\theta^{(j)}$. Wu [1983] proved that, under mild regularity conditions, the sequence of iterates converges to a stationary point of the likelihood function. However, EM algorithm does not guarantee convergence to a global maximum. It can get stuck in suboptimal local maxima, particularly in high-dimensional spaces or poorly conditioned problems. Since the algorithm is deterministic, its output depends heavily on the initial parameter $\theta^{(0)}$. A common practical strategy is to run the EM algorithm multiple times from different initializations and retain the solution with the highest likelihood. For standard parametric distributions, the E- and M-steps often admit closed-form expressions, allowing the algorithm to be implemented efficiently through explicit recursive updates. In the case of HMMs, the two steps of the EM algorithm admit a closed formula. However, these recursive formula depend on the *a posteriori* distribution of the hidden states which can be computed using the celebrated Forward-Backward algorithm Cappé et al. [2005]. For rigorous theoretical guarantees on the convergence of this algorithm, see Balakrishnan et al. [2017].

2.3.2 Spectral estimation

Spectral algorithms are a class of algorithms that estimate the parameters of latent variable models by exploiting the algebraic structure of low-order moment tensors or matrices, such as second and third order moments. These methods typically avoid non-convex optimization by relying on linear algebra tools, such as eigenvalue and singular value decompositions, and tensor decomposition. Recent developments, particularly in Anandkumar et al. [2012], have used this technique for the estimation of the parameters of some parametric Hidden Markov Models. This technique was then refined to cover also the estimation under nonparametric Hidden Markov Models. More precisely, consider a nonparametric HMM with parameter $\theta = (K, \nu, Q, F)$ and assume ν is the stationary distribution of the transition matrix Q of the hidden Markov chain and that the emission distributions have densities with respect to a common dominating measure $\mathcal{L}$ on $(\mathbb{Y}, \mathcal{Y})$. Let $F = (F_k)_{k \in [K]}$ where $F_k = f_k d\mathcal{L}$. Denote by $\left(\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L}), \|.\|\right)$ the Hilbert space of square integrable functions on $\mathbb{Y}$ with respect to the measure $\mathcal{L}$ equipped with the usual inner product $\langle ., . \rangle$

on $\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$. Let $(\varphi_a)_{a\in\mathbb{N}}$ be an orthonormal basis of $\mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$. For $a, b, c \in \mathbb{N}$, consider:

$$\begin{aligned}\mathbf{L}(a) &= \mathbb{E}[\varphi_a(Y_1)]\\ \mathbf{N}(a,b) &= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)]\\ \mathbf{P}(a,c) &= \mathbb{E}[\varphi_a(Y_1)\varphi_c(Y_3)]\\ \mathbf{M}(a,b,c) &= \mathbb{E}[\varphi_a(Y_1)\varphi_b(Y_2)\varphi_c(Y_3)]\end{aligned}$$

Using Equation (2.2), one can express the quantities above in terms of the initial distribution ν, the transition matrix Q and the coordinates of the projections of the emission distributions on the basis $(\varphi_a)_{a\in\mathbb{N}}$. Using simple algebraic operations such as diagonalizations and singular value decompositions and exploiting the assumptions of identifiability, these expressions can be inverted and the model parameters can be expressed in terms of $\mathbf{L}$, $\mathbf{N}$, $\mathbf{P}$ and $\mathbf{M}$. Estimation is then achieved by substituting the true joint distributions with their empirical counterparts derived from the data. This approach has the key advantage of being non-iterative, avoiding local optima and initialization sensitivity inherent in EM-like methods. This framework underlies the estimation procedures developed in Anandkumar et al. [2012], Gassiat et al. [2016], De Castro et al. [2017], where consistent estimation is ensured through spectral and algebraic techniques.
More recently, Abraham et al. [2022] introduced a spectral kernel density estimator that refines earlier spectral methods to control the estimation error in the supremum norm rather than the $\mathbf{L}^2$ norm. This refinement specifically targets the case of Hidden Markov Models with two hidden states. The authors establish the following theorem.

Theorem 2.3.1. *Under standard assumptions ensuring identifiability of the HMM and assuming the emission densities to be $s-$Hölder regular, there exists an estimator $\left(\hat{f}_j\right)_{j\in[K]}$ and a permutation τ such that for C large enough*

$$\mathbb{P}_\theta\left(\|\hat{f}_j - f_{\tau(j)}\|_\infty \geq C\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right) \xrightarrow[n\to+\infty]{} 0$$

Convergence in expectation also holds: for some $C' > 0$,

$$\mathbb{E}\left[\|\hat{f}_j - f_{\tau(j)}\|_\infty\right] \leq C'\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}$$

A refined version of this estimator will be used in Chapter 3 for the construction of a plug-in clustering procedure. Under the same setting of nonparametric HMM with two hidden states, Abraham et al. [2025] proposed a wavelet block thresholding estimator that was shown to be adaptive to the smoothness of each density.

2.3.3 Penalized least squares estimation

The least squares paradigm allows the construction of minimax adaptive estimation procedures in the nonparametric setting. In this section, we describe a method for estimating the parameters of the Hidden Markov Model using the penalized least-squares approach. Define :

$$(\forall k \in [K])\,(\forall M \in \mathbb{N}^\star) \quad f_{M,k} = \sum_{m=1}^{M} \langle f_k, \varphi_m\rangle \varphi_m$$

Let $\theta = (K, \nu, Q, (f_k)_{k\in[K]})$ be the parameters of a Hidden Markov Model. Let g_θ denote the joint density of a triple of observations (Y_1, Y_2, Y_3):

$$g_\theta(y_1, y_2, y_3) = \sum_{k_1,k_2,k_3=1}^{K} \pi_{k_1} Q_{k_1,k_2} Q_{k_2,k_3} f_{k_1}(y_1) f_{k_2}(y_2) f_{k_3}(y_3)$$

For any square-integrable function $t \in \mathbf{L}^2(\mathbb{Y}, \mathcal{Y}, \mathcal{L})$, define the contrast functional:

$$C(t) = \|t - g_\theta\|_2^2 = \|t\|_2^2 - 2\langle t, g_\theta\rangle + \|g_\theta\|_2^2.$$

which is minimal for $t = g_\theta$. Since $\|g_\theta\|_2^2$ is constant with respect to t, minimizing $C(t)$ is equivalent to minimizing $\|t\|_2^2 - 2\langle t, g_\theta\rangle$. Given a sample $(Y_s)_{s\in[N+2]}$ of observations of a HMM, we thus consider the empirical contrast function:

$$\gamma_N(t) = \|t\|_2^2 - \frac{2}{N}\sum_{s=1}^{N} t(Y_s, Y_{s+1}, Y_{s+2}),$$

Assume we are given an estimator $\hat{Q}$ of Q. For example, one may use the spectral estimator of the previous section. Based on $\hat{Q}$, we consider $\mathcal{S}(\hat{Q}, M)$ the collection of functions g_θ such that $\theta = \left(K, \hat{\nu}, \hat{Q}, f_M\right)$ where $\hat{\nu}$ is the stationary distribution of $\hat{Q}$ and $f_M = (f_{M,k})_{k\in[K]}$. We then define

$$\hat{g}_M = \underset{t\in\mathcal{S}(\hat{Q},M)}{\arg\min}\ \gamma_N(t)$$

To choose the optimal value of M, we seek to minimize the squared error $\|\hat{g}_M - g^\star\|_2^2$. Since $\|g^\star\|_2^2$ is fixed, this is approximately equivalent to minimizing $\gamma_N(\hat{g}_M)$. However, to account for the stochastic fluctuations of the empirical process γ_N, we introduce a penalty function $\mathrm{pen}(N, M)$. We then select:

$$\hat{M} = \underset{M=1,\dots,N}{\arg\min}\ \left\{\gamma_N(\hat{g}_M) + \mathrm{pen}(N, M)\right\}.$$

With this selection, the final penalized least-squares estimator is defined by:

$$\hat{g} := \hat{g}_{\hat{M}}.$$

It is important to note that this procedure assumes prior knowledge of the HMM order K. In Lehéricy [2019], the author improves upon this approach by introducing a consistent estimator for the order K. Beyond establishing consistency for the order estimation, the proposed method also yields estimators for the HMM parameters that are minimax adaptive, up to a logarithmic factor, with respect to the global regularity of the model A more refined approach that achieves state-by-state minimax adaptive estimation is presented in Lehéricy [2018].

2.3.4 Penalized Maximum Likelihood Estimator

Under the HMM framework, the maximum likelihood estimator (MLE) is defined as the parameter value maximizing the likelihood of the observations, namely the function

$$L_n : (\theta, y_{1:n}) \longmapsto \mathbb{P}_\theta^{(n)}(y_{1:n}),$$

where $\theta = (K, \nu, Q, F)$ denotes the model parameter and $\mathbb{P}_\theta^{(n)}$ is the joint distribution of $Y_{1:n}$ under θ. The construction of the MLE under the HMM framework follows the

following steps. First, consider a collection of parametric models $\left(\mathcal{M}_M\right)_{M\in I}$, where each $\mathcal{M}_M$ consists of n-fold product measures on the observation space $(\mathbb{Y}, \mathcal{Y})$. For each $M \in I$, define

$$\Theta_{K,M} = \{\theta = (K, \nu, Q, F) : F \in \mathcal{M}_M\}.$$

Let $y_{1:n}$ denote a fixed realization of the random vector $Y_{1:n}$. For fixed K and M, the likelihood maximizer is given by

$$\hat{\theta}_{n,K,M} \in \underset{\theta\in\Theta_{K,M}}{\arg\max}\, L_n(\theta, y_{1:n}).$$

To select among models, introduce a penalization term $\mathrm{pen}_n(K, M)$ and choose

$$(\hat{K}_n, \hat{M}_n) \in \underset{(K,M)\in\mathbb{N}^*\times I}{\arg\max} \left\{\frac{1}{n}\log L_n\left(\hat{\theta}_{n,K,M}, y_{1:n}\right) - \mathrm{pen}_n(K, M)\right\}.$$

The final estimator is then $\hat{\theta} = \hat{\theta}_{n,\hat{K}_n,\hat{M}_n}$.

The first theoretical guarantees for this estimator were established in Vernet [2015], where posterior consistency and concentration rates of a Bayesian nonparametric MLE are derived. Alexandrovich et al. [2016a] proved the consistency of a nonparametric MLE constructed from hidden Markov models with finite state space and nonparametric mixtures of parametric densities. In the misspecified setting, Lehéricy [2021] showed that the nonparametric MLE under the hidden Markov model recovers the best approximation of the underlying misspecified distribution.

2.4 Some inference problems in HMMs and Mixture Models

2.4.1 Clustering

Clustering is defined as the task of recovering the random partition $\Pi_n = \{A_1, \ldots, A_m\}$ of the index set $\{1, \ldots, n\}$, where the partition is induced by the latent states $X_{1:n} = (X_1, \ldots, X_n)$, such that for $k \in [\![1, m]\!]$, $i, j \in A_k$ if and only if $X_i = X_j$. The goal is to infer this partition based solely on the observed data $Y_{1:n} = (Y_1, \ldots, Y_n)$.

Definition 2.4.1. *A n-clusterer g is a measurable map from the space of observed data $\mathbb{Y}^n$ to the set of partitions of $[n]$, denoted by $\mathcal{P}[n]$:*

$$g : \mathbb{Y}^n \to \mathcal{P}[n].$$

The set of all n-clusterers is denoted by $\mathcal{G}_n$.

Below, we provide an overview of several widely used clustering algorithms. For an in-depth discussion and theoretical guarantees on the risk of clustering of these procedures, we direct the reader to Chapter 12 of Giraud [2021].

K-means clustering

Definition 2.4.2 (K-means Clustering)**.** *Let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ and let $K \geq 1$ an integer.* K-means clustering *seeks a partition $\{C_1, \ldots, C_K\}$ and centroids $\{\mu_1, \ldots, \mu_K\} \subset \mathbb{R}^d$ that minimize the total within-cluster variance:*

$$\sum_{k=1}^{K}\sum_{j\in C_k} \|Y_j - \mu_k\|^2, \quad \textit{where } \mu_k = \frac{1}{|C_k|}\sum_{j\in C_k} Y_j.$$

K-means optimizes a non-convex objective and is NP-hard in general. In practice, Lloyd's algorithm is used. It assigns iteratively the points to their nearest centroids and updates the centroids.

Definition 2.4.3 (Lloyd's Algorithm)**.** *Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ and let $K \in \mathbb{N}$ a predefined number of clusters. Lloyd's algorithm iteratively updates a set of cluster centers $\{\mu_1, \dots, \mu_K\}$ and an assignment function $\varphi : \{1, \dots, n\} \to \{1, \dots, K\}$ as follows:*

Algorithm 3: Lloyd's Algorithm

Input: *Observations $Y_1, \dots, Y_n \in \mathbb{R}^d$, number of clusters K*
Output: *Cluster assignments $\varphi^{(t)}$ and centers $\mu_1^{(t)}, \dots, \mu_K^{(t)}$*

Initialize: *Choose initial centers $\mu_1^{(0)}, \dots, \mu_K^{(0)}$;*
repeat
  Assignment step:
  for $i = 1$ **to** n **do**
    $\varphi^{(t)}(i) \leftarrow \arg\min_{k \in \{1,\dots,K\}} \|Y_i - \mu_k^{(t)}\|^2$
  Update step:
  for $k = 1$ **to** K **do**
    $\mu_k^{(t+1)} \leftarrow \frac{1}{|\{i : \varphi^{(t)}(i) = k\}|} \sum_{i : \varphi^{(t)}(i) = k} Y_i$
until *Convergence is reached;*

The algorithm terminates when the assignments $\varphi^{(t)}$ no longer change or the decrease in the within-cluster sum of squares falls below a fixed threshold. Lu and H. Zhou [2016] establish theoretical guarantees for the performance of this algorithm on Sub-Gaussian observations. Theoretical guarantees for other variants of the K-means algorithm have also been established. We refer to Ndaoud [2022] for a variant of Lloyd algorithm initialized by the spectral clustering algorithm and Giraud and Verzelen [2018] for a relaxed version of the K-means algorithm.

Spectral clustering

Spectral clustering is a widely used algorithm for uncovering latent group structures in data. In the context of Gaussian mixture models, it offers a simple yet powerful procedure for estimating clusters based solely on the observed data. The spectral algorithm typically serves as a first-stage clustering step that produces a coarse group assignment, which can later be refined using a more specialized method.

Let $Y_1, \dots, Y_n \in \mathbb{R}^d$ be i.i.d. observations from a GMM with K components. Denote by $\mathbf{Y} = [Y_1, \dots, Y_n]^\top \in \mathbb{R}^{n \times d}$ the data matrix. The core idea of the spectral algorithm is to compute the top K eigenvectors of the Gram matrix $\mathbf{Y}\mathbf{Y}^\top \in \mathbb{R}^{n \times n}$, and then use these eigenvectors to define a new representation of the data that reveals the clustering structure. The theoretical justification stems from the fact that, under a well-separated GMM, the expectation of the Gram matrix admits the following decomposition:

$$\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] = A\Theta\Theta^\top A^\top + \Gamma,$$

where:

- $A \in \mathbb{R}^{n \times K}$ is the membership matrix, defined by $A_{ik} = \mathbf{1}_{\{i \in G_k\}}$ with G_k being the set of indices belonging to cluster k,

- $\Theta \in \mathbb{R}^{K \times d}$ contains the means of the Gaussian components,
- $\Gamma \in \mathbb{R}^{n \times n}$ is a diagonal matrix accounting for the within-cluster covariance.

Thus, $\mathbb{E}[\mathbf{Y}\mathbf{Y}^\top]$ consists of a low-rank (rank K) signal matrix structured by the cluster assignment and an additive noise term. This motivates the use of the best rank-K approximation of $\mathbf{Y}\mathbf{Y}^\top$ as a proxy for the clustering structure.

Let $\mathbf{Y}\mathbf{Y}^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ be the eigendecomposition of the Gram matrix, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Define the rank-$K$ approximation by:

$$\left(\mathbf{Y}\mathbf{Y}^\top\right)_{(K)} = \sum_{k=1}^{K} \lambda_k v_k v_k^\top.$$

The spectral embedding of the data consists of the top K eigenvectors $v_1, \ldots, v_K \in \mathbb{R}^n$. These eigenvectors can be stacked column-wise to form a matrix $V \in \mathbb{R}^{n \times K}$, where each row corresponds to a low-dimensional representation of an observation. A clustering algorithm, such as K-means, is then applied to the rows of V to recover the group structure. The following algorithm summarizes this clustering procedure:

Algorithm 4: Spectral Clustering under a Gaussian Mixture Model

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$, number of clusters K
Output: Group affectation of each observation
Step 1: Compute Gram matrix
$\mathbf{G} \leftarrow \mathbf{Y}\mathbf{Y}^\top$
Step 2: Eigendecomposition of Gram matrix
Compute eigenvalues and eigenvectors:
$\mathbf{G} = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$
Step 3: Construct rank-K approximation
$\mathbf{G}_{(K)} \leftarrow \sum_{k=1}^{K} \lambda_k v_k v_k^\top$
Step 4: Form spectral embedding
$V \leftarrow [v_1 \; v_2 \; \cdots \; v_K] \in \mathbb{R}^{n \times K}$
Step 5: Cluster embedded points
Apply K-means clustering to the rows of V to obtain cluster labels
return *Affectation of each observation to its group*

Löffler et al. [2021] establish theoretical guarantees on the performance of the spectral clustering algorithm in Gaussian mixture models with isotropic covariance. It is shown that the algorithm achieves minimax optimal performance, provided the number of clusters is fixed and the signal-to-noise ratio is sufficiently high.

Model-based clustering

Model-based clustering McNicholas [2016] is a statistical approach that relies on finite mixture models to perform clustering. Let $\mathbf{Y} = [Y_1, \ldots, Y_n]^\top \in \mathbb{R}^{n \times d}$ denote a collection of n observations, assumed to be independently drawn from a finite mixture distribution with density

$$f(y) = \sum_{k=1}^{K} \pi_k f_k(y, \alpha_k),$$

where $\forall k \in [K]$, $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. In typical clustering applications, the component densities $f_k(y, \alpha_k)$ are assumed to share the same functional form, so that $f_k(y, \alpha_k) = f(y, \alpha_k)$ for all $k \in [K]$, where f is a known parametric family. The full set of

model parameters is denoted by $\theta = \big(\pi, (\alpha_k)_{k\in[K]}\big)$. Under this formulation, the likelihood function of the observed data is

$$L_n(\theta; \mathbf{Y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f(Y_i, \alpha_k).$$

Once the parameters are estimated, typically via the Expectation-Maximization (EM) algorithm, each observation is assigned to the component maximizing the posterior probability:

$$\forall i \in [n], \quad \hat{z}_i := \arg\max_{k\in[K]} \frac{\hat{\pi}_k f(Y_i, \hat{\alpha}_k)}{\sum_{j=1}^{K} \hat{\pi}_j f(Y_i, \hat{\alpha}_j)}.$$

Clustering is then performed based on the partition defined by the estimated labels $\hat{z}_i$.

This methodology extends naturally to Hidden Markov Models (HMMs). Once the model parameters have been estimated, computing the posterior distributions of the latent variables becomes more challenging because of the temporal dependencies among hidden states, which complicate the likelihood structure. Nevertheless, the Forward–Backward algorithm Cappé et al. [2005] offers an efficient way to evaluate these posteriors. This in turn allows one to adopt a clustering approach analogous to that of parametric mixture models, by using the partition induced by the inferred hidden states. Concretely, the latent variables are estimated as

$$\forall i \in [n], \quad \hat{z}_i := \arg\max_{k\in[K]} \phi_{\hat{\theta},i|1:n}(k),$$

where $\hat{\theta}$ denotes the estimator of the HMM parameters and $\phi_{\hat{\theta},i|1:n}$ is the posterior distribution of the hidden state associated with the i-th observation given the observed sequence $Y_{1:n}$ under parameter $\hat{\theta}$. Since a clusterer can be obtained from a classifier by retaining only the induced partition, it is natural to use the partition generated by $(\hat{z}_i)_{i\in[n]}$ for clustering. However, this widespread practice among HMM practitioners lacks rigorous decision-theoretic justification: it essentially imitates the Bayes classifier (the risk minimizer for classification) rather than the Bayes clusterer (the risk minimizer for clustering). Formal definitions are provided in Chapter 3. The issue is practically important, since the Bayes classifier has a closed-form expression and is generally more tractable computationally. Whenever the Bayes clusterer can be derived from the Bayes classifier, this yields a clear computational advantage. Yet, the relationship between the Bayes clusterer and the Bayes classifier has not been studied before. A key part of Chapter 3 is devoted to investigating and clarifying this connection.

2.4.2 Other inference problems

In addition to classical tasks such as estimation, classification, and clustering, several other inference problems arise in the context of mixture models and Hidden Markov Models. A central inference task in HMMs is to recover the sequence of hidden states $X_1, \ldots, X_n$ given the observed sequence $Y_1, \ldots, Y_n$. Unlike the Bayes classifier, which assigns to each observation the most likely hidden state via the marginal Maximum a posteriori classifier, recovering the full sequence of hidden states requires a different approach. Due to the dependence structure imposed by the hidden Markov chain, the joint posterior distribution $\mathcal{L}(X_{1:n} \mid Y_{1:n})$ is not simply the product of the individual posteriors $\mathcal{L}(X_t \mid Y_{1:n})$. As a result, a sequence formed by taking the most probable state at each time point may not correspond to a valid path under the model, especially when certain state transitions are not allowed. To address this, the Viterbi algorithm is used and which allows identifying,

through dynamic programming, the sequence maximizing the a posteriori distribution of the joint distribution of the hidden states given the observations. See Cappé et al. [2005]. The algorithm differs from the Maximum a posteriori in the type of optimality that is sought. The Maximum a posteriori classifier maximizes the marginal posterior for each hidden state independently, while the Viterbi algorithm identifies the most probable *entire* sequence of states by maximizing the joint posterior distribution. The latter is particularly important in applications where temporal consistency and model constraints must be respected, as it ensures the inferred path is valid under the HMM's transition dynamics. See Rabiner [1989], Forney [1973].

2.5 Contributions to the problem of clustering

In Chapters 3 and 4, our contributions concern mainly the problem of clustering. We will tackle several fundamental questions which will be answered under the I.I.D and HMM model. While our primary focus is the nonparametric setting, we will also study the Gaussian case in order to identify the exact dependence of the Bayes risk of clustering with respect to the model parameters and the improvement allowed by the dependence structure in terms of the clustering performance (See Chapter 4). In Chapter 3, it will be assumed that there exists δ absolute constant such that $\min_{x,x'} Q_{x,x'} \geq \delta$ and $\min_x \nu_x \geq \delta$ where Q is the transition matrix of the hidden chain and ν is the initial distribution, which guarantees that the hidden process sufficiently explores all states. This assumption will be relaxed in Chapter 4. Our contribution to the problem of clustering can be summarized in what follows:

1. **Relationship between Bayes classifier and Bayes clusterer:** The Bayes classifier and the Bayes clusterer are defined as the minimizers of the risks of classification and clustering, respectively. A central question that arises is whether there exists a clear relationship between these two objects, and if so, under which structural assumptions on the data this relationship can be made precise. Addressing this question is of practical relevance: the Bayes classifier often enjoys a closed-form expression and is therefore significantly easier to compute, while the Bayes clusterer may be more abstract and less directly accessible. Consequently, whenever the Bayes clusterer can be recovered from the Bayes classifier, this not only provides conceptual insight but also yields concrete computational advantages. Our analysis reveals a rather striking result. Clustering induced by the Bayes classifier coincides with that of the Bayes clusterer if and only if the observations are i.i.d. samples from a mixture distribution with exactly two components. In every other setting—namely mixtures with more than two components or dependent data generated by Hidden Markov Models—there always exists a parameter configuration under which the clustering produced by the Bayes classifier differs from that of the Bayes clusterer with positive probability. These findings are rigorously established in Theorems 3.3.1, 3.3.4, 3.3.6, and 3.3.8 of Chapter 3.

2. **Magnitude of Bayes risks:** Under what conditions is the Bayes risk of clustering comparable to the Bayes risk of classification? This question is of particular significance because the Bayes risk of classification admits a closed-form expression, which makes it amenable to thorough analysis. This analytic tractability suggests that, whenever the Bayes risks of classification and clustering are of comparable order, any theoretical bound established for the former can be directly transferred to the latter. Such a connection would allow us to leverage the relatively simpler structure

of the classification problem in order to gain insights into the more intricate clustering problem. Our results show that this comparability holds only in two distinct regimes: either when the number of clusters is restricted to two, or when the Bayes risk of classification does not decay at an exponential rate with respect to the sample size n. Outside of these scenarios, the two risks behave in fundamentally different ways, and extrapolation is no longer valid. These findings are established rigorously for both i.i.d. and Hidden Markov model (HMM) observations. See Theorems 3.3.2, 3.3.5, 3.3.7, 3.3.9 and Corollary 1.

3. **Dependence on the model parameters:** A central aspect of our analysis lies in understanding how the Bayes risks depend on the underlying model parameters, and in particular on the population densities that characterize the distributions of the observations. Such an understanding is crucial because it allows us to isolate the correct notion of *separation* between distributions that effectively captures the intrinsic difficulty of the inference problem at hand. From a practical standpoint, this perspective is highly valuable: once the Bayes risk can be expressed in terms of this separation measure, ensuring that the risk stays within a desired magnitude amounts to verifying a simple condition on how well-separated the underlying distributions are. This provides a direct and intuitive bridge between abstract risk bounds and concrete conditions on the model. In Theorem 3.3.10 and Corollary 4, we succeed in identifying the fundamental quantity that governs the difficulty of both clustering and classification problems, in the settings of i.i.d. mixtures as well as Hidden Markov models (HMMs). This key quantity is given by

$$\Lambda := \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left(\sum_{x \neq x_0} f_x(y) \right) d\mathcal{L}(y),$$

where $(f_x)_{x\in\mathbb{X}}$ denote the population densities of the distributions $(F_x)_{x\in\mathbb{X}}$ with respect to a dominating measure $\mathcal{L}$ on the observation space $\mathbb{Y}$, i.e., $dF_x = f_x d\mathcal{L}$. In particular, we will show that in our regime of interest,

$$\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \approx \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \asymp \Lambda.$$

See Corollary 4 of Chapter 3. This expression reveals, in a unified way, how the structure of the problem and the overlap between population densities dictate the Bayes risks. In the special case of two populations ($J = 2$), the formula simplifies to

$$\Lambda = 1 - \|F_1 - F_2\|_{\text{TV}},$$

where $\|F_1 - F_2\|_{\text{TV}}$ denotes the total variation distance between the two distributions.

4. **Learnability of clustering:** Is it possible to design a clusterer—without knowledge of the true model parameters—whose excess risk vanishes as the number of observations grows? That is, can such a method achieve clustering performance comparable to that of the Bayes clusterer asymptotically? We provide a positive answer to this question by constructing a clustering procedure that achieves near-optimal performance. Specifically, we show that when the emission densities are s-Hölder regular and under additional regularity assumptions, there exists a clustering procedure $\tilde{g}$ such that

$$\mathcal{R}_n^{\text{clust}}(\theta, \tilde{g}) - \inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) = \mathcal{O}\left(\left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}} \right).$$

This result establishes a rigorous theoretical justification for employing the plug-in Bayes classifier in practice. For practitioners dealing with Hidden Markov Models (HMMs), the key message is that, despite the conceptual difference between the Bayes classifier and the Bayes clusterer, the Bayes classifier can be confidently used as a surrogate for clustering. Its excess risk is not only theoretically bounded but also remains negligible in relevant regimes, ensuring that the performance loss compared to the Bayes clusterer is limited. To reinforce this conclusion, we present numerical experiments in a nonparametric framework, which clearly illustrate that clustering procedures based on HMMs are capable of uncovering latent structures even in situations where conventional methods fail to do so.

5. **Tight characterization of the Bayes risk:** How does the dependence structure of the Hidden Markov Model improve the performance of clustering? We will illustrate the added value of dependence in the regime where the hidden Markov chain is slowly mixing and the emission densities are Gaussian. This is due to a tight characterization of the dependence of the Bayes risk of clustering with respect to the model parameters. We will also build Bayes optimal clustering procedures. This is the subject of Chapter 4.

Following our study of clustering, we were naturally led to explore other inference problems in mixture models and HMMs, mainly change-point detection and segmentation. As we were delving into the literature on this subject, we came across a conjecture posted on the arXiv paper Bet et al. [2025] concerning change-point detection in a completely different setting: preferential attachment random graphs. This unexpected encounter triggered our interest in solving the conjecture and we consequently managed to study the problem of change-point detection under the preferential attachment random graph model.

2.6 The preferential attachment random graph model

The preferential attachment (PA) model, along with its various extensions, has become one of the most widely used frameworks for modeling randomly growing graphs in network science. These models are widely used to study real-world networks such as social networks, citation networks, and the World Wide Web, where new nodes tend to connect to previous ones. Over the past two decades, extensive research has rigorously analyzed the structural properties of PA models, resulting in well-established findings regarding asymptotic degree distributions, the local structure, and more. We refer readers to the comprehensive books van der Hofstad [2016, 2024]. First introduced by Barabási and Albert [1999a], the PA model has attracted significant theoretical and practical attention due to its straightforward, local growth process and its ability to naturally produce power-law degree distributions, a feature commonly observed in real-world networks.
The preferential attachment mechanism defines a sequence of random multigraphs $(G_t)_{t\geq 1}$ on vertex sets $\{1,\dots,t\}$. There is no loss of generality in assuming that these graphs are directed, using the convention that the arrows go from vertices with largest labels to vertices with smallest labels. To be somewhat more precise, in the next, a *labeled graph* refers to the following definition.

Definition 2.6.1 (Labeled graph)**.** *A labeled (multi)graph $\mathfrak{g}$ is a couple $(\mathcal{V},\mathcal{E})$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}\subset\mathcal{V}^2$ is the multiset of directed edges, with no loop allowed. For an edge $(u,v)\in\mathcal{E}$, we use the convention that the arrow goes from u to v, and we write for simplicity $u\to_{\mathfrak{g}} v$ for $(u,v)\in\mathcal{E}$.*

We recall that in a multigraph, two vertices can be connected by more than one edge. Many versions of the preferential attachment mechanism exist. The simplest model is the Barabási–Albert model which is defined below.

Definition 2.6.2 (Barabási–Albert model)**.** *The* Barabási–Albert (BA) model *is a model for generating preferential attachment random graphs as follows:*

- *Start from G_1 the graph with one single vertex labeled* 1*.*
- *A vertex labelled* 2 *attaches to vertex* 1 *to form G_2.*
- *For $n \geq 3$, G_n is obtained from G_{n-1} inductively, by attaching the newly added vertex n to a random vertex of label $v \in \{1, \ldots, n-1\}$ with probability proportional to $d_i(G_{n-1})$, the degree of node i in G_{n-1}:*

$$\mathbb{P}\left(n \to_{G_n} v \mid G_{n-1}\right) = \frac{d_v(G_{n-1})}{\sum_{j=1}^{n-1} d_j(G_{n-1})}$$

While the canonical model is the Barabási–Albert (BA) model, several variants have been developed to capture more nuanced properties observed in real-world networks. For example, the Barabási–Albert model can naturally be adapted to generate *multigraphs*, where multiple edges between two nodes are allowed. The core mechanism remains unchanged; the key difference is that, at each time step, the newly added node forms m edges instead of just one. These m connections are made by selecting existing nodes—either simultaneously or one after another—according to the same preferential attachment rule. This allows for repeated edges and reflects more complex connectivity patterns observed in real-world networks. The following model is commonly used for this purpose.

Definition 2.6.3 (Preferential Attachment Model with Out-Degree m)**.** *The preferential attachment (PA) model with out-degree m and attachment function f generates a sequence of graphs $(G_t)_{t \in \mathbb{N}^\star}$ on vertex sets $V_t = \{1, \ldots, t\}$, defined inductively as follows:*

- *Initialize G_1 to be a graph on a single vertex* 1 *with no edges.*
- *Let G_2 be the graph with two vertices* 1 *and* 2 *attached by m edges.*
- *For $t \geq 3$ and given G_{t-1}, define a sequence of intermediate graphs:*

$$G_{t,0}, G_{t,1}, \ldots, G_{t,m}$$

  where:

  - *$G_{t,0}$ is the graph G_{t-1} with a new isolated vertex labeled t.*
  - *For each $i \in [m]$, construct $G_{t,i}$ from $G_{t,i-1}$ by adding an edge between v_t and an existing vertex $v \in V_{t-1}$, sampled according to the distribution:*

$$\mathbb{P}(v_{t,i} = v \mid G_{t,i-1}) = \frac{f(d_v(G_{t,i-1}))}{\sum_{j=0}^{t-1} f(d_j(G_{t,i-1}))}$$

    where $d_v(G)$ denotes the degree of vertex v in graph G.

- *Define $G_t := G_{t,m}$.*

When $f(x) = x$, we talk about the linear preferential attachment with out-degree m.
When $f(x) = x + \delta$, we talk about the affine preferential attachment with out-degree m.
When f is constant, we talk about uniform attachment (UA).

Note that when f is not linear, the model is more challenging to analyze because, unlike linear models, where the denominator in the attachment probability is deterministic and can be computed explicitly, the denominator here does not admit a simple closed-form expression. This is the reason why the major part of the literature on the subject focuses on the linear preferential attachment model. The affine preferential attachment model generalizes the Barabási–Albert model by providing an additional degree of freedom that allows adjusting the bias toward high-degree nodes thanks to a tunable parameter δ. More precisely, the probability of attachment becomes:

$$\mathbb{P}\left(n \to_{G_n} v \mid G_{n-1}\right) = \frac{d_v(G_{n-1}) + \delta}{\sum_{j=1}^{n-1} \left(d_j(G_{n-1}) + \delta\right)}$$

For example, high values of δ mitigate the *preferential effect* because they ensure that even low-degree nodes have a non-trivial chance of receiving new links. When δ is very large, the preferential property disappears, resulting in a uniform attachment model where each node is selected with equal probability at every step. Other variants of the model allow the parameter δ to vary over time, meaning its value can change during the graph construction process.

One of the properties of preferential attachment graphs is the scale-free property. This refers to the fact that the asymptotic degree distribution of the resulting network follows a power-law, meaning the probability that a node has degree k decays like $k^{-\gamma}$ for $\gamma > 1$. This heavy-tailed behavior is a distinctive feature of many real-world networks, and it is one of the main reasons the preferential attachment model has gained prominence in network science. Empirical evidence shows that scale-free structures appear in a wide range of complex systems. For instance, the World Wide Web exhibits a power-law in the number of hyperlinks per page; citation networks are dominated by a few highly cited papers; social networks include individuals with exceptionally high connectivity; and in biological systems, such as protein–protein interaction networks, a small number of proteins participate in many interactions. The PA model captures this broad heterogeneity through a simple yet powerful *rich-get-richer* mechanism: nodes with higher degrees are more likely to attract new links. We illustrate this in Figure 2.3 with different choices of the attachment function. For additional examples and discussion, see Chapter 1 of van der Hofstad [2016].

2.7 Inference problems under the preferential attachment random graph model

In this section, we review the main inference problems that are usually studied under the preferential attachment random graph model. We assume one observes only the unlabeled preferential attachment random graph. See Definition 5.2.3 for a formal definition of the unlabeled graph.

2.7.1 Asymptotic degree distribution

One of the key properties of the PA model is its asymptotic behavior: the degree distribution of the nodes converges to a heavy-tailed, often power-law form. This section reviews the mathematical foundation behind this phenomenon and highlights how different choices in the attachment rule influence the asymptotic distribution.

It is worth noting that the power-law degree distribution does not typically arise in other standard random graph models such as uniform attachment and Erdős–Rényi graphs (see Figure 2.4), a usual model we recall below.

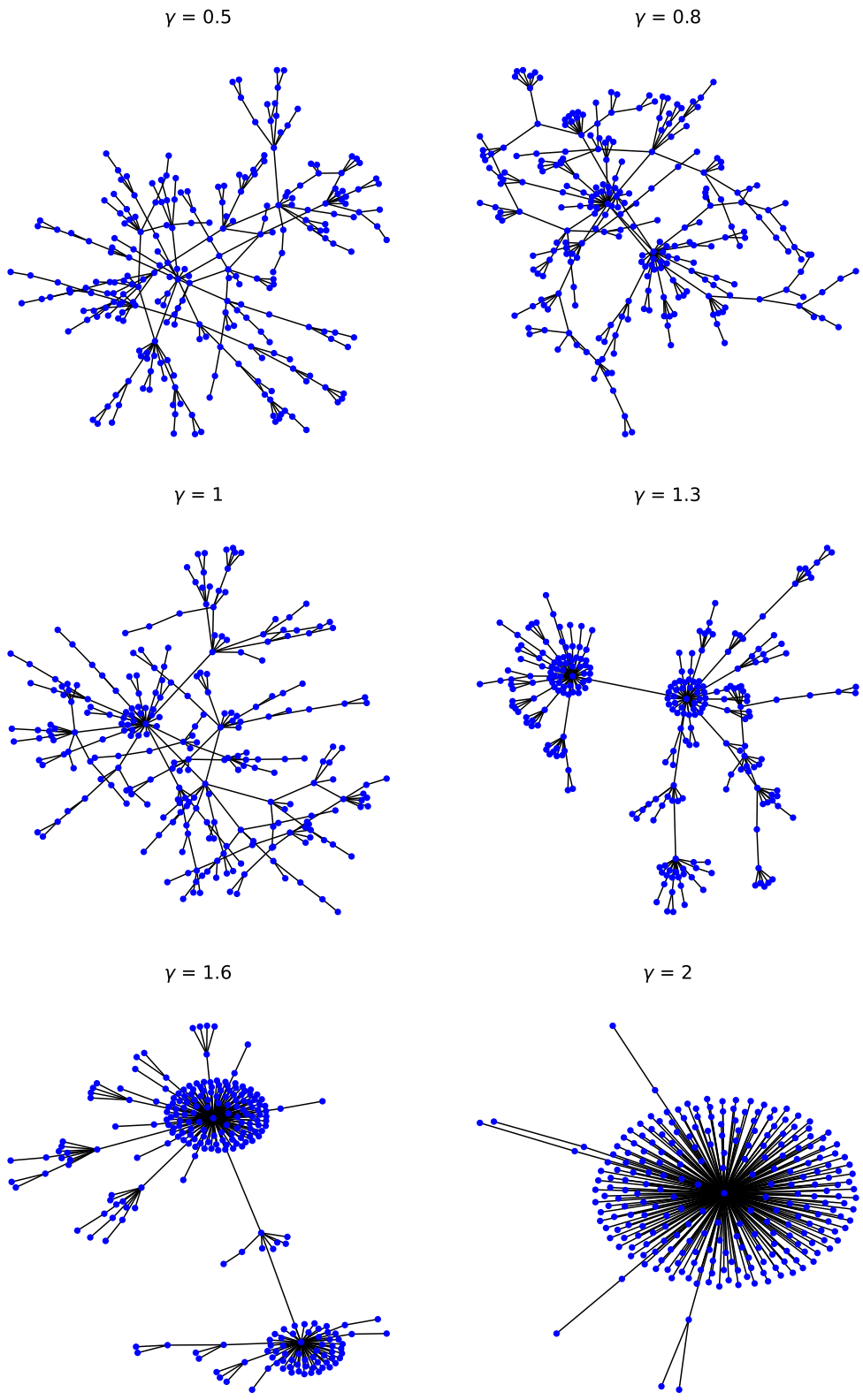


Figure 2.3: Preferential Attachment Graphs with attachment function $f(x) = x^{\gamma}$. Each graph contains 250 nodes.

Definition 2.7.1 (Erdős–Rényi Random Graph Model)**.** *The* Erdős–Rényi *(ER) random graph model $G(n, p)$ is a random undirected graph on n vertices where, for each pair of distinct vertices $i, j \in \{1, \ldots, n\}$, an edge $\{i, j\}$ is present in the graph with probability p, independently of all other edges.*

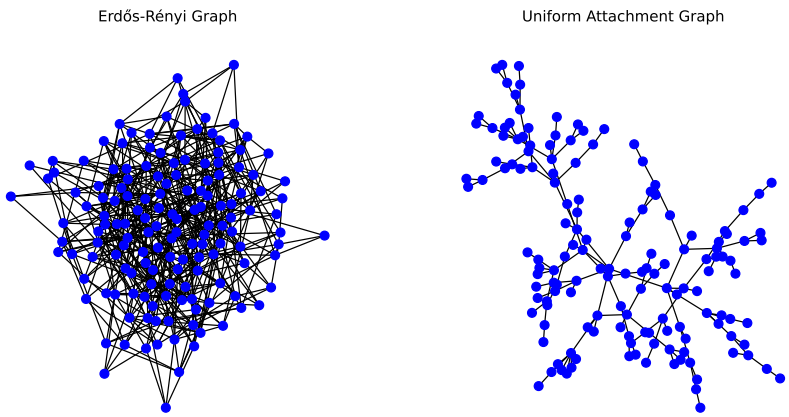


Figure 2.4: Erdős–Rényi graph with parameter $p = 0.05$ and Uniform Attachment graph. Each graph contains $n = 150$ nodes.

Let $P_k(n)$ denote the proportion of nodes in G_n with degree k. The key question is whether this quantity stabilizes as $n \to \infty$, and if so, what form it takes. The following theorem identifies this limiting distribution for the linear preferential attachment model with out-degree m (this corresponds to the attachment function $f(x) = x$. See Definition 2.6.3) and is due to Bollobás et al. [2001]. Clearly, the limiting distribution exhibits a power-law tail with exponent $\gamma = 3$, an empirical evidence that was first reported in Barabási and Albert [1999b].

Theorem 2.7.2 (Bollobás et al. [2001])**.** *For $k \geq m$ and as $n \to +\infty$, $P_k(n) \to p_k$ in probability, where*

$$p_k = \frac{2m(m+1)}{k(k+1)(k+2)}$$

For sublinear attachment (that is $f(x) \ll x$), the rich-get-richer mechanism is weaker. The degree distribution in this case decays faster than any power law, typically as a *stretched exponential* Krapivsky et al. [2000], Dereich and Mörters [2009]:

$$p_k \asymp \exp(-c \cdot k^{\alpha}).$$

with $\alpha \in (0, 1)$. This regime produces more homogeneous graphs, where hubs are rare and most nodes have comparable degrees. In the superlinear regime (that is $f(x) \gg x$), a condensation phenomenon happens: one dominant vertex emerges, connecting to nearly every other vertex, while all other vertices have only finite degrees. This phenomenon is rigorously formalized and proven in Oliveira and Spencer [2005]. The derivation of these results employs several key techniques: recursive relations for tracking the expected number of nodes of degree k, martingale decomposition methods, embedding the discrete preferential attachment process into continuous-time branching or birth processes.

2.7.2 Diameter

Another property that has been extensively studied is that of *diameter*.

Definition 2.7.3 (Diameter of a Graph)**.** *Let $G = (\mathcal{V}, \mathcal{E})$ be a connected graph. The* diameter *of G, denoted by* $\mathrm{diam}(G)$*, is defined as the maximum shortest-path distance between any pair of vertices:*

$$\mathrm{diam}(G) = \max_{u,v \in \mathcal{V}} d_G(u, v),$$

where $d_G(u, v)$ is the shortest-path distance (i.e., the minimal number of edges) between nodes u and v. When G is not connected, its diameter is that of its largest connected component.

2.7.3 Maximal degree

The maximal degree in a random graph refers to the highest degree of any node in the network. It captures the presence and influence of hubs and is a key indicator of heterogeneity. Studying it helps understand how centralized or uneven a network is. In Table 2.1 below, we summarize the key properties of the usual graph models. Let

$$p_k(m, \delta) = (2 + \delta/m) \frac{\Gamma(k + \delta)\Gamma(m + 2 + \delta + \delta/m)}{\Gamma(m + \delta)\Gamma(k + 3 + \delta + \delta/m)}$$

Note that for k large, $p_k(m, \delta) \approx C(m, \delta)k^{-\left(3+\frac{\delta}{m}\right)}$.

Model	Asymp. Deg. Dist.	Maximal Degree	Diameter
ER$(n, \lambda/n)$ $(\lambda > 1)$	Poisson(λ)	$\Theta\left(\frac{\log n}{\log\log n}\right)$ Móri [2005]	$\Theta(\log n)$ Durrett [2006]
UA	$\left(2^{-k}\right)_{k\geq 1}$ Janson [2005]	$\Theta(\log n)$ Devroye and Lu [1995]	$\Theta(\log n)$ Pittel [1994]
BA	$\left(\frac{4}{k(k+1)(k+2)}\right)_{k\geq 1}$	$\Theta(n^{1/2})$	$\Theta(\log n)$ Bollobás and Riordan [2004]
Linear PA $(m > 1)$	$\left(\frac{2m(m+1)}{k(k+1)(k+2)}\right)_{k\geq m}$	$\Theta(n^{1/2})$	$\Theta\left(\frac{\log n}{\log\log n}\right)$ Bollobás and Riordan [2004]
Affine PA $(m, \delta > 0)$	$(p_k(m, \delta))_{k\geq m}$	$\Theta(n^{\frac{1}{2+\delta/m}})$	$\Theta(\log n)$ Dommers et al. [2010]

Table 2.1: Asymptotic properties of some random graph models.

The results with no references in the table above, together with their corresponding proofs, can be found in van der Hofstad [2016].

2.7.4 Network archaeology in recursive random graphs

In recursive random graphs, such as PA graphs and UA graphs, network archaeology refers to the problem of identifying the order of arrival of the nodes (or part of the nodes) of the observed final graph. Of course, this assumes the observed final graph is unlabeled, because otherwise, the order can be inferred from the labels. This problem includes:

- **Root finding:** This corresponds to identifying the first vertex (often denoted as vertex 1) added to the graph during its construction. Since exact root identification is usually impossible, the idea is to identify the smallest confidence set that contains the root with high probability. This problem was studied in Bubeck et al. [2017] in the case of UA and PA. For PA graphs, a minimal size of the confidence set for root finding was identified in Bubeck et al. [2017] and an optimal algorithm was proposed in Contat et al. [2024] and shown to be optimal. Khim and Loh [2016] have studied root finding in the case where the tree is obtained by diffusion on an infinite regular tree and Brandenberger et al. [2022] in the case of a size-conditionned Galton-Watson tree.

- **Arrival time estimation:** While identifying the first vertex offers only partial insight into the graph's evolution, reconstructing the entire arrival order of nodes can be far more informative. This information is particularly valuable for tracing the spread of misinformation, rumors, or even viruses through a network. In Crane and Xu [2021], a general framework for network archaeology is developped and it can be applied to the problem of inferring arrival times. More recently, this problem was studied in Briend et al. [2025] in the case of UA and linear PA where an order estimator was introduced and was shown to be optimal with respect to a family of risk measures.

2.7.5 Change-point detection and localization

First, note that this problem can not be formalised under the uniform attachment mechanism or the linear attachment mechanism because the attachment mechanism remains the very same during the process of graph construction. A simple model where this problem can be studied is the affine preferential attachment (PA) model (see Definition 2.6.3). Under this model, at each time step, a new vertex joins the graph and forms m edges to existing vertices, with the attachment probability influenced by a function $\delta(t)$ that modifies the linear preference based on vertex degree. More formally, the probability that a new vertex at time t connects to an existing vertex v is proportional to $d_{G_{t-1}}(v) + \delta(t)$, where $d_{G_{t-1}}(v)$ is the degree of v in G_{t-1}. To analyze this rigorously, one introduces an intermediate process $((G_{t,i})_{i=1}^{m})_{t\geq 1}$, where each $G_{t,i}$ corresponds to the intermediate state of the graph after i of the m new edges from vertex t have been added. This formulation allows precise modeling of the probabilistic attachment rule at each edge addition. The final graph G_t is then obtained by setting $G_t = G_{t,m}$. The central statistical question addressed in this setting is whether the attachment mechanism, encoded in $\delta(t)$, remains constant over time or undergoes a structural change. The problem of change-point detection can be defined as a hypothesis testing problem, which we recall below:

Definition 2.7.4 (Hypothesis test)**.** *Let $(\mathbb{Y}, \mathcal{Y}, \mathbb{P})$ be a measurable sample space and let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures on $\mathbb{Y}$ corresponding to the null hypothesis H_0 and the alternative hypothesis H_1, respectively. A* statistical test *is a measurable function*

$$\phi : \mathbb{Y} \to \{0, 1\},$$

where $\phi(y) = 1$ indicates rejection of the null hypothesis H_0 based on observation $y \in \mathbb{Y}$.

- *The* Type I error *(false positive) is the probability of rejecting H_0 when it is true:*

$$\alpha(\phi) := \mathbb{P}_0(\phi = 1).$$

- *The* Type II error *(false negative) is the probability of accepting H_0 when H_1 is true:*

$$\beta(\phi) := \mathbb{P}_1(\phi = 0).$$

The change-point detection problem can be formulated as a hypothesis testing problem:

$$(H_0) : \delta(t) = \delta_0, \quad \text{for all } t, \qquad (H_1) : \delta(t) = \delta_0 \mathbf{1}_{\{t \leq \tau_n\}} + \delta_1 \mathbf{1}_{\{t > \tau_n\}}, \quad \text{for all } t$$

where τ_n denotes the unknown change-point, and δ_0, δ_1 are parameters in $(-m, +\infty)$. The objective is to determine, **using only the final unlabeled graph**, whether a change has taken place in the attachment mechanism, and if so, to identify the point in time τ_n at which the transition from δ_0 to δ_1 occurred. This problem has significant implications for understanding the temporal evolution of networks, particularly in identifying moments of abrupt behavioral shifts, such as changes in growth dynamics or preferential tendencies. In Banerjee et al. [2023], the authors studied early change-detection. It corresponds to the situation where $\tau_n = \lfloor cn^\gamma \rfloor$ with $\gamma \in (0, 1)$ and $c > 0$ where n is the number of nodes on the graph. It is shown that the structural properties of the resulting random graph, such as the degree distribution tail, are determined solely by the model parameters prior to the change-point. The test used in this regime is based rather on the maximal degree of the graph, since its distribution still depends on γ asymptotically (see 2.1). Late change-point detection was studied in Bet et al. [2025], where the authors built a test based on low-degree vertices. The test was shown to detect the change only when

$$\frac{n - \tau_n}{n^{1/2}} \xrightarrow[n \to \infty]{} +\infty.$$

The authors also conjectured the following.

Conjecture 2. Let $\tau_n = n - cn^\gamma$ with $c > 0$ and $\gamma < \frac{1}{2}$. Then,

- All tests based on the sequence of degrees are powerless.
- All tests based on the unlabeled final graph are powerless.

Unlike the proof of *possibility of detection*, which relies on constructing a single test whose Type I and Type II errors can both be made arbitrarily small, the proof of the conjecture of the *impossibility of detection* requires a much stronger claim: one must show that for every test based on the observation, it is not possible to simultaneously make both Type I and Type II errors small. In this context, the concept of *contiguity* is particularly important, as it provides a criterion under which no statistical test can reliably distinguish between two sequences of distributions.

Definition 2.7.5 (Contiguity of Probability Measures)**.** *Let $(\Omega_n, \mathcal{F}_n)$ be a sequence of measurable spaces, and let $(\mathbb{P}_n)$ and $(\mathbb{Q}_n)$ be sequences of probability measures defined on these spaces. We say that $(\mathbb{Q}_n)$ is* contiguous *with respect to $(\mathbb{P}_n)$, written $\mathbb{Q}_n \triangleleft \mathbb{P}_n$, if for every sequence of measurable sets $A_n \in \mathcal{F}_n$,*

$$\mathbb{P}_n(A_n) \to 0 \quad \Longrightarrow \quad \mathbb{Q}_n(A_n) \to 0.$$

Applying this to the context of change-point detection in preferential attachment random graphs, if $\mathbb{P}_n$ and $\mathbb{Q}_n$ correspond respectively to the distribution of the observed unlabeled graph under the null hypothesis and under the alternative, then contiguity implies (by Le Cam's first lemma [Vaart, 1998, Section 6.2]) that no (eventually randomized) test made on the basis of the observed unlabeled graph is capable of controlling both Type I and Type II error rates simultaneously: if $(\phi_n)_{n \geq 1}$ is a sequence of tests based on the

observed unlabeled graph such that $\mathbb{E}_{\mathbb{P}_n}(\phi_n) \to 0$, then $\mathbb{E}_{\mathbb{Q}_n}(\phi_n) \to 0$ as well.

In many concrete problems, a usual technique known as the *second moment trick* can be used for the proof of contiguity. The second moment trick is based on computing the second moment of the likelihood ratio under the null distribution. Suppose we are testing between a null hypothesis $\mathbb{P}_n$ and an alternative hypothesis $\mathbb{Q}_n$, both defined on the same measurable space $(\Omega_n, \mathcal{F}_n)$. If $\mathbb{Q}_n$ is absolutely continuous with respect to $\mathbb{P}_n$, then for any $A_n, B_n \in \mathcal{F}_n$

$$\mathbb{Q}_n(A_n) \leq \mathbb{Q}_n(B_n^c) + \mathbb{P}_n(A_n)^{1/2} \mathbb{E}_{\mathbb{P}_n}\left[\left(\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right)^2 \mathbf{1}_{B_n}\right]^{1/2}.$$

Hence, if we build a sequence of events $(B_n)_{n\geq 1}$ in $\mathcal{F}_n$ such that

$$\mathbb{Q}_n(B_n^c) \to 0, \qquad \text{and,} \qquad \limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}_n}\left[\left(\frac{\mathrm{d}\mathbb{Q}_n}{\mathrm{d}\mathbb{P}_n}\right)^2 \mathbf{1}_{B_n}\right] < +\infty,$$

then contiguity of $\mathbb{Q}_n$ with respect to $\mathbb{P}_n$ holds and thus, no powerful test exists. One of the main advantages of the second moment trick is that it avoids the need to verify the definition of contiguity directly, which may involve reasoning about all measurable sets. Instead, it reduces the problem to a manageable moment computation under the null distribution.

2.8 Contribution to the problem of change-point detection

In Chapter 5, we try to solve the conjecture raised in Bet et al. [2025] together with answering other questions related to the problem of change-point detection and localization. More precisely, in light of the framework of Bet et al. [2025], Chapter 5 has two goals: (i) Prove the conjecture holds at least for $n - \tau_n = o(n^{1/3})$ where τ_n is the moment where change occurs, and (ii) Study the problem of change-point detection in the situation where the labeled graph is observed. More precisely, below is an informal statement of our main results.

Theorem 2.8.1 (Informal)**.** *Using the unlabeled preferential attachment random graph, detection of the change-point is not possible when* $n - \tau_n = o(n^{1/3})$.

Theorem 2.8.2 (Informal)**.** *Using the labeled preferential attachment random graph, detection of the change-point is possible if and only if* $n - \tau_n \to \infty$.

The formal statements of these theorems are provided in Chapter 5. When only the unlabeled graph is observed, directly applying the second moment trick is challenging due to the intractability of the likelihood ratio: computing it requires marginalizing over all possible labellings of the observed graph, which is computationally infeasible. To overcome this difficulty, we adopt a reduction strategy in which we assume that a subset of the labels is revealed. This partial labeling simplifies the analysis and allows us to obtain tractable bounds on the likelihood ratio. In the setting where the labeled graph is observed, we also studied the problems of parameter estimation and change-point localization. More precisely, we have shown that the maximum likelihood estimator of the model parameters is consistent and asymptotically normal (see Theorem 5.3.7). We have also shown that the change-point can be localized with an error of polylogarithmic order (see Proposition 5.3.9).

Chapter 3

Clustering and Classification risks in non-parametric Hidden Markov Models

We conduct an in-depth analysis of the Bayes risk of clustering in the context of Hidden Markov and i.i.d. models. In both settings, we identify the situations where this risk is comparable to the Bayes risk of classification and those where its minimizer, the Bayes clusterer, can be derived from the Bayes classifier. While we demonstrate that clustering based on the Bayes classifier does not always match the optimal Bayes clusterer, we show that this difference is primarily theoretical and that the Bayes classifier remains nearly optimal for clustering. A key quantity emerges, capturing the fundamental difficulty of both classification and clustering tasks. Furthermore, by leveraging the identifiability of HMMs, we establish bounds on the clustering excess risk of a plug-in Bayes classifier in the general nonparametric setting, offering theoretical justification for its widespread use in practice. Simulations further illustrate our findings.

This chapter is based on the paper Gassiat et al. [2025], co-authored with Elisabeth Gassiat and Zacharie Naulet and which will appear in *Annals of Statistics*.

Contents

3.1 Introduction

Clustering is the problem of organizing a collection of objects into groups, ensuring that elements within each group exhibit greater *similarity* to each other than to those in other groups. Clustering plays a crucial role in various domains such as machine learning, pattern recognition and image processing, helping to reveal the hidden structure of data without requiring prior labels of the observations. In contrast, classification, while closely related to clustering, seeks not only to group similar objects together but also to assign class labels to them.

Mixture models are a common framework in which the two problems can be formally defined. In these models, observations $\mathbf{Y} = (Y_1, Y_2, \dots)$ are independent conditional on unobserved random variables $\mathbf{X} = (X_1, X_2, \dots)$ taking values in $\mathbb{X} = \{1, \dots, J\}$ that represent the labels of the classes in which observations originated, with J being the total number of classes. In this work, we consider the settings where the latent variables $\mathbf{X}$ are i.i.d. or form a Markov Chain. Since the i.i.d. setting is a strict subcase of the Markovian case, without loss of generality we consider the general model

$$\begin{aligned} Y_i \mid \mathbf{X} &\overset{\text{ind}}{\sim} F_{X_i} \qquad i = 1, 2, \dots \\ \mathbf{X} &\sim \text{Markov}(\nu, \mathrm{Q}) \end{aligned} \tag{3.1}$$

with parameter $\theta = (\nu, Q, (F_x)_{x \in \mathbb{X}})$ where ν is the initial distribution of the chain (in the next identified with probability vectors on $\mathbb{X}$), Q the transition matrix of the hidden chain and $(F_x)_{x \in \mathbb{X}}$ are the emission distributions (*aka.* populations). In particular, the case of i.i.d. latent variables is obtained by restricting the parameters of the previous model to transition matrices Q having identical lines equal to ν. The reason why we emphasize the i.i.d. subcase in the paper is not only because of its wide use, but also because some interesting phenomena emerge only in the absence of dependencies. More details on the model are given in Section 3.2.2.

In this context, clustering is the task of recovering the partition $\{A_1, \dots, A_m\}$ of $\{1, \dots, n\}$ induced by the labels (*aka.* groups), namely $i, j \in A_k \iff X_i = X_j$. Measuring the loss incured by a clusterer g requires defining a notion of similiarity between two partitions, which can be achieved in many ways. Here we make the choice of using the misclassification error metric Meilă and Heckerman [2001], Meilă [2005] which is based on

measuring the best overlap achievable by matching the elements of the two partitions. A formal definition of the loss is stated later in Section 3.2.3. Pursuing the goal of establishing solid decision-theoretic foundations for clustering, we aim at characterizing the *Bayes clusterer*, namely the oracle clusterer g_θ that minimizes the risk of clustering $g \mapsto \mathcal{R}_n^{\text{clust}}(\theta, g)$ when the parameter θ is known. The Bayes clusterer remains relatively unexplored, computationally demanding, and generally does not have an explicit formula. A widely held belief among statisticians is that using the *Bayes classifier* for clustering is a reasonnable approach. The Bayes classifier solves the seemingly related problem of minimizing the risk of classification $h \mapsto \mathcal{R}_n^{\text{class}}(\theta, h)$, defined using the loss function that counts the number of missclassified observations. Note that classification aims at completely identifying the hidden variables $X_1, \dots, X_n$ based on the observed variables $Y_1, \dots, Y_n$, which is infeasible without prior knowledge of some labels (supervised learning). In contrast with the Bayes clusterer, though, the Bayes classifier has a well-defined closed-form solution and has been extensively studied Devroye et al. [1996]. Since a clusterer can be obtained from a classifier by retaining only the partition induced by the groups, it is common-sense to use the Bayes classifier as a clusterer. This practice, however, while common amongst HMM practictionners, does not have solid decision-theoretic grounds. To the best of our knowledge, the relation between Bayes clusterer and Bayes classifier has never been investigated in the past. We fill this gap in the literature, by establishing somewhat curious results. In particular, we address the following questions:

1. Is there any clear link between the Bayes classifier and the Bayes clusterer? If so, under what condition?

2. When can the Bayes risk of clustering be comparable to that of classification? What is the proper measure of separation that quantifies the difficulty of both classification and clustering problems in a general nonparametric context?

3. Given that practitioners of Hidden Markov Models commonly use the plug-in Bayes classifier for clustering, is there any theoretical justification for this approach?

We answer these questions in the context of the model (3.1) with a focus on both the i.i.d. and dependent subcases. We voluntary establish a dichotomy between those two cases, to emphasize certain phenomena that occur only in the dependent case.

To answer Question 1, we show that surprisingly, the clustering using the Bayes classifier equals that of the Bayes clusterer if and only if observations are i.i.d. from a mixture of two components. In the remaining situations (more than two components or dependent HMMs), there is always a set of parameters for which the clusterer obtained from the Bayes classifier and the Bayes clusterer differ with non-zero probability. See Theorems 3.3.1, 3.3.4, 3.3.6 and 3.3.8. Below is an informal statement of these theorems:

Theorem 3.1.1 (Informal)**.** *The Bayes classifier and clusterer coincides in the i.i.d. setting with $J = 2$. If $J \geq 3$ or the labels are dependent, then there exist distributions for which Bayes classifier and Bayes clusterer differ.*

The Question 2 is of interest because, thanks to the closed formula of the Bayes risk of classification, it can be easily analyzed and when the two risks are comparable, any bound on the risk of classification can be extrapolated to the risk of clustering. We show that surprisingly, the Bayes risk of clustering is comparable to the Bayes risk of classification in only two situations: when there are two clusters or when the risk of classification does not decrease exponentially fast in n. These results are proved for i.i.d. and HMM observations. See Corollary 1 and Theorem 3.3.5 for the i.i.d. model, Theorem 3.3.7 and Theorem 3.3.9 for the HMM model. Thanks to this relationship between the risks studied, we can

perform a precise analysis of the Bayes risk of clustering based on the simpler Bayes risk of classification. Understanding the dependence of the Bayes risks with respect to the model parameters (mainly the population densities) is important because this will clearly identify the appropriate notion of separation measuring the difficulty of each problem. From a practical point of view, guaranteeing that the Bayes risk is of a certain magnitude will therefore translate into a simple condition on the separation between densities, which is easily interpretable. In Theorem 3.3.10 and Corollary 4, we identify the key quantity driving the difficulty of clustering and classification which turns out to be

$$\Lambda := \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left(\sum_{x \neq x_0} f_x(y) \right) d\mathcal{L}(y)$$

in both HMMs and i.i.d. models. Here, $(f_x)_{x\in\mathbb{X}}$ are the population densities of the distributions $(F_x)_{x\in\mathbb{X}}$ with respect to a dominating measure $\mathcal{L}$ over the observation space $\mathbb{Y}$, that is $dF_x = f_x d\mathcal{L}$. Notice that when $J = 2$, $\Lambda = 1 - \|F_1 - F_2\|_{\mathrm{TV}}$, where $\|F_1 - F_2\|_{\mathrm{TV}}$ is the total variation distance between the two distributions, unsurprisingly showing that the difficulty of the clustering tasks is governed by the difficulty of the hypothesis testing between F_0 and F_1. An informal answer to Question 2 can be summarized in the following theorem:

Theorem 3.1.2 (Informal)**.** *In both the i.i.d. and the HMM settings, when $J = 2$ or when $J > 2$ and $\Lambda \gtrsim e^{-cn}$ for a positive constant c:*

$$\inf_g \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \approx \inf_h \mathcal{R}_n^{\mathrm{class}}(\theta, h) \asymp \Lambda$$

where $\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g)$ and $\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h)$ are respectively the Bayes risks of clustering and classification.

Finally, we turn our attention to Question 3. As it will be shown in Theorem 3.3.6 and Theorem 3.3.8, the Bayes clusterer can not always be derived from the Bayes classifier. Nevertheless, the plug-in Bayes classifier remains the most commonly used method for clustering under Hidden Markov Models (HMMs); see Khiatani and Ghose [2017], Schliep et al. [2003], Ghassempour et al. [2014]. Despite its popularity, this approach lacks theoretical justification. Such a substitution is not justified unless the Bayes clusterer can be derived from the Bayes classifier or if the clustering risk of the Bayes classifier is comparable to the Bayes risk of clustering. This highlights the need for theoretical guarantees on the performance of this procedure. We show in Corollaries 2 and 3 that the clustering risk incurred by the Bayes classifier closely approximates the Bayes risk of clustering, showing thus the near optimality of the Bayes classifier for clustering. Furthermore, in Theorem 3.3.11, we establish that the clustering excess risk of a plug-in Bayes classifier decreases at the nonparametric estimation rate when using an appropriate estimator of the model parameters. This result provides a theoretical foundation for the practical use of the plug-in Bayes classifier. The main takeaway for practitioners working with HMMs is that, although a conceptual distinction exists between the Bayes classifier and the Bayes clusterer, in practice, the Bayes classifier is a reliable choice for clustering, with its excess risk provably controlled. We support this claim with numerical experiments in the nonparametric setting, demonstrating that HMM-based clustering can successfully recover latent structures in scenarios where standard methods fail. These results can be summarized in the following informal theorem.

Theorem 3.1.3 (Informal)**.** *Let g_θ be the clustering procedure built using the Bayes classifier. The Bayes classifier is nearly optimal:*

$$\mathcal{R}_n^{\mathrm{clust}}(\theta, g_\theta) \approx \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g).$$

Furthermore, under the hidden Markov Model, when the emission densities are s-Hölder regular, and under some additional regularity conditions, there exists an estimator $\hat{\theta}$ of θ such that:

$$\mathcal{R}_n^{\text{clust}}(\theta, g_{\hat{\theta}}) - \inf_g \mathcal{R}_n^{\text{clust}}(\theta, g) = \mathcal{O}\left(\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right).$$

Note that unlike specific cases such as parametric or translation models where many clustering algorithms can be proposed, the general nonparametric setting lacks a clear alternative to the plug-in procedure. This is why the excess risk reflects the nonparametric estimation rate.

The assumptions of the model and the definitions of the various risks we shall use in this work are described in Section 3.2. We state our main results in Section 3.3, the proofs of which are detailed in Section 3.6. Section 3.4 is devoted to simulation experiments, and Section 3.5 to possible further work.

3.2 Setting and definitions

3.2.1 Notations

We note for $i \leq j$ $X_{i:j} := (X_i, \dots, X_j)$ and $Y_{i:j} := (Y_i, \dots, Y_j)$. We denote $[n] \coloneqq \{1, \dots, n\}$ and $\mathcal{P}[n]$ is the set of partitions of $[n]$. The set of permutations of $[J]$ will be denoted by $\mathcal{S}_J$, and for any $\tau \in \mathcal{S}_J$, we consider θ^τ, the model parameter when labels are permuted with τ, that is $\theta^\tau = \left(\nu^\tau, Q^\tau, (f_{\tau(x)})_{x\in\mathbb{X}}\right)$ where $\nu^\tau = (\nu_{\tau(x)})_{x\in\mathbb{X}}$ and $Q^\tau = (Q_{\tau(x),\tau(x')})_{x,x'\in\mathbb{X}}$. For h a measurable function, $\|h\|_\infty$ denotes the essential supremum of h, possibly infinite. Frobenius norm is denoted by $\|.\|_{\text{F}}$ and $\|.\|$ stands for the operator norm. Given two sequences of positive numbers $(a_n)_n$ and $(b_n)_n$, $a_n \gtrsim b_n$ means that there exists an absolute constant $c > 0$ and $n_0 \in \mathbb{N}$ such that $(\forall n \geq n_0) \quad a_n \geq cb_n$. In the HMM modeling, for any parameter θ and any $\tau \in \mathcal{S}_J$, the distribution of the observations under $\mathbb{P}_\theta$ is the same as under $\mathbb{P}_{\theta^\tau}$. In other words, this distribution is invariant up to permutation of the labels given by the hidden states. This is known as the *label-switching* issue.

3.2.2 The model

We consider a Hidden Markov Model with J hidden states taking value in a set of labels $\mathbb{X} = \{1, .., J\}$, and observations in a Polish space endowed with its Borel σ-field $(\mathbb{Y}, \mathcal{Y})$. We denote $\mathbf{X} = (X_1, X_2, \dots)$ and $\mathbf{Y} = (Y_1, Y_2, \dots)$ respectively the sequence of hidden states forming the Markov chain and the observations. We assume that the emission distributions have densities f_x, $x = 1, \dots J$, with respect to a dominating measure $\mathcal{L}$ on $(\mathbb{Y}, \mathcal{Y})$. The HMM assumption boils down to:

$$\mathbf{X} \sim \text{Markov}(\nu, Q)$$
$$\mathbb{P}_\theta(Y_i \in \cdot \mid \mathbf{X}) = f_{X_i}(\cdot) d\mathcal{L}$$

where ν is the initial distribution of the chain and Q the transition matrix of the hidden chain. Throughout the paper, it is assumed that only the beginning $Y_{1:n}$ of the sequence $\mathbf{Y}$ is observed, and nothing else. We set $\theta = \left(\nu, Q, (f_x)_{x\in\mathbb{X}}\right)$ the parameter of the model and Θ denotes the space of all valid parameters. The following assumption is made:

- *Independent case.* In this case we assume that all the lines of Q are identical and equal to the vector ν of weights forming its stationary distribution. This corresponds to the usual mixture model with independent latent variables. The set of these parameters will be denoted Θ^{ind}.

- *Dependent case.* In this case we assume the lines of the transition matrix Q not all equal, so that the $(X_i)_{i\geq 1}$ are not independent. The set of these parameters will be denoted Θ^{dep}.

Throughout this work, we shall consider the transition matrix being fixed, and we will be interested in how the separation of populations, understood as some quantity depending on the populations densities, drives the difficulty of the clustering task. We show that, under the HMM modelling, it is possible to cluster general populations without assuming they belong to some prescribed parametric family. While our primary interest is in the dependent case in which the emission densities can be identified without any further constraint, we obtain, however, results that have interest also in the widely used independent case, in particular the analysis of the Bayes risk and its minimizers. (see Section 3.3).

One main difference between the HMM and the i.i.d. situation in the analysis of the Bayes risk of classification is that in the HMM modeling, the probability of a label X_i given the observations $Y_1, \dots, Y_n$ depends on all the observations, whereas in the i.i.d. case it depends only on the associated observation Y_i. The vector of probabilities of the X_i's given $Y_1, \dots, Y_n$ are called the smoothing distributions, they depend on i, n and the observations. The vectors of probabilities of the X_i's given the observations up to time i, $Y_1, \dots, Y_i$, are called the filtering distributions, they depend on i, n and the involved observations. Recursive formulas verified by the filtering and smoothing distributions make the computations tractable under the HMM modelling. See Section 3 of Cappé et al. [2005] for details.

3.2.3 The problem of clustering

For any $n \geq 1$, the finite sequence $X_{1:n} = (X_1, \dots, X_n)$ induces a random partition $\Pi_n = \{C_1, C_2, \dots\}$ of $[n]$ whose blocks – the so-called *clusters* – are the equivalence classes for the random equivalence relation $i \sim j \iff X_i = X_j$. The goal of clustering is to uncover this partition Π_n on the sole basis of the observation $Y_{1:n} = (Y_1, \dots, Y_n)$. We define a *clusterer*:

Definition 3.2.1 (Clusterer)**.** *A n-clusterer is a measurable map $g : \mathbb{Y}^n \to \mathcal{P}[n]$. We denote by $\mathcal{G}_n$ the set of all n-clusterers.*

We measure the loss incurred by guessing $g(Y_{1:n})$ in place of Π_n via the misclassification error distance Meilă and Heckerman [2001], Meilă [2005]. For two partitions A and B of $[n]$, this loss is defined by:

$$\ell(A, B) = 1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(A,B) \\ M \text{ is a matching}}} \sum_{\{C,C'\} \in M} \text{Card}(C \cap C') \tag{3.2}$$

where the supremum is taken over the set of matchings. To define a matching, we build the complete bipartite graph $(A, B, \mathcal{E}(A, B))$ on vertices A and B with edge set $\mathcal{E}(A, B) := \{\{C, C'\} \ : \ C \in A,\, C' \in B\}$. Then we recall that a matching M is a set $M \subseteq \mathcal{E}(A, B)$ of edges without common vertices (*i.e.* each block of A and B appears in at most one edge of the matching). An example of a matching is depicted in Figure 3.1. Then the risk of a clusterer g can be defined as the expected loss of the partition $g(Y_{1:n})$ with respect to the true partition Π_n which is summarized in the risk function $\mathcal{R}_n^{\text{clust}} : \Theta \times \mathcal{G}_n \to [0, 1]$

$$\mathcal{R}_n^{\text{clust}}(\theta, g) := \mathbb{E}_\theta \Big[\ell\left(g(Y_{1:n}), \Pi_n\right) \Big]. \tag{3.3}$$

A closely related notion is that of a *classifier*:

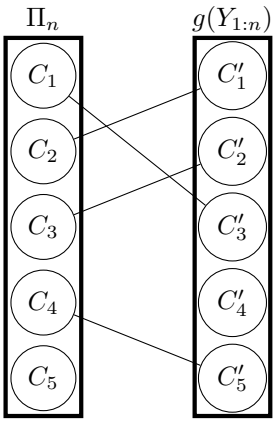


Figure 3.1: Example of a matching. Nodes on the left represent the clusters induced by the partition of Π_n; those on the right are the clusters of $g(Y_{1:n})$. Edges form a matching between the two partitions.

Definition 3.2.2 (Classifier)**.** *A n-classifier is a measurable map $h : \mathbb{Y}^n \to \mathbb{X}^n$. We denote by $\mathcal{H}_n$ the set of all n-classifiers.*

One may find our Definition 3.2.2 different from standard textbook definitions Devroye et al. [1996]: we explain the reason of this choice later in Remark 3.2.3. A classifier differs from a clusterer in that it not only seeks for the hidden partition, but also for the labels of the observations. Hence, the usage of a classifier only makes sense in a supervised framework where access to some labeled data is allowed in some way. In an unsupervised framework such as our model, $\mathbf{Y}$ does not contain any information about the labels and recovering them better than a lucky guess is impossible. It is however true that, to any n-classifier $h \in \mathcal{H}_n$ corresponds a unique n-clusterer $g \in \mathcal{G}_n$ which can be built via the map $\pi_n : \mathbb{X}^n \to \mathcal{P}[n]$ such that

$$g(Y_{1:n}) = \pi_n \circ h(Y_{1:n}) = \{\{i \ : \ h_i(Y_{1:n}) = x\} \ : \ x \in \mathbb{X}\}\backslash\{\varnothing\}$$

and any clusterer can be represented that way by choosing a specific labelling of the clusters. For this reason, the notions of clusterer and classifier are very much often amalgamated in the literature. We argue that it would be better to define them separately in order to avoid confusions between the risk of clustering $\mathcal{R}_n^{\text{clust}}(\theta, \pi_n \circ h)$ and the risk of classification $\mathcal{R}_n^{\text{class}} : \Theta \times \mathcal{H}_n \to [0, 1]$ (relative to the loss counting number of misclassified observations)

$$\mathcal{R}_n^{\text{class}}(\theta, h) \coloneqq \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{h_i(Y_{1:n}) \neq X_i} \right]. \tag{3.4}$$

Here again we insist that the definition of $\mathcal{R}_n^{\text{class}}$ in Equation (3.4) – although mathematically correct – has no statistical interest in an unsupervised model where the classifier h is not allowed to see some of the labels. An easy exercise (see Lemma 3.6.15) shows that the risk of clustering of $\pi_n \circ h$ can be rewritten as

$$\mathcal{R}_n^{\text{clust}}(\theta, \pi_n \circ h) = \mathbb{E}_\theta \left[\min_{\tau \in \mathcal{S}_J} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{h_i(Y_{1:n}) \neq \tau(X_i)} \right] \tag{3.5}$$

and differs from (3.4). Note that it does not depend on the classifier h chosen to represent the clusterer thanks to the infimum over the permutations of the labels. Note also that the infimum inside the expectation reflects the ease of the clustering problem compared to

the classification problem, because unlike classification, clustering does not seek to identify the true labels themselves.

It is customary to compare the performance of a given classifier h to the best performance attainable by an oracle classifier, namely the *Bayes risk of classification*:

$$\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h).$$

In particular, it is well-known that the previous optimization problem is solved by the (albeit non necessarily unique) so-called Bayes classifier $h_\theta^\star = (h_{\theta,1}^\star, \dots, h_{\theta,n}^\star)$ such that

$$\mathbb{P}_\theta(X_i = h_{\theta,i}^\star(Y_{1:n}) \mid Y_{1:n}) = \max_{x \in \mathbb{X}} \mathbb{P}_\theta(X_i = x \mid Y_{1:n}), \qquad i = 1, \dots, n.$$

In an unsupervised learning context, it makes sense to compare the risk of a clusterer to the *Bayes risk of clustering*:

$$\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g).$$

As for the classification risk, the solution to the previous optimization problem exists and is obtained by the Bayes clusterer $g_\theta^\star$ such that $g_\theta^\star(Y_{1:n})$ is the partition that minimizes

$$g \mapsto \mathbb{E}_\theta \left[1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(g(Y_{1:n}), \Pi_n) \\ M \text{ is a matching}}} \sum_{\{C, C'\} \in M} \text{Card}(C \cap C') \middle| Y_{1:n} \right]. \tag{3.6}$$

In contrast with classification, however, the Bayes clusterer has usually no simple expression. It is to be noted that there is no reason that $g_\theta^\star = \pi_n \circ h_\theta^\star$. An analysis will be conducted to determine when the equality $g_\theta^\star = \pi_n \circ h_\theta^\star$ holds. Although the inequality $\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$ is true and easily proved from (3.5), it is not guaranteed that $\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g)$ and $\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$ are equivalent (*ie.* have comparable magnitude). In Sections 3.3.1 and 3.3.2 we show that the equivalence holds both in the independent and in the dependent scenarios as long as the Bayes risk of classification is not exponentially small in n or when there are only two clusters. Though we prove the two risks to be equivalent in some specific contexts, we provide a counter-example showing that this is not always true.

Remark 3.2.3. It is common Devroye et al. [1996] to define a classifier as a function $h : \mathbb{Y} \to \mathbb{X}$ (as opposed to $\mathbb{Y}^n \to \mathbb{X}^n$ in our Definition 3.2.2). This usual definition is motivated by the fact that in the independent scenario the law of $X_i \mid Y_{1:n}$ is that of $X_1 \mid Y_1$. Hence if one is willing to classify the vector $Y_{1:n}$, the Bayes classifier rewrites as $h_\theta^\star = (h_{\theta,1}^\star(Y_1), \dots, h_{\theta,1}^\star(Y_n))$ with $h_{\theta,1}(y)$ maximizing $x \mapsto \mathbb{P}_\theta(X_1 = x \mid Y_1 = y)$ and the Bayes risk of classification equals $\mathbb{P}_\theta(h_{\theta,1}^\star(Y_1) \neq X_1)$. Thus classifying Y_1 or $Y_{1:n}$ is not very different. In the dependent case, however, the situation differs since $\mathbf{X} \mid \mathbf{Y}$ is a inhomogeneous Markov chain. This implies in particular that classifying Y_1 or $Y_{1:n}$ are different problems when $n \geq 2$, and the optimal solution for classifying $Y_{1:n}$ can not be obtained from the optimal solution of classifying Y_1.

