Contributions to ranking problems and high-dimensional change-point detection

Emmanuel Pilliat

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biostatistiques

École doctorale : Information, Structures, Systèmes

Unité de recherche : Institut Montpelliérain Alexander Grothendieck

Contributions to ranking problems and high-dimensional change-point detection

Présentée par Emmanuel Pilliat
le 8 décembre 2023

Sous la direction de Joseph Salmon, Nicolas Verzelen et Alexandra Carpentier

Devant le jury composé de

Sivaraman Balakrishnan	Rapporteur
Claire Boyer	Examinatrice
Cristina Butucea	Examinatrice
Alexandra Carpentier	Co-encadrante
Arnak Dalalyan	Rapporteur
Béatrice Laurent	Présidente du jury
Joseph Salmon	Directeur de thèse
Nicolas Verzelen	Co-encadrant

UNIVERSITÉ DE MONTPELLIER

Remerciements

Je tiens d'abord à remercier ma direction de thèse. Merci Joseph pour tout le temps que tu m'as accordé et pour ton soutien constant tout au long de ce parcours. Alexandra, merci pour tous ces échanges, au cours desquels tu as su me transmettre beaucoup sur la façon d'aborder un problème de recherche. En particulier, je suis maintenant convaincu qu'il faut parfois redécouvrir certains concepts par soi-même pour mieux les intégrer ! J'espère emporter avec moi ta confiance et ta sérénité face aux situations en recherche qui semblent parfois insurmontables. Nicolas, merci pour le temps considérable que tu m'as accordé, que ce soit pour clarifier des concepts théoriques ou pour les aspects plus techniques. Je garde tous tes précieux conseils, qui me guideront dans mes projets futurs, aussi bien sur le plan scientifique que rédactionnel.

Je tiens aussi à remercier Arnak et Sivaraman pour avoir accepté de rapporter mon manuscrit. Merci Arnak pour le temps consacré à une relecture approfondie. Thank you Sivaraman for the time devoted to a thorough review. Je suis également reconnaissant envers Claire, Cristina et Béatrice pour avoir accepté de faire partie de mon jury de thèse et pour votre participation à ma soutenance.

J'exprime aussi mes remerciements envers l'école doctorale EDI2S pour les nombreuses formations enrichissantes auxquelles j'ai pu participer, et à tous les formateurs impliqués. Les méthodes de communication que j'ai apprises me permettront de continuellement remettre en question et d'améliorer mes futures présentations. Je la remercie également pour son soutien dans ma démarche de césure de thèse. Je remercie aussi QRT pour l'opportunité professionnelle offerte durant mes six mois de césure. J'ai pu, notamment grâce à Adrien, découvrir le monde de la finance et développer des compétences pratiques essentielles.

Réaliser ma thèse au sein du département de mathématiques et informatique MISTEA de l'INRAE a été un vrai plaisir. J'ai apprécié partager de nombreux moments conviviaux, que ce soit à la cantine ou aux pauses café, et je remercie tous les collègues de l'unité qui m'ont permis de vivre cette thèse dans les meilleurs conditions. Merci en particulier à Adrien, pour tous les échanges mathématiques ainsi que pour ta relecture de mon introduction.

Je tiens aussi à remercier tous les collègues de l'IMAG avec qui j'ai pu échanger de près ou de loin et qui ont aussi participé au bon déroulement de ma thèse. Je remercie Elodie et Alice pour m'avoir soutenu lors de mes comités de suivi de thèse et pour leur bienveillance. J'ai aussi très apprécié enseigner aux côtés d'Elodie et de Ioan, et je vous remercie pour l'expérience d'enseignement que j'ai eue.

Merci à mes professeurs de classe prépa, Mr. Génaux et Mr. Paviet, ainsi qu'à mon professeur de lycée, Mr. Velikonia. Leurs enseignements m'ont non seulement donné des bases solides, mais également donné l'envie de poursuivre ce parcours académique.

Du côté plus personnel, merci à mes amis Nicolas, Bastien, Julie, Alice, Yoann, Kévin, Lucille et Valentin pour tous les bons moments partagés et toutes les discussions durant ces trois dernières années. Je remercie enfin ma famille pour leur soutien, et en particulier à mes parents, ma soeur ainsi qu'à Simone pour leur participation à ma soutenance. Merci Paloma pour avoir été là dans la durée, et m'avoir rendu heureux papa. Merci enfin Mathis pour cette nouvelle vie, et j'espère te rendre fier le jour où tu pourras lire et comprendre ces lignes !

Contents

Chapter 1

Introduction (French version)

Dans cette thèse, nous explorons deux domaines des statistiques modernes : les problèmes de classement et la détection de points de rupture. Les problèmes de classement ont de nombreuses applications, notamment dans les tournois, les systèmes de vote ou le classement d'experts dans des données de crowdsourcing. De même, la détection des points de rupture joue un rôle crucial dans diverses situations pratiques, telles que le suivi des changements de température, le suivi des cours boursiers ou l'analyse de données génomiques.

Bien que nous abordions ces deux sujets de manière indépendante, notre approche s'inscrit dans un même cadre : les statistiques en grande dimension. En effet, dans les deux cas, le nombre de paramètres inconnus peut être supérieur au nombre d'échantillons. Pour gérer cette complexité et rendre ces problèmes en grande dimension abordable, nous introduisons des hypothèses structurelles spécifiques, ou des contraintes de forme, dans chaque modèle.

La première partie de cette thèse se penche sur les problèmes de classement. Essentiellement, les méthodes de classement visent à trier une collection d'éléments sur la base d'observations bruitées, avec des données potentiellement manquantes. Nous explorerons plusieurs modèles qui diffèrent par leurs contraintes de forme. En particulier, nos contributions se concentrent sur deux modèles où l'objectif est de retrouver une permutation des lignes d'une matrice. Dans Chapter 4, nous supposons que la matrice réordonnée a des colonnes croissante. Ce modèle englobe de nombreux modèles, notamment le classement d'experts au sein d'une foule ou le classement dans les tournois à partir de comparaisons par paire. Dans Chapter 3, en plus de supposer que la matrice réordonnée a ses colonnes croissante, nous supposons que ses lignes le sont aussi. Pour les deux modèles, nous fournissons des algorithmes calculables en temps polynomial qui permettent d'obtenir des garanties optimales dans l'estimation de la permutation.

La deuxième partie porte sur un problème de détection de points de ruptures multiples pour des séries temporelles multivariées. De manière informelle, un point de rupture est un point dans une séquence où les propriétés statistiques des données changent. Dans notre contexte, les observations séquentielles peuvent exister dans un espace en grande dimension et peuvent contenir un nombre arbitraire de points de ruptures. Cela étend en particulier le modèle univarié plus simple, où les données consistent en des observations à valeur réelle. Pour contrôler la complexité du modèle, nous considérons les cas où les variations du signal en grande dimension peuvent être parcimonieuses. Dans un régime parcimonieux, de nombreuses entrées du vecteur représentant les variations sont égales à zéro. Dans Chapter 5, nous établissons les conditions minimales pour que la détection soit possible dans ce cadre. Dans ces conditions, nous fournissons des garanties qui s'adaptent à la parcimonie inconnu et à la distance entre les points de rupture.

Dans le dernier chapitre, nous discutons des axes d'approfondissement de cette thèse, en introduisant trois problèmes. Le premier est lié à l'identification de label en crowdsourcing, le deuxième au classement d'experts lorsque les observations peuvent être choisies séquentiellement, et le dernier à la localisation de points de rupture dans des séries temporelles multivariées.

Dans les sections suivantes, nous explorons d'abord certaines des motivations des statistiques en grande dimension. Ensuite, nous effectuons une revue de certains problèmes de classement, y compris les modèles monotones et bi-monotones-1D étudiés dans Chapter 3 [74] et Chapter 4. Enfin, nous présentons le problème de détection de points de rupture dans le cas univarié, avant de passer au cas multivarié également détaillé dans Chapter 5 [73].

1.1 Statistiques en grande dimension

Au cours des dernières décennies, nous avons assisté à une évolution significative des technologies d'acquisition de données. Comme mentionné dans [42], ces avancées ont donné naissance à des dispositifs capables de capturer

simultanément des milliers de mesures, entraînant la production de données en grande dimension. Ces ensembles de données à grande échelle apparaissent dans de nombreux domaines, allant des sciences naturelles à la finance et aux sciences sociales.

La théorie classique en statistiques traite généralement des ensembles de données où la taille de l'échantillon n est grande par rapport au nombre de paramètres inconnus p du modèle. Dans le régime asymptotique où n tend vers l'infini et où p est fixe, les principales garanties souhaitables pour un estimateur donné sont la consistence et la normalité asymptotique. En termes simples, la consistence signifie que l'estimateur converge en probabilité vers le paramètre inconnu, et la normalité asymptotique caractérise sa vitesse de convergence. Pour établir ces propriétés, les outils standards sont la loi des grands nombres et le théorème central limite.

Cependant, dans les données en grande dimension, p est parfois si grand que le point de vue asymptotique classique échoue à fournir des prédictions utiles. Dans de tels régimes, il devient souvent difficile de distinguer les informations utiles du bruit dans les données. Pour cette raison, de nombreux travaux ont été consacré au développement de nouveaux outils et techniques statistiques. Plus précisément, cela implique d'introduire une certaine structure sur les paramètres inconnus, et de développer des méthodes qui s'adaptent à cette structure. Par exemple, supposons que nous observons un vecteur $y \in \mathbb{R}^p$ qui suit une distribution gaussienne $\mathcal{N}(\theta, \mathbf{I}_p)$, et que nous voulons retrouver le vecteur inconnu $\theta \in \mathbb{R}^p$. Pour un estimateur $\hat{\theta}$ donné, on considère le risque $\mathbb{E}[\|\hat{\theta} - \theta\|_2^2]$. Ceci est explicitement un problème en grande dimension, étant donné qu'il n'y a seul échantillon dans ce cas ($n = 1$). Discutons maintenant de deux structures potentielles que nous pouvons supposer sur θ, et sous lesquelles l'estimation de θ devient plus simple.

Une contrainte structurelle simple est la parcimonie. Si nous supposons que seul un petit nombre d'entrées de θ sont non nulles par rapport à sa dimension p, alors nous pouvons obtenir un estimateur plus précis que de simplement fixer $\hat{\theta} = y$. Par exemple, l'estimateur avec seuil donné par $\hat{\theta}_i = y_i \mathbf{1}|y_i| \geq t$ obtient de bonnes garanties théoriques si le seuil t est fixé à $\sqrt{2\log(p)}$ – voir par exemple [94]. L'idée principale est que le seuillage est particulièrement adapté pour les vecteurs parcimonieux θ, car il réduit l'impact des entrées nulles de θ dans l'erreur des moindres carrés $\sum_i (\hat{\theta}_i - \theta_i)^2$.

Une autre structure possible est de supposer que θ est monotone, c'est-à-dire que $\theta_1 \leq \cdots \leq \theta_p$. Dans ce cas, il peut être démontré que l'estimateur des moindres carrés défini par $\hat{\theta} = \arg\min_{\theta'} \sum_i (\hat{\theta}'_i - y_i)^2$, où l'argmin est pris sur tous les vecteurs monotones, obtient des garanties optimales par rapport à l'estimateur naïf $\hat{\theta} = y$. Retrouver θ sous cette contrainte est appelé le problème de régression monotone [107, 17].

Dans les problèmes de classement étudiés dans Chapter 3 et Chapter 4, nous cherchons à trier les lignes d'une matrice $n \times d$, en nous basant sur moins de $n \times d$ observations bruitées. Dans le problème de détection de points de changement abordé dans Chapter 5, nous cherchons à détecter des points de changement dans une séquence de n vecteurs de dimension p, où p est potentiellement beaucoup plus grand que n. Ces problèmes relèvent de la catégorie des problèmes statistiques en grande dimension, et nous supposerons des contraintes de forme distinctes pour chacun.

1.2 Problèmes de classement

Les problèmes de classement s'inscrivent dans le sujet plus large de l'estimation des permutations. Ces dernières années, ce domaine a suscité une attention significative, notamment dans des problèmes impliquant l'appariement de vecteurs ou l'appariement de graphes – voir par exemple [25, 100]. Bien qu'il y ait quelques idées communes entre les problèmes d'appariement et notre travail, nous n'entrerons pas dans plus de détails. Le but général des problèmes de classement est de trier un ensemble d'éléments sur la base d'observations bruitées. Ces problèmes englobent un large éventail d'applications, telles que le classement d'experts dans des données de crowdsourcing ou le classement de joueurs dans les tournois. Dans ce qui suit, nous donnons d'abord un aperçu des sujets qui peuvent être trouvés dans l'abondante littérature sur le classement. Ensuite, nous passons au cœur de nos contributions, qui sont principalement centrées sur les modèles monotone et bi-monotone-1D.

1.2.1 Revue sélective des modèles de classement

Premièrement, nous commençons par introduire deux modèles paramétriques simples pour la comparaison par paires, à savoir le modèle Bradley-Terry-Luce (BTL) et le modèle noisy sorting (tri bruité). Ensuite, nous approfondissons le modèle non paramétrique SST (strong stochastic transitive), notamment connu pour sa flexibilité dans le traitement des problèmes de tournois. Enfin, nous explorons d'autres modèles non paramétriques, particulièrement motivés par des données de crowdsourcing. Ces modèles non paramétriques incluent en particulier les modèles monotones et bi-monotones-1D.

1.2.1.1 Modèles paramétriques en comparaisons par paires

Considérons un tournoi où nous observons les comparaisons par paires entre n joueurs. Plus formellement, nous observons une matrice $n \times n$ notée Y, dont les coefficients Y_{ij} appartiennent à $0, 1$ et satisfont $Y_{ij} = 1 - Y_{ji}$. Si le joueur i gagne contre le joueur j, alors Y_{ij} est égal à 1. Il est égal à 0 sinon. Dans un cadre sans bruit, supposons qu'il existe une permutation π^* qui classe parfaitement les joueurs selon leurs aptitudes. En d'autres termes, les entrées de Y représentent les comparaisons données par π^*, c'est-à-dire que $Y_{ij} = \mathbf{1}\{\pi^*(i) - \pi^*(j) > 0\}$. Alors, retrouver la permutation π^* revient simplement à appliquer une méthode de tri standard. Cependant, dans de nombreux cas pratiques, nous ne pouvons pas supposer l'existence d'un tel classement déterministe π^*, car des facteurs aléatoires entrent souvent en jeu.

Modèle noisy sorting. Dans le modèle noisy sorting, Y est également une matrice $n \times n$ qui représente les comparaisons par paires entre les joueurs. Supposons que les coefficients de Y soient des variables aléatoires de Bernoulli indépendantes avec des paramètres $M_{ij} = \mathbb{E}[Y_{ij}]$. De plus, supposons que lorsque $\pi^*(i) < \pi^*(j)$, alors $M_{ij} \geq 1/2 + \gamma$ pour un certain $\gamma \in (0, 1/2)$. Alternativement, soit $\mathcal{M}_{\text{NS}}$ l'ensemble de toutes les matrices qui satisfont $M_{ij} \geq 1/2 + \gamma$ lorsque $i < j$. Alors, le modèle de tri bruité revient à supposer que

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\text{NS}} \ , \tag{1.1}$$

où $(M_{\pi^{*-1}\pi^{*-1}})_{ij} = M_{\pi^{*-1}(i)\pi^{*-1}(j)}$. Dit autrement, π^* représente le classement inconnu entre les joueurs, et M_{ij} dénote la probabilité que i gagne contre j. Si i est meilleur que j, alors ses chances de gagner sont supérieures à $1/2+\gamma$. Contrairement au cadre sans bruit, nous n'observons pas directement quel joueur est le meilleur entre i et j. Au lieu de cela, nous obtenons généralement un résultat Y_{ij} qui est en faveur du meilleur joueur en probabilité. Le but principal est de retrouver π^* aussi précisément que possible.

Dans le modèle noisy sorting, les principaux critères pour mesurer la qualité d'un estimateur $\hat{\pi}$ sont basés sur des distances entre les permutations [12, 61, 13]. Plus précisément, considérons la distance de Kendall tau, définie comme

$$d_{KT}(\pi, \sigma) = \sum_{\pi(i)<\pi(j)} \mathbf{1}\{\sigma(i) > \sigma(j)\} \ , \tag{1.2}$$

pour toutes permutations π et σ. Dans l'équation ci-dessus, la somme est prise sur toutes les paires possibles (i, j). $d_{KT}(\pi, \sigma)$ correspond au nombre d'inversions entre π et σ. Ici, le risque minimax associé est donné par $\inf_{\hat{\pi}} \sup_{\pi^*, M} \mathbb{E}[d_{KT}(\pi^*, \hat{\pi})]$, où le supremum est pris sur toutes les permutations possibles π^*, et toutes les matrices M telles que $M \in \mathcal{M}_{\text{NS}}$. Dans ce contexte, Mao et al. [61] ont établi que le risque minimax est de l'ordre de $n/\gamma^2 \wedge n^2$. En particulier, supposons que γ est de l'ordre d'une constante, par exemple $\gamma = 0.01$. Alors, tout estimateur fait au moins un nombre d'inversions de l'ordre de n parmi les $\binom{n}{2}$ paires possibles.

Braverman et Mossel [12] ont initialement considéré la distance empirique de Kendall tau comme critère pour récupérer π^* :

$$\hat{d}_{KT}(Y, \pi) = \sum_{\pi(i)<\pi(j)} Y_{ij} \ . \tag{1.3}$$

Pour une permutation π donnée, la quantité ci-dessus est calculée en comptant le nombre de paires i, j pour lesquelles la comparaison basée sur π est incompatible avec la comparaison Y_{ji}. Les mêmes auteurs ont établi qu'il existe un estimateur $\hat{\pi}$ qui peut être calculé en temps polynomial, et qui minimise la quantité ci-dessus avec une grande probabilité, c'est à dire $\hat{d}_{KT}(Y, \hat{\pi}) = \min_{\pi} \hat{d}_{KT}(Y, \pi)$.

Notamment, l'existence d'une méthode en temps polynomial pour calculer le minimiseur de (1.3) repose fortement sur l'hypothèse probabiliste faite sur les coefficients de Y. En effet, si les coefficients Y_{ij} peuvent prendre n'importe quelle valeur arbitraire de l'ensemble $\{0, 1\}$, le problème de minimiser $\hat{d}_{KT}(Y, \pi)$ sur tous les π possibles devient équivalent à résoudre le problème de l'ensemble des arcs de rétroaction (feedback arc set). Ce problème d'optimisation s'avère être computationnellement difficile, car il peut être réduit à partir d'un problème NP difficile – voir par exemple [1]. De manière intéressante, c'est un exemple d'un problème d'optimisation computationnellement difficile qui devient faisable avec une instance aléatoire pertinente.

Le modèle BTL. Le modèle Bradley-Terry-Luce (BTL) [11] est un cadre statistique légèrement plus complexe pour classer n joueurs sur la base de comparaisons par paires. Chaque joueur i correspond à un paramètre inconnu $\theta_i \in \mathbb{R}$ où, par convention, $\sum_{i=1}^n \theta_i = 0$. θ_i représente l'aptitude du joueur i. La comparaison $Y_{ij} = 1 - Y_{ji} \in \{0, 1\}$ entre i et j est supposée être une variable aléatoire de Bernoulli de paramètre $M_{ij} = \psi(\theta_i - \theta_j)$, où ψ est la fonction logistique $\psi(t) = 1/(1 + e^{-t})$. De manière équivalente, si π^* est une permutation telle que $\theta_{\pi^{*-1}(1)} \leq \cdots \leq \theta_{\pi^{*-1}(n)}$, alors

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\text{BTL}} \ , \tag{1.4}$$

où $\mathcal{M}_{\text{BTL}}$ dénote l'ensemble de toutes les matrices M telles que $M_{ij} = \psi(\theta'_i - \theta'_j)$ pour un certain vecteur croissant $\theta'_1 \leq \cdots \leq \theta'_n$ satisfaisant $\sum_i \theta'_i = 0$. Comme pour le modèle noisy sorting, l'objectif est de retrouver

la permutation inconnue π^*. La distance de Kendall tau a également été considérée comme mesure de qualité dans le modèle BTL, et nous renvoyons le lecteur aux travaux de Chen et al. [20] pour des garanties optimales avec cette distance.

Cependant, contrairement au modèle noisy sorting, l'accent principal dans le modèle BTL a été mis sur l'estimation de θ [66, 22, 21, 32, 39]. Pour comparer ce modèle avec d'autres modèles que nous présentons plus tard, nous considérons plutôt dans ce qui suit le problème de l'estimation de la matrice entière M. Pour ce problème, nous définissons le risque minimax, ou risque minimax, comme suit:

$$\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n) := \inf_{\hat{M}} \sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right] \ , \tag{1.5}$$

où $\|A\|_F = \sum_{i,j} A_{ij}^2$. Dans (1.5), le supremum est pris sur toutes les permutations possibles π^* et toutes les matrices possibles M telles que $M_{\pi^{*-1},\pi^{*-1}} \in \mathcal{M}_{\mathrm{BTL}}$ – voir (1.4). Pour une matrice donnée M, la distance de Frobenius au carré $\|\hat{M} - M\|_F^2$ mesure la perte de l'estimateur $\hat{M}$. La perte moyenne $\mathbb{E}\left[\|\hat{M} - M\|_F^2\right]$ mesure son risque, et le supremum $\sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right]$ son risque maximal, ou pire risque. Nous disons qu'un estimateur $\hat{M}$ qui a un risque maximal de l'ordre de $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$, à constante multiplicative près indépendante de n, est minimax optimal pour le problème de reconstruction. Nous pouvons déduire de [66, 22, 21] que $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$ est de l'ordre de n [1]. En particulier, cela est nettement inférieur à n^2, qui est le nombre d'entrées de M et une limite supérieure triviale sur le risque minimax.

Le taux de l'ordre de n peut être atteint en temps polynomial dans le modèle BTL. En effet, l'estimateur du maximum de vraisemblance (MLE) $\hat{\theta}$ est une méthode efficace qui conduit à un estimateur minimax optimal de M – voir par exemple [21]. Le MLE $\hat{\theta}$ est donné comme le problème de minimisation suivant:

$$\hat{\theta} := \operatorname*{arg\,min}_{\theta' \,:\, \mathbf{1}^T\theta' = 0} \sum_{i<j} Y_{ij} \log\left(\frac{1}{\psi(\theta'_i - \theta'_j)}\right) + (1 - Y_{ij}) \log\left(\frac{1}{1 - \psi(\theta'_i - \theta'_j)}\right) \ , \tag{1.6}$$

qui est un problème de minimisation convexe sur un espace vectoriel. De plus, l'estimateur correspondant $\hat{M}$ de M peut être défini comme $\hat{M}_{ij} = \psi(\hat{\theta}_i - \hat{\theta}_j)$.

1.2.1.2 Le modèle SST dans les contextes de tournoi

Bien que les modèles BTL et noisy sorting soient souvent des modèles étalons, il a été noté qu'ils manquent souvent de réalisme et qu'ils peuvent ne pas bien correspondre aux données. Pour aborder ces problèmes, le modèle fortement stochastiquement transitif (SST) a été introduit comme un modèle beaucoup plus flexible [83, 86, 60, 56]. Plus précisément, il remplace les hypothèses paramétriques strictes des modèles noisy sorting et BTL par des hypothèses non paramétriques avec des contraintes de forme.

De manière similaire aux modèles BTL et noisy sorting, considérons un scénario où nous observons une matrice $n \times n$ Y d'observations de Bernoulli. Les éléments du triangle supérieur de cette matrice sont indépendants, et chaque entrée $Y_{ij} = 1 - Y_{ji}$ a un paramètre de Bernoulli M_{ij}. En particulier, M est anti-symétrique, c'est-à-dire que $M_{ij} = 1 - M_{ji}$. Dans SST, il est également supposé qu'il existe une permutation inconnue π^* telle que lorsque les lignes et les colonnes de la matrice M sont réarrangées selon π^*, la matrice résultante $M_{\pi^{*-1}\pi^{*-1}}$ est bi-monotone – ses lignes et colonnes sont croissantes. De manière équivalente,

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{SST}} \ , \tag{1.7}$$

où $\mathcal{M}_{\mathrm{SST}}$ désigne l'ensemble de toutes les matrices qui sont anti-symétriques et bi-monotones. En particulier, le modèle SST englobe le modèle BTL puisque $\mathcal{M}_{\mathrm{BTL}} \subset \mathcal{M}_{\mathrm{SST}}$. En d'autres termes, la matrice qui a pour coefficients $M'_{ij} = \psi(\theta_i - \theta_j)$ est anti-symétrique ($M'_{ij} = 1 - M'_{ji}$), et elle est bi-monotone à permutation π^* près de ses lignes et de ses colonnes. Nous renvoyons le lecteur à Figure 1.1 pour un exemple d'une matrice 3×3 qui est à la fois anti-symétrique et bi-monotone. Dans le contexte d'un tournoi, l'hypothèse de bi-monotonie peut être expliquée comme suit. Si un joueur i est meilleur en moyenne qu'un autre joueur j, alors i devrait gagner plus en moyenne que j contre tout autre joueur k. Plus formellement, si $M_{ij} \geq 1/2$, alors $M_{ik} \geq M_{jk}$. Dans la littérature, il y a également eu un accent significatif sur le scénario où seule une proportion $\lambda \in [0, 1]$ des paires (i, j) est observée – voir par exemple [19]. Néanmoins, dans cette discussion sur le modèle SST, nous supposons pour simplifier que toutes les observations nd sont disponibles, c'est à dire $\lambda = 1$.

De manière similaire à ce que nous avons présenté pour le modèle BTL, une question naturelle souvent posée dans la littérature sur le modèle SST [83, 19, 60, 59, 56, 71] peut être formulée comme suit : *Quelle est la*

[1] [66, 22, 21] ne fournissent que le risque minimax pour l'estimation de θ, mais nous pourrions déduire le risque minimax optimal pour la reconstruction de M à partir de leurs résultats.

$$M = \begin{pmatrix} 0.5 & 0.6 & 0.8 \\ 0.4 & 0.5 & 0.7 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

Figure 1.1: Un exemple de matrice bi-monotone et antisymétrique ($M_{ij} = 1 - M_{ij}$)

précision avec laquelle nous pouvons reconstruire la matrice réordonnée M à partir des données observées Y ? Pour quantifier cela, considérons à nouveau le risque de reconstruction minimax suivant :

$$\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n) := \inf_{\hat{M}} \sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right] \ , \tag{1.8}$$

où ici, le supremum est pris sur toutes les permutations π^* et sur toutes les matrices M telles que $M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{SST}}$.

Shah et al. [83] ont établi un fait surprenant : $\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n)$ est de l'ordre de n. Par conséquent, $\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n)$ et $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$ sont du même ordre de grandeur. Ainsi, d'un point de vue statistique, il n'est pas beaucoup plus facile de reconstruire M dans le modèle BTL que dans le modèle SST. En revanche, rappelons que le modèle SST a n^2 paramètres inconnus, tandis que le modèle BTL n'en contient que n.