Remark 3.2.4. In Marandon et al. [2023], the authors define the risk of clustering of a n-classifier h in a different way:

$$\mathcal{R}_n^{\text{MRSS}}(\theta, h) := \mathbb{E}_\theta \left[\inf_{\tau \in \mathcal{S}_J} \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\tau(X_i) \neq h_i(Y_{1:n})} \middle| Y_{1:n} \right] \right]$$

Similarly, in Lu and H. Zhou [2016], the risk of clustering is defined by:

$$\tilde{\mathcal{R}}_n^{\text{clust}}(\theta, h) := \inf_{\tau \in \mathcal{S}_J} \mathbb{E}_\theta \left[\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\tau(X_i) \neq h_i(Y_{1:n})} \right]$$

Their definition is mathematically convenient as one can easily show that (see Lemma 3.6.2):

$$\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{MRSS}}(\theta,h) = \inf_{h\in\mathcal{H}_n} \tilde{\mathcal{R}}_n^{\mathrm{clust}}(\theta,h) = \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h)$$

Hence, the Bayes clusterer relative to their risks is $g_\theta^\star = \pi_n \circ h_\theta^\star$ with $h_\theta^\star$ the Bayes classifier relative to the risk $\mathcal{R}_n^{\mathrm{class}}(\theta,.)$. However, contrary to our definition, it seems that there is no suitable statistical interpretation of these risks of clustering.

3.3 Main results

The Bayes risk of classification offers an advantage due to its closed formula, stemming from the explicit identification of the Bayes classifier. In contrast, the Bayes risk of clustering lacks this straightforward formulation. However, thanks to Equation (3.5), the risk of a clustering procedure can be seen as the risk of classification of the associated classifier up to the best random permutation (the one minimizing the sum inside the expectation in Equation (3.5)). This is why a common idea is that risk of clustering and risk of classification are closely related. Following this intuition, authors of Zhang and Zhou [2016], Gao et al. [2018] propose an inequality linking the minimax (over a specific class) risk of clustering to the minimax risk of classification, in the context of community detection under the stochastic block model. Their inequality is applied in Ndaoud [2022] in the context of the mixture of two Gaussian distributions and in Lu and H. Zhou [2016] for general subGaussian mixtures. Although their argument is neat, it relies heavily on two key ingredients: (i) in their model the latent partition is deterministic and viewed as a parameter, and (ii) their point of view is minimax and thus, it is enough to consider some well-chosen parameters in order to lower bound the minimax risk. Thus their argument is not transposable to the situation where one is interested in bounding the Bayes risk of clustering at every possible value of the parameter with random labels.

3.3.1 I.I.D. case

The following theorem shows that in the i.i.d. setting and when there are only two classes ($J = 2$), the clustering resulting from the Bayes classifier coincides with that of the Bayes clusterer, almost-everywhere.

Theorem 3.3.1. *In the case of independent labels, if $J = 2$, then for all $\theta \in \Theta^{\mathrm{ind}}$ and all $n \geq 2$*

$$g_\theta^\star(Y_{1:n}) = \pi_n \circ h_\theta^\star(Y_{1:n}) \quad \mathbb{P}_\theta\text{-a.s.}$$

The proof of Theorem 3.3.1 can be found in Section 3.6.1. Thanks to this result, the difference between the two risks of clustering and classification can be bounded almost tightly, as shown in the next theorem. Let $\varepsilon_{n,\theta} = \frac{1}{2} - \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h)$.

Theorem 3.3.2. *When $J = 2$ and $\theta \in \Theta^{\mathrm{ind}}$, $\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h) = 0$ if and only if $\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta,g) = 0$. If $\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h) \neq 0$ then the difference between the Bayes risks satisfies*

$$\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h) - \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta,g) \leq \min\left(\frac{(1-4\varepsilon_{n,\theta}^2)^{\frac{n}{2}}}{\frac{n}{2}\log\left(\frac{1+2\varepsilon_{n,\theta}}{1-2\varepsilon_{n,\theta}}\right)}, \sqrt{\frac{\pi}{2n}} \right).$$

Furthermore, there exists a universal constant $B > 0$ such that for all $n \geq 100$ and all $\theta \in \Theta^{\mathrm{ind}}$

$$\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,h) - \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta,g) \geq B \min\left(\frac{(1-4\varepsilon_{n,\theta}^2)^{\frac{n}{2}\left(1+\frac{6.8}{1\vee\sqrt{n}\varepsilon_{n,\theta}}\right)}}{\frac{n}{2}\log\left(\frac{1+2\varepsilon_{n,\theta}}{1-2\varepsilon_{n,\theta}}\right)}, \frac{1}{\sqrt{n}} \right).$$

The proof of Theorem 3.3.2 is given in Section 3.6.2. As emphasized by the lower bound, the upper bound on the difference between the two Bayes risks in the previous theorem is essentially tight. The lower bound also shows that the Bayes risk of clustering is always strictly smaller than the Bayes risk of classification, unless they are both zero (which happens when the two emission densities have disjoint support). Also, the difference between the risks decays super-polynomially in n as soon as $\varepsilon_{n,\theta} \gg \sqrt{\log(n)/n}$, and polynomially if $\varepsilon_{n,\theta} = O(\sqrt{\log(n)/n})$ with worst-case rate $\asymp n^{-1/2}$ when $\varepsilon_{n,\theta} = O(n^{-1/2})$.

A direct consequence of this result is that the Bayes risk of classification is equivalent to the Bayes risk of clustering for n large enough as shown in the following corollary.

Corollary 1. *In the case of independent labels with $J = 2$, for all $\theta \in \Theta^{\mathrm{ind}}$ and all $n \geq 2$*

$$(1 - \alpha_n) \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h) \leq \inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h)$$

where $\alpha_n = 2 \min \left(2(1 + \varepsilon_{n,\theta}) \frac{(1 - 4\varepsilon_{n,\theta}^2)^{\frac{n-2}{2}}}{n \log\left(\frac{1+2\varepsilon_{n,\theta}}{1-2\varepsilon_{n,\theta}}\right)}, \frac{1}{1-2\varepsilon_{n,\theta}} \sqrt{\frac{\pi}{2n}} \right)$.

The following proposition shows that, contrary to Corollary 1, the two Bayes risks are not equivalent in general.

Proposition 3.3.3. *Whenever $J > 2$ and $n \geq 1$:*

$$\inf_{\theta \in \Theta^{\mathrm{ind}}} \frac{\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g)}{\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h)} = 0.$$

The proof of Proposition 3.3.3 is given in Section 3.6.9. It sheds light on why the previously established equivalence when there were only two classes no longer holds when the number of classes exceeds two. The proof is based on the following intuition. The observations which are misclustered or misclassified appear only in regions of overlap between the emission densities (cf. Theorem 3.3.10 and Corollary 4). Consider now the situation where n observations are derived from a mixture of $J = 3$ densities F_1, F_2 and F_3 with weights $\nu_1 = 1 - 2\eta$, $\nu_2 = \eta$ and $\nu_3 = \eta$. Assume that the support of F_1 is disjoint from that of F_2 and F_3 and that F_2 and F_3 have overlapping supports over a small region of the space. When the weight η is small, two situations are possible: eventually only one observation belongs to the support of F_2 or F_3, in which case the clustering error is null, or more observations appear in this region of the space which happens with very small probability because η is small. Combining these two insights, the magnitude of the risk of clustering is shown to be negligible with respect to the risk of classification. The use of compactly supported distributions is not essential to the previous argument. For instance, the argument still holds modulo supplementary technicalities if one takes $F_1 = \mathcal{N}(-\alpha x, 1)$, $F_2 = \mathcal{N}(-\alpha, 1)$ and $F_3 = \mathcal{N}(\alpha, 1)$ with $\alpha > 0$ and $x > 0$ large enough.

This dichotomy between $J = 2$ and $J > 2$ concerns also the minimizers of the risks. Contrary to Theorem 3.3.1 (when $J = 2$), in the i.i.d. setting with $J > 2$, one can always find some model parameters for which the result of clustering using the Bayes classifier differs from that of the Bayes clusterer with positive probability, as shown in the next theorem.

Theorem 3.3.4. *If $J > 2$, then for all $\theta \in \Theta^{\mathrm{ind}}$, if*

$$\mathbb{P}_\theta \left(\bigcup_{j=1}^{J} \left\{ 0 < \max_{l \neq j} \nu_l f_l(Y) < \nu_j f_j(Y) \leq \sum_{l \neq j} \nu_l f_l(Y) \right\} \right) > 0$$

then,

$$\forall n \geq 2, \quad \mathbb{P}_\theta \left(g_\theta^\star(Y_{1:n}) \neq \pi_n \circ h_\theta^\star(Y_{1:n}) \right) > 0.$$

The condition above can be ensured easily for many distributions $(f_j)_{j\in\mathbb{X}}$ such as multinomials, mixtures of Gaussians, etc. For example, it is valid when $J = 3$, $(\nu_1, \nu_2, \nu_3) = (0.4, 0.4, 0.2)$, and the emission densities are Gaussians with variance $\sigma^2 = 1$ and means $(\mu_1, \mu_2, \mu_3) = (1, 2, 3)$. The proof of Theorem 3.3.4 is given in Section 3.6.3.

Even though the Bayes risks of clustering and classification are not equivalent uniformly over Θ^{ind} by Proposition 3.3.3, we show in the next theorem that when $J > 2$, they become equivalent when $\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \gtrsim J^2 e^{-n\beta/8}$ where $\beta = \min_{j\neq k}(\nu_j + \nu_k)$.

Theorem 3.3.5. *For all $\theta \in \Theta^{\text{ind}}$ and all $n \geq 1$ the following bounds hold*

$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \geq \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) - \sqrt{\frac{\log(J!)}{2n}},$$
$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \geq (1 - \xi_n) \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) - J^2 e^{-n\beta/8},$$

where $\beta = \min_{j\neq k}(\nu_j + \nu_k)$ and $\xi_n = \frac{4e}{\beta}[\sqrt{\log(J!)/(2n)}]^{1-4/(n\beta)}$.

The proof of Theorem 3.3.5 is given in Section 3.6.5. Even if the second lower bound obtained does not establish an equivalence between the Bayes risks of clustering and classification, it remains useful for analyzing phase transitions. Specifically, under a mild assumption on β, this bound results in similar phase transitions for both risks.

Even if clustering using the Bayes classifier differs sometimes from that of the Bayes clusterer (as shown in Theorem 3.3.4), Theorem 3.3.5 provides guarantees for the risk of clustering using the Bayes classifier as shown by the following corollary.

Corollary 2. *For all $\theta \in \Theta^{\text{ind}}$ and all $n \geq 1$ the following bounds hold*

$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \mathcal{R}_n^{\text{clust}}(\theta, \pi_n \circ h_\theta^\star) \leq \frac{1}{1 - \xi_n}\left(\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) + J^2 e^{-n\beta/8}\right)$$

with ξ_n and β as defined in Theorem 3.3.5. When there are only two classes

$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \mathcal{R}_n^{\text{clust}}(\theta, \pi_n \circ h_\theta^\star) \leq \frac{1}{1 - \alpha_n} \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g)$$

where α_n is defined in Corollary 1.

3.3.2 HMM case

The behavior of the Bayes risks under the HMM modeling exhibits similarities and differences with the i.i.d. case. As we will see, the Bayes risks keep having the same behavior under the HMM setting, while their minimizers, on the other hand, can be different. First, contrary to the i.i.d. case where the result of clustering using the Bayes classifier matches that of the Bayes clusterer, it is always possible to find a set of parameters for which they differ under the HMM setting, as shown in the next theorem.

Theorem 3.3.6. *In the case of dependent labels and when $J = 2$, there exists a subset $\tilde{\Theta} \subset \Theta^{\text{dep}}$ such that*

$$\left(\forall\theta \in \tilde{\Theta}\right)(\forall n \geq 2), \quad \mathbb{P}_\theta\left(g_\theta^\star(Y_{1:n}) \neq \pi_n \circ h_\theta^\star(Y_{1:n})\right) > 0.$$

We illustrate this in the simple situation of $n = 2$. Assume one observes two consecutive observations Y_1 and Y_2 of a HMM with transition matrix

$$Q = \begin{pmatrix} 1 - p & p \\ q & 1 - q \end{pmatrix}$$

and emission densities f_1 and f_2. Denoting for simplicity $\bar{p} = 1 - p$ and $\bar{q} = 1 - q$, one can easily check that the Bayes clusterer puts the two observations in the same cluster when

$$q\bar{p}f_1(Y_1)f_1(Y_2) + p\bar{q}f_2(Y_1)f_2(Y_2) \geq pq\left(f_2(Y_1)f_1(Y_2) + f_1(Y_1)f_2(Y_2)\right) \tag{3.7}$$

On the other hand, the Bayes classifier puts the two observations in the same class when

$$\left(\mathbb{P}_\theta\left(X_1 = 2 \mid Y_{1:2}\right) - \mathbb{P}_\theta\left(X_1 = 1 \mid Y_{1:2}\right)\right)\left(\mathbb{P}_\theta\left(X_2 = 2 \mid Y_{1:2}\right) - \mathbb{P}_\theta\left(X_2 = 1 \mid Y_{1:2}\right)\right) \geq 0,$$

or equivalently,

$$\begin{aligned}&\left(pqf_2(Y_1)f_1(Y_2) + p\bar{q}f_2(Y_1)f_2(Y_2) - q\bar{p}f_1(Y_1)f_1(Y_2) - qpf_1(Y_1)f_2(Y_2)\right)\\ &\quad\times\left(qpf_1(Y_1)f_2(Y_2) + p\bar{q}f_2(Y_1)f_2(Y_2) - q\bar{p}f_1(Y_1)f_1(Y_2) - pqf_2(Y_1)f_1(Y_2)\right) \geq 0.\end{aligned} \tag{3.8}$$

The two conditions (3.7) and (3.8) are not equivalent. In the simple situation of Bernoulli emissions $(f_1, f_2) = (\mathcal{B}(\alpha_1), \mathcal{B}(\alpha_2))$, and $Y_{1:2} = (1, 1)$, the Bayes clustering puts the two observations in the same cluster when

$$q\bar{p}\alpha_1^2 + p\bar{q}\alpha_2^2 > 2pq\alpha_1\alpha_2$$

while the Bayes classifier puts them always in the same cluster since the corresponding condition becomes

$$\left(p\bar{q}\alpha_2^2 - q\bar{p}\alpha_1^2\right)^2 \geq 0.$$

Note that the first condition is not always ensured when p and q are chosen near 1. Note also that in the situation where $p + q = 1$, the dependence structure is lost and the two conditions become equivalent to

$$(qf_1(Y_1) - pf_2(Y_1))(qf_1(Y_2) - pf_2(Y_2)) \geq 0$$

which is coherent with Theorem 3.3.1 in the i.i.d. case. This highlights the strong difference between the dependent and independent setting. The proof of Theorem 3.3.6 can be found in Section 3.6.7.

We now establish the equivalence between the two risks under the Markovian dependence of the labels. We consider the following assumption.

Assumption 1. $\delta = \min_{x,x'} Q_{x,x'} > 0$ and $\min_x \nu_x \geq \delta$.

The positive lower bound δ introduced in Assumption 1 makes the hidden Markov chain irreducible. It will be used in proving deviation inequalities or when forgetting properties of the chain are needed. Even if the minimizers of the risk of classification and the risk of clustering might differ, we are still able to prove the equivalence between the two risks. The following theorem shows this is the case for n large enough.

Theorem 3.3.7. *If $J = 2$, then for all $\theta \in \Theta^{\text{dep}}$ such that Assumption 1 holds and all $n \geq 1$*

$$(1 - \tilde{\alpha}_n) \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \leq \inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h),$$

where $\tilde{\alpha}_n = 2e\left(\frac{1-\delta}{\delta}\right)^4 \left[\frac{1-\delta}{\delta}\sqrt{\log(2)/2n}\right]^{1-2/n}$.

Thanks to Theorem 3.3.7, it suffices to study the Bayes risk of classification, since any bound on this risk can be extrapolated to the risk of clustering, regardless of the magnitude of the Bayes risk of clustering. The proof of Theorem 3.3.7 is given in Section 3.6.6.

Let us now consider the case $J > 2$. As in the i.i.d. setting where the minimizers of the risks do not always coincide, it is always possible to find a set of parameters for which the HMM observations are dependent, but the results of clustering using the Bayes clusterer and the partition induced by the Bayes classifier are not the same, as shown in the next theorem.

Theorem 3.3.8. *In the case of dependent labels, for all $J > 2$ and all $n \geq 2$, there exists a subset $\tilde{\Theta}_{n,J} \subset \Theta^{\text{dep}}$ such that*

$$\forall \theta \in \tilde{\Theta}_{n,J}, \quad \mathbb{P}_\theta \left(g^\star_\theta(Y_{1:n}) \neq \pi_n \circ h^\star_\theta(Y_{1:n}) \right) > 0.$$

The proof of Theorem 3.3.8 can be found in Section 3.6.8.

Finally, we establish a result similar to Theorem 3.3.5 under the HMM setting.

Theorem 3.3.9. *For all $\theta \in \Theta^{\text{dep}}$ such that Assumption 1 holds and all $n \geq 1$, the following bounds are true*

$$\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \geq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) - \frac{1}{1-\rho_0} \sqrt{\frac{\log(J!)}{2n}},$$
$$\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \geq (1 - \tilde{\xi}_n) \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) - (J^2+1) e^{-2n(1-\rho_0)^2 \beta^2 / 25},$$

where $\beta = \min_{i,j \neq k} \mathbb{P}_\theta \left(X_i \in \{j, k\} \right)$, $\rho_0 = \frac{1 - J\delta}{1-(J-1)\delta}$, and $\tilde{\xi}_n = \frac{5}{\beta(1-\rho_0)} \sqrt{\log(J!)/(2n)}$.

The proof of Theorem 3.3.9 is given in Section 3.6.6. Note that when the HMM is stationary, $\beta = \min_{j \neq k}(\nu_j + \nu_k)$ as in the i.i.d. case. Notice also that when $\delta = 1/J$, the observations are i.i.d. with uniform distribution over the set $\mathbb{X}$, and we recover the first inequality for i.i.d. observations. However, we do not recover the inequality for i.i.d. observations in general from that of HMM observations.

As for the i.i.d setting, even if clustering using the Bayes classifier differs sometimes from that of the Bayes clusterer (as shown in Theorem 3.3.6 and Theorem 3.3.8), Theorem 3.3.9 provides guarantees for the risk of clustering using the Bayes classifier as shown by the following corollary.

Corollary 3. *For all $\theta \in \Theta^{\text{dep}}$ and all $n \geq 1$ the following bounds hold*

$$\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \mathcal{R}_n^{\text{clust}}\left(\theta, \pi_n \circ h^\star_\theta\right) \leq \frac{1}{1 - \tilde{\xi}_n} \left(\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) + (J^2+1) e^{-2n(1-\rho_0)^2 \beta^2 / 25} \right)$$

where $\tilde{\xi}_n$, β and ρ_0 are as in Theorem 3.3.9. When there are only two classes

$$\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \mathcal{R}_n^{\text{clust}}\left(\theta, \pi_n \circ h^\star_\theta\right) \leq \frac{1}{1 - \tilde{\alpha}_n} \inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g)$$

where $\tilde{\alpha}_n$ is as in Theorem 3.3.7.

3.3.3 A key quantity for the Bayes risk of clustering for both I.I.D. and HMM

We now state our main result which proves upper and lower bounds on the Bayes risks in function of a quantity measuring the separation between the emission densities up to constants depending on the transition matrix. These bounds translate into bounds on the risk of clustering thanks to the result above. Let $\Lambda := \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left[\sum_{x \neq x_0} f_x(y) \right] d\mathcal{L}(y)$.

Theorem 3.3.10. *Under Assumption 1, for all $n \geq 1$, the Bayes risk of classification satisfies:*

$$\forall \theta \in \Theta^{\text{ind}}, \quad \delta\Lambda \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \leq (1 - (J-1)\delta)\Lambda$$
$$\forall \theta \in \Theta^{\text{dep}}, \quad \frac{\delta^2}{1-(J-1)\delta}\Lambda \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \leq (1 - (J-1)\delta)\Lambda$$

The proof of Theorem 3.3.10 is given in Section 3.6.10. Note that the bounds vanish when all the emission densities have disjoint supports. Note also that upper and lower bounds match when $\delta = \frac{1}{J}$ and the risk corresponds to Λ/J in this case. This situation corresponds to i.i.d. observations derived from a mixture with J components of equal weights. This proves in particular the tightless of the bounds which can not be improved by any absolute multiplicative constant without restricting the parameter space. Thanks to the results comparing the Bayes risks, Λ is the appropriate measure of the difficulty of clustering in many regimes as shown in the following corollary. Recall the definition of α_n from Corollary 1, of ξ_n from Theorem 3.3.5 of $\tilde{\alpha}_n$ from Theorem 3.3.7, of $\tilde{\xi}_n$ from Theorem 3.3.9, and of $\beta = \min_{i\in[n],j\neq k\in\mathbb{X}} \mathbb{P}_\theta\left(X_i \in \{j,k\}\right)$.

Corollary 4. *Under Assumption 1, the following holds.*

- *When $J = 2$:*

$$\forall \theta \in \Theta^{\mathrm{ind}}, \quad (1-\alpha_n)\delta\Lambda \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-\delta)\Lambda,$$
$$\forall \theta \in \Theta^{\mathrm{dep}}, \quad \frac{\delta^2(1-\tilde{\alpha}_n)}{1-\delta}\Lambda \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-\delta)\Lambda.$$

- *When $J > 2$ and $\theta = (\nu, Q, (f_x)_{x\in\mathbb{X}}) \in \Theta^{\mathrm{ind}}$ is such that $\delta\Lambda \geq 4J^2 e^{-n\beta/8}$ and n is sufficiently large to have $\xi_n \leq \frac{1}{2}$:*

$$\frac{\delta}{4}\Lambda \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-(J-1)\delta)\Lambda.$$

- *When $J > 2$ and $\theta = (\nu, Q, (f_x)_{x\in\mathbb{X}}) \in \Theta^{\mathrm{dep}}$ is such that $\delta^2\Lambda \geq 4(1-(J-1)\delta)(J^2+1)e^{-2n(1-\rho_0)^2\beta^2/15}$ and n is sufficiently large to have $\tilde{\xi}_n \leq \frac{1}{2}$:*

$$\frac{\delta^2}{4(1-\delta)}\Lambda \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-(J-1)\delta)\Lambda.$$

In other words, when there are only two classes, we obtain a tight characterization of the Bayes risk of clustering in terms of the separation between the two emission densities (i) covering all the regimes (ii) valid in the parametric and non-parametric setting (iii) without imposing any separation between the emission densities. In both i.i.d. and HMM settings, when there are only two classes, there exists a positive constant $\alpha(\delta)$ depending only on δ such that

$$\alpha(\delta)\int_{\mathbb{Y}} [f_1 \wedge f_2](y)d\mathcal{L}(y) \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-\delta)\int_{\mathbb{Y}} [f_1 \wedge f_2](y)d\mathcal{L}(y).$$

For example, in the case of Gaussian emission distributions with two hidden states, this translates in a clear identification of the signal-to-noise ratio which drives the Bayes risk of clustering. For two Gaussian emission densities with means μ_0 and μ_1 and the same covariance matrix Σ, the Bayes risk of clustering ensures:

$$\frac{\alpha(\delta)}{2}\exp\left(-\frac{\mathrm{SNR}}{4}\right) \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) \leq (1-\delta)\exp\left(-\frac{\mathrm{SNR}}{8}\right)$$

where $\mathrm{SNR} = (\mu_0-\mu_1)^\intercal\Sigma^{-1}(\mu_0-\mu_1)$.

3.3.4 Reaching the Bayes risk

I.I.D. setting

While there is no formal proof of this fact, it seems that there is no way of reaching the Bayes risk of clustering or classification without strong assumptions on the mixture components under the non-parametric i.i.d. mixture model. Without structural assumptions, the associated Bayes classifier can never be learnt nor approximated from the data in the i.i.d. case because the model is not identifiable and thus, the mixture components can not be estimated. The algorithms proposed in the literature perform well in clustering i.i.d. observations only when the clusters are assumed to be separated. This is the case for the k-means and its variants Giraud and Verzelen [2018], Mixon et al. [2017] and the spectral algorithms Shi et al. [2009], Kannan et al. [2005] to cite a few. We refer to Grün [2019] and Chapter 12 of Giraud [2021] for a review of model-based clustering techniques.

HMM setting

Unlike the i.i.d. hypothesis, the HMM hypothesis allows us to identify the model without any assumptions about the emission distributions, apart from the fact that they are distinct Alexandrovich et al. [2016b]. This allows the construction of simple clustering procedures with risk comparable to the Bayes risk of clustering. Given that the transition matrix of the hidden Markov chain is non-singular and ergodic, all parameters – including the number J of hidden states and the emission distributions – can be identified from the distribution of K consecutive observations $(Y_1, \dots, Y_K)$, provided that the emission distributions are distinct and K is sufficiently large compared to J Alexandrovich et al. [2016c]. This allows complete flexibility on the emission distributions, provided a Markovian dependence of labels. Within the framework of this model, non-parametric estimation of the mixture components becomes possible without any restrictions on the population densities, apart from the fact that they are distinct, and various estimation procedures have been proposed De Castro et al. [2016, 2017], Lehéricy [2018, 2021], Abraham et al. [2022, 2025]. Although we have shown in Theorems 3.3.6 and 3.3.8 that the Bayes clusterer $g^\star_\theta$ does not necessarily coincide with $\pi_n \circ h^\star_\theta$, this distinction matters only to establish strong decision-theoretic foundations for clustering. In practice, the Bayes classifier still exhibits a competitive performance in clustering. Corollary 3 shows that the risk of clustering of the Bayes classifier approximates closely the Bayes risk of clustering, proving thus its near-optimality. The widely used method of clustering when the observations are drawn from a HMM consists in approaching the behavior of the Bayes classifier $h^\star_\theta$ by plugging-in estimated parameters $\hat\theta$ and using the induced clusterer $\pi_n \circ h^\star_{\hat\theta}$. It is not worth the effort of computing $g^\star_\theta$ and use $g^\star_{\hat\theta}$ because the price to pay in the excess risk for trading the true θ by its estimate $\hat\theta$ is most likely of several order of magnitude larger than the price to pay in using $\pi_n \circ h^\star_\theta$ in place of $g^\star_\theta$. Thus in this section, we focus on the clustering rule

$$g^\star_{\hat\theta}(Y_{1:n}) \coloneqq \pi_n \circ h^\star_{\hat\theta}(Y_{1:n}) = \pi_n \left(\left(\arg\max_{x \in \mathbb{X}} \mathbb{P}_{\hat\theta} \left(X_i = x \mid Y_{1:n} \right) \right)_{1 \le i \le n} \right)$$

where $\hat\theta = \hat\theta(Y_{1:n})$ is an estimator constructed using the celebrated tensor method Anandkumar et al. [2014], Abraham et al. [2022]. The main advantage in using $\pi_n \circ h^\star_{\hat\theta}$ is that in contrast with $g^\star_{\hat\theta}$ the classifier $h^\star_{\hat\theta}$ is easily computed thanks to the recurrence formulas ensured by the Forward-Backward algorithm Cappé et al. [2005]. Theorem 3.3.11 below controls the excess risk of this clustering procedure. We will make the following assumption:

Assumption 2. The initial distribution ν is the stationary distribution of $\mathbf{X}$.

Notice however that under Assumption 2, the second part of Assumption 1 follows directly from the first part.

Assumption 3. Q is full-rank and aperiodic.

Under Assumptions 1, 2 and 3, the hidden Markov chain is stationary ergodic. The following assumption is sufficient to build estimators using the empirical distribution of the distribution of three consecutive observations.

Assumption 4. The densities $(f_x)_{x\in\mathbb{X}}$ are compactly supported and $C^\star = \int \frac{dy}{\sum_{x\in\mathbb{X}} f_x(y)} < \infty$, the integral is over the union of the supports of the emission densities.

Assumption 5. $(f_x)_{x\in\mathbb{X}}$ are linearly independent and belong to $C^s(\mathbb{R})$ the space of locally Hölder continuous functions.

We now state the theorem. Its proof is given in Section 3.6.13.

Theorem 3.3.11. *Let $D_n \to +\infty$ arbitrarily slowly. There exists a sequence of randomized estimators $(\hat{\theta}_n)_{n\geq 1}$ such that for all $\theta \in \Theta^{\mathrm{dep}}$ satisfying Assumption 1 to 5*

$$\mathbb{E}[\mathcal{R}_n^{\mathrm{clust}}(\theta, \pi_n \circ h^\star_{\hat{\theta}})] - \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta, g) = \mathcal{O}\left(D_n^{5/2}\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right)$$

where the expectation $\mathbb{E}[\cdot]$ is understood with respect to the randomness of the algorithm.

The cornerstone of the proof is the analysis of how errors in estimation of the parameters propagate to errors in the filtering and smoothing distributions proved in De Castro et al. [2017]. To achieve this rate for the excess risk, we must specify an estimator. Here we make the choice of using a modified version of the spectral algorithm of Abraham et al. [2022] which relies on the tensor method developped in Anandkumar et al. [2014]. The full algorithm is given in details in Section 3.6.13. The main difference with the original algorithm of Abraham et al. [2022] is a modification guaranteeing that the algorithm outputs the same permutation for the estimation of the transition matrix $\hat{Q}$ and the emission densities $\hat{f}_1, \ldots, \hat{f}_J$. Indeed, in the previous works, the aim was more to get upper bounds on the risks up to label-switching, and the analysis was done for Q and the emission distributions separately which does not guarantee that the permutation used in the control of the error of estimation of Q and the emission densities is the same. Albeit the modification is rather natural, proof that it works require a substantial effort.

Notice that the rate on the excess risk is exactly the estimation rate of Hölder regular functions in sup-norm. However, since we do not observe realizations of each emission density separately, this rate is not a straightforward consequence of density estimation theory. See Abraham et al. [2022] where the usual non-parametric rate for the estimation of densities in sup-norm is obtained in the HMM context.

Though we analyse the question for the spectral estimator, we believe that results similar to Theorem 3.3.11 hold for most estimation procedures previously proposed in the literature, putting in the upper bounds the bounds on the estimation risk up to label-switching obtained in those works. This is why we shall use in Section 3.4 the least-squares estimation method for which a public and efficient code exists De Castro [2016].

3.4 Numerical simulations

We present here the results of numerical simulations which leverage the added value of non-parametric clustering under hidden Markov modelling. We will consider two examples

in which the hidden states will be generated through the same transition matrix

$$Q = \begin{pmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{pmatrix}.$$

Example 1. A sample of $n = 5.10^4$ observations of a HMM with transition matrix Q and emissions: $F_1 = \frac{1}{2}\left(\mathcal{N}(1.7, 0.2) + \mathcal{N}(7, 0.15)\right)$ and $F_2 = \frac{1}{2}\left(\mathcal{N}(3.5, 0.2) + \mathcal{N}(5, 0.4)\right)$.

Example 2. A sample of $n = 10^5$ observations of a HMM with transition matrix Q and emissions: $F_1 = \frac{1}{2}\left(\mathcal{N}(3, 0.6) + \mathcal{N}(7, 0.4)\right)$ and $F_2 = \frac{1}{2}\left(\mathcal{N}(5, 0.3) + \mathcal{N}(9, 0.4)\right)$.

On these examples, we use the plug-in classifier whose clustering risk has been controlled in Theorem 3.3.11. Recall our procedure works as follows:

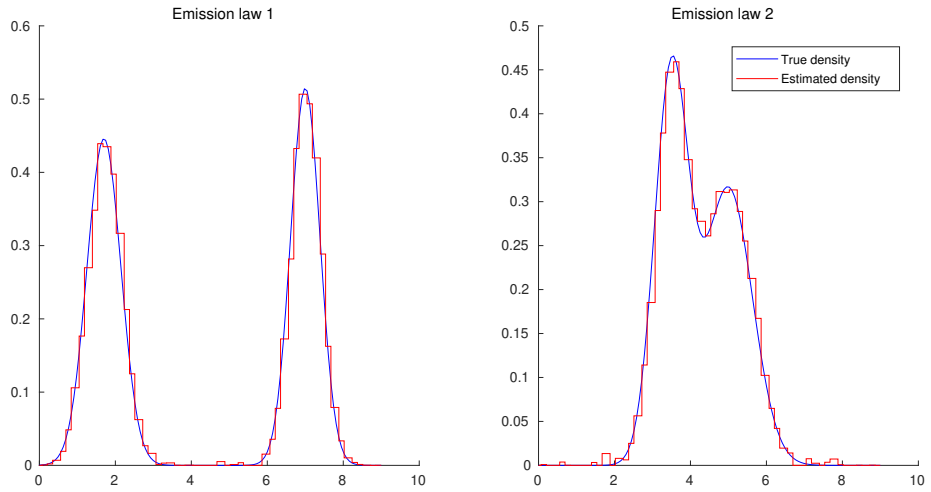
- First, the emission densities are estimated using the observations. We use the penalized least squares estimator proposed in De Castro et al. [2016] whose code has been made public in De Castro [2016]. We use the histogram basis for the estimation.
- Second, the Forward-Backward algorithm is used to compute the a posteriori distributions of the hidden states under the estimated model parameters and given the observations.
- Third, the hidden states are estimated by maximizing the a posteriori distributions.

The results of clustering using the Forward-Backward algorithm will be compared to those using the k-means algorithm. Since we have access to the hidden states, the error of clustering can be estimated by choosing the best permutation. Figures 3.2, 3.3, and 3.4 display the results of estimation and clustering. Performance of Bayes classifier, plug-in classifier and k-means are reported in Table 3.1. If the observations were independent

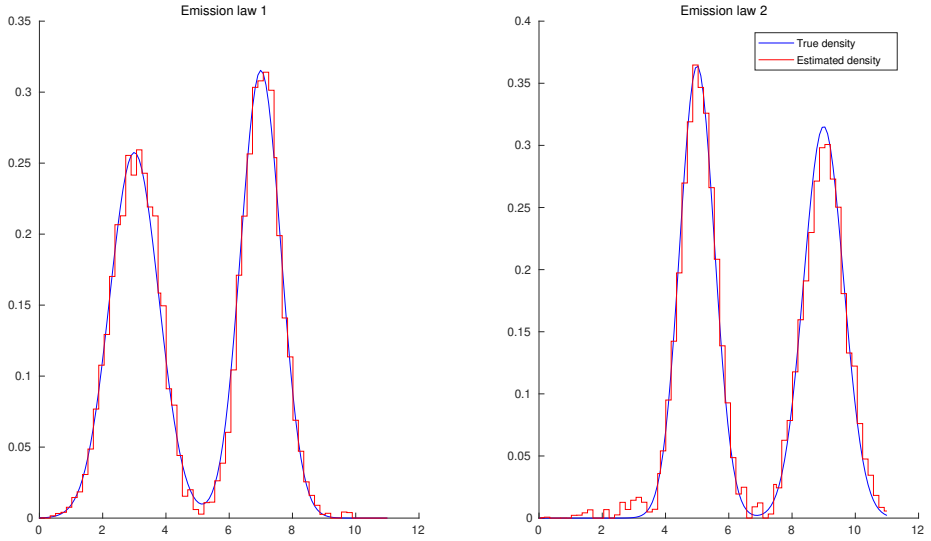
	Bayes classifier	Plug-in classifier	k-means algorithm	Λ
Example 1	1.56%	1.61%	46.7%	0.046
Example 2	6.42%	6.51%	47.3%	0.165

Table 3.1: Errors of clustering using three clustering rules: the Bayes classifier (using the true model parameters), the plug-in classifier (using the estimated parameters) and the k-means algorithm.

and the emissions modelled non-parametrically, the unique quantity that could have been estimated consistently would be the stationary distribution. However, under the HMM assumption, the estimation of each emission density with the minimax rate is possible. This is due to the identifiability of the model which holds even if no assumption is made on the emission densities. This is not possible in the independent case. Figure 3.2 shows the estimation results of the emission densities and confirms the theoretical properties of the estimator. On the other hand, Corollary 4 proves that in the case of a HMM with two classes, clustering errors could appear only in zones where the two emission densities overlap. In Figures 3.3b, and 3.4b, misclustered observations appear only in the overlaps between the emission densities. Compared to the k-means algorithm which is purely geometric and does not exploit the distribution of the observations, the plug-in procedure allows combining together observations even if they are geometrically distant from each other, which is not possible with the k-means algorithm. In this context of Gaussian mixtures, the performance of the k-means algorithm is mediocre as depicted in Table 3.1 and does not improve significantly when the overlap between the emission densities is small. However, for the plug-in procedure, the more separated are the emission densities, the better are the results of clustering.



(a) Estimation results for Example 1



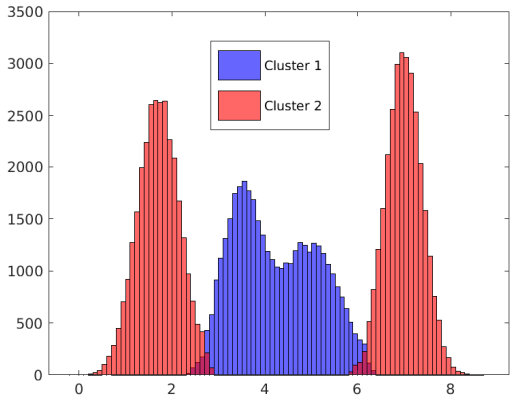
(b) Estimation results for Example 2

Figure 3.2: Non-parametric penalized least squares density estimation using the histogram basis for Example 1 and Example 2

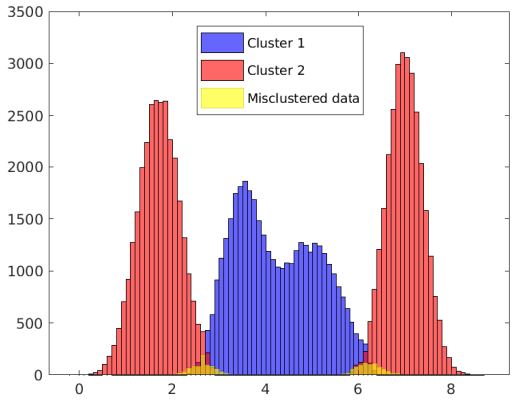
3.5 Discussions and Perspectives

This work focuses on an in-depth study of Bayes risks of clustering and classification. This analysis has led us to prove a form of equivalence between both risks. After identifying the key quantity which measures the difficulty of the classification task, it was extrapolated to the Bayes risk of clustering in several regimes. Finally, the excess risk of the plug-in procedure was studied. Although the analysis is sufficiently detailed to ensure a thorough understanding of both problems, there are still some interesting questions which were not covered by this work. We give a small overview below.

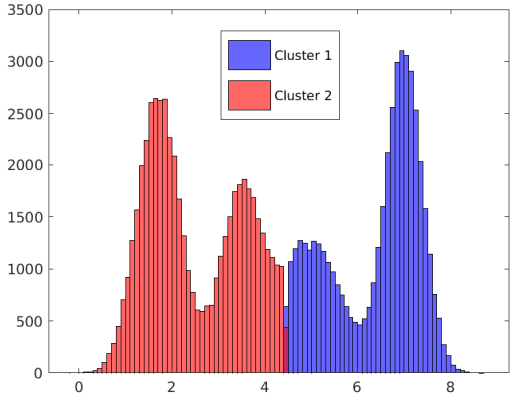
Lower-bound on entries of Q Throughout our analysis of the Bayes risk of clustering, we have used Assumption 1 which was crucial in obtaining the lower-bounds. In the



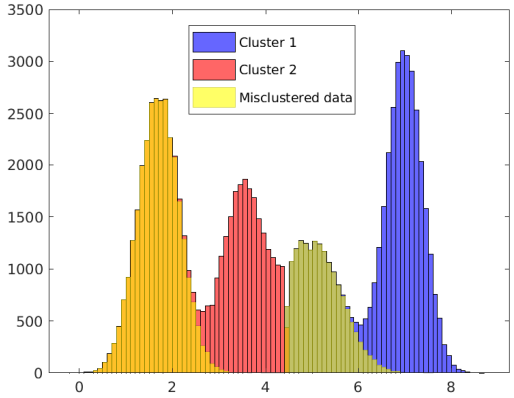
(a) Clustering using plug-in clusterer



(b) Misclustered observations for plug-in clusterer



(c) Clustering using k-means



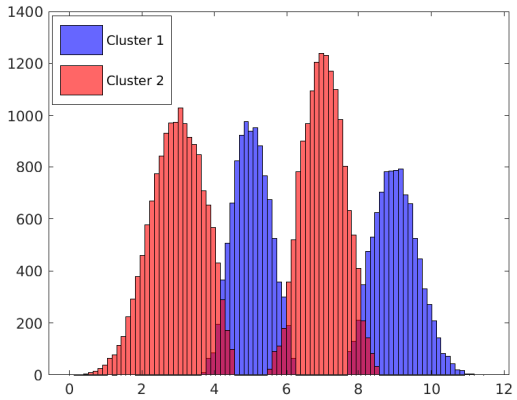
(d) Misclustered observations for k-means

Figure 3.3: Histograms of clusters and clustering errors for Example 1

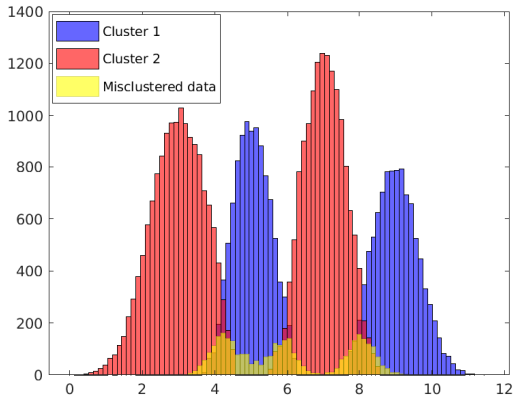
absence of such an assumption, the lower-bound of Theorem 3.3.10 no more matches the upper-bound and the magnitude of the Bayes risk of classification can not be precisely understood. The same thing applies to the Bayes risk of clustering. In addition, the control of the excess risk of the plug-in procedure is no more guaranteed.

Approaching the frontier to independence This situation happens when the emission distributions are nearly similar or when the transition matrix has almost equal lines. In this case, one can hope to improve the coefficient $\frac{\delta^2}{1-(J-1)\delta}$ which appears in the lower-bound on the risk of classification (Theorem 3.3.10) to δ as in the independent case. In fact, in this situation, the dependence between the observations is so weak and the effect of future and past observations on the current classification rule is so negligible that the magnitude of the Bayes risk of classification is the same as in the i.i.d. setting. On the other hand, as shown in Abraham et al. [2023, 2025], estimation of the model parameters no more becomes possible when approaching the frontier. The plug-in procedure should not work as well.

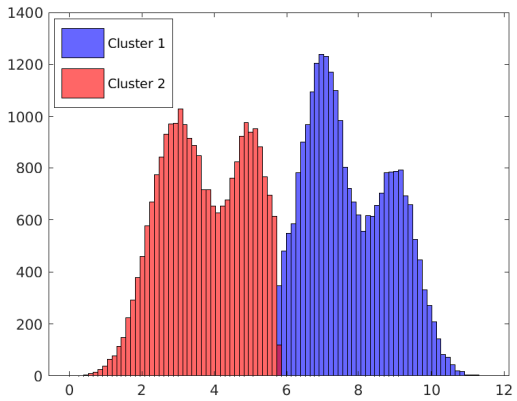
Lower bounds on the Bayes risk of clustering when it is very small Theorems 3.3.2, 3.3.5, 3.3.7 and 3.3.9 establish lower bounds on the Bayes risk of clustering in



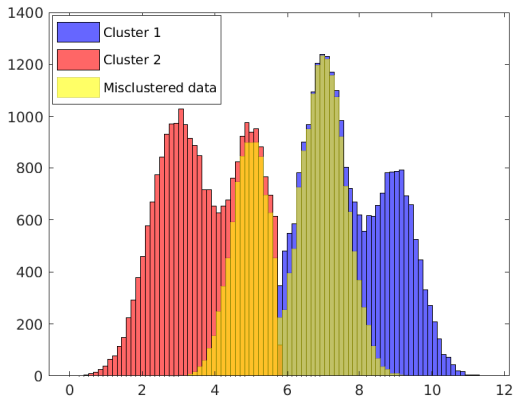
(a) Clustering using plug-in clusterer



(b) Misclustered observations for plug-in clusterer



(c) Clustering using k-means



(d) Misclustered observations for k-means

Figure 3.4: Histograms of clusters and clustering errors for Example 2

terms of the Bayes risk of classification. When $J = 2$ these bounds are meaningful regardless of how small is the Bayes risk of clustering. When $J > 2$, however, these bounds can be vacuous if the Bayes risk of classification gets too small. This is not an artifact of our bounds since we have shown in Proposition 3.3.3 that the two Bayes risks are not uniformly comparable when $J > 2$. Whence from the current work we only know that the Bayes risk of clustering is driven by Λ in the region of parameters for which it is not exponentially small in n and that it can not be driven by Λ otherwise. If it was the case, then it would be equivalent to the Bayes risk of classification in contradiction with the Proposition 3.3.3. Understanding the Bayes risk of clustering in the region of extreme parameters is still an open question.

Fast rates Theorem 3.3.11 has interest mainly in the situation where $\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \gtrsim (\log(n)/n)^{\frac{s}{2s+1}}$. In this regime, Theorem 3.3.11 tells us that the plug-in procedure has a risk of the same magnitude. However, in the situation where the magnitude of the Bayes risk of clustering is much smaller, that is when the emission distributions are very separated, one can hope to obtain faster rates. For example, when $\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) = \mathcal{O}\left(e^{-cn}\right)$ for c a positive absolute constant, one can hope to show that the risk of the plug-in procedure is exponentially small in n. The following lemma represents a first step for the proof of

such a result.

Lemma 3.5.1. *For all $0 < \gamma < 1/2$ and all $\theta \in \Theta$*

$$\mathcal{R}_n^{\text{class}}(\theta, h_{\hat{\theta}}) \leq \frac{\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)}{1/2 - \gamma} + \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_\theta \left(\left\| \phi_{\theta,i|n} - \phi_{\hat{\theta},i|n} \right\|_{\text{TV}} > \gamma \right). \tag{3.9}$$

where $\phi_{\theta,i|n} = \mathbb{P}_\theta \left(X_i \in . \mid Y_{1:n} \right)$ and $h_{\hat{\theta}}$ is the plug-in classifier defined in Section 3.3.4.

The proof of Lemma 3.5.1 is given in Section 3.5.1.

Observe that the second term of the rhs of (3.9) is a large deviation term which may eventually decrease exponentially fast in n. The only price to pay to obtain large deviation type of decay is a constant factor of at least 2 in front of $\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$. In situations where $\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$ is small, this might be advantageous compared to the bound in Theorem 3.3.11. In Proposition 2.2 of De Castro et al. [2017], the authors prove the inequality

$$\begin{aligned} \|\phi_{\theta,i|n}(., Y_{1:n}) - \phi_{\hat{\theta},i|n}(., Y_{1:n})\|_{\text{TV}} \leq \frac{4(1-\delta)}{\delta^2} \Bigg(& \rho^{i-1} \|\nu - \hat{\nu}\|_2 \\ & + (1/(1-\rho) + 1/(1-\hat{\rho})) \|Q - \hat{Q}\|_{\text{F}} + \sum_{l=1}^{n} \delta \frac{(\hat{\rho} \vee \rho)^{|l-i|}}{c^\star(Y_l)} \max_{x \in \mathbb{X}} \left| f_x(Y_l) - \hat{f}_x(Y_l) \right| \Bigg) \end{aligned} \tag{3.10}$$

where $c^\star(y) = \min_{x \in \mathbb{X}} \sum_{x' \in \mathbb{X}} Q(x, x') f_{x'}(y)$, $\rho = 1 - \delta/(1-\delta)$ and $\hat{\rho} = 1 - \hat{\delta}/(1-\hat{\delta})$ where $\hat{\delta}$ is an estimator of δ. (3.10) can be used to control the second term in the rhs of (3.9). This would require to derive large deviation inequalities for all the terms involved in (3.10), which turns out to be a rather challenging problem.

Optimal excess risk Although we obtain upper bounds on the excess risk of clustering of the plug-in procedure (Theorem 3.3.11), we do not know the optimal rate of decay of the excess risk. In particular, it is unknown if the plug-in procedure achieves optimal excess risk. The Lemma 3.5.1 suggests that when the Bayes risk is smaller than $\mathcal{O}(n^{-s/(2s+1)})$ then our upper bounds on the excess risk of the plug-in could be improved. Yet without optimality guarantees. Determining the optimal excess risk is an open and interesting question to investigate.

Alternatives to plug-in Under the hidden Markov modeling, the most straightforward way to take advantage of the identifiability of the model is to estimate the model parameters and use them for clustering through the plug-in procedure. Unlike the i.i.d. case where algorithms such as k-means can be used as an alternative, we do not know of any alternative to the plug-in in the HMM case. It would be very interesting to find clustering procedures that leverage the nonparametric identifiability of HMM without relying on estimating the parameters first.

3.6 Proofs

3.6.1 Proof of Theorem 3.3.1

Let $\theta \in \Theta^{\text{ind}}$. Thanks to Equation (3.5), the Bayes clusterer $g_\theta^\star$ can be defined as the partition $g_\theta^\star = \pi_n \circ \tilde{h}_\theta$ where $\tilde{h}_\theta = \left(\tilde{h}_{\theta,i} \right)_{i \in [n]}$ is the classifier minimizing:

$$h = (h_i)_{1 \leq i \leq n} \mapsto \mathbb{E}_\theta \left[\min_{\tau \in \mathcal{S}_2} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{h_i(Y_{1:n}) \neq \tau(X_i)} \right]. \tag{3.11}$$

Let $Y_{1:n}$ be n i.i.d. observations of the mixture with parameters θ. We have $g^\star_\theta(Y_{1:n}) = \pi_n \circ \tilde{h}_\theta(Y_{1:n})$ a.e, where

$$\tilde{h}_\theta(Y_{1:n}) \in \underset{h=(h_i)_{i\in[n]}}{\arg\min}\ \mathbb{E}_\theta\left[\min_{\tau\in\mathcal{S}_2}\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq\tau(X_i)}\middle| Y_{1:n}\right].$$

Given that:

$$\begin{aligned}\mathbb{E}_\theta\left[\min_{\tau\in\mathcal{S}_2}\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq\tau(X_i)}\middle| Y_{1:n}\right] &= \mathbb{E}_\theta\left[\min\left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq X_i}, 1-\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq X_i}\right)\middle| Y_{1:n}\right]\\ &= \frac{1}{2} - \frac{1}{n}\mathbb{E}_\theta\left[\left|\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq X_i} - \frac{n}{2}\right|\middle| Y_{1:n}\right],\end{aligned}$$

one gets $\tilde{h}_\theta(Y_{1:n}) \in \arg\max_{h=(h_i)_{i\in[n]}} \mathbb{E}_\theta\left[\left|\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq X_i} - \frac{n}{2}\right|\middle| Y_{1:n}\right]$. We now consider the following lemma. Its proof is due to Ziv Scully and can be found in Scully [2024]. We will detail its proof in Section 3.6.16 for the sake of completeness.

Lemma 3.6.1. *Let $(Z_i)_{i\in[n]}$ be a sequence of independent Bernoulli random variables such that $Z_i \sim \mathcal{B}(p_i)$ where $p_i \in \{\alpha_i, 1-\alpha_i\}$. Then the maximum $\max_{(p_i)_{i\in[n]}} \mathbb{E}\left[\left|\sum_{i\in[n]} Z_i - \frac{n}{2}\right|\right]$ is reached at $(p_i)_{i\in[n]} = (\alpha_i \wedge (1-\alpha_i))_{i\in[n]}$ and $(p_i)_{i\in[n]} = (\alpha_i \vee (1-\alpha_i))_{i\in[n]}$.*

We apply this lemma to the random variables $\left(\mathbf{1}_{h_i(Y_{1:n})\neq X_i}\right)_{i\in[n]}$ which are independent conditionally to $Y_{1:n}$ and ensure :

$$\mathbf{1}_{h_i(Y_{1:n})\neq X_i} \mid Y_{1:n} \sim \mathcal{B}(p_i(Y_i))$$

such that $p_i(Y_i) \in \{\mathbb{P}_\theta(X_i = 1 \mid Y_i), \mathbb{P}_\theta(X_i = 2 \mid Y_i)\}$. Two cases occur:

- $(\forall i \in [n])\ \mathbb{P}_\theta(X_i = 1 \mid Y_i) \neq 1/2$, then the Bayes classifier is unique and Lemma 3.6.1 allows us to conclude that $\tilde{h}_\theta(Y_{1:n}) = \left(\arg\max_{x=1,2} \mathbb{P}_\theta(X_i = x \mid Y_i)\right)_{i\in[n]} = h^\star_\theta(Y_{1:n})$. Consequently:
$$g^\star_\theta(Y_{1:n}) = \pi_n \circ h^\star_\theta(Y_{1:n}).$$
- $(\exists i \in [n])\ \mathbb{P}_\theta(X_i = 1 \mid Y_i) = 1/2$, then the same argument yields
$$g^\star_\theta(Y_{1:n}) = \pi_n \circ h^\star_\theta(Y_{1:n}).$$
where we abuse the notation of $h^\star_\theta(Y_{1:n})$ to refer not only to a specific Bayes classifier but to the set of all the Bayes classifiers since it is not unique. The same is done for the Bayes clusterer $g^\star_\theta(Y_{1:n})$. The permutation π_n is then applied to the set of all Bayes classifiers.

3.6.2 Proof of Theorem 3.3.2

We first prove the upper bound. Let define $h^\star_{\theta,i}(Y_i) = \arg\max_{a\in\{1,2\}} \mathbb{P}_\theta(X_i = a \mid Y_i)$. Let $Z_n = \sum_{i=1}^n \mathbf{1}_{X_i \neq h^\star_{\theta,i}(Y_i)}$. One gets:

$$\begin{aligned}\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) - \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) &= \mathbb{E}_\theta\left[\left(\frac{2}{n}Z_n - 1\right)\mathbf{1}_{\frac{2}{n}Z_n > 1}\right]\\ &= \int_0^1 \mathbb{P}_\theta\left(\left(\frac{2}{n}Z_n - 1\right)\mathbf{1}_{\frac{2}{n}Z_n>1} > x\right) dx.\end{aligned}$$

Namely,

$$\inf_{h\in\mathcal{H}_n}\mathcal{R}_n^{\text{class}}(\theta,h)-\inf_{g\in\mathcal{G}_n}\mathcal{R}_n^{\text{clust}}(\theta,g)=\int_0^1\mathbb{P}_\theta\left(Z_n>\frac{n}{2}\,(x+1)\right)dx=:J_n. \tag{3.12}$$

Chernoff bound yields:

$$\begin{aligned}\mathbb{P}_\theta\left(Z_n>\frac{n}{2}\,(x+1)\right)&\le\inf_\lambda\left\{\exp\left(-\frac{n\lambda}{2}(x+1)\right)\mathbb{E}_\theta\left[\exp\left(\lambda Z_n\right)\right]\right\}dx\\&\le\left(\left(\frac{1+x}{1-2\varepsilon_{n,\theta}}\right)^{\frac{1+x}{2}}\left(\frac{1-x}{1+2\varepsilon_{n,\theta}}\right)^{\frac{1-x}{2}}\right)^{-n}.\end{aligned}$$

Let $g(t)=\frac{1+t}{2}\log\left(\frac{1+t}{1-2\varepsilon_{n,\theta}}\right)+\frac{1-t}{2}\log\left(\frac{1-t}{1+2\varepsilon_{n,\theta}}\right)$. Then, $g'(t)=\frac{1}{2}\log\left(\frac{1+t}{1-t}\right)+\frac{1}{2}\log\left(\frac{1+2\varepsilon_{n,\theta}}{1-2\varepsilon_{n,\theta}}\right)$ and $g''(t)=\frac{1}{1-t^2}$. Deduce that $g(t)\ge g(0)+\max(g'(0)t,\frac{t^2}{2})$ for all $t\in(0,1)$. Then,

$$\begin{aligned}J_n&\le e^{-ng(0)}\int_0^1 e^{-n\max(g'(0)t,\frac{t^2}{2})}\mathrm{d}t\\&\le e^{-ng(0)}\min\left(\int_0^1 e^{-ng'(0)t}\mathrm{d}t,\int_0^1 e^{-nt^2/2}\mathrm{d}t\right)\\&\le\min\left(\frac{e^{-ng(0)}}{ng'(0)},\sqrt{\frac{\pi}{2n}}\right)\end{aligned}$$

We now derive the more challenging lower bound. We assume throughout that $n\ge 100$. We also assume that $\inf_{h\in\mathcal{H}_n}\mathcal{R}_n^{\text{class}}(\theta,h)>0$ otherwise the lower bound is zero and holds trivially. Suppose first that $0\le\varepsilon_{n,\theta}\le\frac{1}{10\sqrt{n}}$. Then,

$$\begin{aligned}J_n&\ge\int_0^{\frac{1}{10\sqrt{n}}}\mathbb{P}\Big(S_n>\frac{n(x+1)}{2}\Big)\mathrm{d}x\\&\ge\frac{1}{10\sqrt{n}}\mathbb{P}\Big(S_n>\frac{n}{2}+\frac{\sqrt{n}}{20}\Big)\\&=\frac{1}{10\sqrt{n}}\mathbb{P}\Big(\frac{S_n-n(\frac{1}{2}-\varepsilon_{n,\theta})}{\sqrt{n(1-\varepsilon_{n,\theta}^2)/4)}}>\frac{\frac{\sqrt{n}}{20}+n\varepsilon_{n,\theta}}{\sqrt{n(1-\varepsilon_{n,\theta}^2)/4}}\Big)\\&\ge\frac{1}{10\sqrt{n}}\mathbb{P}\Big(\frac{S_n-n(\frac{1}{2}-\varepsilon_{n,\theta})}{\sqrt{n(1-\varepsilon_{n,\theta}^2)/4}}>\sqrt{\frac{10}{111}}\Big)\end{aligned}$$

because $n\ge 100$ and $\varepsilon_{n,\theta}\le\frac{1}{10\sqrt{n}}\le\frac{1}{100}$. By the theorem of Berry and Esseen,

$$J_n\ge\frac{1}{\sqrt{n}}\left(1-\Phi(\sqrt{\frac{10}{111}})-\frac{0.4748}{\sqrt{n(1-\varepsilon_{n,\theta}^2)/4}}\right)\ge\frac{0.2870}{\sqrt{n}}$$

since $n\ge 100$. Finally, because $\varepsilon_{n,\theta}\le\frac{1}{10\sqrt{n}}$ it must be that

$$\begin{aligned}\frac{\exp\big(-ng(0)\big[1+\frac{6.8}{1\vee 10\sqrt{n\varepsilon_{n,\theta}^2}}\big]\big)}{ng'(0)}&\ge\frac{1}{\sqrt{n}}\frac{\exp\big(-\frac{1}{2}\log(1-0.04/n)[1+6.8]\big)}{\frac{1}{2\sqrt{n}}\log(\frac{1+0.2/\sqrt{n}}{1-0.2/\sqrt{n}})}\\&\ge\frac{500}{\sqrt{n}}.\end{aligned}$$

Deduce that for a universal constant $B > 0$

$$J_n \geq B \min\left(\frac{\exp\big(- ng(0)\big[1 + \frac{6.8}{1 \vee 10\sqrt{n\varepsilon_{n,\theta}^2}}\big]\big)}{ng'(0)}, \frac{1}{\sqrt{n}} \right). \tag{3.13}$$

Now suppose $\frac{1}{10\sqrt{n}} < \varepsilon_{n,\theta} < \frac{1}{2}$. We first lower bound,

$$J_n \geq \int_0^{\frac{1}{10\sqrt{n}}} \mathbb{P}\Big(S_n > \frac{n(x+1)}{2}\Big) \mathrm{d}x.$$

We lower bound the probability $P(S_n > \frac{n}{2}(x+1))$ using Cramér's technique. In the next $0 \leq x \leq \frac{1}{10\sqrt{n}}$. Then for all $\lambda > 0$ and all $0 < \delta < \frac{n}{2} - \frac{\sqrt{n}}{20}$ (which guarantees that $\frac{n(1+x)}{2} + \delta < n$] we have

$$\begin{aligned}
\mathbb{P}\Big(S_n > \frac{n}{2}(x+1)\Big) &= \sum_{\frac{n(x+1)}{2} < y \leq n} \binom{n}{y} r^y (1-r)^{n-y} \\
&\geq \sum_{\frac{n(x+1)}{2} < y < \frac{n(x+1)}{2} + \delta} \binom{n}{y} e^{-\lambda y + n\psi_r(\lambda)} \frac{(re^\lambda)^y (1-r)^{n-y}}{\exp(n\psi_r(\lambda))} w
\end{aligned}$$

where $\psi_r(\lambda) = \log(1 - r + re^\lambda)$ is the cumulant generating function of the Bernoulli distribution with parameter r; where $r \equiv \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h)$ for simplicity. Here one notice that $y \mapsto \binom{n}{y} \frac{(pe^\lambda)^y (1-p)^{n-y}}{\exp(n\psi_r(\lambda))}$ is the density of the Binomial distribution with parameters (n, q_λ) where $q_\lambda = \frac{re^\lambda/(1-r)}{1 + re^\lambda/(1-r)}$. Letting $\tilde{S}_n \sim \mathrm{Binomial}(n, q_\lambda)$, it is seen that

$$\mathbb{P}\Big(S_n > \frac{n}{2}(x+1)\Big) \geq e^{-\lambda(\frac{n(x+1)}{2} + \delta) + n\psi_r(\lambda)} \mathbb{P}\left(\frac{n}{2}(x+1) < \tilde{S}_n < \frac{n}{2}(x+1) + \delta \right).$$

Now we make the choice that $q_\lambda = \frac{1+x}{2} + \delta/n$, which corresponds to $\lambda = -\log\big(1 - \frac{1+x+2\delta/n}{2}\big) + \log\big(\frac{1-r}{r}\big)$. Observe that λ is well-defined and positive since by assumption $0 < r \leq \frac{1}{2}$ and $1 + x + 2\delta/n < 2$; this also guarantees that $0 < q_\lambda < 1$. Then,

$$\begin{aligned}
\mathbb{P}\Big(S_n > \frac{n}{2}(x+1)\Big) &\geq e^{-nI_r(\frac{1+x}{2} + \frac{\delta}{n})} \mathbb{P}\Big(\frac{n(1+x)}{2} < \tilde{S}_n < \frac{n(1+x)}{2} + \delta \Big) w \\
&= e^{-ng(x+2\delta/n)} \mathbb{P}\Big(- \frac{\delta}{\sqrt{nq_\lambda(1-q_\lambda)}} < \frac{\tilde{S}_n - nq_\lambda}{\sqrt{nq_\lambda(1-q_\lambda)}} < 0 \Big)
\end{aligned}$$

By the theorem of Berry and Esseen, we can conclude that

$$\begin{aligned}
\mathbb{P}\Big(S_n > \frac{n}{2}(x+1)\Big) &\geq e^{-ng(x+2\delta/n)} \left(\Phi(0) - \Phi\Big(- \frac{\delta}{\sqrt{nq_\lambda(1-q_\lambda)}} \Big) - 2\frac{0.4748}{\sqrt{nq_\lambda(1-q_\lambda)}} \right) \\
&\geq e^{-ng(x+2\delta/n)} \left(\frac{1}{2} - \Phi\Big(- \frac{2\delta}{\sqrt{n}} \Big) - \frac{1.8992}{\sqrt{n(1-\kappa_x^2)}} \right)
\end{aligned}$$

where $\kappa_x \coloneqq x + \frac{2\delta}{n}$. We choose $\delta = -\frac{\sqrt{n}}{2}\Phi^{-1}(1/4)$. This implies that $\kappa_x \leq \frac{0.1 - \Phi^{-1}(1/4)}{\sqrt{n}} \leq \frac{0.7745}{\sqrt{n}}$ and $\Phi(-2\delta/\sqrt{n}) = \frac{1}{4}$. Consequently for $n \geq 100$,

$$\mathbb{P}\Big(S_n > \frac{n}{2}(x+1)\Big) \geq 0.0595 \cdot e^{-ng(x+2\delta/n)}, \text{ and, } \quad J_n \geq 0.0595 \int_0^{\frac{1}{10\sqrt{n}}} e^{-ng(x+2\delta/n)} \mathrm{d}x.$$

A Taylor expansion of g near zero yields the existence of $u \in (0, x + 2\delta/n)$ such that $g(x + 2\delta/n) = g(0) + g'(0)(x + 2\delta/n) + g''(u)(x + 2\delta/n)^2/2$. But when $x \in (0, \frac{1}{10\sqrt{n}})$ it must be that $0 \le x + 2\delta/n \le \frac{0.1 - \Phi^{-1}(1/4)}{\sqrt{n}} \le \frac{0.7745}{\sqrt{n}}$ and thus $g''(u) = \frac{1}{1-u^2} \le 1.0061$ when $n \ge 100$. Hence,

$$\begin{aligned} J_n &\ge 0.0595 \cdot e^{-ng(0) - 2\delta g'(0) - \frac{1.0061 \cdot 0.7745^2}{2}} \int_0^{\frac{1}{10\sqrt{n}}} e^{-ng'(0)x} \mathrm{d}x \\ &= 0.0440 \cdot \left(1 - e^{-\frac{g'(0)\sqrt{n}}{10}}\right) \frac{e^{-ng(0)[1 + \frac{2\delta g'(0)}{ng(0)}]}}{ng'(0)}. \end{aligned}$$

Here recall that $g(0) = -\frac{1}{2}\log(1 - 4\varepsilon_{n,\theta}^2)$ and $g'(0) = \frac{1}{2}\log\left(\frac{1+2\varepsilon_{n,\theta}}{1-2\varepsilon_{n,\theta}}\right)$. It follows that the function $\varepsilon_{n,\theta} \mapsto g'(0) - \frac{g(0)}{\varepsilon_{n,\theta}}$ admits the derivative $\varepsilon_{n,\theta} \mapsto \frac{-(1-4\varepsilon_{n,\theta}^2)\log(1-4\varepsilon_{n,\theta}^2) - 4\varepsilon_{n,\theta}^2}{\varepsilon_{n,\theta}^2(1-4\varepsilon_{n,\theta}^2)}$ which is negative for $\varepsilon_{n,\theta} > 0$. Deduce that $g'(0) - \frac{g(0)}{\varepsilon_{n,\theta}} \le 0$, or equivalently $\frac{g'(0)}{g(0)} \le \frac{1}{\varepsilon_{n,\theta}}$. Therefore,

$$\frac{2\delta g'(0)}{ng(0)} \le \frac{-\Phi^{-1}(1/4)}{\sqrt{n}\varepsilon_{n,\theta}} \le \frac{6.8}{\max(1, 10\sqrt{n\varepsilon_{n,\theta}^2})}.$$

Similarly since $\varepsilon_{n,\theta} \mapsto g'(0)$ is monotonically increasing, $\frac{g'(0)\sqrt{n}}{10} \ge \frac{\frac{1}{2}\log(\frac{1+0.4/\sqrt{n}}{1-0.4/\sqrt{n}})\sqrt{n}}{10} \ge 0.04$. Hence,

$$J_n \ge 0.0017 \cdot \frac{\exp\left(-ng(0)\left[1 + \frac{6.8}{1 \vee 10\sqrt{n\varepsilon_{n,\theta}^2}}\right]\right)}{ng'(0)}$$

Finally, since $\varepsilon_{n,\theta} > \frac{1}{10\sqrt{n}}$ the above computations show that $ng'(0) \ge 0.4\sqrt{n}$. Therefore there is a universal constant $B > 0$ such that Equation (3.13) is also satisfied when $\frac{1}{10\sqrt{n}} < \varepsilon_{n,\theta} < \frac{1}{2}$.

3.6.3 Proof of Theorem 3.3.4

Given two partitions A and B, we recall the clustering loss defined in Equation (3.2):

$$\ell(A, B) = 1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(A,B) \\ M \text{ is a matching}}} \sum_{\{C,C'\} \in M} \mathrm{Card}(C \cap C')$$

We define

$$(\forall i \in [n])\,(\forall k \in \mathbb{X}) \quad \alpha_k^{(Y_i)} = \frac{\nu_k f_k(Y_i)}{\sum_{j=1}^J \nu_j f_j(Y_i)}$$

and $h_\theta^\star(Y_{1:n}) = \left(h_{\theta,i}^\star(Y_i)\right)_{i \in [n]}$ where $(\forall i \in [n])\ \ h_{\theta,i}^\star(Y_i) = \arg\max_{k \in \mathbb{X}} \alpha_k^{(Y_i)}$.
Consider the event :

$$A_n = \bigcup_{j=1}^J \bigcap_{i=1}^n \left\{ \max_{k \ne j} \nu_k f_k(Y_i) < \nu_j f_j(Y_i) \right\}.$$

Then,

$$A_n \subset \left\{ \pi_n\left(\left(h_{\theta,i}^\star(Y_i)\right)_{i \in [n]}\right) = \pi_n\left((1, \dots, 1)\right) \right\}.$$

Let $\theta \in \Theta^{\text{ind}}$, such that

$$\mathbb{P}_\theta \left(\bigcup_{j=1}^{J} \left\{ 0 < \max_{l \neq j} \nu_l f_l(Y) < \nu_j f_j(Y) \leq \sum_{l \neq j} \nu_l f_l(Y) \right\} \right) > 0.$$

Then $\mathbb{P}_\theta(A_n) > 0$. Since:

$$\begin{aligned} &A_n \cap \left\{ \mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1))) \middle| Y_{1:n} \right] > \mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1,2))) \middle| Y_{1:n} \right] \right\} \\ &\subset \left\{ g_\theta^\star(Y_{1:n}) \neq \pi_n \circ h_\theta^\star(Y_{1:n}) \right\}, \end{aligned}$$

it suffices then to show that

$$\mathbb{P}_\theta \left(\mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1))) \middle| Y_{1:n} \right] > \mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1,2))) \middle| Y_{1:n} \right] \middle| A_n \right) > 0.$$

Let:

$$\begin{aligned} k_n^{(1)}(x_{1:n}) &= \arg\max_{i \in \mathbb{X}} \sum_{k=1}^{n} \mathbf{1}_{x_k = i} \\ N_n^{(1)}(x_{1:n}) &= \max_{i \in \mathbb{X}} \sum_{k=1}^{n} \mathbf{1}_{x_k = i} \\ N_n^{(2)}(x_{1:n}) &= \max_{i \neq k_n^{(1)}(x_{1:n})} \sum_{k=1}^{n} \mathbf{1}_{x_k = i} \\ N_{n,i}(x_{1:n}) &= \sum_{k=1}^{n} \mathbf{1}_{x_k = i} \end{aligned}$$

First, note that for $x_{1:n} \in \mathbb{X}^n$:

$$\ell(\pi_n(x_{1:n}), \pi_n(1,\ldots,1,2)) = \begin{cases} n - N_n^{(1)}(x_{1:n}) - 1 & \text{if } x_n \neq k_n^{(1)}(x_{1:n}) \\ n - N_n^{(1)}(x_{1:n}) + 1 & \text{if } N_n^{(2)}(x_{1:n}) < N_n^{(1)}(x_{1:n}) - 1, x_n = k_n^{(1)}(x_{1:n}) \\ n - N_n^{(1)}(x_{1:n}) & \text{if } N_n^{(2)}(x_{1:n}) = N_n^{(1)}(x_{1:n}) - 1, x_n = k_n^{(1)}(x_{1:n}) \\ n - N_n^{(1)}(x_{1:n}) - 1 & \text{if } N_n^{(2)}(x_{1:n}) = N_n^{(1)}(x_{1:n}), x_n = k_n^{(1)}(x_{1:n}) \end{cases}$$

The inequality

$$\mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1))) \middle| Y_{1:n} \right] > \mathbb{E}_\theta \left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1,2))) \middle| Y_{1:n} \right] \tag{3.14}$$

is equivalent to:

$$\sum_{x_{1:n} \in \mathbb{X}^n} \ell(\pi_n(x_{1:n}), \pi_n((1,...,1))) \prod_{i=1}^{n} \alpha_{x_i}^{(Y_i)} > \sum_{x_{1:n} \in \mathbb{X}^n} \ell(\pi_n(x_{1:n}), \pi_n((1,..,1,2))) \prod_{i=1}^{n} \alpha_{x_i}^{(Y_i)}$$

Conditionally to $Y_{1:n}$, $X_1, ..., X_n$ are independent multinomial random variables such that $X_i \middle| Y_i \sim \left(\alpha_k^{(Y_i)} \right)_{k \in \mathbb{X}}$. Using the expression of $\ell(\pi_n(x_{1:n}), \pi_n((1,\ldots,1,2)))$ and the fact that

$\ell(\pi_n(x_{1:n}), \pi_n((1,\dots,1))) = n - N_n^{(1)}(x_{1:n})$, one obtains:

$$
\begin{aligned}
(3.14) \Leftrightarrow & \sum_{\substack{x_{1:n}\in\mathbb{X}^n \\ x_n \neq k_n^{(1)}(x_{1:n}) \text{ or} \\ x_n = k_n^{(1)}(x_{1:n}) \text{ and } N_n^{(2)}(x_{1:n}) = N_n^{(1)}(x_{1:n})}} \prod_{i=1}^{n} \alpha_{x_i}^{(Y_i)} > \sum_{\substack{x_{1:n}\in\mathbb{X}^n \\ x_n = k_n^{(1)}(x_{1:n}) \text{ and } N_n^{(2)}(x_{1:n}) < N_n^{(1)}(x_{1:n})-1}} \prod_{i=1}^{n} \alpha_{x_i}^{(Y_i)} \\
& \Leftrightarrow \mathbb{P}_\theta\left(X_n \neq k_n^{(1)}(X_{1:n}) \middle| Y_{1:n}\right) + \mathbb{P}_\theta\left(X_n = k_n^{(1)}(X_{1:n}), N_n^{(2)}(X_{1:n}) = N_n^{(1)}(X_{1:n}) \middle| Y_{1:n}\right) \\
& > \mathbb{P}_\theta\left(X_n = k_n^{(1)}(X_{1:n}), N_n^{(2)}(X_{1:n}) < N_n^{(1)}(X_{1:n}) - 1 \middle| Y_{1:n}\right) \\
& \Leftrightarrow \mathbb{P}_\theta\left(X_n \neq k_n^{(1)}(X_{1:n}) \middle| Y_{1:n}\right) - \mathbb{P}_\theta\left(X_n = k_n^{(1)}(X_{1:n}), N_n^{(2)}(X_{1:n}) < N_n^{(1)}(X_{1:n}) - 1 \middle| Y_{1:n}\right) \\
& + \mathbb{P}_\theta\left(X_n = k_n^{(1)}(X_{1:n}), N_n^{(2)}(X_{1:n}) = N_n^{(1)}(X_{1:n}) \middle| Y_{1:n}\right) > 0
\end{aligned}
$$

By marginalization over the possible values of X_n, one gets:

$$
\begin{aligned}
(3.14) \Leftrightarrow & \sum_{j=1}^{J} \Bigg[\mathbb{P}_\theta\left(X_n \neq j, N_{n,j}(X_{1:n}) \geq \max_{k\neq j} N_{n,k}(X_{1:n}) \middle| Y_{1:n}\right) \\
& - \mathbb{P}_\theta\left(X_n = j, N_{n,j}(X_{1:n}) > \max_{k\neq j} N_{n,k}(X_{1:n}) + 1 \middle| Y_{1:n}\right) \\
& + \mathbb{P}_\theta\left(X_n = j, N_{n,j}(X_{1:n}) = \max_{k\neq j} N_{n,k}(X_{1:n}) \middle| Y_{1:n}\right) \Bigg] > 0 \\
\Leftrightarrow & \sum_{j=1}^{J} \sum_{l\neq j} \Bigg[\alpha_l^{(Y_n)} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) \geq \max_{k\neq j,l} N_{n-1,k}(X_{1:n-1}) \vee (N_{n-1,l}(X_{1:n-1}) + 1) \middle| Y_{1:n-1}\right) \\
& - \frac{\alpha_l^{(Y_n)} \alpha_j^{(Y_n)}}{1-\alpha_j^{(Y_n)}} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) > \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) \middle| Y_{1:n-1}\right) \\
& \frac{\alpha_l^{(Y_n)} \alpha_j^{(Y_n)}}{1-\alpha_j^{(Y_n)}} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) = \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) - 1 \middle| Y_{1:n-1}\right) \Bigg] > 0 \\
\Leftrightarrow & \sum_{j=1}^{J} \sum_{l\neq j} \Bigg[\alpha_l^{(Y_n)} \Bigg\{ \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) \geq \max_{k\neq j,l} N_{n-1,k}(X_{1:n-1}) \vee (N_{n-1,l}(X_{1:n-1}) + 1) \middle| Y_{1:n-1}\right) \\
& - \frac{\alpha_j^{(Y_n)}}{1-\alpha_j^{(Y_n)}} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) > \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) \middle| Y_{1:n-1}\right) \Bigg\} \Bigg] \\
& + \sum_{j=1}^{J} \alpha_j^{(Y_n)} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) = \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) - 1 \middle| Y_{1:n-1}\right) > 0.
\end{aligned}
$$

On the one hand, since:

$$
\begin{aligned}
& \left\{ N_{n-1,j}(X_{1:n-1}) > \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) \right\} \\
& \subset \left\{ N_{n-1,j}(X_{1:n-1}) \geq \max_{k\neq j,l} N_{n-1,k}(X_{1:n-1}) \vee (N_{n-1,l}(X_{1:n-1}) + 1) \right\}
\end{aligned}
$$

and $\alpha_k^{(Y_n)} \leq \frac{1}{2} \iff \frac{\alpha_k^{(Y_n)}}{1-\alpha_k^{(Y_n)}} \leq 1$, one gets:

$$\bigcap_{k\in\mathbb{X}} \left\{\alpha_k^{(Y_n)} \leq \frac{1}{2}\right\}$$
$$\subset \left\{\sum_{j=1}^{J}\sum_{l\neq j} \alpha_l^{(Y_n)} \left[\mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) \geq \max_{k\neq j,l} N_{n-1,k}(X_{1:n-1}) \vee (N_{n-1,l}(X_{1:n-1})+1) \middle| Y_{1:n-1}\right)\right.\right.$$
$$\left.\left. - \frac{\alpha_j^{(Y_n)}}{1-\alpha_j^{(Y_n)}} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) > \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) \middle| Y_{1:n-1}\right)\right] \geq 0\right\}.$$

On the other hand,

$$\bigcup_{l\neq j\in\mathbb{X}} \bigcap_{i=1}^{n} \left\{\alpha_l^{(Y_i)} \wedge \alpha_j^{(Y_i)} > 0\right\} \subset$$
$$\left\{\sum_{j=1}^{J} \alpha_j^{(Y_n)} \mathbb{P}_\theta\left(N_{n-1,j}(X_{1:n-1}) = \max_{k\neq j} N_{n-1,k}(X_{1:n-1}) - 1 \middle| Y_{1:n-1}\right) > 0\right\}$$

because for $k \in \mathbb{X}$, $N_{n-1,k}(X_{1:n-1})$ is a sum of Bernoulli random variables. Consequently,

$$\bigcup_{l\neq j\in\mathbb{X}} \bigcap_{i=1}^{n} \left\{\alpha_l^{(Y_i)} \wedge \alpha_j^{(Y_i)} > 0\right\} \bigcap \bigcap_{k\in\mathbb{X}} \left\{\alpha_k^{(Y_n)} \leq \frac{1}{2}\right\} \subset \{(3.14)\}$$

Consequently,

$$\mathbb{P}_\theta\left(\bigcup_{l\neq j\in\mathbb{X}} \bigcap_{i=1}^{n} \left\{\alpha_l^{(Y_i)} \wedge \alpha_j^{(Y_i)} > 0\right\} \bigcap \bigcap_{k\in\mathbb{X}} \left\{\alpha_k^{(Y_n)} \leq \frac{1}{2}\right\} \middle| A_n\right)$$
$$\leq \mathbb{P}_\theta\left(\mathbb{E}_\theta\left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1))) \middle| Y_{1:n}\right] > \mathbb{E}_\theta\left[\ell(\pi_n(X_{1:n}), \pi_n((1,\ldots,1,2))) \middle| Y_{1:n}\right] \middle| A_n\right)$$

To conclude, we only need to prove that the conditional probability in the lower-bound is positive. Finally,

$$\bigcup_{1\leq j\neq k\leq J} \bigcap_{i=1}^{n} \left\{\max_{l\neq j} \nu_l f_l(Y_i) < \nu_j f_j(Y_i) \leq \sum_{l\neq j} \nu_l f_l(Y_i), \nu_k f_k(Y_i) > 0\right\}$$
$$\subset \bigcup_{j=1}^{J} \bigcap_{i=1}^{n} \left\{\max_{l\neq j} \nu_l f_l(Y_i) < \nu_j f_j(Y_i) \leq \frac{1}{2}\sum_{l=1}^{J} \nu_l f_l(Y_i)\right\} \bigcap \bigcup_{1\leq l\neq j\leq J} \bigcap_{i=1}^{n} \{\nu_l f_l(Y_i) > 0, \nu_j f_j(Y_i) > 0\}$$
$$\subset \bigcup_{j=1}^{J} \bigcap_{i=1}^{n} \left\{\max_{l\neq j} \nu_l f_l(Y_i) < \nu_j f_j(Y_i)\right\} \bigcap \bigcup_{1\leq l\neq j\leq J} \bigcap_{i=1}^{n} \{\nu_l f_l(Y_i) > 0, \nu_j f_j(Y_i) > 0\} \bigcap \bigcap_{k=1}^{J} \left\{\alpha_k^{(Y_n)} < \frac{1}{2}\right\}$$
$$\subset A_n \bigcap \bigcup_{1\leq l\neq j\leq J} \bigcap_{i=1}^{n} \left\{\alpha_l^{(Y_i)} \wedge \alpha_j^{(Y_i)} > 0\right\} \bigcap \bigcap_{k=1}^{J} \left\{\alpha_k^{(Y_n)} \leq \frac{1}{2}\right\}$$

Given that the observations $Y_{1:n}$ are i.i.d. following the stationary distribution $\sum_{k=1}^{J} \nu_k f_k$ and that

$$\bigcup_{j=1}^{J}\left\{0<\max_{l\neq j}\nu_l f_l(Y)<\nu_j f_j(Y)\leq\sum_{l\neq j}\nu_l f_l(Y)\right\}$$
$$\subset\bigcup_{1\leq j\neq k\leq J}\left\{\max_{l\neq j}\nu_l f_l(Y)<\nu_j f_j(Y)\leq\sum_{l\neq j}\nu_l f_l(Y),\nu_k f_k(Y)>0\right\}$$

and that by the assumption of the theorem,

$$\mathbb{P}_\theta\left(\bigcup_{j=1}^{J}\left\{0<\max_{l\neq j}\nu_l f_l(Y)<\nu_j f_j(Y)\leq\sum_{l\neq j}\nu_l f_l(Y)\right\}\right)>0,$$

the result follows.