Étant donnée une matrice d'observation Y, un estimateur naturel de π^* et de la matrice triée $M_{\pi^{*-1}\pi^{*-1}}$ est l'estimateur des moindres carrés, défini par

$$(\hat{\pi}^{LS}, \hat{M}^{LS}_{\mathrm{sorted}}) = \operatorname*{arg\,min}_{\pi \in \Pi_n, \tilde{M} \in \mathcal{M}_{\mathrm{biso}}} \|Y_{\pi^{-1}\pi^{-1}} - \tilde{M}\|_F^2 \ , \tag{1.9}$$

où Π_n désigne l'ensemble de toutes les permutations et $\mathcal{M}_{\mathrm{biso}}$ l'ensemble de toutes les matrices bi-monotones. Ensuite, l'estimateur correspondant de M est $\hat{M}^{LS} = (\hat{M}^{LS}_{\mathrm{sorted}})_{\hat{\pi}^{LS}\hat{\pi}^{LS}}$. Shah et al. [83] ont établi que l'estimateur des moindres carrés $\hat{M}^{LS}$ est minimax optimal : son risque maximal de reconstruction est de l'ordre de n. Malheureusement, aucune méthode en temps polynomial connue n'existe pour résoudre le problème de minimisation ci-dessus (1.9), principalement parce qu'il implique une recherche exhaustive sur toutes les $n!$ permutations.

Les mêmes auteurs proposent également un estimateur en temps polynomial, qui consiste à classer les joueurs selon les moyennes des lignes de Y. Essentiellement, la méthode consiste d'abord à estimer π^* par la permutation $\hat{\pi}$ qui trie les moyennes des lignes de Y. Ensuite, une estimation de M est obtenue en minimisant les moindres carrés sur toutes les matrices bi-monotones :

$$\hat{M}_{\mathrm{sorted}} = \operatorname*{arg\,min}_{\tilde{M} \in \mathcal{M}_{\mathrm{biso}}} \|Y_{\hat{\pi}^{-1}\hat{\pi}^{-1}} - \tilde{M}\|_F^2 \ . \tag{1.10}$$

Contrairement à l'estimateur des moindres carrés (1.9), cet estimateur en temps polynomial atteint seulement un taux de l'ordre de $n^{3/2}$.

Beaucoup d'efforts ont depuis été consacrés à réduire l'écart computationnel entre $n^{3/2}$ et n. En bref, Chatterjee et Mukherjee [19] ont fourni une méthode qui s'adapte à la régularité de la matrice $M_{\pi^{*-1}\pi^{*-1}}$ et au cas où il s'agit d'une matrice bloc. Néanmoins, ils n'ont pas amélioré le taux de $n^{3/2}$ dans le pire des cas. Ensuite, Mao et al. [60, 59] ont introduit une méthode en temps polynomial atteignant le taux de $n^{5/4}$. Pour retrouver π^*, l'approche de [60] comporte deux étapes principales. Tout d'abord, les joueurs sont triés selon les moyennes des lignes de Y, comme dans [83]. Cela donne une matrice pré-triée Y'. Ensuite, les joueurs sont comparés selon les moyennes locales des lignes de Y' sur des intervalles inclus dans $\{1, \dots, n\}$. Cependant, comme l'ont souligné Liu et Moitra [56], la méthode présentée dans [60] n'exploite pas les informations globales disponibles dans toute la matrice car elle ne compare les joueurs que deux à deux. En s'appuyant sur cette remarque, Liu et Moitra [56] ont réussi à atteindre le meilleur taux de $n^{7/6+o(1)}$, en utilisant une méthode en temps polynomial dans le cas où au moins $n^{o(1)}$ échantillons indépendants par entrée sont à disposition. Ces résultats sont également discutés dans Chapter 4, où nous retrouvons également un taux de l'ordre de $n^{7/6}$ dans le modèle SST, à facteur polylogarithmique près. Ainsi, dans le modèle SST, le risque minimax est de l'ordre de n, mais la meilleure méthode connue en temps polynomial atteint seulement un taux de l'ordre de $n^{7/6}$. Cependant, si l'écart entre n et $n^{7/6}$ est intrinsèque au modèle SST ou non reste une question ouverte.

Puisque l'estimateur des moindres carrés atteint le risque minimax de l'ordre de n, il pourrait sembler raisonnable de résoudre le problème computationnel par une relaxation convexe de (2.9). Une première idée est de calculer les moindres carrés $|Y - \tilde{M}|F^2$ sur toutes les matrices $\tilde{M}$ dans l'enveloppe convexe de $M \in \mathbb{R}^{n \times d} : M\pi^{*-1} \in \mathcal{M}$. Néanmoins, un tel estimateur est peu susceptible d'atteindre de bonnes garanties théoriques, et nous renvoyons le lecteur à [82] pour certains résultats négatifs – du moins dans le modèle bi-isotone-2D. Une autre idée est

d'utiliser le théorème de Birkhoff-von Neumann. Soit $\mathcal{P}_n$ l'ensemble des matrices $n \times n$ doublement stochastiques et considérons la relaxation suivante de (1.9) :

$$\hat{M}^{REL}_{\text{sorted}} = \underset{P \in \mathcal{P}_n \tilde{M} \in \mathcal{M}}{\arg\min} \|Y - P\tilde{M}\|_F^2 \ . \tag{1.11}$$

Les ensembles $\mathcal{P}_n$ et $\mathcal{M}$ sont convexes et la fonction de minimisation est convexe en P et en $\tilde{M}$. Cependant, cette fonction n'est pas jointement convexe en $(P, \tilde{M})$, et à notre connaissance, l'existence d'une procédure efficace résolvant (1.11) est un problème ouvert. De plus, il n'est pas clair si le minimiseur du problème relaxé (1.11) atteint le taux de convergence optimal de n, comme le fait l'estimateur des moindres carrés.

1.2.1.3 Autres modèles non-paramétriques en crowdsourcing

Au-delà des problèmes de tournoi et motivés par les problèmes de crowdsourcing, il y a eu une augmentation récente dans le développement de nouveaux modèles non-paramétriques [60, 81, 84, 85, 56, 33]. Avant d'explorer ces modèles, décrivons d'abord le cadre général, qui est similaire à celui du modèle SST. Soit M_{ik} une matrice rectangulaire $n \times d$ dont les coefficients sont dans $[0, 1]$. Dans les données de crowdsourcing, n fait référence au nombre d'experts, d représente le nombre de questions et M_{ik} dénote la probabilité que l'expert i fournisse une réponse correcte à la question k. En particulier, $M_{ik} = 1/2$ signifie que l'expert i fait une estimation aléatoire de la question k et $M_{ik} = 1$ signifie qu'il connaît parfaitement la bonne réponse. Supposons que pour chaque paire (i, k) d'expert/question, nous recevons N_{ik} observations indépendantes de Bernoulli

$$Y^{(u)}_{ik} = \text{Bern}(M_{ik}), \quad u = 1, \dots, N_{ik} \ . \tag{1.12}$$

La paire (i, k) est observée si et seulement si $N_{ik} > 0$ et $Y^{(u)}_{ik} = 1$ signifie que l'expert i est correct par rapport à la question k lors de l'essai u. Pour tenir compte des observations partielles possibles, nous utilisons une astuce de poissonisation standard. À savoir, nous supposons que N_{ik} suit une distribution de Poisson de paramètre $\lambda > 0$ – voir par exemple [60]. Le cas intéressant est lorsque $\lambda \leq 1$, car cela correspond à une proportion de données manquantes de l'ordre de $1 - \lambda$. Dans ce qui suit, nous exposons les modèles bi-monotone-2D, bi-monotones-1D et monotones. Chacun de ces modèles inclut une certaine forme de contrainte de forme sur la matrice M, puisque la reconstruction de M n'est pas possible sans autre hypothèse.

Le modèle bi-monotone-2D. Dans le contexte des données de crowdsourcing, le pendant du modèle SST est le modèle bi-monotone-2D. Supposons qu'il existe deux permutations inconnues π^* et η^* telles que $M_{\pi^{*-1}\eta^{*-1}}$ soit bi-monotone, c'est-à-dire que ses lignes et colonnes sont croissantes. Nous écrivons $\mathcal{M}_{\text{biso}}$ pour l'ensemble de toutes les matrices bi-monotones, de sorte que

$$M_{\pi^{*-1}\eta^{*-1}} \in \mathcal{M}_{\text{biso}} \ . \tag{1.13}$$

Pour illustrer, Figure 1.2 offre une représentation visuelle d'une matrice bi-monotone générée aléatoirement en affichant un graphique de chacune de ses lignes. Ce modèle implique qu'il existe un ordre intrinsèque π^* qui classe les experts selon leurs compétences, et un autre ordre intrinsèque η^* qui trie les questions selon leurs difficultés. Ce modèle englobe le modèle SST précité, où les hypothèses supplémentaires sont que $\pi^* = \eta^*$ et que M est antisymétrique.

Comme garantie théorique, la perte de reconstruction pour un estimateur donné $\hat{M}$ est définie comme $\|\hat{M} - M\|_F^2$, ce qui est analogue à la fonction de perte définie dans le modèle SST. Le risque maximum $\sup \mathbb{E}[\|\hat{M} - M\|_F^2]$ pour cette perte est pris sur toutes les permutations π^*, η^* et matrices M telles que $M_{\pi^{*-1},\eta^{*-1}}$ soit bi-monotone. Mao et al. [60] ont montré que lorsque M est une matrice carrée, c'est-à-dire $n = d$, le risque minimax dans ce modèle est de l'ordre de n à facteurs polylogarithmiques près. Plus précisément, le risque minimax est du même ordre que dans le modèle SST. Ainsi, le modèle bi-monotone-2D n'est pas beaucoup plus difficile, statistiquement parlant, que le modèle SST.

Discutons des problèmes computationnels dans ce modèle. Liu et Moitra [56] ont établi que, lorsque $n = d$ et $\lambda = n^{o(1)}$, il existe une méthode en temps polynomial qui atteint un risque dans le pire des cas de l'ordre de $n^{7/6+o(1)}$. Il s'avère que l'hypothèse que $\lambda = n^{o(1)}$ peut être assouplie à $\lambda = 1$, comme nous le montrons dans un corollaire dans Chapter 4 pour le modèle monotone. Le cas $\lambda = 1$ correspond à la situation où, en moyenne, nous avons une observation pour chaque paire (i, k). D'un autre côté, $\lambda = n^{o(1)}$ représente un nombre sous-polynomial d'observations pour chaque paire (i, k). Dans l'ensemble, de manière similaire au modèle SST, le fait que le même écart computationnel-statistique entre n et $n^{7/6}$ soit intrinsèque au problème demeure une question ouverte dans ce modèle également.

Le modèle bi-monotone-1D. Un modèle plus spécifique motivé par les données de crowdsourcing est le modèle bi-monotone-1D. L'hypothèse plus forte est que la matrice M est bi-monotone à une seule permutation près π^* agissant sur ses lignes. De manière équivalente,

$$M_{\pi^{*-1}} \in \mathcal{M}_{\text{biso}} \ , \tag{1.14}$$

où $(M_{\pi^{*-1}})_{ik} = M_{\pi^{*-1}(i),k}$. Alternativement, le modèle bi-monotone-1D peut être considéré comme un cas particulier du modèle bi-monotone-2D, car il correspond au cas où la permutation sur les colonnes η^* est connue et égale à l'identité. En pratique, connaître η^* dans le modèle bi-monotone-2D revient à supposer que le statisticien a accès à la difficulté intrinsèque des questions, ce qui est une hypothèse plus forte. Étonnamment, dans le cas où $n = d$ et $\lambda = 1$, le taux de reconstruction n'est pas meilleur dans le modèle bi-monotone-1D par rapport aux modèles bi-monotone-2D ou SST. En effet, Mao et al. [60] ont établi que le risque minimax pour la reconstruction est de l'ordre de n, comme dans les modèles SST et bi-monotone-2D.

Néanmoins, Mao et al. [60] ont laissé un écart computationnel-statistique dans ce modèle : leur solution optimale est peu probablement calculable en temps polynomial, et leur méthode efficace ne peut reconstruire M qu'à un taux taux de $n^{5/4}$. Dans [60], l'approche principale consiste à comparer les lignes de Y deux à deux. Ceci est fait en calculant des moyennes locales des lignes sur des intervalles locaux inclus dans $\{1, \dots, d\}$. Plus récemment, Liu et Moitra [56] ont presque fermé l'écart computationnel dans le cas $n = d$ et $\lambda = n^{o(1)}$. Les auteurs ont établi une méthode en temps polynomial qui atteint un taux de reconstruction de $n^{1+o(1)}$, ce qui est presque minimax optimal. Contrairement à [60], une idée cruciale introduite par Liu et Moitra est de se concentrer sur des intervalles spécifiques avant de moyenner les observations. À savoir, pour un ensemble donné P de lignes, ils se concentrent sur des régions où la moyenne de toutes les lignes de P change de manière significative.

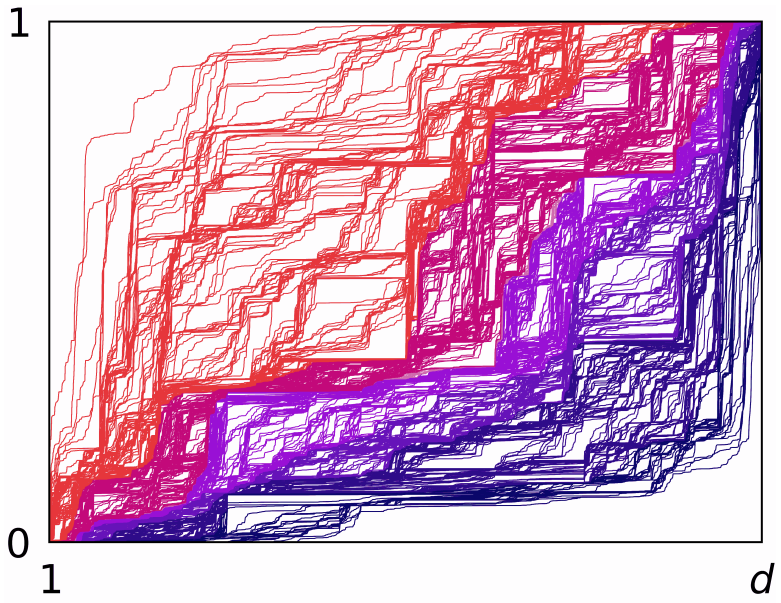


Figure 1.2: Un exemple d'une matrice bi-monotone $M_{\pi^{*-1}}$. Chaque ligne colorée se situe dans $[0, 1]$ et représente une ligne de $M_{\pi^{*-1}}$. Puisque $M_{\pi^{*-1}}$ est bi-monotone, ces lignes sont croissantes.

Le modèle monotone. Enfin, une extension du modèle bi-monotone-2D est le modèle monotone [33]. Dans ce cadre, la seule hypothèse est que toutes les colonnes de M sont croissantes à permutation près π^* des lignes. De manière équivalente, si $\mathcal{M}_{\text{iso}}$ désigne l'ensemble de toutes les matrices monotones, nous avons

$$M_{\pi^{*-1}} \in \mathcal{M}_{\text{iso}} \ . \tag{1.15}$$

En d'autres termes, le modèle monotone est un assouplissement du modèle bi-monotone-2D, sans l'hypothèse que les lignes sont croissantes à permutation η^* près. Notamment, cela rend le modèle monotone plus flexible, même si la reconstruction de M devient statistiquement plus difficile. Pour simplifier les comparaisons avec les autres modèles, supposons que $\lambda = 1$ et que $n = d$. Flammarion et al. [33] ont montré que le risque minimax pour ce modèle est de l'ordre de $n^{4/3}$. Ils ont également introduit RankScore, une méthode computationnellement efficace et qui repose sur une comparaison moyenne globale et une comparaison élément par élément. Cependant, RankScore n'atteint qu'un risque maximal de l'ordre de $n^{3/2}$, ce qui est sous-optimal. Dans le pire des cas,

RankScore atteint des performances similaires à celles obtenues en classant simplement les lignes selon leurs moyennes. Néanmoins, il n'y a pas d'écart computationnel dans ce modèle et le taux $n^{4/3}$ peut être atteint en temps polynomial, à polylogs près. Ceci est prouvé dans l'analyse du modèle monotone, dans Chapter 4. À part quand $n = d$, un autre cas intéressant est lorsque $d = 1$, ce qui revient à supposer que M est un vecteur colonne.

Dans le cas où $d = 1$, ce modèle est étroitement lié à un problème de régression monotone non couplée, qui trouve sa motivation dans les transports optimaux ou les problèmes de sciences sociales [76, 14]. Supposons que nous observions les ensembles non ordonnés $\{x_1, \dots, x_n\}$ et $\{y_1, \dots, y_n\}$, liés par la relation $y_i = f(x_i) + \varepsilon_i$ pour une fonction croissante inconnue f et un bruit ε avec des coefficients indépendants. Comme illustré dans [14], les x_i et y_i peuvent respectivement représenter les données de salaire collectées par une agence gouvernementale et les y_i les prix des logements collectés par une banque. Dans notre cas, $x_i = i$, $f(x_i) = M_{\pi^{*-1}(i)}$, et estimer f revient à estimer le vecteur ordonné $M_{\pi^{*-1}}$. Rigollet et Niles-Weed [76] ont établi que le risque minimax pour l'estimation de f, ou $M_{\pi^{*-1}}$, est de l'ordre de $n(\frac{\log\log(n)}{\log(n)})^2$.

Bien que les trois modèles mentionnés précédemment semblent structurellement similaires, ils diffèrent considérablement d'un point de vue statistique. En particulier, beaucoup plus d'information est disponible lorsque nous supposons que l'ordre des colonnes est connu dans le modèle bi-monotone-1D. Figure 1.3 donne une illustration de la différence entre les modèles monotone et bi-monotone-1D en représentant des matrices M générées aléatoirement.

Dans l'ensemble, les relations entre les trois modèles peuvent être résumées comme suit : le modèle monotone est une extension du modèle bi-monotone-2D, qui est lui-même une extension du modèle bi-monotone-1D. Résumons les écarts computationnelles-statistiques non résolues dans la littérature concernant la reconstruction de M dans ces trois modèles. Dans le modèle monotone, [33] a laissé un écart computationnel entre le taux optimal $n^{4/3}$ et $n^{3/2}$, dans le cas $n = d$. Concernant le modèle bi-monotone-1D, [56] a presque comblé l'écart et a atteint un taux de l'ordre de $n^{1+o(1)}$ dans le cas $n = d$. Cependant, un écart computationnel significatif demeure dans le modèle bi-monotone-1D pour tous n, d, λ tels que $n \ll d$. Enfin, dans le modèle bi-monotone-2D, [56] a réussi à réduire l'écart. Néanmoins, la question de savoir s'il est possible de réduire davantage l'écart entre n et $n^{7/6+o(1)}$ reste ouverte.

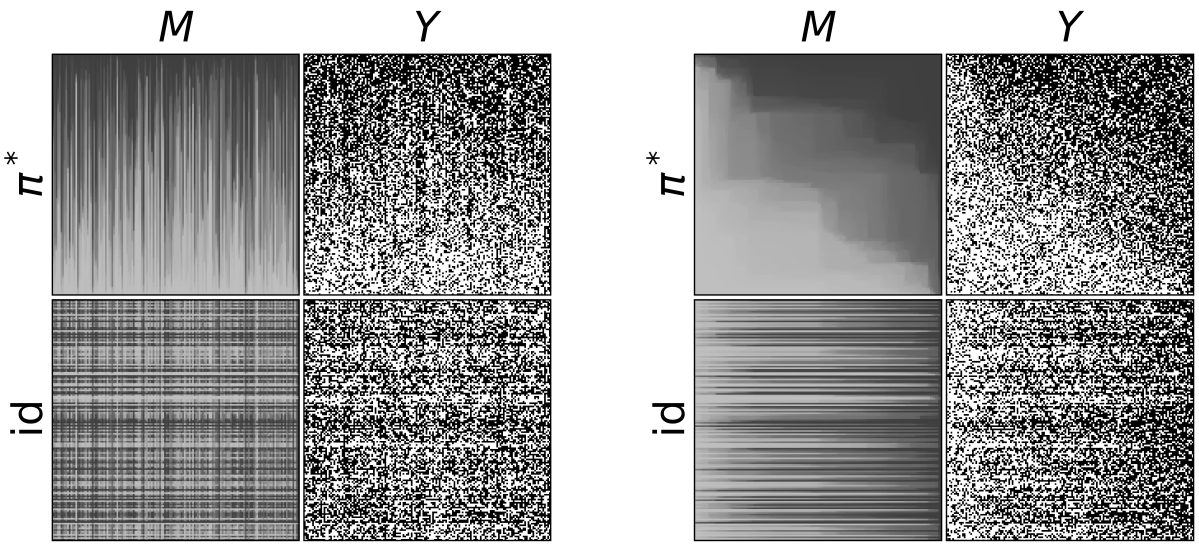


Figure 1.3: Pour chaque modèle – monotone (à gauche) ou bi-monotone-1D (à droite) – les matrices $M_{\pi^{*-1}}$, $Y_{\pi^{*-1}}$, M, Y sont respectivement représentées dans l'ordre de lecture.

1.2.2 Aperçu de notre contribution

Dans Chapter 3 et Chapter 4, nous comblons les écarts computationnels existants dans les modèles bi-monotone-1D (1.14) et monotone (1.15), pour presque toutes les valeurs possibles de n, d et λ. De plus, nous réduisons davantage l'écart dans les modèles bi-monotone-2D (1.13) et SST.

Dans ce qui suit, nous résumons nos principales contributions et nous concentrons notre attention sur les modèles monotone et bi-monotone-1D. Rappelons que $\mathcal{M}_{\text{iso}}$ et $\mathcal{M}_{\text{biso}}$ désignent les ensembles de toutes les matrices monotones et bi-monotones respectivement, et soit $\mathcal{M}$ l'un ou l'autre $\mathcal{M}_{\text{iso}}$ ou $\mathcal{M}_{\text{biso}}$. Comme dans la section précédente, M représente une matrice inconnue dont les entrées sont dans $[0,1]$, et est telle que $M_{\pi^{*-1}} \in \mathcal{M}$ pour une certaine permutation inconnue π^* de ses lignes. Considérons une observation Y donnée par le modèle (1.12).

1.2.2.1 Risque minimax pour l'estimation de la permutation

Comme souligné dans Section 1.2.2, estimer la matrice M est un problème important dans la littérature sur le classement [83, 84, 60, 56, 71]. Cependant, l'objectif principal en classement n'est pas tant de reconstruire l'intégralité de la matrice M, mais plutôt de trouver une bonne estimation de l'ordre original π^*. Par conséquent, nous adoptons une approche différente pour construire un estimateur de π^* et mesurer sa qualité. Étant donné un estimateur $\hat{\pi}$ de π^*, soit $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ la perte de permutation. Contrairement à la perte de reconstruction $\|\hat{M} - M\|_F^2$, cette perte quantifie la distance entre la matrice M triée selon l'estimateur $\hat{\pi}$ et la matrice M triée selon la permutation réelle π^*. En particulier, il n'est pas nécessaire de définir un estimateur $\hat{M}$ de toute la matrice M pour mesurer la qualité d'un estimateur donné $\hat{\pi}$ de π. Nous définissons les risques minimax à la fois pour l'estimation de la permutation et la reconstruction de la matrice comme

$$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}(n,d,\lambda) := \inf_{\hat{\pi}} \sup_{\substack{\pi^* \in \Pi_n \\ M:\, M_{\pi^{*-1}} \in \mathcal{M}}} \mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$$
$$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}(n,d,\lambda) := \inf_{\hat{M}} \sup_{\substack{\pi^* \in \Pi_n \\ M:\, M_{\pi^{*-1}} \in \mathcal{M}}} \mathbb{E}\left[\|\hat{M} - M_{\pi^{*-1}}\|_F^2\right] \quad .$$

La définition de ces deux risques permet en particulier de séparer la difficulté d'estimer π^* de la difficulté d'estimer M.

1.2.2.2 Résultats

Que nous considérions le modèle monotone (1.15) ou le modèle bi-monotone-1D (1.14), il s'avère qu'il existe un estimateur en temps polynomial $\hat{\pi}$ qui est presque minimax optimal, pour presque tous les régimes en n, d et λ. De plus, tout estimateur optimal de π^* peut être utilisé pour construire un estimateur optimal de M. Par conséquent, il n'y a pas d'écart statistique computationnel significatif pour l'estimation de la permutation et la reconstruction de la matrice, contrairement aux modèles SST et bi-monotone-2D.

Considérons, pour simplifier, le régime où $\lambda = 1$, et discutons des risque minimax de permutation résumés dans les tableaux de Figure 1.4. Tout d'abord, la reconstruction de M est statistiquement plus difficile que l'estimation de π^*. En effet, nous déduisons de Figure 1.4 que $\mathcal{R}^{*\mathcal{M}}_{\text{reco}} \gtrsim \mathcal{R}^{*\mathcal{M}}_{\text{perm}}$, à facteurs polylogarithmiques près. Essentiellement, le taux de reconstruction $\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$ peut être décomposé en deux composants : le taux d'estimation de la permutation $\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$ et le taux de reconstruction d'une matrice triée, c'est-à-dire lorsque π^* est connu.

Pour illustrer les taux d'estimation de la permutation, considérons le cas $n = 2$. Dans les deux modèles, la matrice M a deux lignes, l'une étant uniformément au-dessus de l'autre. Sans perte de généralité, supposons que $M_{1,k} \geq M_{2,k}$ pour tout k. Lorsque $n = 2$, ceci est la seule et unique hypothèse dans le modèle monotone. Cependant, dans le modèle bi-monotone-1D, il est en outre supposé que chaque ligne M_1 et M_2 est croissante. Cette hypothèse supplémentaire explique pourquoi le taux de $d^{1/6}$ dans le modèle bi-monotone-1D est beaucoup plus petit que le taux de $\sqrt{d}$ dans le modèle monotone.

Donnons maintenant l'intuition du taux $\sqrt{d}$ dans le modèle monotone. Considérons la méthode simple qui compare les moyennes des deux lignes Y_1 et Y_2. Avec cette méthode, si la moyenne de la ligne 1 est plus grande, alors nous retrouvons la véritable permutation et la perte est égale à 0. Sinon, nous inversons l'ordre des lignes et la perte de permutation est égale à $2\|M_1 - M_2\|_2^2$. Il s'avère que cette méthode simple atteint le taux optimal $\sqrt{d}$. Par la suite, nous fournissons les principaux arguments pour cette affirmation. Étant donné que la ligne 1 est au-dessus de la ligne 2, nous avons que

$$\sum_{k=1}^{d} Y_{1,k} - Y_{2,k} = \|M_1 - M_2\|_1 + \sum_{k=1}^{d} E_{1,k} - E_{2,k} \quad ,$$

où $E_{ik} = Y_{ik} - M_{ik}$. En utilisant l'inégalité de Hoeffding pour les variables aléatoires de Bernoulli, nous déduisons que $|\sum_{k=1}^{d} E_{1,k} - E_{2,k}| \leq C\sqrt{d}$ pour une certaine constante C, avec une probabilité d'au moins $0,99$. De plus, puisque $M_{ik} \in [0,1]$ pour tous i,k, il est vrai que

$$\|M_1 - M_2\|_1 \geq \|M_1 - M_2\|_2^2 \quad .$$

monotone Model $\mathcal{M}_{\text{iso}}$:

	$n \lesssim d^{3/2}$	$d^{3/2} \lesssim n$
$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$	$n^{2/3}\sqrt{d}$	n
$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$	$n^{1/3}d$	n

Bi-monotone Model $\mathcal{M}_{\text{biso}}$:

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$	$nd^{1/3}$	$\sqrt{nd}$	n

Figure 1.4: Taux optimaux dans les modèles monotone et bi-monotone, pour toutes les valeurs possibles de n, d et $\lambda = 1$, à facteur polylogarithmique près en nd. Ces taux sont atteints par des estimateurs en temps polynomial.

Par conséquent, si $\|M_1 - M_2\|_2^2 > C\sqrt{d}$, alors la moyenne de la ligne 1 est au-dessus de la moyenne de la ligne 2 avec une probabilité de $0,99$. Sur cet événement de haute probabilité, nous retrouvons la véritable permutation et la perte en norme de Frobenius carrée est égale à 0. Sinon, lorsque $\|M_1 - M_2\|_2^2 \le C\sqrt{d}$, la perte est limitée à $2C\sqrt{d}$. En fait, cette limite supérieure de l'ordre de $\sqrt{d}$ est optimale au sens minimax, lorsque $n = 2$. De plus, cette limite peut être étendue à des n plus grands, résultant en une limite supérieure sous-optimale de $n\sqrt{d}$. Plus précisément, c'est l'idée sous-jacente derrière le taux sous-optimal de l'ordre de $n^{3/2}$ établi par Shah et al. [83] dans le modèle SST (1.7). Dans le modèle bi-monotone-1D, nous pouvons atteindre le taux $d^{1/6}$ lorsque $n = 2$ en utilisant le fait que les deux lignes M_1 et M_2 sont croissantes. Voir Chapter 3 pour plus de détails.

Lorsque n et d sont égaux, le taux pour estimer la permutation est de l'ordre de $n^{7/6}$ dans le modèle monotone, et il est atteint par un estimateur en temps polynomial $\hat{\pi}$ de π^*. Ce taux a été initialement établi dans le modèle bi-monotone-2D ou SST par Liu et Moitra [56], pour un nombre d'échantillons de l'ordre de $n^{o(1)}$. Cependant, contrairement à [56], notre méthode dans le modèle monotone présentée dans Chapter 4 ne nécessite aucune hypothèse sur les lignes de M. L'estimateur $\hat{\pi}$ peut également être utilisé pour définir un estimateur de la matrice M qui atteint un taux de reconstruction de l'ordre de $n^{4/3}$ dans le modèle monotone – voir Figure 2.4 avec $n = d$. De plus, nous montrons dans Corollary 4.2.5 de Chapter 4 que, dans le modèle bi-monotone-2D ou SST, il est également possible de dériver un estimateur de M à partir de $\hat{\pi}$, qui atteint un taux de reconstruction de l'ordre de $n^{7/6}$. Ce taux de $n^{7/6}$, comme mentionné précédemment, est le meilleur taux connu pour la reconstruction de matrice en temps polynomial ou l'estimation de permutation dans les modèles bi-monotone-2D et SST.

1.2.2.3 Idées générales des procédures

Les procédures que nous décrivons dans Chapter 3 pour le modèle bi-monotone-1D et dans Chapter 4 pour le modèle monotone sont substantiellement différentes. La première repose sur un regroupement hiérarchique avec mémoire, tandis que la seconde est basée sur un graphe de comparaison. Cependant, il vaut la peine de noter que notre analyse du modèle monotone s'appuie sur plusieurs éléments initialement introduits dans notre analyse du modèle bi-monotone-1D. Par la suite, nous donnons un aperçu informel des deux approches.

Dans Chapter 3, nous visons à construire un arbre de tri, comme illustré dans Figure 1.5. En commençant par l'ensemble complet des lignes $[n]$, nous le divisons en trois sous-ensembles (O, P, I) de $[n]$, où O et I contiennent des lignes qui sont probablement en dessous et au-dessus de la ligne médiane, respectivement. Le sous-ensemble P contient des lignes qui ne peuvent pas être classifiées avec une grande confiance. Ensuite, nous divisons récursivement les sous-ensembles O et I, comme le montre la partie gauche de Figure 1.5. Lorsque l'arbre est terminé, nous obtenons un ordre partiel sur toutes les lignes qui peut être utilisé pour estimer π^*.

La principale difficulté de cette procédure est de diviser de manière optimale un ensemble donné $G \subset [n]$ en (O, P, I). Nous y parvenons essentiellement en combinant des techniques allant de la détection de point de ruptures aux méthodes spectrales. Pour calculer les sous-ensembles (O, P, I) d'un ensemble donné G, il est également crucial de garder en mémoire l'arbre de tri. En effet, cette approche nous permet d'utiliser des informations précieuses provenant des autres feuilles de l'arbre pour affiner la division de G. Par exemple, dans Figure 1.5, le groupe $G^{(0)}$ pourrait être davantage divisé en utilisant l'information qu'il est compris entre les deux ensembles de lignes $\mathcal{V}_-$ et $\mathcal{V}_+$.

D'un autre côté, la méthode basée sur un graphe de comparaison dans Chapter 4 consiste à mettre à jour de manière itérative un graphe orienté pondéré. Les arêtes de ce graphe quantifient le niveau de la comparaison entre les lignes de M. Une arête qui pointe d'une ligne i à une autre ligne j signifie que i devrait être au-dessus de j. De plus, nous sommes plus confiants quant à l'ordre entre les lignes pour lesquelles les arêtes ont un poids plus important. À la dernière mise à jour du graphe, nous obtenons un graphe pondéré à partir duquel nous dérivons un estimateur de π^*.

Il est intéressant de noter que la technique basée sur un graphe de comparaison est étroitement liée à la méthode reposant sur le groupement hiérarchique. Le lien principal entre les deux approches peut être résumé comme suit. Dans un graphe de comparaison, l'idée de base est de mettre à jour le poids des arêtes entre une ligne donnée i et les autres lignes dans son voisinage P qui est lui-même calculé à partir du graphe pondéré. A

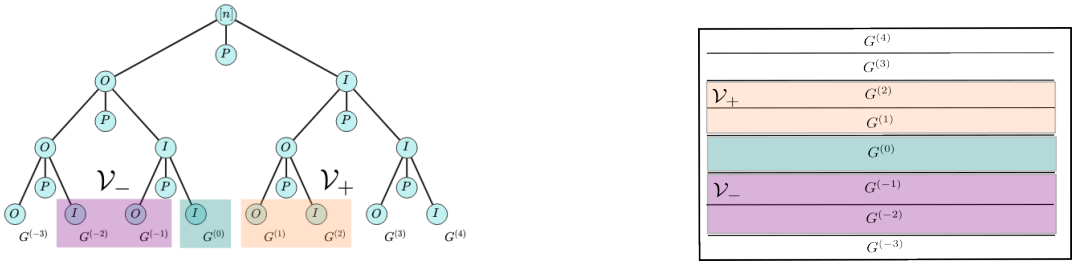


Figure 1.5: Exemple d'un arbre de tri hiérarchique (à gauche), et de la matrice M triée avec l'arbre (à droite). $\mathcal{V}_-$ (resp. $\mathcal{V}_+$) représente un ensemble de groupes de lignes qui sont en dessous (resp. au-dessus) du groupe $G^{(0)}$.

chaque itération est calculé un sous-ensemble de colonnes $\hat{Q} \subset [d]$ qui sert à réduire la dimension et à comparer la ligne i avec les lignes de son voisinage P par des sommes pondérées. De manière similaire, l'approche de groupement hiérarchique implique de comparer les lignes d'un ensemble P en utilisant des sommes pondérées calculées sur des sous-ensembles $\hat{Q}$. Cependant, au lieu de mettre à jour les arêtes entre les lignes, l'approche de groupement hiérarchique calcule une division de P en sous-ensembles. Chaque division consiste à calculer deux sous-ensembles L et U de P, de sorte que les lignes de l'ensemble L soient en dessous des lignes de l'ensemble U. Les fortes relations entre ces deux méthodes suggèrent que les deux pourraient être appliquées aux modèles isotone et bi-isotone-1D pour atteindre les risques minimax à facteurs polylogarithmiques près.

1.3 Détection de points de rupture

La détection de points de rupture a une riche histoire, commençant par les travaux fondateurs de Wald [95], qui ont depuis inspiré des avancées significatives dans le domaine [68, 89]. Comme mentionné précédemment, la détection de points de rupture est cruciale dans un large éventail de situations pratiques, allant de la surveillance des fluctuations quotidiennes de la température et de l'observation des tendances du marché boursier à l'analyse de données génomiques. Dans ce qui suit, nous commençons par discuter du cas univarié, où nous observons une séquence de données réelles. Ensuite, nous introduisons le problème de la détection de points de rupture dans les séries temporelles en grande dimension avant de passer à notre contribution dans ce contexte.

1.3.1 Discussion sur le cas univarié

Dans cette discussion, nous nous concentrons exclusivement sur les séries temporelles univariées. Supposons que nous observons une séquence de variables aléatoires réelles indépendantes $(y_1, \dots, y_n)$, avec des fonctions de distribution cumulatives $(F_1, \dots, F_n)$. Nous disons qu'il y a un point de rupture à une position τ si la fonction de distribution cumulative à τ est différente de la précédente, c'est-à-dire $F_{\tau-1} \neq F_\tau$. En particulier, si nous savons que le nombre de points de rupture K est au plus égal à 1, la question devient de savoir si la distribution des données reste stationnaire au fil du temps, ou s'il y a un point de rupture détectable. Bien que ce modèle non paramétrique englobe de nombreuses situations, il est souvent trop large dans de nombreuses applications pratiques [68]. Pour cette raison, les distributions F_i doivent souvent être paramétrées.

Par la suite, nous supposons que pour chaque $t = 1, \dots, n$, nous avons la décomposition signal/bruit suivante :

$$y_t = \theta_t + \varepsilon_t \in \mathbb{R} \ , \tag{1.16}$$

où la suite déterministe (θ_t) est inconnue, et le bruit $\varepsilon_1, \dots, \varepsilon_n$ est composé de variables indépendantes gaussiennes centrées standard $\mathcal{N}(0,1)$. Dans ce modèle, la suite des points de rupture $(\tau_1, \dots, \tau_K)$ correspond aux positions τ_k où $\theta_{\tau_k - 1}$ diffère de θ_{τ_k}. Pour chaque point de rupture τ_k, nous définissons $D_k \in \mathbb{R}$ comme la différence $\theta_{\tau_k} - \theta_{\tau_{k-1}}$, représentant le rupture moyen dans les données. Nous définissons également r_k comme la distance entre τ_k et son point de rupture adjacent le plus proche, c'est-à-dire $r_k = \min(\tau_k - \tau_{k-1}, \tau_{k+1} - \tau_k)$. Par convention, nous fixons $\tau_0 = 1$ et $\tau_{K+1} = n + 1$. Discutons maintenant des problèmes de détection de ruptures simples et multiples dans ce contexte.

Détection d'un unique point de rupture. Supposons que nous savons qu'il y a au plus un point de rupture, c'est-à-dire $K \leq 1$. Le problème revient alors à tester les deux hypothèses suivantes :

$$H_0 \ : \text{ Il n'y a pas de point de rupture}$$
$$H_1 \ : \text{ Il y a un unique point de rupture à une position inconnue } \tau \ .$$

Nous cherchons à déterminer s'il y a ou non un point de rupture dans la suite (θ_t). Si un point de rupture τ existe, alors θ_t est égal à μ_1 si $t < \tau$ et à $\mu_2 \neq \mu_1$ sinon. Dans ce modèle à un seul point de rupture, l'approche originale de Hinkley [44] consiste à maximiser la valeur absolue de la statistique CUSUM

$$\mathbf{C}_t(y) = \sqrt{\frac{(t-1)(n-t+1)}{n}} \left(\frac{1}{n-t+1} \sum_{i=t}^{n} y_i - \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \right) \ , \tag{1.17}$$

sur toutes les positions possibles $t = 2, \dots, n$. En termes plus simples, $\mathbf{C}_t(y)$ représente la différence redimensionnée entre la moyenne des données sur l'intervalle $[t, n]$ et la moyenne sur $[1, t)$. L'idée principale est que s'il n'y a pas de point de rupture, alors $\mathbf{C}_t(y)$ suit une distribution normale standard $\mathcal{N}(0,1)$ pour tout t. D'un autre côté, s'il y a un point de rupture à la position τ, alors $\mathbf{C}_\tau(y)$ suit une distribution normale avec une espérance égale à $\sqrt{\frac{(\tau-1)(n-\tau+1)}{n}} D$, où $D = \mu_2 - \mu_1$. En particulier, cette quantité satisfait :

$$\tfrac{1}{2} r D^2 \leq \tfrac{(\tau-1)(n-\tau+1)}{n} D^2 \leq r D^2 \ , \tag{1.18}$$

où $r = \min(\tau, n+1-\tau)$.

Détecter τ devient statistiquement plus facile à mesure que rD^2 augmente. Il s'avère que la quantité rD^2 caractérise précisément la limite de détection dans ce problème. Récemment, Gao et al. [38] ont établi que le point de rupture peut être détecté avec une grande probabilité, dès que $rD^2 \geq C\sqrt{\log\log(n)}$, pour une constante C qui dépend uniquement de la probabilité d'erreur désirée. Voir également les travaux antérieurs de Csörgö et Horváth [26] pour un résultat asymptotique connexe. Pour un résultat plus précis remplaçant $\sqrt{\log\log(n)}$ par $\sqrt{\log\log(n/r)}$, nous renvoyons le lecteur aux travaux de Verzelen et al. [92].

Détection de points de rupture multiples. Le problème est plus complexe lorsque K est inconnu et arbitraire. Pour estimer l'un des K points de rupture, il semblerait raisonnable de prendre le maximum sur tous les t de la valeur absolue de $\mathbf{C}_t(y)$, comme dans le cas précédent avec un seul point de rupture. C'est le principe de la segmentation binaire (BS) [80], mais malheureusement, cela ne conduit pas à un estimateur consistent de l'un des points de rupture. Nous présentons brièvement ensuite deux classes de méthodes visant à surmonter les limites de BS.

Une large gamme de méthodes est basée sur une variante de la segmentation binaire : la segmentation binaire sauvage (WBS), qui est une approche de haut en bas introduite par Fryzlewicz [36]. Dans WBS, nous tirons d'abord au hasard certains intervalles aléatoires $[s, e) \subset [n]$. Ensuite, nous maximisons la statistique CUSUM locale

$$\mathbf{C}_{s,t,e}(y) = \sqrt{\frac{(t-s)(e-t)}{e-s}} \left(\frac{1}{e-t} \sum_{i=t}^{e-1} y_i - \frac{1}{t-s} \sum_{i=s}^{t-1} y_i \right) \ ,$$

sur tous les t possibles dans $[s, e)$ et tous les intervalles aléatoirement choisis $[s, e)$. Si le maximum est au-dessus d'un certain seuil en valeur absolue, alors nous prenons le t correspondant comme premier estimateur d'un point de rupture. Ensuite, nous subdivisons $[n]$ en $[1, t-1]$ et $[t, n]$ et nous cherchons de manière récursive d'autres points de rupture potentiels dans ces deux intervalles.

Une autre classe de méthodes est basée sur des critères des moindres carrés pénalisés [92, 98]. L'idée principale est d'estimer une séquence constante par morceaux $\hat{\theta}$ via le problème de minimisation suivant :

$$\hat{\theta} = \operatorname*{arg\,min}_{\theta' \in \mathbb{R}^n} \sum_i (y_i - \theta'_i)^2 + \lambda \mathrm{pen}(\theta') \ , \tag{1.19}$$

où λ est un paramètre de réglage et $\mathrm{pen}(\theta') \geq 0$ est une fonction de θ' visant à pénaliser les variations de θ'. En particulier, Wang et al. [98] définissent $\mathrm{pen}(\theta')$ comme le nombre de positions i où $\theta'_{i-1} \neq \theta'_i$. Bien que le problème de minimisation (1.9) ne soit pas convexe, $\hat{\theta}$ peut encore être calculé efficacement en utilisant des techniques de programmation dynamique – voir par exemple l'algorithme 1 de Friedrich et al. [35].

De manière similaire au problème de détection d'un seul point de rupture, la condition minimale de détection d'un point de rupture donné τ_k dans ce contexte dépend de la quantité $r_k D_k^2$. Wang et al. [98] ont établi que, dès que $r_k D_k^2 \gtrsim \sqrt{\log(n)}$ pour tous les k, nous pouvons détecter tous les points de rupture avec une grande probabilité. Cette découverte a été davantage affinée par Verzelen et al. [98], qui ont montré que la condition minimale nécessaire à la détection du point de rupture τ_k est $r_k D_k^2 \gtrsim \sqrt{\log(n/r_k)}$.

1.3.2 Le cas multivarié

Dans le cas multivarié, les observations $y_1, \dots, y_n$ appartiennent à l'espace vectoriel $\mathbb{R}^p$, de dimension $p \geq 1$ arbitraire. Nous observons, pour $t = 1, \dots, n$

$$y_t = \theta_t + \varepsilon_t \in \mathbb{R}^p \ . \tag{1.20}$$

Contrairement au cas univarié (1.16), θ_t est un vecteur dans $\mathbb{R}^p$ et les variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables gaussiennes multivariées standard i.i.d $\mathcal{N}(0, \mathbf{I}_p)$. Dans ce contexte, le k^{ime} point de changement τ_k est toujours défini par $\theta_{\tau_k} \neq \theta_{\tau_k - 1}$, et $D_k = \theta_{\tau_k} - \theta_{\tau_k - 1}$ est un vecteur dans $\mathbb{R}^p$. Nous notons s_k pour la parcimonie de D_k, c'est-à-dire $s_k = \|D_k\|_0$. En termes plus simples, s_k représente le nombre d'entrées non nulles de D_k.

Parmi la littérature sur les séries temporelles multivariées, beaucoup d'efforts ont été concentrés sur l'adaptation à la parcimonie de D_k [103, 47, 57]. Comme dans le cas univarié, nous cherchons la valeur minimale de $r_k \|D_k\|_2^2$ pour laquelle le point de rupture τ_k peut être détecté avec précision. Dans ce contexte, cette valeur dépend de n, r_k, p et s_k.

Liu et al. [57] ont considéré le cas d'un seul point de rupture ($K \leq 1$), où le but est de détecter un point de rupture potentiel τ. Comme dans le cas univarié, soit $r = \min(\tau, n - \tau)$, $D = \theta_\tau - \theta_{\tau - 1}$ et $s = \|D\|_0$. Les auteurs ont établi que le point de rupture τ peut être détecté avec une grande probabilité dès lors que $r\|D\|_2^2$ est supérieur à $C \min(\sqrt{p \log\log(8n)}, s \log(\frac{p}{s^2} \log\log(8n)))$, où C est une constante qui dépend uniquement de la probabilité de détection souhaitée. En particulier, lorsque $p = 1$, nous retrouvons le résultat de Gao et al. [38] mentionné plus tôt dans le cas univarié.

Dans le cas de points de rupture multiples ($K \geq 1$), Wang et Samworth [103] ont introduit une méthode basée sur des projections parcimonieuses. Cependant, leur procédure ne détecte les points de rupture que sous une condition forte sur $r_k \|D_k\|_2^2$. Plus récemment, Hu et al. [47] ont assoupli cette condition dans un cadre asymptotique spécifique, en utilisant une approche basée sur un score de vraisemblance parcimonieux. Néanmoins, la condition de détection prouvée dans [47] n'est pas optimale si l'on s'intéresse aux facteurs logarithmiques.

En plus de la détection de points de rupture en moyenne dans le cas multivarié, le sujet plus large de la détection de points de rupture dans les séries temporelles en grande dimension englobe également de nombreux autres problèmes. Ceux-ci incluent entre autre des problèmes tels que la détection de points de rupture en covariance [96] ou la détection de points de rupture dans des séquences de réseaux [97]. Dans chaque cas, les données consistent en une suite en grande dimension, où le signal sous-jacent est constant par morceaux avec une structure spécifique.

1.3.3 Aperçu de notre contribution

Dans Chapter 5, nous établissons des conditions minimales de détection des points de rupture dans plusieurs problèmes, y compris la détection de points de rupture en covariance, la détection de points de rupture non paramétrique et la détection de points de rupture en moyenne multivariée sparse. Par la suite, nous nous concentrons sur notre contribution à la détection des points de rupture en moyenne dans des séries temporelles multivariées, et nous fournissons un résumé de notre travail sur ce problème. Nous commençons par décrire la condition minimale de détection avant de discuter des idées de notre approche ascendante qui atteint des garanties optimales minimax.

Comme mentionné précédemment, la condition minimale que les points de rupture doivent satisfaire pour être détectés dépend de $r_k \|D_k\|_2^2$. Intuitivement, si $r_k \|D_k\|_2^2$ est très grand, alors τ_k peut facilement être détecté. D'autre part, la détection de τ_k devient impossible lorsque $r_k \|D_k\|_2^2$ approche 0. Dans Chapter 5, nous montrons que la condition minimale de détection de tous les points de rupture τ_k est donnée par

$$r_k \|D_k\|_2^2 \geq C \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k} \sqrt{\log\left(\frac{n}{r_k}\right)}\right) + \log\left(\frac{n}{r_k}\right) \right] , \tag{1.21}$$

pour tout $k = 1, \ldots, K$. Ici, C est une constante qui ne dépend que de la probabilité souhaitée de détecter tous les points de rupture. Plus précisément, si tous les points de rupture satisfont (1.21), alors il existe un estimateur $(\hat{\tau}_k)$ de (τ_k) qui satisfait avec grande probabilité :

1. Nous retrouvons le vrai nombre de points de rupture, c'est-à-dire $\hat{K} = K$
2. Le point de rupture estimé $\hat{\tau}_k$ n'est pas trop éloigné du vrai point de rupture τ_k, au sens où $\hat{\tau}_k \in [\tau_k - r_k/2, \tau_k + r_k/2]$

Les deux propriétés ci-dessus sont sans doute les garanties minimales souhaitables que l'on pourrait attendre d'un estimateur de points de rupture, lorsqu'ils sont tous supposés être détectables. Notre analyse faite dans Chapter 5 prend en compte les points de rupture qui ne satisfont peut-être pas la condition (1.21). De manière informelle, l'estimateur est tenu de détecter les points de rupture qui satisfont (1.21), et de ne pas détecter deux fois un même point de rupture. De plus, les intervalles de détection fournis dans Chapter 5 sont plus précis que $[\tau_k - r_k/2, \tau_k + r_k/2]$.

De manière intéressante, la condition de détection (1.21) peut être approximée à $r_k \|D_k\|_2^2 \gg s_k \wedge \sqrt{p}$, à facteurs logarithmiques près. Notamment, la parcimonie s_k de D_k rend le problème de détection substantiellement plus facile lorsque $s_k \leq \sqrt{p}$.

Pour détecter les points de rupture lorsqu'ils satisfont la condition (1.21), nous utilisons une approche ascendante dans Chapter 5. De manière informelle, notre méthode est basée sur l'agrégation de plusieurs tests, qui sont effectués à différents emplacements et à différentes échelles. Tout d'abord, nous commençons par essayer de détecter les points de rupture sur les intervalles du type $[l-1, l]$, en nous basant sur l'observation $y_l - y_{l-1} \in \mathbb{R}^p$. Ensuite, pour tout point de rupture potentiel détecté, nous enlevons un petit voisinage autour de lui. Puis, nous essayons une échelle plus grande $r = 2$, et nous effectuons des tests locaux sur les intervalles restants de la forme $[l-r, l+r)$. Nous continuons ce processus pour des échelles r croissantes, jusqu'à ce que nous atteignions une échelle maximale r de l'ordre de $n/2$. À la fin, l'estimateur correspond aux positions au sein des intervalles qui ont été retirés au cours du processus.

Dans Chapter 5, l'analyse de cet estimateur implique une borne d'union sur les événements contrôlant les tests locaux sur les intervalles $[l-r, l+r)$. Pour chacun de ces intervalles, nous contrôlons la statistique CUSUM en grande dimension correspondante avec des techniques similaires à celles utilisées dans la détection d'un signal sparse – voir par exemple [30]. Finalement, cela assure avec une grande probabilité que nous détectons les points de rupture τ_k qui satisfont la condition (1.21), à des échelles r plus petites que $r_k/2$.

Pour conclure, les points de rupture d'une série temporelle en grande dimension peuvent être détectés sous la condition minimale (1.21), qui s'adapte à la parcimonie s_k des points de rupture. De plus, l'approche ascendante décrite dans Chapter 5 est capable de détecter avec une grande probabilité tous les points de rupture qui satisfont cette condition.

Chapter 2

Introduction

In this thesis, we explore two areas in modern statistics: ranking problems and change-point detection. Ranking problems have wide applications, from ranking players in tournaments to voting and organizing experts in crowdsourcing data. Similarly, change-point detection plays a crucial role in various real-world scenarios, such as tracking daily temperature changes, monitoring stock prices, or analyzing genomic data.

While we discuss these two topics independently, our approach to both is grounded in the framework of high-dimensional statistics. Indeed, in both cases, the number of unknown parameters can be larger than the number of samples. To handle this complexity and to make these high-dimensional problems tractable, we introduce specific structural assumptions, or shape constraints, in each model.

The first part of this thesis delves into ranking problems. Essentially, ranking methods aim to sort a collection of items based on noisy observations, with potentially missing data. We will explore various models that differ in their shape constraints. Specifically, our contributions center on two models where the goal is to recover a permutation of the rows of a matrix. In Chapter 3, we assume that the reordered matrix has nondecreasing columns and rows. In Chapter 4, we only assume that it has nondecreasing columns. This model encompasses many models, including crowd-labeling and ranking in tournaments by pairwise comparisons. For both models, we provide computationally tractable algorithms that achieve optimal guarantees in estimating the permutation.

The second part focuses on a multiple change-point detection problem for multivariate time series. Informally, a change-point is a point in time in a sequence where the statistical properties of the data change. In our context, the sequential observations can exist in a high-dimensional space, and may contain an arbitrary number of change-points. This extends in particular the simpler univariate model, where the data consist of real-valued observations. To make the problem more tractable, we consider cases where the variations of the high-dimensional signal can be sparse. In a sparse regime, many entries of the vector representing the variations are equal to zero. In Chapter 5, we establish the minimal conditions for possible detection in this framework. Under these conditions, we provide guarantees that are adaptive to the unknown sparsity and to the distance between the change-points.

In the last chapter, we discuss potential developments for this thesis. Specifically, we explore problems related to ranking and label identification in crowdsourcing, online ranking problems, and localization problems within multivariate time series.

In the following sections, we first explore some motivations of high-dimensional statistics. Next, we make a review of the main ranking problems, including the isotonic and bi-isotonic-1D models studied in Chapter 3 [74] and Chapter 4. Lastly, we present a change-point detection problem in the univariate case, before moving to the multivariate case also detailed in Chapter 5 [73].

2.1 High-dimensional statistics

Over the past few decades, we have witnessed a significant evolution in data acquisition technologies. As mentioned in [42], these advancements have given rise to devices capable of capturing thousands of measurements simultaneously, resulting in the production of high-dimensional data. These large-scale datasets appear in many fields, ranging from natural sciences to finance and social sciences.

Classical theory in statistics typically deals with datasets where sample size n is large with respect to the number of unknown parameters p of the model. In the asymptotic regime where n goes to infinity and where p is fixed, the main desirable guarantees for a given estimator are consistency and asymptotic normality. In simple terms, consistency means that the estimator converges in probability to the unknown parameter, and asymptotic normality characterizes its speed of convergence. To establish these properties, the standard tools are the law of large numbers and the central limit theorem.

However, in high-dimensional data, p is sometimes so large that the classical asymptotic point of view fail to provide useful predictions. In such regimes, it often becomes challenging to distinguish useful information from noise in the data. Hence, considerable work has been devoted to develop new tools and statistical techniques. Specifically, these involve introducing some structure on the unknown parameters, and developing methods that adapt to this structure. For instance, assume that we observe a vector $y \in \mathbb{R}^p$ that follows a Gaussian distribution $\mathcal{N}(\theta, \mathbf{I}_p)$, and that we want to recover the unknown vector $\theta \in \mathbb{R}^p$. For a given estimator $\hat{\theta}$, we consider the L_2 risk $\mathbb{E}[\|\hat{\theta} - \theta\|_2^2]$. We are explicitly dealing with a high-dimensional setting, given that we have just one sample in this case ($n = 1$). Let us now discuss two potential structures that we can assume on θ to make its estimation more feasible.