3.6.4 Common elements to the proof of Theorems 3.3.5, 3.3.7, and 3.3.9

Recall the definition of $\mathcal{R}_n^{\mathrm{MRSS}}$ in Remark 3.2.4: for all $(\theta,h)\in\Theta\times\mathcal{H}_n$,

$$\mathcal{R}_n^{\mathrm{MRSS}}(\theta,h)=\mathbb{E}_\theta\Big[\min_{\tau\in\mathcal{S}_J}\mathbb{E}_\theta\big[U_{n,\tau}(h)\mid Y_{1:n}\big]\Big] \tag{3.15}$$

where $U_{n,\tau}(h)\coloneqq\frac{1}{n}\sum_{i=1}^n\mathbf{1}_{\tau(X_i)\neq h_i(Y_{1:n})}$. We also make use of the notation $\hat{p}_\tau(h)\coloneqq\mathbb{E}_\theta\left[U_{n,\tau}(h)\mid Y_{1:n}\right]$. Let $\hat{\tau}_h$ denote a $Y_{1:n}$-measurable permutation satisfying:

$$\hat{p}_{\hat{\tau}_h}(h)=\mathbb{E}_\theta\left[U_{n,\hat{\tau}_h}(h)\mid Y_{1:n}\right]=\min_\tau\mathbb{E}_\theta\left[U_{n,\tau}(h)\mid Y_{1:n}\right]=\min_\tau\hat{p}_\tau(h).$$

Instead of comparing $\mathcal{R}_n^{\mathrm{clust}}$ and $\mathcal{R}_n^{\mathrm{class}}$, we compare $\mathcal{R}_n^{\mathrm{clust}}$ and $\mathcal{R}_n^{\mathrm{MRSS}}$, which is enough to obtain the result thanks to the following easy lemma.

Lemma 3.6.2. *For all $\theta\in\Theta$ and all $n\geq1$*

$$\inf_{h\in\mathcal{H}_n}\mathcal{R}_n^{\mathrm{class}}(\theta,h)=\inf_{h\in\mathcal{H}_n}\mathcal{R}_n^{\mathrm{MRSS}}(\theta,h).$$

Proof. The optimal permutation $\hat{\tau}_h$ such that $\hat{p}_{\hat{\tau}_h}(h)=\min_{\tau\in\mathcal{S}_J}\mathbb{E}_\theta\left[U_{n,\tau}(h)\mid Y_{1:n}\right]$ is a $Y_{1:n}$-measurable permutation valued random variable. Since any $h\in\mathcal{H}_n$ is also $Y_{1:n}$-measurable, the result is immediate. □

In the next we then focus on comparing $\mathcal{R}_n^{\mathrm{clust}}$ and $\mathcal{R}_n^{\mathrm{MRSS}}$. We let $N_j\coloneqq\sum_{i=1}^n\mathbf{1}_{\{X_i=j\}}$ and $N_{(1)}\leq N_{(2)}\leq\cdots\leq N_{(J)}$ denote the order statistics of $(N_1,\ldots,N_J)$.

Proposition 3.6.3. *A generic lower bound that works for any latent model (i.i.d. or HMM or whatever). For all classifiers h, all ε, all η and all $\theta\in\Theta$*

$$\begin{aligned}\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\Big]\geq{}&\mathbb{E}_\theta\Big[\min_\tau\mathbb{E}_\theta[U_{n,\tau}(h)\mid Y_{1:n}]\Big]\\&-\mathbb{E}_\theta\Big[\mathbb{E}_\theta\Big[\max_\tau(-U_{n,\tau}(h)+\hat{p}_\tau(h))\mid Y_{1:n}\Big]\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)\geq\varepsilon\}}\Big]\\&-\mathbb{E}_\theta\Big[\mathbb{P}_\theta(U_{n,\hat{\tau}_h}(h)>\eta\mid Y_{1:n})\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon\}}\Big]\\&-\mathbb{P}_\theta\big(N_{(1)}+N_{(2)}<2n\eta\big).\end{aligned}$$

Proof. For $\varepsilon \in [0,1]$, we decompose

$$\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\Big] = \mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)\geq\varepsilon\}}\Big] + \mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon\}}\Big]. \tag{3.16}$$

The first term in the rhs of (3.16) is handled via a small deviation principle, remarking that on the event $\{\hat{p}_{\hat{\tau}(h)}(h) \geq \varepsilon\}$

$$\begin{aligned}\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h) \mid Y_{1:n}\Big] &= \mathbb{E}_\theta\Big[\min_\tau(\hat{p}_\tau(h) + U_{n,\tau}(h) - \hat{p}_\tau(h)) \mid Y_{1:n}\Big] \\ &\geq \min_\tau \hat{p}_\tau(h) - \mathbb{E}_\theta\Big[\max_\tau(-U_{n,\tau}(h) + \hat{p}_\tau(h)) \mid Y_{1:n}\Big].\end{aligned}$$

The second term in (3.16) is handled via a large deviation principle. By Lemma 3.6.4, on the event $\{\hat{p}_{\hat{\tau}_h}(h) < \varepsilon\}$

$$\begin{aligned}\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h) \mid Y_{1:n}\Big] &\geq \mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\mathbf{1}_{\{U_{n,\hat{\tau}_h}(h)\leq(N_{(1)}+N_{(2)})/(2n)\}} \mid Y_{1:n}\Big] \\ &= \mathbb{E}_\theta\Big[U_{n,\hat{\tau}_h}(h)\mathbf{1}_{\{U_{n,\hat{\tau}_h}(h)\leq(N_{(1)}+N_{(2)})/(2n)\}} \mid Y_{1:n}\Big] \\ &\geq \hat{p}_{\hat{\tau}_h}(h) - \mathbb{P}_\theta\Big(U_{n,\hat{\tau}_h}(h) > \frac{N_{(1)}+N_{(2)}}{2n} \mid Y_{1:n}\Big).\end{aligned}$$

But,

$$\mathbb{P}_\theta\Big(U_{n,\hat{\tau}_h}(h) > \frac{N_{(1)}+N_{(2)}}{2n} \mid Y_{1:n}\Big) \leq \mathbb{P}_\theta\big(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\big) + \mathbb{P}_\theta\big(N_{(1)}+N_{(2)} < 2n\eta \mid Y_{1:n}\big).$$

The generic lower bound follows. □

Lemma 3.6.4. *If* $U_{n,\tau'}(h) \leq \frac{N_{(1)}+N_{(2)}}{2n}$ *then* $\min_\tau U_{n,\tau}(h) = U_{n,\tau'}(h)$.

Proof. Let $U_{n,\tau'}(h) \leq \frac{N_{(1)}+N_{(2)}}{2n}$ and suppose $\min_\tau U_{n,\tau}(h) < U_{n,\tau'}(h)$. Then, there exists a permutation $\tau'' \neq \tau'$ such that $U_{n,\tau''}(h) < U_{n,\tau'}(h)$. But then, letting $I = \{i \in \{1,\dots,n\} : \tau'(X_i) = h_i(Y_{1:n})\}$:

$$\begin{aligned}n(U_{n,\tau'}(h) - U_{n,\tau''}(h)) &= \sum_{i=1}^n \Big(\mathbf{1}_{\{\tau'(X_i)\neq h_i(Y_{1:n})\}} - \mathbf{1}_{\{\tau''(X_i)\neq h_i(Y_{1:n})\}}\Big) \\ &= \sum_{i=1}^n \Big(\mathbf{1}_{\{\tau''(X_i)= h_i(Y_{1:n})\}} - \mathbf{1}_{\{\tau'(X_i)= h_i(Y_{1:n})\}}\Big) \\ &= -\sum_{i\in I}\mathbf{1}_{\{\tau'(X_i)\neq\tau''(X_i)\}} + \sum_{i\in I^c}\mathbf{1}_{\{\tau'(X_i)=h_i(Y_{1:n})\}} \\ &= -\sum_{i=1}^n\mathbf{1}_{\{\tau'(X_i)\neq\tau''(X_i)\}} + \sum_{i\in I^c}\Big(\mathbf{1}_{\{\tau'(X_i)=h_i(Y_{1:n})\}} + \mathbf{1}_{\{\tau'(X_i)\neq\tau''(X_i)\}}\Big) \\ &\leq -(N_{(1)}+N_{(2)}) + 2|I^c| \\ &= -(N_{(1)}+N_{(2)}) + 2nU_{n,\tau'}(h)\end{aligned}$$

where we have used that since $\tau' \neq \tau''$, it must be that $\sum_{i=1}^n \mathbf{1}_{\tau'(X_i)\neq\tau''(X_i)} \geq N_{(1)}+N_{(2)}$. Rearranging the previous:

$$U_{n,\tau''}(h) \geq \frac{N_{(1)}+N_{(2)}}{n} - U_{n,\tau'}(h) \geq U_{n,\tau'}(h)$$

which contradicts that $U_{n,\tau''}(h) < U_{n,\tau'}(h)$. Hence $\min_\tau U_{n,\tau}(h) \geq U_{n,\tau'}(h)$. □

3.6.5 Proof of Theorem 3.3.5 (independent scenario)

Here we apply the result of Proposition 3.6.3 to the i.i.d. case.

When $J > 2$, the first trivial bound is obtained by choosing $\varepsilon = \eta = 0$. With this choice, Proposition 3.6.3 gives for $J \geq 2$:

$$\begin{aligned}\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\Big] &\geq \mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta[U_{n,\tau}(h) \mid Y_{1:n}]\Big] - \mathbb{E}_\theta\Big[\mathbb{E}_\theta\Big[\max_\tau(-U_{n,\tau}(h) + \hat{p}_\tau(h)) \mid Y_{1:n}\Big]\Big] \\ &\geq \mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta[U_{n,\tau}(h) \mid Y_{1:n}]\Big] - \sqrt{\frac{\log(J!)}{2n}}\end{aligned}$$

by Lemma 3.6.5 below.

When $J > 2$, Lemma 3.6.7 can be used to find that

$$\mathbb{P}_\theta\big(N_{(1)} + N_{(2)} < 2n\eta\big) \leq J^2 e^{-\frac{n(\beta-2\eta)^2}{2\beta}}$$

and the bound is obtained by choosing $\eta = \beta/4$ and $\varepsilon = \frac{\beta}{4e}[\log(J!)/(2n)]^{2/(n\beta)}$.

Lemma 3.6.5. *For all $\theta \in \Theta^{\mathrm{ind}}$, $\mathbb{P}_\theta$-almost-surely*

$$\mathbb{E}_\theta\Big[\max_\tau(-U_{n,\tau}(h) + \hat{p}_\tau(h)) \mid Y_{1:n}\Big] \leq \sqrt{\frac{\log(J!)}{2n}}.$$

Proof. For any $\lambda > 0$,

$$\begin{aligned}\mathbb{E}_\theta\left[\max_\tau\{-U_{n,\tau}(h) + \hat{p}_\tau(h)\}\right] &\leq \frac{1}{\lambda}\log\left(\mathbb{E}_\theta\left[\exp\left(\sup_\tau\{-\lambda\left(U_{n,\tau}(h) - \hat{p}_\tau(h)\right)\}\right)\middle| Y_{1:n}\right]\right) \\ &= \frac{1}{\lambda}\log\left(\mathbb{E}_\theta\left[\sup_\tau \exp\left(-\lambda\left(U_{n,\tau}(h) - \hat{p}_\tau(h)\right)\right)\middle| Y_{1:n}\right]\right) \\ &\leq \frac{1}{\lambda}\log\left(J!\sup_\tau \mathbb{E}_\theta\left[\exp\left(-\lambda\left(U_{n,\tau}(h) - \hat{p}_\tau(h)\right)\right)\middle| Y_{1:n}\right]\right) \\ &\leq \frac{1}{\lambda}\log\left(J!\exp\left(\frac{\lambda^2}{8} \times n \times \left(\frac{1}{n}\right)^2\right)\right) \text{ (Hoeffding's lemma)} \\ &\leq \inf_{\lambda>0}\left\{\frac{\log(J!)}{\lambda} + \frac{\lambda}{8n}\right\} \\ &\leq \sqrt{\frac{\log(J!)}{2n}}.\end{aligned}$$

Hoeffding's lemma applies because conditionally to the sequence of observations $Y_{1:n}$, the labels $X_{1:n}$ are still independent. □

Lemma 3.6.6. *For all $\theta \in \Theta^{\mathrm{ind}}$, $\mathbb{P}_\theta$-almost-surely*

$$\mathbb{P}_\theta\big(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\big) \leq \hat{p}_{\hat{\tau}_h}(h) \cdot \frac{e}{\eta}\Big(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta}\Big)^{n\eta-1} e^{-n\hat{p}_{\hat{\tau}_h}(h)}.$$

Proof. By Chernoff's bound (with $q_i(h) = \mathbb{P}_\theta(h_i(Y_{1:n}) \neq \hat{\tau}_h(X_i) \mid Y_{1:n})$):

$$\begin{aligned}
\mathbb{P}_\theta\big(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\big) &= \mathbb{P}_\theta\Big(\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n})\neq\hat{\tau}_h(X_i)} > n\eta \mid Y_{1:n}\Big) \\
&\leq \inf_{\lambda>0} \exp\Big(-\lambda n\eta + \sum_{i=1}^n \log\Big(q_i(h)e^\lambda + 1 - q_i(h)\Big)\Big) \\
&\leq \inf_{\lambda>0} \exp\Big(-\lambda n\eta + n\hat{p}_{\hat{\tau}_h}(h)(e^\lambda - 1)\Big) \\
&= \Big(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta}\Big)^{n\eta} e^{-n\hat{p}_{\hat{\tau}_h}(h)} \\
&\leq \hat{p}_{\hat{\tau}_h}(h) \cdot \frac{e}{\eta}\Big(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta}\Big)^{n\eta-1} e^{-n\hat{p}_{\hat{\tau}_h}(h)}.
\end{aligned}$$

□

Lemma 3.6.7. *For $\theta \in \Theta^{\mathrm{ind}}$, let $\beta = \min_{j\neq k}(\nu_j + \nu_k)$. If $J > 2$, then*

$$\mathbb{P}_\theta\big(N_{(1)} + N_{(2)} < 2n\eta\big) \leq J^2 e^{-\frac{n(\beta-2\eta)^2}{2\beta}}.$$

Proof. If $J = 2$, remark that $N_{(1)} + N_{(2)} = n$. Now we assume that $J > 2$. It holds that

$$\begin{aligned}
\mathbb{P}_\theta\big(N_{(1)} + N_{(2)} < 2n\eta\big) &= P_\theta\big(\exists\, j \neq k,\ N_j + N_k \leq 2n\eta\big) \\
&\leq J^2 \max_{j\neq k} \mathbb{P}_\theta\big(N_j + N_k \leq 2n\eta\big).
\end{aligned}$$

Then observe that $N_j + N_k = \sum_{i=1}^n (\mathbf{1}_{X_i=j} + \mathbf{1}_{X_i=k}) = \sum_{i=1}^n \mathbf{1}_{X_i\in\{j,k\}}$ whenever $j \neq k$. In other words, when $j \neq k$ the random variables $N_j + N_k$ has a Binomial distribution with parameters $(n, \nu_j + \nu_k)$ under $\mathbb{P}_\theta$. The conclusion follows using Chernoff's bound on the Binomial distribution (recall the Binomial distribution is subGaussian on the left-tail). □

3.6.6 Proof of Theorems 3.3.7 and 3.3.9(dependent scenario)

Preliminary

We first recall basic results for HMMs that can be found in Cappé et al. [2005] about the distribution of the hidden states given a set of observations. For any parameter θ, any integers k, $i \leq j$, the distribution of X_k given $Y_{i:j}$ under $\mathbb{P}_\theta$ will be denoted $\phi_{\theta,k|i:j}(., Y_{i:j})$. For any integers $i \leq n$, we shall simplify the so-called filtering distribution $\phi_{\theta,n|i:n}(., Y_{i:n})$ to $\phi_{\theta,n}(., Y_{i:n})$.

Conditional on observations $Y_{i:n}$, the sequence of the hidden states is an inhomogeneous Markov chain, with transition matrices called *forward kernels*. For each $k \leq n-1$, the forward kernel is denoted $(F_{\theta,k|n}[Y_{k+1:n}])$ to emphasize that it only depends on $Y_{k+1:n}$. When $k \geq n$, the kernel does not depend on the observations and is equal to the transition matrix Q, so that $F_{\theta,k|n}[Y_{k+1:n}] := Q$ for $k \geq n$. In other words, for any $n \in \mathrm{N}$, for any index $i \leq n$ and $k \geq i$ and any real-valued function f on $\mathbb{X}$ (understood as a vector in $\mathbb{R}^J$),

$$\mathbb{E}_\theta[f(X_{k+1}) \mid X_{i:k}, Y_{i:n}] = F_{\theta,k|n}[Y_{k+1:n}]f = \sum_{x\in\mathbb{X}} F_{\theta,k|n}[Y_{k+1:n}](X_k, x)f(x).$$

Conditional on observations $Y_{i:n}$, the reverse time sequence of hidden states is also an inhomogeneous Markov chain with transition matrices $(B_{\theta,k}[Y_{i:k}])_{k\leq n-1}$ called *backward kernels*. In other words, for any $n \in \mathrm{N}$, $i \leq k \leq n-1$ and any function f on $\mathbb{X}$:

$$\mathbb{E}_\theta[f(X_k) \mid X_{k+1:n}, Y_{i:n}] = B_{\theta,k}[Y_{i:k}]f = \sum_{x\in\mathbb{X}} B_{\theta,k}[Y_{i:k}](X_{k+1}, x)f(x).$$

Here, the backward kernel $B_{\theta,k}[Y_{i:k}]$ depends only on the observations up to time k. It is given by:

$$B_{\theta,k}[Y_{i:k}](\tilde{x}, x) = \frac{\phi_{\theta,k}(x, Y_{i:k})Q(x, \tilde{x})}{\sum_{x' \in \mathbb{X}} \phi_{\theta,k}(x', Y_{i:k})Q(x', \tilde{x})}. \tag{3.17}$$

Note that the denominator is always positive thanks to Assumption 1.

For any transition kernel T, we denote $\delta(T)$ is the Dobrushin coefficient of T defined by:

$$\delta(T) = \sup_{(x,x') \in \mathbb{X} \times \mathbb{X}} \|T(x, .) - T(x', .)\|_{\mathrm{TV}}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. We recall the following two lemmas which can be found in Cappé et al. [2005]. To end with, notice that under Assumption 1, for any subset A of $\mathbb{X}$,

$$J\delta\gamma(A) \leq \sum_{x' \in A} Q(x, x') \leq J(1 - (J-1)\delta)\gamma(A)$$

with γ the uniform distribution over $\mathbb{X}$. Using Lemma 4.3.13 in Cappé et al. [2005], this leads to the following lemma.

Lemma 3.6.8. *Under Assumption 1, for any integers k and n, the Dobrushin coefficient of the forward kernel $F_{\theta,k|n}$ satisfies:*

$$\delta(F_{\theta,k|n}) \leq \begin{cases} \rho_0 & k < n \\ \rho_1 & k \geq n \end{cases}$$

with $\rho_0 = 1 - \frac{J\delta}{J(1-(J-1)\delta)} = \frac{1-J\delta}{1-(J-1)\delta}$ and $\rho_1 = 1 - J\delta$.

Using Equation (3.17) we get that, under Assumption 1, for any (possibly non positive) integers $i \leq k$,

$$\forall x \in \mathbb{X}, \qquad B_{\theta,k}[Y_{i:k}](x, .) \geq \frac{\delta}{1 - (J-1)\delta} \phi_{\theta,k}(., Y_{i:k}),$$

so that applying Lemma 4.3.13 in Cappé et al. [2005], one gets

Lemma 3.6.9. *Under Assumption 1, for any (possibly non positive) integers $i \leq k \leq n-1$, the Dobrushin coefficient of the backward kernel $B_{\theta,k}[Y_{i:k}]$ satisfies:*

$$\delta(B_{\theta,k}[Y_{i:k}]) \leq 1 - \frac{\delta}{1 - (J-1)\delta} = \rho_0.$$

Proofs

We apply the result of Proposition 3.6.3 to the HMM case. As in the i.i.d. case, when $J = 2$ it must be that $\mathbb{P}_\theta(N_{(1)} + N_{(2)} < n) = 0$. On the one hand, using Lemma 3.6.10 and Markov's inequality as in the independent case, one gets for $\theta \in \Theta^{\mathrm{dep}}$

$$\begin{aligned} &\mathbb{E}_\theta\Big[\mathbb{E}_\theta\Big[\max_\tau(-U_{n,\tau}(h) + \hat{p}_\tau(h)) \mid Y_{1:n}\Big]\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h) \geq \varepsilon\}}\Big] \\ &\qquad \leq \frac{1}{1-\rho_0}\sqrt{\frac{\log(J!)}{2n}}\mathbb{P}_\theta\big(\hat{p}_{\hat{\tau}_h}(h) \geq \varepsilon\big) \\ &\qquad \leq \frac{1}{\varepsilon(1-\rho_0)}\sqrt{\frac{\log(J!)}{2n}}\mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta[U_{n,\tau}(h) \mid Y_{1:n}]\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h) \geq \varepsilon\}}\Big] \end{aligned}$$

and then using Lemma 3.6.11, we establish that for all $\varepsilon, \eta > 0$

$$\begin{aligned}
&\mathbb{E}_\theta\Big[\mathbb{P}_\theta(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n})\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon\}}\Big] \\
&\qquad \le \mathbb{E}_\theta\Big[\hat{p}_{\hat{\tau}_h}(h) \cdot \frac{e}{\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n}\left(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta}\right)^{n\eta-1} e^{-n\hat{p}_{\hat{\tau}_h}(h)}\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon\}}\Big] \\
&\qquad \le \frac{e}{\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n}\left(\frac{e\varepsilon}{\eta}\right)^{n\eta-1}\mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta[U_{n,\tau}(h) \mid Y_{1:n}]\mathbf{1}_{\{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon\}}\Big].
\end{aligned}$$

Then, Theorem 3.3.7 follows by taking $\eta \to \frac{1}{2}$ (by below) and $\varepsilon = \frac{1}{2e}\left(\frac{\delta}{1-\delta}\right)^4 [\log(J!)/(2n)]^{1/n}$. When $J > 2$, the first trivial bound of Theorem 3.3.9 is obtained by choosing $\varepsilon = \eta = 0$. Proposition 3.6.3 yields:

$$\begin{aligned}
\mathbb{E}_\theta\Big[\min_\tau U_{n,\tau}(h)\Big] &\ge \mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta(U_{n,\tau}(h) \mid Y_{1:n})\Big] - \mathbb{E}_\theta\Big[\mathbb{E}_\theta\Big[\max_\tau(\hat{p}_\tau(h) - U_{n,\tau}(h)) \mid Y_{1:n}\Big]\Big] \\
&\ge \mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta(U_{n,\tau}(h) \mid Y_{1:n})\Big] - \frac{1}{1-\rho_0}\sqrt{\frac{\log(J!)}{2n}}
\end{aligned}$$

by Lemma 3.6.10 below. The inequality follows by taking the infimum over h on both sides. For the remaining inequality, when $J > 2$, Lemma 3.6.12 ensures:

$$\mathbb{P}_\theta\big(N_{(1)} + N_{(2)} < 2n\eta\big) \le J^2 e^{-2n(1-\rho_1)^2(\beta-2\eta)^2}.$$

On the other hand, we have by Lemma 3.6.10:

$$\mathbb{E}_\theta\left[\max_\tau(\hat{p}_\tau(h) - U_{n,\tau}(h)) \mid Y_{1:n}\right] \le \frac{1}{1-\rho_0}\sqrt{\frac{\log(J!)}{2n}}$$

On the event $\{\hat{p}_{\hat{\tau}_h}(h) < \varepsilon\}$, for $\varepsilon < \eta$:

$$\begin{aligned}
\mathbb{P}_\theta\left(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\right) &= \mathbb{P}_\theta\left(U_{n,\hat{\tau}_h}(h) - \hat{p}_{\hat{\tau}_h}(h) > \eta - \hat{p}_{\hat{\tau}_h}(h) \mid Y_{1:n}\right) \\
&\le \mathbb{P}_\theta\left(U_{n,\hat{\tau}_h}(h) - \hat{p}_{\hat{\tau}_h}(h) > \eta - \varepsilon \mid Y_{1:n}\right) \\
&\le e^{-\lambda(\eta-\varepsilon)}\mathbb{E}_\theta\left[e^{\lambda\left\{U_{n,\hat{\tau}_h}(h) - \hat{p}_{\hat{\tau}_h}(h)\right\}} \mid Y_{1:n}\right] \\
&\le e^{-\lambda(\eta-\varepsilon)} e^{\frac{1}{8n}\left(\frac{\lambda}{1-\rho_0}\right)^2}
\end{aligned}$$

where the last inequality is due to the argument using Marton coupling as shown in the proof of Lemma 3.6.10. Taking the minimum over λ, one gets

$$\mathbb{E}_\theta\left[\mathbb{P}_\theta\left(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\right)\mathbf{1}_{\hat{p}_{\hat{\tau}_h}(h)<\varepsilon}\right] \le e^{-2n(1-\rho_0)^2(\eta-\varepsilon)^2}.$$

Using Lemma 3.6.12, the final bound reads:

$$\begin{aligned}
\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h)\right] &\ge \mathbb{E}_\theta\left[\min_\tau \mathbb{E}_\theta\left[U_{n,\tau}(h) \mid Y_{1:n}\right]\right] \\
&\quad - \frac{1}{\varepsilon(1-\rho_0)}\sqrt{\frac{\log(J!)}{2n}}\mathbb{E}_\theta\left[\hat{p}_{\hat{\tau}_h}(h)\mathbf{1}_{\hat{p}_{\hat{\tau}_h}(h)\ge\varepsilon}\right] \\
&\quad - e^{-2n(1-\rho_0)^2(\eta-\varepsilon)^2} - J^2 e^{-2n(1-\rho_1)^2(\beta-2\eta)^2}
\end{aligned}$$

Choosing $\varepsilon = \frac{\eta}{2}$ and $\eta = \frac{2}{5}\beta$ and noting that $\rho_1 < \rho_0$ one obtains:

$$\begin{aligned}
\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h)\right] \ge \left[1 - \frac{5}{\beta(1-\rho_0)}\sqrt{\frac{\log(J!)}{2n}}\right]\mathbb{E}_\theta\left[\min_\tau \mathbb{E}_\theta\left[U_{n,\tau}(h) \mid Y_{1:n}\right]\right] \\
- (J^2+1)e^{-cn(1-\rho_0)^2\beta^2}.
\end{aligned}$$

The result follows by taking the infimum over h.

Lemma 3.6.10. *Under Assumptions 1 and 2, for all $\theta \in \Theta^{\mathrm{dep}}$, $\mathbb{P}_\theta$-almost-surely*

$$\mathbb{E}_\theta\left[\max_\tau(\hat{p}_\tau(h) - U_{n,\tau}(h)) \mid Y_{1:n}\right] \leq \frac{1}{1-\rho_0}\sqrt{\frac{\log(J!)}{2n}}$$

where $\rho_0 = \frac{1-J\delta}{1-(J-1)\delta}$.

Proof. Given that for any $\lambda > 0$,

$$\mathbb{E}_\theta\left[\max_\tau\left(\hat{p}_\tau(h) - (U_{n,\tau}(h))\right] \leq \frac{1}{\lambda}\log\left(J!\max_\tau \mathbb{E}_\theta\left[\exp\left(-\lambda(U_{n,\tau}(h) - \hat{p}_\tau(h))\right)\middle| Y_{1:n}\right]\right)$$

we shall exhibit an upper bound of the rhs term by applying Theorem 2.9 of Paulin [2015], conditional on $Y_{1:n}$. So that for now we consider $Y = Y_{1:n}$ as fixed. Define $f_Y^{h,\tau}$ for any $x = x_{1:n}$ by:

$$f_Y^{h,\tau}(x) = -\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\tau(x_i)\neq h_i(Y_{1:n})}.$$

Then, for any $x = x_{1:n}$ and $x' = x'_{1:n}$,

$$f_Y^{h,\tau}(x) - f_Y^{h,\tau}(x') \leq \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i\neq x'_i}.$$

We thus may apply (2.5) in Theorem 2.9 of Paulin [2015] to get

$$\mathbb{E}_\theta\left[e^{\lambda(f_Y^{h,\tau}(X) - \mathbb{E}_\theta\left[f_Y^{h,\tau}(X)|Y_{1:n}\right])}\middle| Y_{1:n}\right] \leq e^{\frac{\lambda^2}{8n^2}\sum_{i=1}^n\left(\sum_{j=i}^n \Gamma_{i,j}\right)^2}, \tag{3.18}$$

where Γ comes from a Marton coupling (see Definition 2.1 in Paulin [2015]) and is given by:

$$\Gamma_{j,i} := 0, \qquad \Gamma_{i,j} := \sup_{x_1,\ldots,x_i,x'_i\in\mathbb{X}} \mathbb{P}_\theta\left(X_{1,j}^{(x_1,\ldots,x_i,x'_i,Y_{1:n})} \neq X_{2,j}^{(x_1,\ldots,x_i,x'_i,Y_{1:n})}\middle| Y_{1:n}\right)$$

for $1 \leq i < j \leq n$. Now,

$$\Gamma_{i,j} \leq \sup_{x_{1:i-1}\in\mathbb{X}^{i-1}}\Bigg\|\mathbb{P}_\theta\left(X_{1,j}^{(x_{1:i-1},\tilde{a},\tilde{b},Y_{1:n})} \in \cdot\middle| X_{1:i-1} = x_{1:i-1}, Y_{1:n}\right) - \mathbb{P}_\theta\left(X_{2,j}^{(x_{1:i-1},\tilde{a},\tilde{b},Y_{1:n})} \in \cdot\middle| X_{1:i-1} = x_{1:i-1}, Y_{1:n}\right)\Bigg\|_{\mathrm{TV}},$$

ie.,

$$\Gamma_{i,j} \leq \sup_{x_{1:i-1}\in\mathbb{X}^{i-1}}\Bigg\|\mathbb{P}_\theta\left(X_j \in \cdot\middle| X_{1:i-1} = x_{1:i-1}, X_i = \tilde{a}, Y_{1:n}\right) - \mathbb{P}_\theta\left(X_j \in \cdot\middle| X_{1:i-1} = x_{1:i-1}, X_i = \tilde{b}, Y_{1:n}\right)\Bigg\|_{\mathrm{TV}},$$

ie.,

$$\Gamma_{i,j} \leq \left\|\mathbb{P}_\theta\left(X_j \in \cdot\middle| X_i = \tilde{a}, Y_{1:n}\right) - \mathbb{P}_\theta\left(X_j \in \cdot\middle| X_i = \tilde{b}, Y_{1:n}\right)\right\|_{\mathrm{TV}}$$

since conditional on $Y_{1:n}$, the hidden states form a inhomogeneous Markov chain with transition kernels $(F_{k|n}[Y_{k+1:n}])$. Exponential forgetting of the smoothing distributions in HMMs (Proposition 4.3.26 in Cappé et al. [2005]) allows to conclude that

$$\left\| \mathbb{P}_\theta \left(X_j \in \cdot \middle| X_i = \tilde{a}, Y_{1:n} \right) - \mathbb{P}_\theta \left(X_j \in \cdot \middle| X_i = \tilde{b}, Y_{1:n} \right) \right\|_{\mathrm{TV}} \leq \rho_0^{j-i}$$

where $\rho_0 = \frac{1-J\delta}{1-(J-1)\delta}$ (see also Lemma 3.6.8). By inequality (3.18):

$$\max_\tau \mathbb{E}_\theta \left[e^{\lambda(f_Y^{h,\tau}(X) - \mathbb{E}_\theta\left[f_Y^{h,\tau}(X)|Y_{1:n}\right])} \middle| Y_{1:n} \right] \leq e^{\frac{1}{8n}\left(\frac{\lambda}{1-\rho_0}\right)^2}.$$

Thus,

$$\begin{aligned} \mathbb{E}_\theta \left[\max_\tau \left(\hat{p}_\tau(h) - U_{n,\tau}(h) \right) \middle| Y_{1:n} \right] &\leq \inf_{\lambda > 0} \left\{ \frac{\log(J!)}{\lambda} + \frac{\lambda}{8n(1-\rho_0)^2} \right\} \\ &= \frac{1}{1-\rho_0} \sqrt{\frac{\log(J!)}{2n}} \end{aligned}$$

□

Lemma 3.6.11. *For all $\theta \in \Theta^{\mathrm{dep}}$, $\mathbb{P}_\theta$-almost-surely*

$$\mathbb{P}_\theta\left(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\right) \leq \hat{p}_{\hat{\tau}_h}(h) \cdot \frac{e}{\eta} \left(\frac{1-(J-1)\delta}{\delta} \right)^{2n} \left(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta} \right)^{n\eta - 1} e^{-n\hat{p}_{\hat{\tau}_h}(h)}.$$

Proof. Let $S = \sum_{i=1}^n \mathbf{1}_{\hat{\tau}_h(X_i) \neq h_i(Y_{1:n})}$. We consider the following operators defined on $L^\infty(\{0,1\})$, for $i \in [n]$:

$$(M_{\theta,i}.f)(x) \coloneqq f(0) e^{\lambda \mathbf{1}_{\hat{\tau}_h(0) \neq h_i(Y_{1:n})}} B_{\theta,i}(x,0) + f(1) e^{\lambda \mathbf{1}_{\hat{\tau}_h(1) \neq h_i(Y_{1:n})}} B_{\theta,i}(x,1)$$

where $B_{\theta,i}$ is the Backward kernel defined by:

$$B_{\theta,i}(x,y) = \frac{\phi_{\theta,i}(y) Q(y,x)}{\sum_{y'} \phi_{\theta,i}(y') Q(y',x)}.$$

Then observe that,

$$\begin{aligned} \mathbb{E}_\theta[e^{\lambda S} \mid Y_{1:n}] &= \mathbb{E}_\theta \left[e^{\lambda \sum_{i=2}^n \mathbf{1}_{\hat{\tau}_h(X_i) \neq h_i(Y_{1:n})}} \mathbb{E}_\theta \left[e^{\lambda \mathbf{1}_{\hat{\tau}_h(X_1) \neq h_1(Y_{1:n})}} \mid X_{2:n}, Y_{1:n} \right] \mid Y_{1:n} \right] \\ &= \mathbb{E}_\theta \left[e^{\lambda \sum_{i=2}^n \mathbf{1}_{\hat{\tau}_h(X_i) \neq h_i(Y_{1:n})}} (M_{\theta,1}.\mathbf{1})(X_2) \mid Y_{1:n} \right] \end{aligned}$$

Repeating inductively the same trick leads to

$$\mathbb{E}_\theta[e^{\lambda S} \mid Y_{1:n}] = \mathbb{E}_\theta \left[(M_{\theta,n} ... M_{\theta,1}.\mathbf{1})(X_{n+1}) \mid Y_{1:n} \right]$$

Hence

$$\mathbb{E}_\theta[e^{\lambda S} \mid Y_{1:n}] \leq \|(M_{\theta,n} ... M_{\theta,1}.\mathbf{1})\|_\infty \leq \prod_{i=1}^n |||M_{\theta,i}|||_\infty$$

where

$$\begin{aligned} |||M_{\theta,i}|||_\infty &\coloneqq \sup_{f, \|f\|_\infty = 1} \|M_{\theta,i}.f\|_\infty \\ &= \max\left((M_{\theta,i}.f)(0), (M_{\theta,i}.f)(1) \right) \\ &\leq \max\left(\sum_{z \in \{0,1\}} e^{\lambda \mathbf{1}_{\hat{\tau}_h(z) \neq h_i(Y_{1:n})}} B_{\theta,i}(0,z), \sum_{z \in \{0,1\}} e^{\lambda \mathbf{1}_{\hat{\tau}_h(z) \neq h_i(Y_{1:n})}} B_{\theta,i}(1,z) \right). \end{aligned}$$

But, given that

$$B_{\theta,i}(x,y) = \frac{\phi_{\theta,i}(y)Q(y,x)}{\sum_{y'}\phi_{\theta,i}(y')Q(y',x)} \leq \frac{1-(J-1)\delta}{\delta}\phi_{\theta,i}(y)$$

and that

$$\phi_{\theta,i|n}(y) = \phi_{\theta,i+1|n}B_{\theta,i}(y) = \sum_x \frac{\phi_{\theta,i+1|n}(x)\phi_{\theta,i}(y)Q(y,x)}{\sum_{y'}\phi_{\theta,i}(y')Q(y',x)} \geq \frac{\delta\phi_{\theta,i}(y)}{1-(J-1)\delta}$$

one obtains

$$B_{\theta,i}(x,y) \leq \left(\frac{1-(J-1)\delta}{\delta}\right)^2 \phi_{\theta,i|n}(y).$$

Thus,

$$\begin{aligned}
|||M_{\theta,i}|||_\infty &\leq \left(\frac{1-(J-1)\delta}{\delta}\right)^2 \left(e^{\lambda \mathbf{1}_{\hat{\tau}_h(0)\neq h_i(Y_{1:n})}}\phi_{i|n}(0) + e^{\lambda \mathbf{1}_{\hat{\tau}_h(1)\neq h_i(Y_{1:n})}}\phi_{i|n}(1)\right) \\
&= \left(\frac{1-(J-1)\delta}{\delta}\right)^2 \mathbb{E}_\theta\left[e^{\lambda \mathbf{1}_{\hat{\tau}_h(X_i)\neq h_i(Y_{1:n})}} \mid Y_{1:n}\right].
\end{aligned}$$

Finally,

$$\mathbb{E}_\theta[\exp(\lambda S) \mid Y_{1:n}] \leq \left(\frac{1-(J-1)\delta}{\delta}\right)^{2n} \prod_{i=1}^n \mathbb{E}_\theta\left[e^{\lambda \mathbf{1}_{\hat{\tau}_h(X_i)\neq h_i(Y_{1:n})}} \mid Y_{1:n}\right].$$

One can then use Chernoff's bound (with $q_i(h) = \mathbb{P}_\theta(h_i(Y_{1:n}) \neq \hat{\tau}_h(X_i) \mid Y_{1:n})$):

$$\begin{aligned}
\mathbb{P}_\theta\big(U_{n,\hat{\tau}_h}(h) > \eta \mid Y_{1:n}\big) &= \mathbb{P}_\theta\Big(\sum_{i=1}^n 1_{h_i(Y_{1:n})\neq\hat{\tau}_h(X_i)} > n\eta \mid Y_{1:n}\Big) \\
&\leq \inf_{\lambda>0} e^{-\lambda n\eta}\mathbb{E}_\theta\left[e^{\lambda S} \mid Y_{1:n}\right] \\
&\leq \inf_{\lambda>0} e^{-\lambda n\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n} \prod_{i=1}^n \mathbb{E}_\theta\left[e^{\lambda \mathbf{1}_{\hat{\tau}_h(X_i)\neq h_i(Y_{1:n})}} \mid Y_{1:n}\right] \\
&\leq \inf_{\lambda>0} e^{-\lambda n\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n} e^{\sum_{i=1}^n \log\left(q_i(h)e^\lambda + 1 - q_i(h)\right)} \\
&\leq \inf_{\lambda>0} e^{-\lambda n\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n} e^{n\hat{p}_{\hat{\tau}_h}(h)(e^\lambda-1)} \\
&\leq \hat{p}_{\hat{\tau}_h}(h)\cdot\frac{e}{\eta}\left(\frac{1-(J-1)\delta}{\delta}\right)^{2n}\left(\frac{e\hat{p}_{\hat{\tau}_h}(h)}{\eta}\right)^{n\eta-1} e^{-n\hat{p}_{\hat{\tau}_h}(h)}
\end{aligned}$$

□

Lemma 3.6.12. *Let $\beta = \min_{i,j\neq k}\mathbb{P}_\theta\left(X_i \in \{j,k\}\right)$ and $\rho_1 = 1 - J\delta$. If $J \geq 3$, then for $\eta < \frac{\beta}{2}$ and $\theta \in \Theta^{\text{dep}}$*

$$\mathbb{P}_\theta\big(N_{(1)} + N_{(2)} < 2n\eta\big) \leq J^2 e^{-2n(1-\rho_1)^2(\beta-2\eta)^2}.$$

Proof. Let $\lambda > 0$, $\eta < \frac{\beta}{2}$, $j \neq k$, $\beta_i(j,k) = \mathbb{P}_\theta\left(X_i \in \{j,k\}\right)$.

$$\begin{aligned}
\mathbb{P}_\theta\left(N_j + N_k < 2n\eta\right) &= \mathbb{P}_\theta\left(\sum_{i=1}^n (\mathbf{1}_{X_i\in\{j,k\}} - \beta_i(j,k)) < 2n\eta - \sum_{i=1}^n \beta_i(j,k)\right) \\
&= \mathbb{P}_\theta\left(e^{\lambda\sum_{i=1}^n\left(\beta_i(j,k) - \mathbf{1}_{X_i\in\{j,k\}}\right)} > e^{\lambda\left(\sum_{i=1}^n \beta_i(j,k) - 2n\eta\right)}\right) \\
&\leq e^{-\lambda\left(\sum_{i=1}^n \beta_i(j,k) - 2n\eta\right)}\mathbb{E}_\theta\left[e^{\lambda\sum_{i=1}^n\left(\beta_i(j,k) - \mathbf{1}_{X_i\in\{j,k\}}\right)}\right]
\end{aligned}$$

$\mathbb{E}_\theta\left[e^{\lambda\sum_{i=1}^n\left(\beta_i(j,k)-\mathbf{1}_{X_i\in\{j,k\}}\right)}\right]$ can be controlled by the same technique using the Marton coupling that was used in the proof of Lemma 3.6.10:

$$\mathbb{E}_\theta\left[e^{\lambda\sum_{i=1}^n\left(\beta_i(j,k)-\mathbf{1}_{X_i\in\{j,k\}}\right)}\right]\leq e^{\frac{(n\lambda)^2}{8n(1-\rho_1)^2}}=e^{\frac{n\lambda^2}{8(1-\rho_1)^2}}$$

It follows that

$$\begin{aligned}\mathbb{P}_\theta\left(N_j+N_k<2n\eta\right)&\leq\inf_{\lambda>0}e^{-\lambda(\sum_{i=1}^n\beta_i(j,k)-2n\eta)+\frac{n\lambda^2}{8(1-\rho_1)^2}}\\&\leq e^{-2n(1-\rho_1)^2\left(\frac{\sum_{i=1}^n\beta_i(j,k)}{n}-2\eta\right)^2}\\&\leq e^{-2n(1-\rho_1)^2(\beta-2\eta)^2}\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}_\theta\left(N_{(1)}+N_{(2)}<2n\eta\right)&\leq J^2\max_{j\neq k}\mathbb{P}_\theta\left(N_j+N_k<2n\eta\right)\\&\leq J^2e^{-2n(1-\rho_1)^2(\beta-2\eta)^2}.\end{aligned}$$

□

3.6.7 Proof of Theorem 3.3.6

Let $a\in\{1,2\}$ and $n\in\mathbb{N}$. Let $Q=\begin{pmatrix}1-p & p\\ q & 1-q\end{pmatrix}$ be the transition matrix and assume $0<q<\frac{1}{2}<p$ and $p+q<1$. Assume also that the initial distribution is the stationary distribution, that is $\nu=\left(\frac{q}{p+q},\frac{p}{p+q}\right)$.

$$\begin{aligned}\mathbb{P}_\theta\left(X_1=a\mid Y_{1:n}\right)&=\sum_{x_{2:n}\in\{1,2\}}\mathbb{P}_\theta\left(X_1=a,X_{2:n}=x_{2:n}\mid Y_{1:n}\right)\\&\propto\sum_{x_{2:n}\in\{1,2\}}\nu(a)Q_{a,x_2}...Q_{x_{n-1},x_n}f_a(Y_1)..f_{x_n}(Y_n)\\\mathbb{P}_\theta\left(X_2=a\mid Y_{1:n}\right)&=\sum_{x_1,x_{3:n}\in\{1,2\}}\mathbb{P}_\theta\left(X_1=x_1,X_2=a,X_{3:n}=x_{3:n}\mid Y_{1:n}\right)\\&\propto\sum_{x_1,x_{3:n}\in\{1,2\}}\nu(x_1)Q_{x_1,a}...Q_{x_{n-1},x_n}f_{x_1}(Y_1)f_a(Y_2)...f_{x_n}(Y_n)\end{aligned}$$

The Bayes classifier puts the two first observations in the same cluster when:

$$\begin{aligned}&\left[\mathbb{P}_\theta\left(X_1=2\mid Y_{1:n}\right)-\mathbb{P}_\theta\left(X_1=1\mid Y_{1:n}\right)\right]\times\left[\mathbb{P}_\theta\left(X_2=2\mid Y_{1:n}\right)-\mathbb{P}_\theta\left(X_2=1\mid Y_{1:n}\right)\right]\geq0\\&\iff\left[\sum_{x_{2:n}\in\{1,2\}}\left(\nu(2)Q_{2,x_2}f_2(Y_1)-\nu(1)Q_{1,x_2}f_1(Y_1)\right)Q_{x_2,x_3}...Q_{x_{n-1},x_n}f_{x_2}(Y_2)...f_{x_n}(Y_n)\right]\\&\times\left[\sum_{x_1,x_{3:n}\in\{1,2\}}\left(Q_{x_1,2}Q_{2,x_3}f_2(Y_2)-Q_{x_1,1}Q_{1,x_3}f_1(Y_2)\right)\nu(x_1)Q_{x_3,x_4}...Q_{x_{n-1},x_n}f_{x_1}(Y_1)f_{x_3}(Y_3)...f_{x_n}(Y_n)\right]\geq0\end{aligned}$$

A sufficient condition for this to be ensured is:

$$\left(\frac{f_1}{f_2}(Y_1)<\frac{p\min(q,1-q)}{q\max(p,1-p)}\text{ and }\frac{f_1}{f_2}(Y_2)<\frac{\min(p,1-q)\min(q,1-q)}{\max(q,1-p)\max(p,1-p)}\right)$$

or

$$\left(\frac{f_2}{f_1}(Y_1)<\frac{q\min(p,1-p)}{p\max(q,1-q)}\text{ and }\frac{f_2}{f_1}(Y_2)<\frac{\min(q,1-p)\min(p,1-p)}{\max(p,1-q)\max(q,1-q)}\right)$$

Since $q < \frac{1}{2} < p$ and $p + q < 1$, the condition simplifies to:

$$\left(\frac{f_1}{f_2}(Y_1) < 1 \text{ and } \frac{f_1}{f_2}(Y_2) < \frac{q}{1-p}\right) \text{ or } \left(\frac{f_2}{f_1}(Y_1) < \frac{q(1-p)}{p(1-q)} \text{ and } \frac{f_2}{f_1}(Y_2) < \frac{q(1-p)}{(1-q)^2}\right).$$

We consider the event:

$$A = \left\{\frac{f_1}{f_2}(Y_1) < 1, \frac{f_1}{f_2}(Y_2) < \frac{q}{1-p}\right\} \bigcup \left\{\frac{f_2}{f_1}(Y_1) < \frac{q(1-p)}{p(1-q)}, \frac{f_2}{f_1}(Y_2) < \frac{q(1-p)}{(1-q)^2}\right\}$$

In what follows, we seek a sufficient condition under which the Bayes clusterer puts the two first observations in two different clusters. The Bayes clusterer is a partition $F_n^\star$ that minimizes $\mathbb{E}_\theta\left[\ell\left(\pi_n\left(X_{1:n}\right), F_n\right) \mid Y_{1:n}\right]$. Let $L(Y_{1:n})$ be the likelihood of the observations $Y_{1:n}$. Consider the event

$$B_n = \{(\forall i \in [\![3, n]\!]) \quad Y_i \notin \text{Supp}\left(f_2\right)\}$$

Assume B_n has positive probability. Since the hidden Markov chain is mixing, this happens for example when f_1 and f_2 do not have the same support. On this event,

$$\begin{aligned}
&\mathbb{E}_\theta\left[\ell\left(\pi_n(X_{1:n}), F_n\right) \mid Y_{1:n}\right] = \sum_{x_{1:n}\in\{1,2\}^n} \ell\left(\pi_n(x_{1:n}), F_n\right) \mathbb{P}_\theta\left(X_{1:n} = x_{1:n} \mid Y_{1:n}\right) \\
&= \sum_{x_{1:2}\in\{1,2\}^2} \ell\left(\pi_n((x_1, x_2, 1, \ldots, 1), F_n\right) \mathbb{P}_\theta\left(X_{1:2} = x_{1:2}, X_{3:n} = 1 \mid Y_{1:n}\right) \\
&= \frac{1}{L(Y_{1:n})} \sum_{x_{1:2}\in\{1,2\}^2} \ell\left(\pi_n((x_1, x_2, 1, \ldots, 1)), F_n\right) \nu(x_1) Q_{x_1,x_2} Q_{x_2,1} Q_{1,1}^{n-3} f_{x_1}(Y_1) f_{x_2}(Y_2) \prod_{i=3}^{n} f_1(Y_i) \\
&\propto \Big(\ell\left(\pi_n((1, \ldots, 1)), F_n\right) \nu(1) Q_{1,1}^2 f_1(Y_1) f_1(Y_2) + \ell\left(\pi_n((2, 2, 1, \ldots, 1), F_n\right) \nu(2) Q_{2,2} Q_{2,1} f_2(Y_1) f_2(Y_2) \\
&+ \ell\left(\pi_n((1, 2, 1, \ldots, 1)), F_n\right) \nu(1) Q_{1,2} Q_{2,1} f_1(Y_1) f_2(Y_2) + \ell\left(\pi_n((2, 1, \ldots, 1)), F_n\right) \nu(2) Q_{2,1} Q_{1,1} f_2(Y_1) f_1(Y_2)\Big) \\
&\propto \Big(\ell\left(\pi_n((1, \ldots, 1)), F_n\right) q(1-p)^2 f_1(Y_1) f_1(Y_2) + \ell\left(\pi_n((2, 2, 1, \ldots, 1), F_n\right) pq(1-q) f_2(Y_1) f_2(Y_2) \\
&+ \ell\left(\pi_n((1, 2, 1, \ldots, 1)), F_n\right) pq^2 f_1(Y_1) f_2(Y_2) + \ell\left(\pi_n((2, 1, \ldots, 1)), F_n\right) pq(1-p) f_2(Y_1) f_1(Y_2)\Big)
\end{aligned}$$

For $n \geq 5$, $F_n^\star$ is necessarily of the form $F_n^\star = \pi_n((y_1^\star, y_2^\star, 1, .., 1))$ with $y_1^\star$ and $y_2^\star$ in $\{1, 2\}$. Thus,

$$F_n^\star \in \arg\min \mathbb{E}_\theta\left[\ell\left(\pi_n(X_{1:n}), F_n\right) \mid Y_{1:n}\right] \iff (y_1^\star, y_2^\star) \in \arg\min H(y_1, y_2)$$

where

$$\begin{aligned}
H(y_1, y_2) &= (y_1 + y_2 - 2)(1-p)^2 f_1(Y_1) f_1(Y_2) + (4 - y_1 - y_2) p(1-q) f_2(Y_1) f_2(Y_2) \\
&\quad + (1 + y_1 - y_2) pq f_1(Y_1) f_2(Y_2) + (1 - y_1 + y_2) p(1-p) f_2(Y_1) f_1(Y_2) \\
&= y_1\left[(1-p)^2 f_1(Y_1) f_1(Y_2) + pq f_1(Y_1) f_2(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2) - p(1-p) f_2(Y_1) f_1(Y_2)\right] \\
&\quad + y_2\left[(1-p)^2 f_1(Y_1) f_1(Y_2) - pq f_1(Y_1) f_2(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2) + p(1-p) f_2(Y_1) f_1(Y_2)\right] \\
&\quad - 2(1-p)^2 f_1(Y_1) f_1(Y_2) + 4p(1-q) f_2(Y_1) f_2(Y_2) + pq f_1(Y_1) f_2(Y_2) + p(1-p) f_2(Y_1) f_1(Y_2)
\end{aligned}$$

$$\begin{aligned}
y_1^\star \neq y_2^\star &\iff \left[(1-p)^2 f_1(Y_1) f_1(Y_2) + pq f_1(Y_1) f_2(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2) - p(1-p) f_2(Y_1) f_1(Y_2)\right] \\
&\quad \times \left[(1-p)^2 f_1(Y_1) f_1(Y_2) - pq f_1(Y_1) f_2(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2) + p(1-p) f_2(Y_1) f_1(Y_2)\right] < 0 \\
&\iff \left|(1-p)^2 f_1(Y_1) f_1(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2)\right| < p\left|q f_1(Y_1) f_2(Y_2) - (1-p) f_2(Y_1) f_1(Y_2)\right|
\end{aligned}$$

Finally, consider the event:

$$C_n = \left\{ \left| (1-p)^2 f_1(Y_1) f_1(Y_2) - p(1-q) f_2(Y_1) f_2(Y_2) \right| < p \left| q f_1(Y_1) f_2(Y_2) - (1-p) f_2(Y_1) f_1(Y_2) \right| \right\}$$

We finally have:

$$A \cap B_n \cap C_n \subset \left\{ g_\theta^\star(Y_{1:n}) \neq \pi_n \circ h_\theta^\star(Y_{1:n}) \right\}.$$

By choosing appropriately the parameters p and q, one can ensure that $\mathbb{P}_\theta\left(A \cap B_n \cap C_n\right)$ is positive for many emission densities not having the same support. For example, one can ensure that for $p = 0.58$, $q = 0.35$, $f_1(Y_1) = 5$, $f_2(Y_1) = 2.5$, $f_1(Y_2) = 1.8$, $f_2(Y_2) = 1.2$, both inequalities involved in the definition of event C_n and A are ensured. Choosing smooth densities having not exactly the same support, the event $A \cap B_n \cap C_n$ can be ensured to have positive probability. The counterexamples for $n \in \{3, 4\}$ can be proved as for the case $n = 2$ presented in Section 3.3.2.

3.6.8 Proof of Theorem 3.3.8

Let $n \in \mathrm{N}$ and $\theta^\star = \left(\nu^\star, Q^\star, (f_x^\star)_{x \in \mathbb{X}}\right) \in \Theta^{\text{ind}}$ such that the assumption of Theorem 3.3.4 is ensured. It follows that $\mathbb{P}_{\theta^\star}\left(g_{\theta^\star}^\star(Y_{1:n}) \neq \pi_n \circ h_{\theta^\star}^\star(Y_{1:n})\right) > 0$. We also assume the emission densities $(f_x^\star)_{x \in \mathbb{X}}$ to be uniformly continuous. We denote $\tilde{\Theta} \subset \Theta$ the subset of parameters of the form $\theta = (\nu, Q, f_1^\star, \ldots, f_J^\star)$. Consider the two functions

$$\begin{aligned}
H \colon \mathcal{H}_n \times \Theta \times \mathcal{Y}^n &\longrightarrow [0,1] \\
(h, \theta, y_{1:n}) &\longmapsto \mathbb{E}_\theta\left[\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n}) \neq X_i} \middle| Y_{1:n} = y_{1:n} \right] \\
G \colon \mathcal{G}_n \times \Theta \times \mathcal{Y}^n &\longrightarrow [0,1] \\
(g, \theta, y_{1:n}) &\longmapsto \mathbb{E}_\theta\left[\ell\left(\pi_n\left(X_{1:n}\right), g(Y_{1:n})\right) \middle| Y_{1:n} = y_{1:n} \right]
\end{aligned}$$

By uniform continuity of $(f_x^\star)_{x \in \mathbb{X}}$, it follows that for all $g \in \mathcal{G}_n$ and $h \in \mathcal{H}_n$, $H(h, ., .)$ and $G(g, ., .)$ are uniformly continuous. Since $\mathcal{H}_n$ and $\mathcal{G}_n$ are finite, there exists $\mathcal{V}(\theta^\star) \subset \tilde{\Theta}$ a neighborhood of $\theta^\star$ and A_n an open subset of $\mathcal{Y}^n$ such that $\mathbb{P}_{\theta^\star}\left(A_n\right) > 0$ and $(\forall \theta \in \mathcal{V}(\theta^\star))\,(\forall y_{1:n} \in A_n)$

$$\arg\min_h H(h, \theta, y_{1:n}) = \arg\min_h H(h, \theta^\star, y_{1:n}) \text{ and } \arg\min_g G(g, \theta, y_{1:n}) = \arg\min_g G(g, \theta^\star, y_{1:n})$$

Or equivalently $g_{\theta^\star}^\star(y_{1:n}) = g_\theta^\star(y_{1:n})$ and $h_{\theta^\star}^\star(y_{1:n}) = h_\theta^\star(y_{1:n})$. On the other hand, using exactly the same arguments as those of Theorem 3.3.4, one could also have chosen $\theta^\star \in \Theta^{\text{ind}}$ such that not only the probability $\mathbb{P}_{\theta^\star}\left(g_{\theta^\star}^\star(Y_{1:n}) \neq \pi_n \circ h_{\theta^\star}^\star(Y_{1:n})\right) > 0$ but also $\mathbb{P}_{\theta^\star}\left(\left\{g_{\theta^\star}^\star(Y_{1:n}) \neq \pi_n \circ h_{\theta^\star}^\star(Y_{1:n})\right\} \cap A_n\right) > 0$. It follows that

$$(\forall \theta \in \mathcal{V}(\theta^\star))\, \mathbb{P}_\theta\left(\left\{g_\theta^\star(Y_{1:n}) \neq \pi_n \circ h_\theta^\star(Y_{1:n})\right\} \cap A_n\right) = \mathbb{P}_\theta\left(\left\{g_{\theta^\star}^\star(Y_{1:n}) \neq \pi_n \circ h_{\theta^\star}^\star(Y_{1:n})\right\} \cap A_n\right)$$

which is positive when θ approches $\theta^\star$ by continuity of the map

$$\theta \mapsto \mathbb{P}_\theta\left(\left\{g_{\theta^\star}^\star(Y_{1:n}) \neq \pi_n \circ h_{\theta^\star}^\star(Y_{1:n})\right\} \cap A_n\right).$$

The result follows.

3.6.9 Proof of Proposition 3.3.3

The result is straightforward when $n = 1$. We assume in what follows $n \geq 2$. We prove the proposition by showing that when the probability of having small clusters is high, the two risks are not necessarily equivalent; and $\inf_{g \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g)$ may be much smaller than $\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$.

Consider $J = 3$ (similar examples can be constructed for any $J \geq 3$) with $\nu_1 = 1 - 2\eta$, $\nu_2 = \nu_3 = \eta$. We take $F_1 = U(0, 1/2)$, $F_2 = U(3/4, 1)$ and $F_3 = U(3/4 - \varepsilon, 1 - \varepsilon)$ for some $0 < \varepsilon < 1/4$ where $U(a, b)$ is the uniform distribution on the interval (a, b). In this case,

$$\begin{aligned}
\mathbb{P}_\theta(X_i \in \cdot \mid Y_i \in (0, 1/2)) &= \delta_1(\cdot) \\
\mathbb{P}_\theta(X_i \in \cdot \mid Y_i \in (3/4 - \varepsilon, 3/4)) &= \delta_3(\cdot) \\
\mathbb{P}_\theta(X_i \in \cdot \mid Y_i \in (1 - \varepsilon, 1)) &= \delta_2(\cdot) \\
\mathbb{P}_\theta(X_i \in \cdot \mid Y_i \in (3/4, 1 - \varepsilon)) &= \frac{1}{2}\delta_2(\cdot) + \frac{1}{2}\delta_3(\cdot).
\end{aligned}$$

So in this case

$$\begin{aligned}
\mathbb{E}_\theta\left[U_{n,\tau}(h) \mid Y_{1:n}\right] &= \mathbb{E}_\theta\Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n}) \neq \tau(X_i)} \mid Y_{1:n}\Big] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{P}_\theta(h_i(Y_{1:n}) \neq \tau(X_i) \mid Y_{1:n}).
\end{aligned}$$

A Bayes "classifier" minimizing $h \mapsto \mathcal{R}_n^{\text{class}}(\theta, h)$ in this case is given by $h_\theta^\star = \left(h_{\theta,i}^\star\right)_{i \in [n]}$ where

$$h_{\theta,i}^\star = \begin{cases} 1 & \text{if } Y_i \in (0, 1/2) \\ 2 & \text{if } Y_i \in (1 - \varepsilon, 1) \\ 3 & \text{if } Y_i \in (3/4 - \varepsilon, 1 - \varepsilon). \end{cases}$$

With this choice, the optimal permutation is identity and

$$\begin{aligned}
\min_\tau \mathbb{E}_\theta\left[U_{n,\tau}(h_\theta^\star) \mid Y_{1:n}\right] = &\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \in (0,1/2)} \mathbb{P}_\theta(1 \neq X_i \mid Y_i) \\
&+ \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \in (3/4-\varepsilon, 3/4)} \mathbb{P}_\theta(3 \neq X_i \mid Y_i) \\
&+ \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \in (1-\varepsilon, 1)} \mathbb{P}_\theta(2 \neq X_i \mid Y_i) \\
&+ \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i \in (3/4, 1-\varepsilon)} \mathbb{P}_\theta(3 \neq X_i \mid Y_i),
\end{aligned}$$

ie.,

$$\min_\tau \mathbb{E}_\theta\left[U_{n,\tau}(h_\theta^\star) \mid Y_{1:n}\right] = \frac{1}{2n}\sum_{i=1}^n \mathbf{1}_{Y_i \in (3/4, 1-\varepsilon)}.$$

Thus,

$$\begin{aligned}
\mathbb{E}_\theta\Big[\min_\tau \mathbb{E}_\theta(U_{n,\tau}(h_\theta^\star) \mid Y_{1:n})\Big] &= \frac{1}{2}\mathbb{P}_\theta(Y_1 \in (3/4, 1 - \varepsilon)) \\
&= \frac{1}{2}\Big(\eta \cdot (1 - 4\varepsilon) + \eta \cdot (1 - 4\varepsilon)\Big) \\
&= \eta(1 - 4\varepsilon).
\end{aligned}$$

We now investigate $\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\right]$, for the previous Bayes classifier $h^\star_\theta$ (which does not necessarily minimize $h \mapsto \mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h)\right]$). We rewrite,

$$\begin{aligned}\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\right] &= \mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)}=0}\right] \\ &\quad + \sum_{m=1}^n \mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)}=m}\right]\end{aligned}$$

Let first consider $\mathbb{E}_\theta[\min_\tau U_{n,\tau}(h^\star_\theta) \mid Y_{1:n}]$ on the event that $\{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = 0\}$. Let define $N = \sum_{i=1}^n \mathbf{1}_{X_i\neq 1}$:

- if $N = 0$ this means that all the Y_i are in $(0, 1/2)$ and our classifier $h^\star_\theta$ combined with the identity permutation will make zero error, *i.e.* $\min_\tau U_{n,\tau}(h^\star_\theta) = 0$ on this event;
- if $N = 1$ this means that there is only one $j \in \{1, \dots, n\}$ such that $Y_j \in (3/4, 1-\varepsilon)$ [by assumption it can not be in $(3/4-\varepsilon, 3/4)$ or $(1-\varepsilon, 1)$]. Our classifier will predict $h^\star_{\theta,i} = 1$ for all $i \neq j$ and $h^\star_{\theta,j} = 3$. Now, necessarily $X_i = 1$ for $i \neq j$. If $X_j = 3$ then $h^\star_\theta \circ \mathrm{Id}$ will have loss zero, and if $X_j = 2$ then $h^\star_\theta \circ \begin{pmatrix}1 & 2 & 3\\ 1 & 3 & 2\end{pmatrix}$ will have loss zero. So in the event that $\{N = 1\}$ we also have that $\min_\tau U_{n,\tau}(h^\star_\theta) = 0$.
- If $N \geq 2$ our classifier will still classify perfectly all the $Y_i \in (0, 1/2)$ so the loss can not exceed $\min_\tau U_{n,\tau}(h^\star_\theta) \leq N/n$ in this case.

So on the event $\{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = 0\}$:

$$\begin{aligned}\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta) \mid Y_{1:n}\right] &\leq \mathbb{E}_\theta\Big[\frac{N}{n}\mathbf{1}_{N\geq 2} \mid Y_{1:n}\Big] \\ &\leq \sum_{k=2}^n \frac{k}{n}\mathbb{P}_\theta\Big(N = k \mid Y_{1:n}\Big)\end{aligned}$$

But under the law of $X_{1:n} \mid Y_{1:n}$ in the considered event, we have that N is almost-surely equal to the number of $Y_i \in (3/4, 1-\varepsilon)$, so

$$\begin{aligned}\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta) \mid Y_{1:n}\right] &\leq \sum_{k=2}^n \frac{k}{n}\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}=k} \\ &= \Big(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}\Big)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}\geq 2}.\end{aligned}$$

Deduce that,

$$\begin{aligned}&\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)}=0}\right] \\ &\quad\leq \mathbb{E}_\theta\left[\Big(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}\Big)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}\geq 2} \mid \sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = 0\right]\end{aligned}$$

Conditional on $\{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = 0\}$, the random variable $\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4,1-\varepsilon)}$ has a Binomial$(n, 2\eta)$ distribution. Then,

$$\begin{aligned}\mathbb{E}_\theta\left[\min_\tau U_{n,\tau}(h^\star_\theta)\mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i\in(3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)}=0}\right] &\leq \frac{1}{n}\sum_{k=2}^n \binom{n}{k} k\cdot(2\eta)^k(1-2\eta)^{n-k} \\ &= \frac{2\eta(1-(1-2\eta)^n-2\eta)}{1-2\eta} \\ &\asymp 4(n-1)\eta^2\end{aligned}$$

when $\eta \ll 1/n$.

Next (remark that this can be largely improved, but this is indeed for our purpose),

$$\begin{aligned}
&\sum_{m=1}^{n} \mathbb{E}_\theta \left[\min_\tau U_{n,\tau}(h_\theta^\star) \mathbf{1}_{\sum_{i=1}^n \mathbf{1}_{Y_i \in (3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = m} \right] \\
&\qquad \leq \mathbb{P}_\theta \left(\sum_{i=1}^{n} \mathbf{1}_{Y_i \in (3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} \geq 1 \right) \\
&\qquad = 1 - \mathbb{P}_\theta \left(\sum_{i=1}^{n} \mathbf{1}_{Y_i \in (3/4-\varepsilon,3/4)\cup(1-\varepsilon,1)} = 0 \right) \\
&\qquad = 1 - \mathbb{P}_\theta \Big(\forall i,\ Y_i \notin (3/4-\varepsilon, 3/4) \cup (1-\varepsilon, 1) \Big) \\
&\qquad = 1 - \Big((1-2\eta) + 2\eta(1-4\varepsilon) \Big)^n \\
&\qquad = 1 - (1 - 8\eta\varepsilon)^n \\
&\qquad \asymp 8n\eta\varepsilon
\end{aligned}$$

when $\eta \ll n$. So by choosing $\varepsilon \asymp \eta$, we have shown that whenever $\eta \ll 1/n$

$$\inf_h \mathbb{E}_\theta \left[\min_\tau U_{n,\tau}(h) \right] \leq \mathbb{E}_\theta \left[\min_\tau U_{n,\tau}(h_\theta^\star) \right] \lesssim n\eta^2$$

but

$$\inf_h \mathbb{E}_\theta \left[\min_\tau \mathbb{E}_\theta \left[U_{n,\tau}(h) \mid Y_{1:n} \right] \right] = \mathbb{E}_\theta \left[\min_\tau \mathbb{E}_\theta \left[U_{n,\tau}(h_\theta^\star) \mid Y_{1:n} \right] \right] \sim \eta$$

so that

$$\frac{\inf_h \mathbb{E}_\theta \left[\min_\tau U_{n,\tau}(h) \right]}{\inf_h \mathbb{E}_\theta \left[\min_\tau \mathbb{E}_\theta \left[U_{n,\tau}(h) \mid Y_{1:n} \right] \right]} \lesssim n\eta$$

which goes to zero as $\eta \to 0$.

3.6.10 Proof of Theorem 3.3.10

Simple computations lead to the expression of the Bayes risk of classification:

$$\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \mathbb{P}_\theta \left(X_i \neq x_0 \mid Y_{1:n} \right) \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \left(\sum_{x \neq x_0} \phi_{\theta,i|n}(x) \right) \right].$$

3.6.11 Bounds for the independent scenario

First, for $\theta \in \Theta^{\text{ind}}$,

$$\begin{aligned}
\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) &= \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \mathbb{P}_\theta \left(X_1 \neq x_0 \mid Y_1 \right) \right] \\
&= \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \frac{\sum_{x \neq x_0} \nu_x f_x(Y_1)}{\sum_x \nu_x f_x(Y_1)} \right] \\
&= \int \min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \nu_x f_x(y) d\mathcal{L}(y)
\end{aligned}$$

so that using Assumption 1,

$$\delta \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left[\sum_{x \neq x_0} f_x(y) \right] d\mathcal{L}(y) \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h)$$
$$\leq (1 - (J-1)\delta) \int_{\mathbb{Y}} \min_{x_0 \in \mathbb{X}} \left[\sum_{x \neq x_0} f_x(y) \right] d\mathcal{L}(y). \quad (3.19)$$

3.6.12 Bounds for the dependent scenario

Let $\theta \in \Theta^{\text{dep}}$. We first have

$$\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i|1:n}(x) \right] \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right].$$

Now, for any $i \leq n-1$, using the Backward recursions (see Proposition 3.3.9 in Cappé et al. [2005]),

$$\mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i|1:n}(x) \right] = \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \sum_{x' \in \mathbb{X}} \phi_{\theta,i+1|1:n}(x') B_{\theta,i}[Y_{1:i}](x', x) \right].$$

Now, using Assumption 1 and Equation (3.17), we get

$$B_{\theta,i}[Y_{1:i}](x', x) = \frac{\phi_{\theta,i}(x) Q(x, x')}{\sum_{\tilde{x} \in \mathbb{X}} \phi_{\theta,i}(\tilde{x}) Q(\tilde{x}, x')} \geq \frac{\phi_{\theta,i}(x) Q(x, x')}{1 - (J-1)\delta}$$

so that

$$\begin{aligned}
\mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i|1:n}(x) \right] &\geq \frac{\delta}{1 - (J-1)\delta} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x'} \phi_{\theta,i+1}(x') \sum_{x \neq x_0} \phi_{\theta,i}(x) \right] \\
&\geq \frac{\delta}{1 - (J-1)\delta} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right] \\
&= \frac{\delta}{1 - (J-1)\delta} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right].
\end{aligned}$$

Now, for $i = n$, the same inequality obviously holds, so that we get

$$\frac{\delta}{1 - (J-1)\delta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right] \leq \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right].$$

It suffices then to exhibit upper and lower bounds on $\sum_{i=1}^{n} \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \sum_{x \neq x_0} \phi_{\theta,i}(x) \right]$. Using the Forward recursions (see Equation (3.22) in Proposition 3.2.5 of Cappé et al. [2005]), for any $i \geq 2$,

$$\mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \left(\sum_{x \neq x_0} \phi_{\theta,i}(x) \right) \right] = \mathbb{E}_\theta \left[\min_{x_0 \in \mathbb{X}} \left(\frac{\sum_{x \neq x_0} \sum_{x' \in \mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(Y_i)}{\sum_{x \in \mathbb{X}} \sum_{x' \in \mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(Y_i)} \right) \right].$$

Let $A \in \mathcal{Y}$. One has:

$$\begin{aligned}
\mathbb{P}_\theta\left(Y_i \in A \mid Y_{1:i-1}\right) &= \sum_{x\in\mathbb{X}} \mathbb{P}_\theta\left(Y_i \in A, X_i = x \mid Y_{1:i-1}\right) \\
&= \sum_{x\in\mathbb{X}} \mathbb{P}_\theta\left(Y_i \in A \mid X_i = x, Y_{1:i-1}\right) \mathbb{P}_\theta\left(X_i = x \mid Y_{1:i-1}\right) \\
&= \sum_{x\in\mathbb{X}} \mathbb{P}_\theta\left(X_i = x \mid Y_{1:i-1}\right) \int_A f_x(y) d\mathcal{L}(y) \\
&= \sum_{x\in\mathbb{X}} \sum_{x'\in\mathbb{X}} \mathbb{P}_\theta\left(X_{i-1} = x', X_i = x \mid Y_{1:i-1}\right) \int_A f_x(y) d\mathcal{L}(y) \\
&= \sum_{x\in\mathbb{X}} \sum_{x'\in\mathbb{X}} \mathbb{P}_\theta\left(X_i = x \mid X_{i-1} = x', Y_{1:i-1}\right) \phi_{\theta,i-1}(x') \int_A f_x(y) d\mathcal{L}(y) \\
&= \int_A \left(\sum_{x\in\mathbb{X}} \sum_{x'\in\mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(y) \right) d\mathcal{L}(y),
\end{aligned}$$

so that, conditionally on $Y_{1:i-1}$, Y_i has density $\sum_{x\in\mathbb{X}} \sum_{x'\in\mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x$ with respect to the dominating measure $\mathcal{L}$. We thus get :

$$\begin{aligned}
\mathbb{E}_\theta\left[\min_{x_0\in\mathbb{X}} \left(\sum_{x\neq x_0} \phi_{\theta,i}(x) \right) \right] &= \mathbb{E}_\theta\left[\mathbb{E}_\theta\left[\min_{x_0\in\mathbb{X}} \left(\frac{\sum_{x\neq x_0} \sum_{x'\in\mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(Y_i)}{\sum_{x\in\mathbb{X}} \sum_{x'\in\mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(Y_i)} \right) \middle| Y_{1:i-1} \right] \right] \\
&= \mathbb{E}_\theta\left[\int_{\mathbb{Y}} \min_{x_0\in\mathbb{X}} \left(\sum_{x\neq x_0} \sum_{x'\in\mathbb{X}} Q_{x',x} \phi_{\theta,i-1}(x') f_x(y) \right) d\mathcal{L}(y) \right].
\end{aligned}$$

Then, under Assumption 1, for any $i \geq 2$,

$$\begin{aligned}
\delta \int_{\mathbb{Y}} \min_{x_0\in\mathbb{X}} \left[\sum_{x\neq x_0} f_x(y) \right] d\mathcal{L}(y) &\leq \mathbb{E}_\theta\left[\min_{x_0\in\mathbb{X}} \left(\sum_{x\neq x_0} \phi_{\theta,i}(x) \right) \right] \\
&\leq (1-(J-1)\delta) \int_{\mathbb{Y}} \min_{x_0\in\mathbb{X}} \left[\sum_{x\neq x_0} f_x(y) \right] d\mathcal{L}(y), \quad (3.20)
\end{aligned}$$

The result follows.

3.6.13 Proof of Theorem 3.3.11

We first control the excess risk of classification by the errors made in the estimation of the parameters.

Proposition 3.6.13. *For all $\theta \in \Theta^{\mathrm{dep}}$ satisfying Assumptions 1, 2 and 4 and for all $n \geq 1$,*

$$\begin{aligned}
\mathcal{R}_n^{\mathrm{class}}(\theta, \hat{h}) - \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h) &\leq C\mathbb{E}_\theta\left[\frac{1}{n} \|\nu - \hat{\nu}\|_2 + \left(1 + \frac{\delta}{\hat{\delta}}\right) \|Q - \hat{Q}\|_{\mathrm{F}} \right] \\
&\quad + \frac{\delta C \sqrt{C^\star}}{n} \sum_{i=1}^n \sum_{l=1}^n \mathbb{E}_\theta\left[(\hat{\rho} \vee \rho)^{2|l-i|} \|f_x - \hat{f}_x\|_\infty^2 \right]^{1/2},
\end{aligned}$$

where $C = \frac{8(1-\delta)}{\delta^3}$, $\rho = \frac{1-2\delta}{1-\delta}$, $\hat{\rho} = \frac{1-2\hat{\delta}}{1-\hat{\delta}}$ and $\hat{\delta} = \min_{x,x'} \hat{Q}_{x,x'}$.