A simple structural constraint is sparsity. If we assume that only a small number of the entries of θ are non-zero compared to its dimension p, then we can derive a more accurate estimator than simply setting $\hat{\theta} = y$. For example, the hard-thresholding estimator given by $\hat{\theta}_i = y_i \mathbf{1}\{|y_i| \geq t\}$ achieves good theoretical guarantees if the threshold t is set to $\sqrt{2\log(p)}$ – see e.g. [94]. The main idea is that hard-thresholding is particularly adapted for sparse vectors θ, as it reduces the impact of the zero entries of θ in the least square error $\sum_i (\hat{\theta}_i - \theta_i)^2$.

Another possible structure is to assume that θ is isotonic, i.e. that $\theta_1 \leq \cdots \leq \theta_p$. In this case, it can be shown that the least square estimator defined by $\hat{\theta} = \arg\min_{\theta'} \sum_i (\theta'_i - y_i)^2$, where the argmin is taken over all isotonic vectors, achieves optimal guarantees in contrast to the naive estimator $\hat{\theta} = y$. Recovering θ under this constraint is called the isotonic regression problem [107, 17].

In the ranking problems studied in Chapter 3 and Chapter 4, we aim to sort the rows of an $n \times d$ matrix, based on less than $n \times d$ noisy observations. In the change-point detection problem covered in Chapter 5, we aim to detect change-points in a sequence of n vectors of dimension p, where p is potentially much larger than n. Both of these problems fall under the category of high-dimensional statistical problems, and we will assume distinct shape constraints for each.

2.2 Ranking problems

Ranking problems fall within the broader topic of permutation estimation. In recent years, this area has gained significant attention, notably in problems involving vector matching or graph matching – see e.g. [25, 100]. While there are a few common ideas between matching problems and our work, we will not discuss them into further detail. The general purpose in ranking problems is to sort a set of items based on noisy observations. These problems encompass a wide range of applications, such as ranking experts or workers in crowdsourcing data or ranking players in tournaments. In what follows, we first give an overview of the topics that can be found in the extensive body of ranking literature. Next, we shift to the core of our contributions, which are primarily centered on the isotonic and bi-isotonic-1D models.

2.2.1 Selective review of ranking models

First, we start by introducing two simple parametric models for pairwise comparison, namely the Bradley-Terry-Luce (BTL) model [11] and the noisy sorting model [12]. Next, we delve deeper into the non-parametric strong stochastically transitive (SST) model [83], notably known for its flexibility in addressing tournament problems. Finally, we explore other non-parametric models, particularly motivated by crowdsourcing data. These models include in particular the isotonic and the bi-isotonic-1D models.

2.2.1.1 Parametric models in pairwise comparisons

Consider a tournament where we observe the pairwise comparisons between n players. More formally, we observe an $n \times n$ matrix Y whose coefficients Y_{ij} belong to $\{0, 1\}$ and satisfy $Y_{ij} = 1 - Y_{ji}$. If player i wins against player j, then Y_{ij} is equal to 1. It is equal to 0 otherwise. In a noiseless setting, assume that there exists a permutation π^* that perfectly ranks the players according to their abilities. In other words, the entries of Y represent the comparisons given by π^*, that is $Y_{ij} = \mathbf{1}\{\pi^*(i) - \pi^*(j) > 0\}$. Then, recovering the permutation π^* simply amounts to applying a standard sorting method. However, in many real-world scenarios, we cannot assume the existence of such a deterministic ranking π^*, as random factors often come into play.

The noisy sorting model. Within the noisy sorting framework, Y is also an $n \times n$ matrix which represents the pairwise comparisons between players. Assume that the coefficients of Y are independent Bernoulli random variables with parameters $M_{ij} = \mathbb{E}[Y_{ij}]$. Additionally, assume that when $\pi^*(i) < \pi^*(j)$, it holds that $M_{ij} \geq 1/2 + \gamma$ for some $\gamma \in (0, 1/2)$. Alternatively, let $\mathcal{M}_{\text{NS}}$ be the set of all matrices that satisfy $M_{ij} \geq 1/2 + \gamma$ when $i < j$.

Then, the noisy sorting model is equivalent to assume that

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{NS}} \ , \tag{2.1}$$

where we write $(M_{\pi^{*-1}\pi^{*-1}})_{ij} = M_{\pi^{*-1}(i)\pi^{*-1}(j)}$. In simpler terms, π^* represents the unknown ranking between the players, and M_{ij} denotes the probability that i wins against j. If i is better than j, then its chances of winning are greater than $1/2 + \gamma$. In contrast to the noiseless setting, we do not directly observe which player is the best between i and j. Instead, we typically get an outcome Y_{ij} that is in favor of the best player in probability. The main purpose is to accurately recover π^*.

In the noisy sorting model, the main criteria to measure the quality of a given estimator $\hat{\pi}$ are based on distances between permutations [12, 61, 13]. Specifically, consider the Kendall tau distance, defined as

$$d_{KT}(\pi, \sigma) = \sum_{\pi(i)<\pi(j)} \mathbf{1}\{\sigma(i) > \sigma(j)\} \ , \tag{2.2}$$

for any permutations π and σ. In the above equation, the sum is take over all possible pairs of indices (i, j). $d_{KT}(\pi, \sigma)$ corresponds to the number of inversions between π and σ. Then, the associated minimax risk is given by $\inf_{\hat{\pi}} \sup_{\pi^*, M} \mathbb{E}[d_{KT}(\pi^*, \hat{\pi})]$, where the supremum is taken over all possible permutations π^*, and all matrices M such that $M \in \mathcal{M}_{\mathrm{NS}}$. In this context, Mao et al. [61] established that the minimax risk is of the order of $n/\gamma^2 \wedge n^2$. In particular, assume that γ is of the order of a constant, e.g. $\gamma = 0.01$. Then, any estimator makes at least a number of inversions of the order of n among the $\binom{n}{2}$ possible pairs.

Interestingly, Braverman and Mossel [12] originally considered the empirical Kendall tau distance as a criterion to recover π^*:

$$\hat{d}_{KT}(Y, \pi) = \sum_{\pi(i)<\pi(j)} Y_{ij} \ . \tag{2.3}$$

For a given permutation π, the above quantity is calculated by counting the number of pairs i, j for which the comparison based on π is inconsistent with the comparison Y_{ji}. The authors established that there exists an estimator $\hat{\pi}$ that can be computed in polynomial-time, and that minimizes the above quantity with high probability. In other words, $\hat{d}_{KT}(Y, \hat{\pi}) = \min_{\pi} \hat{d}_{KT}(Y, \pi)$.

Notably, the existence of a polynomial-time method to compute the minimizer of (2.3) heavily relies on the probabilistic assumption made on the coefficients of Y. Indeed, if the coefficients Y_{ij} can take any arbitrary value from the set $\{0, 1\}$, the problem of minimizing $\hat{d}_{KT}(Y, \pi)$ over all π becomes equivalent to solving the feedback arc set problem. This optimization problem turns out to be computationally challenging, as it can be reduced from a NP hard problem – see e.g. [1]. Interestingly, this is an example of a computationally hard optimization problem that becomes feasible in a relevant random instance.

The BTL model. The Bradley-Terry-Luce (BTL) model [11] is a slightly more involved statistical framework for ranking n players based on pairwise comparisons. Each player i corresponds to an unknown parameter $\theta_i \in \mathbb{R}$ where, by convention, $\sum_{i=1}^n \theta_i = 0$. θ_i represents the ability of player i. The comparison $Y_{ij} = 1 - Y_{ji} \in \{0, 1\}$ between i and j is assumed to be a Bernoulli random variable with parameter $M_{ij} = \psi(\theta_i - \theta_j)$, where ψ is the logistic function $\psi(t) = 1/(1 + e^{-t})$. Equivalently, if π^* is a permutation such that $\theta_{\pi^{*-1}(1)} \leq \cdots \leq \theta_{\pi^{*-1}(n)}$, then

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{BTL}} \ , \tag{2.4}$$

where $\mathcal{M}_{\mathrm{BTL}}$ denotes the set of all matrices M such that $M_{ij} = \psi(\theta'_i - \theta'_j)$ for some nondecreasing vector $\theta'_1 \leq \cdots \leq \theta'_n$ satisfying $\sum_i \theta'_i = 0$. Similarly to the noisy sorting model, we aim to recover the unknown permutation π^*. The Kendall tau distance has also been considered as a measure of quality in the BTL model, and we refer the reader to the of work Chen et al. [20] for optimal guarantees in Kendall tau distance.

However, unlike in the noisy sorting model, the main focus in the BTL model has been on the estimation of θ [66, 22, 21, 32, 39]. To compare this model with other models that we present later, we rather consider in what follows the problem of estimating the whole matrix M. For this problem, we define the minimax risk as follows:

$$\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n) := \inf_{\hat{M}} \sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right] \ , \tag{2.5}$$

where $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2}$. In (2.5), the supremum is taken over all possible permutations π^* and all possible matrices M such that $M_{\pi^{*-1},\pi^{*-1}} \in \mathcal{M}_{\mathrm{BTL}}$ – see (2.4). For a given matrix M, the squared Frobenius distance $\|\hat{M} - M\|_F^2$ measures the loss of the estimator $\hat{M}$. The expected loss $\mathbb{E}\left[\|\hat{M} - M\|_F^2\right]$ measures its risk, and the supremum $\sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right]$ its maximum risk, or worst risk. We say that an estimator $\hat{M}$ that has a maximum risk of order $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$, up to a multiplicative constant independent of n, is minimax optimal for

the reconstruction problem. Interestingly, we can derive from [66, 22, 21] that $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$ is of order n [1]. In particular, this is significantly smaller than n^2, which is the number of entries of M and a trivial upper bound on the minimax risk.

The rate of order n can be achieved in polynomial-time in the BTL model. Indeed, the Maximum Likelihood Estimator (MLE) $\hat{\theta}$ is an efficient method that leads to a minimax optimal estimator of M – see e.g. [21]. The MLE $\hat{\theta}$ is given as the following minimization problem:

$$\hat{\theta} := \operatorname*{arg\,min}_{\theta' \,:\, \mathbf{1}^T\theta'=0} \sum_{i<j} Y_{ij} \log\left(\frac{1}{\psi(\theta'_i - \theta'_j)}\right) + (1 - Y_{ij}) \log\left(\frac{1}{1 - \psi(\theta'_i - \theta'_j)}\right) \ , \tag{2.6}$$

which is a convex minimization problem on a vector space. Moreover, the corresponding estimator $\hat{M}$ of M can be defined as $\hat{M}_{ij} = \psi(\hat{\theta}_i - \hat{\theta}_j)$.

2.2.1.2 The SST model in tournament contexts

While the BTL and noisy sorting models are practical benchmark models, it has been noted that they often lack realism and that they may not fit the data well. To address these issues, the strong stochastically transitive model (SST) has been introduced as a much more flexible model [83, 86, 60, 56]. Specifically, it replaces the stringent parametric assumptions of the noisy sorting and BTL models by non-parametric assumptions with shape constraints.

Similarly to BTL and noisy sorting models, consider a scenario where we observe an $n \times n$ matrix Y of Bernoulli observations. The upper triangular elements of this matrix are independent, and each entry $Y_{ij} = 1 - Y_{ji}$ has a Bernoulli parameter M_{ij}. In particular, M is skew-symmetric, that is $M_{ij} = 1 - M_{ij}$. In SST, it is additionally assumed that there exists an unknown permutation π^* such that when the rows and columns of the matrix M are rearranged according to π^*, the resulting matrix $M_{\pi^{*-1}\pi^{*-1}}$ is bi-isotonic – its rows and columns are non-decreasing. Equivalently,

$$M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{SST}} \ , \tag{2.7}$$

where $\mathcal{M}_{\mathrm{SST}}$ denotes the set of all matrices that are skew-symmetric and bi-isotonic. In particular, the SST model encompasses the BTL model since $\mathcal{M}_{\mathrm{BTL}} \subset \mathcal{M}_{\mathrm{SST}}$. In other words, the matrix with coefficients $M'_{ij} = \psi(\theta_i - \theta_j)$ is skew-symmetric ($M'_{ij} = 1 - M'_{ji}$), and it is bi-isotonic up to a permutation π^* of its rows and columns. We refer the reader to Figure 2.1 for an example of a 3×3 matrix that is both skew-symmetric and bi-isotonic. In the context of a tournament, the assumption of bi-isotonicity can be explained as follows. If a player i is better on average than another player j, then i should win more on average than j against any other player k. More formally, if $M_{ij} \geq 1/2$, then $M_{ik} \geq M_{jk}$. In the literature, there has also been significant focus on the scenario where only a proportion $\lambda \in [0, 1]$ of the pairs (i, j) is observed – see e.g. [19]. Nevertheless, in this discussion on the SST model, we assume for simplicity that all the nd observations are available.

$$M_{\pi^{*-1}\pi^{*-1}} = \begin{pmatrix} 0.5 & 0.6 & 0.8 \\ 0.4 & 0.5 & 0.7 \\ 0.2 & 0.3 & 0.5 \end{pmatrix}$$

Figure 2.1: An example of a bi-isotonic and skew symmetric matrix.

Similarly to what we presented for the BTL model, a natural question often raised in the literature on the SST model [83, 19, 60, 59, 56, 71] can be framed as follows: *How accurately can we reconstruct the reordered matrix M from the observed data Y ?* To quantify this, consider again the following minimax reconstruction risk:

$$\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n) := \inf_{\hat{M}} \sup_{\pi^*, M} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right] \ , \tag{2.8}$$

where here, the supremum is taken over all permutation π^* and over all matrices M such that $M_{\pi^{*-1}\pi^{*-1}} \in \mathcal{M}_{\mathrm{SST}}$.

Shah et al. [83] established a surprising fact: $\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n)$ is of the order of n. Consequently, $\mathcal{R}^{*\mathrm{SST}}_{\mathrm{reco}}(n)$ and $\mathcal{R}^{*\mathrm{BTL}}_{\mathrm{reco}}(n)$ are of the same order. Hence, from a statistical point of view, it is not much easier to reconstruct M in the BTL model than in the SST model. In contrast, recall that the SST model has n^2 unknown parameters, while the BTL model contains only n unknown parameters.

[1] [66, 22, 21] only provide the minimax risk for the estimation of θ, but we could deduce the minimax optimal rate for the reconstruction M from their result.

Given a matrix of observation Y, a natural estimator of π^* and of the sorted matrix $M_{\pi^{*-1}\pi^{*-1}}$ is the least square estimator, defined by

$$(\hat{\pi}^{LS}, \hat{M}^{LS}_{\text{sorted}}) = \underset{\pi\in\Pi_n, \tilde{M}\in\mathcal{M}_{\text{biso}}}{\arg\min} \|Y_{\pi^{-1}\pi^{-1}} - \tilde{M}\|_F^2 \ , \tag{2.9}$$

where Π_n denotes the set of all permutations and $\mathcal{M}_{\text{biso}}$ the set of all bi-isotonic matrices. Then, the corresponding estimator of M is $\hat{M}^{LS} = (\hat{M}^{LS}_{\text{sorted}})_{\hat{\pi}^{LS}\hat{\pi}^{LS}}$. Shah et al. [83] established that the least square estimator $\hat{M}^{LS}$ is minimax optimal: its maximum risk of reconstruction is of the order n. Unfortunately, no known polynomial-time method exists to solve the above minimization problem (2.9), primarily because it involves an exhaustive search over all $n!$ permutations.

The same authors also propose an efficient estimator, based on global average comparison, which amounts to ranking the players according to the means of the rows of Y. Essentially, the method consists in first estimating π^* by the permutation $\hat{\pi}$ which sorts the row means of Y. Then, an estimation of M is derived by minimizing the least squares over all bi-isotonic matrices:

$$\hat{M}_{\text{sorted}} = \underset{\tilde{M}\in\mathcal{M}_{\text{biso}}}{\arg\min} \|Y_{\hat{\pi}^{-1}\hat{\pi}^{-1}} - \tilde{M}\|_F^2 \ . \tag{2.10}$$

In contrast to the least square estimator (2.9), this polynomial-time estimator only achieves a rate of order $n^{3/2}$.

A lot of effort has since been dedicated to narrowing the computational gap between $n^{3/2}$ and n. In short, Chatterjee and Mukherjee [19] provided a method that adapt to the regularity of the matrix $M_{\pi^{*-1}\pi^{*-1}}$, or to the case where it is a block matrix. Nevertheless, they did not improve the rate $n^{3/2}$ in the worst case. Then, Mao et al. [60, 59] introduced a polynomial-time method achieving the rate $n^{5/4}$. To recover π^*, the approach of [60] involves two main steps. First, the players are sorted according to the means of the rows of Y, as in [83]. This gives a pre-sorted matrix Y'. Then, the players are compared according to local means of the rows of Y' over intervals included in $\{1,\dots,n\}$. However, as pointed out by Liu and Moitra [56], [60] did not leverage the global information available in the whole matrix as they only compare players two by two. Building upon this remark, Liu and Moitra [56] successfully achieved the better rate $n^{7/6+o(1)}$, using a polynomial-time method in the case where we have access to $n^{o(1)}$ independent samples per entry. These findings are also discussed in Chapter 4, where we also recover a rate of order $n^{7/6}$ in the SST model, up to polylogarithmic factor. Hence, in the SST model, the minimax risk is of order n, but the best known polynomial-time method only achieves a rate of order $n^{7/6}$. It is still an open question whether the gap between n and $n^{7/6}$ is intrinsic to the SST model, or if it can be further reduced.

Since the least square estimator achieves the minimax risk of order n, it might be tempting to solve the computational issue with a convex relaxation of (2.9). A first idea is to compute the least-square $\|Y-\tilde{M}\|_F^2$ over all the matrices $\tilde{M}$ in the convex hull of $\{M \in \mathbb{R}^{n\times d} \ : \ M_{\pi^{*-1}} \in \mathcal{M}\}$. Nonetheless, such an estimator is unlikely to achieve good theoretical guarantees, and we refer the reader to [82] for some negative result – at least in the bi-isotonic-2D model. Another idea is to use the Birkhoff-von Neumann theorem. Let $\mathcal{P}_n$ be the set of doubly stochastic $n\times n$ matrices and consider the following relaxation of (2.9):

$$\hat{M}^{REL}_{\text{sorted}} = \underset{P\in\mathcal{P}_n \tilde{M}\in\mathcal{M}}{\arg\min} \|Y - P\tilde{M}\|_F^2 \ . \tag{2.11}$$

Both $\mathcal{P}_n$ and $\mathcal{M}$ are convex sets and the minimization function is convex in P and in $\tilde{M}$. However, this function is not jointly convex in $(P,\tilde{M})$, and up to our knowledge, the existence of an efficient procedure solving (2.11) is an open problem. Moreover, it is unclear whether the minimizer of the relaxed problem (2.11) achieves the optimal convergence rate of n, as does the least square estimator.

2.2.1.3 Other non-parametric models in crowdsourcing

Beyond tournament problems and motivated by crowdsourcing problems, there has been a recent surge in the development of new non-parametric models [60, 81, 84, 85, 56, 33]. Before exploring these models, let us first describe the general framework, which is similar to that of the SST model. Let M_{ik} be any rectangular $n\times d$ matrix whose coefficient are in $[0,1]$. In crowdsourcing data, n refers to the number of experts or workers, d represents the number of questions or tasks and M_{ik} denotes the probability that expert i provides a correct response to question k. In particular, $M_{ik} = 1/2$ means that expert i gives a random guess of question k and $M_{ik} = 1$ means that he perfectly knows the correct answer. Let us assume that for each pair (i,k) of expert/question, we receive N_{ik} independent Bernoulli observations

$$Y^{(u)}_{ik} = \text{Bern}(M_{ik}), \quad u = 1,\dots,N_{ik} \ . \tag{2.12}$$

The pair (i,k) is observed if and only if $N_{ik} > 0$ and $Y_{ik}^{(u)} = 1$ means that expert i is correct to question k at trial u. To take into account possible partial observations, we use a standard poissonization trick. Namely, we assume that N_{ik} follows a Poisson distribution of parameter $\lambda > 0$ – see e.g. [60]. The interesting case is when $\lambda \leq 1$, since it corresponds to a proportion of missing data of order $1 - \lambda$. Subsequently, we expose the bi-isotonic-2D, the isotonic and the bi-isotonic-1D models. Each one of these models include some form of shape constraint on matrix M, since the reconstruction of M is hopeless without any further assumption.

The bi-isotonic-2D model. In the context of crowdsourcing data, the counterpart of the SST model is the bi-isotonic-2D model. Assume that there exists two unknown permutations π^* and η^* such that $M_{\pi^{*-1}\eta^{*-1}}$ is bi-isotonic, i.e. that has non-decreasing rows and columns. We write $\mathcal{M}_{\text{biso}}$ for the set of all bi-isotonic matrices, so that we have

$$M_{\pi^{*-1}\eta^{*-1}} \in \mathcal{M}_{\text{biso}} \ . \tag{2.13}$$

To illustrate, Figure 2.2 offers a visual representation of a randomly generated bi-isotonic matrix by displaying a plot of each of its rows. This model implies that there exists an intrinsic order π^* that ranks the experts according to their abilities, and another intrinsic order η^* that sorts the questions according to their difficulties. This model encompasses the aforementioned SST model, where the additional assumptions are that $\pi^* = \eta^*$ and that M is skew-symmetric.

As a theoretical guarantee, the reconstruction loss for a given estimator $\hat{M}$ is defined as $\|\hat{M} - M\|_F^2$, which is analogous to the loss function defined in the SST model. The worst case risk $\sup \mathbb{E}[\|\hat{M} - M\|_F^2]$ for this loss is taken over all permutation π^*, η^* and matrices M such that $M_{\pi^{*-1},\eta^{*-1}}$ is bi-isotonic. Mao et al. [60] have shown that when M is a square matrix, that is $n = d$, the minimax risk in this model is of order n up to polylogarithmic factors. Specifically, it is of the same order as in the SST model and the bi-isotonic-2D model is therefore not much statistically harder than SST.

Let us discuss the computational issues in this model. Liu and Moitra [56] established that, when $n = d$ and $\lambda = n^{o(1)}$, there exists a polynomial-time method that achieves a maximum risk of order $n^{7/6+o(1)}$. It turns out that the assumption that $\lambda = n^{o(1)}$ can be relaxed to $\lambda = 1$, as we show in a corollary in Chapter 4 for the isotonic model. The case $\lambda = 1$ corresponds to the situation where, on average, we have one observation for each pair (i,k). Meanwhile, $\lambda = n^{o(1)}$ represents a sub-polynomial number of observations for each pair. Overall, similarly to the SST model, whether the same computational-statistical gap between n and $n^{7/6}$ is intrinsic to the problem remains an open question in this model as well.

The bi-isotonic-1D model. A more specific model motivated by crowdsourcing data is the bi-isotonic-1D model. The stronger assumption is that the matrix M is bi-isotonic up to a single permutation π^* acting on its rows. Equivalently,

$$M_{\pi^{*-1}} \in \mathcal{M}_{\text{biso}} \ , \tag{2.14}$$

where $(M_{\pi^{*-1}})_{ik} = M_{\pi^{*-1}(i),k}$. Alternatively, the bi-isotonic-1D model can be viewed as a special case of the bi-isotonic-2D model, since it corresponds to the case where the permutation on the columns η^* is known and equal to the identity. In practice, knowing η^* in the bi-isotonic-2D model amounts to assuming that the statistician has access to the intrinsic difficulty of the questions, which is a stronger assumption. Surprisingly, in the case where $n = d$ and $\lambda = 1$, the reconstruction rate is not better in the bi-isotonic-1D model compared to the bi-isotonic-2D or SST models. Indeed, Mao et al. [60] established that the minimax risk for reconstruction is of the order of n, as in the SST and in the bi-isotonic-2D models.

Nevertheless, Mao et al. [60] left a computational-statistical gap in this model: their optimal solution is unlikely to be computable in polynomial-time, and their efficient method can only reconstruct M at a rate $n^{5/4}$. In [60], the main approach is to compare the rows of Y two by two. This is done by computing local means of the rows over local intervals included in $\{1, \dots, d\}$. More recently, Liu and Moitra [56] almost closed the computational gap in the case $n = d$ and $\lambda = n^{o(1)}$. The authors established a polynomial-time method that achieves a reconstruction rate of $n^{1+o(1)}$, which is nearly minimax optimal. In contrast to [60], a crucial idea introduced by Liu and Moitra is to focus on specific intervals before averaging the observations. Namely, for a given set P of rows, they focus on regions where the mean of all the rows of P changes significantly.

The isotonic model. Lastly, an extension of the bi-isotonic-2D model is the isotonic model [33]. Within this framework, the only assumption is that all the columns of M are non-decreasing up to an unknown permutation of the rows π^*. Equivalently, if $\mathcal{M}_{\text{iso}}$ denotes the set of all isotonic matrices, we have that

$$M_{\pi^{*-1}} \in \mathcal{M}_{\text{iso}} \ . \tag{2.15}$$

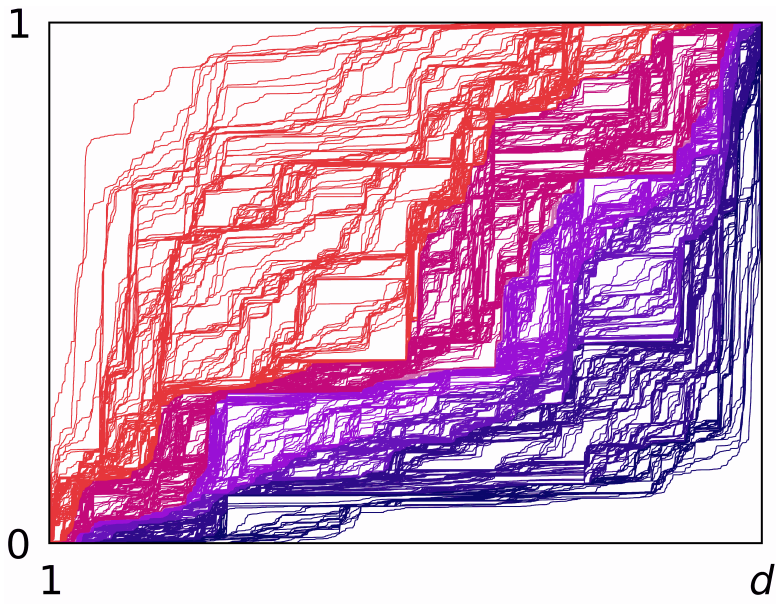


Figure 2.2: An example of a bi-isotonic matrix $M_{\pi^{*-1}}$. Each colored line lies in $[0,1]$ and represents a row of $M_{\pi^{*-1}}$. Since $M_{\pi^{*-1}}$ is bi-isotonic, these lines are nondecreasing.

In other words, the isotonic model is a relaxation of the bi-isotoni-2D model, without the assumption that the rows are non-decreasing up to a permutation η^*. Notably, this makes the isotonic model more flexible, even if the reconstruction of M becomes statistically harder. To simplify the comparisons with the other models, assume that $\lambda = 1$ and that $n = d$. Flammarion et al. [33] have shown that the minimax risk for this model is of the order of $n^{4/3}$. They also introduced RankScore, a computationally efficient method that is based on global average comparison and entrywise comparison. However, RankScore only achieves a maximum risk of the order of $n^{3/2}$, which is suboptimal. In the worst case, RankScore reaches similar performances to simply ranking the rows according to their averages. Nevertheless, there is no computational gap in this model and the rate $n^{4/3}$ can be reached in polynomial-time, up to polylogs. This is proven in the analysis of the isotonic model, in Chapter 4. Aside from when $n = d$, another interesting case is when $d = 1$, which amounts to assuming that M is a column vector.

In the case $d = 1$, this model closely related to a problem of uncoupled isotonic regression, which finds motivation in optimal transports or social science problems [76, 14]. Assume that we observe the unordered sets $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, linked by the relation $y_i = f(x_i) + \varepsilon_i$ for some unknown non-decreasing function f and noise ε with independent coefficients. As illustrated in [14], the x_i and y_i can respectively represent the wage data collected by a governmental agency and the y_i the housing prices collected by a bank. In our case, $x_i = i$, $f(x_i) = M_{\pi^{*-1}(i)}$, and estimating f comes down to estimating the sorted vector $M_{\pi^{*-1}}$. Rigollet and Niles-Weed [76] have established that the minimax risk for the estimation of f, or $M_{\pi^{*-1}}$, is of the order of $n(\frac{\log\log(n)}{\log(n)})^2$.

Although the three aforementioned models seem structurally similar, they substantially differ from a statistical point of view. In particular, much more information is available when we assume that the order of the columns is known in the bi-isotonic-1D model. Figure 2.3 gives an illustration of the difference between the isotonic and bi-isotonic-1D models by depicting randomly generated matrices M.

Overall, the relations between the three models can be summarized as follows: the isotonic model is an extension of the bi-isotonic-2D model, which in turn is an extension of the bi-isotonic-1D model. Let us address unresolved computational-statistical gaps in the literature concerning the reconstruction of M across these three models. In the isotonic model, [33] left a computational gap between the optimal rate $n^{4/3}$ and $n^{3/2}$, in the case $n = d$. Concerning the bi-isotonic-1D model, [56] almost closed the gap and achieved a rate of order $n^{1+o(1)}$ in the case $n = d$. However, a significant computational gap remained in the bi-isotonic-1D model for all n, d, λ such that $n \ll d$. Finally, in the bi-isotonic-2D model, [56] managed to narrow the gap. Nevertheless, whether it is possible to further reduce the gap between n and $n^{7/6+o(1)}$ remains an open question.

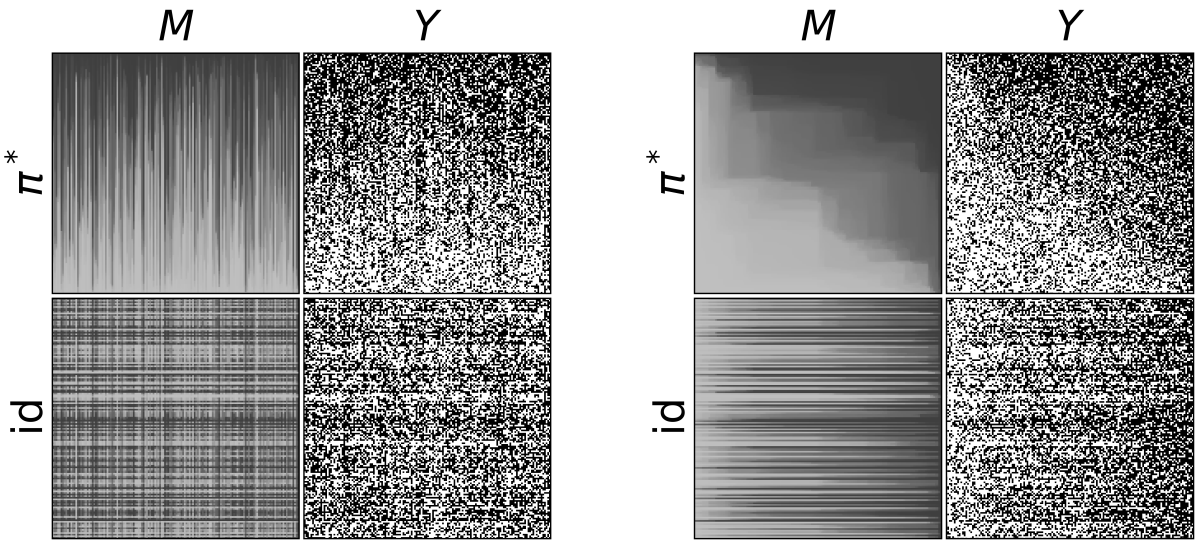


Figure 2.3: For each model – isotonic (left) or bi-isotonic-1D (right) – the matrices $M_{\pi^{*-1}}$,$Y_{\pi^{*-1}}$, M, Y are respectively represented in the reading order.

2.2.2 Overview of our contribution

In Chapter 3 and Chapter 4, we close the existing computational gaps in both the bi-isotonic-1D (2.14) and isotonic (2.15) models, for almost all possible values of n, d and λ. Additionally, we further narrow the gap in the bi-isotonic-2D (2.13) and SST models.

In what follows, we summarize our main contributions, and we focus our attention on the isotonic and bi-isotonic-1D models. Recall that $\mathcal{M}_{\text{iso}}$ and $\mathcal{M}_{\text{biso}}$ denote the sets of all isotonic and bi-isotonic matrices respectively, and let $\mathcal{M}$ be either $\mathcal{M}_{\text{iso}}$ or $\mathcal{M}_{\text{biso}}$. As in the previous section, M represents an unknown matrix with entries in $[0,1]$, and is such that $M_{\pi^{*-1}} \in \mathcal{M}$ for some unknown permutation π^* of the rows. Consider some observation Y given by model (2.12).

2.2.2.1 Minimax risk for permutation estimation

As highlighted in Section Section 2.2.1, estimating the matrix M is a significant problem within the ranking literature [83, 84, 60, 56, 71]. However, the primary focus in ranking is not so much to reconstruct the entire matrix M, but rather to find a good estimation of the original order π^*. Therefore, we take a different approach to build an estimator of π^* and to measure its quality. Given an estimator $\hat{\pi}$ of π^*, let $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ be the permutation loss. In contrast to the reconstruction loss $\|\hat{M} - M\|_F^2$, this loss quantifies the distance between the matrix M reordered according to the estimator $\hat{\pi}$ and the matrix M sorted according to the ground truth permutation π^*. In particular, it is not necessary to define an estimator $\hat{M}$ of the whole matrix M to measure the quality of a given estimator $\hat{\pi}$ of π. We define the minimax risks for both permutation estimation and matrix reconstruction as

$$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}(n,d,\lambda) := \inf_{\hat{\pi}} \sup_{\substack{\pi^* \in \Pi_n \\ M:\, M_{\pi^{*-1}} \in \mathcal{M}}} \mathbb{E}\big[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2\big]$$

$$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}(n,d,\lambda) := \inf_{\hat{M}} \sup_{\substack{\pi^* \in \Pi_n \\ M:\, M_{\pi^{*-1}} \in \mathcal{M}}} \mathbb{E}\big[\|\hat{M} - M\|_F^2\big] \quad .$$

The definition of both of these two risks allows to decipher the difficulty of estimating the permutation π^* from the difficulty of reconstructing the whole matrix M.

2.2.2.2 Results

Whether we consider the isotonic (2.15) or bi-isotonic-1D (2.14) model, it turns out that there exists a polynomial-time estimator $\hat{\pi}$ that is nearly minimax optimal, for almost all regimes in n, d and λ. Moreover, any optimal estimator of π^* can be used to build an optimal estimator of M. Hence, there is no significant

Bi-isotonic-1D Model (2.14):

	$n \lesssim d^{1/3}$	$d^{1/3} \lesssim n \lesssim d$	$d \lesssim n$
$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$	$nd^{1/3}$	$\sqrt{nd}$	n

Isotonic Model (2.15):

	$n \lesssim d^{3/2}$	$d^{3/2} \lesssim n$
$\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$	$n^{2/3}\sqrt{d}$	n
$\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$	$n^{1/3}d$	n

Figure 2.4: Optimal rates in the bi-isotonic-1D and isotonic models, for all possible values of n, d and $\lambda = 1$, up to a polylogarithmic factor in nd. These rates are achieved by polynomial-time estimators. For further details, see Chapters 3 and 4, respectively.

computational statistical gap for both permutation estimation and matrix reconstruction, unlike in the SST and the bi-isotonic-2D models.

Consider for simplicity the regime where $\lambda = 1$, and let us discuss the minimax permutation rates summarized in the tables of Figure 2.4. First, the reconstruction of M is statistically harder than the estimation of π^*. Indeed, we deduce from Figure 2.4 that $\mathcal{R}^{*\mathcal{M}}_{\text{reco}} \gtrsim \mathcal{R}^{*\mathcal{M}}_{\text{perm}}$, up to polylogarithmic factors. Essentially, the rate of reconstruction $\mathcal{R}^{*\mathcal{M}}_{\text{reco}}$ can be decomposed into two components: the rate of permutation estimation $\mathcal{R}^{*\mathcal{M}}_{\text{perm}}$ and the rate of reconstructing a sorted matrix, that is when π^* is known.

To illustrate the rates of permutation estimation, let us consider the case $n = 2$. In both models, the matrix M has two rows, with one uniformly above the other. Without loss of generality, assume that $M_{1,k} \geq M_{2,k}$ for all k. When $n = 2$, this is the only assumption in the isotonic model. However, in the bi-isotonic-1D model, it is further assumed that each row M_1 and M_2 is non-decreasing. This additional assumption explains why the rate of $d^{1/6}$ in the bi-isotonic-1D model is much smaller than the rate of $\sqrt{d}$ in the isotonic model.

Let us give the intuition of the rate $\sqrt{d}$ in the isotonic model. Consider the simple method which compares the means of the two rows Y_1 and Y_2. With this method, if the mean of row 1 is larger, then we recover the true permutation and the loss is equal to 0. Otherwise, we reverse the order of the rows and the permutation loss is equal to $2\|M_1 - M_2\|_2^2$. It turns out that this simple method achieves the optimal rate $\sqrt{d}$. Subsequently, we provide the main arguments for this claim. Since row 1 is above row 2, we have that

$$\sum_{k=1}^{d} Y_{1,k} - Y_{2,k} = \|M_1 - M_2\|_1 + \sum_{k=1}^{d} E_{1,k} - E_{2,k} \ ,$$

where $E_{ik} = Y_{ik} - M_{ik}$. Using the Hoeffding inequality for Bernoulli random variables, we deduce that $|\sum_{k=1}^{d} E_{1,k} - E_{2,k}| \leq C\sqrt{d}$ for some constant C, with probability at least 0.99. Moreover, since $M_{ik} \in [0,1]$ for all i, k, it holds that

$$\|M_1 - M_2\|_1 \geq \|M_1 - M_2\|_2^2 \ .$$

Hence, if $\|M_1 - M_2\|_2^2 > C\sqrt{d}$, then the mean of row 1 is above the mean of row 2 with probability 0.99. On this high probability event, we recover the true permutation and the loss in square Frobenius norm is equal to 0. Otherwise, when $\|M_1 - M_2\|_2^2 \leq C\sqrt{d}$, the loss is bounded by $2C\sqrt{d}$. In fact, this upper bound of order $\sqrt{d}$ is optimal in a minimax sense, when $n = 2$. Moreover, this bound can be extended to larger n, resulting in a suboptimal upper-bound of $n\sqrt{d}$. Specifically, this is the underlying idea behind the suboptimal rate of order $n^{3/2}$ established by Shah et al. [83] in the SST model (2.7). In the bi-isotonic-1D model, we can achieve the rate $d^{1/6}$ when $n = 2$ by using the fact that the two rows M_1 and M_2 are nondecreasing. See Chapter 3 for more details.

When both n and d are equal, the rate for estimating the permutation is of order $n^{7/6}$ in the isotonic model, and it is achieved by a polynomial time estimator $\hat{\pi}$ of π^*. This rate was originally established in the bi-isotonic-2D or SST model by Liu and Moitra [56], up to a number of samples of the order of $n^{o(1)}$. However, in contrast to [56], our method in the isotonic model presented in Chapter 4 does not require any assumption on the rows of M. The estimator $\hat{\pi}$ can also be used to define an estimator of the matrix M which achieves a reconstruction rate of order $n^{4/3}$ in the isotonic model – see Figure 2.4 with $n = d$. Additionally, we show in Corollary 4.2.5 of Chapter 4 that, in the bi-isotonic-2D or SST model, it is also possible to derive an estimator of M from $\hat{\pi}$ that achieves a reconstruction rate of order $n^{7/6}$. This $n^{7/6}$ rate, as previously mentioned, is the best known rate for polynomial-time matrix reconstruction or permutation estimation in the bi-isotonic-2D and SST models.

2.2.2.3 General ideas of the procedures

The procedures that we describe in Chapter 3 for the bi-isotonic-1D model and in Chapter 4 for the isotonic model are substantially different. The former relies on hierarchical clustering with memory, while the latter is based on a comparison graph. However, it is worth noting that our analysis of the isotonic model builds upon

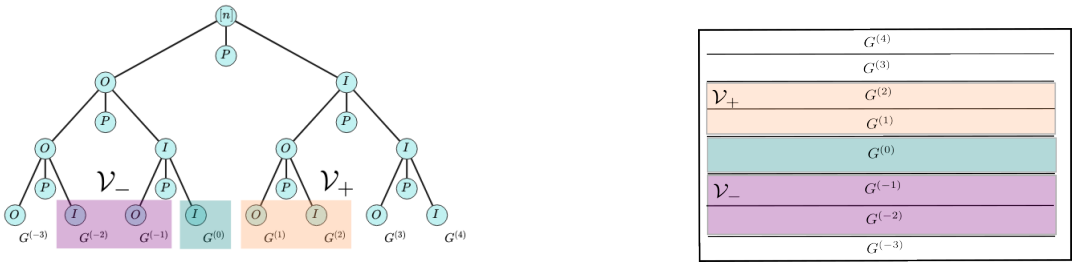


Figure 2.5: Example of a hierarchical sorting tree (left), and of the matrix M sorted with the tree (right). $\mathcal{V}_-$ (resp. $\mathcal{V}_+$) represents a set of group of rows that are below (resp. above) the group $G^{(0)}$.

several elements originally introduced in our analysis of the bi-isotonic-1D model. Subsequently, we give an informal overview of the two approaches.

In Chapter 3, we aim at building a sorting tree using a top-down strategy, as illustrated in Figure 2.5. Starting with the complete set of rows $[n]$, we divide it into three subsets (O, P, I) of $[n]$, where O and I contain rows that are likely to be below and above the median row, respectively. The subset P contains rows which cannot be classified with high confidence. Then, we recursively divide the subsets O and I, as shown in the left part of Figure 2.5. When the tree is completed, we obtain a partial ordering on all the rows which can be used to estimate π^*.

The main difficulty in this procedure is to optimally divide a given set $G \subset [n]$ into (O, P, I). We obtain this by essentially combining techniques ranging from change-point detection to spectral methods. To compute the subsets (O, P, I) of a given set G, it is also crucial to keep in memory the sorting tree. Indeed, this approach allows us to use valuable information from the other leaves of the tree to refine the division of G. For instance, in Figure 2.5, the group $G^{(0)}$ could be further divided using the information that it is sandwiched between the two sets of rows $\mathcal{V}_-$ and $\mathcal{V}_+$.

In contrast, the method based on a comparison graph in Chapter 4 amounts to iteratively updating a weighted directed graph. The edges of this graph quantify the level of the comparison between the rows of M. An edge that points from a row i to another row j means that i should be above j. Moreover, we are more confident about the order between the rows for which the edges have a larger weight. At the last update of the graph, we obtain a weighted graph from which we derive an estimator of π^*.

Interestingly, the technique based on a comparison graph is closely related to the method relying on hierarchical clustering. The main connection between the two approaches can be summarized as follows. In a comparison graph, the core idea is to update the weights of the edges between a given row i and the other rows in its neighborhood P which is itself computed from the weighted graph. From a broad perspective, at each iteration, a subset of columns $\hat{Q} \subset [d]$ is computed to reduce the dimension and to compare row i with the other rows in P using weighted sums. Similarly, the hierarchical clustering approach involves comparing rows in a set P using weighted sums computed over subsets $\hat{Q}$. However, instead of updating edges between rows, the hierarchical clustering approach involves a clustering of P. Each clustering step consists in computing two subsets L and U of P, such that the rows in set L are provably below the rows in set U. The strong relations between these two methods suggests that both could be applied to both the isotonic and bi-isotonic-1D models to achieve the minimax risks up to polylogarithmic factors.

2.3 Change-point detection

Change-point detection has a rich historical background, starting with Wald's seminal work [95], which has since inspired significant advancements in the field [68, 89]. As mentioned earlier, detecting change-point is crucial in a wide range of practical situations, from monitoring daily temperature fluctuations and observing stock market trends to examining genomic information. In what follows, we start by discussing the univariate case, where we observe a sequence of real-valued data. Then, we introduce the problem of change-point detection in high-dimensional time series before moving to our contribution in this setting.

2.3.1 Discussion on the univariate case

In this discussion, we focus exclusively on univariate time series. Assume that we observe a sequence of independent real-valued random variables $(y_1, \dots, y_n)$, with cumulative distribution functions $(F_1, \dots, F_n)$. We say that a change-point occurs at a position τ if the cumulative distribution functions at τ is different from the previous one, i.e., $F_{\tau-1} \neq F_\tau$. In particular, if we know that the number of change-points K is at most equal

to 1, the question becomes whether the distribution of the data remains stationary over time, or if there is a detectable change in it. While this non-parametric model encompasses many situations, it is often too broad in many practical applications [68]. For this reason, the distributions F_i often need to be parametrized.

Henceforth, we assume that for each $t = 1, \dots, n$, we have the following signal/noise decomposition:

$$y_t = \theta_t + \varepsilon_t \in \mathbb{R} \ , \tag{2.16}$$

where (θ_t) is an unknown deterministic sequence, and the noise $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered standard Gaussian variable $\mathcal{N}(0,1)$. In this model, the sequence of change-points $(\tau_1, \dots, \tau_K)$ corresponds to positions τ_k where $\theta_{\tau_k - 1}$ differs from θ_{τ_k}. For each change-point τ_k, we define $D_k \in \mathbb{R}$ as the difference $\theta_{\tau_k} - \theta_{\tau_{k-1}}$, representing the average change in the data. We also define r_k as the distance between τ_k and its closest adjacent change-point, that is $r_k = \min(\tau_k - \tau_{k-1}, \tau_{k+1} - \tau_k)$. By convention, we set $\tau_0 = 1$ and $\tau_{K+1} = n + 1$. Let us now discuss the single and multiple change-point detection problems in this setting.

Single change-point detection. Assume that we know that there is at most one change-point, i.e. $K \leq 1$. We are essentially dealing with a hypothesis testing problem, and we aim to test the two following hypotheses:

$$\begin{aligned} H_0 &: \text{ There is no change-point} \\ H_1 &: \text{ There is a single change-point at an unknown position } \tau \ . \end{aligned}$$

We aim to determine whether there is one change-point or not in the sequence (θ_t). If a change-point τ exists, then θ_t is equal to μ_1 if $t < \tau$ and to $\mu_2 \neq \mu_1$ otherwise. In this single change-point model, the original approach of Hinkley [44] is to maximize the absolute value of the CUSUM statistic

$$\mathbf{C}_t(y) = \sqrt{\frac{(t-1)(n-t+1)}{n}} \left(\frac{1}{n-t+1} \sum_{i=t}^{n} y_i - \frac{1}{t-1} \sum_{i=1}^{t-1} y_i \right) \ , \tag{2.17}$$

over all possible positions $t = 2, \dots, n$. In simpler terms, $\mathbf{C}_t(y)$ represents the rescaled difference between the average of the data over the interval $[t, n]$ and the average over $[1, t)$. The primary idea is that if there is no change-point, then $\mathbf{C}_t(y)$ follows a standard normal distribution $\mathcal{N}(0,1)$ for all t. On the other hand, if there is a change-point at position τ, then $\mathbf{C}_\tau(y)$ follows a normal distribution with expectation equal to $\sqrt{\frac{(\tau-1)(n-\tau+1)}{n}} D$, where $D = \mu_2 - \mu_1$. In particular, this quantity satisfies:

$$\tfrac{1}{2} r D^2 \leq \tfrac{(\tau-1)(n-\tau+1)}{n} D^2 \leq r D^2 \ , \tag{2.18}$$

where $r = \min(\tau, n + 1 - \tau)$.

Detecting τ becomes statistically easier as rD^2 increases. It turns out that the quantity rD^2 precisely characterizes the limit of detection in this problem. Recently, Gao et al. [38] established that the change-point can be detected with high probability, as soon as $rD^2 \geq C\sqrt{\log\log(n)}$, for some constant C that only depends on the desired probability of error. See also the earlier work of Csörgö and Horváth [26] for a related asymptotic result. For a more precise result replacing $\sqrt{\log\log(n)}$ by $\sqrt{\log\log(n/r)}$, we refer the reader to the work of Verzelen et al. [92].

Multiple change-point detection. The problem is more challenging in the case where K is unknown and arbitrary. To estimate one of the K change-points, one could be tempted to take the maximum over all t of the absolute value of $\mathbf{C}_t(y)$, as in the previous setting with a single change-point. This is the principle of binary segmentation (BS) [80], but unfortunately, this does not lead to a consistent estimator of one of the change-points. Subsequently, we briefly outline two classes of methods that aim to overcome the limitations of BS.

A wide range of methods are based on a variant of binary segmentation: wild binary segmentation (WBS), which is a top-down approach that was introduced by Fryzlewicz [36]. In WBS, we first draw at random some random intervals $[s, e) \subset [n]$. Then, we maximize the local CUSUM statistic

$$\mathbf{C}_{s,t,e}(y) = \sqrt{\frac{(t-s)(e-t)}{e-s}} \left(\frac{1}{e-t} \sum_{i=t}^{e-1} y_i - \frac{1}{t-s} \sum_{i=s}^{t-1} y_i \right) \ ,$$

over all possible $t \in [s, e)$ and all randomly chosen intervals $[s, e)$. If the maximum is above some threshold in absolute value, then we take the corresponding t as the first estimator of a change-point. Then, we subdivide $[n]$ into $[1, t-1]$ and $[t, n]$, and we recursively look for other potential change-points in these two intervals.

Another class of methods is based on penalized least square criteria [92, 98]. The primary idea is to estimate a piecewise constant sequence $\hat{\theta}$ through the following minimization problem

$$\hat{\theta} = \underset{\theta' \in \mathbb{R}^n}{\arg\min} \sum_i (y_i - \theta'_i)^2 + \lambda \text{pen}(\theta') \ , \tag{2.19}$$

where λ is a tuning parameter and $\text{pen}(\theta') \geq 0$ is a function of θ' that aims at penalizing the variations of θ'. In particular, Wang et al. [98] define $\text{pen}(\theta')$ as the number of positions i where $\theta'_{i-1} \neq \theta'_i$. While the minimization problem (2.9) is not convex, $\hat{\theta}$ can still be computed efficiently using dynamic programming techniques – see e.g. algorithm 1 of Friedrich et al. [35].

Similarly to the single change-point detection problem, the minimal condition of detection of a given change-point τ_k in this context depends on the quantity $r_k D_k^2$, where we recall that, when $K \geq 2$, $r_k = \min(\tau_k - \tau_{k-1}, \tau_{k+1} - \tau_k)$ and $D_k = \theta_{\tau_k} - \theta_{\tau_{k-1}}$. Wang et al. [98] established that, as soon as $r_k D_k^2 \gtrsim \sqrt{\log(n)}$ for all k, we can detect all the change points with high probability. This finding was further refined by Verzelen et al. [98], who showed that the minimal condition needed of detection of change-point τ_k is $r_k D_k^2 \gtrsim \sqrt{\log(n/r_k)}$.

2.3.2 The multivariate case

In the multivariate case, the observations $y_1, \dots, y_n$ belong to the vector space $\mathbb{R}^p$, with arbitrary dimension $p \geq 1$. We observe, for $t = 1, \dots, n$

$$y_t = \theta_t + \varepsilon_t \in \mathbb{R}^p \ . \tag{2.20}$$

In contrast to the univariate case (2.16), θ_t is a vector in $\mathbb{R}^p$ and the random variables $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. multivariate standard Gaussian random variables $\mathcal{N}(0, \mathbf{I}_p)$. In this context, the k^{th} change-point τ_k is still defined as $\theta_{\tau_k} \neq \theta_{\tau_k - 1}$, and $D_k = \theta_{\tau_k} - \theta_{\tau_k - 1}$ is a vector in $\mathbb{R}^p$. We write s_k for the sparsity of D_k, that is $s_k = \|D_k\|_0$. In simpler terms, s_k represents the number of non-zero entries of D_k.

Among the literature on multivariate time series, a lot of effort has been focused on adapting to the sparsity of D_k [103, 47, 57]. Similarly to the univariate case, we look for the minimal value of $r_k \|D_k\|_2^2$ such that the change-point τ_k can be accurately detected. In this context, this value depends on n, r_k, p and s_k.

Liu et al. [57] considered the single change-point case ($K \leq 1$), where the purpose is to detect a potential change-point τ. Similarly to the univariate case, let $r = \min(\tau, n - \tau)$, $D = \theta_\tau - \theta_{\tau - 1}$ and $s = \|D\|_0$. The authors established that the change-point τ can be detected with high probability as soon as $r\|D\|_2^2$ is larger than $C \min(\sqrt{p \log\log(8n)}, s \log(\frac{p}{s^2} \log\log(8n)))$, where C is a constant that only depends on the desired probability of detection. In particular, when $p = 1$, we recover the result of Gao et al. [38] mentioned earlier in the univariate case.

In the case of multiple change-points ($K \geq 1$), Wang and Samworth [103] introduced a method based on sparse projections. However, their procedure provably detects the change-points only under a strong condition on $r_k \|D_k\|_2^2$. More recently, Hu et al. [47] relaxed this condition within a specific asymptotic framework, using an approach based on a sparse likelihood score. Nevertheless, the condition of detection proven in [47] is not optimal, at least up to a logarithmic factor.

In addition to multivariate mean change-point detection, the broader topic of change-point detection in high-dimensional time series also encompasses many other problems. These include problems such as covariance change-point detection [96] or network change-point detection [97]. In each case, the data consist of a high-dimensional sequence, where the underlying signal is piecewise constant with a specific structure.

2.3.3 Overview of our contribution

In Chapter 5, we establish minimal conditions of detection of change-points in several problems, including covariance change-point detection, non-parametric change-point detection and sparse multivariate mean change-point detection. Subsequently, we focus on our contribution to mean change-point detection in multivariate time series, and we provide a summary of our work on this problem. We start by describing the minimal condition of detection before discussing the ideas of our bottom-up approach that achieves minimax optimal guarantees.

As mentioned earlier, the minimal condition that the change-points have to satisfy to be detected depends on $r_k \|D_k\|_2^2$. Intuitively, if $r_k \|D_k\|_2^2$ is very large, then τ_k can easily be detected. On the other hand, the detection of τ_k becomes impossible when $\|D_k\|_2$ approaches 0. In Chapter 5, we show that the minimal condition of detection of all the change-points τ_k is given by

$$r_k \|D_k\|_2^2 \geq C \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k} \sqrt{\log\left(\frac{n}{r_k}\right)}\right) + \log\left(\frac{n}{r_k}\right) \right] \ , \tag{2.21}$$

for all $k = 1, \dots, K$. Here, C is a constant that only depends on the desired probability of detecting all the change-points. More precisely, if all the change-points satisfy (2.21), then there exists an estimator $(\hat{\tau}_k)$ of (τ_k) that satisfies with high probability:

1. We recover the true number of change-points, that is $\hat{K} = K$
2. The estimated change-point $\hat{\tau}_k$ is not too far from the true change-point τ_k, in the sense that $\hat{\tau}_k \in [\tau_k - r_k/2, \tau_k + r_k/2]$

The above two properties are arguably the minimal desirable guarantees could be expected from an estimator of change-points, when they are all assumed to be detectable. Our analysis made in Chapter 5 takes into account change-points that possibly do not meet condition (2.21). Informally, the estimator is required to detect the change-points that satisfy (2.21), and not to detect any change-point twice. Additionally, the intervals of detection provided in Chapter 5 are more precise than $[\tau_k - r_k/2, \tau_k + r_k/2]$.