Proof. Recall that $\phi_{\theta,i|n}(\cdot) = \mathbb{P}_\theta(X_i \in \cdot \mid Y_{1:n})$. Then,

$$
\begin{aligned}
&\mathcal{R}_n^{\mathrm{class}}(\theta, \hat{h}) - \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h) \\
&\quad = \mathbb{E}_\theta\Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \neq \hat{h}_i}\Big] - \mathbb{E}_\theta\Big[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{X_i \neq h^\star_{\theta,i}}\Big] \\
&\quad = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta\Big[\mathbb{P}_\theta(X_i = h^\star_{\theta,i} \mid Y_{1:n}) - \mathbb{P}_\theta(X_i = \hat{h}_i \mid Y_{1:n})\Big] \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta\Big[\mathbb{P}_\theta(X_i = h^\star_{\theta,i} \mid Y_{1:n}) - \mathbb{P}_{\hat{\theta}}(X_i = \hat{h}_i \mid Y_{1:n}) + \|\phi_{\hat{\theta},i|n} - \phi_{\theta,i|n}\|_{\mathrm{TV}}\Big] \\
&\quad \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_\theta\Big[\mathbb{P}_\theta(X_i = h^\star_{\theta,i} \mid Y_{1:n}) - \mathbb{P}_{\hat{\theta}}(X_i = h^\star_{\theta,i} \mid Y_{1:n}) + \|\phi_{\hat{\theta},i|n} - \phi_{\theta,i|n}\|_{\mathrm{TV}}\Big] \\
&\quad \leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}_\theta\Big[\|\phi_{\hat{\theta},i|n} - \phi_{\theta,i|n}\|_{\mathrm{TV}}\Big]
\end{aligned}
$$

where the penultimate line follows because by definition $\hat{h}_i$ maximizes $x \mapsto \mathbb{P}_{\hat{\theta}}(X_i = x \mid Y_{1:n})$. Then, under Assumption 4 and by application of Proposition 2.2 of De Castro et al. [2017] (see also Equation ((3.10))), one has

$$
\begin{aligned}
&\mathcal{R}_n^{\mathrm{class}}(\theta, \hat{h}) - \inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta, h) \\
&\leq \frac{8(1-\delta)}{\delta^2}\mathbb{E}_\theta\left[\frac{1}{n\delta}\|\nu - \hat{\nu}\|_2 + \left(\frac{1}{\delta} + \frac{1}{\hat{\delta}}\right)\|Q - \hat{Q}\|_{\mathrm{F}} + \frac{1}{n}\sum_{i=1}^n\sum_{l=1}^n \delta\frac{(\hat{\rho}\vee\rho)^{|l-i|}}{c^\star(Y_l)}\max_{x\in\mathbb{X}}\|f_x - \hat{f}_x\|_\infty\right] \\
&\leq C\mathbb{E}_\theta\left[\frac{1}{n}\|\nu - \hat{\nu}\|_2 + \left(1 + \frac{\delta}{\hat{\delta}}\right)\|Q - \hat{Q}\|_{\mathrm{F}} + \frac{\delta^2}{n}\sum_{i=1}^n\sum_{l=1}^n \frac{(\hat{\rho}\vee\rho)^{|l-i|}}{c^\star(Y_l)}\max_{x\in\mathbb{X}}\|f_x - \hat{f}_x\|_\infty\right] \\
&\leq C\mathbb{E}_\theta\left[\frac{1}{n}\|\nu - \hat{\nu}\|_2 + \left(1 + \frac{\delta}{\hat{\delta}}\right)\|Q - \hat{Q}\|_{\mathrm{F}}\right] + \frac{\delta^2 C}{n}\sum_{i=1}^n\sum_{l=1}^n \mathbb{E}_\theta\left[\frac{(\hat{\rho}\vee\rho)^{|l-i|}}{c^\star(Y_l)}\max_{x\in\mathbb{X}}\|f_x - \hat{f}_x\|_\infty\right] \\
&\leq C\mathbb{E}_\theta\left[\frac{1}{n}\|\nu - \hat{\nu}\|_2 + \left(1 + \frac{\delta}{\hat{\delta}}\right)\|Q - \hat{Q}\|_{\mathrm{F}}\right] + \frac{\delta^2 C}{n}\sum_{i=1}^n\sum_{l=1}^n \mathbb{E}_\theta\left[\frac{1}{c^\star(Y_l)^2}\right]^{1/2}\mathbb{E}_\theta\left[(\hat{\rho}\vee\rho)^{2|l-i|}\|f_x - \hat{f}_x\|_\infty^2\right]^{1/2} \\
&\leq C\mathbb{E}_\theta\left[\frac{1}{n}\|\nu - \hat{\nu}\|_2 + \left(1 + \frac{\delta}{\hat{\delta}}\right)\|Q - \hat{Q}\|_{\mathrm{F}}\right] + \frac{\delta C\sqrt{C^\star}}{n}\sum_{i=1}^n\sum_{l=1}^n \mathbb{E}_\theta\left[(\hat{\rho}\vee\rho)^{2|l-i|}\|f_x - \hat{f}_x\|_\infty^2\right]^{1/2}
\end{aligned}
$$

□

In order to obtain a rate on the excess risk, we make use of Algorithm 5 which will yield the estimates used in the statement of the corollary. This algorithm merges the spectral algorithms of De Castro et al. [2017], Abraham et al. [2022] with some slight modifications. Note that all the expectations and probabilities of this proof are with respect to the observations and the random unit matrices. Also note that the algorithm outputs estimates of the densities that are not necessarily *bona-fide* densities. This is not problematic for the plug-in procedure as one typically uses the Forward-Backward algorithm Cappé et al. [2005] which works even if the emissions are not correctly normalized.

First, we start by controlling $\mathbb{E}_\theta\left[\|Q - \hat{Q}\|_{\mathrm{F}}^2\right]$, $\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right]$ and $\mathbb{E}_\theta\left[\max_{x\in\mathbb{X}}\|f_x - \hat{f}_x\|_\infty^2\right]$ using the estimates yielded by the algorithm. Thanks to step 7 of the algorithm, one can

Algorithm 5: Non-parametric spectral estimation of the transition matrix and the emission laws

Input

- Number of states J, integers D and r.
- Data $(Y_i)_{i\leq n+2}$ drawn from a HMM with J states.
- Functions $(\varphi_d)_{d\in\mathbb{N}}$ **uniformly bounded** such that $O = (\mathbb{E}_\theta[\varphi_d(Y_1) \mid X_1 = j])_{1\leq d\leq D, 1\leq j\leq J}$ is of rank J with $\sigma_J(O)$ **bounded away from 0 uniformly** in D, at least for D large enough.
- K a Lipschitz-continuous kernel

Output

- Spectral estimators $\hat{Q}$ and $(\hat{f}_j)_{1\leq j\leq J}$

Estimation

[Step 1] For all $a, b, c \in [\![1, D]\!]$, consider the following empirical estimators:

$$\hat{L}(a) = \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)$$

$$\hat{N}(a,b) = \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})$$

$$\hat{P}(a,c) = \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_c(Y_{s+2})$$

$$\hat{M}(a,b,c) = \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)\varphi_b(Y_{s+1})\varphi_c(Y_{s+2})$$

$$\hat{M}^{x,L}(a,b) = \frac{1}{n}\sum_{s=1}^{n} \varphi_a(Y_s)K_L(x, Y_{s+1})\varphi_b(Y_{s+2})$$

[Step 2] Let $\hat{V}$ be the $D \times J$ matrix of orthonormal right singular vectors of $\hat{P}$ corresponding to its top J singular values.

[Step 3] For all $d \in [\![1, D]\!]$, set $\hat{B}(d) = (\hat{V}^\top \hat{P}\hat{V})^{-1}\hat{V}^\top \hat{M}(., d, .)\hat{V}$

[Step 4] Generate Ω a $J \times J$ unit matrix uniformly drawn, for all $x \in [\![1, J]\!]$, $\hat{C}(x) = \sum_{d=1}^{D}(\hat{V}\Omega)(d, x)\hat{B}(d)$

[Step 5] Compute $\hat{R}_1$ a $J \times J$ unit Euclidean norm columns matrix that diagonalizes the matrix $\hat{C}(1)$:

$$\hat{R}_1^{-1}\hat{C}(1)\hat{R}_1 = \mathcal{D}iag[(\hat{\Lambda}(1,1), ..., \hat{\Lambda}(1,J))]$$

[Step 6] For all $x, x' \in [\![1, J]\!]$, $\hat{\Lambda}(x, x') = (\hat{R}_1^{-1}\hat{C}(x)\hat{R}_1)_{x',x'}$.

[Step 7] Repeat steps 4 to 6 r times and take Ω_r maximizing $i \mapsto \min_{k\leq J} \min_{k_1\neq k_2} |\hat{\Lambda}_i(k, k_1) - \hat{\Lambda}_i(k, k_2)|$

[Step 8] Set $\hat{O} = \hat{V}\Omega_r\hat{\Lambda}$, $\tilde{\nu} = (\hat{V}^\top\hat{O})^{-1}\hat{V}^\top\hat{L}$ and $\hat{Q} = \Pi_{TM}\left((\hat{V}^\top\hat{O}\mathcal{D}iag[\tilde{\nu}])^{-1}\hat{V}^\top\hat{N}\hat{V}(\hat{O}^\top\hat{V})^{-1}\right)$ where Π_{TM} denotes the projection (with respect to the scalar product given by the Frobenius norm) onto the convex set of transition matrices.

[Step 9] For $x \in \mathbb{R}$, set $\hat{B}^x = \hat{B}^{x,D,L} = (\hat{V}^\top\hat{P}\hat{V})^{-1}\hat{V}^\top\hat{M}^{x,L}\hat{V}$

[Step 10] Set $\hat{R}_2 = \hat{Q}\hat{O}^\top\hat{V}$ and take $\tilde{f}_j(x) = (\hat{R}_2\hat{B}^x\hat{R}_2^{-1})_{j,j}$

[Step 11] $\hat{f}_j(x) = \begin{cases} \tilde{f}_j(x) & \text{si } |\tilde{f}_j(x)| \leq n^\beta \\ n^\beta sign(\tilde{f}_j(x)) & \text{otherwise} \end{cases}$ for $\beta > 0$ fixed (but arbitrary).

obtain a slightly different version of Theorem 3.1 of De Castro et al. [2017] (Note that this version is used in the proof of Corollary 3.2 in De Castro et al. [2017]). It ensures the existence of positive constants C, x_0, y_0, D_0 and n_1 such that for all $D \geq D_0$ there exist a permutation $\tau_D \in \mathcal{S}_J$ such that for all $n \geq n_1 \eta_3^2(\varphi_D) x(y + \log(r)) e^{y/r}$, $x \geq x_0$, $y \geq y_0$ and $r \geq 1$, with probability at least $1 - 4e^{-x} - 2e^{-y}$:

$$\begin{aligned} \|Q^{\tau_D} - \hat{Q}\|_{\mathrm{F}}^2 &\leq C \eta_3^2(\varphi_D) x(y + \log(r)) e^{y/r}/n \\ \max_{x \in \mathbb{X}} \|f_{D,\tau_D(x)} - \hat{f}_{D,x}^{(r)}\|_2^2 &\leq C \eta_3^2(\varphi_D) x(y + \log(r)) e^{y/r}/n \end{aligned} \tag{3.21}$$

where:

- $(\varphi_k)_{k \in \mathbb{N}}$ is the basis used in Algorithm 5
- $\eta_3^2(\varphi_D) = \sup_{y,y' \in \mathcal{Y}^3} \sum_{a,b,c=1}^{D} \left(\varphi_a(y_1)\varphi_b(y_2)\varphi_c(y_3) - \varphi_a(y_1')\varphi_b(y_2')\varphi_c(y_3')\right)^2$
- $f_{D,x} = \sum_{d=1}^{D} \langle f_x, \varphi_d \rangle \varphi_d$ the projection of the density f_x on the subspace spanned by the first D components of the basis.
- $\hat{f}_{D,x}^{(r)} = \sum_{d=1}^{D} \hat{O}_{d,x} \varphi_d$ where $\hat{O}$ is the matrix constructed at step 8 of Algorithm 5.

As detailed in De Castro et al. [2017], when using a wavelet basis or trigonometric polynomials basis, $\eta_3(\varphi_D)$ ensures for a constant $C_\eta > 0$:

$$\eta_3(\varphi_D) \leq C_\eta D^{3/2} \text{ and } \max_{l \in \mathbb{N}} \|\varphi_l\|_\infty < \infty \tag{3.22}$$

We assume a similar basis is used.

It is important to note that the estimator $\hat{f}_{D,x}^{(r)}$ is not the one yielded by the Algorithm 5 but it is rather the one used in De Castro et al. [2017]. We do not use it for the estimation because it does not allow obtaining the appropriate rate in infinite norm (see De Castro et al. [2017] for more details). However, we will use it in our proof because $\|\hat{O}(., k) - O(., \tau_{D_n}(k))\|_2 = \|\hat{f}_{D_n,k}^{(r)} - f_{D_n,\tau_{D_n}(k)}\|_2$.

Assume that the parameters $x = x_n$, $y = y_n$, $D = D_n$ and $r = r_n$ are increasing with respect to n and that $n \geq n_1 C_\eta^2 D_n^3 x_n (y_n + \log(r_n)) e^{y_n/r_n}$.

Control of $\mathbb{E}_\theta \left[\|Q^{\tau_{D_n}} - \hat{Q}\|_{\mathrm{F}}^2 \right]$ The control of $\|Q^{\tau_{D_n}} - \hat{Q}\|_{\mathrm{F}}^2$ in expectation is already proved in Corollary 3.2 of De Castro et al. [2017] using Inequality (3.21). The proof chooses $r_n \propto \log(n)$ and $\eta_3(\varphi_{D_n}) = o(\sqrt{n}/\log(n))$ and yields:

$$\mathbb{E}_\theta \left[\|Q^{\tau_{D_n}} - \hat{Q}\|_{\mathrm{F}}^2 \right] = \mathcal{O}(\eta_3^2(\varphi_{D_n}) \log(n)/n) = \mathcal{O}(D_n^3 \log(n)/n) \quad \text{(by (3.22))}$$

for a sequence $(\tau_{D_n})_n$ of permutations. We will keep the same values of r_n and D_n in what follows.

One of the advantages of this algorithm with respect to the previous versions is that it allows obtaining the appropriate rate on the errors in the estimation of all the model parameters. This is done thanks to the use of the kernel estimator of the emission densities for which the error of approximation is tuned (through the parameter L) independently of the error of estimation of the transition matrix. The shortcoming of the algorithm proposed in De Castro et al. [2017] is that it does not allow controlling the rate on the emission densities without altering that of the transition matrix as it is clear in Corollary 3.3 of that paper.

Control of $\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right]$ Let $\tilde{\delta} = \frac{\delta}{2}$. Then,

$$\begin{aligned}\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right] &= \mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\mathbf{1}_{\hat{\delta}<\tilde{\delta}}\right] + \mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\mathbf{1}_{\hat{\delta}\geq\tilde{\delta}}\right] \\ &\leq \mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\mathbf{1}_{\hat{\delta}<\tilde{\delta}}\right] + \frac{1}{\tilde{\delta}^2}\end{aligned}$$

On the other hand:

$$\tilde{\delta} - \hat{\delta} = \delta - \hat{\delta} - \frac{\delta}{2} \leq \left|\delta - \hat{\delta}\right| - \frac{\delta}{2} \leq \max_{i,j}\left|Q_{i,j}^{\tau_{D_n}} - \hat{Q}_{i,j}\right| - \frac{\delta}{2} \leq \|Q^{\tau_{D_n}} - \hat{Q}\|_F - \frac{\delta}{2}$$

where we have used the inequality $|\min_{i,j} Q_{i,j} - \min_{i,j}\hat{Q}_{i,j}| \leq \max_{i,j}|Q_{i,j} - \hat{Q}_{i,j}|$.

We assume that all the entries of $\hat{Q}$ are between $n^{-\alpha/2}$ and $1 - n^{-\alpha/2}$ for $\alpha \geq 2$. If it is not the case, modifying the entries of $\hat{Q}$ to obtain a similar property induces an error of order $n^{-\alpha/2}$ which is negligible with respect to the rates we seek and all the subsequent results remain unchanged. It follows that $\hat{\delta} \geq n^{-\alpha/2}$ and:

$$\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right] \leq n^\alpha \mathbb{P}_\theta\left(\|Q^{\tau_{D_n}} - \hat{Q}\|_F > \frac{\delta}{2}\right) + \frac{1}{\tilde{\delta}^2}$$

Choosing for example $x_n = y_n = r_n = \alpha\log(n)$, then one obtains for n large enough:

$$\begin{aligned}\mathbb{P}_\theta\left(\|Q^{\tau_{D_n}} - \hat{Q}\|_F > \frac{\delta}{2}\right) &\leq \mathbb{P}_\theta\left(\|Q^{\tau_{D_n}} - \hat{Q}\|_F > C\eta_3^2(\varphi_{D_n})x_n(y_n + \log(r_n))e/n\right) \\ &\leq 4e^{-x_n} + 2e^{-y_n}\end{aligned}$$

It follows that $\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right]$ is upper-bounded by an absolute constant.

The values x_n, y_n and r_n will be kept the same in what follows.

Control of $\mathbb{E}_\theta\left[\max_{x\in\mathbb{X}}\|f_{\tau_{D_n}(x)} - \hat{f}_x\|_\infty^2\right]$ The difficulty of the control of this quantity lies in ensuring that the same permutation τ_{D_n} used for the control of Q still works for the control of the emission densities. In the spectral algorithm of Abraham et al. [2022], the matrix $\hat{R}$ is chosen independently of Q or $\hat{Q}$ (In fact, this algorithm does not even estimate $\hat{Q}$). Had we used this matrix, there would not be any reason for which the same permutation τ_{D_n} works for the control of the emission densities. To solve this problem, we choose a matrix $\hat{R}$ that depends explicitly on $\hat{Q}$ so that the same permutation that works for the control of Q works also for that of the emission densities. We follow here the steps of the proof of Theorem 5 in Abraham et al. [2022].

Let $M^{x,L}, P, O$ be the quantities estimated by $\hat{M}^{x,L}, \hat{P}, \hat{O}$ and construct $\hat{f}_j$, $\tilde{f}_j$ using Algorithm 5. Let E_n be the event with probability greater than $1 - 4e^{-x_n} - 2e^{-y_n}$ on which the control of (3.21) holds when $x_n = y_n = r_n = \alpha\log(n)$ and $D = D_n$. For $\gamma > 0$ there exists $c = c(\gamma)$ such that the event:

$$\mathcal{A}_n = \left\{\|\hat{P} - P\| \leq cD_n\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}, \quad \sup_{x\in\mathbb{R}}\|\hat{M}^{x,L} - M^{x,L}\| \leq cD_n^2\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right\}$$

is measurable and has probability $n^{-\gamma}$ (Cf. Lemma 25.a of Abraham et al. [2022]). Given that E_n has probability greater than $1 - 4e^{-x_n} - 2e^{-y_n} = 1 - 6n^{-\alpha}$, it follows that $\mathcal{A}_n \cap E_n$ has probability greater than $1 - n^{-\gamma} - 6n^{-\alpha}$. Note that the difference with the original proof is that we use the event $\mathcal{A}_n \cap \mathcal{E}_n$ instead of the event $\mathcal{A}_n$. This is compulsory to control the errors of $\hat{Q}$ and $\hat{f}$ simultaneously.

On the event $\mathcal{A}_n \cap E_n$, and at step 10 of Algorithm 5, instead of using the matrix $\hat{R}$ appearing in the spectral algorithm in Abraham et al. [2022], we rather use the matrix $\hat{R}_2 = \hat{Q}\hat{O}^\top \hat{V}$ where the components $\hat{V}$, $\hat{Q}$ and $\hat{O}$ are constructed in the Algorithm 5. On the other hand, since the columns of $\hat{R}$ are not normalized, we choose $\tilde{R}_2 = QO^\top \hat{V}$ on the contrary to what is done in the proof of Theorem 5 in Abraham et al. [2022]. By denoting $Q^{\tau_{D_n}} = P_{\tau_{D_n}} Q P_{\tau_{D_n}}^{-1}$, one obtains:

$$\hat{R}_2 - P_{\tau_{D_n}} \tilde{R}_2 = \hat{Q}\hat{O}^\top \hat{V} - P_{\tau_{D_n}} Q O^\top \hat{V} = \left(\hat{Q}(\hat{O} - O P_{\tau_{D_n}}^\top)^\top + (\hat{Q} - Q^{\tau_{D_n}}) P_{\tau_{D_n}} O^\top \right) \hat{V}$$

It follows by using operator norm:

$$\|\hat{R}_2 - P_{\tau_{D_n}} \tilde{R}_2\| \leq \left(\|\hat{Q}\| \|\hat{O} - O P_{\tau_{D_n}}^\top\| + \|P_{\tau_{D_n}} O^\top\| \|\hat{Q} - Q^{\tau_{D_n}}\| \right) \|\hat{V}\| \tag{3.23}$$

First, note that on the event $\mathcal{A}_n \cap E_n$:

$$\begin{aligned}
\|\hat{Q}\| &\leq 1 \\
\|\hat{V}\| &\leq \|\hat{V}\|_F = J^{1/2} \quad \text{(columns are normalized)} \\
\|\hat{Q} - Q^{\tau_{D_n}}\| &\leq \|\hat{Q} - Q^{\tau_{D_n}}\|_{\mathrm{F}} \\
&\leq \left(C C_\eta^2 D_n^3 x_n (y_n + \log(r_n)) e^{y_n / r_n} / n \right)^{1/2} \quad \text{(by (3.21) and (3.22))} \\
\|P_{\tau_{D_n}} O^\top\| = \|O\| &= \sup_{\|v\|=1} \left(\sum_{j=1}^{J} \left(\sum_{l=1}^{D_n} v_l \langle f_j, \varphi_l \rangle \right)^2 \right)^{1/2} \\
&\leq (J D_n)^{1/2} \max_{1 \leq l \leq D_n} \|\varphi_l\|_\infty \\
\|\hat{O} - O P_{\tau_{D_n}}^\top\| \leq \|\hat{O} - O P_{\tau_{D_n}}^\top\|_F &= \left(\sum_{k=1}^{K} \|\hat{O}(., k) - O(., \tau_{D_n}(k))\|_2^2 \right)^{1/2} \\
&= \left(\sum_{k=1}^{K} \|\hat{f}_{D_n,k}^{(r)} - f_{D_n, \tau_{D_n}(k)}\|_2^2 \right)^{1/2} \\
&\leq \left(K C C_\eta^2 D_n^3 x_n (y_n + \log(r_n)) e^{y_n / r_n} / n \right)^{1/2} \quad \text{(by (3.21) and (3.22))}.
\end{aligned}$$

By keeping the previous choices of x_n, y_n and r_n then by Inequality (3.23), there exists a constant C' such that:

$$\mathbf{1}_{\mathcal{A}_n \cap E_n} \|\hat{R}_2 - P_{\tau_{D_n}} \tilde{R}_2\| \leq C' D_n^{1/2} D_n^{3/2} \frac{\log(n)}{\sqrt{n}}.$$

By Lemma 25.b of Abraham et al. [2022], for n large enough, $\hat{P}$ has rank J, $(\hat{V}^T \hat{P} \hat{V})$ and $(\hat{V}^T P \hat{V})$ are invertible and the matrices $(\hat{B}(d))_{1 \leq d \leq D_n}$ appearing in Algorithm 5 are then well-defined. By Lemma 11 and 25.b of Abraham et al. [2022], $\tilde{B}^x = (\hat{V}^T P \hat{V})^{-1} \hat{V}^T M^x \hat{V}$ satisfies:

$$\tilde{B}^x = (QO^T \hat{V})^{-1} D^x (QO^T \hat{V}) = \tilde{R}_2^{-1} D^x \tilde{R}_2 \tag{3.24}$$

where $D^x = (K_L[f_j](x))_{j \leq J}$. Using the fact that $\sigma_J(\hat{V}) = 1$ (the columns of $\hat{V}$ are orthonormal) and $\sigma_J(QO^T) \geq \sigma_J(Q)\sigma_J(O) > 0$ (because Q in full rank and $\sigma_J(O)$ is bounded from below by an absolute constant by assumption of the algorithm), it follows that $\|\tilde{R}_2^{-1}\|^{-1}$ is bounded from below by an absolute constant because:

$$\begin{aligned}
\|\tilde{R}_2^{-1}\|^{-1} &= \frac{1}{\sigma_1(\tilde{R}_2^{-1})} = \sigma_J(\tilde{R}_2) \\
&= \sigma_J(QO^T \hat{V}) \geq \sigma_J(QO^T)\sigma_J(\hat{V}) = \sigma_J(QO^T) \geq \sigma_J(Q)\sigma_J(O) > 0.
\end{aligned} \tag{3.25}$$

Thus, for n large enough, the assumption of Lemma 3.6.14 (stated below) is verified with $A_t = \tilde{B}^t$, $\hat{A}_t = \hat{B}^t$, $R = \tilde{R}_2 = QO^T\hat{V}$. This ensures that:

$$\mathbf{1}_{\mathcal{A}_n \cap E_n} \max_{x \in \mathbb{X}} \|\tilde{f}_x - K_L[f_{\tau_{D_n}(x)}]\|_\infty$$
$$\leq 4\kappa(\tilde{R}_2) \left[\mathbf{1}_{\mathcal{A}_n \cap E_n} \sup_t \|\tilde{B}^t - \hat{B}^t\| + \lambda_{\max}\kappa(\tilde{R}_2)\|\tilde{R}_2^{-1}\|\mathbf{1}_{\mathcal{A}_n \cap E_n}\|\hat{R}_2 - P_{\tau_{D_n}}\tilde{R}_2\| \right]$$

where $\lambda_{\max} = \sup_t \max_j |\lambda_{t,j}| = \sup_t \max_j |K_L[f_j]| < \infty$ and $\lambda_{t,j}$ is the j-th eigenvalue of $\tilde{B}^t$. By Lemma 35.c of Abraham et al. [2022], one has for some constant C:

$$\mathbf{1}_{\mathcal{A}_n \cap E_n} \sup_t \|\tilde{B}^t - \hat{B}^t\| \leq C D_n^2 \left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}}.$$

By Lemma 35.b in Abraham et al. [2022], one obtains: $\kappa(\tilde{R}_2) \leq \tilde{C} D_n^{1/2}$ and $\|\tilde{R}_2^{-1}\|$ upper-bounded by an absolute constant by (3.25). We finally obtain that:

$$\mathbf{1}_{\mathcal{A}_n \cap E_n} \max_{x \in \mathbb{X}} \|\tilde{f}_x - K_L[f_{\tau_{D_n}(x)}]\|_\infty \leq c' D_n^{1/2} \left[D_n^2 \left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}} + D_n^{1/2} D_n^{1/2} D_n^{3/2} \frac{\log(n)}{\sqrt{n}} \right]$$
$$\leq c'' D_n^{5/2} \left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}}$$

for come constants c', c''. The choice of L (cf. Algorithm 5) allows obtaining then:

$$\mathbf{1}_{\mathcal{A}_n \cap E_n} \max_{x \in \mathbb{X}} \|\tilde{f}_x - f_{\tau_{D_n}(x)}\|_\infty \leq c'' D_n^{5/2} \left(\frac{\log(n)}{n} \right)^{\frac{s}{2s+1}}.$$

Given that for n large enough $\|f_j\|_\infty \leq n^\beta$, it follows that: $\|\hat{f}_j - f_{\tau_{D_n}(j)}\|_\infty \leq \|\tilde{f}_j^L - f_{\tau_{D_n}(j)}\|_\infty$.

Finally, thanks to the truncation of the emission densities, it is possible to obtain the same rate in expectation:

$$\mathbb{E}_\theta \left[\|\hat{f}_j - f_{\tau_{D_n}(j)}\|_\infty^2 \right] \leq c'' D_n^{5/2} \left(\frac{\log(n)}{n} \right)^{\frac{2s}{2s+1}} + 2n^\beta \mathbb{P}((\mathcal{A}_n \cap E_n)^c)$$
$$\leq c'' D_n^{5/2} \left(\frac{\log(n)}{n} \right)^{\frac{2s}{2s+1}} + 2n^\beta (n^{-\gamma} + 6n^{-\alpha}).$$

By choosing α and γ sufficiently large, one obtains:

$$\mathbb{E}_\theta \left[\|\hat{f}_j - f_{\tau_{D_n}(j)}\|_\infty^2 \right] = \mathcal{O}\left(D_n^{5/2} \left(\frac{\log(n)}{n} \right)^{\frac{2s}{2s+1}} \right).$$

Control of $\frac{1}{n}\sum_{i=1}^n \sum_{l=1}^n \mathbb{E}_\theta \left[(\rho \vee \hat{\rho})^{2|l-i|} \max_{x \in \mathbb{X}} \|f_x - \hat{f}_x\|_\infty^2 \right]^{1/2}$ Let $\tilde{\rho} = \frac{1-2\tilde{\delta}}{1-\tilde{\delta}}$ where $\tilde{\delta} = \frac{\delta}{2}$.

$$\mathbb{E}_\theta \left[(\rho \vee \hat{\rho})^{2|l-i|} \max_{x \in \mathbb{X}} \|f_x - \hat{f}_x\|_\infty^2 \right]^{1/2} \leq \tilde{\rho}^{|l-i|} \mathbb{E}_\theta \left[\max_{x \in \mathbb{X}} \|f_x - \hat{f}_x\|_\infty^2 \right]^{1/2}$$
$$+ \mathbb{E}_\theta \left[\max_{x \in \mathbb{X}} \|f_x - \hat{f}_x\|_\infty^2 \mathbf{1}_{\hat{\rho} \geq \tilde{\rho}} \right]^{1/2}.$$

The term $\mathbb{E}_\theta\left[\max_{x\in\mathbb{X}}\|f_x-\hat{f}_x\|_\infty^2\mathbf{1}_{\hat{\rho}\geq\tilde{\rho}}\right]^{1/2}$ can be made of order $n^{-\alpha}$ for arbitrary $\alpha>0$ by the same large deviation argument used in the control of $\mathbb{E}_\theta\left[\frac{1}{\hat{\delta}^2}\right]$. Then, summing up over i and n yields:

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n}\mathbb{E}_\theta\left[(\rho\vee\hat{\rho})^{2|l-i|}\max_{x\in\mathbb{X}}\|f_x-\hat{f}_x\|_\infty^2\right]^{1/2}=\mathcal{O}\left(D_n^{5/2}\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right)$$

Finally, using Proposition 3.6.13 and the previous controls of the errors of estimation of the model parameters, one gets:

$$\mathbb{E}[\mathcal{R}_n^{\text{class}}(\theta,\hat{h}^{\tau_n})]-\inf_{h\in\mathcal{H}_n}\mathcal{R}_n^{\text{class}}(\theta,h)=\mathcal{O}\left(D_n^{5/2}\left(\frac{\log(n)}{n}\right)^{\frac{s}{2s+1}}\right).$$

A similar rate holds for the excess risk of clustering thanks to the relationship between the Bayes risk of classification and the Bayes risk of clustering established in Theorems 3.3.7 and 3.3.9.

Lemma 3.6.14. *Suppose $(A_t, t\in\mathbb{R})$ are $J\times J$ matrices simultaneously diagonalized by a matrix R:*

$$RA_tR^{-1}=diag(\lambda_{t,1},...,\lambda_{t,J}), t\in\mathbb{R}.$$

Let $\hat{R}$ be a matrix such that for some permutation τ of $\{1,...,J\}$, we have:

$$\|\hat{R}-P_\tau R\|=\varepsilon_R\leq\frac{\|R^{-1}\|^{-1}}{2}$$

Assume $\lambda_{\max}=\sup_t\max_j|\lambda_{t,j}|<\infty$. For matrices $(\hat{A}_t)_{t\in\mathbb{R}}$, write $\varepsilon_A=\sup_t\|A_t-\hat{A}_t\|$ and define

$$\hat{\lambda}_{t,j}=e_j^T\hat{R}\hat{A}_t\hat{R}^{-1}e_j\quad\text{and}\quad\lambda_{t,\tau(j)}=e_j^T(P_\tau R)A_t(P_\tau R)^{-1}e_j.$$

Then

$$\sup_t\max_j|\hat{\lambda}_{t,j}-\lambda_{t,\tau(j)}|\leq 4\kappa(R)\left[\varepsilon_A+\lambda_{\max}\kappa(R)\|R^{-1}\|\varepsilon_R\right].$$

Proof. Let $\hat{\zeta}_j^T=e_j^T\hat{R}$, let $\hat{\xi}_j=\hat{R}^{-1}e_j$ and define $\zeta_j^T=e_j^TP_\tau R$ and $\xi_j=(P_\tau R)^{-1}e_j$. Then, $\lambda_{t,\tau(j)}=\zeta_j^TA_t\xi_j$, $\hat{\lambda}_{t,j}=\hat{\zeta}_j^T\hat{A}_t\hat{\xi}_j$ and we have:

$$\begin{aligned}
|\hat{\lambda}_{t,j}-\lambda_{t,\tau(j)}|&=|\hat{\zeta}_j^T\hat{A}_t\hat{\xi}_j-\zeta_j^TA_t\xi_j|\\
&=|\hat{\zeta}_j^T\hat{A}_t(\hat{\xi}_j-\xi_j)+(\hat{\zeta}_j^T-\zeta_j^T)A_t\xi_j+\hat{\zeta}_j^T(\hat{A}_t-A_t)\xi_j|\\
&\leq\|\hat{\zeta}_j^T\|\|\hat{A}_t\|\|\hat{\xi}_j-\xi_j\|+\|\hat{\zeta}_j^T-\zeta_j^T\|\|A_t\xi_j\|+\|\hat{\zeta}_j^T\|\|\xi_j\|\varepsilon_A\\
\|\zeta_j^T\|&=\|e_j^TP_\tau R\|\leq\|P_\tau R\|=\|R\|\\
\|\hat{\zeta}_j^T-\zeta_j^T\|&=\|e_j^T(\hat{R}-P_\tau R)\|\leq\|\hat{R}-P_\tau R\|=\varepsilon_R\\
\|\xi_j\|&=\|(P_\tau R)^{-1}e_j\|\leq\|(P_\tau R)^{-1}\|=\|R^{-1}\|\\
\|\hat{\xi}_j-\xi_j\|&=\|(\hat{R}^{-1}-(P_\tau R)^{-1})e_j\|\leq\|\hat{R}^{-1}-(P_\tau R)^{-1}\|\\
&\leq\frac{\|R^{-1}\|^2\varepsilon_R}{1-\varepsilon_R\|R^{-1}\|}\text{ (by Lemma 37 of Abraham et al. [2022])}\\
\|A_t\|&=\|R^{-1}diag(\lambda_{t,.})R\|\leq\lambda_{\max}\kappa(R)\\
\|A_t\xi_j\|&=\|\lambda_{t,\tau(j)}\xi_j\|\leq\lambda_{\max}\|R^{-1}\|\\
\|\hat{\zeta}_j^T\|&=\|\hat{\zeta}_j^T-\zeta_j^T+\zeta_j^T\|\leq\|\hat{\zeta}_j^T-\zeta_j^T\|+\|\zeta_j^T\|\leq\varepsilon_R+\|R\|\\
\|\hat{A}_t\|&\leq\varepsilon_A+\|A_t\|\leq\varepsilon_A+\lambda_{\max}\kappa(R).
\end{aligned}$$

Thus, by the inequalities above:

$$
\begin{aligned}
|\hat{\lambda}_{t,j} - \lambda_{t,\tau(j)}| &\le (\varepsilon_R + \|R\|)(\varepsilon_A + \lambda_{\max}\kappa(R))\frac{\|R^{-1}\|^2\varepsilon_R}{1-\varepsilon_R\|R^{-1}\|} \\
&\quad + \lambda_{\max}\varepsilon_R\|R^{-1}\| + \|R^{-1}\|(\varepsilon_R + \|R\|)\varepsilon_A \\
&\le 2(\varepsilon_R + \|R\|)(\varepsilon_A + \lambda_{\max}\kappa(R))\|R^{-1}\|^2\varepsilon_R \\
&\quad + \lambda_{\max}\|R^{-1}\|\varepsilon_R + \|R^{-1}\|(\varepsilon_R + \|R\|)\varepsilon_A \\
&\le 2\|R^{-1}\|^2\varepsilon_R^2\varepsilon_A + 2\kappa(R)\|R^{-1}\|\varepsilon_A\varepsilon_R + 2\lambda_{\max}\kappa(R)\|R^{-1}\|^2\varepsilon_R^2 \\
&\quad + 2\lambda_{\max}\kappa(R)^2\|R^{-1}\|\varepsilon_R + \lambda_{\max}\|R^{-1}\|\varepsilon_R + \kappa(R)\varepsilon_A + \|R^{-1}\|\varepsilon_R\varepsilon_A \\
&\le 2\|R^{-1}\|^2\varepsilon_R^2\varepsilon_A + (2\kappa(R)+1)\|R^{-1}\|\varepsilon_A\varepsilon_R \\
&\quad + 2\lambda_{\max}\kappa(R)\left((\|R^{-1}\|\varepsilon_R)^2 + \kappa(R)\left(\|R^{-1}\|\varepsilon_R\right)\right) + \lambda_{\max}\|R^{-1}\|\varepsilon_R + \kappa(R)\varepsilon_A \\
&\le 2\left(\frac{1}{2}\right)^2\varepsilon_A + \frac{3}{2}\kappa(R)\varepsilon_A + 2\lambda_{\max}\kappa(R)\|R^{-1}\|\varepsilon_R\left(\kappa(R) + \frac{1}{2}\right) \\
&\quad + \lambda_{\max}\|R^{-1}\|\varepsilon_R + \kappa(R)\varepsilon_A \\
&\le 3\kappa(R)\varepsilon_A + \lambda_{\max}(1+3\kappa(R)^2)\|R^{-1}\|\varepsilon_R \\
&\le 4\kappa(R)\left[\varepsilon_A + \lambda_{\max}\kappa(R)\|R^{-1}\|\varepsilon_R\right]
\end{aligned}
$$

where we have used $\|R^{-1}\|\varepsilon_R \le 1/2$ and $\kappa(R) \ge 1$. □

3.6.14 Proof of Lemma 3.5.1

Let define $p_{n,i}(\theta) := \max_k \mathbb{P}_\theta\left(X_i = k \mid Y_{1:n}\right) = \mathbb{P}_\theta\left(X_i = h^\star_{\theta,i} \mid Y_{1:n}\right)$; defining $h^\star_{\theta,i}$ realizing the maximum. Suppose $p_{n,i}(\theta) \ge \frac{1}{2} + \gamma$ for some $0 < \gamma \le 1/2$. Then,

$$
\begin{aligned}
p_{n,i}(\theta) - \max_{k \neq h^\star_{\theta,i}} \mathbb{P}_\theta\left(X_i = k \mid Y_{1:n}\right) &\ge p_{n,i}(\theta) - \sum_{k \neq h^\star_{\theta,i}} \mathbb{P}_\theta\left(X_i = k \mid Y_{1:n}\right) \\
&= p_{n,i}(\theta) - [1 - p_{n,i}(\theta)] \\
&= 2p_{n,i}(\theta) - 1 \\
&\ge 2\gamma.
\end{aligned}
$$

Consequently if $p_{n,i}(\theta) \ge \frac{1}{2} + \gamma$,

$$
\begin{aligned}
\mathbb{P}_{\hat{\theta}}\left(X_i = h^\star_{\theta,i} \mid Y_{1:n}\right) &\ge p_{n,i}(\theta) - \left\|\phi_{\theta,i|n} - \phi_{\hat{\theta},i|n}\right\|_{\mathrm{TV}} \\
&\ge \max_{k \neq h^\star_{\theta,i}} \mathbb{P}_\theta\left(X_i = k \mid Y_{1:n}\right) + 2\gamma - \left\|\phi_{\theta,i|n} - \phi_{\hat{\theta},i|n}\right\|_{\mathrm{TV}} \\
&\ge \max_{k \neq h^\star_{\theta,i}} \mathbb{P}_{\hat{\theta}}\left(X_i = k \mid Y_{1:n}\right) + 2\gamma - 2\left\|\phi_{\theta,i|n} - \phi_{\hat{\theta},i|n}\right\|_{\mathrm{TV}}
\end{aligned}
$$

We have shown that on the intersection of the two events

$$
E_{n,i} := \left\{p_{n,i}(\theta) \ge \frac{1}{2} + \gamma\right\}, \quad F_{n,i} := \left\{\left\|\phi_{\theta,i|n} - \phi_{\hat{\theta},i|n}\right\|_{\mathrm{TV}} < \gamma\right\}
$$

the plug-in rule $h^\star_{\hat\theta}$ maximizing $k \mapsto \mathbb{P}_{\hat\theta}(X_i = k \mid Y_{1:n})$ is unique and it must be that $h^\star_{\hat\theta,i} = h^\star_{\theta,i}$. Then, we bound the risk as follows,

$$\begin{aligned}
\mathcal{R}_n^{\text{class}}(\theta, h^\star_{\hat\theta}) &\le \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{h^\star_{\hat\theta,i} \neq X_i\right\}} \mathbf{1}_{E_{n,i}\cap F_{n,i}}\right] + \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{h^\star_{\hat\theta,i} \neq X_i\right\}} \mathbf{1}_{E^c_{n,i}}\right] \\
&\quad + \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\left\{h^\star_{\hat\theta,i} \neq X_i\right\}} \mathbf{1}_{F^c_{n,i}}\right] \\
&\le \mathbb{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n \mathbb{P}_\theta\left(h^\star_{\theta,i} \neq X_i\right) \mathbf{1}_{E_{n,i}}\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{P}_\theta\left(E^c_{n,i}\right) + \frac{1}{n}\sum_{i=1}^n \mathbb{P}_\theta\left(F^c_{n,i}\right)
\end{aligned}$$

Finally, notice that,

$$\begin{aligned}
\mathbb{P}_\theta\left(E^c_{n,i}\right) &= \mathbb{P}_\theta\left(p_{n,i}(\theta) < \frac{1}{2} + \gamma\right) \\
&= \mathbb{P}_\theta\left(\mathbb{P}_\theta\left(X_i = h^\star_{\theta,i} \mid Y_{1:n}\right) < \frac{1}{2} + \gamma\right) \\
&= \mathbb{P}_\theta\left(\mathbb{P}_\theta\left(X_i \neq h^\star_{\theta,i} \mid Y_{1:n}\right) > \frac{1}{2} - \gamma\right) \\
&\le \frac{1}{1/2 - \gamma}\mathbb{E}_\theta\left(\mathbb{P}_\theta\left(h^\star_{\theta,i} \neq X_i \mid Y_{1:n}\right) \mathbf{1}_{E^c_{n,i}}\right).
\end{aligned}$$

Hence the result.

3.6.15 Equivalence of the definitions of the risk of clustering

Lemma 3.6.15. *The risk of clustering of $\pi_n \circ h$ can be rewritten as*

$$\mathcal{R}_n^{\text{clust}}(\theta, \pi_n \circ h) \coloneqq \mathbb{E}_\theta\left[\min_{\tau \in \mathcal{S}_J} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n}) \neq \tau(X_i)}\right] \tag{3.26}$$

Proof. It suffices to show that

$$\sup_{\substack{M \subseteq \mathcal{E}(\pi_n \circ h(Y_{1:n}), \Pi_n) \\ M \text{ is a matching}}} \sum_{\{C,C'\} \in M} \text{Card}(C \cap C') = \sup_{\tau \in \mathcal{S}_J} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(Y_{1:n}) = \tau(X_i)}$$

Let $C_k = \{i \in [n] \mid h_i(Y_{1:n}) = k\}$ and $C'_k = \{i \in [n] \mid X_i = k\}$. Since the two partitions Π_n and $\pi_n \circ h(Y_{1:n})$ have the same number of clusters J (with possibly empty clusters), the supremum is reached on matchings with J edges. Using this fact, it follow that the matching reaching the supremum is of the form:

$$M = \left\{(C_k, C'_{\tau(k)}) \mid 1 \le k \le J\right\}$$

where τ is a permutation of $\{1, .., J\}$. One obtains:

$$\begin{aligned}
\sup_{\substack{M \subseteq \mathcal{E}(\pi_n \circ h(Y_{1:n}), \Pi_n) \\ M \text{ is a matching}}} \sum_{\{C,C'\} \in M} \text{Card}(C \cap C') &= \sup_{\tau \in \mathcal{S}_J} \sum_{k=1}^J \text{Card}(C_k \cap C'_{\tau(k)}) \\
&= \sup_{\tau \in \mathcal{S}_J} \sum_{k=1}^J \sum_{i=1}^n \mathbf{1}_{\tau^{-1}(X_i) = h_i(Y_{1:n}) = k} \\
&= \sup_{\tau \in \mathcal{S}_J} \sum_{i=1}^n \mathbf{1}_{\tau(X_i) = h_i(Y_{1:n})}
\end{aligned} \tag{3.27}$$

□

3.6.16 Proof of Lemma 3.6.1

Without loss of generality, let $\alpha_i \geq \frac{1}{2}$ for all i. When $p_i = \alpha_i$, we will say that i is given positive bias, and similarly when $p_i = 1 - \alpha_i$ we say it has negative bias. We show that unless we give all positive bias or all negative bias, we can flip the bias of some i to increase the expectation. This suffices to conclude. Let $(Z_i)_{i\in[n]}$ be a sequence of independent Bernoulli random variables such that $Z_n \sim \mathcal{B}(\alpha_i)$, and let $\beta_i \in \{-1, +1\}$ be the bias we give i. We then let

$$Y_i = 2Z_i - 1 \in \{-1, +1\} \text{ and } X_i = \frac{1 + \beta_i Y_i}{2} \in \{0, 1\}.$$

Consequently, $X_i \sim \mathcal{B}\left(\alpha_i \mathbf{1}_{\beta_i=1} + (1 - \alpha_i)\mathbf{1}_{\beta_i=-1}\right)$ and $\sum_{i=1}^n X_i - \frac{n}{2} = \frac{1}{2}\sum_{i=1}^n \beta_i Y_i$. Letting $S_n = \sum_{i=1}^n \beta_i Y_i$, we intend to choose the β_i to maximize $\mathbb{E}[|S_n|]$.
Let $S_{\neq k} = \sum_{i\neq k} \beta_i Y_i$ and define:

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x > 0 \end{cases}$$

Then,

$$\begin{aligned} |S_n| &= S_n \text{sign}(S_n) \\ &= S_n \text{sign}\left(S_{\neq k}\right) + S_n\left(\text{sign}(S_n) - \text{sign}(S_{\neq k})\right) \\ &= S_{\neq k}\text{sign}\left(S_{\neq k}\right) + \beta_k Y_k \text{sign}\left(S_{\neq k}\right) + S_n\left(\text{sign}\left(S_n\right) - \text{sign}\left(S_{\neq k}\right)\right) \\ &= |S_{\neq k}| + \beta_k Y_k \text{sign}\left(S_{\neq k}\right) + \mathbf{1}_{S_{\neq k}=0} \end{aligned}$$

By the computation above and the fact that $S_{\neq k}$ and Y_k are independent:

$$\mathbb{E}[|S_n|] = \mathbb{E}[|S_{\neq k}|] + \mathbb{P}(S_{\neq k} = 0) + \mathbb{E}[\text{sign}(S_{\neq k})]\beta_k \mathbb{E}[Y_k].$$

Since $\mathbb{E}[Y_k] \geq 0$, we conclude that if we fix the values of $(\beta_i)_{i\neq k}$, then the value of β_k that maximizes $\mathbb{E}[|S_n|]$ is:

$$\beta_k = \text{sign}\left(\mathbb{E}[\text{sign}(S_{\neq k})]\right). \tag{3.28}$$

Assume for the moment that all the α_i are distinct and obey $\alpha_i > \frac{1}{2}$. This assumption will guarantee that for non-same-sign biases, there is always at least one bias we can flip to strictly increase $\mathbb{E}[|S_n|]$. We will remove this assumption at the end of the proof. Consider any assignment of the biases $\beta_1, \ldots, \beta_n$.

Lemma 3.6.16. *If all the α_i are distinct and ensure $\alpha_i > \frac{1}{2}$, the values $(\mathbb{E}[sign(S_{\neq j})])_{j\in[n]}$ are distinct, so there exists k such that $\mathbb{E}[sign(S_{\neq k})] \neq 0$.*

Lemma 3.6.17. *Suppose $\beta_j = 1$ and $\beta_k = -1$. Then $\mathbb{E}[sign(S_{\neq k})] \geq \mathbb{E}[sign(S_{\neq j})]$.*

Both facts will be proved at the end of this section. With these facts in hand, consider any non-same-sign biases $\beta_1, \ldots, \beta_n$. By Lemma 3.6.16, there exists k such that:

$$\mathbb{E}[\text{sign}(S_{\neq k})] \neq 0.$$

Without loss of generality, suppose $\beta_k = -1$. There are two cases to consider.

- If $\mathbb{E}[\text{sign}(S_{\neq k})] > 0$, then β_k disobeys the condition (3.28), so swapping to $\beta_k = 1$ increases $\mathbb{E}_\theta[|S_n|]$.

- If $\mathbb{E}[\text{sign}(S_{\neq k})] < 0$, then β_k obeys the condition (3.28), so we need to find another bias to swap. Because the assignment is non-same-sign, there exists j such that $\beta_j = 1$. And by Lemma 3.6.17 , $\mathbb{E}[\text{sign}(S_{\neq j})] < 0$. This means β_j disobeys the condition 3.28, so swapping to $\beta_j = -1$ increases $\mathbb{E}[|S_n|]$.

We have shown that any non-same-sign bias assignment is suboptimal for maximizing $\mathbb{E}[|S_n|]$, so only the same-sign cases can be optimal. And it is clear by symmetry that they both yield the same $\mathbb{E}[|S_n|]$ value, so both are optimal. Note that $\mathbb{E}[|S_n|]$ is a polynomial in the parameters $(\alpha_i)_{i\in[n]}$. When the assumption that all α_i are distinct and obey $\alpha_i > \frac{1}{2}$ does not hold, the result is still true thanks to continuity with respect to $(\alpha_i)_{i\in[n]}$.

Proof of Lemma 3.6.16

Let $S_{\notin\{j,k\}} = \sum_{i\notin\{j,k\}} \beta_i Y_i$ and $j \neq k \in [n]$.

$$\begin{aligned}
&\mathbb{E}\left[\text{sign}(S_{\neq j})\right] - \mathbb{E}\left[\text{sign}(S_{\neq k})\right] = \mathbb{P}\left(S_{\neq j} > 0\right) - \mathbb{P}\left(S_{\neq j} < 0\right) - \mathbb{P}\left(S_{\neq k} > 0\right) + \mathbb{P}\left(S_{\neq k} < 0\right) \\
&= \mathbb{P}\left(S_{\notin\{j,k\}} + \beta_k Y_k > 0\right) - \mathbb{P}\left(S_{\notin\{j,k\}} + \beta_k Y_k < 0\right) - \mathbb{P}\left(S_{\notin\{j,k\}} + \beta_j Y_j > 0\right) + \mathbb{P}\left(S_{\notin\{j,k\}} + \beta_j Y_j < 0\right) \\
&= \alpha_k \mathbb{P}\left(S_{\notin\{j,k\}} > -\beta_k\right) + (1-\alpha_k)\mathbb{P}\left(S_{\notin\{j,k\}} > \beta_k\right) - \alpha_k \mathbb{P}\left(S_{\notin\{j,k\}} < -\beta_k\right) - (1-\alpha_k)\mathbb{P}\left(S_{\notin\{j,k\}} < \beta_k\right) \\
&- \alpha_j \mathbb{P}\left(S_{\notin\{j,k\}} > -\beta_j\right) - (1-\alpha_j)\mathbb{P}\left(S_{\notin\{j,k\}} > \beta_j\right) + \alpha_j \mathbb{P}\left(S_{\notin\{j,k\}} < -\beta_j\right) + (1-\alpha_j)\mathbb{P}\left(S_{\notin\{j,k\}} < \beta_j\right)
\end{aligned}$$

When $\beta_j = \beta_k = 1$,

$$\mathbb{E}\left[\text{sign}(S_{\neq j})\right] - \mathbb{E}\left[\text{sign}(S_{\neq k})\right] = (\alpha_k - \alpha_j)\left(\mathbb{P}\left(S_{\notin\{j,k\}} \in \{0,1\}\right) + \mathbb{P}\left(S_{\notin\{j,k\}} \in \{0,-1\}\right)\right) \neq 0.$$

The remaining cases can be analyzed similarly.

Proof of Lemma 3.6.17

We want to show that:

$$\mathbb{E}\left[\text{sign}(S_{\notin\{j,k\}} + Y_j)\right] \geq \mathbb{E}\left[\text{sign}(S_{\notin\{j,k\}} - Y_k)\right]$$

The key observation is that we have a stochastic dominance $+Y_j \geq_{st} -Y_k$. This means that for any nondecreasing function f:

$$\mathbb{E}\left[f(Y_j)\right] \geq \mathbb{E}\left[f(-Y_k)\right].$$

The desired result follows by applying the above to

$$f(x) = \mathbb{E}\left[\text{sign}(S_{\notin\{j,k\}} + x)\right]$$

which is clearly nondecreasing.

Chapter 4

Clustering in slowly-mixing Gaussian HMMs

We study the problem of clustering under the hidden Markov model with Gaussian emissions, focusing on the regime where the hidden chain mixes slowly. We provide a precise characterization of how the Bayes risk depends on the model parameters and construct a Bayes-optimal clustering procedure. Notably, our analysis reveals surprising and non-standard behavior of the Bayes risk in certain parameter regimes, offering new insights into the interplay between signal strength and temporal dependence.

This chapter is a collaboration with Mohamed Ndaoud.

Contents

4.1 Introduction

Clustering is a central problem in unsupervised learning, widely used to uncover latent structure in data, particularly those generated by mixture models. It has been extensively studied from both a practical standpoint, leading to a variety of clustering algorithms, and from a theoretical perspective, where rigorous guarantees have been established for the clustering performance in various settings. In this work, we focus on the problem of clustering in hidden Markov models (HMMs) with Gaussian emission distributions and a slowly mixing latent Markov chain. More precisely, we consider observations $(X_i)_{i\in[n]}$ generated according to

$$X_i = \eta_i \theta + \xi_i, \tag{4.1}$$

where $\theta \in \mathbb{R}^d$ is an unknown signal vector, the noise vectors $\xi_i \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent, and $(\eta_i)_{i\in[n]}$ is a stationary, symmetric binary Markov chain on $\{-1, +1\}$ with transition matrix

$$Q = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix}.$$

Our focus is on the *slowly mixing* regime, where the transition probability δ is small, so the latent labels (η_i) tend to persist for longer periods in the same state. The purpose of this work is to identify precisely the dependence of the Bayes risk of clustering with respect to the model parameters δ and $\|\theta\|$. The Bayes risk of clustering is defined by

$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, \delta, g)$$

where $\mathcal{G}_n$ is the set of clusterers and $\mathcal{R}_n^{\text{clust}}(\theta, \delta, g)$ is the risk of clustering of clusterer g when clustering n observations generated through model (4.1). We will study both the online and offline risks of clustering. See Section 4.3 for the formal definition of these risks.

4.1.1 Notations

Throughout this work, we use the following notations. The operator $\|\cdot\|$ is the ℓ_2-norm in $\mathbb{R}^d$, and $\|\cdot\|_{\text{op}}$ is the operator norm with respect to the ℓ_2-norm. $\mathbf{I_d}$ is the identity matrix of dimension d. For any symmetric matrix A, $\lambda_{\max}(A)$ will denote the top eigenvalue of A and $v_{\max}(A)$ the corresponding unit eigenvector. For given quantities u and v that might depend on some parameters, we write $u \lesssim v$ ($u \gtrsim v$) when $u \le cv$ ($u \ge cv$) for some absolute constant $c > 0$. When $u \lesssim v$ and $v \lesssim u$, we write $u \asymp v$.

4.2 Related literature

4.2.1 Estimation

A key question explored in the literature of parameter estimation concerns the characterization of the minimax estimation error rate for the parameter θ up to a sign flip. The loss function under consideration is

$$\ell\left(\theta, \theta'\right) = \min\left\{\|\theta - \theta'\|,\ \|\theta + \theta'\|\right\},$$

and the corresponding minimax estimation error is defined by.

$$M(n, d, \delta, t) = \inf_{\hat{\theta}(X_{1:n})} \sup_{\|\theta\|=t} \mathbb{E}\left[\ell\left(\hat{\theta}(X_{1:n}),\ \theta\right)\right]$$

where $\mathbb{E}$ is the expectation under the HMM model (4.1). The objective is to characterize the exact dependence of this risk with respect to the model parameters. This problem was

first studied in the context of Gaussian location models, which corresponds to the setting where $\delta = 0$. Under these models, one observes independent samples from the distribution $\mathcal{N}(\theta, \mathbf{I}_d)$. It was shown in Wu [2017] that the minimax estimation error in this setting satisfies:

$$M(n, d, 0, t) = \begin{cases} t & t \le \sqrt{\frac{d}{n}}, \\ \sqrt{\frac{d}{n}} & t \ge \sqrt{\frac{d}{n}}. \end{cases}$$

This rate is achieved by the empirical mean estimator when $t \ge \sqrt{d/n}$, and by the trivial estimator $\hat{\theta} = 0$ when $t \le \sqrt{d/n}$. In the case where the norm $\|\theta\|$ is unknown, the global minimax risk is given by

$$\inf_{\hat{\theta}(X_{1:n})} \sup_{\theta} \mathbb{E}\left[\ell\left(\hat{\theta}(X_{1:n}),\ \theta\right)\right] = \sqrt{\frac{d}{n}}.$$

Another setting where the problem of estimation was studied is the Gaussian Mixture model. In this setting, one observes an i.i.d. sample from the mixture distribution $\mathbf{P}_\theta = \frac{1}{2}\mathcal{N}(\theta, \mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\theta, \mathbf{I}_d)$ and aims to estimate the parameter θ. This setting corresponds to the case where $\delta = \frac{1}{2}$. It was shown in Wu and Zhou [2021] that in this case, the minimax estimation error satisfies

$$M(n, d, \tfrac{1}{2}, t) = \min\left\{\sqrt{\frac{d}{n}} + \frac{1}{t}\left(\frac{d}{n} + \sqrt{\frac{d}{n}}\right),\ t\right\}.$$

This expression implies the following asymptotic behavior (for $d \le n$):

$$M(n, d, \tfrac{1}{2}, t) \asymp \begin{cases} t & t \le \left(\frac{d}{n}\right)^{1/4}, \\ \frac{1}{t}\sqrt{\frac{d}{n}} & \left(\frac{d}{n}\right)^{1/4} \le t \le 1, \\ \sqrt{\frac{d}{n}} & t \ge 1. \end{cases}$$

In contrast, for the high-dimensional regime $d \ge n$, the minimax rate simplifies to

$$M(n, d, \tfrac{1}{2}, t) = \begin{cases} t & t \le \sqrt{\frac{d}{n}}, \\ \sqrt{\frac{d}{n}} & t \ge \sqrt{\frac{d}{n}}. \end{cases}$$

The minimax rate in this case is achieved by the spectral estimator, as demonstrated in Appendix B of Wu and Zhou [2021]. Additionally, the Expectation-Maximization (EM) algorithm is shown to attain the minimax rate in certain regimes, up to logarithmic factors. The general case, where the latent labels follow a Markovian dependence structure, was studied in Zhang and Weinberger [2022]. It was shown that when the dimension satisfies $d \ge \delta n$, the minimax estimation rates are comparable to those in the Gaussian location model. However, when $d \le \delta n$, an improvement in the estimation rate is achievable. Specifically, the minimax estimation error satisfies

$$M(n, d, \delta, t) \asymp \begin{cases} t & t \le \left(\frac{\delta d}{n}\right)^{1/4}, \\ \frac{1}{t}\sqrt{\frac{\delta d}{n}} & \left(\frac{\delta d}{n}\right)^{1/4} \le t \le \sqrt{\delta}, \\ \sqrt{\frac{d}{n}} & t \ge \sqrt{\delta}. \end{cases}$$

A spectral clustering algorithm achieving this rate was proposed in Zhang and Weinberger [2022], but the analysis contained an error. This was corrected in Karagulyan and Ndaoud

[2024], where a revised spectral estimator was introduced and shown to match the optimal rates. The proposed estimator is adaptive to $\|\theta\|$, and can also be made adaptive to the parameter δ. Importantly, the fully adaptive minimax rate remains an open question. The estimator used differs from the standard spectral method of Wu and Zhou [2021]; it involves partitioning the sample into multiple "buckets", averaging observations within each bucket, and then applying the spectral technique. The appropriate choice of bucket size depends critically on both δ and $\|\theta\|$, which is the source of the adaptation challenge.

4.2.2 Clustering

Theoretical guarantees for clustering algorithms have been extensively studied in the i.i.d. setting. For instance, Lu and H. Zhou [2016] investigated the performance of Lloyd's algorithm for clustering sub-Gaussian mixtures and identified the signal-to-noise ratio with respect to which the clustering error of Lloyd's algorithm decays exponentially after $\log(n)$ iterations. Similarly, relaxed versions of the K-means algorithm were studied in Giraud and Verzelen [2018], Royer [2017], Fei and Chen [2018] and the signal-to-noise ratio governing the associated clustering risk was identified, clarifying thus the dependence of the clustering error with respect to the model parameters in many regimes. In Ndaoud [2022], the minimax risk of clustering was tightly analyzed in the setting of Gaussian mixtures and the optimal separation conditions for exact recovery were identified. The problem was formulated as follows: a statistician observes an i.i.d. sample $(X_i)_{1\leq i\leq n}$ from the mixture distribution

$$\mathbf{P}_\theta = \frac{1}{2}\mathcal{N}(-\theta, \sigma^2\mathbf{I}_d) + \frac{1}{2}\mathcal{N}(\theta, \sigma^2\mathbf{I}_d),$$

with the goal of recovering the hidden labels (up to a global sign flip). The loss function used is

$$r(\hat{\eta}, \eta) = \min\left\{|\hat{\eta} - \eta|, |\hat{\eta} + \eta|\right\},$$

where $|.|$ is the Hamming distance defined by:

$$|\hat{\eta} - \eta| := \sum_{i=1}^{n} |\hat{\eta}_i - \eta_i| = 2\sum_{i=1}^{n} \mathbf{1}_{\hat{\eta}_i \neq \eta_i}.$$

The corresponding minimax risk is defined by

$$\inf_{\hat{\eta}(X_{1:n})} \sup_{(\theta,\eta)\in\Omega_\Delta} \frac{1}{n}\mathbb{E}\left[r\left(\hat{\eta}(X_{1:n}),\ \eta\right)|\eta\right],$$

where $\Omega_\Delta = \left\{\theta \in \mathbb{R}^d : \|\theta\| \geq \Delta\right\} \times \{-1, 1\}^n$. It was established in Ndaoud [2022] that:

$$\inf_{\hat{\eta}(X_{1:n})} \sup_{(\theta,\eta)\in\Omega_\Delta} \frac{1}{n}\mathbb{E}\left[r\left(\hat{\eta}(X_{1:n}),\ \eta\right)|\eta\right] \asymp e^{-cr_n^2}, \tag{4.2}$$

for absolute constant c, where $r_n^2 = \frac{\Delta^2}{\sigma^2} \cdot \frac{1}{1+\frac{d\sigma^2}{n\Delta^2}}$. This provides a tight understanding of the minimax risk of clustering and allows the identification of the optimal separation conditions required for recovering the clusters. A minimax optimal clustering procedure based on Lloyd's iterations is also provided. Building on this line of work, Gassiat et al. [2025] considered the problem of clustering in the general nonparametric HMM framework. There, the fundamental limits of clustering were characterized through the Bayes risk of clustering, defined as the best performance achievable by an oracle with access to the model parameters:

$$\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, \delta, g)$$

where $\mathcal{G}_n$ is the set of clusterers. They provided a sharp characterization of this Bayes risk in the regime where the hidden Markov chain is strongly mixing and constructed a nonparametric plug-in estimator achieving near-optimal performance. However, their results primarily address the *strong mixing* regime, where the parameter δ of the transition matrix is of constant order. In the *slowly mixing* regime, where δ is small or even vanishing, their bounds on the Bayes risk of clustering do not match. Specifically, they established that

$$\frac{\delta^2(1-\tilde{\alpha}_n)}{1-\delta}\Lambda \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta,\delta,g) \leq (1-\delta)\Lambda,$$

where $\Lambda = \int_{\mathbb{R}^d} f_\theta \wedge f_{-\theta}$, with f_θ denoting the Gaussian density centered at θ, and $\tilde{\alpha}_n = \mathcal{O}\left(\frac{1}{n\delta^5}\right)$. As $\delta \to 0$, these bounds fail to provide a precise understanding of how the interplay between the signal strength $\|\theta\|$ and the dependence level δ governs the optimal clustering performance. Also, the optimal conditions for almost full and exact recovery of the clusters are still unclear in the slowly mixing regime. The objective of the present work is to study the Bayes risk of clustering in the slowly mixing regime in order to elucidate how temporal dependencies translate in terms of the clustering performance.

4.3 Setting and definitions

4.3.1 Offline clustering

For any $n \geq 1$, the finite sequence $\eta_{1:n} = (\eta_1, \dots, \eta_n)$ induces a random partition $\Pi_n = \{C_1, C_2, \dots\}$ of $[n]$ whose blocks – the so-called *clusters* – are the equivalence classes for the random equivalence relation $i \sim j \iff \eta_i = \eta_j$. The goal of clustering is to uncover this partition Π_n on the sole basis of the observation $X_{1:n} = (X_1, \dots, X_n)$. We recall the definition of a *clusterer* in this context:

Definition 4.3.1 (Clusterer)**.** *A n-clusterer is a measurable map $g : \left(\mathbb{R}^d\right)^n \to \mathcal{P}[n]$ where $\mathcal{P}[n]$ is the set of partitions of $[n]$. We denote by $\mathcal{G}_n$ the set of all n-clusterers.*

We measure the loss incurred by guessing $g(X_{1:n})$ in place of Π_n via the maximum overlap of a *matching* between the two partitions. To define a matching, we build the complete bipartite graph $(g(X_{1:n}), \Pi_n, \mathcal{E}(g(X_{1:n}), \Pi_n))$ on vertices $g(X_{1:n})$ and Π_n with edge set $\mathcal{E}(g(X_{1:n}), \Pi_n) \coloneqq \{\{C, C'\} \ : \ C \in g(X_{1:n}),\, C' \in \Pi_n\}$. Then we recall that a matching M is a set $M \subseteq \mathcal{E}(g(X_{1:n}), \Pi_n)$ of edges without common vertices (*i.e.* each block of Π_n and $g(X_{1:n})$ appears in at most one edge of the matching). Then, we consider the following loss (see Meilă and Heckerman [2001], Meilă [2005]) of $g(X_{1:n})$ with respect to the true partition Π_n

$$(g(X_{1:n}), \Pi_n) \mapsto 1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(g(X_{1:n}),\Pi_n) \\ M \text{ is a matching}}} \sum_{\{C,C'\}\in M} \text{Card}(C \cap C').$$

with associated risk function $\mathcal{R}_n^{\text{clust}} : \Theta \times \mathcal{G}_n \to [0,1]$

$$\mathcal{R}_n^{\text{clust}}(\theta,\delta,g) \coloneqq \mathbb{E}\left[1 - \frac{1}{n} \sup_{\substack{M \subseteq \mathcal{E}(g(X_{1:n}),\Pi_n) \\ M \text{ is a matching}}} \sum_{\{C,C'\}\in M} \text{Card}(C \cap C')\right]. \tag{4.3}$$

where $\mathcal{G}_n$ is the set of all $n-$*clusterers*. The difficulty of clustering under model (4.1) is measured through the Bayes risk of clustering $\inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta,\delta,g)$. A closely related notion is that of a *classifier*:

Definition 4.3.2 (Classifier). *A n-classifier is a measurable map $h : \left(\mathbb{R}^d\right)^n \to \{0,1\}^n$. We denote by $\mathcal{H}_n$ the set of all n-classifiers.*

Note that, from any n-classifier $h \in \mathcal{H}_n$ corresponds a unique n-clusterer $g \in \mathcal{G}_n$ which can be built via the map $\pi_n : \{0,1\}^n \to \mathcal{P}[n]$ such that

$$g(X_{1:n}) = \pi_n \circ h(X_{1:n}) = \{\{i \ : \ h_i(X_{1:n}) = x\} \ : \ x \in \{0,1\}\}\backslash\{\varnothing\}$$

and any clusterer can be represented that way by choosing a specific labelling of the clusters. For this reason, the notions of clusterer and classifier are very much often amalgamated in the literature. We argue that it would be better to define them separately in order to avoid confusions between the risk of clustering $\mathcal{R}_n^{\text{clust}}(\theta, \delta, \pi_n \circ h)$ and the risk of classification $\mathcal{R}_n^{\text{class}} : \Theta \times \mathcal{H}_n \to [0,1]$ (relative to the loss counting number of misclassified observations)

$$\mathcal{R}_n^{\text{class}}(\theta, \delta, h) \coloneqq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(X_{1:n}) \neq \eta_i}\right]. \tag{4.4}$$

It is easy to show that the risk of clustering of $\pi_n \circ h$ can be rewritten as

$$\mathcal{R}_n^{\text{clust}}(\theta, \delta, \pi_n \circ h) \coloneqq \mathbb{E}\left[\min\left\{\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(X_{1:n}) \neq \eta_i},\ 1 - \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{h_i(X_{1:n}) \neq \eta_i}\right\}\right] \tag{4.5}$$

and differs from (4.4). It is customary to compare the performance of a given classifier h to the best performance attainable by an oracle classifier, namely the *Bayes risk of classification*:

$$\inf_{h \in \mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, \delta, h).$$

4.3.2 Online clustering

We are also interested in the situation where the observations must be clustered *sequentially*; *ie.* estimating $(\Pi_k)_{k\geq 1}$ sequentially. In this situation, we are aiming for a sequence of clusterers $\mathfrak{g} = (g^{(k)})_{k\geq 1}$ with $g^{(k)} \in \mathcal{G}_k$ for every $k \geq 1$ (thus each $g^{(k)}$ is $X_{1:k}$-measurable). The major difference with offline clustering is that, the sequence must satisfy the constraint that the partition $g^{(i+1)}(X_{1:i+1})$ is obtained from $g^{(i)}(X_{1:i})$ by either assigning X_{i+1} to an existing block, or a new block, but no other change is allowed. Another way to put this is that the sequence $\mathfrak{g}$ must be *consistent* in the sense that for all $1 \leq m \leq n$, the partition $g^{(n)}(X_{1:n})$ restricted to $[m]$ must equal $g^{(m)}(X_{1:m})$, $\mathbb{P}_\theta$- a.s, *ie.*

$$\{C \cap [m] \ : \ C \in g^{(n)}(X_{1:n})\}\backslash\{\varnothing\} = g^{(m)}(X_{1:m}) \quad \mathbb{P}\text{- a.s} \tag{4.6}$$

Note that the interest of studying online methods lies in the advantages they offer in terms of theoretical and computational properties: online methods are generally easier to implement in the HMM framework and they have plausible theoretical properties making their theoretical study less complicated compared to their offline counterparts.

Definition 4.3.3 (Online clusterer). *An online clusterer is a sequence $\mathfrak{g} \in \prod_{k\geq 1} \mathcal{G}_k$ that satisfies the consistency constraint (4.6). We denote by $\mathcal{G}^{\text{on}}$ the set of all online clusterers.*

For each time horizon $n \geq 1$, we consider the risk function $\mathcal{R}_n^{\text{clust,on}} : \Theta \times \mathcal{G}^{\text{on}} \to [0,1]$

$$\mathcal{R}_n^{\text{clust,on}}(\theta, \delta, \mathfrak{g}) \coloneqq \mathcal{R}_n^{\text{clust}}(\theta, \delta, g^{(n)}) = \mathbb{E}\left[1 - \frac{1}{n}\sup_{\substack{M \subseteq \mathcal{E}(g^{(n)}(X_{1:n}), \Pi_n) \\ M \text{ is a matching}}} \sum_{\{C,C'\} \in M} \text{Card}(C \cap C')\right].$$

Since $\mathcal{R}_n^{\text{clust,on}}(\theta, \mathfrak{g})$ depends solely on $g^{(n)}$, we will use the abuse of notation $\mathcal{R}_n^{\text{clust,on}}(\theta, g^{(n)})$ instead in order to align with notations used in the offline context. Since each g_k in the sequence $\mathfrak{g} \in \mathcal{G}^{\text{on}}$ can be represented (in a non-unique way) as $\pi_k \circ h^{(k)}$ for some classifier $h^{(k)} \in \mathcal{H}_k$, the whole sequence can be represented as $\mathfrak{g} = (\pi_k \circ h^{(k)})_{k\geq 1} \equiv \pi \circ \mathfrak{h}$. The consistency constraint (4.6) then translates on the sequence $\mathfrak{h} = (h^{(k)})_{k\geq 1}$ as

$$\forall\ 1 \leq m \leq n, \quad h_m^{(m)}(X_{1:m}) = h_m^{(n)}(X_{1:n}) \quad \mathbb{P}\text{- a.s} \tag{4.7}$$

Although online classifiers have no admitted definition Hoi et al. [2021], we adopt the following definition which is consistent with our framework.

Definition 4.3.4 (Online classifier)**.** *An online classifier is a sequence* $\mathfrak{h} \in \prod_{k\geq 1} \mathcal{H}_k$ *that satisfies the consistency constraint* (4.7)*. We denote by* $\mathcal{H}^{\text{on}}$ *the set of all online classifiers.*

Then, for a classifier $\mathfrak{h} \in \mathcal{H}^{\text{on}}$, we define the risk of online classification by:

$$\mathcal{R}_n^{\text{class,on}}(\theta, \delta, \mathfrak{h}) \coloneqq \mathcal{R}_n^{\text{class}}(\theta, \delta, h^{(n)}) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h_i^{(n)}(X_{1:n})\neq \eta_i}\right]. \tag{4.8}$$

As for clustering, we will use the abuse of notation $\mathcal{R}_n^{\text{class,on}}(\theta, h^{(n)})$ because the risk depends only on $h^{(n)}$. Here again, the same remarks we made in the offline context still apply: the risk of online classification is of no statistical interest in an unsupervised context where labels are not observed. However, its analysis would be crucial in obtaining results concerning the risk of clustering.
In view of the definition of $\mathcal{R}_n^{\text{clust,on}}$, of (4.5) and (4.7), we can easily show that

$$\forall \mathfrak{h} \in \mathcal{H}^{\text{on}}, \quad \mathcal{R}_n^{\text{clust,on}}(\theta, \delta, \pi\circ\mathfrak{h}) = \mathbb{E}\left[\min\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h_i^{(n)}(X_{1:n})\neq \eta_i},\ 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h_i^{(n)}(X_{1:n})\neq \eta_i}\right\}\right]. \tag{4.9}$$

Similarly to the offline framework, the *Bayes risk of online classification* can then be defined as :

$$\inf_{\mathfrak{h}\in\mathcal{H}^{\text{on}}} \mathcal{R}_n^{\text{class,on}}(\theta, \delta, \mathfrak{h}) = \inf_{\mathfrak{h}\in\mathcal{H}^{\text{on}}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{h_i^{(n)}(X_{1:n})\neq \eta_i}\right]$$

It was shown in Gassiat et al. [2025] that the Bayes risks of clustering and classification ensure:

$$(1 - \tilde{\alpha}_n) \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h) \leq \inf_{g\in\mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, g) \leq \inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\text{class}}(\theta, h),$$

where $\tilde{\alpha}_n = \mathcal{O}\left(\frac{1}{n^{1/2}\delta^5}\right)$. This ensures that as long as $\delta \gtrsim n^{-1/10}$, the two Bayes risks are of the same magnitude. This applies also to the online bayes risks. Since it is easier to study the Bayes risk of classification, we will focus in what follows on this risk in the regime $\delta \gtrsim n^{-\frac{1}{10}}$: in this case, the bounds on the Bayes risk of classification can be extrapolated to the Bayes risk of clustering.

4.4 Main results

We first study the Bayes risk of clustering.

4.4.1 Online setting

In the case of online clustering, the following proposition provides a lower-bound on the Bayes risk of clustering.

Proposition 4.4.1. *The Bayes risk of online clustering ensures:*

$$\inf_{\mathfrak{g}\in\mathcal{G}^{\text{on}}} \mathcal{R}_n^{\text{clust,on}}(\theta,\delta,\mathfrak{g}) \gtrsim \begin{cases} 1 & \|\theta\|^2 \leq 2\delta, \\ \frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right) & 2\delta < \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right), \\ \frac{1}{\|\theta\|}\exp\left(-\frac{\|\theta\|^2}{2}\left(1+\frac{\log\left(\frac{1-\delta}{\delta}\right)}{2\|\theta\|^2}\right)^2\right) & \|\theta\|^2 > \log\left(\frac{1}{\delta}\right). \end{cases}$$

Now, we seek a clustering procedure that matches this lower-bound, up to multiplicative constants. Consider the online clustering procedure $\mathfrak{g} = \pi_n \circ \hat{\eta}$ where for $i \in [\![1,n]\!]$:

- If $\|\theta\|^2 > \log\left(\frac{1}{\delta}\right)$,

$$\hat{\eta}_i(X_{1:n}) = \text{sign}\left(\langle X_i, \theta\rangle\right)$$

- If $\|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$,

$$\hat{\eta}_i(X_{1:n}) = \begin{cases} \text{sign}(\langle X_i, \theta\rangle) & i \in [\![1,k]\!] \\ \text{sign}\left(\left\langle \frac{1}{k}\sum_{j=i-k+1}^{i} X_j, \theta\right\rangle\right) & i > k \end{cases} \tag{4.10}$$

with $k = \left\lceil \frac{2}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)\right\rceil$.

The following proposition controls the risk of clustering of the clustering procedure $\mathfrak{g} = \pi_n \circ \hat{\eta}$.

Proposition 4.4.2. *The Bayes risk of online clustering ensures:*

$$\inf_{\mathfrak{g}\in\mathcal{G}^{\text{on}}} \mathcal{R}_n^{\text{clust,on}}(\theta,\delta,\mathfrak{g}) \leq \mathcal{R}_n^{\text{clust,on}}(\theta,\delta,\pi_n\circ\hat{\eta}) \leq \begin{cases} \frac{1}{2} & \|\theta\|^2 \leq 2\delta, \\ \frac{4\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right) & 2\delta < \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right), \\ \frac{1}{\sqrt{2\pi}\|\theta\|}\exp\left(-\frac{\|\theta\|^2}{2}\right) & \|\theta\|^2 > \log\left(\frac{1}{\delta}\right). \end{cases}$$

Disregarding multiplicative constants, Propositions 4.4.1 and 4.4.2 together yield a sharp characterization of the online Bayes risk of clustering across the three regimes. Interestingly, the intermediate regime

$$2\delta < \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$$

exhibits a polynomial decay of the Bayes risk with respect to the signal strength—a phenomenon that, to the best of our knowledge, has not been previously documented in the literature on the Bayes risk. The behavior of the Bayes risk of online clustering is summarized in Figure 4.1. Note that the values shown in the figure are not exact, but are accurate up to constant factors.