Interestingly, the condition of detection (2.21) can be approximated to $r_k \|D_k\|_2^2 \gg s_k \wedge \sqrt{p}$, up to the log factors. Notably, the sparsity s_k of D_k makes the detection problem substantially easier when $s_k \leq \sqrt{p}$.

To detect the change-points when they satisfy condition (2.21), we use a bottom-up approach in Chapter 5. Informally, our method is based on the aggregation of several tests, that are performed at different location and at different scale. First, we start by trying to detect change-points on the intervals of the type $[l-1, l]$, based on the observation $y_l - y_{l-1} \in \mathbb{R}^p$. Then, for every potential change-point detected, we remove a small neighborhood around it. Next, we try a greater scale $r = 2$, and we perform local tests on the remaining intervals of the form $[l - r, l + r)$. We continue this process for increasing scales r, until we reach a maximum scale r of order $n/2$. In the end, the estimator corresponds to positions within the intervals that have been removed through the process.

In Chapter 5, the analysis of this estimator involves a union bound on events controlling the local tests on the intervals $[l-r, l+r)$. For each of these intervals, we control the corresponding high-dimensional CUSUM statistic with techniques that are similar to those used in the detection of a sparse signal – see e.g. [30]. Ultimately, this ensures with high probability that we detect the change-points τ_k that satisfy condition (2.21), at scales r smaller than $r_k/2$.

To conclude, the change-points of a high-dimensional time series can be detected under the minimal condition (2.21), which adapts to the sparsity s_k of the change-points. Moreover, the bottom-up approach described in Chapter 5 is able to accurately detect with high probability all the change-points that satisfy this condition.

Chapter 3

Ranking a permuted matrix under the bi-isotonic-1D model

Motivated by crowdsourcing applications, we consider a model where we have partial observations from a bivariate isotonic $n \times d$ matrix with an unknown permutation π^* acting on its rows. Focusing on the twin problems of recovering the permutation π^* and estimating the unknown matrix, we introduce a polynomial-time procedure achieving the minimax risk for these two problems, this for all possible values of n, d, and all possible sampling efforts. Along the way, we establish that, in some regimes, recovering the unknown permutation π^* is considerably simpler than estimating the matrix.

This chapter is based on [74].

3.1 Introduction

We consider a crowdsourcing problem with n experts and d questions. For an unknown matrix M, $M_{i,j} \in [0,1]$ stands for the ability of expert i at question j. For the purpose of calibrating the model, we receive partial and noisy observations of the matrix M and our goal is to rank the experts according to their ability. Earlier models in crowd-labelling problems or in the related problems of pairwise comparisons typically assumed that the matrix M belongs to a parametric model [11, 58, 87, 27, 12], a prominent example being Bradley-Luce-Terry model. While there has been significant progress in this direction, such models do not tend to fit well real-world data [63, 4].

To address this issue, there has been a recent interest in the class of permutation-based models [16, 81, 83, 60, 19, 33, 72, 85] where it is only assumed that the matrix M satisfies some shape-constrained conditions before one (or two) permutations acts on the rows (and possibly on the columns) of M. Quite surprisingly, it has been established in [83] that, at least in some settings, the matrix M can be estimated at the same rate in those non-parametric models as in classical parametric models by relying on the least-square estimator on the class of permuted bi-isotonic matrices. Unfortunately, the corresponding class of matrices is highly non-convex and no polynomial-time algorithm is known for computing this least-square estimator. Furthermore, known computationally efficient procedures such as spectral estimators [16, 18] only achieve suboptimal convergence rates. This has led several authors to conjecture the existence of computational-statistical trade-offs [33, 84]. Despite recent progress in this direction [60, 56], the fundamental limits of polynomial-time algorithms for this class of problems remain largely unknown.

Arguably, for most applications, the primary objective is to recover the underlying permutation π^* acting on the rows or equivalently to rank the experts accordingly. While estimation of the full matrix M is closely related to ranking, it is also of a quite different nature as argued below. In this work, we investigate the estimation of the permutation π^* by characterizing the minimax risk for estimating π^* in a permuted shape-constrained model, introducing a polynomial-time procedure nearly achieving this risk bound. As a byproduct, we also disprove the existence of a computational-statistical gap for the reconstruction of the matrix M.

3.1.1 Problem formulation

A bounded matrix $B \in [0,1]^{n \times d}$ is said to be bi-isotonic if it satisfies $B_{i,j} \le B_{i+1,j}$ and $B_{i,j} \le B_{i,j+1}$ for any $i \in [n-1]$ and $j \in [d-1]$. Henceforth, we write $\mathbb{C}_{\mathrm{BISO}}$ for the collection of such $n \times d$ bounded bi-isotonic matrices.

In this work, we assume that the matrix M is a row-permuted bi-isotonic matrix as in [60, 56]. In other words, up to a single permutation π^* of $[n]$, the matrix $M_{\pi^{*-1}}$ defined by $(M_{\pi^{*-1}})_{i,j} = (M_{\pi^{*-1}(i),j})$ is bi-isotonic.

From a modeling viewpoint, this amounts to assuming that the d questions are ordered from the most difficult question to the most simple question. The permutation π^* is not necessarily unique, but the corresponding permuted matrix $M_{\pi^{*-1}}$ is unique. Despite that, we refer, with a slight abuse of terminology, to π^* as the oracle permutation. With this definition, $\pi^{*-1}(i)$ corresponds to any i-th smallest row (or equivalently expert to use the crowdsourcing terminology) in the matrix M. In the following, the i^{th} row of M is referred to as expert i, whereas the k^{th} column is referred to as question k.

We consider an observation-scheme where the statistician has partial access to noisy observations Y of M such that

$$Y = M + E \ , \tag{3.1}$$

where the entries of E are centered, independent, subGaussian - see definition 2.2 of [94] - with Orlicz norm at most ζ, but are not necessarily identically distributed. In particular, this model encompasses binary observations $Y_{i,k} \sim Ber(M_{i,k})$ which arise in crowd-labelling problems, in which case we have $\zeta = 1$. In the following, we refer to ζ as the noise level.

As usual in the literature –e.g. [60], we use the Poissonization trick to model the partial observations. Given some $\lambda > 0$, which is henceforth referred as the sampling effort, we have $N = Poi(\lambda nd)$ observations of the form

$$(x_t, y_t), \quad t =, 1 \dots, N, \tag{3.2}$$

where the position x_t is sampled uniformly in $[n] \times [d]$, and $y_t = M_{x_t} + E_{x_t}$ is an independent observation of matrix Y of (3.1) at position x_t. Conditionally to N, this scheme is equivalent to a uniform sampling scheme with replacement [61]. If $\lambda < 1$, then a specific entry of Y is sampled at least once with probability $1 - e^{-\lambda}$ which is close to λ. More generally, λ corresponds to the expected number of times a specific entry of Y is observed, so that $\lambda > 1$ would correspond to the situation where entries are sampled multiple times.

Since our aim is to recover the permutation π^* from the partial observations (x_t, y_t), we consider, for some estimator $\hat{\pi}$, the following error metric

$$l(\hat{\pi}; \pi^*) = \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \ , \tag{3.3}$$

where $\|.\|_F$ stands for the Frobenius norm. This loss quantifies the distance between the matrix M ordered according to the oracle permutation π^* and the matrix M ordered according to the estimated permutation. When π^* is not unique, the error $l(\pi^*, \pi^{*'})$ between any two oracle permutations is zero. If $\hat{\pi}$ and π^* only differ by a transposition or equivalently if the ranking $\hat{\pi}$ and π^* only differ on two experts, then $l(\hat{\pi}; \pi^*)$ is twice the square Euclidean distance between the corresponding rows of M. More generally, $l(\hat{\pi}; \pi^*)$ interprets as the sum over all $i = 1, \dots, n$ of the square Euclidean distance between the i-th smallest row of M according to $\hat{\pi}$ and according to the oracle ranking π^*. In constrast to other metrics between permutations such as the Kendall tau distance, – see e.g. [12, 61, 13] – the loss function l depends on M. Our choice of this loss function l results in a higher penalty for inverting a pair of rows that are distant in L_2 norm compared to inverting another pair of rows that are closer in L_2 norm.

The loss (3.3) is ubiquitous when one aims at estimating the matrix M in Frobenius norm, that is building an estimator $\widehat{M}$ such that $\|\widehat{M} - M\|_F^2$ is as small as possible –see e.g. [83, 60, 56]. Indeed, estimating π^* by $\hat{\pi}$ is a first step towards building an estimator of M by doing as if $M_{\hat{\pi}^{-1}}$ was bi-isotonic. It turns out that the error in $\|\widehat{M} - M\|_F^2$ decomposes as the sum of two terms, one of them being $l(\hat{\pi}, \pi^*)$ while the other one does not really depend on $\hat{\pi}$. Conversely, an estimator $\widehat{M}$ can be easily transformed into an estimator $\hat{\pi}$ whose loss $l(\hat{\pi}, \pi^*)$ is controlled by $\|\widehat{M} - M\|_F^2$. See [83, 60] for further discussions. In summary, controlling $l(\hat{\pi}; \pi^*)$ is important in order to evaluate to what extent π^* is well estimated, but it is also the key stepping stone towards a good estimation of the matrix M.

In some works, the authors directly consider distances on the symmetric group of permutations. Examples include the Kendall tau distance $d_{KT}(\pi, \pi') = \sum_{(i,j):\pi(i)<\pi(j)} \mathbf{1}\{\pi'(i) > \pi'(j)\}$ or the l_∞ distance $\|\pi - \pi'\|_\infty = \max_{i\in[n]} |\pi(i) - \pi'(i)|$ –see [12, 61] in the noisy sorting model. However, those distances are not well suited to handle the non-parametric class of bi-isotonic matrices, because to control them we would need to make assumptions on the separation between the rows of the matrix M –see Appendix A of [83].

Equipped with this notation, we consider the minimax risk of permutation recovery as a function of the number n of experts, the number d of question, the sampling effort λ, and the noise level ζ.

$$\mathcal{R}^*[n, d, \lambda, \zeta] = \inf_{\hat{\pi}} \sup_{\pi^* \in \Pi_n} \sup_{M:\, M_{\pi^{*-1}} \in \mathbb{C}_{\text{BISO}}} \mathbb{E}\left[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2\right] \ , \tag{3.4}$$

where Π_n stands for the collection of all permutations of $[n]$. In particular, our general aim is to tightly control this minimax risk and, if possible, to provide a computationally efficient procedure achieving this minimax risk. Although our primary interest lies in the estimation of π^*, we also consider the minimax estimation risk of M

$$\mathcal{R}^*_{est}[n, d, \lambda, \zeta] = \inf_{\hat{M}} \sup_{\pi^* \in \Pi_n} \sup_{M:\, M_{\pi^{*-1}} \in \mathbb{C}_{\text{BISO}}} \mathbb{E}\left[\|\hat{M} - M\|_F^2\right] \ . \tag{3.5}$$

as studied in [83, 56, 60, 71] in order to assess the performances of our computationally efficient procedures.

3.1.2 Related work and open questions

The most relevant body of work to the current chapter is that on estimating square matrices M satisfying the so-called strong stochastic transitivity class (SST) [16, 83]. A matrix M belongs to the SST class if (i) M is skew-symmetric that is $M + M^T = ee^T$ where e is the constant vector of size n and (ii) there exists a common permutation π^* of $[n]$ such that row and column-permuted matrix $M_{\pi^{*-1}\pi^{*-1}}$ is bi-isotonic. This class is suited for considering pairwise comparisons problems. Shah et al. [83] consider the full observation setting, namely a setting where each entry of the matrix M is observed once in noise - which is in some sense akin to $\lambda = 1$ in our Poissonian scheme[1]. They proved that the minimax risk for estimating M in square Frobenius distance is, up to logarithmic terms, of the order of n and is achieved by the corresponding least-square estimator over the SST class. Unfortunately, this estimator cannot be efficiently computed. They also analyzed an efficient spectral estimator achieving the rate $n^{3/2}$. This rate is also achieved [83] by the near-linear time Borda count algorithm CRL that simply ranks the individuals according to the row sums of the observations and then plugs the corresponding permutation to estimate M. See also [19] for related results. This led some authors [33, 84] to conjecture the existence of a $\sqrt{n}$ computational gap for SST matrices and for other shape-constrained matrices with unknown permutation.

In crowdsourcing problems where $M \in [0,1]^{n\times d}$, non-parametric models [60] assume that the matrix M is bi-isotonic up to a permutation π^* of the rows (experts) - and sometimes also up to a permutation τ^* of the columns (questions)[2]. In this chapter as in this literature review, we focus however solely on the case where M is bi-isotonic up to a permutation π^* of the rows (experts). Mao et al. [60] have established the minimax risk $\mathcal{R}^*_{est}[n,d,\lambda,1]$ for estimating M in the specific case where $n \geq d$. In the arguably most interesting regime of partial observations $\lambda \leq 1$, they prove that this minimax risk is of the order of $n/\lambda \wedge (nd)$. This rate is achieved by the inefficient least-square estimator. Furthermore, Mao et al. [60] were the first to narrow the conjectured computational gap by introducing a new efficient procedure called one-dimensional sorting. In the square case $n = d$ with full observations, these procedures achieve (up to log terms) the rate $n^{5/4}$ for estimating the matrix M, thereby improving over the previous $n^{3/2}$ barrier.

Recently, this rate was improved by Liu and Moitra [56] in a specific instance of the problem where $n = d$ and one has access to a sub-polynomial number of noisy independent samples of the complete matrix M from (3.1) – which is akin to our Poissonian scheme for λ being sub-polynomial in n, d. They introduce a polynomial-time procedure achieving the rate $n^{1+o(1)}$ for permutation recovery and matrix estimation which, up to the factor $n^{o(1)}$, turns out to be minimax optimal for both problems. As a consequence, in this very specific instance, the computational gap turns out to be nonexistent.

There remain important open problems to characterize the estimation of π^* and M in crowdsourcing problems.

- Beyond the case $n \geq d$ handled by Mao et al. [60], the minimax risk $\mathcal{R}^*[n,d,\lambda,1]$ of estimation of the permutation π^* - as well as the minimax risk $\mathcal{R}^*_{est}[n,d,\lambda,1]$ of estimation of the matrix M - are unknown. In particular, in the rectangular case where $n \ll d$, the number of questions exceeds the number of experts is both relevant from a practical [85] and a conceptual perspective. Indeed, the analysis of the least-square estimator of [60] and related works is based on entropy calculation of the class of permuted bi-isotonic matrices. While the minimax risk $\mathcal{R}^*_{est}[n,d,\lambda,1]$ turns out to be (up to logarithm terms), characterized by this entropy, this is not always the case for the estimation of π^* as many matrices M share the same permutation π^*. As a consequence, even if we leave aside computational constraints, pinpointing the optimal risk $\mathcal{R}^*[n,d,\lambda,1]$ for estimating π^* requires quite different arguments.
- Beyond the toy "over-complete" observation model in the square case $n = d$ of Liu and Moitra [56], where each entry is sampled at least $n^{o(1)}$ times, it remains unclear whether there is a computational gap for general rectangular settings with partial observations.

.

3.1.3 Our contributions

Echoing with these open problems, we make the following contributions in this work:

- First, we characterize (up to polylogarithmic multiplicative terms) the minimax risk $\mathcal{R}^*[n,d,\lambda,\zeta]$ of permutation recovery, this, for all possible number of experts $n \geq 1$, number of questions $d \geq 1$, noise level

[1] In the Poissonian scheme, each entry is observed at least once with probability $1 - e^{-\lambda}$.
[2] This would correspond to the situation where the corresponding ordering of the questions is also unknown.

	$n \le d^{1/3}$	$d^{1/3} \le n \le d$	$n \ge d$
Permutation estimation: $\mathcal{R}^*[n,d,1,1]$	$nd^{1/6}$	$n^{3/4}d^{1/4}$	n
Matrix estimation: $\mathcal{R}^*_{est}[n,d,1,1]$	$nd^{1/3}$	$\sqrt{nd}$	n

Figure 3.1: Summary of the minimax risks (up to poly-logarithmic terms) for permutation estimation ($\mathcal{R}^*[n,d,1,1]$) and matrix estimation ($\mathcal{R}^*_{est}[n,d,1,1]$) in the specific cases where $\lambda, \zeta = 1$.

$\zeta \ge 0$, and almost all sampling efforts $\lambda > 0$. When $n \ll d$, we prove in particular that $\mathcal{R}^*[n,d,\lambda,1] \ll \mathcal{R}^*_{est}[n,d,\lambda,1]$ in all non-trivial regimes, highlighting that when $n \ll d$, the problem of permutation recovery is statistically easier than the problem of matrix estimation.

- Moreover, we introduce a polynomial-time procedure achieving this risk bound, thereby establishing that there does not exist any significant computational-statistical trade-off for the problem of recovering a single permutation π^*. While our procedure borrows some of the ingredients of Liu and Moitra [56], we need to introduce several new ideas to deal with the significantly more involved case $n \ll d$. Since an estimator $\hat{\pi}$ of π^* can be easily combined with a least-square estimator of a bi-isotonic matrix to estimate the matrix M –see e.g. [83, 60] – we also deduce a polynomial time estimator $\widehat{M}$ which nearly achieves the minimax estimation risk $\mathcal{R}^*_{est}[n,d,\lambda,1]$, thereby proving that this problem does not either exhibit any computational-statistical trade-off, thereby answering the open problem of [60].

To provide a glimpse of our results, let us describe the minimax risks on the arguably most interesting case where the noise level ζ is of order 1 as in the Bernoulli observation setting and where $\lambda < 1$ which corresponds to a partially observed matrix. In Section 3.4, we establish that the minimax risk $\mathcal{R}^*[n,d,\lambda,1]$ of permutation recovery is (up to polylogarithmic multiplicative terms) of the order of

$$\left[\frac{nd^{1/6}}{\lambda^{5/6}} \bigwedge \frac{n^{3/4}d^{1/4}}{\lambda^{3/4}} + \frac{n}{\lambda}\right] \bigwedge nd, \tag{3.6}$$

whereas the minimax reconstruction risk $\mathcal{R}^*_{est}[n,d,\lambda,1]$ is of the order

$$\left[\sqrt{\frac{nd}{\lambda}} \bigwedge \frac{nd}{\lambda^{2/3}(n \vee d)^{2/3}} + \frac{n}{\lambda}\right] \bigwedge nd. \tag{3.7}$$

We display in Figure 3.1 a summary of our results in the specific case where we also have $\lambda = 1$ on top of $\zeta = 1$, and will discuss this case more in details, as it highlights one of our main findings.

A first comment is that the minimax risk of matrix estimation $\mathcal{R}^*_{est}[n,d,1,1]$ can be interpreted through the covering numbers of the space of permuted bi-isotonic matrices as in [60]. For $n \ge d$ both minimax risks - $\mathcal{R}^*[n,d,1,1]$, $\mathcal{R}^*_{est}[n,d,1,1]$ - are of the order of n so that recovering the permutation π^* is as hard as estimating the matrix M (up to logarithmic factors). This is the regime studied in the literature, see [60, 56]. When the number d of questions is large - $n \ll d$ - then the regimes are more tricky. There are two of them, depending on whether n is larger than $d^{1/3}$ or not, and in both regimes $\mathcal{R}^*_{est}[n,d,1,1]$ is significantly larger than $\mathcal{R}^*[n,d,1,1]$. More regimes appear when we do not restrict ourselves to $\lambda = 1$, $\zeta = 1$. This complex picture, as well as the fact that $\mathcal{R}^*[n,d,1,1] \ll \mathcal{R}^*_{est}[n,d,1,1]$ for $n \ll d$ - and also in many other configurations of λ, ζ - highlights the fact that the difficulty of estimating π^* is not governed by the size of the space of permuted bi-isotonic matrices. As a consequence, even if we leave computational aspects aside, it is not clear that the least-square estimator of [60] achieves optimal risk for estimating the permutation π^* and, in any case, entropy-based arguments would lead to suboptimal bounds, at least if we use the same arguments as in [60].

As a byproduct of our results, we also establish the minimax risk - and prove that it is achievable in polynomial time - for another loss function termed $l_\infty(\hat{\pi}, \pi^*)$ (see (3.30)) put forward in [19, 84, 60] - and we also disprove a conjecture regarding a computational-statistical gap for this loss. See Subsection 3.4.4.

As our minimax results remain valid in the noiseless case ($\zeta = 0$) where one has access to partial observation of the matrix M itself, we are able to tightly decipher the approximation error which is due to the partial sampling of the matrix M from the stochastic error stemming from noisy observations. In some way, this complements the works of Pananjady et al. [71] on the effect of the design in the specific case where the sampled entries are sampled uniformly.

3.1.4 Proof techniques and further comparison with the literature

In order to build a polynomial-time procedure nearly achieving the minimax permutation risk in the partial observation setting (3.2), we first consider the so-called full observation setting where one has access to poly-

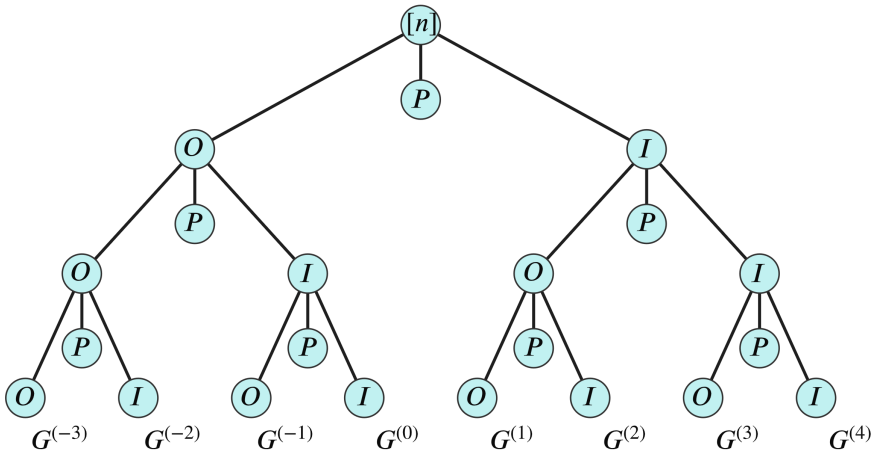


Figure 3.2: Example of a hierarchical sorting tree.

logarithmic number Υ of samples $Y^{(0)}, Y^{(1)}, \dots, Y^{(\Upsilon)}$ of the complete $n \times d$ matrix. This setting is akin to that of Liu and Moitra [56] when they handled the specific square case $n = d$ with noise level $\zeta = 1$.

For this reason, our estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$ introduced in Section 3.3 share some features with the procedure of [56]. From a broad perspective, our procedure and theirs build a hierarchical sorting tree using a top-down approach as depicted in Figure 3.2. We start from the complete set $[n]$ of all experts and build a trisection (O, P, I) of $[n]$, where O (resp. I) contains experts that provably are below (resp. above) the median expert, whereas P contains all the experts for which we cannot certify with high confidence whether they are above or below the median. Then, we recursively trisect the sets O and I as depicted in Figure 3.2. At the end of the algorithm, we obtain a partial ordering on all the experts which can be used to estimate the oracle permutation π^*.

Then, the problem of building a suitable estimator boils down to introducing a suitable trisection procedure. We could naively do this by comparing the row-sums of the observed matrix Y which amounts to comparing the mean ability of each expert, but this is well known to lead to suboptimal performances by a factor $\sqrt{d}$ –see e.g. [60]. To improve over this rate, we need to compare the experts according to convex combinations of suitable questions. As in [56], we start by selecting suitable blocks of questions by detecting the high-variation regions of the mean empirical expert and combine them with spectral algorithms to select suitable convex combinations of questions. Still, we have to refine significantly their spectral procedure to handle the rectangular case $n \ll d$. Equipped with these refinements, which are involved technically, but are built on the ideas developed in [56], we arrive at the estimator $\hat{\pi}_{HT}$ (see Section 3.3) that turns out to be minimax optimal in some regimes of (n, d, ζ).

Unfortunately, this method turns out to be suboptimal in many regimes, for instance for mild values of $n \in [d^{1/3}, d]$. Informally, this is due to the fact that our first estimator $\hat{\pi}_{HT}$ as well as that of Liu and Moitra [56] build an oblivious hierarchical sorting tree. This means that the trisection method decomposes a group $G^{(0)}$ of experts in the hierarchical sorting tree in (O, P, I) only using the experts in $G^{(0)}$ of the matrix Y. In the related problem of hierarchical clustering, most top-down procedures also share this feature. It turns out that the observations of other experts can help improving the trisection of $G^{(0)}$. In particular, sets of experts that are close in the ordering –such as $G^{(1)}$ and $G^{(-1)}$ in Figure 3.2– are sometimes valuable to improve the selection of a suitable convex combination of questions. We emphasize this phenomenon and provide more intuition on it in Section 3.3, when we introduce a new estimator $\hat{\pi}_{WM}$ that builds upon the memory of the sorting tree. This new procedure $\hat{\pi}_{WM}$ turns out to be near minimax optimal for all values of (n, d, ζ).

Coming back to the partial observation setting (3.2), we introduce in Section 3.4 a reduction scheme which boils down to reducing the number of questions in order to come back to a full observation model for a sub-matrix of size $n \times d_-$ where d_- is possibly much smaller than d. Then, relying on the full observation setting described above, we estimate the permutation π^* based on the corresponding reduced matrix. In comparison to the full observation model, we can suffer from an additional bias term which arises in the reduction process. To handle this, we develop a slight variant $\hat{\pi}_{WM-SR}$ of $\hat{\pi}_{WM}$ –see 3.5.7 for details. The resulting procedure turns out to nearly achieve minimax permutation recovery risk for all values (n, d, ζ) and all values of λ. Plugging our procedure to estimating the matrix M, we close all the computational gaps pointed out in Mao et al. [60]

for the problem of matrix estimation with a single unknown permutation - see Subsection 3.4.3.

3.1.5 Notation and organization of the chapter

In the following, c, c_1, ... stand for numerical positive constants that may change from line to line. Given a vector u and $p \in [1,\infty]$, we write $\|u\|_p$ for its l_p norm. For a matrix A, $\|A\|_F$ and $\|A\|_{\mathrm{op}}$ stand for its Frobenius and its operator norm. We write $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) for the largest (resp. smallest) integer smaller than (resp. larger than) or equal to x.

Although M stands for an $n \times d$ matrix, we extend it sometimes in an infinite matrix by setting $M_{i,k} = 0$ when either $i \le 0$ or $k \le 0$ and $M_{i,k} = 1$ when either $i \ge n+1$ and $k > 0$ or $k \ge d+1$ and $i > 0$. The corresponding infinite matrix $M_{\pi^{*-1}}$ which is obtained by permuting the n original rows is still bi-isotonic and takes values in $[0,1]$. We shall often work with sub-matrices of M that are restricted to a subset $P \subset [n]$ and $Q \subset [d]$ of rows and columns, in which case we write that the corresponding matrix M' belongs to $\mathbb{R}^{P\times Q}$. More precisely, M' is such that, $M'_{i,j} = M_{i,j}$ for any $i \in P$ and any $j \in Q$.

In the following, we write that two sequences or functions u and v satisfy $u \lesssim v$, if there exists a universal constant such that $u \le cv$.

In Section 3.2, we first consider the complete observation problem, where one has access to a poly-logarithmic number of independent samples of the complete noisy matrix Y. We characterize the minimax risk for permutation recovery and prove that it is achieved by a polynomial-time procedure. In section 3.3, we describe the corresponding polynomial-time procedure. In Section 3.4, we deal with the problem of partially observed matrix in the model (3.2).

3.2 Analysis of the full observation problem

As explained in the introduction, and following [56], we first consider a slightly different problem where we fully observe a Υ-sample $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(\Upsilon-1)})$ of the noisy matrix according to the model $Y = M + E$ in (3.1). Here, Υ should be considered as a polylogarithms in n and d. This is of course not very realistic in applications, but it is simpler to first present our algorithmic procedure in this setting, and it also enables more direct comparison to [56]. We will explain later in Section 3.4, how one can transform data in the more realistic partial observation scheme from (3.2) to this full observation scheme. We will then prove that the algorithm applied to the transformed data is near minimax optimal.

We recall that M is a bi-isotonic matrix, up to an unknown permutation π^* of its rows. Besides, the noise matrix E is made of independent mean zero subGaussian entries, with Orlicz norm less than or equal to ζ.

3.2.1 Minimax lower bounds

Before considering ranking procedures, we characterize the minimax risk for the problem of ranking with full observations. For the purpose of the minimax lower bound, we assume that the noise matrix E in (3.1) is made of independent normal random variables with variance ζ^2. For a permutation π^* and a matrix M such that $M_{\pi^*} \in \mathbb{C}_{\mathrm{BISO}}$, we respectively denote $\mathbb{P}_{(\pi^*,M)}$ and $\mathbb{E}_{(\pi^*,M)}$ the corresponding probability and expectations with respect to the Υ independent observations of Y. Define

$$\mathcal{R}_F(n,d,\zeta) = \zeta^2 \left[\frac{nd^{1/6}}{\zeta^{1/3}} \wedge \frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}} + n \right] . \tag{3.8}$$

The following minimax lower bound is stated in a setting where one has access to a polylogarithmic number Υ of full samples to be in line with the analysis of the next subsection. Still, we can forget about the dependency in Υ at first reading.

Theorem 3.2.1. *There exists a universal constant c such that the following holds for any $n \ge 2$, $d \ge 1$, $\zeta > 0$, and $\kappa > 2$. Provided that the sample size Υ is less than or equal to $\log^{\kappa}(2nd/\zeta)$, we have*

$$\inf_{\hat\pi} \sup_{\pi^* \in \Pi_n} \sup_{M : M_{\pi^{*-1}} \in \mathbb{C}_{BISO}} \mathbb{E}_{(\pi^*,M)} \|M_{\hat\pi^{-1}} - M_{\pi^{*-1}}\|_F^2 \ge c \left[\log^{-\kappa}(nd/\zeta) \mathcal{R}_F[n,d,\zeta] \bigwedge nd \right] . \tag{3.9}$$

In fact, this theorem turns out to be a consequence of the minimax lower bound in the partial observation scheme –see Section 3.4. Together with the risk upper bounds of the next section, (3.9) characterizes, up to polylogarithmic terms, the minimax risk for estimating π^*. The term nd in (3.9) is related to the fact that the loss $\|M_{\hat\pi^{-1}} - M_{\pi^{*-1}}\|_F^2$ cannot be larger than nd because the entries of M are in $[0,1]$.

Mild noise level. The risk bound $\mathcal{R}_F(n,d,\zeta)$ involves five different terms, some of them being significant only when ζ is small in comparison to n and d. As these regimes with very small ζ are arguably quite specific, and to simplify the discussion, we will now detail the minimax lower bound in the specific case when $\zeta = 1$.

$$\overline{\mathcal{R}}_F(n,d) = \mathcal{R}_F(n,d,1) = (nd^{1/6}) \bigwedge (n^{3/4}d^{1/4}) + n \ . \tag{3.10}$$

In particular, we recognize three main regimes in (3.10) that depend on n and d. When the number of experts is relatively small ($n \leq d^{1/3}$), the risk is proportional to $nd^{1/6}$. Specifying the result to $n = 2$, one checks that a square distance $d^{1/6}$ is necessary to distinguish two experts. As a consequence, a suitable estimator $\hat{\pi}$ should be able to coherently rank experts that are distant by more than $d^{1/6}$ in squared Frobenius norm, and then to achieve a risk smaller than $nd^{1/6}$. For larger $n \geq d^{1/3}$, it is in fact possible to build upon the large number of experts to improve the comparisons between experts using in particular spectral methods. For this reason, the optimal risk is proportional to n for $n \geq d$. For an intermediary number of experts $n \in [d^{1/3}, d]$, the risk is of the order of $n^{3/4}d^{1/4}$. Our main contribution is the construction of a polynomial-time procedure that achieves these risk bounds, see below.

Low noise level. For mild values of ζ, the minimax risk $\mathcal{R}_F(n,d,\zeta)$ has the same form as $\overline{\mathcal{R}}_F(n,d)$, up to some factors that depend on ζ. However, for very small ζ, the risk becomes qualitatively different. For example, we have $\mathcal{R}_F(n,d,\zeta) \asymp \zeta^2 n\sqrt{d}$ when $\zeta \in (0, \frac{1}{n \vee d}]$. In fact, this rate is quite easy to achieve by a polynomial time algorithm in this extreme case. It is proven in various works – see e.g. [83] that ranking the experts according to the row sum of the matrix correctly compares two experts as long as their square distance is at least $\zeta^2\sqrt{d}$ (up to logarithmic terms). As a consequence, this simple procedure leads to an error $\zeta^2 n\sqrt{d}$. While $\zeta^2 n\sqrt{d}$ is highly suboptimal in most realistic regimes, it turns out to be tight for extremely low level of noise. Finally, the intermediary rate $\mathcal{R}_F(n,d,\zeta) \asymp \zeta^{5/3}n^{2/3}\sqrt{d}$ is achieved for slightly larger values of ζ, but it is less clear how to interpret it.

3.2.2 Minimax upper bounds

In the following, we fix a parameter $\delta \in (0,1)$ that will correspond to a small probability. We write $\zeta_- = \zeta \wedge 1$, where ζ is the noise level. In this section, we analyze two estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$ of π^* that are described in Section 3.3 and more formally defined in Section 3.4.5. The first estimator $\hat{\pi}_{HT}$ is based on the construction of an oblivious hierarchical sorting tree. We will later explain all the ingredients of this procedure. In contrast, the second estimator $\hat{\pi}_{WM}$ relies on the construction of a hierarchical sorting tree with memory. Both procedures have a computational complexity of the order of $\log^c(\frac{nd}{\zeta_-\delta})(n^3 + nd^2)$, for some $c > 0$, which makes them polynomial time - unlike the least square procedure e.g. from [60].

Theorem 3.2.2. *There exist three numerical constants c, c', and c_0 such that the following holds. Fix $\delta > 0$ and assume that $\Upsilon \geq c_0 \log^8(nd/\delta)$. For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{BISO}$, the oblivious hierarchical sorting tree estimator $\hat{\pi}_{HT}$ defined in the next section satisfies*

$$\|M_{\hat{\pi}_{HT}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq c\zeta^2 \log^{10.5}\left(\frac{2nd}{\delta\zeta_-}\right)\left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] \ ,$$

with probability at least $1 - c'n\log^9(\frac{nd}{\delta\zeta_-})\delta$.

If we take $\delta = \zeta^2(nd)^{-1}$ in the above expression, we easily deduce - reminding that the entries of M are in $[0,1]$ - the following risk bound

$$\mathbb{E}\left[\|M_{\hat{\pi}_{HT}^{-1}} - M_{\pi^{*-1}}\|_F^2\right] \leq c\zeta^2 \log^{10.5}\left(\frac{2nd}{\zeta_-}\right)\left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] \ .$$

Comparing this bound with (3.10) in the specific case where $\zeta = 1$, we observe that $\hat{\pi}_{HT}$ achieves the optimal risk $nd^{1/6}$ for small $n \leq d^{1/3}$ and the optimal risk n for large $n \geq d$. Unfortunately, for mild $n \in [d^{1/3}, d]$, the risk bound is of the order of $n^{2/3}d^{1/3}$, which is significantly higher than the minimax lower bound $n^{3/4}d^{1/4}$. To close this gap, we turn to the more refined estimator $\hat{\pi}_{WM}$.

Theorem 3.2.3. *There exist three numerical constants c, c', and c_0 such that the following holds. Fix $\delta > 0$ and assume that $\Upsilon \geq c_0 \log^8(nd/(\delta\zeta_-))$. For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{BISO}$, the hierarchical sorting tree estimator with memory $\hat{\pi}_{WM}$ satisfies*

$$\|M_{\hat{\pi}_{\mathrm{WM}}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq \left[c\log^{11}\left(\frac{2nd}{\delta\zeta_-}\right)\mathcal{R}_F[n,d,\zeta]\right] \bigwedge nd \ , \tag{3.11}$$

with probability at least $1 - c'n\log^9(\frac{nd}{\delta\zeta_-})\delta$.

As for the previous theorem, this high probability result can be turned into a risk bound by taking $\delta = \zeta^2/(nd)$. In particular, this risk bound matches, up to polylogarithmic terms, the minimax lower bound (3.9) for all possible values of n, d, and ζ. As a consequence, the estimator $\hat{\pi}_{WM}$ is nearly minimax and this ranking problem does not exhibit any computational gap.

In [56], the polynomial-time estimator $\hat{\pi}_{LM}$ of Liu and Moitra achieves the minimax risk in the specific square where $n = d$ and $\zeta = 1$. In all the other regimes, no polynomial-time procedure was previously proved to achieve the minimax risk. In fact, even if we do not restrict our attention to polynomial-time procedures, least-square type procedures studied e.g. in [60] provably achieve the minimax risk only in the regime when $n \geq d$. As alluded in the introduction - see Equations (3.6) and (3.7), the minimax risks for estimating π^* and M differ when $n \leq d$, so that achieving the optimal risk for π^* is not possible using the classical entropy arguments as in [83, 60]. This highlights the fact that estimating the permutation π^* is significantly more challenging in the regime $n \leq d$ - both from a statistical and computational perspective - than in the regime $n \geq d$ handled in [56, 60].

Consequences for the estimation of the matrix M. Provided that we have estimated π^* with $\Upsilon - 1$ independent samples, we could use the last sample $Y^{(\Upsilon)}$ to estimate the matrix M by minimizing the least-square criterion $\widehat{B} = \arg\min_{B \in \mathbb{C}_{\mathrm{BISO}}} \|Y^{(\Upsilon)}_{\hat{\pi}_{WM}} - B\|_F^2$ and setting $\widehat{M} = \widehat{B}_{\hat{\pi}^{-1}_{WM}}$. Since the set of bi-isotonic matrices is convex, this estimator is computable efficiently [53]. As argued in Proposition 3.3 of [60] and often used in the ranking literature [84, 19, 72], it turns out that, with high probability, the reconstruction error $\|M - \widehat{M}\|_F^2$ is (up to polylogarithmic terms) the sum of the expected permutation loss $\mathbb{E}\left[\|M_{\hat{\pi}_{WM}} - M_{\pi^*}\|_F^2\right]$ and the minimax reconstruction risk of a bi-isotonic matrix $\inf_{\widehat{B}} \sup_{B \in \mathbb{C}_{\mathrm{BISO}}} \mathbb{E}[\|\widehat{B} - B\|_F^2]$ where $Y = B + E'$ and E' is made of independent subGaussian random variables. Hence, based on $\hat{\pi}_{WM}$ and Theorem 3.2.3, it is easy to construct a polynomial-time estimator of M that is also near minimax-optimal in the sense of Equation (3.5). We will further build upon this remark in Section 3.4 when we come back to the problem of partial observations of the matrix.

3.3 Description of the hierarchical sorting estimators

Let us now describe the construction of the estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$ of π^*. The construction is quite long and involves several subroutines. For this reason and to ease the understanding of proof details, we also provide a more formal and longer definition in Section 3.4.5. Afterwards, we comment on the different steps of the procedure and on their connection to the literature in Subsection 3.3.3.

Define $\tau_\infty = \lceil 4 \cdot 10^7 \log^7(\frac{nd}{\delta(\zeta_-)^2}) \rceil$ and $t_\infty = \lceil \log(n)/\log(2) \rceil$. We define $\Upsilon^* = 6\tau_\infty t_\infty$ for the total number of independent samples required for the computation of these two estimators.

Hence, we are given independent samples $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(\Upsilon^*-1)})$. From a broad perspective, both procedures are based on the construction of the recursive sorting tree as illustrated in Figure 3.2. Starting from the root of the tree which corresponds to the set $[n]$ of all experts, we build a partition O, P, I, of $[n]$ in such a way that, with high probability, all the experts in O are below the median expert of $[n]$, all the experts in I are above the median expert of $[n]$, while the remaining experts in P are those for which we are not able to decipher whether they are below or above the median expert of $[n]$.

Having trisected $[n]$, we recursively trisect the subsets O and I- see Figure 3.2. Each time, the size of the groups O and I is divided by at least 2. Hence, at depth t_∞, all the groups of O and I have size at most 1. For each depth $t = 0, \dots, t_\infty - 1$, we use $6\tau_\infty$ new samples. The construction of the tree is described in **TreeSort** –see Algorithm 1 and is based on the routine **BlockSort** which performs the trisection of a group into (O, P, I).

Let us now explain how to deduce an estimator $\hat{\pi}$ from the final hierarchical sorting tree $\mathcal{T}$. Indeed, the hierarchical sorting tree $\mathcal{T}$ induces an order on its leaves as follows. For any groups (O, P, I) sharing the same parent, we say that any descendent of O in the tree $\mathcal{T}$ is below P, which, in turn, is below any descendent of I in $\mathcal{T}$. This endows a complete ordering on the leaves of the tree $\mathcal{T}$. Denote $\mathcal{G} = (G_1, \dots, G_\alpha)$ the sequence of leaves of the final tree ranked according to this complete order. For any $a \in [\alpha]$, we define the lower bound $\pi^-_{\mathcal{G}}(G_a)$ and the upper bound $\pi^+_{\mathcal{G}}(G_a)$ of the ranks of experts in G_a by $\pi^-_{\mathcal{G}}(G_a) := \sum_{a' < a} |G_{a'}|$ and $\pi^+_{\mathcal{G}}(G_a) := \sum_{a' \leq a} |G_{a'}|$. Finally, we sample $\hat{\pi}$ arbitrarily in such a way that

$$\hat{\pi}(G_a) = [\pi^-_{\mathcal{G}}(a) + 1, \pi^+_{\mathcal{G}}(a)] \ . \tag{3.12}$$

In other words, the estimator $\hat{\pi}$ ranks the groups G_a according to the ordering of the groups endowed by $\mathcal{T}$ and, given that, ranks the experts G_a uniformly at random. See Section 3.4.5 for a more formal definition of the ordering.

3.3.1 Description of the trisection of a leaf G into (O, P, I) with BlockSort

The purpose of **BlockSort** is to build a trisection of a group G of experts into (O, P, I) where O is made of experts that are, with high probability, below the median expert in G and I is made of experts which are, with high probability, above this median expert. It turns out that this construction is based on τ_∞ iterations of a procedure called **DoubleTrisection** which is the backbone of our procedure. Intuitively, we shall iteratively detect subgroups of experts that are below (resp. above) the median expert of G which, after τ_∞ iterations, will allow us to obtain O and I.

For technical reasons, our definition is slightly more intricate. We shall simultaneously build two collections (O_τ, I_τ) and $(\overline{O}_\tau, \overline{I}_\tau)$ of groups, the second one being more conservative. We start with empty sets for $(O_0, I_0, \overline{O}_0, \overline{I}_0) = \varnothing$. Then, at each step τ, we will consider the remaining set of experts $G \setminus (\overline{O}_\tau \cup \overline{I}_\tau)$. Define $\gamma = \lfloor |G|/2 \rfloor - |\overline{O}_\tau|$ for the presumed rank of the median expert of G inside $G \setminus (\overline{O}_\tau \cup \overline{I}_\tau)$. Then, using 6 independent samples, we apply **DoubleTrisection**$(\mathcal{Y}, \mathcal{T}, G \setminus (\overline{O}_\tau \cup \overline{I}_\tau), \gamma)$ to compute four subsets (L_τ, U_τ) and $(\overline{L}_\tau, \overline{U}_\tau)$. With high probability, it turns out that $\overline{L}_\tau \subset L_\tau$ is made of experts below the median expert of G and $\overline{U}_\tau \subset U_\tau$ is made of experts above the median expert of G. This allows us to update as follows

$$O_{\tau+1} = O_\tau \cup L_\tau, \quad I_{\tau+1} = I_\tau \cup U_\tau, \quad \overline{O}_{\tau+1} = \overline{O}_\tau \cup \overline{L}_\tau, \quad \overline{I}_{\tau+1} = \overline{I}_\tau \cup \overline{U}_\tau \ . \tag{3.13}$$

The procedure is summarized in Algorithm 2 below.

Algorithm 1 **TreeSort**$(\mathcal{Y})$

Require: $6\tau_\infty t_\infty$ samples $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(6\tau_\infty t_\infty - 1)})$
Ensure: A tree $\mathcal{T}$ and an estimator $\hat{\pi}$

Initialize $\mathcal{T}$ as the tree with only the root $[n]$
for $t = 0, \dots, t_\infty - 1$ **do**
  Take $6\tau_\infty$ samples $\mathcal{Y}_t = (Y^{(6t\tau_\infty)}, \dots, Y^{(6(t+1)\tau_\infty - 1)})$
  Initialize $\mathcal{T}' = \mathcal{T}$
  for All the leaves G at depth t corresponding to O or I as in Figure 3.2 **do**
    Set $(O_G, P_G, I_G) =$ **BlockSort**$(\mathcal{Y}_t, \mathcal{T}, G)$
    Add (O_G, P_G, I_G) to the tree $\mathcal{T}'$
  end for
  Update $\mathcal{T} = \mathcal{T}'$
end for
Set $\hat{\pi} := \hat{\pi}(\mathcal{T})$ as in (3.12)
return $\mathcal{T}$ and $\hat{\pi}$

Algorithm 2 **BlockSort**$(\mathcal{Y}, \mathcal{T}, G)$

Require: $6\tau_\infty$ samples $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(6\tau_\infty - 1)})$, the tree $\mathcal{T}$, a leaf G in $\mathcal{T}$
Ensure: A partition of G into (O, P, I)

Set $\gamma = \lfloor |G|/2 \rfloor$ and $O_0, I_0, \overline{O}_0, \overline{I}_0 = \varnothing$
for $\tau = 0, \dots, \tau_\infty - 1$ **do**
  Take 6 samples $\mathcal{Y}_\tau = (Y^{(6\tau)}, \dots, Y^{(6\tau+5)})$
  set $\gamma = \lfloor |G|/2 \rfloor - |\overline{O}_\tau|$
  $(L_\tau, U_\tau), \quad (\overline{L}_\tau, \overline{U}_\tau) =$ **DoubleTrisection**$(\mathcal{Y}_\tau, \mathcal{T}, G \setminus (\overline{O}_\tau \cup \overline{I}_\tau), \gamma)$ as in Algorithm 3
  Update $O_{\tau+1} = O_\tau \cup L_\tau, \quad I_{\tau+1} = I_\tau \cup U_\tau, \quad \overline{O}_{\tau+1} = \overline{O}_\tau \cup \overline{L}_\tau, \quad \overline{I}_{\tau+1} = \overline{I}_\tau \cup \overline{U}_\tau$
end for
return $(O_{\tau_\infty}, G \setminus (O_{\tau_\infty} \cup I_{\tau_\infty}), I_{\tau_\infty})$

3.3.2 Description of the double trisection procedure

We now describe the trisection procedure **DoubleTrisection**. For this purpose, we first provide a few definitions.

3.3.2.1 Definitions

In this subsection, we write Y for one data set sampled according to (3.1). For the sake of simplicity, we often omit the dependence of Y in the definitions. We write $\mathcal{D}$ for the set of all dyadic numbers: $\mathcal{D} = \{2^k : k \in \mathbb{Z}\}$ and we define the sets $\mathcal{R} = \mathcal{D} \cap [1, d]$ and $\mathcal{H} = \mathcal{D} \cap \left[\frac{\zeta^2}{nd}, 1\right]$. The collection $\mathcal{R}$ corresponds to the possible scales, that is the number of questions under consideration, whereas the collection $\mathcal{H}$ corresponds to the possible heights of variations.

For all $r \in \mathcal{R}$, we write $\mathcal{Q}_r = \{1, r+1, 2r+1, \dots, \left\lfloor \frac{d}{r} \right\rfloor r + 1\}$ for the regular grid of questions with spacing r. If $\overline{P} \subset [n]$ is a set of experts, we denote $\overline{y}(\overline{P})$ as the mean of the vectors $Y_{i,\cdot}$ for $i \in \overline{P}$, that is, for all $k \in [d]$, we have $\overline{y}_k(\overline{P}) = \frac{1}{|\overline{P}|} \sum_{i \in \overline{P}} Y_{i,k}$. For any $r \in \mathcal{R}$, we define $Z(Y, \overline{P}, r)$ as the aggregation of the matrix Y on blocks of questions of size r and with lines restricted to $\overline{P}$. More formally, for any $i \in \overline{P}$ and $l \in \mathcal{Q}_r$, we have

$$Z_{i,l}(Y, \overline{P}, r) = \frac{1}{\sqrt{r}} \sum_{k=l}^{l+r-1} Y_{i,k} \quad \text{and} \quad \overline{Z}_{i,l}(Y, \overline{P}, r) = \frac{1}{\sqrt{r}} \sum_{k=l}^{l+r-1} \overline{y}_k(\overline{P}) \ . \tag{3.14}$$

Both matrices are of size $|\overline{P}| \times |\mathcal{Q}_r|$. Note that, in the above definition, $Z_{i,l}(Y, \overline{P}, r)$ and $\overline{Z}_{i,l}(Y, \overline{P}, r)$ are rescaled by $\sqrt{r}$ so that the subGaussian norm remains at most ζ. For any subset $Q \subset \mathcal{Q}_r$, we also write $Z(Y, \overline{P}, Q, r)$ for the sub-matrix of $Z(Y, \overline{P}, r)$ restricted to columns in Q.

Given a matrix $Z \in \mathbb{R}^{\overline{P}\times Q}$, a vector $w \in \mathbb{R}_+^Q$ with non-negative components and i, j in $\overline{P}$, we say i is (Z,w)-above j (or equivalently that j is (Z,w)-below i) if the projection of $Z_{i,\cdot}$ on the direction w is larger than the projection of $Z_{j,\cdot}$ on w, that is $\langle Z_{i,\cdot} - Z_{j,\cdot}, w\rangle > 0$, where $\langle .,. \rangle$ stands for the standard inner product between vectors. Now, for $\gamma \in \{1,\dots,|\overline{P}|\}$, we can consider the γ-th expert $i_\gamma \in \overline{P}$ such that there are exactly $\gamma-1$ experts which are (Z,w)-below i_γ. Given a tuning parameter $\beta > 0$ to be fixed below, we then define the (Z,w,γ,β)-trisection of $\overline{P}$ on direction w with respect to pivot index γ and matrix Z as the sets:

$$\begin{cases} U_w := U(Z,w,\gamma,\beta) &= \left\{ i \in \overline{P} \ : \ \langle Z_{i,\cdot} - Z_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle \geq \beta\sqrt{\log\left(2\frac{|\overline{P}|}{\delta}\right)} \right\} \\ L_w := L(Z,w,\gamma,\beta) &= \left\{ i \in \overline{P} \ : \ \langle Z_{i,\cdot} - Z_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle \leq -\beta\sqrt{\log\left(2\frac{|\overline{P}|}{\delta}\right)} \right\} \end{cases} . \tag{3.15}$$

Hence a (Z,w,γ,β)-trisection on direction w and pivot γ consists of two possibly empty disjoint subsets U and L which are respectively taken among the $\gamma-1$ experts (resp. the $|\overline{P}|-\gamma$) which are (Z,w)-above (resp. (Z,w)-below) the expert i_γ, with a margin of the order of $\sqrt{\log(|\overline{P}|/\delta)}$. Remark that if $\beta < \overline{\beta}$ then $U(Z,w,\gamma,\overline{\beta}) \subset U(Z,w,\gamma,\beta)$, which means that the trisection of $\overline{P}$ on direction w becomes more conservative as β increases.

In fact, (3.15) turns out to be the cornerstone or our procedure. Since the coordinates of w are non-negative, the corresponding row-wise weighted sums of the aggregation $\mathbb{E}[Z(Y,\overline{P},r)]$ of the signal matrix M are also ordered according to the oracle permutation. In other words, the k-th expert in $\overline{P}$ has the k-th highest value of the expectation of this weighted sum.

For $r \in \mathcal{R}$ and $Q \subset \mathcal{Q}_r$, choosing $w = \mathbf{1}_Q$ in (3.15) amounts to trisecting $\overline{P}$ according to the average of the observations over all questions in $\bigcup_{l\in Q}[l, l+r)$. In that case, we write for simplicity $(L_Q, U_Q) = (L_{\mathbf{1}_Q}, U_{\mathbf{1}_Q})$. When $Q = \mathcal{Q}_r$ and $w = \mathbf{1}_Q$, then (3.15) simply amounts to ranking experts according to their average over all the questions. As explained in the introduction, the global average does not lead to optimal performances. This is why most following steps in the algorithm amount to selecting suitable blocks Q of questions and directions w.

In the following, the tuning parameters β are set as follows.

$$\beta_{\text{tris}} = 4\sqrt{2}\zeta\,, \qquad \overline{\beta}_{\text{tris}} = 8\sqrt{2}\zeta \ . \tag{3.16}$$

3.3.2.2 Description of the double trisection procedure

Recall that the purpose of **DoubleTrisection** is to select subsets (L,U) and $(\overline{L},\overline{U})$ of a group $\overline{P}$ of experts in such a way that $\overline{L} \subset L$, $\overline{U} \subset U$, and experts in L (resp. in U) are with high probability below (resp. above) the γ-th expert of $\overline{P}$.

For that purpose, we have 6 independent samples $(Y^{(s)})_{s=1,\dots,6}$ sampled from (3.1) at our disposal. Fix any height $h \in \mathcal{H}$ and any scale $r \in \mathcal{R}$. **DoubleTrisection** relies on the following steps also described in Algorithm 3.

1. **Selection of a suitable subset of questions.** Using the first sample $Y^{(1)}$, we first select a subset $\widehat{Q} \subset \mathcal{Q}_r$. We postpone the definition of the selection procedure to the next subsection. We will introduce two approaches for this $\widehat{Q} := \widehat{Q}_{\text{cp}}(\overline{P},h,r)$ as in (3.20) or $\widehat{Q} := \widehat{Q}_{WM}(\mathcal{T},\overline{P},h,r)$ as in (3.26). These two definitions respectively correspond to the oblivious estimator $\hat{\pi}_{HT}$ and to the estimator with memory $\hat{\pi}_{\text{WM}}$.