The behavior of the Bayes risk of online clustering exhibits three regimes:

- $\|\theta\|^2 \leq \delta$: This corresponds to the weak signal regime in which it is not possible to distinguish the two clusters. The Bayes risk of clustering is of constant order.

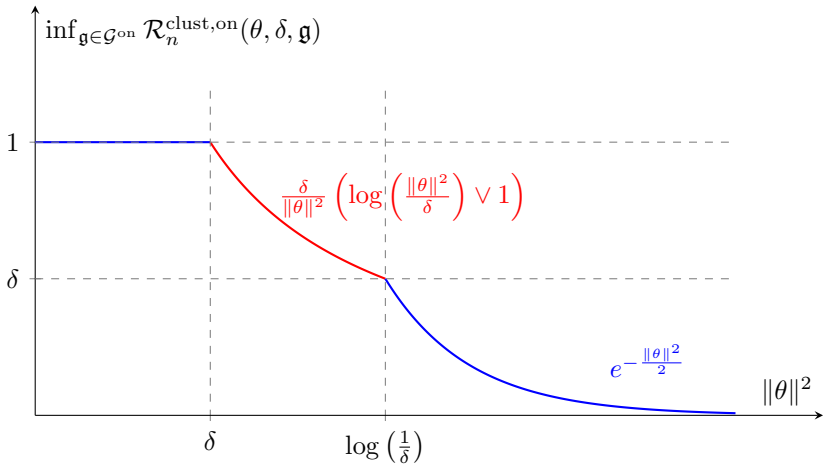


Figure 4.1: Behavior of the Bayes risk of online clustering in the slowly mixing regime

- $\delta < \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$: This is the intermediate signal regime, where the signal is strong enough to begin separating the clusters. In this regime, the Bayes risk decays slowly at a hyperbolic rate, approximately like $\delta/\|\theta\|^2$.
- $\|\theta\|^2 > \log\left(\frac{1}{\delta}\right)$: This is the strong signal regime, where the separation between clusters is large. The Bayes risk decays exponentially fast in $\|\theta\|^2$.

Note that in the strongly mixing regime, that is when δ is of constant order, the intermediate signal regime disappears and one retrieves the behavior displayed in Figure 4.2.

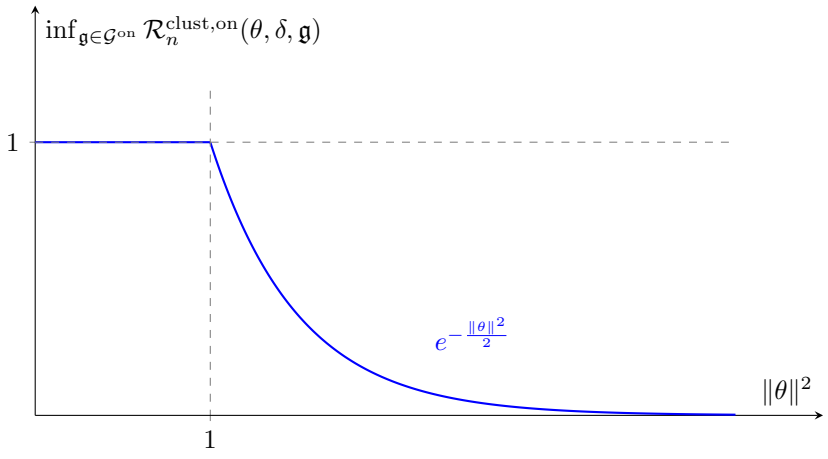


Figure 4.2: Behavior of the Bayes risk of online clustering in the strong mixing regime.

4.4.2 Offline setting

We first have the following lower-bound on the Bayes risk of clustering:

Proposition 4.4.3. *The Bayes risk of offline clustering ensures:*

$$\inf_{\mathfrak{g}\in\mathcal{G}_n} \mathcal{R}_n^{\mathrm{clust}}(\theta,\delta,\mathfrak{g}) \gtrsim \begin{cases} 1 & \|\theta\|^2 \leq 2\delta, \\ \frac{\delta}{\|\theta\|^2} & 2\delta < \|\theta\|^2 \leq 1, \\ \sqrt{\frac{2}{\pi}}\frac{\|\theta\|}{\|\theta\|^2+1}\delta\exp\left(-\frac{\|\theta\|^2}{2}\right) & \|\theta\|^2 > 1. \end{cases}$$

In what follows, we seek a clustering procedure that matches (up to logarithmic terms) the lower-bound of Proposition 4.4.3. Motivated by the proof of the lower bound, we consider the following classification procedure, which uses estimators of the neighboring labels. For $a \in [\![1, n]\!]$, set $\hat{\eta}$ as follows:

- If $\|\theta\|^2 < \log\left(\frac{1}{\delta}\right)$, then set $k = \left\lceil \frac{2}{\|\theta\|^2} \log\left(\frac{\|\theta\|^2}{2\delta}\right) \right\rceil$ and

$$\hat{\eta}_a(\tilde{\eta}_{a-1}, X_{a-k:a+k}, \tilde{\eta}_{a+1}) = \begin{cases} \tilde{\eta}_{a-1} & \text{if } \tilde{\eta}_{a-1} = \tilde{\eta}_{a+1} \text{ and } a \in [\![k+1, n-k]\!] \\ \operatorname{sign}\left(\langle X_a, \theta \rangle\right) & \text{else} \end{cases}$$

  where $\tilde{\eta}_{a-1}$ and $\tilde{\eta}_{a+1}$ are the online estimators defined by:

$$\tilde{\eta}_{a-1}(X_{a-k:a-1}) = \operatorname{sign}\left(\langle \frac{1}{k} \sum_{j=a-k}^{a-1} X_j, \theta \rangle\right), \; \tilde{\eta}_{a+1}(X_{a+1:a+k}) = \operatorname{sign}\left(\langle \frac{1}{k} \sum_{j=a+1}^{a+k} X_j, \theta \rangle\right)$$

- If $\|\theta\|^2 \geq \log\left(\frac{1}{\delta}\right)$, then set $k = 1$ and

$$\hat{\eta}_a(\tilde{\eta}_{a-1}, X_{a-k:a+k}, \tilde{\eta}_{a+1}) = \begin{cases} \operatorname{sign}\left(\langle X_a, \theta \rangle + \tilde{\eta}_{a-1} \log\left(\frac{1-\delta}{\delta}\right)\right) & \text{if } \tilde{\eta}_{a-1} = \tilde{\eta}_{a+1} \text{ and } a \in [\![k+1, n-k]\!] \\ \operatorname{sign}\left(\langle X_a, \theta \rangle\right) & \text{else} \end{cases}$$

The following proposition controls the risk of the clustering procedure associated to this classifier.

Proposition 4.4.4. *The Bayes risk of offline clustering ensures:*

$$\inf_{\mathfrak{g} \in \mathcal{G}_n} \mathcal{R}_n^{\text{clust}}(\theta, \delta, \mathfrak{g}) \leq \mathcal{R}_n^{\text{clust}}(\theta, \delta, \pi_n \circ \hat{\eta}) \lesssim \begin{cases} \frac{1}{2} & \|\theta\|^2 \leq 2\delta, \\ \frac{2\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right) & 2\delta \leq \|\theta\|^2 \leq 1, \\ \frac{2\delta}{\|\theta\|^2} e^{-\frac{\|\theta\|^2}{2}}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right) & 1 \leq \|\theta\|^2 \leq 2\log(\frac{1}{\delta}), \\ \delta \exp\left(-\frac{\|\theta\|^2}{2}\left(1 - \frac{1}{\|\theta\|^2}\log\left(\frac{1-\delta}{\delta}\right)\right)^2\right) & \|\theta\|^2 > 2\log\left(\frac{1}{\delta}\right). \end{cases}$$

The approximate behavior of the Bayes risk in this regime is summarized in Figure 4.3.

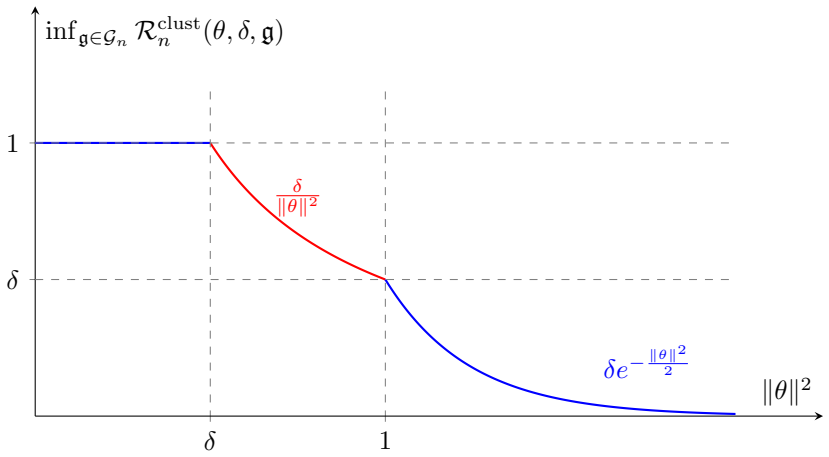


Figure 4.3: Behavior of the Bayes risk of offline clustering in the slowly mixing regime

4.4.3 Adaptation to θ

The clustering procedures described above rely strongly on the knowledge of both θ and δ, as the choice of the bucket size k depends on $\|\theta\|$, and the procedures involve projections along $\frac{\theta}{\|\theta\|}$. Without loss of generality, assume $n = k\ell$ where ℓ is the number of buckets and k is the size of each bucket. For each bucket $i \in [\![1, \ell]\!]$, consider the sample mean of k observations inside the bucket

$$\tilde{X}_i = \frac{1}{k} \sum_{j=k(i-1)+1}^{ki} X_j$$

Note that

$$\tilde{X}_i = \theta \bar{\eta}_i + \frac{\xi_i}{\sqrt{k}}$$

where $\bar{\eta}_i = \frac{1}{k} \sum_{j=k(i-1)+1}^{ki} \eta_j$ and ξ_i is a standard Gaussian vector. We stack the ℓ terms in an $\mathbb{R}^{d\times\ell}$ matrix form as follows:

$$\tilde{X} = \theta \bar{\eta}^\top + \frac{\xi}{\sqrt{k}}$$

We then consider the Gram matrix of observations $\tilde{X}\tilde{X}^\top$. It is easy to see that:

$$\mathbb{E}\left[\frac{1}{\ell}\tilde{X}\tilde{X}^\top\right] = \frac{\mathbb{E}\left[\|\bar{\eta}\|^2\right]}{\ell}\theta\theta^\top + \frac{1}{k}\mathbf{I}_d$$

As established in [Karagulyan and Ndaoud, 2024, Lemma 1], provided that $\ell \geq n\delta$, we have $\mathbb{E}\left[\|\bar{\eta}\|^2\right] \geq \frac{\ell}{3}$. Since our goal is adaptation up to multiplicative constants, it is natural to consider the following estimator of $\|\theta\|$:

$$\|\hat{\theta}(\ell)\| = \left\|\frac{1}{\ell}\tilde{X}\tilde{X}^\top - \frac{1}{k}\mathbf{I}_d\right\|_{op}^{1/2}$$

Similarly, a natural estimator of $u = \frac{\theta}{\|\theta\|}$ is $\hat{u}(\ell)$ the top eigenvector of matrix $\frac{1}{\ell}\tilde{X}\tilde{X}^\top - \frac{1}{k}\mathbf{I}_d$. The following proposition controls the deviation of $\|\hat{\theta}(\ell)\|$ with respect to $\|\theta\|$ and the alignment between $\hat{u}(\ell)$ and u.

Proposition 4.4.5. *There exists $c > 0$ and $\tilde{C} > 0$ such that for $\varepsilon > 0$ and $\ell \geq \left(n\|\theta\|^2 \vee \frac{\tilde{C}^2}{\varepsilon^2}d \vee \frac{2\tilde{C}}{3\varepsilon}n\delta\right) \wedge n$,*

$$\mathbb{P}\left(\left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\right| \geq \varepsilon\left(\frac{\ell}{n} \vee \|\theta\|^2\right)\right) \leq e^{-\frac{c\varepsilon^2\ell}{2}}$$

$$\mathbb{P}\left(\min_{\nu\{-1,1\}} \|\hat{u}(\ell) - \nu u\| \geq \varepsilon\left(\frac{\ell}{n\|\theta\|^2} \vee 1\right)\right) \leq e^{-\frac{c\varepsilon^2\ell}{2}}.$$

Let $\varepsilon > 0$. Define now

$$\mathbf{I}_\ell = \left[\|\hat{\theta}(\ell)\|^2 - \varepsilon\frac{\ell}{n},\ \|\hat{\theta}(\ell)\|^2 + \varepsilon\frac{\ell}{n}\right]$$

$$\hat{\ell} = \min\left\{\ell' \middle| \bigcap_{\ell\geq\ell'} \mathbf{I}_\ell \neq \varnothing\right\}.$$

Let $s^2 \in \cap_{\ell\geq\hat{\ell}}\mathbf{I}_\ell$. The following proposition shows that s^2 is an appropriate estimator of $\|\theta\|^2$ up to multiplicative constants.

Proposition 4.4.6. *There exist positive absolute constants c, c' and C such that in the regime where $\|\theta\|^2 \geq c'\left(\frac{d}{n} \vee \delta\right)$,*

$$\mathbb{P}\left(\left|s^2 - \|\theta\|^2\right| \geq \frac{\|\theta\|^2}{2}\right) \leq Ce^{-cn\left(\|\theta\|^2 \wedge 1\right)}$$

$$\mathbb{P}\left(\min_{\nu \in \{-1,1\}} \|\hat{u} - \nu u\| \geq \varepsilon\right) \leq Ce^{-c\varepsilon^2 n\left(\|\theta\|^2 \wedge 1\right)}$$

where $\hat{u} = \hat{u}(\tilde{\ell})$ with $\tilde{\ell} = 2^{\tilde{m}}$ and $\tilde{m}$ is the largest integer such that $ns^2 \geq 2^{\tilde{m}}$.

Thus, when δ is known, one can use the plug-in the values of $\tilde{\theta}$ and $\hat{u}$ to mimic the behavior of the clustering procedures presented above.

4.4.4 Lower-bound on the minimax risk of clustering

We now turn to the analysis of the minimax risk of clustering in the presence of a Markovian dependence among the labels. Our objective is to understand how the interplay between the parameters δ, d, and n influences the overall clustering performance. The dimension d does not directly affect the Bayes risk, since the latter corresponds to the performance of an oracle that knows the model parameters exactly, thereby avoiding the challenges inherent in high-dimensional estimation. To quantify the impact of not knowing the parameters, we instead consider the minimax risk, defined by

$$\inf_{\mathfrak{g} \in \mathcal{G}_n} \sup_{\|\theta\| \geq \Delta} \mathcal{R}_n^{\text{clust}}(\theta, \delta, \mathfrak{g}).$$

Our goal is to compare the magnitude of this risk with the minimax clustering risk studied by Ndaoud [2022] in the i.i.d. setting. Although the definition of the Bayes risk in Ndaoud [2022] differs slightly from ours (since their supremum also ranges over the possible label values), their bounds remain applicable to the minimax risk in the i.i.d. case, which can be written as

$$\Psi_{1/2,\Delta} = \inf_{\mathfrak{g} \in \mathcal{G}_n} \sup_{\|\theta\| \geq \Delta} \mathcal{R}_n^{\text{clust}}(\theta, 1/2, \mathfrak{g}).$$

In Ndaoud [2022], it is shown that $\Psi_{1/2,\Delta} \asymp e^{-cr_n^2}$ with $r_n = \frac{\Delta^2}{\sqrt{\Delta^2 + \frac{d}{n}}}$. For now, we leave aside the issue of the equivalence between the minimax risks of clustering and classification, and focus on the quantity

$$\Psi_{\delta,\Delta} = \inf_{h \in \mathcal{H}_n} \sup_{\|\theta\| \geq \Delta} \mathcal{R}_n^{\text{class}}(\theta, \delta, h)$$

which corresponds to the minimax risk of offline classification. The following theorem provides a lower-bound on this minimax risk of clustering.

Theorem 4.4.7. *There exists an absolute constant $c > 0$ such that*

$$\Psi_{\delta,\Delta} \geq c\delta\Phi^c\left(r_n\right)$$

where $r_n = \frac{\frac{\Delta^2}{\sigma^2}}{\sqrt{\frac{\Delta^2}{\sigma^2} + \frac{d}{n}}}$.

4.5 Proofs

We first recall the following lemma.

Lemma 4.5.1. *Let X be a Gaussian random variable.*

$$(\forall x > 0) \quad \frac{x}{x^2+1}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \leq \mathbb{P}\left(X > x\right) \leq \left(\frac{1}{\sqrt{2\pi}x} \wedge 1\right)e^{-\frac{x^2}{2}}$$

4.5.1 Proof of Proposition 4.4.1

By stationarity of the hidden chain, we extend the process $(X_i)_{i\in\mathbb{N}}$ to $(X_i)_{i\in\mathbb{Z}}$.

$$\begin{aligned}
\inf_{\mathfrak{h}\in\mathcal{H}^{\text{on}}} \mathcal{R}_n^{\text{class,on}}(\theta,\delta,\mathfrak{h}) &= \inf_{(\hat{\eta}_i(X_{1:i}))_{1\le i\le n}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\eta_i\neq\hat{\eta}_i(X_{1:i})}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \inf_{\hat{\eta}_i(X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(X_{1:i})}\right] \\
&= \frac{1}{n}\sum_{i=1}^{k} \inf_{\hat{\eta}_i(X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(X_{1:i})}\right] + \frac{1}{n}\sum_{i=k+1}^{n} \inf_{\hat{\eta}_i(X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(X_{1:i})}\right] \\
&\ge \frac{1}{n}\sum_{i=k+1}^{n} \inf_{\hat{\eta}_i(\eta_{i-k},X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(\eta_{i-k},X_{1:i})}\right]
\end{aligned}$$

Let $i\in[\![k+1,n]\!]$. Since $\inf_{\hat{\eta}_i(\eta_{i-k},X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(\eta_{i-k},X_{1:i})}\right] = \inf_{\hat{\eta}_i(\eta_{i-k},X_{i-k+1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i\neq\hat{\eta}_i(\eta_{i-k},X_{i-k+1:i})}\right]$, we consider in what follows an estimator of the form $\hat{\eta}_i(\eta_{i-k},X_{i-k+1:i})$.

$$\begin{aligned}
&\mathbb{E}\left[1_{\eta_i\neq\hat{\eta}_i(\eta_{i-k},X_{i-k+1:i})}\right] \ge \sum_{a\in\{1,-1\}} \Big(\mathbb{E}\left[1_{a\neq\hat{\eta}_i(a,X_{i-k+1:i})}1_{\eta_{i-k:i}=a}\right] \\
&+ \sum_{j=i-k+1}^{i} \mathbb{E}\left[1_{-a\neq\hat{\eta}_i(a,X_{i-k+1:i})}1_{\eta_{i-k:j-1}=a,\eta_{j:i}=-a}\right]\Big) \\
&\ge \frac{(1-\delta)^{k-1}}{2} \sum_{a\in\{1,-1\}} \Big((1-\delta)\mathbb{P}\left(\hat{\eta}_i(a,X_{i-k+1:i})=-a|\eta_{i-k:i}=a\right) \\
&+ \delta \sum_{j=i-k+1}^{i} \mathbb{P}\left(\hat{\eta}_i(a,X_{i-k+1:i})=a|\eta_{i-k:j-1}=a,\eta_{j:i}=-a\right)\Big) \\
&\ge \frac{(1-\delta)^{k-1}}{2} \sum_{a\in\{1,-1\}} \Big((1-\delta)\mathbb{P}\left(\hat{\eta}_i(a,X_{i-k+1:i})=-a|\eta_{i-k:i}=a\right) \\
&+ \delta \sum_{j=i-k+1}^{i} \mathbb{P}\left(\hat{\eta}_i(a,X_{i-k+1:i})=a|\eta_{i-k:j-1}=a,\eta_{j:i}=-a\right)\Big) \\
&\ge \frac{(1-\delta)^{k-1}}{2} \sum_{a\in\{1,-1\}} \int_{(\mathbb{R}^d)^k} \Bigg((1-\delta)1_{\hat{\eta}_i(a,x_{i-k+1:i})=-a} \prod_{s=i-k+1}^{i} f_{a\theta}(x_s) \\
&+ \delta 1_{\hat{\eta}_i(a,x_{i-k+1:i})=a} \sum_{j=i-k+2}^{i} \prod_{s=i-k+1}^{j-1} f_{a\theta}(x_s) \prod_{s=j}^{i} f_{-a\theta}(x_s)\Bigg) dx_{i-k+1:i} \\
&\ge \frac{(1-\delta)^{k-1}}{2} \sum_{a\in\{1,-1\}} \int_{(\mathbb{R}^d)^k} \Bigg((1-\delta) \prod_{s=i-k+1}^{i} f_{a\theta}(x_s) \\
&+ 1_{\hat{\eta}_i(a,x_{i-k+1:i})=a}\left(\delta \sum_{j=i-k+2}^{i} \prod_{s=i-k+1}^{j-1} f_{a\theta}(x_s) \prod_{s=j}^{i} f_{-a\theta}(x_s) - (1-\delta) \prod_{s=i-k+1}^{i} f_{a\theta}(x_s)\right)\Bigg) dx_{i-k+1:i}
\end{aligned}$$

The lower bound reaches its smaller value for $\eta_i^\star$ defined by:

$$\eta_i^\star(a,x_{i-k+1:i}) = \text{sign}\left((1-\delta)\prod_{s=i-k+1}^{i} f_{a\theta}(x_s) - \delta \sum_{j=i-k+2}^{i} \prod_{s=i-k+1}^{j-1} f_{a\theta}(x_s) \prod_{s=j}^{i} f_{-a\theta}(x_s)\right)$$

Denoting $C = \frac{(1-\delta)^{k-1}}{2}$, it follows that

$$\begin{aligned}
&\inf_{\hat{\eta}_i(\eta_{i-k}, X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i \neq \hat{\eta}_i(\eta_{i-k}, X_{1:i})}\right] \geq C\delta \sum_{a\in\{1,-1\}} \sum_{j=i-k+1}^{i} \mathbb{P}\left(\eta_i^\star(a, X_{i-k:i}) = a | \eta_{i-k:j-1} = a, \eta_{j:i} = -a\right) \\
&\geq C\delta \sum_{j=i-k+1}^{i} \mathbb{P}\left((1-\delta)\prod_{s=i-k+1}^{i} f_\theta(X_s) > \delta \sum_{j'=i-k+2}^{i} \prod_{s=i-k+1}^{j'-1} f_\theta(X_s) \prod_{s=j'}^{i} f_{-\theta}(X_s) \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} = -1\right) \\
&\geq Ck\delta \left(1 - \frac{1}{k}\sum_{j=i-k+1}^{i} \mathbb{P}\left(\delta \sum_{j'=i-k+2}^{i} \prod_{s=i-k+1}^{j'-1} f_\theta(X_s) \prod_{s=j'}^{i} f_{-\theta}(X_s) \geq (1-\delta) \prod_{s=i-k+1}^{i} f_\theta(X_s) \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} =\right.\right. \\
&\geq Ck\delta \left(1 - \frac{\delta}{k(1-\delta)} \sum_{j=i-k+1}^{i} \sum_{j'=i-k+2}^{i} \mathbb{E}\left[\frac{\prod_{s=i-k+1}^{j'-1} f_\theta(X_s) \prod_{s=j'}^{i} f_{-\theta}(X_s)}{\prod_{s=i-k+1}^{i} f_\theta(X_s)} \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} = -1\right]\right) \\
&\geq Ck\delta \left(1 - \frac{k\delta}{1-\delta} e^{4k\|\theta\|^2}\right)
\end{aligned}$$

where the last line follows from the fact that

$$\frac{\prod_{s=i-k+1}^{j'-1} f_\theta(X_s) \prod_{s=j'}^{i} f_{-\theta}(X_s)}{\prod_{s=i-k+1}^{i} f_\theta(X_s)} = e^{-2\langle \theta, \sum_{s=j'}^{i} X_s\rangle}$$

and

$$\begin{aligned}
\mathbb{E}\left[e^{-2\langle \theta, \sum_{s=j'}^{i} X_s\rangle} \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} = -1\right] &\leq \mathbb{E}\left[e^{-2\|\theta\|^2 \sum_{s=j'}^{i} \eta_s - 2\|\theta\|\langle \frac{\theta}{\|\theta\|}, \sum_{s=j'}^{i} \xi_s\rangle} \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} = -1\right] \\
&\leq e^{2k\|\theta\|^2} \mathbb{E}\left[e^{-2\|\theta\|\langle \frac{\theta}{\|\theta\|}, \sum_{s=j'}^{i} \xi_s\rangle} \Big| \eta_{i-k:j-1} = 1, \eta_{j:i} = -1\right] \\
&\leq e^{4k\|\theta\|^2}
\end{aligned}$$

- For $2\delta < \|\theta\|^2 \leq 1$, we choose $k = \left\lceil \frac{1}{8\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)\right\rceil$. In this case, one has:

$$\begin{aligned}
\frac{k\delta}{1-\delta} e^{4k\|\theta\|^2} &\leq \frac{\delta}{1-\delta}\left(\frac{1}{8\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right) + 1\right) e^{4\|\theta\|^2 + \frac{1}{2}} \left(\frac{\|\theta\|^2}{2\delta}\right)^{\frac{1}{2}} \\
&\leq \frac{\delta}{1-\delta}\left(\frac{1}{8\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right) + \frac{1}{\|\theta\|^2} + 1\right)\left(\frac{\|\theta\|^2}{2\delta}\right)^{\frac{1}{2}} \\
&\lesssim \left((\delta\|\theta\|^2)^{\frac{1}{2}} + \left(\frac{\delta}{\|\theta\|^2}\right)^{1/2} + \left(\frac{\delta}{\|\theta\|^2}\right)^{1/2} \log\left(\frac{\|\theta\|^2}{2\delta}\right)\right)
\end{aligned}$$

  which can be made small in the regime where $2\delta < \|\theta\|^2 < 1$. It follows that in this regime,

$$\inf_{\hat{\eta}_i(\eta_{i-k}, X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i \neq \hat{\eta}_i(\eta_{i-k}, X_{1:i})}\right] \gtrsim \frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)$$

- For $1 \leq \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$, we choose $k = \left\lfloor \frac{1}{8\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)\right\rfloor$. In this case

$$\begin{aligned}
\frac{k\delta}{1-\delta} e^{4k\|\theta\|^2} &\leq \frac{\delta}{1-\delta}\frac{1}{8\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)\left(\frac{\|\theta\|^2}{2\delta}\right)^{1/2} \\
&\leq \frac{1}{16(1-\delta)}\left(\frac{2\delta}{\|\theta\|^2}\right)^{1/2} \log\left(\frac{\|\theta\|^2}{2\delta}\right)
\end{aligned}$$

which can be made smaller than $1/2$ in the regime where $1 \leq \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$. It follows that

$$\inf_{\hat{\eta}_i(\eta_{i-k}, X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i \neq \hat{\eta}_i(\eta_{i-k}, X_{1:i})}\right] \gtrsim \delta\left(\frac{1}{8\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right) - 1\right)$$

but when $1 \leq \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$, $\frac{1}{8\|\theta\|^2}\log\left(\frac{\|\theta\|}{2\delta}\right) - 1 \gtrsim \frac{1}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)$ and one gets

$$\inf_{\hat{\eta}_i(\eta_{i-k}, X_{1:i})} \mathbb{E}\left[\mathbf{1}_{\eta_i \neq \hat{\eta}_i(\eta_{i-k}, X_{1:i})}\right] \gtrsim \frac{\delta}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right) \gtrsim \frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)$$

These lower bounds are valid uniformly for $i \in [\![k+1, n]\!]$. It follows that in the regime where $2\delta < \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$

$$\inf_{\mathfrak{h} \in \mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta, \delta, \mathfrak{h}) \gtrsim \left(1 - \frac{k}{n}\right)\frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)$$

When $\frac{1}{n} \lesssim 2\delta < \|\theta\|^2$, $\frac{k}{n} \lesssim \frac{\delta}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)$ and can be made very small when $\|\theta\|^2 > 2\delta$. It follows that

$$2\delta \leq \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right) \implies \inf_{\mathfrak{h} \in \mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta, \delta, \mathfrak{h}) \gtrsim \frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)$$

For $\|\theta\|^2 \leq 2\delta$, the Bayes risk is of constant order because it is the case for $\|\theta\|^2 \asymp \delta$ and thus the Bayes risk is still of the same order for a weaker signal.
We now focus on the regime where $\|\theta\|^2 > \log\left(\frac{1}{\delta}\right)$.

$$\begin{aligned}
\inf_{\mathfrak{h} \in \mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta, \delta, \mathfrak{h}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\min\left\{\mathbb{P}\left(\eta_i = 1 \mid X_{1:i}\right), \mathbb{P}\left(\eta_i = -1 \mid X_{1:n}\right)\right\}\right] \\
&\geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\min\left\{\mathbb{P}\left(\eta_i = 1 \mid \eta_{i-1}, X_i\right), \mathbb{P}\left(\eta_i = -1 \mid \eta_{i-1}, X_i\right)\right\}\right] \\
&\geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{\min\left\{Q_{\eta_{i-1},1} f_\theta(X_i), Q_{\eta_{i-1},-1} f_{-\theta}(X_i)\right\}}{Q_{\eta_{i-1},1} f_\theta(X_i) + Q_{\eta_{i-1},-1} f_{-\theta}(X_i)}\right] \\
&\geq \int_{\mathbb{R}^d} \delta f_\theta \wedge (1-\delta) f_{-\theta} \\
&\geq (1-\delta)\int_{\langle x, \theta\rangle > \frac{1}{2}\log\left(\frac{1-\delta}{\delta}\right)} f_{-\theta}(x) \\
&\geq (1-\delta)\mathbb{P}_0\left(\langle X, \theta\rangle > \frac{1}{2}\log\left(\frac{1-\delta}{\delta}\right) + \|\theta\|^2\right) \\
&\geq (1-\delta)\mathbb{P}_0\left(\xi > \|\theta\| + \frac{1}{2\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)
\end{aligned}$$

Applying Lemma 4.5.1, one obtains,

$$\begin{aligned}
\inf_{\mathfrak{h} \in \mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta, \delta, \mathfrak{h}) &\geq \frac{1-\delta}{\sqrt{2\pi}} \frac{\|\theta\| + \frac{1}{2\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)}{1 + \left(\|\theta\| + \frac{1}{2\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2} \exp\left(-\frac{1}{2}\left(\|\theta\| + \frac{1}{2\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2\right) \\
&\gtrsim \frac{1}{\|\theta\|} e^{-\frac{\|\theta\|^2}{2}\left(1 + \frac{\log\left(\frac{1-\delta}{\delta}\right)}{2\|\theta\|^2}\right)^2}
\end{aligned}$$

Finally, one obtains the lower-bound:

$$\inf_{\mathfrak{h}\in\mathcal{H}^{\text{on}}} \mathcal{R}_n^{\text{class,on}}(\theta,\delta,\mathfrak{h}) \gtrsim \begin{cases} 1 & \|\theta\|^2 \le 2\delta, \\ \frac{\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right) & 2\delta < \|\theta\|^2 \le \log\left(\frac{1}{\delta}\right), \\ \frac{1}{\|\theta\|}e^{-\frac{\|\theta\|^2}{2}\left(1+\frac{\log\left(\frac{1-\delta}{\delta}\right)}{2\|\theta\|^2}\right)^2} & \|\theta\|^2 > \log\left(\frac{1}{\delta}\right). \end{cases}$$

4.5.2 Proof of Proposition 4.4.2

Let $\eta = (\eta_i)_{i\in[n]}$ be the sequence of hidden states, and define an estimator $\hat{\eta}(X_{1:n}) = (\hat{\eta}_i(X_{1:i}))_{i\in[n]}$. Recall that the observations are generated as

$$X_i = \theta\eta_i + \xi_i, \quad \text{with } \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d),$$

where $(\eta_i)_{i\in[n]}$ is a two-state Markov chain with transition matrix

$$Q = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix},$$

initialized at stationarity. We first focus on the regime $2\delta \le \|\theta\|^2 \le \log\left(\frac{1}{\delta}\right)$ because the bounds for the other regimes are straightforward (see Gassiat et al. [2025]). Without loss of generality, assume $n = k\ell$, where ℓ is the number of blocks (or buckets) and k is the size of each block. Consider the following *oracle* estimator defined by:

- For $a \in \{1, ..., \ell-1\}$, $i \in [\![1,k]\!]$:

$$\hat{\eta}_{ka+i}(X_{1:n}) = \text{sign}\left(\left\langle \frac{1}{k}\sum_{j=i-k+1}^{i} X_{ka+j}, \theta \right\rangle\right)$$

- For $a = 0$ and $i \in [\![1,k]\!]$:

$$\hat{\eta}_i(X_{1:n}) = \text{sign}\left(\langle X_i, \theta\rangle\right)$$

$$\begin{aligned}
\inf_{\mathfrak{h}\in\mathcal{H}^{\text{on}}} \mathcal{R}_n^{\text{class,on}}(\theta,\delta,\mathfrak{h}) &\le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}_{\hat{\eta}_i(X_{1:i})\neq\eta_i}\right] \\
&\le \frac{1}{n}\sum_{i=1}^{k} \mathbb{E}\left[\mathbf{1}_{\hat{\eta}_i(X_{1:i})\neq\eta_i}\right] + \frac{1}{n}\sum_{i=k+1}^{n} \mathbb{E}\left[\mathbf{1}_{\hat{\eta}_i(X_{1:i})\neq\eta_i}\right] \\
&\le \frac{k}{n}\exp\left(-\frac{\|\theta\|^2}{2}\right) + \frac{1}{n}\sum_{a=1}^{\ell-1}\sum_{i=1}^{k} \mathbb{P}\left(\langle \frac{1}{k}\sum_{j=i-k+1}^{i} X_{ka+j}, \theta\rangle \eta_{ka+i} < 0\right)
\end{aligned}$$

For $a \in [\![1, \ell-1]\!]$ and $i \in [\![1,k]\!]$:

$$\begin{aligned}
\mathbb{P}\left(\langle \frac{1}{k}\sum_{j=i-k+1}^{i} X_{ka+j}, \theta\rangle \eta_{ka+i} < 0\right) &\le \mathbb{P}\left(\left\{\eta_{k(a-1)+i+1} = \cdots = \eta_{ka+i}\right\}^c\right) \\
&\quad + \mathbb{P}\left(\langle \frac{1}{k}\sum_{j=i-k+1}^{i} \xi_{ka+j}, \theta\rangle \eta_{ka+i} + \|\theta\|^2 < 0\right) \\
&\le 1 - (1-\delta)^{k-1} + \exp\left(-\frac{k\|\theta\|^2}{2}\right) \\
&\le \delta(k-1) + \exp\left(-\frac{k\|\theta\|^2}{2}\right)
\end{aligned}$$

Consequently,

$$\inf_{\mathfrak{h}\in\mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta,\delta,\mathfrak{h}) \leq \frac{k}{n}\exp\left(-\frac{\|\theta\|^2}{2}\right) + \delta(k-1) + \exp\left(-\frac{k\|\theta\|^2}{2}\right).$$

Choosing $k^\star = \left\lceil \frac{2}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)\right\rceil$ and using the fact that $\delta \geq \frac{1}{n}$,

$$\begin{aligned}
\frac{k^\star}{n}\exp\left(-\frac{\|\theta\|^2}{2}\right) &\leq \frac{1}{n}\exp\left(-\frac{\|\theta\|^2}{2}\right)\left(\frac{2}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right)\\
&\leq \delta\exp\left(-\frac{\|\theta\|^2}{2}\right) + \frac{2\delta}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)\\
&\leq \frac{2\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right)
\end{aligned}$$

In addition,

$$\delta(k^\star-1) + \exp\left(-\frac{k^\star\|\theta\|^2}{2}\right) \leq \frac{2\delta}{\|\theta\|^2}\left(\log\left(\frac{\|\theta\|^2}{2\delta}\right)+1\right)$$

The bound follows. In the regime where $\|\theta\|^2 \geq \log\left(\frac{1}{\delta}\right)$,

$$\begin{aligned}
\inf_{\mathfrak{h}\in\mathcal{H}^{\mathrm{on}}} \mathcal{R}_n^{\mathrm{class,on}}(\theta,\delta,\mathfrak{h}) &\leq \mathcal{R}_n^{\mathrm{class,on}}(\theta,\delta,\pi_n\circ\hat{\eta})\\
&\leq \mathbb{P}\left(\eta_1\langle X_1,\theta\rangle \leq 0\right)\\
&\leq \frac{1}{\sqrt{2\pi}\|\theta\|}\exp\left(-\frac{\|\theta\|^2}{2}\right)
\end{aligned}$$

by Lemma 4.5.1.

4.5.3 Proof of Proposition 4.4.3

Let $k \in [\![1, \lfloor \frac{n}{4} \rfloor]\!]$. The lower-bound in the regime $\|\theta\|^2 > 1$ follows from Proposition 4.5.2. We first focus on the regime $2\delta < \|\theta\|^2 \leq 1$.

$$\begin{aligned}
\inf_{h\in\mathcal{H}_n} \mathcal{R}_n^{\mathrm{class}}(\theta,\delta,h) &= \inf_{(\hat{\eta}_i(X_{1:n}))_{1\leq i\leq n}} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n 1_{\eta_i\neq\hat{\eta}_i(X_{1:n})}\right]\\
&= \frac{1}{n}\sum_{i=1}^n \inf_{\hat{\eta}_i(X_{1:n})} \mathbb{E}\left[1_{\eta_i\neq\hat{\eta}_i(X_{1:n})}\right]\\
&\geq \frac{1}{n}\sum_{i=2k+1}^{n-2k} \inf_{\hat{\eta}_i(X_{1:n})} \mathbb{E}\left[1_{\eta_i\neq\hat{\eta}_i(X_{1:n})}\right]\\
&\geq \frac{1}{n}\sum_{i=2k+1}^{n-2k} \inf_{\hat{\eta}_i(X_{1:n},\eta_{i-2k},\eta_{i+2k})} \mathbb{E}\left[1_{\eta_i\neq\hat{\eta}_i(X_{1:n},\eta_{i-2k},\eta_{i+2k})}\right]\\
&\geq \frac{1}{n}\sum_{i=2k+1}^{n-2k} \inf_{\hat{\eta}_i(X_{i-2k:i+2k},\eta_{i-2k},\eta_{i+2k})} \mathbb{E}\left[1_{\eta_i\neq\hat{\eta}_i(X_{i-2k:i+2k},\eta_{i-2k},\eta_{i+2k})}\right]
\end{aligned}$$

Let $i \in [\![2k+1, n-2k]\!]$ and denote $A_i = \inf_{\hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})} \mathbb{E}\left[1_{\eta_i \neq \hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})}\right]$.

$$
\begin{aligned}
A_i &\geq \sum_{j=i-2k}^{i-k-1} \mathbb{E}\left[1_{\eta_i \neq \hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})} 1_{\eta_{i-2k:j}=1, \eta_{j+1:i+2k}=-1}\right] \\
&+ \sum_{j=i+k}^{i+2k-1} \mathbb{E}\left[1_{\eta_i \neq \hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})} 1_{\eta_{i-2k:j}=1, \eta_{j+1:i+2k}=-1}\right] \\
&\geq \frac{\delta(1-\delta)^{4k-1}}{2} \sum_{j=i-2k}^{i-k-1} \mathbb{E}\left[1_{\eta_i \neq \hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})} \middle| \eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right] \\
&+ \frac{\delta(1-\delta)^{4k-1}}{2} \sum_{j=i+k}^{i+2k-1} \mathbb{E}\left[1_{\eta_i \neq \hat{\eta}_i(X_{i-2k:i+2k}, \eta_{i-2k}, \eta_{i+2k})} \middle| \eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right] \\
&\geq \frac{\delta(1-\delta)^{4k-1}}{4} \left(\sum_{j=i-2k}^{i-k-1} \mathbb{E}\left[1 + \hat{\eta}_i(X_{i-2k:i+2k}, 1, -1) | \eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right] \right. \\
&\left. + \sum_{j=i+k}^{i+2k-1} \mathbb{E}\left[1 - \hat{\eta}_i(X_{i-2k:i+2k}, 1, -1) | \eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right] \right) \\
&\geq \frac{\delta(1-\delta)^{4k-1}}{4} \left(2k + \int \hat{\eta}_i(x_{i-2k:i+2k}, 1, -1) \left(\sum_{j=i-2k}^{i-k-1} \prod_{a=i-2k}^{j} f_\theta(x_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(x_b) \right. \right. \\
&\left. \left. - \sum_{j=i+k}^{i+2k-1} \prod_{a=i-2k}^{j} f_\theta(x_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(x_b) \right) dx_{i-2k:i+2k} \right)
\end{aligned}
$$

which is minimized by:

$$
\eta_i^\star(x_{i-2k:i+2k}, 1, -1) = \operatorname{sign}\left(\sum_{j=i+k}^{i+2k-1} \prod_{a=i-2k}^{j} f_\theta(x_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(x_b) - \sum_{j=i-2k}^{i-k-1} \prod_{a=i-2k}^{j} f_\theta(x_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(x_b) \right)
$$

In order to ease the notations in what follows, we will denote $\mathbb{E}\left[.|\eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right]$ and $\mathbb{P}\left[.|\eta_{i-2k:j} = 1, \eta_{j+1:i+2k} = -1\right]$ respectively by $\mathbb{E}_j$ and $\mathbb{P}_j$. It follows that,

$$
\begin{aligned}
A_i &\geq \frac{\delta(1-\delta)^{4k-1}}{4} \left(\sum_{j=i-2k}^{i-k-1} \mathbb{E}_j\left[1 + \eta_i^\star(X_{i-2k:i+2k}, 1, -1)\right] + \sum_{j=i+k}^{i+2k-1} \mathbb{E}_j\left[1 - \eta_i^\star(X_{i-2k:i+2k}, 1, -1)\right] \right) \\
&\geq \frac{\delta(1-\delta)^{4k-1}}{2} \sum_{j'=i-2k}^{i-k-1} \mathbb{P}_{j'} \left(\sum_{j=i+k}^{i+2k-1} \prod_{a=i-2k}^{j} f_\theta(X_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(X_b) > \sum_{j=i-2k}^{i-k-1} \prod_{a=i-2k}^{j} f_\theta(X_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(X_b) \right) \\
&+ \frac{\delta(1-\delta)^{4k-1}}{2} \sum_{j'=i+k}^{i+2k-1} \mathbb{P}_{j'} \left(\sum_{j=i+k}^{i+2k-1} \prod_{a=i-2k}^{j} f_\theta(X_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(X_b) < \sum_{j=i-2k}^{i-k-1} \prod_{a=i-2k}^{j} f_\theta(X_a) \prod_{b=j+1}^{i+2k} f_{-\theta}(X_b) \right)
\end{aligned}
$$

Recall the model $(\forall i \in [\![1, n]\!]) \quad X_i = \theta \eta_i + \xi_i$. One has

$$\left\{\sum_{j=i+k}^{i+2k-1}\prod_{a=i-2k}^{j} f_\theta(X_a)\prod_{b=j+1}^{i+2k} f_{-\theta}(X_b) > \sum_{j=i-2k}^{i-k-1}\prod_{a=i-2k}^{j} f_\theta(X_a)\prod_{b=j+1}^{i+2k} f_{-\theta}(X_b)\right\}$$
$$= \left\{\sum_{j=i+k}^{i+2k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} X_b - \sum_{a=i-2k}^{j} X_a,\theta\rangle\right) > \sum_{j=i-2k}^{i-k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} X_b - \sum_{a=i-2k}^{j} X_a,\theta\rangle\right)\right\}$$
$$= \left\{\sum_{j=i+k}^{i+2k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} \xi_b - \sum_{a=i-2k}^{j} \xi_a,\theta\rangle\right)\exp\left(\left(\sum_{b=j+1}^{i+2k}\eta_b - \sum_{a=i-2k}^{j}\eta_a\right)\|\theta\|^2\right)\right.$$
$$\left.> \sum_{j=i-2k}^{i-k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} \xi_b - \sum_{a=i-2k}^{j} \xi_a,\theta\rangle\right)\exp\left(\left(\sum_{b=j+1}^{i+2k}\eta_b - \sum_{a=i-2k}^{j}\eta_a\right)\|\theta\|^2\right)\right\}$$
$$\supseteq \left\{\sum_{j=i+k}^{i+2k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} \xi_b - \sum_{a=i-2k}^{j} \xi_a,\theta\rangle\right) > \sum_{j=i-2k}^{i-k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k} \xi_b - \sum_{a=i-2k}^{j} \xi_a,\theta\rangle + 2(4k+1)\|\theta\|^2\right)\right\}$$

It follows that

$$A_i \geq \frac{k\delta(1-\delta)^{4k-1}}{2}\mathbb{P}\left(\frac{\sum_{j=i+k}^{i+2k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k}\xi_b - \sum_{a=i-2k}^{j}\xi_a,\theta\rangle\right)}{\sum_{j=i-2k}^{i-k-1}\exp\left(\langle\sum_{b=j+1}^{i+2k}\xi_b - \sum_{a=i-2k}^{j}\xi_a,\theta\rangle\right)} > \exp\left(2(4k+1)\|\theta\|^2\right)\right)$$
$$\geq \frac{k\delta(1-\delta)^{4k-1}}{2}\mathbb{P}\left(\frac{\sum_{j=i+k}^{i+2k-1}\exp\left(-2\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\right)}{\sum_{j=i-2k}^{i-k-1}\exp\left(2\langle\sum_{b=j+1}^{i-k}\xi_b,\theta\rangle\right)} > \exp\left(2(4k+1)\|\theta\|^2 + 2\langle\sum_{a=i-k+1}^{i+k-1}\xi_a,\theta\rangle\right)\right)$$

Let $\gamma > 0$. Denote

$$Z = \frac{\sum_{j=i+k}^{i+2k-1}\exp\left(-2\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\right)}{\sum_{j=i-2k}^{i-k-1}\exp\left(2\langle\sum_{b=j+1}^{i-k}\xi_b,\theta\rangle\right)} \text{ and } W = \exp\left(2\langle\sum_{a=i-k+1}^{i+k-1}\xi_a,\theta\rangle\right).$$

Note that $W \perp\!\!\!\perp Z$. One gets,

$$A_i \geq \frac{k\delta(1-\delta)^{4k-1}}{2}\mathbb{P}\left(W\exp\left(2(4k+1)\|\theta\|^2\right) < \gamma\right)\mathbb{P}\left(Z > \gamma\right)$$
$$\geq \frac{k\delta(1-\delta)^{4k-1}}{2}\Phi\left(\frac{\log(\gamma) - 2(4k+1)\|\theta\|^2}{2\|\theta\|\sqrt{2k-1}}\right)\mathbb{P}\left(Z > \gamma\right)$$

where Φ is the cdf of a standard Gaussian.

$$\begin{aligned}
\mathbb{P}\left(Z>\gamma\right) &= \mathbb{P}\left(\frac{\sum_{j=i+k}^{i+2k-1}\exp\left(-2\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\right)}{\sum_{j=i-2k}^{i-k-1}\exp\left(2\langle\sum_{b=j+1}^{i-k}\xi_b,\theta\rangle\right)}>\gamma\right)\\
&\geq \mathbb{P}\left(\left\{\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right\}\bigcap\left\{\frac{\frac{k}{2}}{\sum_{j=i-2k}^{i-k-1}\exp\left(2\langle\sum_{b=j+1}^{i-k}\xi_b,\theta\rangle\right)}>\gamma\right\}\right)\\
&\geq \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right)\mathbb{P}\left(\sum_{j=i-2k}^{i-k-1}\exp\left(2\langle\sum_{b=j+1}^{i-k}\xi_b,\theta\rangle\right)<\frac{k}{2\gamma}\right)\\
&\geq \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right)\left(1-\frac{2\gamma}{k}\sum_{j=i-2k}^{i-k-1}\exp\left(2k\|\theta\|^2\right)\right)\\
&\geq \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right)\left(1-2\gamma\exp\left(2k\|\theta\|^2\right)\right)
\end{aligned}$$

Since the processes $\left(\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\right)_{i+k\leq j\leq i+2k-1}$ and $\left(-\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\right)_{i+k\leq j\leq i+2k-1}$ have the same covariance function, it follows that:

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right) &= \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{-\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right)\\
&= \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}\left(1-1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\leq 0}\right)>\frac{k}{2}\right)\\
&= \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle\leq 0}<\frac{k}{2}\right)\\
&= \mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}<\frac{k}{2}\right)
\end{aligned}$$

Consequently, when k is odd $\mathbb{P}\left(\sum_{j=i+k}^{i+2k-1}1_{\langle\sum_{a=i+k}^{j}\xi_a,\theta\rangle<0}>\frac{k}{2}\right)=1/2$. Choosing $k=2\left\lceil\frac{1}{\|\theta\|^2}\right\rceil+1$ and $\gamma=\frac{1}{4}e^{-2k\|\theta\|^2}$, it follows that in the regime where $2\delta<\|\theta\|^2\leq 1$, one has uniformly for all $i\in[\![2k+1,n-2k]\!]$

$$A_i\gtrsim\frac{\delta}{\|\theta\|^2}$$

Consequently,

$$\mathcal{R}_n^{\star,\text{class}}(\theta,\delta)\gtrsim\frac{n-\frac{4}{\|\theta\|^2}}{n}\frac{\delta}{\|\theta\|^2}\gtrsim\frac{n-\frac{2}{\delta}}{n}\frac{\delta}{\|\theta\|^2}\gtrsim\frac{\delta}{\|\theta\|^2}$$

because $\delta\gtrsim\frac{1}{n}$. the result follows.
We now focus on the regime where $\|\theta\|>1$, the following proposition provides the sought lower-bound.

Proposition 4.5.2.

$$\inf_{\mathfrak{h}\in\mathcal{H}}\mathcal{R}_n^{\text{class}}(\theta,\delta,\mathfrak{h})\geq\delta\left(\frac{1}{2}e^{-2\|\theta\|^2}\vee\sqrt{\frac{2}{\pi}}\frac{\|\theta\|}{\|\theta\|^2+1}e^{-\frac{\|\theta\|^2}{2}}\right)$$

Proof. In what follows η_{-i} stands for $(\eta_j)_{j\in\{1,..,i-1,i+1,..,n\}}$.

$$\begin{aligned}
\inf_{\mathfrak{h}\in\mathcal{H}} \mathcal{R}_n^{\text{class}}(\theta,\delta,\mathfrak{h}) &\geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\min\left\{\mathbb{P}\left(\eta_i=1|X_{1:n},\eta_{-i}\right),\ \mathbb{P}\left(\eta_i=-1|X_{1:n},\eta_{-i}\right)\right\}\right] \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\min\left\{\mathbb{P}\left(\eta_i=1|X_i,\eta_{-i}\right),\ \mathbb{P}\left(\eta_i=-1|X_i,\eta_{-i}\right)\right\}\right] \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\min\left\{\mathbb{P}\left(\eta_i=1,X_i,\eta_{-i}\right),\ \mathbb{P}\left(\eta_i=-1,X_i,\eta_{-i}\right)\right\}}{\mathbb{P}\left(\eta_i=1,X_i,\eta_{-i}\right)+\mathbb{P}\left(\eta_i=-1,X_i,\eta_{-i}\right)}\right] \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\min\left\{f_\theta(X_i)\mathbb{P}\left(\eta_i=1|\eta_{-i}\right),\ f_{-\theta}(X_i)\mathbb{P}\left(\eta_i=-1|\eta_{-i}\right)\right\}}{f_\theta(X_i)\mathbb{P}\left(\eta_i=1|\eta_{-i}\right)+f_{-\theta}(X_i)\mathbb{P}\left(\eta_i=-1|\eta_{-i}\right)}\right]
\end{aligned}$$

Let

$$\xi_i = \mathbb{E}\left[\frac{\min\left\{f_\theta(X_i)\mathbb{P}\left(\eta_i=1|\eta_{-i}\right),\ f_{-\theta}(X_i)\mathbb{P}\left(\eta_i=-1|\eta_{-i}\right)\right\}}{f_\theta(X_i)\mathbb{P}\left(\eta_i=1|\eta_{-i}\right)+f_{-\theta}(X_i)\mathbb{P}\left(\eta_i=-1|\eta_{-i}\right)}\right].$$

- For $i=n$:

$$\xi_n = \mathbb{E}\left[\frac{\min\left\{f_\theta(X_n)Q_{\eta_{n-1},1},\ f_{-\theta}(X_n)Q_{\eta_{n-1},-1}\right\}}{f_\theta(X_n)Q_{\eta_{n-1},1}+f_{-\theta}(X_n)Q_{\eta_{n-1},-1}}\right]$$

  Since $\mathcal{L}\left(X_n|\eta_{n-1}\right) = f_\theta(X_n)Q_{\eta_{n-1},1}+f_{-\theta}(X_n)Q_{\eta_{n-1},-1}$, one gets

$$\xi_n = \mathbb{E}\left[\int_{\mathbb{R}^d} \min\left\{Q_{\eta_{n-1},1}f_\theta(x),\ Q_{\eta_{n-1},-1}f_{-\theta}(x)\right\}dx\right]$$

- For $i=1$:

$$\xi_1 = \mathbb{E}\left[\frac{\min\left\{f_\theta(X_1)Q_{1,\eta_2},\ f_{-\theta}(X_1)Q_{-1,\eta_2}\right\}}{f_\theta(X_1)Q_{1,\eta_2}+f_{-\theta}(X_1)Q_{-1,\eta_2}}\right]$$

  Since $\mathcal{L}\left(X_1|\eta_2\right) = f_\theta(X_1)Q_{1,\eta_2}+f_{-\theta}(X_1)Q_{-1,\eta_2}$ (here we exploit stationarity of the chain), one gets

$$\xi_1 = \mathbb{E}\left[\int_{\mathbb{R}^d} \min\left\{Q_{1,\eta_2}f_\theta(x),\ Q_{-1,\eta_2}f_{-\theta}(x)\right\}dx\right]$$

- For $2\leq i\leq n-1$:

$$\xi_i = \mathbb{E}\left[\frac{\min\left\{f_\theta(X_i)Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}},\ f_{-\theta}(X_i)Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}\right\}}{f_\theta(X_i)Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}+f_{-\theta}(X_i)Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}}\right]$$

Since

$$\mathcal{L}(X_i \mid \eta_{i-1},\eta_{i+1}) = \frac{f_\theta(X_i)Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}+f_{-\theta}(X_i)Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}}{Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}+Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}},$$

$$\begin{aligned}
\xi_i &= \mathbb{E}\left[\int_{\mathbb{R}^d} \frac{\min\left\{Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}f_\theta(x),\ Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}f_{-\theta}(x)\right\}}{Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}+Q_{\eta_{i-1},-1}Q_{-1,\eta_{i+1}}}dx\right] \\
&= \frac{1}{2}\int_{\mathbb{R}^d} Q_{-1,1}Q_{1,1}f_\theta(x)\wedge\ Q_{-1,-1}Q_{-1,1}f_{-\theta}(x)dx + \frac{1}{2}\int_{\mathbb{R}^d} Q_{-1,1}Q_{1,-1}f_\theta(x)\wedge\ Q_{-1,-1}^2 f_{-\theta}(x)dx \\
&+ \frac{1}{2}\int_{\mathbb{R}^d} Q_{1,1}Q_{1,-1}f_\theta(x)\wedge\ Q_{1,-1}Q_{-1,-1}f_{-\theta}(x)dx + \frac{1}{2}\int_{\mathbb{R}^d} Q_{1,1}^2 f_\theta(x)\wedge\ Q_{1,-1}Q_{-1,1}f_{-\theta}(x)dx \\
&= \frac{1}{2}\int_{\mathbb{R}^d} (1-\delta)^2 f_\theta(x)\wedge\ \delta^2 f_{-\theta}(x)dx + \frac{1}{2}\int_{\mathbb{R}^d} \delta^2 f_\theta(x)\wedge(1-\delta)^2 f_{-\theta}(x)dx \\
&\quad + \delta(1-\delta)\int_{\mathbb{R}^d} f_\theta(x)\wedge\ f_{-\theta}(x)dx
\end{aligned}$$

It follows that for all $i \in [\![1, n]\!]$,

$$\xi_i \geq \delta \int_{\mathbb{R}^d} f_\theta(x) \wedge f_{-\theta}(x) dx$$

and consequently,

$$\inf_{\mathfrak{h} \in \mathcal{H}} \mathcal{R}_n^{\text{class}}(\theta, \delta, \mathfrak{h}) \geq \delta \int_{\mathbb{R}^d} f_\theta(x) \wedge f_{-\theta}(x) dx \geq \frac{\delta}{2} e^{-2\|\theta\|^2}$$

By Lemma 4.5.1, one also has:

$$\inf_{\mathfrak{h} \in \mathcal{H}^{\text{on}}} \mathcal{R}_n^{\text{class}}(\theta, \delta, \mathfrak{h}) \geq \delta \sqrt{\frac{2}{\pi}} \frac{\|\theta\|}{\|\theta\|^2 + 1} e^{-\frac{\|\theta\|^2}{2}}$$

□

4.5.4 Proof of Proposition 4.4.4

First note that

$$\begin{aligned}
&\mathbb{P}\left(\{\eta_{a-k} = .. = \eta_{a-1}\}^c \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) = 1 - \mathbb{P}\left(\{\eta_{a-k} = .. = \eta_{a-1}\} \cup \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&= 1 - \mathbb{P}\left(\eta_{a-k} = .. = \eta_{a-1}\right) - \mathbb{P}\left(\eta_{a+1} = \eta_{a-1}\right) + \mathbb{P}\left(\eta_{a-k} = .. = \eta_{a-1} = \eta_{a+1}\right) \\
&= 1 - (1-\delta)^{k-1} - (1-\delta)^2 - \delta^2 + \mathbb{P}\left(\eta_{a-k} = .. = \eta_{a-1} \neq \eta_a \neq \eta_{a+1}\right) \\
&= 1 - (1-\delta)^{k-1} - (1-\delta)^2 - \delta^2 + (1-\delta)^{k+1} + \delta^2(1-\delta)^{k-1} \\
&= 2\delta(1-\delta)(1-(1-\delta)^{k-1}) \\
&\leq 2\delta^2(1-\delta)(k-1) \\
&\mathbb{P}\left(\{\text{at least two changes occur in } \eta_{a-k:a+k}\}\right) = 1 - (1-\delta)^{2k} - 2k\delta(1-\delta)^{2k-1} \leq 4k^2\delta^2
\end{aligned}$$

We first focus on the regime $2\delta \leq \|\theta\|^2 \leq \log\left(\frac{1}{\delta}\right)$. Let $a \in [\![k+1, n-k]\!]$. One has:

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\eta}_a \neq \eta_a\right) \\
&\leq \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\}\right) + \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\eta_a \langle X_a, \theta\rangle < 0\}\right) \\
&\leq \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\eta_{a-1} \neq \eta_a, \eta_a \neq \eta_{a+1}\right) + \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\} \cap \{\eta_{a+1} = \eta_a = \eta_{a-1}\}\right) \\
&\leq \delta^2 + \mathbb{P}\left(\{\text{at least two changes occur in } \eta_{a-k:a+k}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a-1} \neq \eta_{a-1}\} \cap \{\text{only one change occurs in } \eta_{a:a+k}\} \cap \{\eta_{a-k} = ... = \eta_a\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \eta_{a+1}\} \cap \{\text{only one change occurs in } \eta_{a-k:a}\} \cap \{\eta_a = ... = \eta_{a+k}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \eta_{a+1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_{a-1}\} \cap \{\eta_{a-k} = .. = \eta_{a+k}\}\right) \\
&\leq \delta^2 + 4\left(k\delta\right)^2 + 2e^{-\frac{k\|\theta\|^2}{2}}\delta(1-\delta)^{k-1} + e^{-k\|\theta\|^2}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_a\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\tilde{\eta}_{a-1} \neq \eta_{a-1}\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) + \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \eta_{a+1}\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq 2\mathbb{P}\left(\{\eta_{a-k} = .. = \eta_{a-1}\}^c \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&+ 2\mathbb{P}\left(\{\eta_{a-k} = .. = \eta_{a-1}\} \cap \{\eta_{a+1} \neq \eta_{a-1}\} \cap \{\tilde{\eta}_{a-1} \neq \eta_{a-1}\}\right) \\
&\leq 4\delta^2(k-1) + 2\delta(1-\delta)^2 e^{-\frac{k\|\theta\|^2}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\tilde{\eta}_{a+1} \neq \eta_{a+1}\}\right) + \mathbb{P}\left(\{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\tilde{\eta}_{a-1} \neq \eta_{a-1}\}\right) \\
&\leq e^{-\frac{\|\theta\|^2}{2}}\left(\delta(k-1) + e^{-\frac{k\|\theta\|^2}{2}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\eta_a \langle X_a, \theta\rangle < 0\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq 4\delta(1-\delta)^2 e^{-\frac{\|\theta\|^2}{2}}
\end{aligned}
$$

Choosing $k^\star = \left\lceil \frac{2}{\|\theta\|^2} \log\left(\frac{\|\theta\|^2}{2\delta}\right)\right\rceil$ as in the proof of Proposition 4.4.2, one gets that in the regime $2\delta \leq \|\theta\|^2 < \log\left(\frac{1}{\delta}\right)$,

$$
\mathbb{P}\left(\hat{\eta}_a \neq \eta_a\right) \leq 9\delta^2 + \frac{2\delta}{\|\theta\|^2}\left[16\frac{\delta}{\|\theta\|^2}\log\left(\frac{\|\theta\|^2}{2\delta}\right)^2 + 4\delta + \frac{2\delta}{\|\theta\|^2} + 4\delta\log\left(\frac{\|\theta\|^2}{2\delta}\right) + e^{-\frac{\|\theta\|^2}{2}}\left(1 + \log\left(\frac{\|\theta\|^2}{2\delta}\right)\right)\right]
$$

Then it is easy to check that there exists an absolute constant $C > 0$ such that for $2\delta \leq \|\theta\|^2 \leq 8\log(\left(\frac{1}{\delta}\right)$,

$$
\mathbb{P}\left(\hat{\eta}_a \neq \eta_a\right) \leq C\frac{\delta}{\|\theta\|^2}e^{-\frac{\|\theta\|^2}{2}}\left(1 + \log\left(\frac{\|\theta\|^2}{2\delta}\right)\right)
$$

For $2\delta \leq \|\theta\|^2 \leq 8 \log\left(\frac{1}{\delta}\right)$, the final bound reads:

$$\begin{aligned}
\mathcal{R}_n^{\text{clust}}(\theta, \delta, \pi_n \circ \hat{\eta}) &\leq \frac{2k^\star}{n} e^{-\frac{\|\theta\|^2}{2}} + C \frac{\delta}{\|\theta\|^2} e^{-\frac{\|\theta\|^2}{2}} \left(1 + \log\left(\frac{\|\theta\|^2}{2\delta}\right)\right) \\
&\lesssim \frac{2\delta}{\|\theta\|^2} e^{-\frac{\|\theta\|^2}{2}} \left(\log\left(\frac{\|\theta\|^2}{2\delta}\right) + 1\right)
\end{aligned}$$

We now focus on the regime $\|\theta\|^2 > 2\log(\frac{1}{\delta})$. Note that $e^{-\frac{\|\theta\|^2}{2}} \leq \delta$ in this regime.

$$
\begin{aligned}
&\mathbb{P}\left(\hat{\eta}_a \neq \eta_a\right) \\
&\leq \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\}\right) + \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\}\right) \\
&\leq \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&+ \mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\hat{\eta}_{a+1} \neq \eta_{a+1}\} \cap \{\hat{\eta}_{a-1} \neq \eta_{a-1}\}\right) + \mathbb{P}\left(\{\tilde{\eta}_{a-1} = \eta_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\}\right) \\
&\leq \mathbb{P}\left(\{\eta_{a-1}\langle X_{a-1}, \theta\rangle \leq 0\} \cap \{\eta_{a+1}\langle X_{a+1}, \theta\rangle \leq 0\}\right) + \mathbb{P}\left(\eta_a\left(\langle X_a, \theta\rangle + \eta_{a-1}\log\left(\frac{1-\delta}{\delta}\right)\right) \leq 0\right) \\
&\leq e^{-\|\theta\|^2} + \mathbb{P}\left(\eta_a\langle \xi_a, \theta\rangle \leq -\|\theta\|^2 - \eta_a\eta_{a-1}\log\left(\frac{1-\delta}{\delta}\right)\right) \\
&\leq e^{-\|\theta\|^2} + (1-\delta)\mathbb{P}\left(\eta_a\langle \xi_a, \frac{\theta}{\|\theta\|}\rangle \leq -\|\theta\| - \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right) \\
&\quad + \delta\mathbb{P}\left(\eta_a\langle \xi_a, \frac{\theta}{\|\theta\|}\rangle \leq -\|\theta\| + \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right) \\
&\leq e^{-\|\theta\|^2} + (1-\delta)e^{-\frac{1}{2}\left(\|\theta\| + \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2} + \delta e^{-\frac{1}{2}\left(\|\theta\| - \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} = \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\eta_{a+1} \neq \eta_{a-1}\} \cap \left\{\eta_a\left(\langle X_a, \theta\rangle + \eta_{a-1}\log\left(\frac{1-\delta}{\delta}\right)\right) \leq 0\right\}\right) \\
&\quad + \mathbb{P}\left(\{\tilde{\eta}_{a-1} \neq \eta_{a-1}\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\eta_{a+1} \neq \eta_{a-1} \neq \eta_a\} \cap \left\{\eta_a\langle \xi_a, \frac{\theta}{\|\theta\|}\rangle \leq -\|\theta\| + \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right\}\right) + 4\delta(1-\delta)e^{-\frac{\|\theta\|^2}{2}} \\
&+ \mathbb{P}\left(\{\eta_{a+1} \neq \eta_{a-1} = \eta_a\} \cap \left\{\eta_a\langle \xi_a, \frac{\theta}{\|\theta\|}\rangle \leq -\|\theta\| - \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right\}\right) \\
&\leq \delta^2 e^{-\frac{1}{2}\left(\|\theta\| - \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2} + \delta(1-\delta)e^{-\frac{1}{2}\left(\|\theta\| + \frac{1}{\|\theta\|}\log\left(\frac{1-\delta}{\delta}\right)\right)^2} + 4\delta(1-\delta)e^{-\frac{\|\theta\|^2}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\eta_a\langle X_a, \theta\rangle < 0\} \cap \{\tilde{\eta}_{a+1} \neq \eta_{a+1}\}\right) + \mathbb{P}\left(\{\eta_a\langle X_a, \theta\rangle < 0\} \cap \{\tilde{\eta}_{a-1} \neq \eta_{a-1}\}\right) \\
&\leq 2e^{-\frac{\|\theta\|^2}{2}}e^{-\frac{\|\theta\|^2}{2}}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}\left(\{\tilde{\eta}_{a+1} \neq \tilde{\eta}_{a-1}\} \cap \{\hat{\eta}_a \neq \eta_a\} \cap \{\eta_{a+1} = \eta_{a-1}\}\right) \\
&\leq \mathbb{P}\left(\{\eta_a\langle X_a, \theta\rangle < 0\} \cap \{\eta_{a+1} \neq \eta_{a-1}\}\right) \\
&\leq 4\delta(1-\delta)^2 e^{-\frac{\|\theta\|^2}{2}}
\end{aligned}
$$

It follows that in this regime, there exists an absolute positive constant $C > 0$ such that

$$
\mathbb{P}\left(\hat{\eta}_a \neq \eta_a\right) \leq C\delta e^{-\frac{\|\theta\|^2}{2}\left(1 - \frac{1}{\|\theta\|^2}\log\left(\frac{1-\delta}{\delta}\right)\right)^2}
$$

Finally, for $\|\theta\|^2 > 2\log\left(\frac{1}{\delta}\right)$,

$$\begin{aligned}\mathcal{R}_n^{\text{clust}}(\theta,\delta,\pi_n\circ\hat{\eta}) &\leq \frac{2k^\star}{n}e^{-\frac{\|\theta\|^2}{2}} + C\delta e^{-\frac{\|\theta\|^2}{2}\left(1-\frac{1}{\|\theta\|^2}\log\left(\frac{1-\delta}{\delta}\right)\right)^2} \\ &\lesssim \delta e^{-\frac{\|\theta\|^2}{2}\left(1-\frac{1}{\|\theta\|^2}\log\left(\frac{1-\delta}{\delta}\right)\right)^2}\end{aligned}$$

4.5.5 Proof of Proposition 4.4.5

As shown in [Karagulyan and Ndaoud, 2024, Lemma 1], $\mathbb{E}\left[\|\bar{\eta}\|^2\right] \geq \ell\left(1-\frac{2k\delta}{3}\right)$. Thus, for $\varepsilon \in (0,1)$,

$$\ell \geq \frac{2}{3\varepsilon}n\delta \implies (1-\varepsilon)\,\ell \leq \mathbb{E}\left[\|\bar{\eta}\|^2\right] \leq \ell$$

Consequently, for $\ell \geq \frac{2}{3\varepsilon}n\delta$,

$$\begin{aligned}\left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\right| &\leq \left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\frac{\mathbb{E}\left[\|\bar{\eta}\|^2\right]}{\ell}\right| + \left(1-\frac{\mathbb{E}\left[\|\bar{\eta}\|^2\right]}{\ell}\right)\|\theta\|^2 \\ &\leq \left|\left\|\frac{1}{\ell}\tilde{X}\tilde{X}^\top - \frac{1}{k}\mathbf{I}_d\right\| - \left\|\mathbb{E}\left[\frac{1}{\ell}\tilde{X}\tilde{X}^\top\right] - \frac{1}{k}\mathbf{I}_d\right\|\right| + \varepsilon\|\theta\|^2 \\ &\leq \left\|\frac{1}{\ell}\tilde{X}\tilde{X}^\top - \mathbb{E}\left[\frac{1}{\ell}\tilde{X}\tilde{X}^\top\right]\right\| + \varepsilon\|\theta\|^2 \\ &\leq \frac{\|\theta\|^2}{\ell}\left|\|\bar{\eta}\|^2 - \mathbb{E}\left[\|\bar{\eta}\|^2\right]\right| + \frac{1}{k}\left\|\frac{\omega\omega^\top}{\ell} - \mathbf{I}_d\right\| + \frac{2\|\theta\|\|\bar{\eta}\|}{\ell\sqrt{k}}\left\|\frac{\omega\bar{\eta}}{\|\bar{\eta}\|}\right\| + \varepsilon\|\theta\|^2\end{aligned}$$

Using [Karagulyan and Ndaoud, 2024, Lemma 17], it follows that, for c and C absolute positive constants, with probability greater than $1-e^{-ct}$,

$$\left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\right| \leq \frac{\|\theta\|^2}{\ell}\sqrt{2\ell t} + \frac{2\|\theta\|\sqrt{\ell}}{\sqrt{\ell n}}\left(\sqrt{d} + C(\sqrt{t}+(dt)^{1/4})\right) + \frac{C}{k}\left(\frac{t}{\ell}+\frac{d}{\ell}+\sqrt{\frac{t}{\ell}}+\sqrt{\frac{d}{\ell}}\right) + \varepsilon\|\theta\|^2$$

Assume $\ell \geq \frac{d}{\varepsilon^2}\vee\frac{2}{3\varepsilon}n\delta$ and let $t = \frac{\varepsilon^2\ell}{2}$. It follows that there exists absolute positive constants C, C', C'' and c such that for all $\varepsilon > 0$, with probability greater than $1-e^{-\frac{c\varepsilon^2\ell}{2}}$,

$$\begin{aligned}\left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\right| &\leq 2\varepsilon\|\theta\|^2 + \frac{2\|\theta\|}{\sqrt{n}}\left(\varepsilon\sqrt{\ell} + C\left(\varepsilon\sqrt{\frac{\ell}{2}} + \left(\frac{\ell^2\varepsilon^4}{2}\right)^{1/4}\right)\right) + \frac{C\ell}{n}\left(\frac{\varepsilon^2}{2} + \sqrt{\frac{\varepsilon^2}{2}} + \varepsilon^2 + \varepsilon\right) \\ &\leq 2\varepsilon\|\theta\|^2 + C\varepsilon\|\theta\|\sqrt{\frac{\ell}{n}} + C'\varepsilon\frac{\ell}{n} \\ &\leq C''\varepsilon\left(\frac{\ell}{n}\vee\|\theta\|^2\right)\end{aligned}$$

Thus, there exists $c > 0$ and $C'' > 0$ such that for $\varepsilon > 0$ and $\ell \geq \left(\frac{C''^2}{\varepsilon^2}d \vee \frac{2C''}{3\varepsilon}n\delta\right)\wedge n$,

$$\mathbb{P}\left(\left|\|\hat{\theta}(\ell)\|^2 - \|\theta\|^2\right| \geq \varepsilon\left(\frac{\ell}{n}\vee\|\theta\|^2\right)\right) \leq e^{-\frac{c\varepsilon^2\ell}{2}}.$$

After the application of the Davis-Kahan theorem, the very same proof applies to the estimator of the direction u.

4.5.6 Proof of Proposition 4.4.6

We first consider the case where $\|\theta\| \geq 1$. In this case Proposition 4.4.5 ensures that with probability greater than $1 - e^{-\frac{c\varepsilon^2 n}{2}}$,

$$\left|s^2 - \|\theta\|^2\right| \leq \left|s^2 - \|\hat{\theta}(n)\|^2\right| + \left|\|\hat{\theta}(n)\|^2 - \|\theta\|^2\right| \leq \varepsilon + \varepsilon\|\theta\|^2 \leq 2\varepsilon\|\theta\|^2$$

Assume now that $\|\theta\| < 1$. Let $\ell^\star = \lceil n\|\theta\|^2 \rceil$ and $\mathbf{I}_\ell = \left[\|\hat{\theta}(\ell)\|^2 - \varepsilon\frac{\ell}{n}, \|\hat{\theta}(\ell)\|^2 + \varepsilon\frac{\ell}{n}\right]$. Let $\varepsilon > 0$ and $\gamma > 0$. Assume $\|\theta\|^2 \geq \frac{2\tilde{C}}{3\varepsilon}\delta \vee \frac{\tilde{C}}{n\varepsilon^2}d$.

$$\begin{aligned}
\mathbb{P}\left(\left|s^2 - \|\theta\|^2\right| \geq \gamma\|\theta\|^2\right) &\leq \mathbb{P}\left(\tilde{\ell} > \ell^\star\right) + \mathbb{P}\left(\left\{\tilde{\ell} \leq \ell^\star\right\} \bigcap \left\{\left|s^2 - \|\theta\|^2\right| \geq \gamma\|\theta\|^2\right\}\right) \\
&\leq \mathbb{P}\left(\|\theta\|^2 \notin \bigcap_{\ell \geq \ell^\star} \mathbf{I}_\ell\right) + \mathbb{P}\left(\left\{s^2 \in \mathbf{I}_{\ell^\star}\right\} \cap \left\{\left|s^2 - \|\theta\|^2\right| \geq \gamma\|\theta\|^2\right\}\right) \\
&\leq \sum_{\ell \geq \ell^\star} \mathbb{P}\left(\|\theta\|^2 \notin \mathbf{I}_\ell\right) \\
&\quad + \mathbb{P}\left(\left\{\left|s^2 - \|\hat{\theta}(\ell^\star)\|^2\right| \leq \varepsilon\frac{\lceil n\|\theta\|^2\rceil}{n}\right\} \bigcap \left\{\left|s^2 - \|\theta\|^2\right| \geq \gamma\|\theta\|^2\right\}\right) \\
&\leq \sum_{\ell \geq \ell^\star} e^{-c\ell} + \mathbb{P}\left(\left|\|\theta\|^2 - \|\hat{\theta}(\ell^\star)\|^2\right| \geq \left(\gamma - \varepsilon\frac{\lceil n\|\theta\|^2\rceil}{n\|\theta\|^2}\right)\|\theta\|^2\right)
\end{aligned}$$

For $\gamma = 4\varepsilon$, one gets:

$$\mathbb{P}\left(\left|s^2 - \|\theta\|^2\right| \geq 4\varepsilon\|\theta\|^2\right) \leq \frac{e^{-c\ell^\star}}{1 - e^{-c}} + e^{-\frac{c\varepsilon^2\ell^\star}{2}}.$$

Now we move to the adaptation to $u = \frac{\theta}{\|\theta\|}$. Let $\varepsilon > 0$, $\hat{\ell} = 2^{\hat{m}}$ where $\hat{m}$ is the largest integer such that $ns^2 \geq 2^{\hat{m}}$. First note that on the event $\left\{\left|s^2 - \|\theta\|^2\right| \leq \frac{\|\theta\|^2}{2}\right\}$,

$$\frac{3}{2}n\|\theta\|^2 \geq ns^2 \geq 2^{\hat{m}} \geq \frac{ns^2}{2} \geq \frac{n\|\theta\|^2}{4}$$

which ensures that on this event, $\hat{m} \in [\![\underline{m}, \overline{m}]\!]$ where $\underline{m} = \lfloor\frac{\log(n\|\theta\|^2/4)}{\log(2)}\rfloor$ and $\overline{m} = \lceil\frac{\log(3n\|\theta\|^2/2)}{\log(2)}\rceil$. Note that there are at most four integer values in this interval. It follows that

$$\begin{aligned}
&\mathbb{P}\left(\min_{\nu\in\{-1,1\}}\|\hat{u}(\hat{\ell}) - \nu u\| \geq \varepsilon\right) \leq c_2 e^{-c_3 n\|\theta\|^2} + \mathbb{P}\left(\min_{\nu\in\{-1,1\}}\|\hat{u}(\hat{\ell}) - \nu u\| \geq \frac{3}{2}\varepsilon, \left|s^2 - \|\theta\|^2\right| \leq \frac{\|\theta\|^2}{2}\right) \\
&\leq c_2 e^{-c_3 n\|\theta\|^2} + \sum_{m=\underline{m}}^{\overline{m}} \mathbb{P}\left(\hat{m} = m, \min_{\nu\in\{-1,1\}}\|\hat{u}(2^m) - \nu u\| \geq \varepsilon, \left|s^2 - \|\theta\|^2\right| \leq \frac{\|\theta\|^2}{2}\right) \\
&\leq c_2 e^{-c_3 n\|\theta\|^2} + \sum_{m=\underline{m}}^{\overline{m}} \mathbb{P}\left(\hat{m} = m, \frac{s^2}{\|\theta\|^2} \in \left[\frac{1}{2}, \frac{3}{2}\right], \min_{\nu\in\{-1,1\}}\|\hat{u}(2^m) - \nu u\| \geq \varepsilon\right) \\
&\leq c_2 e^{-c_3 n\|\theta\|^2} + \sum_{m=\underline{m}}^{\overline{m}} \mathbb{P}\left(\hat{m} = m, \frac{s^2}{\|\theta\|^2} \in \left[\frac{1}{2}, \frac{3}{2}\right], \min_{\nu\in\{-1,1\}}\|\hat{u}(2^m) - \nu u\| \geq \frac{2\varepsilon}{3}\frac{2^m}{n\|\theta\|^2}\right) \\
&\leq C e^{-c\varepsilon^2 n\|\theta\|^2}
\end{aligned}$$

for c and C positive absolute constants.

4.5.7 Proof of Theorem 4.4.7

Let π_θ be a Gaussian prior for θ over $\mathbb{R}^d$ with isotropic covariance $\alpha^2 I_d$ and π_η a stationary Markovian prior for η over $\{-1,1\}$ with transition matrix Q. Let $\pi = \pi_\theta \times \pi_\eta$ be a prior on (θ, η). Looking at the proof of the lower-bound of Theorem 1 in Ndaoud [2022], it still applies to the minimax risk $\Psi_{\delta,\Delta}$ and one has:

$$\begin{aligned}\Psi_{\delta,\Delta} &\geq \frac{1}{n}\sum_{i=1}^{n} \inf_{\hat{T}_i(X)\in[-1,1]} \mathbb{E}_{(\theta,\eta)\sim\pi}\left[\mathbb{E}\left[\left|\hat{T}_i(X) - \eta_i\right| \mid (\theta,\eta)\right]\right] - \pi_\theta\left(\|\theta\| < \Delta\right) \\ &\geq \frac{1}{n}\sum_{i=2}^{n-1} \inf_{\hat{T}_i(X,\eta_{1:i-1},\eta_{i+1:n})\in[-1,1]} \mathbb{E}_{(\theta,\eta)\sim\pi}\left[\mathbb{E}\left[\left|\hat{T}_i(X,\eta_{1:i-1},\eta_{i+1:n}) - \eta_i\right| \mid (\theta,\eta)\right]\right] - \pi_\theta\left(\|\theta\| < \Delta\right)\end{aligned}$$

We first recall the following lemma.

Lemma 4.5.3. *Let $X \sim \chi^2_p$ (equivalently $X = \sum_{i=1}^p Z_i^2$ with $Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$). Then for every $x \geq 0$*

$$\Pr\big(X - p \geq 2\sqrt{px} + 2x\big) \leq e^{-x}, \qquad \Pr\big(p - X \geq 2\sqrt{px}\big) \leq e^{-x}.$$

In particular, for every $t > 0$ the two-sided concentration

$$\Pr\big(|X - p| \geq t\big) \leq 2\exp\Big(-\frac{1}{4}\min\Big\{\frac{t^2}{p}, t\Big\}\Big)$$

holds.

It follows that for $\alpha^2 = \frac{\Delta^2}{d(1-\nu_n)}$ with $\nu_n \in (0,1)$,

$$\pi_\theta\left(\|\theta\| < \Delta\right) \leq \exp\left(-\frac{d}{4}\nu_n^2\right)$$

We now focus on the other terms. Let $i \in [\![2, n-1]\!]$.

$$\begin{aligned}\mathbb{E}\left[\left|\hat{T}_i - \eta_i\right| \middle| X, \eta_{1:i-1}, \eta_{i+1:n}\right] &\geq \mathbb{E}\left[\left(1 - \hat{T}_i\right)\mathbf{1}_{\eta_i=1} \middle| X, \eta_{1:i-1}, \eta_{i+1:n}\right] \\ &\quad + \mathbb{E}\left[\left(1 + \hat{T}_i\right)\mathbf{1}_{\eta_i=-1} \middle| X, \eta_{1:i-1}, \eta_{i+1:n}\right]\end{aligned}$$

This lower-bound is minimized by

$$\hat{T}_i(X, \eta_{1:i-1}, \eta_{i+1:n}) = \text{sign}\left(\mathbb{P}\left(\eta_i = 1 | X, \eta_{1:i-1}, \eta_{i+1:n}\right) - \mathbb{P}\left(\eta_i = -1 | X, \eta_{1:i-1}, \eta_{i+1:n}\right)\right)$$

where $\mathbb{P}$ is the joint distribution of θ, η and X.

$$\begin{aligned}&\mathbb{P}\left(\eta_i = 1, X, \eta_{1:i-1}, \eta_{i+1:n}\right) = \mathbb{E}_{\theta\sim\pi_\theta}\left[\mathbb{P}\left(\xi_{1:n} = X_{1:n} - \theta\eta_{1:n}, \eta_i = 1, \eta_{1:i-1}, \eta_{i+1:n} | \theta\right)\right] \\ &\propto \mathbb{E}_{\theta\sim\pi_\theta}\left[\exp\left(-\frac{\|X_i - \theta\|^2 + \sum_{j\neq i}\|X_j - \theta\eta_j\|^2}{2\sigma^2}\right)\mathbb{P}\left(\eta_i = 1, \eta_{1:i-1}, \eta_{i+1:n}\right)\right] \\ &\propto Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}\mathbb{E}_{\theta\sim\pi_\theta}\left[\exp\left(-\frac{n\|\theta\|^2 - 2\langle X_i, \theta\rangle - 2\langle\sum_{j\neq i}\eta_j X_j, \theta\rangle}{2\sigma^2}\right)\right] \\ &\propto Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}\exp\left(\frac{\|X_i + \sum_{j\neq i}\eta_j X_j\|^2}{2\left(n\sigma^2 + \frac{\sigma^4}{\alpha^2}\right)}\right)\end{aligned}$$

It follows that the supervised minimizer of this lower-bound is:

$$T_i^\star\left(X, \eta_{1:i-1}, \eta_{i+1:n}\right) = \operatorname{sign}\left(\langle X_i, \sum_{j\neq i}\eta_j X_j\rangle - \frac{n\sigma^2+\frac{\sigma^4}{\alpha^2}}{2}\log\left(\frac{(1-Q_{\eta_{i-1},1})(1-Q_{1,\eta_{i+1}})}{Q_{\eta_{i-1},1}Q_{1,\eta_{i+1}}}\right)\right)$$

$$\begin{aligned}
&\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-1},\eta_{i+1:n})-\eta_i|\right]\\
&\geq 2\mathbb{E}\left[\mathbf{1}_{\eta_i=1}\mathbf{1}_{\langle\eta_i X_i,\sum_{j\neq i}\eta_j X_j\rangle<\frac{n\sigma^2+\frac{\sigma^4}{\alpha^2}}{2}\log\left(\frac{(1-Q_{\eta_{i-1},\eta_i})(1-Q_{\eta_i,\eta_{i+1}})}{Q_{\eta_{i-1},\eta_i}Q_{\eta_i,\eta_{i+1}}}\right)}\right]\\
&+2\mathbb{E}\left[\mathbf{1}_{\eta_i=-1}\mathbf{1}_{\langle\eta_i X_i,\sum_{j\neq i}\eta_j X_j\rangle<\frac{n\sigma^2+\frac{\sigma^4}{\alpha^2}}{2}\log\left(\frac{(1-Q_{\eta_{i-1},\eta_i})(1-Q_{\eta_i,\eta_{i+1}})}{Q_{\eta_{i-1},\eta_i}Q_{\eta_i,\eta_{i+1}}}\right)}\right]\\
&\geq 2\mathbb{P}\left(\langle\eta_i X_i,\sum_{j\neq i}\eta_j X_j\rangle<\frac{n\sigma^2+\frac{\sigma^4}{\alpha^2}}{2}\log\left(\frac{(1-Q_{\eta_{i-1},\eta_i})(1-Q_{\eta_i,\eta_{i+1}})}{Q_{\eta_{i-1},\eta_i}Q_{\eta_i,\eta_{i+1}}}\right)\right)
\end{aligned}$$

Let

- $Z=\frac{1}{\sqrt{n-1}}\sum_{j\neq i}\eta_j\xi_j$, $\varepsilon=-\langle\eta_i\xi_i,\frac{\frac{\theta}{\sigma}+\frac{Z}{\sqrt{n-1}}}{\|\frac{\theta}{\sigma}+\frac{Z}{\sqrt{n-1}}\|}\rangle$
- $V=\frac{n\sigma^2+\frac{\sigma^4}{\alpha^2}}{2(n-1)}$, $A_\eta=V\log\left(\frac{(1-Q_{\eta_{i-1},\eta_i})(1-Q_{\eta_i,\eta_{i+1}})}{Q_{\eta_{i-1},\eta_i}Q_{\eta_i,\eta_{i+1}}}\right)$
- $\mathcal{A}=\left\{\frac{\|Z\|^2}{n-1}\geq\frac{d}{n-1}(1-\xi_n)\right\}\bigcap\left\{\frac{|\langle Z,\frac{\theta}{\sigma}\rangle|}{\sqrt{n-1}}\leq\beta_n\frac{\|\theta\|^2}{\sigma^2}\right\}$ for ξ_n and β_n in $(0,1)$

$$\begin{aligned}
&\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-1},\eta_{i+1:n})-\eta_i|\right]\\
&\geq 2\mathbb{E}\left[\mathbb{P}\left(\langle\theta+\sigma\eta_i\xi_i,\theta+\frac{\sigma Z}{\sqrt{n-1}}\rangle\leq A_\eta\middle|(\theta,\eta)\right)\right]\\
&\geq 2\mathbb{E}\left[\mathbb{P}\left(\langle\frac{\theta}{\sigma}+\eta_i\xi_i,\frac{\theta}{\sigma}+\frac{Z}{\sqrt{n-1}}\rangle\leq\frac{A_\eta}{\sigma^2}\middle|(\theta,\eta)\right)\right]\\
&\geq 2\mathbb{E}\left[\mathbb{P}\left(\mathcal{A}\bigcap\left\{\frac{\|\theta\|^2-A_\eta}{\sigma^2}+\frac{1}{\sqrt{n-1}}\langle Z,\frac{\theta}{\sigma}\rangle\leq\varepsilon\sqrt{\frac{\|\theta\|^2}{\sigma^2}+\frac{\|Z\|^2}{n-1}+\frac{2}{\sqrt{n-1}}\langle Z,\frac{\theta}{\sigma}\rangle}\right\}\middle|(\theta,\eta)\right)\right]\\
&\geq 2\mathbb{E}\left[\mathbb{P}\left(\mathcal{A}\bigcap\left\{\frac{(1+\beta_n)\|\theta\|^2-A_\eta}{\sigma^2}\leq\varepsilon\sqrt{\frac{\|\theta\|^2}{\sigma^2}(1-2\beta_n)+\frac{d}{n-1}(1-\xi_n)}\right\}\middle|(\theta,\eta)\right)\right]
\end{aligned}$$

Conditionally on (θ,η), the two events $\mathcal{A}$ and $\left\{\frac{(1+\beta_n)\|\theta\|^2-A_\eta}{\sigma^2}\leq\varepsilon\sqrt{\frac{\|\theta\|^2}{\sigma^2}(1-2\beta_n)+\frac{d}{n-1}(1-\xi_n)}\right\}$ are independent. It follows that:

$$\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-1},\eta_{i+1:n})-\eta_i|\right]\geq 2\mathbb{E}\left[\mathbb{P}\left(\mathcal{A}|(\theta,\eta)\right)\mathbb{P}\left(\varepsilon\geq\frac{\|\theta\|^2(1+\beta_n)-A_\eta}{\sigma^2\sqrt{\frac{\|\theta\|^2}{\sigma^2}(1-2\beta_n)+\frac{d}{n-1}(1-\xi_n)}}\middle|(\theta,\eta)\right)\right]$$

It is easy to see that $\mathbb{P}\left(\mathcal{A}|(\theta,\eta)\right)\geq 1-e^{-\frac{d\xi_n^2}{4}}-e^{-\frac{\beta_n^2\|\theta\|^2}{2\sigma^2}(n-1)}$.

$$\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-1},\eta_{i+1:n})-\eta_i|\right]$$
$$\geq 2\mathbb{E}\left[\mathbf{1}_{\eta_{i-1}=\eta_i\neq\eta_{i+1}}\mathbb{P}\left(\mathcal{A}|(\theta,\eta)\right)\mathbb{P}\left(\varepsilon\geq\frac{\|\theta\|^2\left(1+\beta_n\right)}{\sigma^2\sqrt{\frac{\|\theta\|^2}{\sigma^2}(1-2\beta_n)+\frac{d}{n-1}(1-\xi_n)}}\middle|(\theta,\eta)\right)\right]$$

For $\gamma_n\in(0,1)$, consider the event: $\mathcal{B}=\left\{\left|\frac{\|\theta\|^2}{\Delta^2}-1\right|\leq\gamma_n\right\}$. On this event:

$$\mathbb{P}\left(\varepsilon\geq\frac{\|\theta\|^2\left(1+\beta_n\right)}{\sigma^2\sqrt{\frac{\|\theta\|^2}{\sigma^2}(1-2\beta_n)+\frac{d}{n-1}(1-\xi_n)}}\middle|(\theta,\eta)\right)\geq\Phi^c\left(\frac{\frac{\Delta^2}{\sigma^2}\left(1+\beta_n\right)\left(1+\gamma_n\right)}{\sqrt{\frac{\Delta^2}{\sigma^2}(1-2\beta_n)(1-\gamma_n)+\frac{d}{n-1}(1-\xi_n)}}\right)$$

Let $\alpha^2=\frac{\Delta^2}{d(1-\nu_n)}$. For $\gamma_n\geq 4\nu_n$, Lemma 4.5.3 implies

$$\mathbb{P}\left(\mathcal{B}^c\right)\leq\mathbb{P}\left(\left|\frac{\|\theta\|^2}{\alpha^2}-d\right|\geq d\left(\gamma_n(1-\nu_n)-\nu_n\right)\right)\leq\mathbb{P}\left(\left|\frac{\|\theta\|^2}{\alpha^2}-d\right|\geq\frac{d\gamma_n}{2}\right)\leq 2\exp\left(-\frac{d\gamma_n^2}{16}\right).$$

It follows that $\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-k},\eta_{i+k:n})-\eta_i|\right]$ is thus lower-bounded by:

$$\frac{\delta}{2}\Phi^c\left(\frac{\frac{\Delta^2}{\sigma^2}\left(1+\beta_n\right)\left(1+\gamma_n\right)}{\sqrt{\frac{\Delta^2}{\sigma^2}(1-2\beta_n)(1-\gamma_n)+\frac{d}{n-1}(1-\xi_n)}}\right)\left(1-e^{-\frac{d\xi_n^2}{4}}-e^{-c\beta_n^2(1+\gamma_n)\frac{\Delta^2}{2\sigma^2}(n-1)}\right)\left(1-2\exp\left(-\frac{d\gamma_n^2}{16}\right)\right)$$

- When $\frac{\Delta^2}{\sigma^2}\geq\frac{d\log(n)}{n}$, as in Proposition 4.5.2, minimizing the minimax risk by the Bayes risk, one gets:
$$\Psi_{\delta,\Delta}\geq\delta\Phi^c\left(\frac{\Delta}{\sigma}\right).$$
In this regime, $\frac{\Delta}{\sigma}\leq r_n\left(1+\frac{1}{\log(n)}\right)$. Consequently, with $\varepsilon_n=\frac{1}{\log(n)}$,
$$\Psi_{\delta,\Delta}\geq\delta\Phi^c\left(r_n\left(1+\varepsilon_n\right)\right).$$

- When $\frac{\Delta^2}{\sigma^2}\leq\frac{\log(n)}{n}$, $\Phi^c\left(r_n\right)\geq\Phi^c\left(\frac{\Delta^2}{\sigma^2}\right)\geq C$ where C is an absolute constant. It follows that
$$\Psi_{\delta,\Delta}\geq\delta\Phi^c\left(\frac{\Delta}{\sigma}\right)\geq c\delta\Phi^c\left(r_n\right)$$
for an absolute constant c.

- When $\frac{\log^4(n)}{n}\leq\frac{\Delta^2}{\sigma^2}\leq\frac{d\log(n)}{n}$, one might choose β_n and ξ_n and γ_n vanishing to 0 such that $\beta_n\geq\frac{\sigma}{\Delta\sqrt{n}}$ and $\xi_n\geq\frac{1}{\sqrt{d}}$. Let $\nu_n=\sqrt{\frac{n\Delta^2}{\log^2(n)d\sigma^2}}$ and $\gamma_n=4\nu_n$. Note that $\gamma_n\geq\frac{4\log(n)}{\sqrt{d}}$ in this regime. With this choice, one obtains for all $i\in[\![2,n-1]\!]$,
$$\mathbb{E}\left[|T_i^\star(X,\eta_{1:i-k},\eta_{i+k:n})-\eta_i|\right]\geq\frac{\delta}{8}\Phi^c\left(\frac{\frac{\Delta^2}{\sigma^2}\left(1+\beta_n\right)\left(1+\gamma_n\right)}{\sqrt{\frac{\Delta^2}{\sigma^2}(1-2\beta_n)(1-\gamma_n)+\frac{d}{n-1}(1-\xi_n)}}\right)$$
It follows that
$$\Psi_{\delta,\Delta}\geq c\delta\Phi^c\left(r_n\left(1+\varepsilon_n\right)\right)-e^{-\frac{d\nu_n^2}{4}}$$
for $\varepsilon_n=o(1)$ and $c>0$ a positive absolute constant. As in Ndaoud [2022], it is easy to show that in this regime, $e^{-\frac{d\nu_n^2}{4}}=o\left(\delta\Phi^c\left(r_n\left(1+\varepsilon_n\right)\right)\right)$. The result follows.

Chapter 5

Late change-point detection in the preferential attachment random graph model

We consider the problem of late change-point detection under the preferential attachment random graph model with time dependent attachment function. This can be formulated as a hypothesis testing problem where the null hypothesis corresponds to a preferential attachment model with a constant affine attachment parameter δ_0 and the alternative corresponds to a preferential attachment model where the affine attachment parameter changes from δ_0 to δ_1 at a time $\tau_n = n - \Delta_n$ where $0 \leq \Delta_n \leq n$ and n is the size of the graph. It was conjectured in Bet et al. [2025] that when observing only the unlabeled graph, detection of the change is not possible for $\Delta_n = o(n^{1/2})$. In this work, we make a step towards proving the conjecture by proving the impossibility of detecting the change when $\Delta_n = o(n^{1/3})$. We also study change-point detection in the case where the labeled graph is observed and show that change-point detection is possible if and only if $\Delta_n \to \infty$, thereby exhibiting a strong difference between the two settings.

Contents

This chapter is based on the paper Kaddouri et al. [2025], co-authored with Elisabeth Gassiat and Zacharie Naulet and which will appear in *Bernoulli*.

5.1 Introduction

Empirical studies carried out on networks modeling different types of interactions have revealed striking similarities between them. In many situations, these networks are scale-free, i.e. their empirical degree distribution generally follows a power law. This was observed in many networks such as citation networks Barabási et al. [2002], Newman [2001], internet Faloutsos et al. [1999] and the World Wide Web Adamic and Huberman [2000]. On the other hand, the typical distances between vertices in these networks are small (see the books Watts and Strogatz [2006], Watts [1999]). This is generally referred to as the small-world phenomenon. Motivated by these observations, the preferential attachment random graph model was proposed to mathematically model scale-free networks. It provides a simple and intuitive mechanism for generating networks with a power-law degree distribution. The model helps in understanding how networks evolve over time by showing that vertices with higher degrees tend to attract more links, leading to the rich-get-richer phenomenon. This mirrors many real-world situations where popular entities tend to become even more popular over time. The first preferential attachment model to emerge was the Barabási-Albert model Barabási and Albert [1999a]. In this model, new vertices are added to the network one at a time, and each new vertex gets attached to existing vertices with a probability proportional to their current degree. In Bollobás et al. [2003], Cirkovic et al. [2023a], Gao et al. [2017], variants of this model were proposed and they depend mainly on the attachment function which can be linear, nonlinear, constant in time or time-varying. Recently, there has been notable interest in investigating time-varying networks Zhao et al. [2019], Wang et al. [2021], Medo et al. [2011], Holme and Saramäki [2019], i.e. networks where the attachment function is not constant over time. These networks usually involve a set of parameters that describe the time evolution of the network. Within this framework, an important question is to understand the effect of abrupt changes in these parameters on the degree distribution and how these changes can be detected and localized. Our work focuses on the situation where the growth dynamics of the network might undergo at most one change at some point of time. To model this, a time-inhomogeneous affine preferential attachment model is used. In this model, a new vertex entering the graph at time $t \in [\![2, n]\!]$ connects to an existing vertex with degree k with probability proportional to $f(k) = k + \delta(t)$ where $\delta(t)$ is the parameter likely to change at a given time. In particular, we are interested in late change-point detection, specifically when the change-point is given by $\tau_n = n - \Delta_n$ with $\Delta_n = o(n)$. This scenario is important for detecting changes as quickly as possible. Understanding this context will highlight the fundamental limits of change point detection and provide an estimate of the minimum number of vertices that must be

observed between the moment the change took place and the moment it is detected. In Bet et al. [2025], the authors built a test based on low degree vertices which was shown to detect the change only when $\frac{\Delta_n}{n^{1/2}} \to \infty$. They conjectured that when $\Delta_n = o(n^{1/2})$ and based only on the unlabeled graph, detection of the change is not possible. In light of this framework, this paper has two goals: (i) Prove the conjecture holds at least for $\Delta_n = o(n^{1/3})$, and (ii) Study the problem of change-point detection in the situation where the labeled graph is observed. More precisely, below is an informal statement of our main results.

Theorem 5.1.1 (Informal)**.** *Using the unlabeled preferential attachment random graph, detection of the change-point is not possible when* $\Delta_n = o(n^{1/3})$.

Theorem 5.1.2 (Informal)**.** *Using the labeled preferential attachment random graph, detection of the change-point is possible if and only if* $\Delta_n \to \infty$.

The formal statement of Theorem 5.1.1 is given later in the paper by Theorem 5.3.1, while the formal statement of Theorem 5.1.2 is given by Theorem 5.3.6.

In what follows, Section 5.2 introduces the notations and defines the model and the attachment mechanism. Section 5.3 presents the main results of the paper. Section 5.4 is devoted to the discussions and perspectives while Sections 5.5 and 5.6 detail the proofs for the unlabeled model.

5.1.1 Related work

This work is a continuation of Bet et al. [2025] where the problem of late change-point detection ($\tau_n = n - \lfloor cn^\gamma \rfloor$) was studied and a test was built for detecting the change when $\gamma \in (1/2, 1)$. The idea behind this test is that the variations in the number of vertices with minimal degree around its asymptotic value exhibit different magnitudes under the null hypothesis compared to the alternative hypothesis. They also conjectured that no test is capable of detecting the change when $\gamma \in (0, 1/2)$. In Bhamidi et al. [2018], Banerjee et al. [2023], the authors considered the problems of change-point detection and localization, but they focused mainly on the situation of early change-point, that is when changes occur at $\tau_n = \alpha n$ with $\alpha \in (0, 1)$. Detection of the change was shown to be always possible in this setting and a non-parametric consistent estimator of α was devised, allowing in addition to detection for localization of the change-point. Similarly, a likelihood-based methodology for change-point localization was proposed in Cirkovic et al. [2023b]. In Banerjee et al. [2023], a different regime of early change-detection was studied. It corresponds to the situation where $\tau_n = \lfloor cn^\gamma \rfloor$ with $\gamma \in (0, 1)$ and $c > 0$. Unlike the case of late change, the test used in this regime is based on maximal degrees. This is because, while the asymptotic degree distribution does not depend on the parameter γ, the distribution of the maximal degree does. A similar phenomenon was noted in Bubeck et al. [2015], which demonstrated that the influence of the seed graph (the initial subgraph from which the preferential attachment graph originates) persists as the number of vertices increases to infinity. In the absence of any change-point, the general problem of estimation of general attachment functions was already studied in Gao et al. [2017]. This problem reduces to a simple parametric estimation in the case of affine preferential attachment. The estimation can be done using the MLE as shown in Bet et al. [2025]. Consistency and asymptotic normality of this estimator were proved in the more general setting of random initial degrees in Gao and van der Vaart [2017]. Finally, let us mention that some ideas underlying our proof are reminiscent to the arguments employed by Briend et al. [2025] to derive minimax lower bounds for the problem of estimating the order of arrival of the vertices in the unlabeled preferential attachment tree.

5.2 Setting, definitions and notations

5.2.1 Labeled versus unlabeled graphs, structure

The preferential attachment mechanism introduced in Section 5.1 (see also next section for more details) defines a sequence of random multigraphs $(G_t)_{t\geq 1}$ on vertex sets $\{0,\dots,t\}$. There is no loss of generality in assuming that these graphs are directed, using the convention that the arrows go from vertices with largest labels to vertices with smallest labels. To be somewhat more precise, in the next, a *labeled graph* refers to the following definition:

Definition 5.2.1 (Labeled graph)**.** *A labeled (multi)graph $\mathfrak{g}$ is a couple $(\mathcal{V},\mathcal{E})$ where $\mathcal{V}$ is the set of vertices and $\mathcal{E}\subset\mathcal{V}^2$ is the multiset of directed edges, with no loop allowed. For an edge $(u,v)\in\mathcal{E}$, we use the convention that the arrow goes from u to v, and we write for simplicity $u\to_{\mathfrak{g}} v$ for $(u,v)\in\mathcal{E}$.*

Given a labeled graph $\mathfrak{g}=(\mathcal{V},\mathcal{E})$, we define for convenience $\mathsf{V}(\mathfrak{g})=\mathcal{V}$ the vertex set of $\mathfrak{g}$, and $\mathsf{E}(\mathfrak{g})=\mathcal{E}$ the edge multiset of $\mathfrak{g}$. Note that in a multigraph, two vertices can be connected by more than one edge. We count the multiplicity of edges via the function $\mu_{\mathfrak{g}}:\mathsf{V}(\mathfrak{g})^2\to\mathbb{Z}_+$, such that $\mu_{\mathfrak{g}}(u,v)=k$ means that there are k directed edges $u\to v$ in $\mathfrak{g}$ (with possibly $k=0$). The set $\mathsf{P}_{\mathfrak{g}}(u)=\{v\in\mathsf{V}(\mathfrak{g}) \ : \ v\to_{\mathfrak{g}} u\}$ are the in-neighbors of vertex $u\in\mathsf{V}(\mathfrak{g})$ (*aka.* parents) and $\mathsf{C}_{\mathfrak{g}}(u)=\{v\in\mathsf{V}(\mathfrak{g}) \ : \ u\to_{\mathfrak{g}} v\}$ are the out-neighbors of vertex $u\in\mathsf{V}(\mathfrak{g})$ (*aka.* children). The in-degree of $u\in\mathsf{V}(\mathfrak{g})$ is written $\mathsf{d}^{\text{in}}_{\mathfrak{g}}(u)=\sum_{v\in\mathsf{P}_{\mathfrak{g}}(u)}\mu_{\mathfrak{g}}(v,u)$, the out-degree is $\mathsf{d}^{\text{out}}_{\mathfrak{g}}(u)=\sum_{v\in\mathsf{C}_{\mathfrak{g}}(u)}\mu_{\mathfrak{g}}(u,v)$, and the degree is $\mathsf{d}_{\mathfrak{g}}(u)=\mathsf{d}^{\text{in}}_{\mathfrak{g}}(u)+\mathsf{d}^{\text{out}}_{\mathfrak{g}}(u)$. For a subset $S\subset\mathrm{V}(\mathfrak{g})$ we denote by $\mathfrak{g}[S]$ the induced sub(multi)graph of S, *ie.* $\mathfrak{g}[S]=(S,\mathcal{E}[S])$ where $\mathcal{E}[S]$ is the multiset obtained from $\mathsf{E}(\mathfrak{g})$ by deleting edges that have an endpoint not in S.

In order to define unlabeled graphs, we require the following definition of an isomorphism of multigraphs.

Definition 5.2.2 (Graph isomorphism)**.** *Let $\mathfrak{g}$ and $\mathfrak{g}'$ be two labeled graphs. An isomorphism ϕ between $\mathfrak{g}$ and $\mathfrak{g}'$ is a bijective map $\phi:\mathsf{V}(\mathfrak{g})\mapsto\mathsf{V}(\mathfrak{g}')$ that preserves the set of neighbors of each vertex. More precisely, for vertices $v,w\in\mathsf{V}(\mathfrak{g})$, and $k\in\mathbb{Z}_+$:*

$$\mu_{\mathfrak{g}}(u,v)=k \iff \mu_{\mathfrak{g}'}(\phi(u),\phi(w))=k.$$

In the next, $\mathfrak{g}\cong\mathfrak{g}'$ will denote the fact that $\mathfrak{g}$ and $\mathfrak{g}'$ are isomorphic, *ie.* there exists an isomorphism between $\mathfrak{g}$ and $\mathfrak{g}'$. We are now in position to define *unlabeled graphs*.

Definition 5.2.3 (Unlabeled graph)**.** *An unlabeled graph $\mathfrak{u}$ is an isomorphism class of labeled graphs (for the relation $\cong$ defined above).*

An important aspect in our work is that we consider the model where only the unlabeled version of the preferential attachment is observed; *ie.* only the *structure* of the graph is available to the statistician:

Definition 5.2.4 (Structure)**.** *Let $\mathfrak{g}$ be a labeled graph. The unlabeled graph associated to $\mathfrak{g}$, which will be denoted $s(\mathfrak{g})$, is the equivalence class of labeled graphs that are isomorphic to $\mathfrak{g}$, ie.*

$$s(\mathfrak{g})=\{\mathfrak{g}' \ : \ \mathfrak{g}'\cong\mathfrak{g}\}.$$

5.2.2 Formal statement of the problem

Using the vocabulary defined in Section 5.2.1, the preferential attachment model produces a sequence $(G_t)_{t\geq 1}$ of random labeled graphs, which we now intend to define rigorously.

Let $m \in \mathbb{N} = \{1, 2, \dots\}$ and $\delta \colon \mathbb{N} \to (-m, +\infty)$. The process $(G_t)_{t \geq 1}$ of interest is better described by introducing the intermediate process $((G_{t,i})_{i=0}^m)_{t \geq 1}$, constructed as follows. For $t = 1$ let $G_{1,0}$ be the graph consisting of two isolated vertices labeled 0 and 1. Then for $i = 1, \dots, m$, $G_{1,i}$ is obtained from $G_{1,i-1}$ by adding an edge between vertices 0 and 1. For $t \geq 2$, the sequence $(G_{t,i})_{i=0}^m$ is obtained by letting $G_{t,0}$ be the graph $G_{t-1,m}$ together with an isolated vertex with label t; and, for $i = 1, \dots, m$, $G_{t,i}$ is obtained from $G_{t,i-1}$ by adding an edge directed from t towards a randomly chosen vertex $V_{t,i}$ in $\{0, \dots, t-1\}$ sampled according to the probabilities that $V_{t,i} = v$ (conditionally to $G_{t,i-1}$) given by

$$\frac{\mathsf{d}_{G_{t,i-1}}(v) + \delta(t)}{\sum_{v'=0}^{t-1} \left(\mathsf{d}_{G_{t,i-1}}(v') + \delta(t)\right)} = \frac{\mathsf{d}_{G_{t,i-1}}(v) + \delta(t)}{2m(t-1) + \delta(t)t + (i-1)}. \tag{5.1}$$

Finally, the process $(G_t)_{t \geq 1}$ is obtained from the process $((G_{t,i})_{i=0}^m)_{t \geq 1}$ by setting $G_t = G_{t,m}$ for each $t \geq 1$. Otherwise said, the process $(G_t)_{t \geq 1}$ is obtained from the intermediate process by forgetting the order of arrivals of the m edges added at every time step $t \geq 1$.

The aim of this work is to find evidence in the preferential attachment graph that the value of δ has changed at a given time or not, using solely the information contained in the unlabeled graph $s(G_n)$ at time n. We are interested in the situation where the value of δ changes at most once. This can be formulated as a simple hypothesis testing problem:

$$(H_0) \ : \ \delta(t) = \delta_0, \qquad\qquad (H_1) \ : \ \delta(t) = \delta_0 \mathbf{1}_{t \leq \tau_n} + \delta_1 \mathbf{1}_{t > \tau_n}$$

where $1 \leq \tau_n \leq n$, $\delta_0 \in (-m, +\infty)$ and $\delta_1 \in (-m, +\infty)$ are known. As in Bet et al. [2025], we are interested only in the situation of late change-points, that is the situation where $\tau_n = n - \Delta_n$ for $\Delta_n = o(n)$. Bet et al. [2025] constructed a sequence of tests $(\phi_n)_{n \geq 1}$ with vanishing Type I and Type II errors when $\frac{\Delta_n}{n^{1/2}} \to \infty$. They conjectured that using *only the unlabeled random graph*, change point detection becomes impossible when $\Delta_n = o(n^{1/2})$. This work proves the conjecture holds at least for $\Delta_n = o(n^{1/3})$. We prove that even if the model parameters τ_n, δ_0 and δ_1 are known, detection of the change is still not possible (and hence also impossible when they are unknown).

In the sequel, for each $n \geq 1$, $(\Omega_n, \mathcal{F}_n, \mathbb{P}_0^n)$ (respectively $(\Omega_n, \mathcal{F}_n, \mathbb{P}_1^n)$) is a probability space that is rich enough to define the beginning of the sequence of intermediate graphs $((G_{t,i})_{i=0}^m)_{t=1}^n$ under the hypothesis H_0 (resp. H_1). Expectation under $\mathbb{P}_0^n$ (respectively $\mathbb{P}_1^n$) is denoted by $\mathbb{E}_0^n$ (resp. $\mathbb{E}_1^n$).

5.2.3 Further Notations

Besides the notations and conventions defined in previous sections, we make use of the following. For real numbers x, y we write $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$. For sequences of real numbers, $a_n \sim b_n$ means that a_n / b_n converges to 1, $a_n = o(b_n)$ means that a_n / b_n converges to 0, $a_n = O(b_n)$ means that a_n / b_n is asymptotically bounded, $a_n \lesssim b_n$ is equivalent to $a_n = O(b_n)$ and $a_n \asymp b_n$ means that $(\exists \alpha, \beta > 0) \ \alpha a_n \leq b_n \leq \beta a_n$. We write $\sigma(X_1, \dots, X_n)$ the σ-field generated by random variables $(X_1, \dots, X_n)$.

5.3 Main results

5.3.1 The observation is the unlabeled graph

We first consider the situation where only the unlabeled graph is observed. The following theorem establishes the conjecture in some regimes of the parameters Δ_n and (δ_0, δ_1) which are assumed to be known.

Theorem 5.3.1. *If $\delta_0 > 0$ and $\Delta_n = o(n^{1/3})$ [for $\delta_0 = 0$ and $\Delta_n = o\big(\frac{n^{1/3}}{\log(n)}\big)$], then for every sequence of events $(A_n)_{n\geq 1}$ with $A_n \in \sigma(s(G_n))$ for all $n \geq 1$,*

$$\mathbb{P}_0^n(A_n) \to 0 \implies \mathbb{P}_1^n(A_n) \to 0.$$

In other words, under the assumptions of the theorem, the laws of $(s(G_n))_{n\geq 1}$ under H_1 are *contiguous* to those under H_0. By Le Cam's first lemma [Vaart, 1998, Section 6.2], no (eventually randomized) test made on the basis of observing $s(G_n)$ is capable of controlling both Type I and Type II error rates simultaneously: if $(\phi_n)_{n\geq 1}$ is a sequence of $s(G_n)$-measurable tests such that $\mathbb{E}_0^n(\phi_n) \to 0$ then $\mathbb{E}_1^n(\phi_n) \to 0$ as well. Note that a consequence of this result is that even if the model parameters are known, detection is still not possible which is a stronger result than if the model parameters are unknown. A sketch of the proof of the theorem is given in the next section.

5.3.2 Sketch of proof of Theorem 5.3.1

Difficulties in proving contiguity

Let us for simplicity denote $Q_j^{n,s} = \mathbb{P}_j^n \circ (s \circ G_n)^{-1}$ the law of $s(G_n)$ under hypothesis H_j. The statement in Theorem 5.3.1 is equivalent to the contiguity of $(Q_1^{n,s})_{n\geq 1}$ with respect to $(Q_0^{n,s})_{n\geq 1}$. A well-known sufficient condition for establishing contiguity is that the second moment of the likelihood ratio $\frac{\mathrm{d}Q_1^{n,s}}{\mathrm{d}Q_0^{n,s}}$ remains bounded as $n \to \infty$. Understanding this likelihood ratio is, however, not a simple task. To see why, observe that for a given unlabeled graph $\mathfrak{u}_n$ on $n+1$ vertices we do have

$$\mathbb{P}_\ell^n(s(G_n) = \mathfrak{u}_n) = \sum_{\substack{\mathfrak{g} \in \mathfrak{u}_n \\ \mathsf{V}(\mathfrak{g}) = [\![0,n]\!]}} \mathbb{P}_\ell^n(G_n = \mathfrak{g}), \qquad \ell = 0, 1. \tag{5.2}$$

Though $\mathbb{P}_\ell^n(G_n = \mathfrak{g})$ is easy to evaluate when $\mathbb{P}_\ell^n(G_n = \mathfrak{g}) > 0$ (see lemmas 5.5.2 and 5.5.3), it is much more delicate for an arbitrary unlabeled graph $\mathfrak{u}_n$ to understand which of the terms in the summation of (5.2) is non-zero. Indeed, if $\mathbb{P}_\ell^n(G_n = \mathfrak{g}) > 0$ and there is an edge $u \to_{\mathfrak{g}} v$, then the graph $\mathfrak{g}'$ obtained from $\mathfrak{g}$ by swapping the labels u and v has the same structure as $\mathfrak{g}$ while $\mathbb{P}_\ell^n(G_n = \mathfrak{g}') = 0$. This is because in the preferential attachment mechanism, arrows can only go from the largest label to the smallest. So to understand the likelihood of $s(G_n)$, it is required to understand the intersection of $\mathfrak{u}_n$ with the support of the law of G_n, which turns out to be rather challenging. Instead, we prefer to reduce the problem to a simpler one, as we explain in the next section.

Problem reduction

Informally, problem reduction consists of analyzing a simpler problem where the observation is richer than the structure, but where detection is still not possible. The first natural reduction to examine is the situation where the labeled graph G_n is observed. Unfortunately, we will show in Section 5.3.3 that in this case, change detection is always possible whenever $\Delta_n \to +\infty$ and $n - \Delta_n \to +\infty$ and that this reduction is therefore useless for our proof. Consequently, We are bound to look for an intermediate problem where the observation is richer than the structure, but not as informative as the labeled graph. The main idea of the proof is that if we show that change-point detection is impossible in this (easier) problem, then this should imply that it is also impossible in the original problem where only the structure is observed. Such an intermediate problem is offered by certain random permutations of the graph. We use the following definition of a permuted graph.

Definition 5.3.2 (Permutation of a labeled graph)**.** *Let $\mathfrak{g}$ be a labeled graph and π a permutation of $\mathsf{V}(\mathfrak{g})$. We call $\pi(\mathfrak{g})$ the labeled graph obtained by the application of permutation π to the vertices of the graph $\mathfrak{g}$. In other words $\mathsf{V}(\pi(\mathfrak{g})) = \mathsf{V}(\mathfrak{g})$ and for vertices $u, v \in \mathsf{V}(\mathfrak{g})$ and $k \in \mathbb{Z}_+$*

$$\mu_{\mathfrak{g}}(u, v) = k \iff \mu_{\pi(\mathfrak{g})}(\pi(u), \pi(v)) = k.$$

One might question the reasoning behind reducing the problem to a random permutation of the graph. In fact, for $\mathfrak{g}$ a preferential attachment graph, the likelihood of $s(\mathfrak{g})$ coincides up to a multiplicative term with that of $\pi(\mathfrak{g})$ which is the graph $\mathfrak{g}$ to which a random uniform permutation π is applied. Of course reducing to a uniform permutation of the graph is helpless since it is statistically equivalent to observe the unlabeled graph, but this motivates the use of other random, simpler, permutations. In particular, our random permutation will be chosen to be uniform over a distinguished subset of vertices. Permuting part of the vertices is statistically equivalent to hiding their labels, which is the purpose sought through the reduction of the original problem.

From now on, it is assumed that the spaces $(\Omega_n, \mathcal{F}_n, \mathbb{P}_0^n)$ and $(\Omega_n, \mathcal{F}_n, \mathbb{P}_1^n)$ are rich enough to define $((G_{t,i})_{i=0}^m)_{t=1}^n$ jointly with a random permutation π_n of $[\![0, n]\!]$; the details of which are given below. The situations where one observe G_n, $\pi_n(G_n)$, or $s(G_n)$, are of increasing difficulty since one observes less and less information related to the labeled graph. Detection of the change should become more and more difficult. The following lemma confirms this insight, provided the conditional distributions of π_n given G_n are the same under $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$.

Lemma 5.3.3. *Let $\mathcal{G}_n$ denote the set of all labeled graphs on vertex set $[\![0, n]\!]$ and $\mathcal{S}_n$ denote the set of all permutations of $[\![0, n]\!]$. Suppose there is a Markov Kernel $K_n : \mathcal{G}_n \times 2^{\mathcal{S}_n} \to [0, 1]$ such that both $\mathbb{P}_0^n$ and $\mathbb{P}_1^n$ admit K_n as conditional distribution of π_n given G_n. Consider the following propositions:*

1. *For every sequence $(A_n)_{n\geq 1}$ of G_n-measurable sets, $\mathbb{P}_0^n(A_n) \to 0 \implies \mathbb{P}_1^n(A_n) \to 0$.*
2. *For every sequence $(A_n)_{n\geq 1}$ of $\pi_n(G_n)$-measurable sets, $\mathbb{P}_0^n(A_n) \to 0 \implies \mathbb{P}_1^n(A_n) \to 0$.*
3. *For every sequence $(A_n)_{n\geq 1}$ of $s(G_n)$-measurable sets, $\mathbb{P}_0^n(A_n) \to 0 \implies \mathbb{P}_1^n(A_n) \to 0$.*

Then $1 \implies 2 \implies 3$.

See Appendix 5.6.1 for the proof of Lemma 5.3.3. In what follows, we will consider the reduction where one observes $\pi_n(G_n)$ in place of $s(G_n)$. Letting $Q_0^{n,p}$ (respectively $Q_1^{n,p}$) denote the law of $\pi_n(G_n)$ under the null hypothesis (resp. the alternative hypothesis), a mere change of variable followed by an application of Cauchy-Schwarz shows that for any events $A_n, B_n \in \sigma(\pi_n(G_n))$

$$\mathbb{P}_1^n(A_n) \leq \mathbb{P}_1^n(B_n^c) + \mathbb{P}_0^n(A_n)^{1/2} \mathbb{E}_0^n \left[\left(\frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}(\pi_n(G_n)) \right)^2 \mathbf{1}_{B_n} \right]^{1/2}.$$

Hence, if we build a sequence of kernels $(K_n)_{n\geq 1}$ and events $(B_n)_{n\geq 1}$ in $\sigma(\pi_n(G_n))$ such that

$$\mathbb{P}_1^n(B_n^c) \to 0, \qquad \text{and,} \qquad \limsup_{n\to\infty} \mathbb{E}_0^n \left[\left(\frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}(\pi_n(G_n)) \right)^2 \mathbf{1}_{B_n} \right] < +\infty,$$

then 2 of Lemma 5.3.3 holds, which by said lemma implies the validity of our theorem. We build $(K_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ in the next section.

Construction of the Markov kernel K_n and the event B_n

We first remark that, when building $K_n(\mathfrak{g}_n, \cdot)$, it is enough to consider $\mathfrak{g}_n$ in $\mathfrak{S}_n = \{\mathfrak{g}'_n \ : \ \mathbb{P}_0^n(G_n = \mathfrak{g}'_n) \neq 0\} = \{\mathfrak{g}'_n \ : \ \mathbb{P}_1^n(G_n = \mathfrak{g}'_n) \neq 0\}$. We give a characterization of the set $\mathfrak{S}_n$ in Lemma 5.5.1. Remark that all graphs in $\mathfrak{S}_n$ have vertex set $[\![0, n]\!]$.

To construct K_n and B_n, we first define the following set of vertices of a labeled graph $\mathfrak{g}_n$, which corresponds to the vertices illustrated in *bold* in Figure 5.1:

$$\tilde{\mathcal{V}}(\mathfrak{g}_n) = \Big\{ v \in [\![\tau'_n + 1, n]\!] \ : \ \mathsf{d}_{\mathfrak{g}_n}(v) = m, \ \forall w \in \mathsf{C}_{\mathfrak{g}_n}(v) \ w \leq \tau'_n \text{ and } \mathsf{P}_{\mathfrak{g}_n}(w) \backslash \{v\} \subset [\![0, \tau'_n]\!] \Big\}.$$

In the previous definition $(\tau'_n)_{n \geq 1}$ is a sequence of integer numbers to be chosen accordingly later, but satisfying $0 \leq \tau'_n < \tau_n$. In other words, $\tilde{\mathcal{V}}(G_n)$ contains the late vertices of G_n which have minimal degree and are the unique late parent of their children. We then consider permutations which leave invariant the labels not in $\tilde{\mathcal{V}}(G_n)$; *ie.* letting $\mathcal{S}_n$ the set of all permutations of $[\![0, n]\!]$ we define

$$\Pi_n(\mathfrak{g}_n) = \Big\{ \pi \in \mathcal{S}_n \ : \ \forall i \notin \tilde{\mathcal{V}}(\mathfrak{g}_n), \ \pi(i) = i \Big\},$$

and we define $K(\mathfrak{g}_n, \cdot)$ as the uniform distribution over $\Pi_n(\mathfrak{g}_n)$, for all $\mathfrak{g}_n \in \mathfrak{S}_n$. We note that one of the advantages of this permutation scheme is that $\pi(\mathfrak{g}_n) \in \mathfrak{S}_n$ for any $\mathfrak{g}_n \in \mathfrak{S}_n$ and $\pi \in \Pi_n(\mathfrak{g}_n)$ (see Lemma 5.6.2), which precludes the issues mentioned in Section 5.3.2. Then, we consider the sequence of events

$$B_n = \Big\{ |\tilde{\mathcal{V}}(G_n)| \geq \Delta'_n \Big(1 - \frac{\alpha_n \Delta'_n}{\tau'_n}\Big), \quad [\![\tau_n + 1, n]\!] \subset \tilde{\mathcal{V}}(G_n) \Big\}$$

for a sequence $(\alpha_n)_{n \geq 1}$ diverging slowly to infinity, and where $\Delta'_n = n - \tau'_n$. We note that by construction $\tilde{\mathcal{V}}(G_n) = \tilde{\mathcal{V}}(\pi_n(G_n))$, so that $B_n \in \sigma(\pi_n(G_n))$ as required (see previous section). On the event B_n all the late vertices of G_n are eventually permuted and are indistinguishable from the earlier vertices in $\tilde{\mathcal{V}}(G_n)$. This informally tells why the change-point cannot be detected. More formally, the Theorem 5.3.1 is an immediate consequence of the two following propositions, choosing $\Delta'_n \asymp n^{2/3}$ (implying $\tau'_n \sim n$), $\Delta_n = O\big(\frac{n^{1/3}}{\alpha_n}\big)$, and $\alpha_n \to \infty$ arbitrarily slowly if $\delta_0 > 0$ or $\frac{\alpha_n}{\log(n)} \to \infty$ arbitrarily slowly if $\delta_0 = 0$.

Proposition 5.3.4. *There exist constants $c_1, c_2 > 0$ depending only on δ_0, δ_1, and m, such that for all $n \geq 4$, if $3 \leq \tau'_n < \tau_n$, $\Delta'_n \leq \tau_n$, $\frac{\alpha_n \Delta'_n}{\tau'_n} \leq \frac{1}{2}$ and $\frac{\Delta_n}{\Delta'_n} \leq \frac{1}{4}$ then*

$$\log \mathbb{E}_0^n \left[\left(\frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}(\pi_n(G_n)) \right)^2 \mathbf{1}_{B_n} \right] \leq \frac{4\alpha_n \Delta_n \Delta'_n}{\tau'_n} + \frac{22 m \Delta_n^2}{\Delta'_n} + \frac{2}{3\Delta'_n} + \sqrt{\frac{c_1 \Delta_n^2}{\Delta'_n}} e^{\frac{c_2 \Delta_n^2}{\Delta'_n}}.$$

See Appendix 5.6.2 for the proof of Proposition 5.3.4.

Proposition 5.3.5. *There exists a constant $C > 0$ depending only on δ_0, δ_1, and m, such that for all $2 \leq \tau'_n \leq n$*

$$\mathbb{P}_1^n(B_n^c) \leq \frac{C}{\alpha_n} \Big(1 + \frac{\alpha_n \Delta_n \Delta'_n}{\tau'_n}\Big) \cdot \begin{cases} \log(\tau'_n) & \text{if } \delta_0 = 0, \\ 1 & \text{if } \delta_0 > 0. \end{cases}$$

See Appendix 5.6.3 for the proof of Proposition 5.3.5.

We believe the Theorem 5.3.1 can not be improved beyond $\Delta_n = o(n^{1/3})$ using the permutation π_n we designed. In fact, we think that the upper bound in Proposition 5.3.4 is essentially tight and could only be made bounded by choosing a smaller event B_n if Δ_n is made larger than $o(n^{1/3})$. However, using a smaller event B_n makes unlikely that a result such as Proposition 5.3.5 hold.

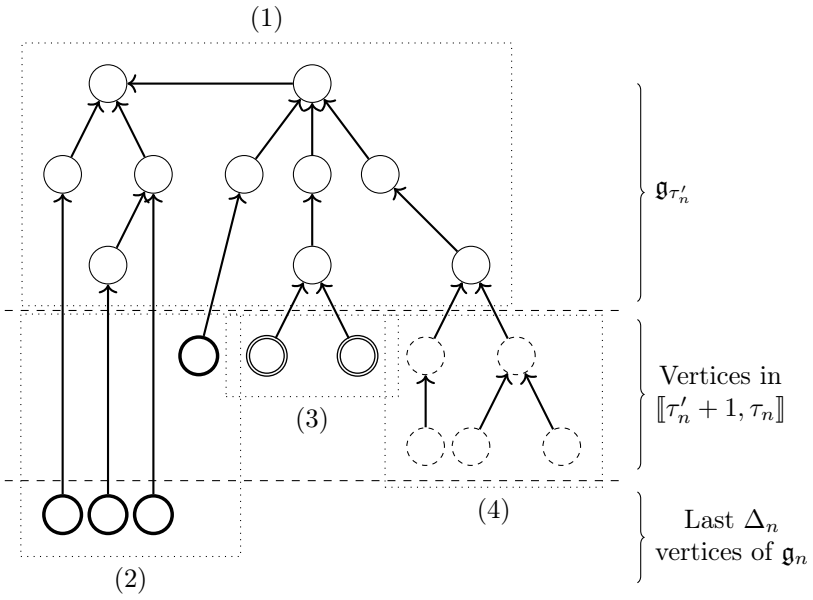


Figure 5.1: Typical preferential attachment graph $\mathfrak{g}_n$ with $m = 1$ when $\Delta_n = o(n^{1/3})$. Four types of vertices emerge: normal vertices (1), bold vertices (2), double circle vertices (3) and dotted vertices (4). Our random permutation π_n is built to permute only vertices represented in bold.

5.3.3 The observation is the labeled graph

We consider now the model where the observation is the labeled graph G_n. The main purpose of this section is to emphasize the difference between the labeled and the unlabled model, by showing that in the labeled model the change-point can be detected as soon as $\Delta_n \to \infty$; in contrast with the unlabeled model for which $\frac{\Delta_n}{n^{1/2}} \to \infty$ is sufficient by Bet et al. [2025] and $\frac{\Delta_n}{n^{1/3}} \to \infty$ is necessary by our previous result. This also shows that a reduction scheme to a problem where the labeled graph undergoes a transformation is unavoidable to obtain a non trivial lower bound in the unlabeled model.

We first assume that the model parameters (δ_0, δ_1) and τ_n are known to be consistent with our Theorem 5.3.1. We then state additional results covering the case where (δ_0, δ_1) are unknown as well as the localization of the change-point (*ie.* estimating τ_n). In particular these additional results show that not knowing the parameters does not affect the capability of detecting the change-point as soon as $\Delta_n \to \infty$.

Theorem 5.3.6. *Let $Q_0^n = \mathbb{P}_0^n(G_n \in \cdot)$ and $Q_1^n = \mathbb{P}_1^n(G_n \in \cdot)$. If $\tau_n \to \infty$ and $\Delta_n \to \infty$, detection of the change is possible: the likelihood-ratio test $T_n = \mathbf{1}\big(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n) > 1\big)$ satisfies*

$$\mathbb{E}_0^n(T_n) + \mathbb{E}_1^n(1 - T_n) \to 0.$$

When $\limsup_{n\to\infty} \Delta_n < +\infty$ detection of the change is not possible: $(Q_1^n)_{n\geq 1}$ is contiguous with respect to $(Q_0^n)_{n\geq 1}$.

See Section 5.7.2 for the proof of Theorem 5.3.6. Observe that Theorem 5.3.6 identifies the exact phase transition for detection when the labeled graph is observed and the model parameters are known.

Maximum Likelihood Estimation of (δ_0, δ_1)

In Gao and van der Vaart [2017], the estimation of δ_0 was done under the null hypothesis. We now investigate the estimation of δ_0 and δ_1 in the model where there is a change-point from δ_0 to δ_1 at instant $\tau_n = n - \Delta_n$. Here τ_n is assumed to be known. As shown in the expression of the likelihood in Lemma A.3, the likelihood factorizes in two parts each of those involving only δ_0 or δ_1; *ie.* letting (here $G_{\tau_n} = G_n[\![0, \tau_n]\!]$)

$$\ell_{1:\tau_n}(\delta_0) = \log\left(\frac{\prod_{k=m}^{\tau_n m}(k+\delta)^{N_{>k}(G_{\tau_n})}}{\prod_{t=2}^{\tau_n}\prod_{i=1}^{m}\left[(2m+\delta_0)\,t - 2m + i - 1\right]}\right)$$
$$\ell_{\tau_n+1:n}(\delta_1) = \log\left(\frac{\prod_{k=m}^{nm}(k+\delta_1)^{N_{>k}(G_n)-N_{>k}(G_{\tau_n})}}{\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\left[(2m+\delta_1)\,t - 2m + i - 1\right]}\right)$$

the log-likelihood of (δ_0, δ_1) writes as $\ell_{1:\tau_n}(\delta_0) + \ell_{\tau_n+1:n}(\delta_1)$. Then, building on the work of Gao and van der Vaart [2017] in the no change-point model, we obtain in the next theorem the asymptotic normality of the MLE in the model with a change-point.

As it will be useful in the next, we recall the expression of the limiting degree distribution of the affine preferential attachment model with parameter δ (see [van der Hofstad, 2016, Sections 8.6.1 and 8.6.2] for details):

$$p_k(\delta) = (2 + \delta/m)\frac{\Gamma(k+\delta)\Gamma(m+2+\delta+\delta/m)}{\Gamma(m+\delta)\Gamma(k+3+\delta+\delta/m)}. \tag{5.3}$$

Theorem 5.3.7. *For all $(\delta_0, \delta_1) \in (-m, \infty)^2$, if $\tau_n \to \infty$ and $\Delta_n \to \infty$, then $(\delta, \delta') \mapsto \ell_{1:\tau_n}(\delta) + \ell_{\tau_n+1:n}(\delta')$ has a unique maximizer $(\hat{\delta}_{0,n}, \hat{\delta}_{1,n})$ with probability going to one under $(\mathbb{P}_1^n)_{n\geq 1}$, and*

$$\begin{pmatrix}\sqrt{\tau_n} & 0\\ 0 & \sqrt{\Delta_n}\end{pmatrix}\begin{pmatrix}\hat{\delta}_{0,n} - \delta_0\\ \hat{\delta}_{1,n} - \delta_1\end{pmatrix} \overset{\mathbb{P}_1^n}{\rightsquigarrow} \mathcal{N}\left(0, \begin{pmatrix}\nu_0 & 0\\ 0 & \nu_1\end{pmatrix}^{-1}\right)$$

where $\rightsquigarrow$ stands for convergence in distribution under $(\mathbb{P}_1^n)_{n\geq 1}$ and where for $j = 0, 1$

$$\nu_j = \frac{m}{2m+\delta_j}\left(\sum_{k=m}^{\infty}\frac{p_k(\delta_0)}{k+\delta_j} - \frac{1}{2m+\delta_j}\right).$$

See Section 5.7.3 for the proof of Theorem 5.3.7. Remark that in Theorem 5.3.7 we do not require that the MLE is restricted to a compact set as in Gao and van der Vaart [2017]. This condition was imposed by Gao and van der Vaart [2017] to avoid issues in controlling the score function near the boundary $-m$. Here we circumvent this issue by showing that the score cannot have a zero close to the boundary (and hence the likelihood a maximum).

Change-point detection when only τ_n is known

We now consider the situation where the change-point τ_n is known but the model parameters δ_0 and δ_1 are unknown. Theorem 3.6 suggests that change-point detection is possible when Δ_n diverges to $+\infty$ and that the likelihood ratio test guarantees that type I and type II error rates decay to 0. However, it requires the knowledge of the parameters δ_0 and δ_1 and τ_n. If these two parameters are unknown in advance, we can always try to estimate them and then consider the likelihood-ratio test with plugin estimates of δ_0 and δ_1. One can use the Maximum Likelihood Estimator (MLE) of (δ_0, δ_1) derived in the previous section.

In the next we let $Q^n_{(\tau_n,\delta_0,\delta_1)}$ be the distribution of the preferential attachment graph G_n when $\delta(t) = \delta_0 \mathbf{1}_{t\leq\tau_n} + \delta_1 \mathbf{1}_{t>\tau_n}$. The following theorem shows that when the model parameters (δ_0, δ_1) are unknown, the test

$$T'_n = \mathbf{1}\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_{0,n},\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_{0,n},\hat{\delta}_{0,n})}}(G_n) > 1\right)$$

using the MLE $(\hat{\delta}_{0,n}, \hat{\delta}_{1,n})$ is ensured to have vanishing error rates. In other words plug-in estimates of the parameters δ_0 and δ_1 allow to mimic the asymptotic behavior of the likelihood ratio test.

Theorem 5.3.8. *For every increasing sequence τ_n such that $\tau_n \to \infty$ and $n - \tau_n = \Delta_n \to \infty$, detection of the change is possible using the test T'_n:*

$$\mathbb{E}^n_0(T'_n) + \mathbb{E}^n_1(1 - T'_n) \to 0.$$

See Section 5.7.4 for the proof of Theorem 5.3.8.

5.3.4 Localization of τ_n

Finally, we consider the situation where parameters δ_0 and δ_1 are known while τ_n is unknown. The purpose is to localize the parameter τ_n. The following proposition shows that τ_n can be localized with an error of order $O\big(\log(n)^3\big)$ using the maximum likelihood estimator

$$\hat{\tau}_n \in \operatorname*{arg\,max}_{\tau\in[\![0,n]\!]} Q_{(\tau_n,\delta_0,\delta_1)}(\{G_n\}).$$

Proposition 5.3.9. *For $C > 0$ a large enough constant,*

$$\mathbb{P}^n_1\left(|\hat{\tau}_n - \tau_n| \leq C\log(n)^3\right) \to 1.$$

See Section 5.7.5 for the proof of Proposition 5.3.9.

Finally, let us mention that we were essentially interested in the model where the unlabeled graph is observed. The case where the labeled graph is observed has only been studied to justify the choice of the reduction of the original problem. We were not interested in obtaining the sharpest results in the labeled graph model. For example, we believe that the result of Proposition 5.3.9 can be generalized to include the simultaneous localisation of all model parameters δ_0, δ_1 and τ_n, and eventually reduced from $\log(n)^3$ to constant, but at the cost of some tedious calculations that are outside the scope of this paper.

5.4 Discussions and perspectives

While the original conjecture in Bet et al. [2025] had two parts, one concerning the impossibility of detection using the sequence of degrees and the other concerning the impossibility of detection using the unlabeled graph, our work focuses only on the second part, which is more general as it implies the impossibility of detection using the degrees. Although we believe the conjecture to be true, our proof of Theorem 5.3.1 does not cover all the regimes of the conjecture in terms of Δ_n and (δ_0, δ_1). As explained in Section 5.3.2, the main step of our proof of Theorem 5.3.1 resides in showing that the second moment of the likelihood-ratio of the permuted graph is bounded by an absolute constant. We were able to exhibit such a bound only in the regime where $\Delta_n = o(n^{1/3})$ and $\delta_0 \geq 0$. To put it simply, our proof works when all the last Δ_n vertices are in $\tilde{\mathcal{V}}(G_n)$ (*ie.* bold in

the Figure 5.1): the expression of the likelihood-ratio is easier to handle in this case and its second moment can be bounded by an absolute constant. This situation is illustrated in Figure 5.1 by a typical example. However, in the regime $n^{1/3} \lesssim \Delta_n \lesssim n^{1/2}$ and as illustrated in Figure 5.2, "double circle" and "dotted vertices" start appearing amongst the last Δ_n vertices, making it more difficult to choose an appropriate permutation. If we keep the same permutation (the one modifying only the labels of bold vertices) in the regime $n^{1/3} \lesssim \Delta_n \lesssim n^{1/2}$, the labels of the "dotted" and "double circle" vertices appearing amongst the last Δ_n vertices will be kept invariant and the second moment of the likelihood-ratio will diverge to infinity. One possible way of generalizing the result to the remaining regime is to construct a permutation that modifies the labels of almost $O(n)$ vertices, including all the last Δ_n vertices, while at the same time still be able to uniformly bound the second moment of the likelihood-ratio. There is a trade-off between the complexity of the chosen permutation (how many labels are modified and how they are modified) and the ease in bounding the second moment of the likelihood-ratio. For a similar reason, the regime $\delta_0 < 0$ was not covered in the proof. The main shortcoming of our proof is that we choose permutations that modify only labels in $\tilde{\mathcal{V}}(G_n)$. The reason behind this choice is that given a preferential attachment random graph, every permutation affecting only $\tilde{\mathcal{V}}(G_n)$ results in a labeled graph having positive probability under preferential attachment (Lemma 5.6.2). This facilitates the explicit writing of the likelihood-ratio. However, if we were to allow the permutations to modify the labels of "dotted" and "double circle" vertices, then one needs to be much more careful to ensure that after the application of the permutation, the labeled graph still has positive probability under preferential attachment; or find another way to circumvent the issues discussed in Section 5.3.2.

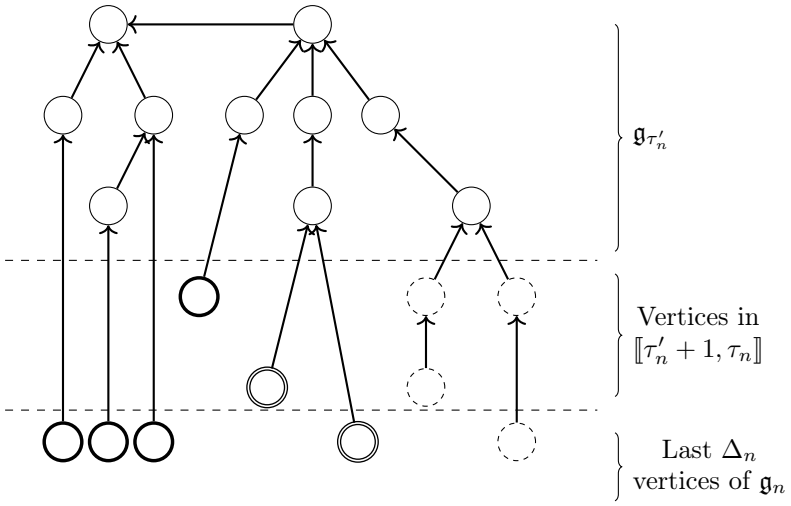


Figure 5.2: Typical preferential attachment graph $\mathfrak{g}_n$with $m = 1$ when $n^{1/3} \lesssim \Delta_n \lesssim n^{1/2}$.

5.5 Proof elements common to both labeled and unlabeled graphs

5.5.1 A result on the support of the general preferential attachment model

Anticipating that we will need to compute the likelihood under both the null hypothesis and the alternative, we first derive the likelihood in the most general case of a Preferential Attachment Model (PAM) with an arbitrary parameter function $\delta_n : \mathbb{N} \to (-m, +\infty)$,

which is allowed to change with n. We let $\mathbb{F}^n_{\delta_n}$ denote the distribution of a partial sequence of random graph $(G_0, G_1, \dots, G_n)$ distributed according to the PAM with parameter δ_n.

Lemma 5.5.1. *For $n \geq 0$, let*

$$\mathfrak{S}_n = \Big\{ \mathfrak{g}_n \ : \ \mathsf{V}(\mathfrak{g}_n) = [\![0,n]\!],\ \mathsf{C}_{\mathfrak{g}_n}(0) = \varnothing,\ \forall v \in [\![1,n]\!]\ \mathsf{C}_{\mathfrak{g}_n}(v) \subset [\![0, v-1]\!] \text{ and } \mathsf{d}^{\mathrm{out}}_{\mathfrak{g}_n}(v) = m \Big\}.$$

Then $\mathbb{F}^n_{\delta_n}(G_n = \mathfrak{g}_n) > 0 \iff \mathfrak{g}_n \in \mathfrak{S}_n$. Furthermore, for any $\mathfrak{g}_n \in \mathfrak{S}_n$,

$$\mathbb{F}^n_{\delta_n}(G_n = \mathfrak{g}_n) = C(\mathfrak{g}_n) \frac{\prod_{j=2}^n \prod_{w \in \mathsf{C}_{\mathfrak{g}_n}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_n}(j,w)} \big(\mathsf{d}_{\mathfrak{g}_n[\![0,j-1]\!]}(w) + k - 1 + \delta_n(j) \big)}{\prod_{j=2}^n \prod_{i=1}^m S_{j,i-1}(\delta_n(j))}$$

where $C(\mathfrak{g}_n) = \frac{(m!)^{n-1}}{\prod_{j=2}^n \prod_{w \in \mathsf{C}_{\mathfrak{g}_n}(j)} \mu_{\mathfrak{g}_n}(j,w)!}$ and $S_{j,i-1}(\delta) = 2m(j-1) + i - 1 + \delta j$ (defined as in Gao and van der Vaart [2017]).

Proof. Suppose $n \geq 2$, otherwise the result is trivial. By construction $\mathbb{F}^n_{\delta_n}((G_0, G_1) = (\mathfrak{g}_0, \mathfrak{g}_1)) = 1$ iff $\mathfrak{g}_0$ is the labeled graph with a unique vertex with label zero and no edge, and $\mathfrak{g}_1$ is the graph with two vertices zero and one with m edges going from one to zero. Let $1 \leq j \leq n$ and suppose that

$$\forall k \in [\![0,j]\!],\ \mathfrak{g}_k \in \mathfrak{S}_k \text{ and } \mathfrak{g}_k[\![0,k-1]\!] = \mathfrak{g}_{k-1} \iff \mathbb{F}^n_{\delta_n}\big((G_0, \dots, G_j) = (\mathfrak{g}_0, \dots, \mathfrak{g}_j)\big) > 0, \tag{5.4}$$

which has been shown to be verified for $j = 1$. The graph G_j is obtained from G_{j-1} by sampling m edges according to the PA rule. In other word

$$\mathbb{F}^n_{\delta_n}((G_0, \dots, G_j) = (\mathfrak{g}_0, \dots, \mathfrak{g}_j)) = \mathbb{F}^n_{\delta_n}((G_0, \dots, G_{j-1}) = (\mathfrak{g}_0, \dots, \mathfrak{g}_{j-1})) K_{\delta_n, j}(\mathfrak{g}_j \mid \mathfrak{g}_{j-1})$$

for a Markov kernel $K_{\delta_n,j}(\mathfrak{g}_j \mid \mathfrak{g}_{j-1})$ that assigns non-zero probability to $\mathfrak{g}_j$ iff $\mathsf{V}(\mathfrak{g}_j) = [\![0,j]\!]$ and $\mathfrak{g}_j[\![0,j-1]\!] = \mathfrak{g}_{j-1}$ and $\mathsf{d}^{\mathrm{out}}_{\mathfrak{g}_j}(j) = m$ and $\mathsf{C}_{g_j}(j) \subset [\![0,j-1]\!]$. By induction (5.4) is then verified for all $1 \leq j \leq n$. Observe that (5.4) implies that the law of $(G_0, \dots, G_n)$ is entirely determined by G_n since it must be that $G_k = G_n[\![0,k]\!]$ $\mathbb{F}^n_{\delta_n}$-almost-surely for all $k \in [\![0,n]\!]$.

Next, let $(\mathfrak{g}_{j-1}, \mathfrak{g}_j) \in \mathfrak{S}_{j-1} \times \mathfrak{S}_j$ with $\mathfrak{g}_{j-1} = \mathfrak{g}_j[\![0,j-1]\!]$. A rapid computation using equation (5.1) shows that if we enumerate $v_1 < \dots < v_\ell$ the elements of $\mathsf{C}_{\mathfrak{g}_j}(j)$ and denote by $\mu_1, \dots, \mu_\ell$ the associated edge multiplicities:

$$\begin{aligned} K_{\delta_n,j}(\mathfrak{g}_j \mid \mathfrak{g}_{j-1}) &= \sum_{(e_1,\dots,e_m)} \prod_{i=1}^m \frac{\mathsf{d}_{\mathfrak{g}_{j-1}}(v_{e_i}) + \sum_{1 \leq k < i} \mathbf{1}_{e_k = v_{e_i}} + \delta_n(j)}{\sum_{w=0}^{j-1} \big(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + \sum_{1 \leq k < i} \mathbf{1}_{e_k = w} + \delta_n(j) \big)} \\ &= \sum_{(e_1,\dots,e_m)} \frac{\prod_{w \in \mathsf{C}_{\mathfrak{g}_j}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_j}(j,w)} \big(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + k - 1 + \delta_n(j) \big)}{\prod_{i=1}^m \sum_{w=0}^{j-1} \big(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + \sum_{1 \leq k < i} \mathbf{1}_{e_k = w} + \delta_n(j) \big)} \end{aligned}$$

where the summation over $(e_1, \dots, e_m)$ is understood over sequences in $[\![1,\ell]\!]^m$ with μ_k elements equal to k for each $k = 1, \dots, \ell$ (*ie.* over all the possible ways of assigning the m edges to the ℓ children with the multiplicity constraint taken into account). We observe that exactly m edges are added at each step of the construction, so

$$\sum_{w=0}^{j-1} \Big(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + \sum_{1 \leq k < i} \mathbf{1}_{e_k = w} + \delta_n(j) \Big) = 2m(j-1) + i - 1 + j\delta_n(j) = S_{j,i-1}(\delta_n(j)).$$

It follows that

$$K_{\delta_n,j}(\mathfrak{g}_j \mid \mathfrak{g}_{j-1}) = \frac{m!}{\prod_{w\in \mathsf{C}_{\mathfrak{g}_j}(j)} \mu_{\mathfrak{g}_j}(j,w)!} \frac{\prod_{w\in \mathsf{C}_{\mathfrak{g}_j}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_j}(j,w)} \left(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + k - 1 + \delta_n(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta_n(j))}.$$

Consequently, for all $\mathfrak{g}_n \in \mathfrak{S}_n$, writing abusively $\mathfrak{g}_j = \mathfrak{g}_n[\![0,j]\!]$ (which is justified by the above discussion),

$$\begin{aligned}\mathbb{F}^n_{\delta_n}(G_n = \mathfrak{g}_n) &= \prod_{j=2}^{n} \frac{m!}{\prod_{w\in \mathsf{C}_{\mathfrak{g}_j}(j)} \mu_{\mathfrak{g}_j}(j,w)!} \frac{\prod_{w\in \mathsf{C}_{\mathfrak{g}_j}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_j}(j,w)} \left(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + k - 1 + \delta_n(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta_n(j))} \\ &= C(\mathfrak{g}_n) \frac{\prod_{j=2}^{n} \prod_{w\in \mathsf{C}_{\mathfrak{g}_n}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_n}(j,w)} \left(\mathsf{d}_{\mathfrak{g}_{j-1}}(w) + k - 1 + \delta_n(j)\right)}{\prod_{j=2}^{n} \prod_{i=1}^{m} S_{j,i-1}(\delta_n(j))}.\end{aligned}$$

This concludes the proof. □

5.5.2 The likelihood of a labeled graph under the null and the alternative hypotheses

In this section we compute the likelihood of the labeled graph under the null hypothesis (Lemma 5.5.2), under the alternative hypothesis (Lemma 5.5.3), as well as the likelihood-ratio (Lemma 5.5.4).

Lemma 5.5.2. *Let $\mathfrak{S}_n$, $C(\mathfrak{g}_n)$, and $S_{t,i-1}(\delta)$ as defined in Lemma 5.5.1. Then for all $\mathfrak{g}_n \in \mathfrak{S}_n$*

$$\mathbb{P}^n_0\left(G_n = \mathfrak{g}_n\right) = C(\mathfrak{g}_n) \frac{\prod_{v=0}^{n-1} \prod_{k=0}^{\mathsf{d}^{\mathrm{in}}_{\mathfrak{g}_n}(v)-1} (m + \delta_0 + k)}{\prod_{t=2}^{n} \prod_{i=1}^{m} S_{t,i-1}(\delta_0)} = C(\mathfrak{g}_n) \frac{\prod_{k=m}^{nm} (k+\delta_0)^{N_{>k}(\mathfrak{g}_n)}}{\prod_{t=2}^{n} \prod_{i=1}^{m} S_{t,i-1}(\delta_0)}$$

where $N_{>k}(\mathfrak{g}_n)$ is the number of vertices in $\mathfrak{g}_n$ which have degree strictly greater than k.

Proof. The first expression comes from swapping the product over parents and children in the expression given in Lemma 5.5.1 and using that the parameter is constant over time:

$$\begin{aligned}\mathbb{P}^n_0(G_n = \mathfrak{g}_n) &= C(\mathfrak{g}_n) \frac{\prod_{j=2}^{n} \prod_{w\in \mathsf{C}_{\mathfrak{g}_n}(j)} \prod_{k=1}^{\mu_{\mathfrak{g}_n}(j,w)} \left(\mathsf{d}_{\mathfrak{g}_n[\![0,j-1]\!]}(w) + k - 1 + \delta_0\right)}{\prod_{j=2}^{n} \prod_{i=1}^{m} S_{j,i-1}(\delta_0)} \\ &= C(\mathfrak{g}_n) \frac{\prod_{t=0}^{n-1} \prod_{s\in \mathsf{P}_{\mathfrak{g}_n}(t)} \prod_{k=1}^{\mu_{\mathfrak{g}_n}(s,t)} \left(\mathsf{d}_{\mathfrak{g}_n[\![0,s-1]\!]}(t) + k - 1 + \delta_0\right)}{\prod_{j=2}^{n} \prod_{i=1}^{m} S_{j,i-1}(\delta_n(s))}.\end{aligned}$$

Now for each vertex t contributing to the above product, order its parents in increasing time of arrivals and see that the product over s and k is in fact equal to $(m+\delta_0)(m+1+\delta_0)\ldots(m+\mathsf{d}^{\mathrm{in}}_{\mathfrak{g}_n}(t)-1+\delta_0)$. Thus,

$$\mathbb{P}^n_0(G_n = \mathfrak{g}_n) = C(\mathfrak{g}_n) \frac{\prod_{v=0}^{n-1} \prod_{k=0}^{\mathsf{d}^{\mathrm{in}}_{\mathfrak{g}_n}(v)-1} (m + \delta_0 + k)}{\prod_{t=2}^{n} \prod_{i=1}^{m} S_{t,i-1}(\delta_0)}$$

which is the first expression in the statement of the Lemma. For the second expression,

notice that

$$\begin{aligned}\prod_{v=0}^{n-1}\prod_{k=0}^{\mathsf{d}^{\text{in}}_{\mathfrak{g}_n}(v)-1}(m+\delta_0+k) &= \prod_{v=0}^{n-1}\prod_{k=0}^{(n-1)m}(m+\delta_0+k)\mathbf{1}_{k\leq \mathsf{d}^{\text{in}}_{g_n}(v)-1}\\ &= \prod_{k=0}^{(n-1)m}(m+\delta_0+k)^{\sum_{v=0}^{n-1}\mathbf{1}_{\mathsf{d}^{\text{in}}_{\mathfrak{g}_n}(v)>k}}\\ &= \prod_{k=m}^{nm}(k+\delta_0)^{\sum_{v=0}^{n-1}\mathbf{1}_{\mathsf{d}^{\text{in}}_{\mathfrak{g}_n}(v)+m>k}}\\ &= \prod_{k=m}^{nm}(k+\delta_0)^{\sum_{v=0}^{n-1}\mathbf{1}_{\mathsf{d}_{\mathfrak{g}_n}(v)>k}}.\end{aligned}$$

Hence the result. □

Note that under the null hypothesis, the likelihood of the graph does not depend on the labels of the vertices. It depends only on the structure $s(\mathfrak{g}_n)$ since $N_{>k}(\cdot)$ is constant over $s(\mathfrak{g}_n)$.

Lemma 5.5.3. *Let $\mathfrak{S}_n$, $C(\mathfrak{g}_n)$, and $S_{t,i-1}(\delta)$ as defined in Lemma 5.5.1. Also define $H^{\leq\tau_n}_{\mathfrak{g}_n}(v)=\sum_{u\in\mathsf{P}_{\mathfrak{g}_n}(v)}\mu_{\mathfrak{g}_n}(u,v)\mathbf{1}_{u\leq\tau_n}$ and $H^{>\tau_n}_{\mathfrak{g}_n}(v)=\sum_{u\in\mathsf{P}_{\mathfrak{g}_n}(v)}\mu_{\mathfrak{g}_n}(u,v)\mathbf{1}_{u>\tau_n}$. Then for all $\mathfrak{g}_n\in\mathfrak{S}_n$*

$$\begin{aligned}\mathbb{P}^n_1(G_n=\mathfrak{g}_n) &= C(\mathfrak{g}_n)\frac{\prod_{v=0}^{n-1}\left[\prod_{k=0}^{H^{\leq\tau_n}_{\mathfrak{g}_n}(v)-1}(m+\delta_0+k)\prod_{k=H^{\leq\tau_n}_{\mathfrak{g}_n}(v))}^{H^{\leq\tau_n}_{\mathfrak{g}_n}(v))+H^{>\tau_n}_{\mathfrak{g}_n}(v)-1}(m+\delta_1+k)\right]}{\prod_{t=2}^{n}\prod_{i=1}^{m}S_{t,i-1}(\delta(t))}\\ &= C(\mathfrak{g}_n)\frac{\prod_{k=m}^{nm}(k+\delta_0)^{N_{>k}(\mathfrak{g}_{\tau_n})}}{\prod_{t=2}^{\tau_n}\prod_{i=1}^{m}S_{t,i-1}(\delta_0)}\frac{\prod_{k=m}^{nm}(k+\delta_1)^{N_{>k}(\mathfrak{g}_n)-N_{>k}(\mathfrak{g}_{\tau_n})}}{\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}S_{t,i-1}(\delta_1)}\end{aligned}$$

with $\mathfrak{g}_{\tau_n}=\mathfrak{g}_n[[0,\tau_n]]$ and $\delta(t)=\delta_0\mathbf{1}_{t\leq\tau_n}+\delta_1\mathbf{1}_{t>\tau_n}$.

Proof. The first expression comes from the Lemma 5.5.1 and using the same arguments as in Lemma 5.5.2. For second expression, notice that

$$\begin{aligned}\prod_{v=0}^{n-1}\left[\prod_{k=0}^{H^{\leq\tau_n}(v)-1}(m+\delta_0+k)\prod_{k=H^{\leq\tau_n}(v)}^{H(v)-1}(m+\delta_1+k)\right] &= \prod_{v=0}^{n-1}\prod_{k=0}^{H^{\leq\tau_n}(v)-1}(m+\delta_0+k)\frac{\prod_{k=0}^{\mathsf{d}^{\text{out}}_{\mathfrak{g}_n}(v)-1}(m+\delta_1+k)}{\prod_{k=0}^{H^{\leq\tau_n}(v)-1}(m+\delta_1+k)}\\ &= \prod_{v=0}^{n-1}\left(\prod_{k=0}^{H^{\leq\tau_n}(v)-1}\frac{m+\delta_0+k}{m+\delta_1+k}\prod_{k=0}^{\mathsf{d}^{\text{out}}_{\mathfrak{g}_n}(v)-1}(m+\delta_1+k)\right)\\ &= \prod_{k=m}^{nm}\left(\frac{k+\delta_0}{k+\delta_1}\right)^{N_{>k}(\mathfrak{g}_{\tau_n})}(k+\delta_1)^{N_{>k}(\mathfrak{g}_n)}\\ &= \prod_{k=m}^{nm}(k+\delta_0)^{N_{>k}(\mathfrak{g}_{\tau_n})}(k+\delta_1)^{N_{>k}(\mathfrak{g}_n)-N_{>k}(\mathfrak{g}_{\tau_n})}\end{aligned}$$

which concludes the proof. □

Lemma 5.5.4. *Let $Q^n_\ell=\mathbb{P}^n_\ell(G_n\in\cdot)$ for $\ell=0,1$. Let $S_{t,i-1}(\delta)$ as defined in Lemma 5.5.1. Then, for every $\mathfrak{g}_n\in\mathfrak{S}_n$*

$$\frac{\mathrm{d}Q^n_1}{\mathrm{d}Q^n_0}(\mathfrak{g}_n)=\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)}\prod_{k=m}^{nm}\left(\frac{k+\delta_1}{k+\delta_0}\right)^{N_{>k}(\mathfrak{g}_n)-N_{>k}(\mathfrak{g}_n[[0,\tau_n]])}.$$

Furthermore, almost-surely under $\mathbb{P}_0^n$

$$\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n) = \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)}\left(\frac{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_1}{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_0}\right).$$

where $V_{t,i}$ *is defined as in Section 5.2.2.*

Proof. The first expression in the statement of the lemma is an immediate consequence of Lemmas 5.5.2 and 5.5.3. Regarding the second statement, it suffices to observe that $N_{>k}(G_n)$ depends only on $s(G_n)$, so that [recall $G_t = G_{t,m}$]

$$N_{>k}(G_n) - N_{>k}(G_{\tau_n}) = \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\mathbf{1}_{\mathsf{d}_{G_{t,i-1}}(V_{t,i})=k}.$$

It follows that

$$\begin{aligned}
\prod_{k=m}^{nm}\left(\frac{k+\delta_1}{k+\delta_0}\right)^{N_{>k}(G_n)-N_{>k}(G_{\tau_n})} &= \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\prod_{k=m}^{nm}\left(\frac{k+\delta_1}{k+\delta_0}\right)^{\mathbf{1}_{\mathsf{d}_{G_{t,i-1}}(V_{t,i})=k}} \\
&= \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\prod_{k=m}^{nm}\left(\frac{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_1}{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_0}\right)^{\mathbf{1}_{\mathsf{d}_{G_{t,i-1}}(V_{t,i})=k}} \\
&= \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\left(\frac{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_1}{\mathsf{d}_{G_{t,i-1}}(V_{t,i})+\delta_0}\right).
\end{aligned}$$

This concludes the proof. □

The following lemma will also be used several times when analyzing likelihood ratios.

Lemma 5.5.5. *Suppose* $\tau_n \geq 3$. *Let* $S_{t,i-1}(\delta)$ *as defined in Lemma 5.5.1. Then for every* $\delta_0, \delta_1 > -m$

$$e^{-\frac{6m\Delta_n}{\tau_n}}\left(\frac{2m+\delta_0}{2m+\delta_1}\right)^{m\Delta_n} \leq \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)} \leq e^{\frac{6m\Delta_n}{\tau_n}}\left(\frac{2m+\delta_0}{2m+\delta_1}\right)^{m\Delta_n}$$

Proof. By definition of $S_{t,i-1}$

$$\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)} = \left(\frac{2m+\delta_0}{2m+\delta_1}\right)^{m\Delta_n}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{1+\frac{-2m+i-1}{t(2m+\delta_0)}}{1+\frac{-2m+i-1}{t(2m+\delta_1)}}$$

But for $j = 0, 1$, $\tau_n + 1 \leq t \leq n$, $1 \leq i \leq m$ and $\delta_j > -m$

$$1 - \frac{2}{\tau_n} \leq 1 + \frac{-2m+i-1}{t(2m+\delta_j)} \leq 1.$$

Thus,

$$\left(1-\frac{2}{\tau_n}\right)^{m\Delta_n} \leq \prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{1+\frac{-2m+i-1}{t(2m+\delta_0)}}{1+\frac{-2m+i-1}{t(2m+\delta_1)}} \leq \left(\frac{1}{1-2/\tau_n}\right)^{m\Delta_n}.$$

The conclusion follows because $\log(1-2/\tau_n) \geq -\frac{2}{\tau_n-2} \geq -\frac{6}{\tau_n}$ when $\tau_n \geq 3$. □

5.6 Proofs when the observation is the unlabeled graph

5.6.1 Proof of Lemma 5.3.3

$1 \implies 2$. Let $(A_n)_{n\geq 1}$ a sequence of $\pi_n(G_n)$-measurable sets such that $\mathbb{P}_0^n(A_n) \to 0$ and let $\varepsilon > 0$ arbitrary. Because $\mathbb{P}_0^n(A_n) = \mathbb{E}_0^n[\mathbb{E}_0^n(\mathbf{1}_{A_n} \mid G_n)] = \mathbb{E}_0^n[K_n(G_n, A_n)]$, it must be that $\mathbb{P}_0^n(K_n(G_n, A_n) > \varepsilon) \to 0$. But $\{\omega \in \Omega_n \ : \ K_n(G_n(\omega), A_n) > \varepsilon\} \in \sigma(G_n)$, so by 1 $\mathbb{P}_1^n(K_n(G_n, A_n) > \varepsilon) \to 0$. Since $\mathbb{P}_1^n(A_n) = \mathbb{E}_1^n[K_n(G_n, A_n)] \leq \varepsilon + \mathbb{P}_1^n(K_n(G_n, A_n) > \varepsilon)$, and since ε is arbitrary, the result follows.

$2 \implies 3$. Let $(E_n)_{n\geq 1}$ be a sequence such that $\mathbb{P}_0^n(s(G_n) \in E_n) \to 0$. Remark that $\mathbb{P}_0^n(s(G_n) \in E_n) = \mathbb{P}_0^n(s(\pi_n(G_n)) \in E_n) = \mathbb{P}_0^n(\pi_n(G_n) \in s^{-1}(E_n))$. So $\mathbb{P}_1^n(s(G_n) \in E_n) = \mathbb{P}_1^n(s(\pi_n(G_n)) \in E_n) = \mathbb{P}_1^n(\pi_n(G_n) \in s^{-1}(E_n))$ goes to zero by 2.

5.6.2 Proof of Proposition 5.3.4

Derivation of the expression of the likelihood-ratio

In this section we determine the expression of the likelihood ratio $\frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}$.

Lemma 5.6.1. *Let $S_{t,i-1}$ as defined in Lemma 5.5.1. $\mathbb{P}_0^n$-almost-surely:*

$$Y_n = \frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}(\pi_n(G_n)) = \frac{1}{|\Pi_n(G_n)|} \sum_{\bar{\pi}\in\Pi_n(G_n)} \prod_{t=\tau_n+1}^{n} \prod_{i=1}^{m} \frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)} \frac{\mathsf{d}_{G_{\bar{\pi}(t),i-1}}(V_{\bar{\pi}(t),i}) + \delta_1}{\mathsf{d}_{G_{\bar{\pi}(t),i-1}}(V_{\bar{\pi}(t),i}) + \delta_0}.$$

Proof. Let $\mathfrak{g}_n \in \mathfrak{S}_n$ and $\pi_0 \in \Pi_n(\mathfrak{g}_n)$. Then, for $j = 0, 1$

$$\begin{aligned}
\mathbb{P}_j^n\big(\pi_n(G_n) = \pi_0(\mathfrak{g}_n)\big) &= \mathbb{E}_j^n\left[\mathbb{E}_0^n\left(\sum_{\bar{\pi}\in\Pi_n(G_n)} \mathbf{1}_{\bar{\pi}(G_n)=\pi_0(\mathfrak{g}_n),\pi_n=\bar{\pi}} \middle| G_n\right)\right] \\
&= \mathbb{E}_j^n\left[\sum_{\bar{\pi}\in\Pi_n(G_n)} \mathbf{1}_{\bar{\pi}(G_n)=\pi_0(\mathfrak{g}_n)} \mathbb{P}_0^n(\pi_n = \bar{\pi} \mid G_n)\right] \\
&= \mathbb{E}_j^n\left[\frac{1}{|\Pi_n(G_n)|} \sum_{\bar{\pi}\in\Pi_n(G_n)} \mathbf{1}_{\bar{\pi}(G_n)=\pi_0(\mathfrak{g}_n)}\right]
\end{aligned}$$

Now remark that $\bar{\pi} \in \Pi_n(G_n)$ leaves invariant $\tilde{\mathcal{V}}(G_n)$ and $\pi_0 \in \Pi_n(\mathfrak{g}_n)$ leaves invariant $\tilde{\mathcal{V}}(\mathfrak{g}_n)$, thus

$$\begin{aligned}
\bar{\pi}(G_n) = \pi_0(\mathfrak{g}_n) &\implies \tilde{\mathcal{V}}(\bar{\pi}(G_n)) = \tilde{\mathcal{V}}(\pi_0(\mathfrak{g}_n)) \\
&\implies \tilde{\mathcal{V}}(G_n) = \tilde{\mathcal{V}}(\mathfrak{g}_n) \\
&\implies \Pi_n(G_n) = \Pi_n(\mathfrak{g}_n).
\end{aligned}$$

It follows

$$\begin{aligned}\mathbb{P}_j^n\big(\pi_n(G_n)=\pi_0(\mathfrak{g}_n)\big) &= \mathbb{E}_j^n\left[\frac{1}{|\Pi_n(\mathfrak{g}_n)|}\sum_{\bar\pi\in\Pi_n(\mathfrak{g}_n)}\mathbf{1}_{\bar\pi(G_n)=\pi_0(\mathfrak{g}_n)}\right]\\ &= \frac{1}{|\Pi_n(\mathfrak{g}_n)|}\sum_{\bar\pi\in\Pi_n(\mathfrak{g}_n)}\mathbb{P}_j^n\left(\bar\pi(G_n)=\pi_0(\mathfrak{g}_n)\right)\\ &= \frac{1}{|\Pi_n(\mathfrak{g}_n)|}\sum_{\bar\pi\in\Pi_n(\mathfrak{g}_n)}\mathbb{P}_j^n\left(\bar\pi(G_n)=\mathfrak{g}_n\right)\\ &= \frac{1}{|\Pi_n(\mathfrak{g}_n)|}\sum_{\bar\pi\in\Pi_n(\mathfrak{g}_n)}\mathbb{P}_j^n\big(G_n=\bar\pi^{-1}(\mathfrak{g}_n)\big)\end{aligned}$$

To simplify the notations in what follows, we denote respectively $\bar\pi^{-1}(G_n)$ and $\bar\pi^{-1}(\mathfrak{g}_n)$ by $\bar G_n$ and $\bar{\mathfrak{g}}_n$. Note that the advantage of permuting only bold vertices is that the set $\Pi_n(G_n)$ is a group, which makes the expression of the likelihood ratio easier to handle. As shown in Lemma 5.5.2, the likelihood of the labeled graph $\bar{\mathfrak{g}}_n$ under the null hypothesis does not depend on the permutation $\bar\pi$ when $\bar\pi\in\Pi_n(\mathfrak{g}_n)$. It follows that $\mathbb{P}_0^n(\pi_n(G_n)=\pi_0(\mathfrak{g}_n)) = \mathbb{P}_0^n\big(G_n=\bar{\mathfrak{g}}_n\big)$ for every $\bar\pi\in\Pi_n(\mathfrak{g}_n)$. Furthermore, the Lemma 5.6.2 below guarantees that $\bar{\mathfrak{g}}_n\in\mathfrak{S}_n$ whenever $\mathfrak{g}_n\in\mathfrak{S}_n$. Then by Lemma 5.5.4

$$\begin{aligned}Y_n &= \frac{1}{|\Pi_n(G_n)|}\sum_{\bar\pi\in\Pi_n(G_n)}\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(\bar G_n)\\ &= \frac{1}{|\Pi_n(G_n)|}\sum_{\bar\pi\in\Pi_n(G_n)}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)}\prod_{k=m}^{nm}\left(\frac{k+\delta_1}{k+\delta_0}\right)^{N_{>k}(\bar G_n)-N_{>k}((\bar G_n)_{\tau_n})}\end{aligned}$$

with $(\bar G_n)_t\equiv\bar G_n[\![0,t]\!]$ for all $t\in[\![1,n]\!]$. Let $\bar\pi\in\Pi_n(\mathfrak{g}_n)$ be arbitrary. Then,

$$N_{>k}(\bar G_n)-N_{>k}\big(\big(\bar G_n\big)_{\tau_n}\big)=\sum_{t=\tau_n+1}^{n}\sum_{s\in\mathsf{C}_{(\bar G_n)_t}(t)}\mathbf{1}\Big(k+1-\mu_{(\bar G_n)_t}(t,s)\le\mathsf{d}_{(\bar G_n)_{t-1}}(s)\le k\Big).$$

Remark that for $s\in\mathsf{C}_{(\bar G_n)_t}(t)$ is must be that $\mathsf{d}_{G_n}(s)>m$ and hence $\bar\pi(s)=s$. In particular $\mathsf{C}_{(\bar G_n)_t}(t)=\mathsf{C}_{G_{\bar\pi(t)}}(\bar\pi(t))$ and $\mu_{(\bar G_n)_t}(t,s)=\mu_{G_{\bar\pi(t)}}(\bar\pi(t),s)$. In addition, using the Lemma 5.6.3, we deduce that

$$\begin{aligned}N_{>k}(\bar G_n)-N_{>k}\big(\big(\bar G_n\big)_{\tau_n}\big) &= \sum_{t=\tau_n+1}^{n}\sum_{s\in\mathsf{C}_{G_{\bar\pi(t)}}(\bar\pi(t))}\mathbf{1}\Big(k+1-\mu_{G_{\bar\pi(t)}}(\bar\pi(t),s)\le\mathsf{d}_{G_{\bar\pi(t)-1}}(s)\le k\Big)\\ &= \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\mathbf{1}\Big(\mathsf{d}_{G_{\bar\pi(t),i-1}}(V_{\bar\pi(t),i})=k\Big)\end{aligned}$$

where the last line follows because the fact that vertex $\bar\pi(t)$ has a child whose degree is $\le k$ at instant $\bar\pi(t)-1$ but $>k$ at instant $\bar\pi(t)$ is equivalent to the fact that vertex $\bar\pi(t)$ choose a vertex $V_{\bar\pi(t),i}$ of degree k in $G_{\bar\pi(t),i-1}$ for some $i=1,\dots,m$. Consequently

$$Y_n=\frac{1}{|\Pi_n(G_n)|}\sum_{\bar\pi\in\Pi_n(G_n)}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{S_{t,i-1}(\delta_0)}{S_{t,i-1}(\delta_1)}\frac{\mathsf{d}_{G_{\bar\pi(t),i-1}}(V_{\bar\pi(t),i})+\delta_1}{\mathsf{d}_{G_{\bar\pi(t),i-1}}(V_{\bar\pi(t),i})+\delta_0}.$$ □

The following lemma shows that the permutations appearing in the proof of Lemma 5.6.1 leave invariant the support of preferential attachment graphs.

Lemma 5.6.2. *Let $\mathfrak{g}_n \in \mathfrak{S}_n$ and let $\bar{\pi} \in \Pi_n(\mathfrak{g}_n)$. Then $\bar{\pi}^{-1}(\mathfrak{g}_n) \in \mathfrak{S}_n$.*

Proof. In view of Lemma 5.5.1, the set $\mathfrak{S}_n$ is the set of directed labeled graphs on vertex set $[\![0, n]\!]$ where each non-zero vertex has out-degree exactly m and arrows are all directed from largest to smallest label. Since $\bar{\pi} \in \Pi_n(\mathfrak{g}_n)$, it permutes only the labels of vertices in $\tilde{\mathcal{V}}(\mathfrak{g}_n)$. But any $v \in \tilde{\mathcal{V}}(\mathfrak{g}_n)$ must satisfy $v > \tau_n'$ and have all of its children $c_1, \dots, c_k$ in $[\![0, \tau_n']\!]$. So $\bar{\pi}^{-1}(v) > \tau_n'$ as well and $\bar{\pi}^{-1}(c_j) = c_j$ for all its children. In other words, the out-degree of any vertex in $\bar{\pi}^{-1}(\mathfrak{g}_n)$ is also m and the arrows are all directed from largest to smallest label, as required. □

The following lemma shows that the degrees of the children of the $n - \tau_n$ late vertices remains invariant after the application of permutation $\bar{\pi}$.

Lemma 5.6.3. *Let $\mathfrak{g}_n \in \mathfrak{S}_n$, $\bar{\pi} \in \Pi_n(\mathfrak{g}_n)$ and let $\bar{\pi}^{-1}(\mathfrak{g}_n)_t = \bar{\pi}^{-1}(\mathfrak{g}_n)[[\![0, t]\!]]$ and $\mathfrak{g}_t = \mathfrak{g}_n[[\![0, t]\!]]$ for all $t \in [\![1, n]\!]$. Then for all $t \in [\![\tau_n + 1, n]\!]$ and all $s \in \mathsf{C}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_t}(t)$:*

$$\mathsf{d}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_{t-1}}(s) = \mathsf{d}_{\mathfrak{g}_{\bar{\pi}(t)-1}}(s).$$

Proof. Let $\bar{\pi} \in \Pi_n(\mathfrak{g}_n)$, $t \in [\![\tau_n + 1, n]\!]$ and $s \in \mathsf{C}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_t}(t)$. Observe that since $s \in \mathsf{C}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_t}(t)$ it is necessary that $\mathsf{d}_{\mathfrak{g}_n}(s) > m$ and then $\bar{\pi}(s) = s$.

Suppose first that for all $t' \in \mathsf{P}_{\bar{\pi}^{-1}(\mathfrak{g}_n)}(s)$ we have $\bar{\pi}(t') = t'$. Then $s \in \mathsf{C}_{\mathfrak{g}_t}(t)$ and $\mathsf{d}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_{t-1}}(s) = \mathsf{d}_{\mathfrak{g}_{t-1}}(s) = \mathsf{d}_{\mathfrak{g}_{\bar{\pi}(t)-1}}(s)$.

Second, suppose there exists $t' \in \mathsf{P}_{\bar{\pi}^{-1}(\mathfrak{g}_n)}(s)$ such that $\bar{\pi}(t') \neq t'$. It is necessary that $\bar{\pi}(t') > \tau_n'$ since $\bar{\pi}$ permute only the labels in $\tilde{\mathcal{V}}(\mathfrak{g}_n) \subset [\![\tau_n' + 1, n]\!]$. Furthermore $\bar{\pi}(t') \in \tilde{\mathcal{V}}(\mathfrak{g}_n)$ so it must be that $\mathsf{P}_{\mathfrak{g}_n}(s) \backslash \{\bar{\pi}(t')\} \subset [\![0, \tau_n']\!]$. Let enumerate $v_1 < \cdots < v_r$ the elements of $\mathsf{P}_{\mathfrak{g}_n}(s) \backslash \{\bar{\pi}(t')\}$. Hence the elements of $\mathsf{P}_{g_n}(s)$ are $v_1 < \cdots < v_r < \bar{\pi}(t')$. Since $v_1 < \cdots < v_r \leq \tau_n'$ they are not in $\tilde{\mathcal{V}}(\mathfrak{g}_n)$ and thus $\bar{\pi}(v_j) = v_j$ for all $j = 1, \dots, r$. It follows that the elements of $\mathsf{P}_{\bar{\pi}^{-1}(\mathfrak{g}_n)}(s)$ are $v_1, \dots, v_r, t'$ and satisfy

$$v_1 < \cdots < v_r \leq \tau_n' < t'$$

because $\bar{\pi}(t') > \tau_n' \implies t' > \tau_n'$. Therefore $t' = t$ and $\mathsf{d}_{\bar{\pi}^{-1}(\mathfrak{g}_n)_{t-1}}(s) = \mathsf{d}_{\mathfrak{g}_{\tau_n'}}(s) = \mathsf{d}_{\mathfrak{g}_{\bar{\pi}(t)-1}}(s)$. □

Bound on the second moment of the likelihood ratio

As in Lemma 5.6.1 we let $Y_n \equiv \frac{\mathrm{d}Q_1^{n,p}}{\mathrm{d}Q_0^{n,p}}(\pi_n(G_n))$ for simplicity. Then by Lemmas 5.6.1 and 5.5.5, since $\tau_n > \tau_n' \geq 3$

$$\mathbb{E}_0^n\big(Y_n^2 \mathbf{1}_{B_n}\big) \leq e^{12m\Delta_n/\tau_n} \left(\frac{2m+\delta_0}{2m+\delta_1}\right)^{2m\Delta_n} \times \mathbb{E}_0^n\left[\frac{\sum_{\pi,\bar{\pi}\in\Pi_n(G_n)}}{|\Pi_n(G_n)|^2} \prod_{t=\tau_n+1}^{n} \prod_{i=1}^{m} \frac{\mathsf{d}_{G_{\bar{\pi}(t),i-1}}(\mathrm{V}_{\bar{\pi}(t),i}) + \delta_1}{\mathsf{d}_{G_{\bar{\pi}(t),i-1}}(\mathrm{V}_{\bar{\pi}(t),i}) + \delta_0} \frac{\mathsf{d}_{G_{\pi(t),i-1}}(\mathrm{V}_{\pi(t),i}) + \delta_1}{\mathsf{d}_{G_{\pi(t),i-1}}(\mathrm{V}_{\pi(t),i}) + \delta_0} \mathbf{1}_{\mathrm{B}_n}\right].$$

Observe that $|\Pi_n(G_n)| = \big|\tilde{\mathcal{V}}(G_n)\big|!$. Moreover, on the event B_n we have forced that $[\![\tau_n + 1, n]\!] \subset \tilde{\mathcal{V}}(G_n)$. This implies that on B_n we have $\bar{\pi}(t) \in \tilde{\mathcal{V}}(G_n)$ for all $t \in [\![\tau_n + 1, n]\!]$.

Consequently on B_n,

$$\frac{\sum_{\pi,\bar\pi\in\Pi_n(G_n)}}{|\Pi_n(G_n)|^2}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{\bar\pi(t),i-1}}(\mathrm{V}_{\bar\pi(t),i})+\delta_1}{\mathsf{d}_{G_{\bar\pi(t),i-1}}(\mathrm{V}_{\bar\pi(t),i})+\delta_0}\frac{\mathsf{d}_{G_{\pi(t),i-1}}(\mathrm{V}_{\pi(t),i})+\delta_1}{\mathsf{d}_{G_{\pi(t),i-1}}(\mathrm{V}_{\pi(t),i})+\delta_0}$$
$$\leq\frac{1}{\left|\tilde{\mathcal{V}}(G_n)\right|!^2}\sum_{\substack{k'_{\tau_n+1}\neq\ldots\neq k'_n\in\tilde{\mathcal{V}}(G_n)\\ k_{\tau_n+1}\neq..\neq k_n\in\tilde{\mathcal{V}}(G_n)}}\sum_{\substack{\pi,\bar\pi\in\Pi_n(G_n)\\ (\pi(\tau_n+1),\ldots,\pi(n))=(k_{\tau_n+1},\ldots,k_n)\\ (\bar\pi(\tau_n+1),\ldots,\bar\pi(n))=(k'_{\tau_n+1},\ldots,k'_n)}}$$
$$\times\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_1}{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_0}\frac{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_1}{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_0}$$

which can be further bounded above by

$$\leq\frac{\left(\left|\tilde{\mathcal{V}}(G_n)\right|-\Delta_n\right)!^2}{\left|\tilde{\mathcal{V}}(G_n)\right|!^2}\sum_{\substack{k'_{\tau_n+1}\neq\ldots\neq k'_n\in\tilde{\mathcal{V}}(G_n)\\ k_{\tau_n+1}\neq\ldots\neq k_n\in\tilde{\mathcal{V}}(G_n)}}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_1}{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_0}\frac{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_1}{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_0}$$
$$\leq\frac{\left(\left|\tilde{\mathcal{V}}(G_n)\right|-\Delta_n\right)!^2}{\left|\tilde{\mathcal{V}}(G_n)\right|!^2}\sum_{\substack{k'_{\tau_n+1},\ldots,k'_n\in\tilde{\mathcal{V}}(G_n)\\ k_{\tau_n+1},\ldots,k_n\in\tilde{\mathcal{V}}(G_n)}}\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_1}{\mathsf{d}_{G_{k'_t,i-1}}(\mathrm{V}_{k'_t,i})+\delta_0}\frac{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_1}{\mathsf{d}_{G_{k_t,i-1}}(\mathrm{V}_{k_t,i})+\delta_0}$$
$$=\frac{\left(\left|\tilde{\mathcal{V}}(G_n)\right|-\Delta_n\right)!^2}{\left|\tilde{\mathcal{V}}(G_n)\right|!^2}\left(\sum_{k\in\tilde{\mathcal{V}}(G_n)}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}\right)^{2\Delta_n}$$
$$\leq\frac{\left(\left|\tilde{\mathcal{V}}(G_n)\right|-\Delta_n\right)!^2}{\left|\tilde{\mathcal{V}}(G_n)\right|!^2}\left(\sum_{k=\tau'_n+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}\right)^{2\Delta_n}$$

Next, we use that for any non-negative integer $\sqrt{2\pi n}(n/e)^n<n!<\sqrt{2\pi n}(n/e)^n e^{1/(12n)}$ (see for instance [Temme, 1996, Section 3.6]) which entails that for any $\nu>k\geq 1$

$$\frac{(\nu-k)!}{\nu!}\leq\frac{\sqrt{\nu-k}\left(\frac{\nu-k}{e}\right)^{\nu-k}e^{\frac{1}{12(\nu-k)}}}{\sqrt{\nu}\left(\frac{\nu}{e}\right)^{\nu}}=\left(1-\frac{k}{\nu}\right)^{\nu-k+\frac12}\left(\frac{e}{\nu}\right)^{k}e^{\frac{1}{12(\nu-k)}}\leq\nu^{-k}e^{k^2/\nu+\frac{1}{12(\nu-k)}}.$$

Since on the event B_n it holds that $|\tilde{\mathcal{V}}(G_n)|\geq\Delta'_n\left(1-\frac{\alpha_n\Delta'_n}{\tau'_n}\right)$, we deduce that

$$\frac{\left(\left|\tilde{\mathcal{V}}(G_n)\right|-\Delta_n\right)!}{\left|\tilde{\mathcal{V}}(G_n)\right|!}\leq\frac{1}{|\tilde{\mathcal{V}}(G_n)|^{\Delta_n}}\exp\left(\frac{\Delta_n^2}{|\tilde{\mathcal{V}}(G_n)|}+\frac{1}{12(|\tilde{\mathcal{V}}(G_n)|-\Delta_n)}\right)$$
$$\leq\frac{1}{(\Delta'_n)^{\Delta_n}}\exp\left(-\Delta_n\log\left(1-\frac{\alpha_n\Delta'_n}{\tau'_n}\right)+\frac{\Delta_n^2}{|\tilde{\mathcal{V}}(G_n)|}+\frac{1}{12(|\tilde{\mathcal{V}}(G_n)|-\Delta_n)}\right)$$
$$\leq\frac{1}{(\Delta'_n)^{\Delta_n}}\exp\left(\frac{\alpha_n\Delta_n\Delta'_n}{\tau'_n-\alpha_n\Delta'_n}+\frac{\Delta_n^2}{|\tilde{\mathcal{V}}(G_n)|}+\frac{1}{12(|\tilde{\mathcal{V}}(G_n)|-\Delta_n)}\right)$$
$$\leq\frac{1}{(\Delta'_n)^{\Delta_n}}\exp\left(\frac{2\alpha_n\Delta_n\Delta'_n}{\tau'_n}+\frac{2\Delta_n^2}{\Delta'_n}+\frac{1}{3\Delta'_n}\right)$$

where in the last line we have used the assumptions that $\frac{\alpha_n\Delta'_n}{\tau'_n}\leq\frac12$ and $\Delta_n\leq\frac14\Delta'_n$, which imply that $|\tilde{\mathcal{V}}(G_n)|\geq\frac{\Delta'_n}{2}$ and $|\tilde{\mathcal{V}}(G_n)|-\Delta_n\geq\frac{\Delta'_n}{4}$. Hence one obtains the bound [here we

use that $\frac{6m\Delta_n}{\tau_n} + \frac{4\Delta_n^2}{\Delta_n'} \leq \frac{10m\Delta_n^2}{\Delta_n'}$]

$$\mathbb{E}_0^n\big(Y_n^2 \mathbf{1}_{B_n}\big) \leq e^{\frac{4\alpha_n\Delta_n\Delta_n'}{\tau_n'} + \frac{10m\Delta_n^2}{\Delta_n'} + \frac{2}{3\Delta_n'}} \Big(\frac{2m+\delta_0}{2m+\delta_1}\Big)^{2m\Delta_n} \mathbb{E}_0^n\Bigg(\bigg(\frac{1}{\Delta_n'}\sum_{k=\tau_n'+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}\bigg)^{2\Delta_n}\Bigg).$$

Letting Z_n and m_n as in Lemma 5.6.4, we deduce from said lemma that

$$\begin{aligned}
\mathbb{E}_0^n\Bigg(\bigg(\frac{1}{\Delta_n'}\sum_{k=\tau_n'+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}\bigg)^{2\Delta_n}\Bigg) &\leq\ m_n^{2\Delta_n}\mathbb{E}_0^n\Bigg(\Big(1+\frac{Z_n-m_n}{m_n}\Big)^{2\Delta_n}\Bigg)\\
&= m_n^{2\Delta_n}\int_0^\infty \mathbb{P}_0^n\Bigg(\Big(1+\frac{Z_n-m_n}{m_n}\Big)^{2\Delta_n} > x\Bigg)\mathrm{d}x\\
&= m_n^{2\Delta_n}\int_0^\infty \mathbb{P}_0^n\Big(Z_n - m_n > m_n\big(x^{\frac{1}{2\Delta_n}}-1\big)\Big)\mathrm{d}x\\
&\leq m_n^{2\Delta_n}\Bigg(1+\int_1^\infty \mathbb{P}_0^n\Big(Z_n-m_n > \frac{m_n\log(x)}{2\Delta_n}\Big)\mathrm{d}x\Bigg)\\
&\leq m_n^{2\Delta_n}\Bigg(1+\int_1^\infty \exp\Big(-\frac{c\Delta_n' m_n^2}{4\Delta_n^2}\log(x)^2\Big)\mathrm{d}x\Bigg).
\end{aligned}$$

Using Lemma 5.6.5 to upper bound the last integral, together with Lemma 5.6.6 implying that $m_n \geq e^{-3m}\big(\frac{2m+\delta_1}{2m+\delta_0}\big)^m$ since $\tau_n' \geq 3$, it is found that there are constants $c_1, c_2 > 0$ depending only on δ_0, δ_1, and m, such that

$$\mathbb{E}_0^n\Bigg(\bigg(\frac{1}{\Delta_n'}\sum_{k=\tau_n'+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}\bigg)^{2\Delta_n}\Bigg) \leq m_n^{2\Delta_n}\Bigg(1+\sqrt{\frac{c_1\Delta_n^2}{\Delta_n'}e^{\frac{c_2\Delta_n^2}{\Delta_n'}}}\Bigg).$$

Finally, summarizing everything and using Lemma 5.6.6 to get an upper bound on m_n, we find that [here we use that $\frac{12m\Delta_n}{\tau_n'} + \frac{10m\Delta_n^2}{\Delta_n'} \leq \frac{22m\Delta_n^2}{\Delta_n'}$]

$$\log\mathbb{E}_0^n\big(Y_n^2\mathbf{1}_{B_n}\big) \leq \frac{4\alpha_n\Delta_n\Delta_n'}{\tau_n'} + \frac{22m\Delta_n^2}{\Delta_n'} + \frac{2}{3\Delta_n'} + \sqrt{\frac{c_1\Delta_n^2}{\Delta_n'}e^{\frac{c_2\Delta_n^2}{\Delta_n'}}}.$$

Auxiliary results used to prove the Proposition 5.3.4

Lemma 5.6.4. *Let*

$$Z_n = \frac{1}{\Delta_n'}\sum_{k=\tau_n'+1}^{n}\prod_{i=1}^{m}\frac{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(\mathrm{V}_{k,i})+\delta_0}, \qquad m_n = \frac{1}{\Delta_n'}\sum_{k=\tau_n'+1}^{n}\prod_{i=1}^{m}\frac{S_{k,i-1}(\delta_1)}{S_{k,i-1}(\delta_0)}.$$

Then there exists a constant $c > 0$ depending only on δ_0, δ_1, and m, such that for all $x \geq 0$

$$\mathbb{P}_0^n\big(Z_n - m_n \geq x\big) \leq e^{-c\Delta_n' x^2}.$$

Proof. In the proof we let $\mathcal{F}_t = \sigma(G_1,\dots,G_t)$ and $\mathcal{F}_{t,i} = \sigma(G_1,\dots,G_{t-1},G_{t,1},\dots,G_{t,i})$ for $t = 1,\dots,n$ and $i = 1,\dots,m$. Let $W_k = \prod_{i=1}^m \frac{\mathsf{d}_{G_{k,i-1}}(V_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(V_{k,i})+\delta_0}$ for $k = \tau_n'+1,\dots,n$.

Clearly $\mathbb{E}_0^n(W_k \mid \mathcal{F}_t) = W_k$ for all $t \geq k$. Also,

$$\begin{aligned}
\mathbb{E}_0^n\big(W_k \mid \mathcal{F}_{k-1}\big) &= \mathbb{E}_0^n\big(\mathbb{E}_0^n(W_k \mid \mathcal{F}_{k,m-1}) \mid \mathcal{F}_{k-1}\big) \\
&= \mathbb{E}_0^n\left(\prod_{i=1}^{m-1} \frac{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_1}{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_0} \mathbb{E}_0^n\Big(\frac{\mathsf{d}_{G_{k,m-1}}(V_{k,m}) + \delta_1}{\mathsf{d}_{G_{k,m-1}}(V_{k,m}) + \delta_0} \mid \mathcal{F}_{k,m-1}\Big) \mid \mathcal{F}_{k-1}\right) \\
&= \mathbb{E}_0^n\left(\prod_{i=1}^{m-1} \frac{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_1}{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_0} \sum_{u=0}^{k-1} \frac{\mathsf{d}_{G_{k,m-1}}(u) + \delta_1}{\mathsf{d}_{G_{k,m-1}}(u) + \delta_0} \frac{\mathsf{d}_{G_{k,m-1}}(u) + \delta_0}{S_{k,m-1}(\delta_0)} \mid \mathcal{F}_{k-1}\right) \\
&= \mathbb{E}_0^n\left(\prod_{i=1}^{m-1} \frac{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_1}{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_0} \frac{S_{k,m-1}(\delta_1)}{S_{k,m-1}(\delta_0)} \mid \mathcal{F}_{k-1}\right).
\end{aligned}$$

Continuing inductively, it is found that

$$\mathbb{E}_0^n\big(W_k \mid \mathcal{F}_{k-1}\big) = \prod_{i=1}^{m} \frac{S_{k,i-1}(\delta_1)}{S_{k,i-1}(\delta_0)}$$

and then $\mathbb{E}_0^n\big(W_k \mid \mathcal{F}_\ell\big) = \prod_{i=1}^m \frac{S_{k,i-1}(\delta_1)}{S_{k,i-1}(\delta_0)}$ for all $\ell < k$. Deduce that for all $k = \tau_n' + 1, \ldots, n$ and all $\ell = \tau_n' + 1, \ldots, n$

$$\mathbb{E}_0^n\big(W_k \mid \mathcal{F}_\ell\big) - \mathbb{E}_0^n\big(W_k \mid \mathcal{F}_{\ell-1}\big) = \begin{cases} 0 & \text{if } \ell \neq k, \\ \prod_{i=1}^m \frac{\mathsf{d}_{G_{k,i-1}}(V_{k,i})+\delta_1}{\mathsf{d}_{G_{k,i-1}}(V_{k,i})+\delta_0} - \prod_{i=1}^m \frac{S_{k,i-1}(\delta_1)}{S_{k,i-1}(\delta_0)} & \text{if } \ell = k. \end{cases} \tag{5.5}$$

Build the Doob martingale $M_j = \Delta_n' \mathbb{E}(Z_n \mid \mathcal{F}_j)$ and observe that

$$\sum_{j=\tau_n'+1}^{n} (M_j - M_{j-1}) = \Delta_n'\Big(Z_n - \mathbb{E}_0^n(Z_n \mid \mathcal{F}_{\tau_n'})\Big).$$

Furthermore for every $j = \tau_n' + 1, \ldots, n$, by equation (5.5)

$$\begin{aligned}
|M_j - M_{j-1}| &= \left| \sum_{k=\tau_n'+1}^{n} \Big(\mathbb{E}_0^n(W_k \mid \mathcal{F}_j) - \mathbb{E}_0^n(W_k \mid \mathcal{F}_{j-1})\Big) \right| \\
&= |W_j - \mathbb{E}_0^n(W_j \mid \mathcal{F}_{j-1})| \\
&\leq \max\Big(1, \frac{m+\delta_1}{m+\delta_0}\Big)^m
\end{aligned}$$

because

$$W_j = \prod_{i=1}^{m} \Big(1 + \frac{\delta_1 - \delta_0}{\mathsf{d}_{G_{k,i-1}}(V_{k,i}) + \delta_0}\Big) \leq \max\Big(1, \frac{m+\delta_1}{m+\delta_0}\Big)^m.$$

By Hoeffding-Azuma's inequality, for all $x \geq 0$, almost-surely

$$\begin{aligned}
\mathbb{P}_0^n\Big(Z_n - \mathbb{E}_0^n(Z_n \mid \mathcal{F}_{\tau_n'}) \geq \frac{x}{\Delta_n'} \mid \mathcal{F}_{\tau_n'}\Big) &= \mathbb{P}_0^n\big(M_n - M_{\tau_n'} \geq x \mid \mathcal{F}_{\tau_n'}\big) \\
&\leq \exp\left(-\frac{x^2}{2\Delta_n' \max\big(1, \frac{m+\delta_1}{m+\delta_0}\big)^m}\right).
\end{aligned}$$

Then the result follows by taking the expectation both sides of the last display and by noticing that $\mathbb{E}_0^n(Z_n \mid \mathcal{F}_{\tau_n'}) = m_n$ almost-surely. □

Lemma 5.6.5. *For every $\beta > 0$*

$$0 \le \int_1^\infty e^{-\beta \log(x)^2} \mathrm{d}x \le \sqrt{\frac{\pi e^{1/(2\beta)}}{\beta}}.$$

Proof. It is found after a straightforward change of variable that

$$\int_1^\infty e^{-\beta \log(x)^2} \mathrm{d}x = \frac{1}{\sqrt{2\beta}} \int_0^\infty e^{-\frac{1}{2}y^2} e^{\frac{1}{\sqrt{2\beta}}y} \mathrm{d}y = \frac{e^{\frac{1}{4\beta}}}{\sqrt{2\beta}} \int_0^\infty e^{-\frac{1}{2}(y-\frac{1}{\sqrt{2c}})^2} \mathrm{d}y \le \sqrt{\frac{\pi e^{1/(2\beta)}}{\beta}}. \quad \square$$

Lemma 5.6.6. *For every $\tau'_n \ge 3$*

$$e^{-\frac{6m}{\tau'_n}} \Big(\frac{2m+\delta_1}{2m+\delta_0}\Big)^m \le m_n \le e^{\frac{6m}{\tau'_n}} \Big(\frac{2m+\delta_1}{2m+\delta_0}\Big)^m.$$

Proof. As in Lemma 5.5.5, we have whenever $k > \tau'_n$ that

$$(1-2/\tau'_n)^m \Big(\frac{2m+\delta_1}{2m+\delta_0}\Big)^m \le \prod_{i=1}^m \frac{S_{k,i-1}(\delta_1)}{S_{k,i-1}(\delta_0)} \le \Big(\frac{2m+\delta_1}{2m+\delta_0}\Big)^m \frac{1}{(1-2/\tau'_n)^m}.$$

Hence the result follows since $\log(1-2/\tau'_n) \ge -\frac{2}{\tau'_n - 2} \ge -\frac{6}{\tau'_n}$ for $\tau'_n \ge 3$. $\square$

5.6.3 Proof of Proposition 5.3.5

Upper bound on the probabilities

By Markov's inequality

$$\begin{aligned}
\mathbb{P}_1^n\Big(|\tilde{\mathcal{V}}(G_n)| < \Delta'_n\Big(1 - \frac{\alpha_n \Delta'_n}{\tau'_n}\Big)\Big) &= \mathbb{P}_1^n\Big(|[\![\tau'_n + 1, n]\!] \backslash \tilde{\mathcal{V}}(G_n)| > \frac{\alpha_n (\Delta'_n)^2}{\tau'_n}\Big) \\
&\le \frac{\tau'_n}{\alpha_n (\Delta'_n)^2} \mathbb{E}_1^n\big(|[\![\tau'_n + 1, n]\!] \backslash \tilde{\mathcal{V}}(G_n)|\big) \\
&= \frac{\tau'_n}{\alpha_n (\Delta'_n)^2} \Big(\Delta'_n - \mathbb{E}_1^n\big(|\tilde{\mathcal{V}}(G_n)|\big)\Big).
\end{aligned}$$

Hence by Lemma 5.6.7 below

$$\mathbb{P}_1^n\Big(|\tilde{\mathcal{V}}(G_n)| < \Delta'_n\Big(1 - \frac{\alpha_n \Delta'_n}{\tau'_n}\Big)\Big) \le \frac{C}{\alpha_n} \begin{cases} \log(\tau'_n) & \text{if } \delta_0 = 0, \\ 1 & \text{if } \delta_0 > 0. \end{cases}$$

Similarly,

$$\begin{aligned}
\mathbb{P}_1^n\Big([\![\tau_n + 1, n]\!] \not\subset \tilde{\mathcal{V}}(G_n)\Big) &= \mathbb{P}_1^n\Big(|[\![\tau_n + 1, n]\!] \backslash \tilde{\mathcal{V}}(G_n)| \ge 1\Big) \\
&\le \mathbb{E}_1^n\big(|[\![\tau_n + 1, n]\!] \backslash \tilde{\mathcal{V}}(G_n)|\big) \\
&= \Delta_n - \mathbb{E}_1^n\big(|\tilde{\mathcal{V}}(G_n) \cap [\![\tau_n + 1, n]\!]|\big).
\end{aligned}$$

Hence by Lemma 5.6.8 below

$$\mathbb{P}_1^n\Big([\![\tau_n + 1, n]\!] \not\subset \tilde{\mathcal{V}}(G_n)\Big) \le \frac{C \Delta_n \Delta'_n}{\tau'_n} \begin{cases} \log(\tau'_n) & \text{if } \delta_0 = 0, \\ 1 & \text{if } \delta_0 > 0. \end{cases}$$

Computation of expectations of $|\tilde{\mathcal{V}}(G_n)|$ and $|\tilde{\mathcal{V}}(G_n) \cap [\![\tau_n + 1, n]\!]|$

In this section we derive estimates on the expectations of $|\tilde{\mathcal{V}}(G_n)|$ and $|\tilde{\mathcal{V}}(G_n) \cap [\![\tau_n + 1, n]\!]|$ which are crucial elements in bounding the probability $\mathbb{P}_1^n(B_n^c)$.

Lemma 5.6.7. *There exists a constant $B > 0$ depending only on m, δ_0 and δ_1, such that for all $2 \le \tau_n' \le n$*

$$\Delta_n' \ge \mathbb{E}_1^n\big(|\tilde{\mathcal{V}}(G_n)|\big) \ge \Delta_n' - \frac{B(\Delta_n')^2}{\tau_n'} \begin{cases} (\tau_n')^{-\delta_0/(2m+\delta_0)} & \text{if } \delta_0 < 0, \\ \log(\tau_n') & \text{if } \delta_0 = 0, \\ 1 & \text{if } \delta_0 > 0. \end{cases}$$

Proof. First observe that the upper bound is trivial since $\tilde{\mathcal{V}}(G_n) \subset [\![\tau_n' + 1, n]\!]$ almost-surely. We now focus on the lower bound.

Let us write $X_n = |\tilde{\mathcal{V}}(G_n)|$ for simplicity. In the whole proof we use the convention that an empty product equals one. Note that

$$\begin{aligned} X_n &= \sum_{j=\tau_n'+1}^{n} \mathbf{1}\Big(\mathsf{d}_{G_n}(j) = m,\ \forall k \in \mathsf{C}_{G_n}(j),\ k \le \tau_n' \text{ and } \forall \ell \in \mathsf{P}_{G_n}(k)\backslash\{j\},\ \ell \le \tau_n'\Big) \\ &= \sum_{\substack{\tau_n' < j \le n \\ 1 \le \ell \le m}} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau_n'} \sum_{\substack{y_1, \ldots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \Bigg(\prod_{j < k \le n} \mathbf{1}(k \not\to_{G_k} j) \Bigg) \Bigg(\prod_{i=1}^{\ell} \mathbf{1}(\mu_{G_j}(j, x_i) = y_i) \Bigg) \Bigg(\prod_{\substack{\tau_n' < k \le n \\ k \ne j}} \prod_{i=1}^{\ell} \mathbf{1}(k \not\to_{G_k} x \end{aligned}$$

Indeed, the previous can be rewritten more conveniently as [here $\boldsymbol{x}_\ell = (x_1, \ldots, x_\ell)$ and $\boldsymbol{y}_\ell = (y_1, \ldots, y_\ell)$]

$$X_n = \sum_{\tau_n' < j \le n} \sum_{\ell=1}^{m} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau_n'} \sum_{\substack{y_1, \ldots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} Y_n^{\boldsymbol{x}_\ell, \boldsymbol{y}_\ell, j}$$

with

$$Y_n^{\boldsymbol{x}_\ell, \boldsymbol{y}_\ell, j} = \prod_{\tau_n' < k < j} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell\}) \prod_{i=1}^{\ell} \mathbf{1}(\mu_{G_j}(j, x_i) = y_i) \prod_{j < k \le n} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell, j\}).$$

Letting $\mathcal{F}_\ell = \sigma(G_1, \ldots, G_\ell)$ and $\delta(j) = \delta_0 \mathbf{1}(j \le \tau_n) + \delta_1 \mathbf{1}(j > \tau_n)$, it is seen that [assuming $j < n$, otherwise the result is trivial]

$$\begin{aligned} \mathbb{E}_1^n(Y_n^{\boldsymbol{x}_\ell, \boldsymbol{y}_\ell, j} \mid \mathcal{F}_{n-1}) &= \prod_{\tau_n' < k < j} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell\}) \prod_{i=1}^{\ell} \mathbf{1}(\mu_{G_j}(j, x_i) = y_\ell) \prod_{j < k \le n-1} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell, j\}) \\ &\quad \times \prod_{i=1}^{m} \left(1 - \frac{[\mathsf{d}_{G_{n-1}}(x_1) + \delta(n)] + \cdots + [\mathsf{d}_{G_{n-1}}(x_\ell) + \delta(n)] + [m + \delta(n)]}{S_{n,i-1}(\delta(n))} \right) \\ &= \prod_{\tau_n' < k < j} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell\}) \prod_{i=1}^{\ell} \mathbf{1}(\mu_{G_j}(j, x_i) = y_\ell) \prod_{j < k \le n-1} \mathbf{1}(k \not\to_{G_k} \{x_1, \ldots, x_\ell, j\}) \\ &\quad \times \prod_{i=1}^{m} \left(1 - \frac{[\mathsf{d}_{G_{\tau_n'}}(x_1) + y_1 + \delta(n)] + \cdots + [\mathsf{d}_{G_{\tau_n'}}(x_\ell) + y_\ell + \delta(n)] + [m + \delta(n)]}{S_{n,i-1}(\delta(n))} \right) \end{aligned}$$

where the second line follows because if the product of indicators is non-zero, then at instant $n - 1$ no vertex other than j has connected to one of the $x_1, \ldots, x_\ell$ on the time

interval $[\![\tau'_n+1, n-1]\!]$, and j has edge multiplicity y_ℓ with x_ℓ. Defining for simplicity $D^{\boldsymbol{x}_\ell}_{\tau'_n} = \sum_{i=1}^{\ell} \mathsf{d}_{G_{\tau'_n}}(x_i)$, and taking conditional expectation of the previous inductively with respect to $\mathcal{F}_{n-2}, \dots, \mathcal{F}_{\tau'_n}$, it is found that [here the combinatorial factor comes from enumerating all the possibilities of connecting j to $x_1, \dots, x_\ell$ with edges multiplicities $y_1, \dots, y_\ell$]

$$\mathbb{E}_1^n(Y_n^{\boldsymbol{x}_\ell, \boldsymbol{y}_\ell, j} \mid \mathcal{F}_{\tau'_n}) = \frac{m!}{\prod_{i=1}^{\ell} y_i!} \prod_{\tau'_n < k < j} \prod_{i=1}^{m} \left(1 - \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + \ell\delta(k)}{S_{k,i-1}(\delta(k))}\right)$$
$$\times \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))} \prod_{j < k \le n} \prod_{i=1}^{m} \left(1 - \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + 2m + (\ell+1)\delta(k)}{S_{k,i-1}(\delta(k))}\right).$$

Hence, for $c = \frac{m!}{\prod_{i=1}^{\ell} y_i!}$,

$$\begin{aligned}
\mathbb{E}_1^n(Y_n^{\boldsymbol{x}_\ell, \boldsymbol{y}_\ell, j} \mid \mathcal{F}_{\tau'_n}) &\ge c \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))} \prod_{\tau'_n < k \le n} \prod_{i=1}^{m} \left(1 - \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + 2m + (\ell+1)\delta(k)}{S_{k,i-1}(\delta(k))}\right) \\
&\ge c \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))} \left(1 - \sum_{\tau'_n < k \le n} \sum_{i=1}^{m} \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + 2m + (\ell+1)\delta(k)}{S_{k,i-1}(\delta(k))}\right) \\
&\ge c \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))} \left(1 - m \sum_{\tau'_n < k \le n} \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + 2m + (m+1)(\delta(k) \vee 0)}{S_{k,0}(\delta(k))}\right).
\end{aligned}$$

We now define two random variables

$$\tilde{X}_n = \sum_{\tau'_n < j \le n} \sum_{\ell=1}^{m} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau'_n} \sum_{\substack{y_1, \dots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \frac{m!}{\prod_{i=1}^{\ell} y_i!} \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))}$$

and

$$R_n = m \sum_{\tau'_n < j,k \le n} \sum_{\ell=1}^{m} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau'_n} \sum_{\substack{y_1, \dots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \frac{m!}{\prod_{i=1}^{\ell} y_i!}$$
$$\times \sum_{\substack{y_1, \dots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \frac{m!}{\prod_{i=1}^{\ell} y_i!} \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{j,i-1}(\delta(j))} \frac{D^{\boldsymbol{x}_\ell}_{\tau'_n} + 2m + (m+1)\left(\delta(k) \vee 0\right)}{S_{k,0}(\delta(k))}$$

so that $\mathbb{E}_1^n(X_n \mid \mathcal{F}_{\tau'_n}) \ge \tilde{X}_n - R_n$ almost-surely. To compute the expectations of $\tilde{X}_n$ and R_n, we use the following trick. For a fixed $j > \tau'_n$ we define on the same probability space a sequence of random graphs $((\tilde{G}^j_{t,i})_{i=0}^m)_{t \ge 1}$ such that $\tilde{G}^j_{t,i} = G_{t,i}$ for $1 \le t \le \tau'_n$ and $0 \le i \le m$, and then $(\tilde{G}^j_{t,i})_{i=0}^m$ evolves independently of $(G_{t,i})_{i=0}^m$ according to the preferential attachment rule with parameter $\delta(t) = \delta(j)$ for all $t > \tau'_n$. Then, we see that

$$\begin{aligned}
1 &= \mathbb{E}_1^n \left(\sum_{\ell=1}^{m} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau'_n} \sum_{\substack{y_1, \dots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \mathbf{1}\left(\forall i = 1, \dots, \ell,\ \mu_{\tilde{G}^j_{\tau'_n + 1}}(\tau'_n + 1, x_i) = y_i\right) \,\middle|\, \mathcal{F}_{\tau'_n} \right) \\
&= \sum_{\ell=1}^{m} \sum_{0 \le x_1 < \cdots < x_\ell \le \tau'_n} \sum_{\substack{y_1, \dots, y_\ell \ge 1 \\ y_1 + \cdots + y_\ell = m}} \frac{m!}{\prod_{i=1}^{\ell} y_i!} \frac{\prod_{i=1}^{\ell} \prod_{i'=1}^{y_i} \left(\mathsf{d}_{G_{\tau'_n}}(x_i) + i' - 1 + \delta(j)\right)}{\prod_{i=1}^{m} S_{\tau'_n + 1, i-1}(\delta(j))}.
\end{aligned}$$

So indeed,

$$\begin{aligned}
\mathbb{E}_1^n(\tilde{X}_n) &= \sum_{\tau_n'<j\le n}\prod_{i=1}^m \frac{S_{\tau_n'+1,i-1}(\delta(j))}{S_{j,i-1}(\delta(j))}\\
&= \sum_{\tau_n'<j\le n}\prod_{i=1}^m \frac{(2m+\delta(j))(\tau_n'+1)-2m+i-1}{(2m+\delta(j))j-2m+i-1}\\
&= \sum_{\tau_n'<j\le n}\prod_{i=1}^m \Big(1-\frac{(2m+\delta(j))(j-\tau_n'-1)}{(2m+\delta(j))j-2m+i-1}\Big)\\
&\ge \sum_{\tau_n'<j\le n}\Big(1-\frac{(2m+\delta(j))(j-\tau_n'-1)}{(2m+\delta(j))j-2m}\Big)^m\\
&\ge \Delta_n'\Big(1-\frac{m\Delta_n'}{n-2}\Big).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}_1^n\Bigg(\sum_{v\in \mathsf{C}_{\tilde{G}^j_{\tau_n'+1}}(\tau_n'+1)} \mathsf{d}_{\tilde{G}^j_{\tau_n'}}(v)\ \Big|\ \mathcal{F}_{\tau_n'}\Bigg)\\
&\quad=\sum_{\ell=1}^m \sum_{0\le x_1<\cdots<x_\ell\le\tau_n'}\ \sum_{\substack{y_1,\dots,y_\ell\ge 1\\ y_1+\cdots+y_\ell=m}} \frac{m!}{\prod_{i=1}^\ell y_i!}\frac{\prod_{i=1}^\ell\prod_{i'=1}^{y_i}\big(\mathsf{d}_{G_{\tau_n'}}(x_i)+i'-1+\delta(j)\big)}{\prod_{i=1}^m S_{\tau_n'+1,i-1}(\delta(j))}D^{\boldsymbol{x}_\ell}_{\tau_n'}
\end{aligned}$$

from which we deduce that

$$\mathbb{E}_1^n(R_n)=m\sum_{\tau_n'<j,k\le n}\frac{1}{S_{k,0}(\delta(k))}\Bigg(\mathbb{E}_1^n\Bigg(\sum_{v\in \mathsf{C}_{\tilde{G}^j_{\tau_n'+1}}(\tau_n'+1)} \mathsf{d}_{\tilde{G}^j_{\tau_n'}}(v)\Bigg)+2m+(m+1)\delta(k)\vee 0\Bigg)\prod_{i=1}^m\frac{S_{\tau_n'+1,i-1}(\delta(j))}{S_{j,i-1}(\delta(j))}$$

But

$$\begin{aligned}
\mathbb{E}_1^n\Bigg(\sum_{v\in \mathsf{C}_{\tilde{G}^j_{\tau_n'+1}}(\tau_n'+1)} \mathsf{d}_{\tilde{G}^j_{\tau_n'}}(v)\Bigg) &\le \sum_{i=1}^m \mathbb{E}_1^n\big(\mathsf{d}_{\tilde{G}^j_{\tau_n'+1,i-1}}(\tilde{V}_{t,i})\big)\\
&= \sum_{i=1}^m \mathbb{E}_1^n\Bigg(\sum_{v=0}^{\tau_n'}\mathsf{d}_{\tilde{G}^j_{\tau_n'+1,i-1}}(v)\frac{\mathsf{d}_{\tilde{G}^j_{\tau_n'+1,i-1}}(v)+\delta(j)}{S_{\tau_n'+1,i-1}(\delta(j))}\Bigg)\\
&= \sum_{i=1}^m\sum_{v=0}^{\tau_n'}\frac{\mathbb{E}_1^n\big(\mathsf{d}_{\tilde{G}^j_{\tau_n'+1,i-1}}(v)^2\big)}{S_{\tau_n'+1,i-1}(\delta(j))}+\delta(j)\sum_{i=1}^m\frac{2m\tau_n'+i-1}{S_{\tau_n'+1,i-1}(\delta(j))}\\
&\le 2m\sum_{v=0}^{\tau_n'}\frac{\mathbb{E}_1^n\big((\mathsf{d}_{G_{\tau_n'}}(v)+\delta_0)^2\big)}{S_{\tau_n'+1,0}(\delta(j))}+\frac{2m(m\delta(j)\vee 0+(m-\delta_0)^2)(\tau_n'+1)}{S_{\tau_n'+1,0}(\delta(j))}
\end{aligned}$$

where we have used that only m edges can be added between τ_n' and $\tau_n'+1$, so the difference between the degree of v in $\tilde{G}^j_{\tau_n'+1}$ and its degree in $G_{\tau_n'}$ cannot exceed m. Remarking that in time interval $[\![0,\tau_n']\!]$ the process $(\tilde{G}_t)_{t\ge1}$ evolves according to the preferential attachment rule with parameter δ_0, and remarking that $\frac{\tau_n'+1}{S_{\tau_n'+1,0}(\delta(j))}$ is bounded by a constant, it follows

letting $\underline{\delta} = \delta_0 \wedge \delta_1$

$$\mathbb{E}_1^n(R_n) \leq \frac{\mathbb{E}_1^n(\tilde{X}_n)\Delta_n'}{S_{\tau_n'+1,0}(\underline{\delta})}\left(C + \frac{2m^2}{S_{\tau_n'+1,0}(\underline{\delta})}\sum_{v=0}^{\tau_n'}\mathbb{E}_0^n\big((\mathsf{d}_{G_{\tau_n'}}(v)+\delta_0)^2\big)\right)$$

for a constant $C > 0$ depending only on δ_0, δ_1, and m. By Lemmas 5.6.10 and 5.6.11, there are constants $C, C' > 0$ depending solely on m and δ_0 such that

$$\begin{aligned}\sum_{v=0}^{\tau_n'}\mathbb{E}_0^n\big((\mathsf{d}_{G_{\tau_n'}}(v)+\delta_0)^2\big) &\leq C'\sum_{v=0}^{\tau_n'}\Big(\frac{\tau_n'}{1\vee v}\Big)^{2m/(2m+\delta_0)}\\ &\leq C'(\tau_n')^{2m/(2m+\delta_0)}\left(4 + \int_1^{\tau_n'}\frac{1}{x^{2m/(2m+\delta_0)}}\mathrm{d}x\right)\\ &\leq C''\begin{cases}(\tau_n')^{2m/(2m+\delta_0)} & \text{if } \delta_0 < 0,\\ \tau_n'\log(\tau_n') & \text{if } \delta_0 = 0,\\ \tau_n' & \text{if } \delta_0 > 0.\end{cases}\end{aligned}$$

The conclusion follows because $S_{\tau_n'+1,0}(\underline{\delta}) = (2m+\underline{\delta})(\tau_n'+1) - 2m \geq m\tau_n' + \underline{\delta}$. □

Lemma 5.6.8. *There exists a constant $B > 0$ depending only on m, δ_0 and δ_1, such that for all $2 \leq \tau_n' < \tau_n \leq n$*

$$\Delta_n \geq \mathbb{E}_1^n\big(|\tilde{\mathcal{V}}(G_n)\cap[\![\tau_n+1,n]\!]|\big) \geq \Delta_n - \frac{B\Delta_n\Delta_n'}{\tau_n'}\begin{cases}(\tau_n')^{-\delta_0/(2m+\delta_0)} & \text{if } \delta_0 < 0,\\ \log(\tau_n') & \text{if } \delta_0 = 0,\\ 1 & \text{if } \delta_0 > 0.\end{cases}$$

Proof. The lemma follows by remarking that $|\tilde{\mathcal{V}}(G_n)\cap[\![\tau_n+1,n]\!]|$ can be rewritten as

$$\sum_{j=\tau_n+1}^{n}\mathbf{1}\Big(\mathsf{d}_{G_n}(j) = m,\ \forall k\in\mathsf{C}_{G_n}(j),\ k\leq\tau_n' \text{ and } \forall\ell\in\mathsf{P}_{G_n}(k)\backslash\{j\},\ \ell\leq\tau_n'\Big).$$

Then the rest of the proof is identical to Lemma 5.6.7 *mutatis mutandis*. □

Auxiliary results used to prove the Proposition 5.3.5

Lemma 5.6.9. *Let $\gamma_t = \prod_{i=1}^m\big(1+\frac{1}{S_{t,i-1}(\delta_0)}\big)$. For every $0 \leq u < t \leq n$*

$$\mathbb{E}_0^n(\mathsf{d}_{G_t}(u)+\delta_0) = \gamma_t\mathbb{E}_0^n(\mathsf{d}_{G_{t-1}}(u)+\delta_0).$$

Proof. These are standard computations, see for instance [van der Hofstad, 2016, Chapter 8]. □

Lemma 5.6.10. *For every $2 \leq t \leq n$ and $0 \leq u \leq t$*

$$\mathbb{E}_0^n[(\mathsf{d}_{G_t}(u)+\delta_0)^2] = \xi_{1\vee u}^t(m+\delta_0)^2 + \kappa_{1\vee u}^t(m+\delta_0)$$

where for all $r = 1,\dots,t$:

$$\xi_r^t = \prod_{r+1\leq j\leq t}\prod_{i=1}^m\Big(1+\frac{2}{S_{j,i-1}(\delta_0)}\Big)$$

and

$$\kappa_r^t = \sum_{r+1\le j\le t}\Bigg(\prod_{j+1\le p\le t}\prod_{i=1}^{m}\Big(1+\frac{2}{S_{p,i-1}(\delta_0)}\Big)\Bigg)\Bigg(\prod_{r+1\le p\le j-1}\prod_{i=1}^{m}\Big(1+\frac{1}{S_{p,i-1}(\delta_0)}\Big)\Bigg)$$
$$\times\Bigg(\sum_{k=1}^{m}\frac{1}{S_{j,k-1}(\delta_0)}\prod_{1\le i\le k-1}\Big(1+\frac{1}{S_{j,i-1}(\delta_0)}\Big)\prod_{k+1\le i\le m}\Big(1+\frac{2}{S_{j,i-1}(\delta_0)}\Big)\Bigg).$$

Proof. Let $\mathcal{F}_t=\sigma(G_1,\dots,G_t)$. We first compute $\mathbb{E}_0^n[(\mathsf{d}_{G_t}(u)+\delta_0)^2\mid\mathcal{F}_t]=\mathbb{E}_0^n[(\mathsf{d}_{G_{t,m}}(u)+\delta_0)^2\mid\mathcal{F}_{t-1}]$. We define the coefficients $(\alpha_{t,i})_{i=1}^m$ and $(\beta_{t,i})_{i=1}^m$ such that $\alpha_{t,m}=1$ and $\beta_{t,m}=0$, and satisfying the recurrence for $i=m,\dots,1$

$$\alpha_{t,i-1}=\alpha_{t,i}\Big(1+\frac{2}{S_{t,i-1}(\delta_0)}\Big),\qquad\qquad \beta_{t,i-1},=\beta_{t,i}\Big(1+\frac{1}{S_{t,i-1}(\delta_0)}\Big)+\frac{\alpha_{t,i}}{S_{t,i-1}(\delta_0)}.$$

It is seen that for every $r=1,\dots,m$ (using the convention that empty product equals one and empty sum equals zero):

$$\begin{aligned}\alpha_{t,r}&=\prod_{r+1\le j\le m}\Big(1+\frac{2}{S_{t,j-1}(\delta_0)}\Big)\\ \beta_{t,r}&=\sum_{r+1\le k\le m}\frac{\alpha_{t,k}}{S_{t,k-1}(\delta_0)}\prod_{r+1\le j\le k-1}\Big(1+\frac{1}{S_{t,j-1}(\delta_0)}\Big)\\ &=\sum_{r+1\le k\le m}\frac{1}{S_{t,k-1}(\delta_0)}\prod_{r+1\le j\le k-1}\Big(1+\frac{1}{S_{t,j-1}(\delta_0)}\Big)\prod_{k+1\le j\le m}\Big(1+\frac{2}{S_{t,j-1}(\delta_0)}\Big).\end{aligned}$$

Then we consider the random variable

$$M_{t,i}=\alpha_{t,i}\big(\mathsf{d}_{G_{t,i}}(u)+\delta\big)^2+\beta_{t,i}\big(\mathsf{d}_{G_{t,i}}(u)+\delta\big).$$

We claim that $(M_{t,i})_{i=1}^m$ is a martingale with respect to $(\mathcal{F}_{t,i})_{i=1}^m$, where $\mathcal{F}_{t,i}=\sigma(G_{t,0},\dots,G_{t,i})$; *ie.* we claim that $\mathbb{E}(M_{t,i}\mid\mathcal{F}_{t,i-1})=M_{t,i-1}$ for $i=1,\dots,m$. Indeed for $i=1,\dots,m$

$$\begin{aligned}\mathbb{E}_0^n\big((\mathsf{d}_{G_{t,i}}(u)+\delta_0)^2\mid\mathcal{F}_{t,i-1}\big)&=(\mathsf{d}_{G_{t,i-1}}(u)+1+\delta_0)^2\frac{\mathsf{d}_{G_{t,i-1}}(u)+\delta_0}{S_{t,i-1}(\delta_0)}\\ &\quad+(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)^2\Big(1-\frac{\mathsf{d}_{G_{t,i-1}}(u)+\delta_0}{S_{t,i-1}(\delta_0)}\Big)\\ &=(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)^2\Big(1+\frac{2}{S_{t,i-1}(\delta_0)}\Big)+\frac{\mathsf{d}_{G_{t,i-1}}(u)+\delta_0}{S_{t,i-1}(\delta_0)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_0^n\big(\mathsf{d}_{G_{t,i}}(u)+\delta_0\mid\mathcal{F}_{t,i-1}\big)&=(\mathsf{d}_{G_{t,i-1}}(u)+1+\delta_0)\frac{\mathsf{d}_{G_{t,i-1}}(u)+\delta_0}{S_{t,i-1}(\delta_0)}\\ &\quad+(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)\Big(1-\frac{\mathsf{d}_{G_{t,i-1}}(u)+\delta_0}{S_{t,i-1}(\delta_0)}\Big)\\ &=(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)\Big(1+\frac{1}{S_{t,i-1}(\delta_0)}\Big)\end{aligned}$$

so that

$$\begin{aligned}
\mathbb{E}_0^n(M_{t,i} \mid \mathcal{F}_{t,i-1}) &= \alpha_{t,i}\mathbb{E}_0^n\big((\mathsf{d}_{G_{t,i}}(u)+\delta_0)^2 \mid \mathcal{F}_{t,i-1}\big) + \beta_{t,i}\mathbb{E}_0^n\big((\mathsf{d}_{G_{t,i}}(u)+\delta_0)\mathcal{F}_{t,i-1}\big) \\
&= \alpha_{t,i}\Big(1+\frac{2}{S_{t,i-1}(\delta_0)}\Big)(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)^2 \\
&\quad + \left(\frac{\alpha_{t,i}}{S_{t,i-1}(\delta_0)} + \beta_{t,i}\Big(1+\frac{1}{S_{t,i-1}(\delta_0)}\Big)\right)(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0) \\
&= \alpha_{t,i-1}(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0)^2 + \beta_{t,i-1}(\mathsf{d}_{G_{t,i-1}}(u)+\delta_0) \\
&= M_{t,i-1}.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}_0^n[(\mathsf{d}_{G_{t,m}}(u)+\delta)^2 \mid \mathcal{F}_{t-1}] &= \mathbb{E}_1^n[M_{t,m} \mid \mathcal{F}_{t-1}] \\
&= M_{t,0} \\
&= \alpha_{t,0}\big(\mathsf{d}_{G_{t,0}}(u)+\delta\big)^2 + \beta_{t,0}\big(\mathsf{d}_{G_{t,0}}(u)+\delta\big) \\
&= \alpha_{t,0}\big(\mathsf{d}_{G_{t-1}}(u)+\delta\big)^2 + \beta_{t,0}\big(\mathsf{d}_{G_{t-1}}(u)+\delta\big).
\end{aligned}$$

Next, let $(\xi_j^t)_{j=1}^t$ and $(\kappa_j^t)_{j=1}^t$ as in the statement of the lemma, and $\gamma_j = \prod_{i=1}^m \big(1+\frac{1}{S_{j,i-1}(\delta_0)}\big)$ (as in Lemma 5.6.9). It is straightforward to show that $(\xi_j^t)_{j=1}^t$ and $(\kappa_j^t)_{j=1}^t$ satisfy $\xi_t^t = 1$ and $\kappa_t^t = 0$ and the recurrence

$$\xi_{j-1}^t = \xi_j^t\alpha_{j,0}, \qquad \kappa_{j-1}^t = \xi_j^t\beta_{j,0} + \kappa_j^t\gamma_j.$$

Indeed, for $r = 1, \dots, t$

$$\begin{aligned}
\xi_r^t &= \prod_{r+1\le j\le t}\alpha_{j,0} = \prod_{r+1\le j\le t}\prod_{i=1}^m\Big(1+\frac{2}{S_{j,i-1}(\delta_0)}\Big) \\
\kappa_r^t &= \sum_{r+1\le j\le t}\xi_j^t\beta_{j,0}\prod_{r+1\le k\le j-1}\gamma_k
\end{aligned}$$

which are equal to the expression given in the statement of the lemma. Let now define for $j = 1\vee u, \dots, t$

$$M_j' = \xi_j^t\big(\mathsf{d}_{G_j}(u)+\delta_0\big)^2 + \kappa_j^t\big(\mathsf{d}_{G_j}(u)+\delta_0\big).$$

The claim is that $(M_j')_{j\ge 1}$ is a martingale with respect to $(\mathcal{F}_j)_{j=1\vee u}^t$. Indeed, using Lemma 5.6.9 and the above computations

$$\begin{aligned}
\mathbb{E}_0^n(M_j' \mid \mathcal{F}_{j-1}) &= \xi_j^t\mathbb{E}_0^n\big((\mathsf{d}_{G_j}(u)+\delta_0)^2 \mid \mathcal{F}_{j-1}\big) + \kappa_j^t\mathbb{E}_0^n\big(\mathsf{d}_{G_j}(u)+\delta_0 \mid \mathcal{F}_{j-1}\big) \\
&= \xi_j^t\Big(\alpha_{j,0}(\mathsf{d}_{G_{j-1}}(u)+\delta_0)^2 + \beta_{j,0}(\mathsf{d}_{G_{j-1}}(u)+\delta_0)\Big) + \kappa_j^t\gamma_j(\mathsf{d}_{G_{j-1}}(u)+\delta_0) \\
&= \xi_j^t\alpha_{j,0}(\mathsf{d}_{G_{j-1}}(u)+\delta_0)^2 + \big(\xi_j^t\beta_{j,0}+\kappa_j^t\gamma_j\big)(\mathsf{d}_{G_{j-1}}(u)+\delta_0) \\
&= \xi_{j-1}^t(\mathsf{d}_{G_{j-1}}(u)+\delta_0)^2 + \kappa_{j-1}^t(\mathsf{d}_{G_{j-1}}(u)+\delta_0) \\
&= M_{j-1}'.
\end{aligned}$$

This implies that (because $\mathsf{d}_{G_{1\vee u}}(u) = m$ almost-surely)

$$\mathbb{E}_0^n\big((\mathsf{d}_{G_t}(u)+\delta)^2\big) = \mathbb{E}_0^n(M_t') = \mathbb{E}_0^n(M_{1\vee u}') = \xi_{1\vee u}(m+\delta)^2 + \kappa_{1\vee u}(m+\delta). \qquad \square$$

Lemma 5.6.11. *Let $\xi_{1\vee u}^t$ and $\kappa_{1\vee u}^t$ as in the statement of Lemma 5.6.10. There exists a constant $B > 0$ depending only on m and δ_0 such that for all $0 \le u < t$ such that*

$$\max(\xi_{1\vee u}^t,\, \kappa_{1\vee u}^t) \le B\Big(\frac{t}{1\vee u}\Big)^{2m/(2m+\delta_0)}.$$

Proof. Let define the function $g : \mathbb{N} \to \mathbb{R}_+$ as $g(1) = 0$ and $g(n) = \sum_{j=2}^{n} \sum_{i=1}^{m} \frac{1}{S_{j,i-1}(\delta_0)}$ for $n \geq 2$. Observe that for any $n \geq 2$

$$\begin{aligned} g(n) &= \sum_{j=2}^{n} \sum_{i=1}^{m} \frac{1}{(2m+\delta_0)j - 2m + i - 1} \\ &= \frac{1}{2m+\delta_0} \sum_{j=2}^{n} \sum_{i=1}^{m} \frac{1}{j + \frac{-2m+i-1}{2m+\delta_0}} \\ &= \frac{m}{2m+\delta_0} \sum_{j=2}^{n} \frac{1}{j} + \frac{1}{2m+\delta_0} \sum_{j=1}^{n} \sum_{i=1}^{m} \Big(\frac{1}{j + \frac{-2m+i-1}{2m+\delta_0}} - \frac{1}{j} \Big) \\ &= \frac{m}{2m+\delta_0} \sum_{j=2}^{n} \frac{1}{j} + \frac{1}{(2m+\delta_0)^2} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{-2m+i-1}{j\big(j + \frac{-2m+i-1}{2m+\delta_0}\big)}. \end{aligned}$$

Thus letting $H(n) = \sum_{j=1}^{n} \frac{1}{j}$ denote the j-th harmonic number, we deduce that there is a constant $C > 0$ depending only on m and δ_0 such that for all $n \geq 2$:

$$\frac{m}{2m+\delta_0}\big(\gamma + \log(n)\big) - C \leq \frac{m}{2m+\delta_0} H(n) - C \leq g(n) \leq \frac{m}{2m+\delta_0}\big(H(n) - 1\big) \leq \frac{m}{2m+\delta_0}\big(\gamma + \log(n)\big) \tag{5.6}$$

with γ the Euler constant, using well known bounds on the harmonic numbers. It follows from (5.6) that

$$\begin{aligned} \xi^t_{1\vee u} &\leq \exp\big(2g(t) - 2g(1 \vee u)\big) \\ &\leq \exp\Big(\frac{2m}{2m+\delta_0} \log\Big(\frac{t}{1 \vee u}\Big) + C\Big). \end{aligned}$$

Next, since $\max_{2 \leq j \leq t} \max_{1 \leq \leq m} \prod_{1 \leq i \leq k-1} \big(1 + \frac{1}{S_{j,i-1}(\delta_0)}\big) \prod_{k+1 \leq i \leq m} \big(1 + \frac{2}{S_{j,i-1}(\delta_0)}\big)$ is finite, we find that for some constants $C', C'', C''' > 0$ depending only on m and δ_0

$$\begin{aligned} \kappa^t_{1\vee u} &\leq C' \sum_{1\vee u+1 \leq j \leq t} \frac{1}{S_{j,0}(\delta_0)} e^{2g(t)-2g(j)} e^{g(j-1)-g(1\vee u)} \\ &\leq C'' \sum_{1\vee u+1 \leq j \leq t} \frac{1}{j} \Big(\frac{t}{j}\Big)^{2m/(2m+\delta_0)} \Big(\frac{j}{1 \vee u}\Big)^{m/(2m+\delta_0)} \\ &\leq C'' \frac{t^{2m/(2m+\delta_0)}}{(1 \vee u)^{m/(2m+\delta_0)}} \int_{1\vee u}^{t} \frac{1}{x^{1+m/(2m+\delta_0)}} \mathrm{d}x \\ &\leq C''' \Big(\frac{t}{1 \vee u}\Big)^{2m/(2m+\delta_0)}. \end{aligned}$$

This concludes the proof. □

5.7 Proofs when the labeled graph is observed

5.7.1 Supplementary notations

We use the same conventions as in the main paper. We furthermore make use of the following supplementary notations in the subsequent proofs. We write $a_n \lesssim b_n$ to denote $a_n = O(b_n)$. We say that $a_n \asymp b_n$ if there exist constants $c_1, c_2 > 0$ such that $c_1 a_n \leq b_n \leq c_2 a_n$. For sequence of real-valued random variables $(X_n)_{n \geq 1}$ with respective distributions $(P_n)_{n \geq 1}$ and real numbers c we write $X_n \xrightarrow{P_n} c$, $j = 0, 1$, to say that $\lim_n \mathbb{P}^n_j(|X_n - c| >$

$\varepsilon) = 0$ for all $\varepsilon > 0$ and we abusively say that $(X_n)_{n\geq 1}$ converges in probability to c, even though the random variables X_n may not be necessarily defined on the same probability space. The notation $X_n \overset{P_n}{\rightsquigarrow} X$ stands for convergence in distribution of $(X_n)_{n\geq 1}$ to a random variable X.

5.7.2 Proof of Theorem 5.3.6

Bounding the sum of the two errors of the likelihood-ratio test

Let Q_0^n (respectively Q_1^n) denote the law of G_n under $\mathbb{P}_0^n$ (resp. $\mathbb{P}_1^n$). The limiting behaviour of $\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}$ under $\mathbb{P}_0^n$ is characterized below in Proposition 5.7.1, while Proposition 5.7.2 characterizes its behaviour under $\mathbb{P}_1^n$. Using Proposition 5.7.1, it is found that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}_0^n\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n) > 1\right) &= \mathbb{P}_0^n\left(\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) > 0\right)\\ &= \limsup_{n\to\infty} \mathbb{P}_0^n\left(\frac{1}{\Delta_n}\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) + \ell_\infty^0 > \ell_\infty^0\right)\\ &= 0.\end{aligned}$$

Using Proposition 5.7.2, we prove similarly that for any $K > 0$

$$\limsup_{n\to\infty} \mathbb{P}_1^n\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n) \leq 1\right) = 0.$$

Regime of contiguity

Using the Lemma A.4, it is clear that $\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}\right)_{n\geq 1}$ is uniformly bounded below and above when $\limsup_n \Delta_n < +\infty$, and thus $(Q_1^n)_{n\geq 1}$ is contiguous to $(Q_0^n)_{n\geq 1}$.

Estimates on the behaviour of the likelihood-ratio under the null and alternative hypothesis

The following propositions are used for the proof of Theorem 5.3.6. Recall that Q_j^n denote the law of G_n under $\mathbb{P}_j^n$, for $j = 0, 1$. We recall that $p(\delta)$ is the limiting distribution of the degree distribution of the affine preferential attachment graph with parameter δ (see also equation (5.3)).

Proposition 5.7.1. *Let $\delta_0, \delta_1 > -m$ with $\delta_0 \neq \delta_1$. For every increasing sequence $(\tau_n)_{n\geq 1}$ of integer numbers satisfying $0 \leq \tau_n < n$ and $\Delta_n = n - \tau_n \to \infty$, one has*

$$\frac{1}{\Delta_n}\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) \xrightarrow{\mathbb{P}_0^n} -\ell_\infty^0$$

where [letting $X \sim p(\delta_0)$]

$$\ell_\infty^0 = \frac{m}{2m+\delta_0}\left((2m+\delta_0)\log\left(1+\frac{\delta_1-\delta_0}{2m+\delta_0}\right) - \mathbb{E}\left[(X+\delta_0)\log\left(1+\frac{\delta_1-\delta_0}{X+\delta_0}\right)\right]\right) > 0.$$

Proof. In what follows, we introduce the random variables $D_{t,i} = \mathsf{d}_{G_{t,i-1}}(\mathrm{V}_{t,i})$ and the filtrations $\mathcal{F}_t = \sigma(G_0, \ldots, G_t)$ and $\mathcal{F}_{t,i-1} = \sigma(G_{t,0}, \ldots, G_{t,i-1})$ as in Gao and van der Vaart [2017] to simplify the notations. We recall that the expression of the likelihood-ratio

has been established in Lemma A.4. Normalizing the log-likelihood ratio by $n-\tau_n$, one obtains:

$$\frac{\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right)}{n-\tau_n} = \frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\log\left(1+\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right)-\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right) \tag{5.7a}$$

$$+\frac{(\delta_1-\delta_0)}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{1}{D_{t,i}+\delta_0}-\frac{t}{S_{t,i-1}(\delta_0)}\right) \tag{5.7b}$$

$$-\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\log\left(1+\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right)-\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right). \tag{5.7c}$$

We will control each of the three terms involved in the previous display separately.

First term (5.7a). This term can be written as:

$$\sum_{k=m}^{\infty}\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\left(\log\left(1+\frac{\delta_1-\delta_0}{k+\delta_0}\right)-\frac{\delta_1-\delta_0}{k+\delta_0}\right)\frac{\mathbf{1}_{D_{t,i}=k}}{n-\tau_n}=\sum_{k=m}^{\infty}\left(\log\left(\frac{k+\delta_1}{k+\delta_0}\right)-\frac{\delta_1-\delta_0}{k+\delta_0}\right)A_{n,m}$$

and

$$\begin{aligned}
A_{n,m} &= \frac{\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\mathbf{1}_{D_{t,i}=k}}{n-\tau_n}\\
&= \frac{\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_0^n\left[\mathbf{1}_{\mathtt{d}_{V_{t,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{n-\tau_n}+\frac{\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\mathbb{P}_0^n(D_{t,i}=k\mid\mathcal{F}_{t,i-1})}{n-\tau_n}\\
&= \frac{\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_0^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{n-\tau_n}+\frac{k+\delta_0}{n-\tau_n}\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_0)}.
\end{aligned}$$

where $N_k(G_{t,i-1})$ is the number of vertices of degree k in the graph after attaching the $(i-1)$-th edge to the vertex t of the graph. On the one hand, by Hoeffding-Azuma inequality, the first term of the equality above converges to 0 in probability. On the other hand, we have that for all (t,i):

$$|N_k(G_{t,i})-N_k(G_n)|\le(n-\tau_n)(m+1)$$

It follows that:

$$m\frac{N_k(G_n)-(n-\tau_n)(m+1)}{S_{n,m}(\delta_0)}\le\frac{1}{n-\tau_n}\sum_{\substack{\tau_n<t\le n\\1\le i\le m}}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_0)}\le m\frac{N_k(G_n)+(n-\tau_n)(m+1)}{S_{\tau_n+1,0}(\delta_0)}$$

where both sides converge in probability to $\frac{m}{2m+\delta_0}p_k(\delta_0)=\frac{p_{>k}(\delta_0)}{k+\delta_0}$. Thanks to the dominated convergence theorem (Note that the dominated convergence theorem holds also when convergence takes place only in probability):

$$(5.7a)\xrightarrow{\mathbb{P}_0^n}\sum_{k=m}^{\infty}\left(\log\left(1+\frac{\delta_1-\delta_0}{k+\delta_0}\right)-\frac{\delta_1-\delta_0}{k+\delta_0}\right)p_{>k}(\delta_0).$$

Second term (5.7b). First, note that:

$$\mathbb{E}_0^n\left[\frac{1}{D_{t,i}+\delta_0}\mid\mathcal{F}_{t,i-1}\right]=\sum_{k=m}^{\infty}\frac{1}{k+\delta_0}\frac{(k+\delta_0)N_k(G_{t,i-1})}{S_{t,i-1}(\delta_0)}=\frac{t}{S_{t,i-1}(\delta_0)}.$$

Given that $0 < \frac{1}{D_{t,i}+\delta_0} \leq \frac{1}{m+\delta_0}$, one can apply the Hoeffding-Azuma inequality and obtain:

$$(5.7b) \xrightarrow{\mathbb{P}_0^n} 0.$$

Third term (5.7c). Assume $\delta_1 > \delta_0$. Since the function $t \mapsto \log\left(1 + \frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right)$ is non-decreasing for every choice of $t \in [\![1, n]\!]$ and $i \in [\![1, m]\!]$, one has:

$$\log\left(1 + \frac{(\tau_n+1)(\delta_1-\delta_0)}{S_{\tau_n+1,0}(\delta_0)}\right) \leq \log\left(1 + \frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right) \leq \log\left(1 + \frac{n(\delta_1-\delta_0)}{S_{n,0}(\delta_0)}\right).$$

It follows that:

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\log\left(1 + \frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right) \to m\log\left(1 + \frac{\delta_1-\delta_0}{2m+\delta_0}\right).$$

This convergence holds also when $\delta_1 < \delta_0$. A similar argument yields:

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)} \to \frac{m(\delta_1-\delta_0)}{2m+\delta_0}.$$

To sum up, one has:

$$(5.7c) \to -m\left(\log\left(1 + \frac{\delta_1-\delta_0}{2m+\delta_0}\right) - \frac{\delta_1-\delta_0}{2m+\delta_0}\right).$$

Gathering all of the above estimates, it follows that :

$$\frac{\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right)}{n-\tau_n} \xrightarrow{\mathbb{P}_0^n} -\ell_\infty^0$$

where

$$\begin{aligned}
\ell_\infty^0 &= m\log\left(1 + \frac{\delta_1-\delta_0}{2m+\delta_0}\right) - \sum_{k=m}^{\infty} p_{>k}(\delta_0)\log\left(1 + \frac{\delta_1-\delta_0}{k+\delta_0}\right) \\
&= \frac{m}{2m+\delta_0}\left((2m+\delta_0)\log\left(1 + \frac{\delta_1-\delta_0}{2m+\delta_0}\right) - \sum_{k=m}^{\infty}(k+\delta_0)p_k(\delta_0)\log\left(1 + \frac{\delta_1-\delta_0}{k+\delta_0}\right)\right) \\
&= \frac{m}{2m+\delta_0}\left((2m+\delta_0)\log\left(1 + \frac{\delta_1-\delta_0}{2m+\delta_0}\right) - \mathbb{E}\left[(X+\delta_0)\log\left(1 + \frac{\delta_1-\delta_0}{X+\delta_0}\right)\right]\right)
\end{aligned}$$

where $X \sim p(\delta_0) = (p_k(\delta_0))_k$. Since $\mathbb{E}(X) = 2m$ (see for instance [van der Hofstad, 2016, Exercise 8.16]) and when $\delta_0 \neq \delta_1$ the map $x \mapsto x\log\left(1 + \frac{\delta_1-\delta_0}{x}\right)$ is concave and non-affine on $\mathbb{R}^+$ and $p(\delta_0)$ is not a Dirac distribution, it follows that $\ell_\infty^0 > 0$. □

Proposition 5.7.2. *Let $\delta_0, \delta_1 > -m$ with $\delta_0 \neq \delta_1$. For every increasing sequence $(\tau_n)_{n\geq 1}$ of integer numbers satisfying $0 \leq \tau_n < n$ and $\Delta_n = n - \tau_n \to \infty$, one has*

$$\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) \xrightarrow{\mathbb{P}_1^n} \ell_\infty^1$$

where [letting $X \sim p(\delta_0)$]

$$\ell_\infty^1 = -\frac{m}{2m+\delta_1}\left(\mathbb{E}\left[(X+\delta_1)\log\left(1 + \frac{\delta_0-\delta_1}{X+\delta_1}\right)\right] - (2m+\delta_1)\log\left(1 + \frac{\delta_0-\delta_1}{2m+\delta_1}\right)\right) < 0.$$

Proof. In what follows, we introduce the random variables $D_{t,i} = \mathsf{d}_{G_{t,i-1}}(\mathrm{V}_{t,i})$ and the filtrations $\mathcal{F}_t = \sigma(G_0, \dots, G_t)$ and $\mathcal{F}_{t,i-1} = \sigma(G_{t,0}, \dots, G_{t,i-1})$ as in Gao and van der Vaart [2017] to simplify the notations. Recall the expression of the likelihood-ratio has been established in Lemma A.4. Normalizing the log-likelihood ratio by $n-\tau_n$, one obtains:

$$\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) = \frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\log\left(1+\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right)-\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right) \tag{5.8a}$$

$$+\frac{(\delta_1-\delta_0)}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{1}{D_{t,i}+\delta_0}-\frac{t}{S_{t,i-1}(\delta_0)}\right) \tag{5.8b}$$

$$-\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\log\left(1+\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right)-\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}\right). \tag{5.8c}$$

We will control each of the three terms involved in the previous display separately.

First term (5.8a).

$$\sum_{\substack{\tau_n<t\le n\\ 1\le i\le m}}\frac{\left(\log\left(1+\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right)-\frac{\delta_1-\delta_0}{D_{t,i}+\delta_0}\right)}{n-\tau_n} = \sum_{k=m}^{+\infty}\left(\log\left(\frac{k+\delta_1}{k+\delta_0}\right)-\frac{\delta_1-\delta_0}{k+\delta_0}\right)\frac{\sum_{\substack{\tau_n<t\le n\\ 1\le i\le m}}\mathbf{1}_{D_{t,i}=k}}{n-\tau_n}.$$

On the one hand with $\mathcal{F}_{t,i-1} = \sigma(G_{t,0}, \dots, G_{t,i-1})$

$$\frac{\sum_{\substack{\tau_n<t\le n\\ 1\le i\le m}}\mathbf{1}_{D_{t,i}=k}}{n-\tau_n} = \sum_{\substack{\tau_n<t\le n\\ 1\le i\le m}}\frac{\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{n-\tau_n}+\frac{1}{n-\tau_n}\sum_{\substack{\tau_n<t\le n\\ 1\le i\le m}}\frac{(k+\delta_1)N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}.$$

The first term converges to 0 in probability $\mathbb{P}_1^n$ using Hoeffding-Azuma inequality. On the other hand, we have that for all (t,i):

$$|N_k(G_{t,i}) - N_k(G_n)| \le (n-\tau_n)(m+1).$$

It follows that:

$$m\frac{N_k(G_n)-(n-\tau_n)(m+1)}{S_{n,m}(\delta_0)} \le \frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_0)} \le m\frac{N_k(G_n)+(n-\tau_n)(m+1)}{S_{\tau_n+1,0}(\delta_0)}.$$

Thus:

$$\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\mathbf{1}_{D_{t,i}=k}}{n-\tau_n} \xrightarrow{\mathbb{P}_1^n} \frac{(k+\delta_1)m}{2m+\delta_1}p_k(\delta_0).$$

Using dominated convergence theorem, one has:

$$(5.8a) \xrightarrow{\mathbb{P}_1^n} \frac{m}{2m+\delta_1}\sum_{k=m}^{\infty}(k+\delta_1)p_k(\delta_0)\left(\log\left(1+\frac{\delta_1-\delta_0}{k+\delta_0}\right)-\frac{\delta_1-\delta_0}{k+\delta_0}\right).$$

Second term (5.8b). The term

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t(\delta_1-\delta_0)}{S_{t,i-1}(\delta_0)}$$

converges clearly to $\frac{m(\delta_1-\delta_0)}{2m+\delta_0}$. On the other hand:

$$\begin{aligned}
\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{1}{D_{t,i}+\delta_0} &= \frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{1}{D_{t,i}+\delta_0}-\mathbb{E}_1^n\left[\frac{1}{D_{t,i}+\delta_0}\mid\mathcal{F}_{t,i-1}\right]\right)\\
&\quad+\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\mathbb{E}_1^n\left[\frac{1}{D_{t,i}+\delta_0}\mid\mathcal{F}_{t,i-1}\right]\\
&= \frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{1}{D_{t,i}+\delta_0}-\mathbb{E}_1^n\left[\frac{1}{D_{t,i}+\delta_0}\mid\mathcal{F}_{t,i-1}\right]\right)\\
&\quad+\sum_{k=m}^{nm}\frac{k+\delta_1}{k+\delta_0}\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}.
\end{aligned}$$

The first term converges in probability to 0. We will show that:

$$\sum_{k=m}^{nm}\frac{k+\delta_1}{k+\delta_0}\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\xrightarrow{\mathbb{P}_1^n}\frac{m}{2m+\delta_1}\sum_{k=m}^{+\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0).$$

For positive K, we have:

$$\begin{aligned}
&\left|\sum_{k=m}^{nm}\frac{k+\delta_1}{k+\delta_0}\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}-\frac{m}{2m+\delta_1}\sum_{k=m}^{+\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)\right|\\
&\quad\le\sum_{k=m}^{K}\frac{k+\delta_1}{k+\delta_0}\left|\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}-\frac{m}{2m+\delta_1}p_k(\delta_0)\right|\\
&\qquad+\frac{m}{2m+\delta_1}\sum_{k=K+1}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)+\frac{C(\delta_0,\delta_1)}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{\sum_{k=K+1}^{nm}N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\\
&\quad\le\sum_{k=m}^{K}\frac{k+\delta_1}{k+\delta_0}\left|\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}-\frac{m}{2m+\delta_1}p_k(\delta_0)\right|\\
&\qquad+\frac{m}{2m+\delta_1}\sum_{k=K+1}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)+\frac{C(\delta_0,\delta_1)}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_{>K}(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\\
&\quad\le\sum_{k=m}^{K}\frac{k+\delta_1}{k+\delta_0}\left|\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}-\frac{m}{2m+\delta_1}p_k(\delta_0)\right|\\
&\qquad+\frac{m}{2m+\delta_1}\sum_{k=K+1}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)+mC(\delta_0,\delta_1)\frac{N_{>K}(G_n)+(n-\tau_n)(m+1)}{S_{\tau_n+1,0}(\delta_1)}\\
&\quad\le\sum_{k=m}^{K}\frac{k+\delta_1}{k+\delta_0}\left|\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}-\frac{m}{2m+\delta_1}p_k(\delta_0)\right|\\
&\qquad+\frac{m}{2m+\delta_1}\sum_{k=K+1}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)+mC(\delta_0,\delta_1)\frac{\frac{2mn}{K}+(n-\tau_n)(m+1)}{S_{\tau_{n+1},0}(\delta_1)}
\end{aligned}$$

where $C(\delta_0,\delta_1)$ is a constant depending solely on δ_0 and δ_1. The upper-bound converges in probability to:

$$\frac{m}{2m+\delta_1}\sum_{k=K+1}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0)+\frac{2m^2C(\delta_0,\delta_1)}{K(2m+\delta_1)}$$

which can be made arbitrarily small for large values of K. We deduce that:

$$(5.8b) \xrightarrow{\mathbb{P}_1^n} \frac{m(\delta_1-\delta_0)}{2m+\delta_1}\sum_{k=m}^{\infty}\frac{k+\delta_1}{k+\delta_0}p_k(\delta_0) - \frac{m(\delta_1-\delta_0)}{2m+\delta_0}.$$

Third term (5.8a). Finally, the last term is shown to converge to:

$$(5.8c) \to m\left(\frac{\delta_1-\delta_0}{2m+\delta_0} - \log\left(1+\frac{\delta_1-\delta_0}{2m+\delta_0}\right)\right).$$

It follows that:

$$\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q_1^n}{\mathrm{d}Q_0^n}(G_n)\right) \xrightarrow{\mathbb{P}_1^n} \ell_\infty^1$$

where

$$\begin{aligned}\ell_\infty^1 &= -\frac{m}{2m+\delta_1}\left(\sum_{k=m}^{\infty}(k+\delta_1)p_k(\delta_0)\log\left(1+\frac{\delta_0-\delta_1}{k+\delta_1}\right) - (2m+\delta_1)\log\left(1+\frac{\delta_0-\delta_1}{2m+\delta_1}\right)\right)\\ &= -\frac{m}{2m+\delta_1}\left(\mathbb{E}\left[(X+\delta_1)\log\left(1+\frac{\delta_0-\delta_1}{X+\delta_1}\right)\right] - (2m+\delta_1)\log\left(1+\frac{\delta_0-\delta_1}{2m+\delta_1}\right)\right)\end{aligned}$$

which can be shown to be positive by a similar argument to that used in Proposition 5.7.1. □

5.7.3 Proof of Theorem 5.3.7

We first remark that the fact that $\hat{\delta}_{0,n}$ and $\hat{\delta}_{1,n}$ are asymptotically independent is an immediate consequence of the fact that the likelihood factorizes as the product of a function depending solely on δ_0 and another function depending solely on δ_1. Hence, it is enough to consider separately $\ell_{1:\tau_n}$ and $\ell_{\tau_{n+1}:n}$. But remark that $\ell_{1:\tau_n}$ is the log-likelihood of the model without change-point at size τ_n. Hence, the existence, uniqueness, and asymptotic normality of $(\hat{\delta}_{0,n})_{n\geq 1}$ follows immediately from the results in Gao and van der Vaart [2017][1].

Thus in the next we focus on the analysis of the sequence of maximizers of $\ell_{\tau_n+1,n}$. The proof is standard and mimicks the steps in Gao and van der Vaart [2017]. First define the score as

$$\dot{\ell}_{\tau_n+1:n}(\delta') = \sum_{k=m}^{nm}\frac{N_{>k}(G_n)-N_{>k}(G_{\tau_n})}{k+\delta'} - \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{(2m+\delta')t-2m+i-1}.$$

The Proposition 5.7.3 below establishes that $\dot{\ell}_{\tau_n+1:n}(\cdot)$ converges uniformly over $(-m+\varepsilon,+\infty)$ in probability to a function ι_1' (whose expression is given in said proposition) that is monotone decreasing with a unique zero. The Proposition 5.7.4 shows that with high probability $\dot{\ell}_{\tau_n+1:n}$ has no zero in $(-m,\varepsilon)$. These facts are exploited hereafter in Proposition 5.7.5 to establish the existence and uniqueness (with high probability) and the consistency of $(\delta_{1,n})_{n\geq 1}$. Finally, given the consistency of $(\delta_{1,n})_{n\geq 1}$, we deduce in Section the asymptotic normality using the standard machinery.

Proposition 5.7.3. *For every $\varepsilon > 0$*

$$\sup_{\delta\geq -m+\varepsilon}\left|\frac{\dot{\ell}_{\tau_n+1:n}(\delta)}{n-\tau_n} - \iota_1'(\delta)\right| \xrightarrow{\mathbb{P}_1^n} 0$$

[1] We note however that for practical reasons Gao and van der Vaart [2017] restricts the MLE to some compact set $[-a,b]$ but this is in fact not required, by the argument we develop in Proposition 5.7.4.

where [with $X \sim p(\delta_0)$]

$$\iota_1'(\delta) = \frac{2m+\delta_0}{2m+\delta_1}\sum_{k=m}^{\infty}\frac{k+\delta_1}{k+\delta_0}\frac{p_{>k}(\delta_0)}{k+\delta} - \frac{m}{2m+\delta} = \frac{m}{2m+\delta_1}\left(\mathbb{E}\left[\frac{X+\delta_1}{X+\delta}\right] - \frac{\mathbb{E}[X]+\delta_1}{\mathbb{E}[X]+\delta}\right).$$

The proof of Proposition 5.7.3 is delayed to Section 5.7.3.

Proposition 5.7.4. *There exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leq \varepsilon_0$ it holds*

$$\mathbb{P}_1^n\Big(\inf_{\delta' \in (-m,\varepsilon)} \dot{\ell}_{\tau_n+1:n}(\delta') \geq \Delta_n\Big) \to 1.$$

The proof or Proposition 5.7.4 is delayed to Section 5.7.3

Proposition 5.7.5. *For every $(\delta_0, \delta_1) \in (-m, +\infty)$, if $\Delta_n \to \infty$:*

$$\hat{\delta}_{1,n} \xrightarrow{\mathbb{P}_1^n} \delta_1.$$

The proof of Proposition 5.7.5 is delayed to Section 5.7.3.

Asmptotic normality of $(\hat{\delta}_{1,n})_{n\geq 1}$

In what follows, we introduce the random variables $D_{t,i} = \mathsf{d}_{G_{t,i-1}}(\mathrm{V}_{t,i})$ and the filtrations $\mathcal{F}_t = \sigma(G_0, \ldots, G_t)$ and $\mathcal{F}_{t,i-1} = \sigma(G_{t,0}, \ldots, G_{t,i-1})$ as in Gao and van der Vaart [2017] to simplify the notations. By definition of $\hat{\delta}_{1,n}$, one has:

$$\sum_{k=m}^{nm}\frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{k+\hat{\delta}_{1,n}} = \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{(2m+\hat{\delta}_{1,n})t - 2m + i - 1}.$$

It follows that:

$$\begin{aligned}
&\sum_{k=m}^{nm}\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k} - \mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k} \mid \mathcal{F}_{t,i-1}\right]\right)}{k+\hat{\delta}_{1,n}} \\
&\quad= \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{t}{S_{t,i-1}(\hat{\delta}_{1,n})} - \sum_{k=m}^{nm}\frac{k+\delta_1}{k+\hat{\delta}_{1,n}}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right) \\
&\quad= \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{t}{S_{t,i-1}(\hat{\delta}_{1,n})} - \frac{t}{S_{t,i-1}(\delta_1)}\right) \\
&\qquad+ \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\frac{t}{S_{t,i-1}(\delta_1)} - \sum_{k=m}^{nm}\frac{k+\delta_1}{k+\hat{\delta}_{1,n}}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right) \\
&\quad= (\delta_1 - \hat{\delta}_{1,n})\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t^2}{S_{t,i-1}(\delta_1)S_{t,i-1}(\hat{\delta}_{1,n})} \\
&\qquad+ \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{1}{S_{t,i-1}(\delta_1)}\sum_{k=m}^{nm}N_k(G_{t,i-1})\left(1 - \frac{k+\delta_1}{k+\hat{\delta}_{1,n}}\right) \\
&\quad= (\delta_1 - \hat{\delta}_{1,n})\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left[\frac{t^2}{S_{t,i-1}(\delta_1)S_{t,i-1}(\hat{\delta}_{1,n})} - \frac{1}{S_{t,i-1}(\delta_1)}\sum_{k=m}^{nm}\frac{N_k(G_{t,i-1})}{k+\hat{\delta}_{1,n}}\right] \\
&\quad= (\delta_1 - \hat{\delta}_{1,n})\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left[\frac{t}{S_{t,i-1}(\hat{\delta}_{1,n})} - \sum_{k=m}^{nm}\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}\right].
\end{aligned}$$

Thus:

$$\sqrt{n-\tau_n}(\delta_1-\hat{\delta}_{1,n})=\frac{\frac{1}{\sqrt{n-\tau_n}}\sum_{k=m}^{nm}\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}|\mathcal{F}_{t,i-1}\right]\right)}{k+\hat{\delta}_{1,n}}}{\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left[\frac{t}{S_{t,i-1}(\hat{\delta}_{1,n})}-\sum_{k=m}^{\infty}\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}\right]}.$$

We will show that the numerator of the previous display is asymptotically normal and the denominator converges to a positive constant. Asymptotic normality of the estimator follows. Let

$$\begin{aligned}
A(G_n)&=\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left[\frac{t}{S_{t,i-1}(\hat{\delta}_{1,n})}-\sum_{k=m}^{\infty}\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}\right],\\
B(G_n)&=\frac{1}{\sqrt{n-\tau_n}}\sum_{k=m}^{nm}\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{k+\hat{\delta}_{1,n}},\\
\tilde{B}(G_n)&=\frac{1}{\sqrt{n-\tau_n}}\sum_{k=m}^{nm}\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{k+\delta_1}.
\end{aligned}$$

Convergence of $A(G_n)$. First, note that

$$\begin{aligned}
\frac{t}{(2m+\hat{\delta}_{1,n})t-2m+i-1}-\sum_{k=m}^{\infty}\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}&=\left(\frac{t}{(2m+\hat{\delta}_{1,n})t-2m+i-1}-\frac{1}{2m+\hat{\delta}_{1,n}}\right)\\
&+\left(\frac{1}{2m+\hat{\delta}_{1,n}}-\frac{1}{2m+\delta_1}\right)\\
&-\sum_{k=m}^{\infty}\left(\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}-\frac{N_k(G_{t,i-1})}{t(k+\delta_1)}\right)\\
&-\sum_{k=m}^{nm}\left(\frac{N_k(G_{t,i-1})}{t(k+\delta_1)}-\frac{p_k(\delta_0)}{k+\delta_1}\right)\\
&\frac{1}{2m+\delta_1}-\sum_{k=m}^{nm}\frac{p_k(\delta_0)}{k+\delta_1}.
\end{aligned}$$

We have :

$$\frac{\tau_n+1}{(2m+\hat{\delta}_{1,n})n-m-1}\leq\frac{t}{(2m+\hat{\delta}_{1,n})t-2m+i-1}\leq\frac{n}{(2m+\hat{\delta}_{1,n})(\tau_n+1)-2m}.$$

It follows that:

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left[\frac{t}{(2m+\hat{\delta}_{1,n})t-2m+i-1}-\frac{1}{2m+\hat{\delta}_{1,n}}\right]\xrightarrow{\mathbb{P}_1^n}0.$$

Similarly:

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left[\frac{1}{2m+\hat{\delta}_{1,n}}-\frac{1}{2m+\delta_1}\right]\xrightarrow{\mathbb{P}_1^n}0.$$

On the other hand:

$$\begin{aligned}
\sum_{k=m}^{\infty}\left(\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}-\frac{N_k(G_{t,i-1})}{t(k+\delta_1)}\right)&=(\delta_1-\hat{\delta}_{1,n})\sum_{k=m}^{\infty}\frac{N_k(G_{t,i-1})}{t(k+\delta_1)(k+\hat{\delta}_{1,n})}\\
&\leq\frac{\delta_1-\hat{\delta}_{1,n}}{(m+\delta_1)(m+\hat{\delta}_{1,n})}.
\end{aligned}$$

It follows that:

$$\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\left(\frac{N_k(G_{t,i-1})}{t(k+\hat{\delta}_{1,n})}-\frac{N_k(G_{t,i-1})}{t(k+\delta_1)}\right)\xrightarrow{\mathbb{P}_1^n}0.$$

The fourth term is smaller than:

$$\begin{aligned}&\frac{1}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\sum_{k=m}^{nm}\frac{1}{k+\delta}\left|\frac{N_k(G_{t,i-1})}{t}-p_k(\delta_0)\right|\\&\leq\frac{\log(nm+\delta)}{n-\tau_n}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{t}{S_{t,i-1}(\delta_1)}\max_{m\leq k\leq nm}\left|\frac{N_k(G_{t,i-1})}{t}-p_k(\delta_0)\right|\end{aligned}$$

which converges to 0 in probability by Theorem 1.3 in Deijfen et al. [2009]. One can then deduce that:

$$A(G_n)\xrightarrow{\mathbb{P}_1^n}\frac{m}{2m+\delta_1}\left[\frac{1}{2m+\delta_1}-\sum_{k=m}^{\infty}\frac{p_k(\delta_0)}{k+\delta_1}\right]$$

Asymptotic normality of $\tilde{B}(G_n)$. Now we turn our attention to $\tilde{B}$. One has

$$\tilde{B}(G_n)=\sum_{t=1}^{n-\tau_n}\sum_{i=1}^{m}\frac{\sum_{k=m}^{nm}\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{\sqrt{n-\tau_n}(k+\delta_1)}.$$

Let $Y_{t,i}^{(n)}=\sum_{k=m}^{nm}\frac{\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{\sqrt{n-\tau_n}(k+\delta_1)}$. We now need to show that $\tilde{B}(G_n)$ is asymptotically normal. We will apply proposition 3 of Gao and van der Vaart [2017]. To do this it is enough to prove that

$$\begin{aligned}&\sum_{t=1}^{n-\tau_n}\sum_{i=1}^{m}\mathbb{E}_1^n\left[\left(\sum_{k=m}^{nm}\frac{\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{\sqrt{n-\tau_n}(k+\delta_1)}\right)^2\mathbf{1}_{|Y_{t,i}^{(n)}|>\varepsilon}\mid\mathcal{F}_{t,i-1}\right]\xrightarrow{\mathbb{P}_1^n}0,\\&\sum_{t=1}^{n-\tau_n}\sum_{i=1}^{m}\mathbb{E}_1^n\left[\left(\sum_{k=m}^{nm}\frac{\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{\sqrt{n-\tau_n}(k+\delta_1)}\right)^2\mid\mathcal{F}_{t,i-1}\right]\xrightarrow{\mathbb{P}_1^n}\nu_1.\end{aligned}$$

The first convergence result is straightforward since for all $t\in[\![1,n-\tau_n]\!]$ and $i\in[\![1,m]\!]$, the random variables $Y_{t,i}^{(n)}$ are uniformly bounded by $\frac{2}{\sqrt{n-\tau_n}(m+\delta_1)}$. For the second one, we

start by computing the expectations:

$$
\begin{aligned}
&\mathbb{E}_1^n\left[\left(\sum_{k=m}^{nm}\frac{\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)}{\sqrt{n-\tau_n}(k+\delta_1)}\right)^2\mid\mathcal{F}_{t,i-1}\right]\\
&=\sum_{k=m}^{nm}\frac{\mathbb{E}_1^n\left[\left(\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{\mathsf{d}_{\mathrm{V}_{t+\tau_n,i}}=k}\mid\mathcal{F}_{t,i-1}\right]\right)^2\mid\mathcal{F}_{t,i-1}\right]}{(k+\delta_1)^2(n-\tau_n)}\\
&-\frac{1}{n-\tau_n}\sum_{k\neq k'}\frac{\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k'}\mid\mathcal{F}_{t,i-1}\right]}{(k'+\delta_1)(k+\delta_1)}\\
&=\sum_{k=m}^{nm}\frac{\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]}{(k+\delta_1)^2(n-\tau_n)}-\frac{1}{n-\tau_n}\left(\sum_{k=m}^{nm}\frac{\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]}{k+\delta_1}\right)^2\\
&=\sum_{k=m}^{nm}\frac{N_k(G_{t,i-1})}{(n-\tau_n)(k+\delta_1)[S_{t,i-1}(\delta_1)]}-\frac{1}{n-\tau_n}\left(\sum_{k=m}^{nm}\frac{N_k(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right)^2\\
&=\sum_{k=m}^{nm}\frac{N_k(G_{t,i-1})}{(n-\tau_n)(k+\delta_1)[S_{t,i-1}(\delta_1)]}-\frac{1}{n-\tau_n}\left(\frac{t}{S_{t,i-1}(\delta_1)}\right)^2.
\end{aligned}
$$

Using arguments exactly similar to those of the previous paragraphs, we have that the second sum converges under $\mathbb{P}_1^n$ to:

$$
\nu_1=\sum_{k=m}^{\infty}\frac{mp_k(\delta_0)}{(k+\delta_1)(2m+\delta_1)}-\frac{m}{(2m+\delta_1)^2}=\frac{m}{2m+\delta_1}\left(\sum_{k=m}^{\infty}\frac{p_k(\delta_0)}{k+\delta_1}-\frac{1}{2m+\delta_1}\right).
$$

By application of Proposition 3 of Gao and van der Vaart [2017], one has:

$$
\tilde{B}(G_n)\overset{\mathbb{P}_1^n}{\rightsquigarrow}\mathcal{N}(0,\nu_1).
$$

Asymptotic normality of $B(G_n)$. We start by showing that $\tilde{B}(G_n)-B(G_n)$ converges to 0 in probability.

$$
\tilde{B}(G_n)-B(G_n)=\sum_{k=m}^{nm}\frac{1}{(k+\delta_1)(k+\hat{\delta}_{1,n})}\left(\frac{(\hat{\delta}_{1,n}-\delta_1)}{\sqrt{n-\tau_n}}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)\right).
$$

First we show that for all $k\geq m$:

$$
\frac{(\hat{\delta}_{1,n}-\delta_1)}{\sqrt{n-\tau_n}}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)\overset{\mathbb{P}_1^n}{\longrightarrow}0.
$$

Let $\epsilon>0$. One has:

$$
\begin{aligned}
&\mathbb{P}_1^n\left(\left|\frac{(\hat{\delta}_{1,n}-\delta_1)}{\sqrt{n-\tau_n}}\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\left(\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]\right)\right|\geq\epsilon^2\right)\\
&\leq\mathbb{P}_1^n\left(\left|\hat{\delta}_{1,n}-\delta_1\right|\geq\frac{\epsilon}{a}\right)\\
&+\mathbb{P}_1^n\left(\left|\frac{\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\mathbf{1}_{D_{t,i}=k}-\mathbb{E}_1^n\left[\mathbf{1}_{D_{t,i}=k}\mid\mathcal{F}_{t,i-1}\right]}{\sqrt{n-\tau_n}}\right|\geq a\epsilon\right)\\
&\leq\mathbb{P}_1^n\left(\left|\hat{\delta}_{1,n}-\delta_1\right|\geq\frac{\epsilon}{a}\right)+2e^{-\frac{2a^2\epsilon^2}{m}}.
\end{aligned}
$$

Taking the limit of n to $+\infty$ and then the same for a, one obtains the desired convergence. Given that $\hat{\delta}_{1,n}$ is far from $-m$- with probability tending to 1, one obtains by application of the dominated convergence theorem that:

$$\tilde{B}(G_n) - B(G_n) \xrightarrow{\mathbb{P}_1^n} 0.$$

Finally, we apply Slutsky lemma to obtain:

$$\sqrt{n-\tau_n}(\delta_1 - \hat{\delta}_{1,n}) \xrightarrow{\mathbb{P}_1^n} \mathcal{N}(0, \nu_1^{-1})$$

where

$$\nu_1 = \sum_{k=m}^{\infty} \frac{m p_k(\delta_0)}{(k+\delta_1)(2m+\delta_1)} - \frac{m}{(2m+\delta_1)^2}.$$

Proof of Proposition 5.7.3

First, one can easily show that:

$$\sup_{\delta \geq -m+\epsilon} \left| \frac{1}{n-\tau_n} \sum_{t=\tau_n+1}^{n} \sum_{i=1}^{m} \frac{t}{(2m+\delta)t - 2m + i - 1} - \frac{m}{2m+\delta} \right| \to 0.$$

For the other sum, we write:

$$\begin{aligned}
&\left| \frac{1}{n-\tau_n} \sum_{k=m}^{+\infty} \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{k+\delta} - \frac{m}{2m+\delta_1} \sum_{k=m}^{+\infty} \frac{k+\delta_1}{k+\delta} p_k(\delta_0) \right| \\
&\leq \sum_{k=m}^{K} \frac{1}{k+\delta} \left| \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{n-\tau_n} - \frac{m}{2m+\delta_1}(k+\delta_1) p_k(\delta_0) \right| \\
&+ \sum_{k=K+1}^{+\infty} \frac{1}{k+\delta} \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{n-\tau_n} + \frac{m}{2m+\delta_1} \sum_{k=K+1}^{+\infty} \frac{k+\delta_1}{k+\delta} p_k(\delta_0)
\end{aligned}$$

Taking the supremum over δ on both sides, one obtains for $K > 0$:

$$\begin{aligned}
&\sup_{\delta \geq -m+\epsilon} \left| \frac{1}{n-\tau_n} \sum_{k=m}^{+\infty} \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{k+\delta} - \frac{m}{2m+\delta_1} \sum_{k=m}^{+\infty} \frac{k+\delta_1}{k+\delta} p_k(\delta_0) \right| \\
&\leq \sum_{k=m}^{K} \frac{1}{k-m+\epsilon} \left| \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{n-\tau_n} - \frac{m}{2m+\delta_1}(k+\delta_1) p_k(\delta_0) \right| \\
&+ \frac{1}{n-\tau_n} \sum_{t=\tau_n+1}^{n} \sum_{i=1}^{m} \sum_{k=K+1}^{+\infty} \frac{1}{k-m+\epsilon} \mathbf{1}_{D_{t,i}=k} + \frac{m}{2m+\delta_1} \sum_{k=K+1}^{+\infty} \frac{k+\delta_1}{k-m+\epsilon} p_k(\delta_0) \\
&\leq \sum_{k=m}^{K} \frac{1}{k-m+\epsilon} \left| \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{n-\tau_n} - \frac{m}{2m+\delta_1}(k+\delta_1) p_k(\delta_0) \right| \\
&+ \frac{m}{K+1+\epsilon-m} + \frac{m}{2m+\delta_1} \sum_{k=K+1}^{+\infty} \frac{k+\delta_1}{k-m+\epsilon} p_k(\delta_0)
\end{aligned}$$

The first term in the upper-bound converges in probability to 0 and the remaining terms can be made arbitrarily small by choosing a large value for K. The convergence in probability result follows:

$$\sup_{\delta \geq -m+\epsilon} \left| \frac{\dot{\ell}_{\tau_n+1:n}(\delta)}{n-\tau_n} - \iota_1'(\delta) \right| \xrightarrow{\mathbb{P}_1^n} 0$$

Proof of Proposition 5.7.4

Based on the expression of the score, we remark that

$$\begin{aligned}\dot{\ell}_{\tau_{n+1}:n}(\delta') &= \sum_{k=m}^{nm} \frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{k+\delta'} - \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m} \frac{t}{(2m+\delta')t - 2m + i - 1} \\ &\geq \frac{N_{>m}(G_n) - N_{>m}(G_{\tau_n})}{m+\delta'} - \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m} \frac{t}{mt} \\ &= \frac{N_{>m}(G_n) - N_{>m}(G_{\tau_n})}{m+\delta'} - \Delta_n.\end{aligned}$$

Further notice that almost-surely letting $D_{t,i} = \mathsf{d}_{G_{t,i-1}}(V_{t,i})$

$$N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m} \mathbf{1}(D_{t,i} = m).$$

Hence with $\mathcal{F}_{t,i-1} = \sigma(G_{t,i-1})$

$$\begin{aligned}\mathbb{E}_1^n\big(N_{>m}(G_n) - N_{>m}(G_{\tau_n})\big) &= \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{n} \mathbb{E}_1^n\Big(\mathbb{P}_1^n(D_{t,i} = m \mid \mathcal{F}_{t,i-1})\Big) \\ &= \sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m} \mathbb{E}_1^n\Big(N_m(G_{t,i-1})\frac{m+\delta_1}{S_{t,i-1}(\delta_1)}\Big) \\ &= (m+\delta_1)\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m} \frac{\mathbb{E}_1^n(N_m(G_{t,i-1}))}{S_{t,i-1}(\delta_1)}\end{aligned}$$

Since only m edges are added at each instant t, we deduce that $N_m(G_{t,i-1}) \geq N_m(G_t) - m$ for all $t \in [\![\tau_n + 1, n]\!]$ and all $i = 1, \ldots, m$. Furthermore $\mathbb{E}(N_m(G_t)) \asymp t p_m(\delta_0)$ (see for instance the computations in Bet et al. [2025]). Hence we deduce that $\mathbb{E}_1^n\big(N_{>m}(G_n) - N_{>m}(G_{\tau_n})\big) \geq C\Delta_n$ for a constant C depending only on (δ_0, δ_1) and m. A standard concentration argument shows that $\dot{\ell}_{\tau_n+1:n}(\delta_\prime) \geq \Delta_n$ with probability going to one provided $\varepsilon > 0$ is taken small enough.

Proof of Proposition 5.7.5

We first show that:

$$\mathbb{P}_1^n\left(N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = 0\right) \to 0.$$

The starting point is

$$\begin{aligned}\mathbb{P}_1^n\left(N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = 0 \mid \mathcal{F}_{\tau_n}\right) &= \mathbb{P}_1^n\left(\cap_{t=\tau_n+1}^{n}\cap_{i=1}^{m}\left\{\mathsf{d}_{G_{t,i-1}}(V_{t,i}) \neq m\right\} \mid \mathcal{F}_{\tau_n}\right) \\ &= \mathbb{E}_1^n\left[\prod_{t=\tau_n+1}^{n}\prod_{i=1}^{m}\left(1 - \frac{(m+\delta_1)N_m(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right) \mid \mathcal{F}_{\tau_n}\right] \\ &\leq \mathbb{E}_1^n\left[\exp\left(-\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{(m+\delta_1)N_m(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right) \mid \mathcal{F}_{\tau_n}\right].\end{aligned}$$

It follows that:

$$\mathbb{P}_1^n\left(N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = 0\right) \leq \mathbb{E}_1^n\left[\exp\left(-\sum_{t=\tau_n+1}^{n}\sum_{i=1}^{m}\frac{(m+\delta_1)N_m(G_{t,i-1})}{S_{t,i-1}(\delta_1)}\right)\right] \to 0.$$

Note that when $N_{>m}(G_n) - N_{>m}(G_{\tau_n}) \geq 1$, there exists a deterministic $\eta_0 > 0$ such that $\hat{\delta}_{1,n} > -m + \eta_0$. Finally,

$$\begin{aligned}\mathbb{P}_1^n\left(\left|\iota_1'(\hat{\delta}_{1,n})\right| \geq \epsilon\right) &\leq \mathbb{P}_1^n\left(\left|\iota_1'(\hat{\delta}_{1,n})\right| \geq \epsilon, N_{>m}(G_n) - N_{>m}(G_{\tau_n}) \geq 1\right) \\ &\quad + \mathbb{P}_1^n\left(N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = 0\right) \\ &\leq \mathbb{P}_1^n\left(\sup_{\delta \geq -m+\eta_0}\left|\iota_1'(\delta) - \frac{\dot{\ell}_{\tau_n+1\to n}(\delta)}{n-\tau_n}\right| \geq \epsilon\right) + \mathbb{P}_1^n\left(N_{>m}(G_n) - N_{>m}(G_{\tau_n}) = 0\right).\end{aligned}$$

It follows by Proposition 5.7.3 that $\iota_1'(\hat{\delta}_{1,n}) \xrightarrow{\mathbb{P}_1^n} 0$.

5.7.4 Proof of Theorem 5.3.8

Theorem 5.3.8 is a direct consequence of the following proposition.

Proposition 5.7.6. *Let $\delta_0 \neq \delta_1$. For every increasing sequence $\tau_n < n$ such that $\Delta_n \to \infty$, one has:*

$$\begin{aligned}\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_0)}}\right) &\xrightarrow[n\to+\infty]{\mathbb{P}_0^n} -\ell_\infty^0 < 0 \\ \frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_0)}}\right) &\xrightarrow[n\to+\infty]{\mathbb{P}_1^n} \ell_\infty^1 > 0\end{aligned}$$

where ℓ_∞^0 and ℓ_∞^1 are defined in Propositions 5.7.1 and 5.7.2.

Proof. Thanks to Proposition 5.7.1 and Proposition 5.7.2, one only needs to show that:

$$\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_0)}}(G_n)\right) - \frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_0)}}(G_n)\right) \xrightarrow{\mathbb{P}_\ell^n} 0$$

for $\ell \in \{0,1\}$. Using the expression of the likelihood ratio in Lemma A.4, one has:

$$\begin{aligned}&\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_0)}}(G_n)\right) - \frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_0)}}(G_n)\right) \\ &\sim m\log\left(\frac{2m+\hat{\delta}_0}{2m+\delta_0}\frac{2m+\delta_1}{2m+\hat{\delta}_{1,n}}\right) + \sum_{k=m}^{nm}\frac{N_{>k}(G_n) - N_{>k}(G_{\tau_n})}{n-\tau_n}\log\left(\frac{k+\hat{\delta}_{1,n}}{k+\delta_1}\frac{k+\delta_0}{k+\hat{\delta}_0}\right).\end{aligned}$$

Using Proposition 5.7.5 and the arguments used in the proof of Proposition 5.7.1 and Proposition 5.7.2, one can easily deduce that:

$$\frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_{1,n})}}{\mathrm{d}Q^n_{(\tau_n,\hat{\delta}_0,\hat{\delta}_0)}}(G_n)\right) - \frac{1}{n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_0)}}(G_n)\right) \xrightarrow{\mathbb{P}_\ell^n} 0$$

for $\ell \in \{0,1\}$. □

5.7.5 Proof of Proposition 5.3.9

In what follows, we introduce the random variables $D_{t,i} = \mathsf{d}_{G_{t,i-1}}(\mathrm{V}_{t,i})$ and the filtrations $\mathcal{F}_t = \sigma(G_0, \ldots, G_t)$ and $\mathcal{F}_{t,i-1} = \sigma(G_{t,0}, \ldots, G_{t,i-1})$ as in Gao and van der Vaart [2017] to

simplify the notations. Let $\bar{\tau}_n > \tau_n$, then

$$\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}} = \prod_{t=\tau_n+1}^{\bar{\tau}_n}\prod_{l=1}^{m}\left(\frac{(2m+\delta_1)t-2m+l-1}{(2m+\delta_0)t-2m+l-1}\right)\prod_{k=m}^{nm}\left(\frac{k+\delta_0}{k+\delta_1}\right)^{N_{>k}(G_{\bar{\tau}_n})-N_{>k}(G_{\tau_n})}$$

$$\frac{1}{\bar{\tau}_n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\right) = m\log\left(\frac{2m+\delta_1}{2m+\delta_0}\right)+\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\frac{N_{>k}(G_{\bar{\tau}_n})-N_{>k}(G_{\tau_n})}{\bar{\tau}_n-\tau_n}+O\left(\frac{1}{n}\right).$$

On the other hand,

$$\begin{aligned}
&\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\left(N_{>k}(G_{\bar{\tau}_n})-N_{>k}(G_{\tau_n})\right) = \sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}^n_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)\\
&+\sum_{k=m}^{nm}\left(\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t}-\frac{(k+\delta_1)p_k}{2m+\delta_1}\right)\\
&+\sum_{k=m}^{nm}\left(\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t-2m+l-1}-\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t}\right)\\
&+(\delta_0-\delta_1)\sum_{k=m}^{nm}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{N_k(t,l-1)}{(2m+\delta_1)t-2m+l-1}-\frac{p_k}{2m+\delta_1}\right)\\
&+m(\bar{\tau}_n-\tau_n)\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\frac{(k+\delta_1)p_k}{2m+\delta_1}.
\end{aligned}$$

Since

$$\begin{aligned}
&\left|\sum_{k=m}^{nm}\left(\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t-2m+l-1}-\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t}\right)\right|\\
&\le\sum_{k=m}^{nm}\left|\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right|(k+\delta_1)N_k(t,l-1)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\frac{2m-l+1}{(2m+\delta_1)t\left((2m+\delta_1)t-2m+l-1\right)}\\
&\lesssim\sum_{k=m}^{nm}\left|\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right|\frac{n(\bar{\tau}_n-\tau_n)}{\tau_n^2}\\
&=O\left(\frac{\bar{\tau}_n-\tau_n}{n}\right)
\end{aligned}$$

and

$$\begin{aligned}
\sum_{k=m}^{nm}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{N_k(t,l-1)}{(2m+\delta_1)t-2m+l-1}-\frac{p_k}{2m+\delta_1}\right) &= \sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{t}{(2m+\delta_1)t-2m+l-1}-\frac{1}{2m+\delta_1}\right)\\
&=O\left(\frac{\bar{\tau}_n-\tau_n}{n}\right)
\end{aligned}$$

and

$$m(\bar{\tau}_n-\tau_n)\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\frac{(k+\delta_1)p_k}{2m+\delta_1} = m(\bar{\tau}_n-\tau_n)\sum_{k=m}^{+\infty}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\frac{(k+\delta_1)p_k}{2m+\delta_1}+O\left(\frac{\bar{\tau}_n-\tau_n}{n}\right),$$

one obtains

$$\frac{1}{\bar{\tau}_n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\right)=\frac{1}{\bar{\tau}_n-\tau_n}\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}^n_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)$$
$$+\sum_{k=m}^{nm}\left(\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right)\frac{1}{\bar{\tau}_n-\tau_n}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t}-\frac{(k+\delta_1)p_k}{2m+\delta_1}\right)$$
$$+m\log\left(\frac{2m+\delta_1}{2m+\delta_0}\right)+m\sum_{k=m}^{+\infty}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\frac{(k+\delta_1)p_k}{2m+\delta_1}+O\left(\frac{1}{n}\right).$$

Let

$$\ell_\infty=m\log\left(\frac{2m+\delta_1}{2m+\delta_0}\right)+\sum_{k=m}^{+\infty}\frac{m(k+\delta_1)}{2m+\delta_1}p_k\log\left(\frac{k+\delta_0}{k+\delta_1}\right)<0.$$

One has:

$$\frac{1}{\bar{\tau}_n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\right)-\ell_\infty=\frac{1}{\bar{\tau}_n-\tau_n}\sum_{k=m}^{nm}\log\left(\frac{k+\delta_0}{k+\delta_1}\right)\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}^n_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)$$
$$+\sum_{k=m}^{nm}\left(\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right)\frac{1}{\bar{\tau}_n-\tau_n}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\frac{(k+\delta_1)N_k(t,l-1)}{(2m+\delta_1)t}-\frac{(k+\delta_1)p_k}{2m+\delta_1}\right)+O\left(\frac{1}{n}\right)$$
$$\leq\log\left(\frac{nm+\delta_0}{m+\delta_1}\right)\sup_{m\leq k\leq nm}\left|\frac{1}{\bar{\tau}_n-\tau_n}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}^n_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)\right|$$
$$+\sup_{t,l,k}\left|\frac{N_k(t,l-1)}{t}-p_k\right|\sum_{k=m}^{nm}\left|\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right|\frac{k+\delta_1}{2m+\delta_1}+O\left(\frac{1}{n}\right)$$

Since $(\exists C=C(m,\delta_0,\delta_1)>0)$ such that $\sum_{k=m}^{nm}\left|\log\left(\frac{k+\delta_0}{k+\delta_1}\right)-\frac{\delta_0-\delta_1}{k+\delta_1}\right|\frac{k+\delta_1}{2m+\delta_1}\leq C\log(n)$ and $\log\left(\frac{nm+\delta_0}{m+\delta_1}\right)\leq C\log(n)$ one gets:

$$\frac{1}{\bar{\tau}_n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\right)\leq\ell_\infty+C\log(n)\sup_{t,l,k}\left|\frac{N_k(t,l-1)}{t}-p_k\right|+O\left(\frac{1}{n}\right)$$
$$+C\log(n)\sup_{m\leq k\leq nm}\left|\frac{1}{\bar{\tau}_n-\tau_n}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}^n_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)\right|$$

where $\sup_{t,l,k}$ corresponds to the supremum over $t\in[\![\tau_n+1,\bar{\tau}_n]\!]$, $l\in[\![1,m]\!]$ and $k\in[\![m,nm]\!]$. It follows that there exists $C'>0$ such that:

$$\mathbb{P}^n_1\left(\sup_{\bar{\tau}_n-\tau_n\geq\kappa_n}\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\geq1\right)\leq\sum_{\bar{\tau}_n-\tau_n\geq\kappa_n}\mathbb{P}^n_1\left(\frac{1}{\bar{\tau}_n-\tau_n}\log\left(\frac{\mathrm{d}Q^n_{(\bar{\tau}_n,\delta_0,\delta_1)}}{\mathrm{d}Q^n_{(\tau_n,\delta_0,\delta_1)}}\right)\geq0\right)$$
$$\leq\sum_{\bar{\tau}_n-\tau_n\geq\kappa_n}\mathbb{P}^n_1\left(\sup_{m\leq k\leq nm}\left|\frac{1}{\bar{\tau}_n-\tau_n}\sum_{t=\tau_n+1}^{\bar{\tau}_n}\sum_{l=1}^{m}\left(\mathbf{1}_{D_{t,l}=k}-\mathbb{E}_1\left[\mathbf{1}_{D_{t,l}=k}\mid\mathcal{F}_{t,l-1}\right]\right)\right|\geq\frac{-\ell_\infty+O(\frac{1}{n})}{2C\log(n)}\right)$$
$$+\sum_{\bar{\tau}_n-\tau_n\geq\kappa_n}\mathbb{P}^n_1\left(\sup_{t,l,k}\left|\frac{N_k(t,l-1)}{t}-p_k\right|\geq\frac{-\ell_\infty+O(\frac{1}{n})}{2C\log(n)}\right)$$
$$\lesssim n\left(n\exp\left(-C'\frac{\bar{\tau}_n-\tau_n}{\log(n)^2}\right)+o\left(\frac{1}{n}\right)\right)$$
$$\lesssim n\left(n\exp\left(-C'\frac{\kappa_n}{\log(n)^2}\right)+o\left(\frac{1}{n}\right)\right)$$

where we have used Hoeffding-Azuma inequality for the first term and Theorem 8.3 of van der Hofstad [2016] (or Proposition 2.1 of Deijfen et al. [2009]) for the second term. Note that these results were stated only in the case of no change. However, they should remain valid for our model. Using an exactly similar argument for $\bar{\tau}_n - \tau_n < -\kappa_n$, we finally obtain:

$$\tilde{\mathbb{P}}_1^n \left(\sup_{|\bar{\tau}_n - \tau_n| \geq \kappa_n} \frac{\mathrm{d}Q^n_{(\bar{\tau}_n, \delta_0, \delta_1)}}{\mathrm{d}Q^n_{(\tau_n, \delta_0, \delta_1)}} \geq 1 \right) = O\left(n^2 \exp\left(-C' \frac{\kappa_n}{\log(n)^2} \right) + o(1) \right)$$

Taking $\kappa_n \asymp \log(n)^3$, one obtains:

$$\tilde{\mathbb{P}}_1^n \left(\sup_{|\bar{\tau}_n - \tau_n| \gtrsim \kappa_n} \frac{\mathrm{d}Q^n_{(\bar{\tau}_n, \delta_0, \delta_1)}}{\mathrm{d}Q^n_{(\tau_n, \delta_0, \delta_1)}} \geq 1 \right) \to 0.$$

One obtains a localization error smaller than $\log(n)^3$.

Chapter 6

Conclusion and perspectives

In this thesis, we investigate a range of inference problems across different types of data. We began by examining clustering under both i.i.d. and HMM models in a non-parametric framework, then refined these results in the slowly mixing regime within the Gaussian setting. Finally, we turned to the problems of change-point detection and localization in the affine preferential attachment random graph model. Below, we highlight possible extensions and future research directions that build on the contributions of this work.

- **The problem of clustering:** In the context of clustering under a general non-parametric HMM model, our findings on the Bayes risk of clustering are valid only in the mixing regime (Chapter 3). While these results were further refined in the slowly mixing regime for the Gaussian case (Chapter 4), a comprehensive understanding of the Bayes risk of clustering in the non-parametric slowly mixing regime is still an open problem. Many other problems remain open in this setting:
  - **Optimal excess risk:** While we upper-bound the excess risk of the plug-in procedure through the estimation error (inducing thus a polynomial decay of the excess risk), we believe this rate should be improved to an exponential decay. Lemma 3.5.1 presents the first steps for this proof. The optimal excess risk is still unclear in the general non-parametric case.
  - **Alternatives to plug-in:** Plug-in clustering depends heavily on parameter estimation, with its performance tied to the quality of the estimates. A natural next step is to design clustering methods that exploit the nonparametric identifiability of HMMs directly, without first estimating the parameters.
  - **High dimension:** In our analysis, we restrict attention to the Bayes risk, which is minimized by an oracle with full knowledge of the model parameters. As a consequence, the dimension of the observation space does not appear in the expression of the bounds. To capture the impact of high dimensionality on the complexity of clustering, one must instead consider the minimax risk of clustering (see Ndaoud [2022]), which does not assume access to the true parameters. Characterizing how this risk depends on the emission distribution, the mixing properties of the hidden Markov chain, the observation dimension and the sample size would shed light on the interplay among all model parameters and provide a thorough understanding of the difficulty of the problem of clustering.
  - **Estimation of the mixing parameter:** A very interesting question is to characterize the minimax rate of estimation of the mixing parameter δ in the high dimensional two component Gaussian Hidden Markov Model. When the

signal θ is known, δ can be estimated at the parametric rate as shown in Zhang and Weinberger [2022]. However, the optimal rate of estimation is still an open question.

- **Full adaptation:** The procedures of clustering proposed in Chapter 4 are not adaptive to δ and when this parameter is unknown, one has to pay a price of $\mathcal{O}(\log(n))$. It is important to come up with a fully adaptive clustering procedure with optimal performance.

- **The problems of change-point detection and localization:** In the context of change-point detection for preferential attachment random graph models, our earlier work Kaddouri et al. [2025] partially resolved a conjecture that was later fully settled in Du et al. [2025]. Their proof builds on and refines our approach, introducing two new key ideas:

  - **Interpolation:** Let $\mathbb{P}_n$ denote the distribution of the unlabeled preferential attachment graph with n vertices under the null hypothesis, and let $\mathbb{Q}_{n,\tau_n}$ be its distribution under the alternative, where the change occurs at $\tau_n = n - \Delta_n$. The interpolation method constructs a chain of intermediate distributions linking the null and alternative:

$$\mathbb{P}_n = \mathbb{Q}_{n,n} \;\to\; \mathbb{Q}_{n,n-1} \;\to\; \mathbb{Q}_{n,n-2} \;\to\; \cdots \;\to\; \mathbb{Q}_{n,\tau_n}.$$

  Applying the triangle inequality along this path yields

$$\mathrm{TV}(\mathbb{P}_n, \mathbb{Q}_{n,\tau_n}) \le \sum_{k=1}^{n-\tau_n} \mathrm{TV}(\mathbb{Q}_{n,n-k+1}, \mathbb{Q}_{n,n-k})\,.$$

  By the data-processing inequality,

$$\mathrm{TV}(\mathbb{P}_n, \mathbb{Q}_{n,\tau_n}) \le \sum_{k=1}^{n-\tau_n} \mathrm{TV}(\mathbb{P}_{n-k+1}, \mathbb{Q}_{n-k+1,n-k})\,.$$

  Hence, the task is to prove that

$$\forall n' \in [\![\tau_n + 1, n]\!], \quad \mathrm{TV}\big(\mathbb{P}_{n'}, \mathbb{Q}_{n',n'-1}\big) = o\Big(\tfrac{1}{\Delta_n}\Big)\,.$$

  Without loss of generality, it suffices to consider the simplified case $n' = n$ and $\tau_n = n - 1$, where the change-point occurs exactly one step before the terminal time. This special case also leads to a more tractable expression for the likelihood ratio.

  - **Second-moment bound via the Efron–Stein inequality:** The Efron–Stein inequality asserts that for independent random variables $Y_1, \ldots, Y_k$ and any measurable function $f : \mathbb{R}^k \to \mathbb{R}$,

$$\mathrm{Var}[f(Y_1, \ldots, Y_k)] \le \sum_{i=1}^{k} \mathbb{E}\Big[\big(f(Y_1, \ldots, Y_i, \ldots, Y_k) - f(Y_1, \ldots, \tilde{Y}_i, \ldots, Y_k)\big)^2\Big]\,,$$

  where $\tilde{Y}_i$ is an independent copy of Y_i. To apply this inequality in bounding the second moment of the likelihood ratio, one must first express the likelihood ratio as a functional of independent random variables. Since the preferential attachment process admits such a representation, the Efron–Stein inequality

becomes a powerful tool in the analysis. Figure 6.1 displays a typical preferential attachment graph with $m = 1$ and $n^{1/3} \lesssim \Delta_n \lesssim n^{1/2}$. Contrary to our proof which relies on revealing the order of arrival of all the nodes except those in bold, the proof of the conjecture in Du et al. [2025] reveals less information: only the order of arrival of normal nodes is revealed. This is the reason for which their proof works.

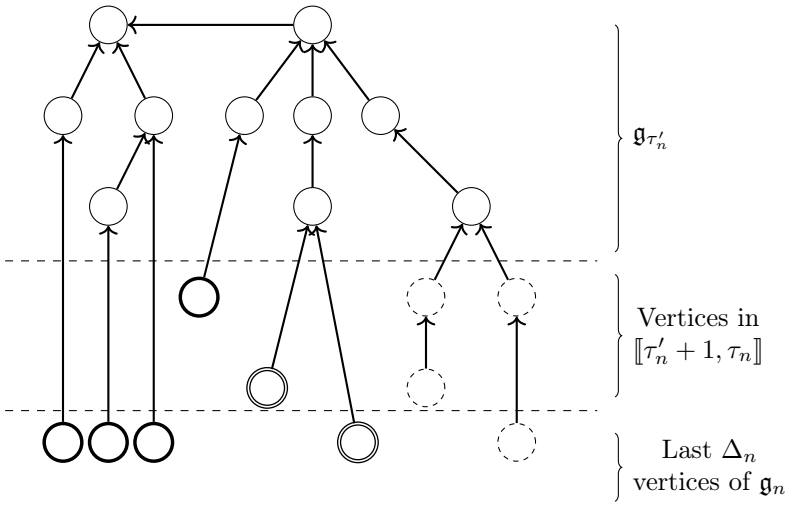


Figure 6.1: Typical preferential attachment graph $\mathfrak{g}_n$with $m = 1$ when $n^{1/3} \lesssim \Delta_n \lesssim n^{1/2}$.

Future research on changepoint detection in random graphs could therefore move in several promising directions. One natural extension is to consider models where the probability of connexion depends on more general attachment functions, rather than being restricted to an affine form. One might investigate general non-parametric transitions from one function to another, thereby modeling very general structural changes in the network's evolution. Another direction concerns the initial degree distribution. Most studies on changepoint detection have concentrated on models with fixed initial degrees m. This turns out to be a very restrictive assumption for modelling many real-world networks. The degree sequence of such graphs was studied in Deijfen et al. [2009] and extending the study of the problem of change-point detection to the case of random initial degrees would be an interesting direction too. Concerning the problem of change-point localization and parameter estimation, we believe that the result of Proposition 5.3.9 can be generalized to include the simultaneous localisation of all model parameters δ_0, δ_1 and τ_n, and eventually reduce the error of localization from $\log(n)^3$ to a constant error, but at the cost of some tedious calculations.

Bibliography

K. Abraham, I. Castillo, and E. Gassiat. Multiple testing in nonparametric hidden markov models: An empirical bayes approach. *J. Mach. Learn. Res.*, 23(1):4061–4117, 2022.

K. Abraham, E. Gassiat, and Z. Naulet. Fundamental limits for learning hidden Markov model parameters. *IEEE Trans. Inform. Theory*, 69(3):1777–1794, 2023. ISSN 0018-9448,1557-9654. doi: 10.1109/tit.2022.3213429.

K. Abraham, E. Gassiat, and Z. Naulet. Frontiers to the learning of nonparametric hidden markov models. *Journal of Machine Learning Research*, 26(155):1–75, 2025. URL http://jmlr.org/papers/v26/24-2230.html.

L. A. Adamic and B. A. Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000. doi: 10.1126/science.287.5461.2115a.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden markov models. *Biometrika*, 103(2):423–434, 03 2016a. ISSN 0006-3444. doi: 10.1093/biomet/asw001.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016b. ISSN 0006-3444. doi: 10.1093/biomet/asw001.

G. Alexandrovich, H. Holzmann, and A. Leister. Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, 103(2):423–434, 2016c. ISSN 0006-3444. doi: 10.1093/biomet/asw001.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In S. Mannor, N. Srebro, and R. C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 33.1–33.34, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, M. Telgarsky, et al. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435. URL https://doi.org/10.1214/16-AOS1435.

S. Banerjee, S. Bhamidi, and I. Carmichael. Fluctuation bounds for continuous time branching processes and evolution of growing trees with a change point. *Ann. Appl. Probab.*, 33(4):2919–2980, 2023. doi: 10.1214/22-AAP1881. URL https://doi.org/10.1214/22-AAP1881.

A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Phys. A: Stat. Mech. Appl.*, 311(3):590–614, 2002. ISSN 0378-4371. doi: https://doi.org/10.1016/S0378-4371(02)00736-7.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999a. doi: 10.1126/science.286.5439.509.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999b.

O. Barndorff-Nielsen. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, 12:115–121, 1965.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

G. Bet, K. Bogerd, R. M. Castro, and R. van der Hofstad. Detecting a late changepoint in the preferential attachment model. *Bernoulli, to appear*, 2025.

S. Bhamidi, J. Jin, and A. Nobel. Change point detection in network models: Preferential attachment and long range dependence. *Ann. Appl. Probab.*, 28(1):35–78, 2018. doi: 10.1214/17-AAP1297. URL https://doi.org/10.1214/17-AAP1297.

B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24:5–34, 2004.

B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Combinatorica*, 21(3):311–340, 2001. doi: 10.1007/s004930100012.

B. Bollobás, C. Borgs, J. Chayes, and O. Riordan. Directed scale-free graphs. In *Proc. Annu. ACM-SIAM Symp. Discrete Algorithms*, SODA '03, pages 132–139, USA, 2003. Society for Industrial and Applied Mathematics. ISBN 0898715385.

L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Ann. Statist.*, 34(3):1204–1232, 2006. doi: 10.1214/009053606000000353.

A. M. Brandenberger, L. Devroye, and M. K. Goh. Root estimation in galton–watson trees. *Random Structures & Algorithms*, 61(3):520–542, 2022.

S. Briend, C. Giraud, G. Lugosi, and D. Sulem. Estimating the history of a random recursive tree. *Bernoulli, to appear*, 2025.

S. Bubeck, E. Mossel, and M. Z. Rácz. On the influence of the seed graph in the preferential attachment model. *IEEE TNSE*, 2(1):30–39, 2015. doi: 10.1109/TNSE.2015.2397592.

S. Bubeck, L. Devroye, and G. Lugosi. Finding adam in random growing trees. *Random Structures & Algorithms*, 50(2):158–172, 2017.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, 2005. ISBN 0387402640.

D. Cirkovic, T. Wang, and S. I. Resnick. Preferential attachment with reciprocity: properties and estimation. *J. Complex Netw.*, 11(5), 09 2023a. ISSN 2051-1329. doi: 10.1093/comnet/cnad031. URL https://doi.org/10.1093/comnet/cnad031.

D. Cirkovic, T. Wang, and X. Zhang. Likelihood-based inference for random networks with changepoints. *arXiv preprint arXiv:2206.01076*, 2023b.

A. Contat, N. Curien, P. Lacroix, E. Lasalle, and V. Rivoirard. Eve, adam and the preferential attachment tree. *Probability Theory and Related Fields*, 190(1–2):321–336, 2024. doi: 10.1007/s00440-023-01253-1. URL https://doi.org/10.1007/s00440-023-01253-1.

H. Crane and M. Xu. Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):639–668, 2021. doi: https://doi.org/10.1111/rssb.12428. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12428.

Y. De Castro. Nonparametric hmm. https://github.com/ydecastro/nonparametric-hmm, 2016.

Y. De Castro, É. Gassiat, and C. Lacour. Minimax adaptive estimation of nonparametric hidden markov models. *J. Mach. Learn. Res.*, 17(111):1–43, 2016. URL http://jmlr.org/papers/v17/15-381.html.

Y. De Castro, É. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden markov models. *IEEE Trans. Inform. Theory*, 63(8):4758–4777, 2017. doi: 10.1109/TIT.2017.2696959.

M. Deijfen, H. van den Esker, R. van der Hofstad, and G. Hooghiemstra. A preferential attachment model with random initial degrees. *Ark. Mat.*, 47(1):41–72, 2009. doi: 10.1007/s11512-007-0067-4. URL https://doi.org/10.1007/s11512-007-0067-4.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

S. Dereich and P. Mörters. Random networks with sublinear preferential attachment: Degree evolutions. *Electronic Journal of Probability*, 14:1222–1267, 2009. doi: 10.1214/EJP.v14-647. URL https://doi.org/10.1214/EJP.v14-647.

L. Devroye and J. Lu. The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures & Algorithms*, 7(1):1–14, 1995. doi: 10.1002/rsa.3240070102.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, New York, NY, 1 edition, 1996. ISBN 978-0-387-94618-4. doi: 10.1007/978-1-4612-0711-5. URL https://doi.org/10.1007/978-1-4612-0711-5.

S. Dommers, R. v. d. Hofstad, and G. Hooghiemstra. Diameters in preferential attachment models. *J. Stat. Phys.*, 139(1):72–107, 2010. ISSN 0022-4715.

H. Du, S. Gong, and J. Xu. A proof of the changepoint detection threshold conjecture in preferential attachment models. In N. Haghtalab and A. Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1559–1563. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/du25a.html.

R. Durrett. *Random Graph Dynamics*. Cambridge University Press, 2006.

M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, SIGCOMM '99, pages 251–262, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131356. doi: 10.1145/316188.316229. URL https://doi.org/10.1145/316188.316229.

Y. Fei and Y. Chen. Hidden integrality of sdp relaxations for sub-gaussian mixture models. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1931–1965. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/fei18a.html.

G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973. doi: 10.1109/PROC.1973.9030.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, 2006. ISBN 9780387357680. doi: 10.1007/0-387-35768-2. URL https://doi.org/10.1007/0-387-35768-2.

S. Frühwirth-Schnatter, G. Celeux, and C. P. Robert, editors. *Handbook of Mixture Analysis*. Chapman and Hall/CRC, New York, 2019. ISBN 9780429055911. doi: 10.1201/9780429055911. URL https://doi.org/10.1201/9780429055911.

C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *Ann. Statist.*, 46(5):2153–2185, 2018. ISSN 0090-5364,2168-8966. doi: 10.1214/17-AOS1615.

F. Gao and A. van der Vaart. On the asymptotic normality of estimating the affine preferential attachment network models with random initial degrees. *Stochastic Process. Appl.*, 127(11):3754–3775, 2017. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2017.03.008.

F. Gao, A. van der Vaart, R. Castro, and R. van der Hofstad. Consistent estimation in general sublinear preferential attachment trees. *Electron. J. Stat.*, 11(2):3979–3999, 2017. doi: 10.1214/17-EJS1356.

E. Gassiat and J. Rousseau. Nonparametric finite translation hidden markov models and extensions. *Bernoulli*, 22(1):193–212, 2016. doi: 10.3150/14-BEJ656.

E. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71, 2016. ISSN 0960-3174. doi: 10.1007/s11222-014-9523-8.

E. Gassiat, I. Kaddouri, and Z. Naulet. Clustering risk in non-parametric hidden markov and i.i.d. models. *The Annals of Statistics, to appear*, 2025.

S. Ghassempour, F. Girosi, and A. Maeder. Clustering multivariate time series using hidden markov models. *International Journal of Environmental Research and Public Health*, 3 (11), 2014. doi: 10.3390/ijerph110302741.

C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, second edition, 2021.

C. Giraud and N. Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Math. Stat. Learn.*, 3(1):317–374, 2018.

B. Grün. Model-based clustering. In *Handbook of mixture analysis*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 157–192. CRC Press, Boca Raton, FL, 2019.

H. Guo, Q. Deng, W. Jia, L. Wang, and C. Sui. Hidden markov model-based modeling and prediction for implied volatility surface. *Journal of Intelligent & Fuzzy Systems*, 45 (6):12381–12394, 2023. doi: 10.3233/JIFS-232139.

P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31(1):201–224, 2003. doi: 10.1214/aos/1046294456.

P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92(3):667–678, 09 2005. ISSN 0006-3444. doi: 10.1093/biomet/92.3.667.

S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

P. Holme and J. Saramäki. *Temporal Network Theory*, volume 2. Springer, 2019.

D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 35(1):224–251, 2007. doi: 10.1214/009053606000001258.

S. Janson. Asymptotic degree distribution in random recursive trees. *Random Structures & Algorithms*, 26(1-2):69–83, 2005. doi: https://doi.org/10.1002/rsa.20046. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20046.

I. Kaddouri, Z. Naulet, and E. Gassiat. On the impossibility of detecting a late changepoint in the preferential attachment random graph model. *Bernoulli, to appear*, 2025.

R. Kannan, H. Salmasian, and S. Vempala. The spectral method for general mixture models. In *International conference on computational learning theory*, pages 444–457. Springer, 2005.

V. Karagulyan and M. Ndaoud. Adaptive mean estimation in the hidden markov subgaussian mixture model. *arXiv preprint arXiv:2406.12446*, 2024.

D. Khiatani and U. Ghose. Weather forecasting using hidden markov model. In *2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)*, pages 220–225, 2017. doi: 10.1109/IC3TSN.2017.8284480.

J. Khim and P.-L. Loh. Confidence sets for the source of a diffusion in regular trees. *IEEE Transactions on Network Science and Engineering*, 4(1):27–40, 2016.

P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Physical Review Letters*, 85(21):4629–4632, 2000.

J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and its Applications*, 18 (2):95–138, 1977. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(77)90069-6.

L. Lehéricy. State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *J. Mach. Learn. Res.*, 19:Paper No. 39, 46, 2018. ISSN 1532-4435,1533-7928.

L. Lehéricy. Consistent order estimation for nonparametric hidden Markov models. *Bernoulli*, 25(1):464–498, 2019.

L. Lehéricy. Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. *Electron. J. Stat.*, 15(2):4916–4965, 2021. ISSN 1935-7524. doi: 10.1214/21-ejs1890.

M. Löffler, A. Y. Zhang, and H. H. Zhou. Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506 – 2530, 2021. doi: 10.1214/20-AOS2044. URL https://doi.org/10.1214/20-AOS2044.

Y. Lu and H. H. Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

A. Marandon, T. Rebafka, and E. Roquain. False clustering rate control in mixture models. *arXiv preprint arXiv:2203.02597*, 2023.

G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000. ISBN 9780471006268.

P. D. McNicholas. Model-based clustering. *Journal of Classification*, 33:331–373, 2016.

M. c. v. Medo, G. Cimini, and S. Gualdi. Temporal effects in the growth of networks. *Phys. Rev. Lett.*, 107:238701, 12 2011. doi: 10.1103/PhysRevLett.107.238701.

M. Meilă. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, pages 577–584, 2005.

M. Meilă and D. Heckerman. An experimental comparison of model-based clustering methods. *Machine learning*, 42:9–29, 2001.

D. G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

T. F. Móri. The maximum degree of the barabási–albert random tree. *Combinatorics, Probability and Computing*, 14(3):339–348, 2005. doi: 10.1017/S0963548304006133.

M. Ndaoud. Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.*, 50(4):2096 – 2126, 2022. doi: 10.1214/22-AOS2178.

M. E. J. Newman. The structure of scientific collaboration networks. *PNAS*, 98(2):404–409, 2001. doi: 10.1073/pnas.98.2.404.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997. doi: 10.1017/CBO9780511810633.

R. Oliveira and J. Spencer. Connectivity transitions in networks with super-linear preferential attachment. *Internet Mathematics*, 2(2):175–185, 2005. doi: 10.1080/15427951.2005.10129101. URL https://doi.org/10.1080/15427951.2005.10129101.

V. Pakštaitė, E. Filatovas, M. Juodis, and R. Paulavičius. Bitcoin price regime shifts: A bayesian mcmc and hidden markov model analysis of macroeconomic influence. *Mathematics*, 13(10):1577, May 2025. doi: 10.3390/math13101577.

D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Prob.*, 20(none):1 – 32, 2015. doi: 10.1214/EJP.v20-4039.

T. Petrie. Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969. doi: 10.1214/aoms/1177697802.

B. Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures & Algorithms*, 5(2):253–374, 1994. doi: 10.1002/rsa.3240050207.

Z. S. Qin, J. Yu, J. Shen, C. A. Maher, M. Hu, S. Kalyana-Sundaram, J. Yu, and A. M. Chinnaiyan. Hpeak: an hmm-based algorithm for defining read-enriched regions in chip-seq data. *BMC Bioinformatics*, 11:369, Jul 2010. doi: 10.1186/1471-2105-11-369.

L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

M. Royer. Adaptive clustering through semidefinite programming. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3a15c7d0bbe60300a39f76f8a5ba6896-Paper.pdf.

R. B. Scharpf, G. Parmigiani, J. Pevsner, and I. Ruczinski. Hidden markov models for the assessment of chromosomal alterations using high-throughput snp arrays. *The Annals of Applied Statistics*, 2(2):687–713, Jun 2008. doi: 10.1214/07-AOAS155.

A. Schliep, A. Schönhuth, and C. Steinhoff. Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, 2003. doi: 10.1093/bioinformatics/btg1036.

Z. Scully. Maximizing the expected deviation of sum of independent bernoulli. *Mathematics Stack exchange*, 2024. URL https://math.stackexchange.com/questions/4961641/maximizing-the-expected-deviation-of-sum-of-independent-bernoulli.

T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, 37(6B):3960 – 3984, 2009. doi: 10.1214/09-AOS700. URL https://doi.org/10.1214/09-AOS700.

H. Teicher. Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 34 (4):1265 – 1269, 1963. doi: 10.1214/aoms/1177703862.

N. M. Temme. *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*. Wiley, 1996. doi: 10.1002/9781118032572.

A. W. v. d. Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

R. van der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016. doi: 10.1017/9781316779422.

R. van der Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2024.

E. Vernet. Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics*, 9(1):717 – 752, 2015. doi: 10.1214/15-EJS1017.

D. Wang, Y. Yu, and A. Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *Ann. Stat.*, 49(1):203–232, 2021. doi: 10.1214/20-AOS1953.

D. J. Watts. *Small Worlds: the Dynamics of Networks Between Order and Randomness*. Princeton University Press, USA, 1999. ISBN 0691005419.

D. J. Watts and S. H. Strogatz. *Collective Dynamics of 'Small-World' Networks*, pages 301–303. Princeton University Press, Princeton, 2006. ISBN 9781400841356. doi: doi: 10.1515/9781400841356.301. URL https://doi.org/10.1515/9781400841356.301.

C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Y. Wu. Lecture notes on information-theoretic methods for high-dimensional statistics, 2017. URL https://yihongwu.stat.yale.edu/teaching/ece598yw-fall17/notes.html. Lecture Notes for ECE598YW (UIUC), Version 16.

Y. Wu and H. H. Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in o $(\sqrt{n})$ iterations. *Mathematical Statistics and Learning*, 4(3), 2021.

S. J. Yakowitz and J. D. Spragins. On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 39(1):209 – 214, 1968. doi: 10.1214/aoms/1177698520. URL https://doi.org/10.1214/aoms/1177698520.

A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block models. *Ann. Statist.*, 44(5):2252–2280, 2016. ISSN 0090-5364,2168-8966. doi: 10.1214/15-AOS1428.

Y. Zhang and N. Weinberger. Mean estimation in high-dimensional binary markov gaussian mixture models. *Advances in Neural Information Processing Systems*, 35:19673–19686, 2022.

Z. Zhao, L. Chen, and L. Lin. Change-point detection in dynamic networks via graphon estimation. *arXiv preprint arXiv:1908.01823*, 2019.

List of Figures

List of Tables