2. **Average-based trisection.** Using the second sample $Y^{(2)}$, we consider the corresponding aggregated matrix $Z^{(2)} := Z(Y^{(2)},\overline{P},\widehat{Q},r)$ as defined in (3.14) which focuses on the selected blocks of questions $\widehat{Q}$. Then, we consider experts whose corresponding row sums on $Z^{(2)}$ is unusually large or small. More formally, we compute the $(Z^{(2)},\mathbf{1}_{\widehat{Q}},\gamma,\beta_{\text{tris}})$-trisection and the $(Z^{(2)},\mathbf{1}_{\widehat{Q}},\gamma,\overline{\beta}_{\text{tris}})$-trisection of $\overline{P}$ as defined in (3.15) and where the tuning parameters β_{tris} and $\overline{\beta}_{\text{tris}}$ are defined in (3.16). This allows us to obtain $(L_{\widehat{Q}},U_{\widehat{Q}})$ and $(\overline{L}_{\widehat{Q}},\overline{U}_{\widehat{Q}})$.

3. **PCA-based trisection.** Then, we focus on the conservative subset of remaining experts $\widetilde{P} = \overline{P}\setminus\overline{L}_{\widehat{Q}}\cup\overline{U}_{\widehat{Q}}$. Relying on the samples $Y^{(3)}$, $Y^{(4)}$, $Y^{(5)}$, $Y^{(6)}$, we build the corresponding aggregated matrices $Z^{(s)} := Z(Y^{(s)},\widetilde{P},\widehat{Q},r)$ restricted to the subset $\widetilde{P}$ for $s = 3,4,5$. In principle, we would like to aim at the right singular value of $\mathbb{E}[Z^{(3)} - \overline{Z}^{(3)}]$ as this would give us a nice direction w on which we could apply (3.15). For technical reasons to be explained later, we take a roundabout way, by first computing a vector $\hat{v}$ indexed by $\widetilde{P}$ which, in principle, is not too far from the left singular value of $\mathbb{E}[Z^{(3)} - \overline{Z}^{(3)}]$. More precisely, we compute $\hat{v}$ as follows

$$\hat{v} := \hat{v}(\widetilde{P},\widehat{Q},r) = \underset{\|v\|_2\leq 1}{\arg\max}\left[\|v^T(Z^{(3)} - \overline{Z}^{(3)})\|_2^2 - \frac{1}{2}\|v^T(Z^{(3)} - \overline{Z}^{(3)} - Z^{(4)} + \overline{Z}^{(4)})\|_2^2\right] . \tag{3.17}$$

The right-hand side term in (3.17) allows us to deal with the fact that the entries of the noise matrix E in (3.1) are possibly heteroskedastic. Although there exist more elegant workarounds for heteroskedastic noise (e.g. PCA [106]), the analysis in those works does not apply in our non-parametric setting. Moreover, $\hat{v}$ in (3.17) corresponds to the leading eigenvector of a square symmetric matrix and can therefore be computed efficiently. Then, we consider the image $\hat{z} = \hat{v}^T(Z^{(5)} - \overline{Z}^{(5)}) \in \mathbb{R}^{\widehat{Q}}$ of $\hat{v}$. After this, we threshold $\hat{z}$ and take the absolute values of the components. Thus, we get $\hat{w}^+ \in \mathbb{R}^Q$ defined by $(\hat{w}^+)_l = |\hat{z}_l| \mathbf{1}_{|\hat{z}_l| \geq 2\zeta\sqrt{2\log(2|\widehat{Q}|/\delta)}}$ for any $l \in \widehat{Q}$. Finally, we consider the last aggregated sample $Z^{(6)} := Z(Y^{(6)}, \overline{P}, \widehat{Q}, r)$ on the set $\overline{P} \supset \widetilde{P}$ of experts. We apply these weights $\hat{w}^+$ to compute the row-wise weighted sums of $Z^{(6)}$ and discard experts whose corresponding weighted sums is unusually small or large. More formally, we apply $(Z^{(6)}, \hat{w}^+, \gamma, \beta)$-trisection and $(Z^{(6)}, \hat{w}^+, \gamma, \overline{\beta})$-trisection of $\overline{P}$ as defined in (3.15). Doing so we obtain $(L_{\hat{w}^+}, U_{\hat{w}^+})$ and $(\overline{L}_{\hat{w}^+}, \overline{U}_{\hat{w}^+})$ respectively.

In the definition of $Z^{(6)}$ we consider the whole set of experts $\overline{P}$ instead of the remaining of experts $\widetilde{P}$ that have not been discarded because otherwise we should have needed to update the value of γ when applying (3.15).

Finally, we define the trisections (L, U) (resp. $(\overline{L}, \overline{U})$) as the union of the corresponding discarded subsets of experts based on $\mathbf{1}_{\widehat{Q}}$ and $\hat{w}^+$, this for all possible height $h \in \mathcal{H}$ and scale $r \in \mathcal{R}$. We recall that the definition of $\widehat{Q}$ was depending on h and r.

$$\begin{cases} (L, U) &= \left(\bigcup_{(h,r)\in\mathcal{H}\times\mathcal{R}} L_{\widehat{Q}}(h,r) \cup L_{\hat{w}^+}(h,r), \bigcup_{(h,r)\in\mathcal{H}\times\mathcal{R}} U_{\widehat{Q}}(h,r) \cup U_{\hat{w}^+}(h,r)\right) \\ (\overline{L}, \overline{U}) &= \left(\bigcup_{(h,r)\in\mathcal{H}\times\mathcal{R}} \overline{L}_{\widehat{Q}}(h,r) \cup \overline{L}_{\hat{w}^+}(h,r), \bigcup_{(h,r)\in\mathcal{H}\times\mathcal{R}} \overline{U}_{\widehat{Q}}(h,r) \cup \overline{U}_{\hat{w}^+}(h,r)\right). \end{cases} \tag{3.18}$$

This whole routine for computing (L, U), $(\overline{L}, \overline{U})$ is referred to as **DoubleTrisection** and is summarized in Algorithm 3. We underline that $\overline{L} \subset L \subset \overline{P}$ and $\overline{U} \subset U \subset \overline{P}$ as we took $\beta_{\text{tris}} < \overline{\beta}_{\text{tris}}$.

Algorithm 3 DoubleTrisection$((Y^{(s)})_{s=1,\dots,6}, \mathcal{T}, \overline{P}, \gamma)$

Require: 6 samples $(Y^{(s)})_{s=1,\dots,6}$, a set $\overline{P}$, a tree $\mathcal{T}$, a pivot index $\gamma \in [1, \dots, |\overline{P}|]$
Ensure: Two trisections (L, U) and $(\overline{L}, \overline{U})$ of $\overline{P}$

1: Start from $L, U, \overline{L}, \overline{U} = \varnothing$
2: **for** $h \in \mathcal{H}$, $r \in \mathcal{R}$ **do**
3: Compute $\widehat{Q} := \widehat{Q}_{\text{cp}}(\overline{P}, h, r)$ as in (3.20) or $\widehat{Q} := \widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r)$ as in (3.26) using sample $Y^{(1)}$
4: Let $Z^{(s)} := Z(Y^{(s)}, \overline{P}, \widehat{Q}, r)$, for $s \in \{2, 6\}$ be the aggregated matrices of samples defined as in (3.14)
5: Let $(L_{\widehat{Q}}, U_{\widehat{Q}})$, $(\overline{L}_{\widehat{Q}}, \overline{U}_{\widehat{Q}})$ be resp. the $(Z^{(2)}, \mathbf{1}_{\widehat{Q}}, \gamma, \beta)$ and the $(Z^{(2)}, \mathbf{1}_{\widehat{Q}}, \gamma, \overline{\beta})$-trisections of $\overline{P}$ as in (3.15)
6: Define $\widetilde{P} = \overline{P} \setminus (\overline{L}_{\widehat{Q}} \cup \overline{U}_{\widehat{Q}})$ and the aggregated samples $Z^{(s)} := Z^{(s)}(Y^{(s)}, \widetilde{P}, \widehat{Q}, r)$ for $s \in \{3, 4, 5\}$
7: Compute the PCA-like direction $\hat{v} := \hat{v}(\widetilde{P}, \widehat{Q}, r)$ as in (3.17)
8: Compute $\hat{z} = \hat{v}^T(Z^{(5)} - \overline{Z}^{(5)})$ and define $\hat{w}^+$ by $(\hat{w}^+)_l = |\hat{z}_l| \mathbf{1}_{|\hat{z}_l| \geq 2\zeta\sqrt{2\log(2|\widehat{Q}|/\delta)}}$ for any $l \in \widehat{Q}$
9: Let $(L_{\hat{w}^+}, U_{\hat{w}^+})$, $(\overline{L}_{\hat{w}^+}, \overline{U}_{\hat{w}^+})$ be resp. the $(Z^{(6)}, \hat{w}^+, \gamma, \beta)$ and the $(Z^{(6)}, \hat{w}^+, \gamma, \overline{\beta})$-trisections of $\overline{P}$ as in (3.15)
10: Update $L = L \cup L_{\hat{w}^+} \cup L_{\widehat{Q}}$, $U = U \cup U_{\hat{w}^+} \cup U_{\widehat{Q}}$, $\overline{L} = \overline{L} \cup \overline{L}_{\hat{w}^+} \cup \overline{L}_{\widehat{Q}}$, $\overline{U} = \overline{U} \cup \overline{U}_{\hat{w}^+} \cup \overline{U}_{\widehat{Q}}$
11: **end for**
12: **return** (L, U), $(\overline{L}, \overline{U})$

To finish the definition of the estimator, it remains to describe the selection procedures for the suitable blocks of questions that are used in Line 3 of Algorithm 3. As explained above, we consider two procedures $\widehat{Q} := \widehat{Q}_{\text{cp}}(\overline{P}, h, r)$ as in (3.20) or $\widehat{Q} := \widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r)$ as in (3.26) - which respectively apply to the oblivious estimator $\hat{\pi}_{HT}$ and to the estimator $\hat{\pi}_{WM}$ with memory.

3.3.2.3 Definition of $\widehat{Q}_{\text{cp}}$

We start with $\widehat{Q}_{\text{cp}}(\overline{P}, h, r)$. The corresponding estimator $\hat{\pi}_{HT}$ is called an oblivious hierarchical sorting tree estimator because $\widehat{Q}_{\text{cp}}$ only depends on the restriction of the data to $\overline{P}$. As a consequence, the corresponding **BlockSort** procedure (see Algorithm 2) which builds a trisection of a group G of experts into three subgroups (O, P, I) only depends on the observations on this set G of experts. In other words, the recursive construction of the hierarchical sorting tree estimator is completely oblivious of the rest of the tree. Up to our knowledge, this feature is shared by most hierarchical clustering algorithms.

Fix some height $h \in \mathcal{H}$ and $r \in \mathcal{R}$. Intuitively, $\widehat{Q}_{\mathrm{cp}}(\overline{P}, h, r)$ amounts to focusing on the subset of questions around which the empirical mean expert $\overline{y}(\overline{P})$ has a high-variation. We provide some intuition on the rationale of this approach in the next subsection. More precisely, we define the CUSUM statistic:

$$\tilde{r} = 8\left\lceil \left(\frac{32\zeta^2}{|\overline{P}|h^2} \log\left(\tfrac{2d}{\delta}\right) \right) \vee r \right\rceil \quad \text{and} \quad \widehat{\mathbf{C}}_{k,\tilde{r}}(\overline{P}) = \frac{1}{\tilde{r}} \left(\sum_{k'=k}^{k+\tilde{r}-1} \overline{y}_{k'}(\overline{P}) - \sum_{k'=k-\tilde{r}}^{k-1} \overline{y}_{k'}(\overline{P}) \right) . \tag{3.19}$$

In a nustshell, $\widehat{\mathbf{C}}_{k,\tilde{r}}(\overline{P})$ is the empirical variation of $\overline{y}(\overline{P})$ at question k and at scale $\tilde{r} \geq r$. Then, we define $\widehat{D}_{\mathrm{cp}}$ as the set of questions where the CUSUM statistic is larger than $h/4$, and $\widehat{Q}_{\mathrm{cp}} \subset \mathcal{Q}_r$ for the corresponding subset of blocks or questions of size r.

$$\widehat{D}_{\mathrm{cp}} = \left\{ k \in [d] \ : \ \widehat{\mathbf{C}}_{k,\tilde{r}}(\overline{y}(\overline{P})) \geq \frac{h}{4} \right\} \quad \text{and} \quad \widehat{Q}_{\mathrm{cp}} = \left\{ l \in \mathcal{Q}_r \ : \ \widehat{D}_{\mathrm{cp}} \cap [l, l+r) \neq \varnothing \right\} . \tag{3.20}$$

In (3.19), the choice of $\tilde{r} \geq r$ is due to the fact that we need to compute an empirical mean $\mathbf{C}_{k,\tilde{r}}(\overline{y}(\overline{P}))$ on enough questions so that its standard deviation is small compared to h.

3.3.2.4 Definition of $\widehat{Q}_{WM}$

Finally, we describe $\widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r)$ which corresponds to the estimator $\hat{\pi}_{WM}$. The set $\overline{P}$ is a subset of a leaf G of the tree $\mathcal{T}$ and we write t for its depth. As illustrated in Figure 3.2, there is a natural order on the nodes of $\mathcal{T}$ at depth t that have been either obtained as subsets of type O or I in **BlockSort** (Algorithm 2). We can index these nodes according to the ordering by setting $G^{(0)} = G$ and then $G^{(1)}$, $G^{(2)}$,... as the following groups. Similarly, $G^{(-1)}$, $G^{(-2)}$,... stand for the groups preceding $G^{(0)}$. See Figure 3.2 for an illustration. In fact, with high probability, for any a, all the experts in $G^{(a+1)}$ are above the expert in $G^{(a)}$. As a consequence, the observations in $G^{(1)}$ and $G^{(-1)}$ can bring some information on the behavior of the experts in $\overline{P} \subset G^{(0)}$.

Fix $r \in \mathcal{R}$ and $h \in \mathcal{H}$. Define $\tilde{r} \in \mathcal{R}$ as $\tilde{r} = 4(\lceil 2^9 \log(4d|\mathcal{R}|/\delta) \frac{\zeta^2}{|\overline{P}|h^2} \rceil^{dya} \vee r)$, where $\lceil x \rceil^{dya} = 2^{\lceil \log_2(x) \rceil}$. As before, $\tilde{r} \geq r$ stands for the scale which is required if we want to estimate the variation of $\overline{y}(\overline{P})$ with a standard error small compared to h.

Now, we consider any scale $r_{\mathrm{cp}} \in [4r, 2\tilde{r}] \cap \mathcal{R}$. The rationale is that, if $r_{\mathrm{cp}} < \tilde{r}$, we can reduce the standard deviations of the empirical means by considering an average over experts in neighboring groups. Define the upper neighborhood $\mathcal{V}^+_{r_{\mathrm{cp}}}$ and lower neighborhood $\mathcal{V}^-_{r_{\mathrm{cp}}}$ as the set of groups above G and below G that are necessary to have enough experts at scale r_{cp}.

$$a^+_{WM} = \min\{a \ : \ |G^{(1)}| + \cdots + |G^{(a)}| \geq \tfrac{2^{11}\zeta^2 \log(4d|\mathcal{R}|/\delta)}{r_{\mathrm{cp}} h^2}\} \quad \text{and} \quad \mathcal{V}^+_{r_{\mathrm{cp}}} = \bigcup_{a=1}^{a^+_{WM}} G^{(a)} \ ; \tag{3.21}$$

$$a^-_{WM} = \min\{a \ : \ |G^{(-1)}| + \cdots + |G^{(-a)}| \geq \tfrac{2^{11}\zeta^2 \log(4d|\mathcal{R}|/\delta)}{r_{\mathrm{cp}} h^2}\} \quad \text{and} \quad \mathcal{V}^-_{r_{\mathrm{cp}}} = \bigcup_{a \in -a^-_{WM}}^{-1} G^{(a)} . \tag{3.22}$$

For a given subset $\overline{P} \subset G$, we define the corresponding CUSUM statistic $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ computed on the questions $[k - r_{\mathrm{cp}}, k + r_{\mathrm{cp}})$ and using the empirical mean observations in $\mathcal{V}^+_{r_{\mathrm{cp}}} \cup \mathcal{V}^-_{r_{\mathrm{cp}}}$ if $r < 2\tilde{r}$ and in $\overline{P}$ if $r_{\mathrm{cp}} = 2\tilde{r}$:

$$\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} = \frac{1}{r_{\mathrm{cp}}} \begin{cases} \sum_{k'=k}^{k+r_{\mathrm{cp}}-1} \overline{y}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}} \cup \mathcal{V}^-_{r_{\mathrm{cp}}}) - \sum_{k'=k-r_{\mathrm{cp}}}^{k-1} \overline{y}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}} \cup \mathcal{V}^-_{r_{\mathrm{cp}}}) & \text{if } r_{\mathrm{cp}} \in [8r, 2\tilde{r}) \\ \sum_{k'=k}^{k+r_{\mathrm{cp}}-1} \overline{y}_{k'}(\overline{P}) - \sum_{k'=k-r_{\mathrm{cp}}}^{k-1} \overline{y}_{k'}(\overline{P}) & \text{if } r_{\mathrm{cp}} = 2\tilde{r} \end{cases} \tag{3.23}$$

If $r_{\mathrm{cp}} = 2\tilde{r}$, this new definition of the CUSUM with memory matches the definition (3.19) in the previous paragraph. For $r_{\mathrm{cp}} < 2\tilde{r}$, we are not able to average on enough expert in $\overline{P}$. To deal with this issue, we average on a suitable number of neighboring experts.

Beside considering questions around which the variations of $\overline{y}(\overline{P})$ are large enough, we also check whether, on the corresponding regions, the width of $\overline{P}$, that is the difference between the best expert and the worst expert in $\overline{P}$ is high enough. Given a question $k \in [d]$, we define $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ as the difference between the locals average on $[k - r_{\mathrm{cp}}, k + r_{\mathrm{cp}})$ of the neighborhoods of G that is

$$\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} = \frac{1}{2r_{\mathrm{cp}}} \sum_{k'=k-r_{\mathrm{cp}}}^{k+r_{\mathrm{cp}}-1} \overline{y}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{y}_{k'}(\mathcal{V}^-_{r_{\mathrm{cp}}}) . \tag{3.24}$$

Since the groups $G^{(1)}, \ldots, G^{(2)}$ are above the best expert in $\overline{P}$ and since the groups $G^{(-1)}$, $G^{(-2)}$,... are below the worst expert in $\overline{P}$, this statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ overestimates the width of $\overline{P}$. In the next subsection, we will explain why it is relevant to consider the width of $\overline{P}$.

We are now equipped to define the subsets $\widehat{D}_{WM}(\mathcal{T}, \overline{P}, h, r)$ of suitable questions and the corresponding $\widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r)$ of corresponding blocks.

$$\widehat{D}_{WM} = \Big\{k \in [d] \ : \ \exists r_{\mathrm{cp}} \in [4r, \tilde{r}] \cap \mathcal{R} \text{ s.t. } \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} \geq \frac{h}{16} \text{ and } \widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} \geq \frac{h}{16}\Big\} \ ; \tag{3.25}$$

$$\widehat{Q}_{WM} = \{l \in \mathcal{Q}_r \ : \ \widehat{D}_{WM} \cap [l, l+r) \neq \varnothing\} \ . \tag{3.26}$$

In other words, $\widehat{D}_{WM}$ is made of questions for which there exists a scale r_{cp} such that simultaneously the empirical variations $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ at scale $2r_{\mathrm{cp}}$ is at least of order h and the empirical width at scale r_{cp} is at least of order h.

3.3.3 Comments on the procedure and relation to the literature

These twin procedure are quite involved and combine several ingredients, some of them being already used by Liu and Moitra [57]. In particular, they introduced the key ideas of localization of the suitable blocks of questions through change-point detection on the mean expert and of a spectral clustering scheme for dividing blocks of experts. Still, we need to add several key elements in order to deal with the arguably more involved setting where $n \leq d$. We describe below how our procedure compares to [57] and highlight also the main differences and new ideas. Also, despite the fact that our procedure is very involved, it remains computationally efficient. Overall, the full procedure requires $O[\log^c(\frac{nd}{\zeta-\delta})(nd^2+n^3)]$ operations for some $c>0$. Indeed, each of the main steps of the algorithm correspond to matrix multiplications and computations of the largest eigenvector of a square symmetric matrix.

In this subsection, we discuss three key steps of the algorithm: (i) the selection of blocks of questions corresponding to the high-variation regions of the average expert in the group as in the definition of $\widehat{Q}_{\mathrm{cp}}$, (ii) construction of the weights vector w^+ by a spectral procedure, (iii) the use of neighboring groups in $\widehat{Q}_{WM}$.

3.3.3.1 Detecting high-variation regions of the average expert

Recall that, for a fixed $r \in \mathcal{R}$ and $h \in \mathcal{H}$, $\widehat{Q}_{\mathrm{cp}}$ selects blocks of questions in which the variations (3.19) $\widehat{\mathbf{C}}_{k,\tilde{r}}(\overline{P})$ of the average $\overline{y}(\overline{P})$ at question k and at scale $\tilde{r} \geq r$ is higher than $h/4$.

To explain the rationale behind this choice, let us first consider a toy example depicted in Figure 3.3. Assume that the group $\overline{P}$ is made of two subgroups of experts U^* and L^* and that all the experts in U^* and all the experts in L^* are identical. Also, assume that the corresponding rows only differ on r consecutive questions by h and are otherwise identical. As illustrated in Figure 3.3, it turns out that the expected average expert $\overline{m}(\overline{P}) = \mathbb{E}[\overline{y}(\overline{P})]$ needs to vary by h at scale r near the block of questions on which the two groups of experts are differing. This is due to the fact that both the rows corresponding to U^* and L^* are isotonic and that the row of U^* is always larger or equal to that of L^*. As a consequence, by restricting our attention to the blocks of questions corresponding to high-variation regions of $\overline{m}(\overline{P})$ (or in practice $\overline{y}(\overline{P})$), we are able to much reduce the dimension of the problem and thereby to improve our ability to distinguish different experts.

Beyond this toy example, we show in Lemma 3.5.10 that there exists a suitable scale $r \in \mathcal{R}$ and a suitable height $h \in \mathcal{H}$ such that, by restricting our attention to blocks of questions of size r such that the expected average expert $\overline{m}(\overline{P})$ varies by at least $h/2$, we are able to retain a significant proportion of the differences between experts in $\overline{P}$. In other words, focusing on regions of high-variation of $\overline{y}(\overline{P})$ in the blocks $\widehat{Q}_{\mathrm{cp}}$ is, at least for some scale and some height, a suitable dimension reduction technique. This phenomenon was already observed in [56] and their procedure also uses such dimension detection techniques. In this chapter, we also build upon this idea, which has also important consequences, in a related yet different manner, in the rectangular case where $n \leq d$.

If we do not apply the spectral clustering sorting steps in $\hat{\pi}_{HT}$, that is, if we do not compute $\hat{v}$ and $\widehat{w}^+$ in **DoubleTrisection**, then we would get a risk bound for $\hat{\pi}_{HT}$ of the order of $\zeta^{5/3}nd^{1/6}$ instead of that of Theorem 3.2.2. In other words, the dimension reduction in $\widehat{Q}_{\mathrm{cp}}$ is alone sufficient to recover the optimal risk in the case where d is quite large and ζ is mild - namely $n \leq \zeta^{2/3}d^{1/3}$ and $\zeta \in [1/d, \sqrt{d}]$.

3.3.3.2 On the spectral estimation of the weights

In this subsection, we explain how the computation of $\hat{z}$ in (3.17) and the corresponding weights $\hat{w}^+$ allow to improve over the $\zeta^{5/3}nd^{1/6}$ rate. Again, we start with a motivating toy example depicted in Figure 3.4. As previously, we consider a situation where $\overline{P}$ can be decomposed into two subgroups U^* and L^* of the same size. The corresponding rows L^* are block-constant with blocks of questions of size r and increased by h at the end of each block of questions. On the other hand, the corresponding lines of U^* are, in each block of questions, either equal to the rows of L^*, or are exactly at a distance h above. These last blocks of questions

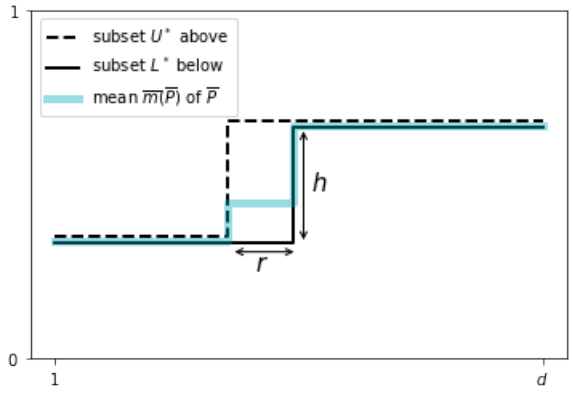
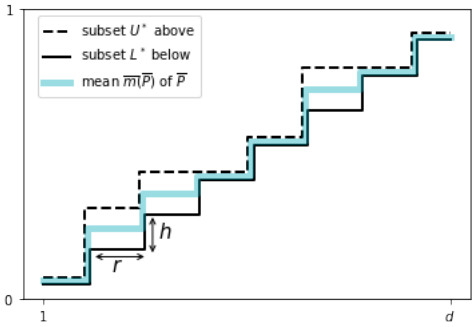


Figure 3.3: In this toy example, the group $\overline{P}$ is only made of two types of experts, those in U^* and those in L^*. The high-variation region of $\overline{m}(\overline{P})$ corresponds to the questions on which U^* and L^* differ.

are the only ones which are informative when it comes to distinguishing the best experts in the group - namely U^* - from the worst experts in the group - namely L^*. Some of the blocks corresponding to high-variation regions of the expected average row $\overline{m}(\overline{P})$ do not convey any information on the difference between U^* and L^* – see Figure 3.4. In this example, at scale r, all the blocks of size r are to be detected by the high variation dimension reduction step, that is $\widehat{Q}_{\mathrm{cp}} = \mathcal{Q}_r$. At the second step, we consider the corresponding aggregated matrix $Z - \overline{Z}$ at scale r as defined in (3.14). To be more specific, let us assume that $|L^*| = |U^*| = 3$. Then, $Z - \overline{Z}$ is a 6×8 matrix whose expectation is of the form of the right panel in Figure 3.4.



$$\mathbb{E}[Z - \overline{Z}] = \frac{1}{2}\sqrt{r}h \begin{pmatrix} 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Figure 3.4: In this toy example, the group $\overline{P}$ is only made of two types of experts U^* and L^* with $|L^*| = |U^*| = 3$.

In this specific example, the rank of this expected matrix is one and some of its columns are completely useless to decipher experts in U^* from experts in L^*. In contrast, taking w as a right singular vector associated to the largest singular value of this matrix would allow us to select the significant blocks of questions while discarding the irrelevant ones. While this example is very specific, this still sheds some light on why spectral clustering procedure can be of interest for this problem and how it can help recover blocks of questions that are the most informative for dividing the experts.

Let us come back to a general matrix M and to the spectral step of **DoubleTrisection** as described in the previous section. Up to a permutation of its rows, the expectation $\Theta^{(3)} - \overline{\Theta}^{(3)}$ of $Z^{(3)} - \overline{Z}^{(3)}$ is isotonic in each column. It turns out that the entries of any left singular vector associated to the largest singular value of $\Theta^{(3)} - \overline{\Theta}^{(3)}$ is, up to the permutation, either non-increasing or non-decreasing. As a consequence, the left-singular value of $\Theta^{(3)} - \overline{\Theta}^{(3)}$ can bring information on the underlying ranking. This property is at the heart of spectral ranking algorithms [93]. Unfortunately, contrary to the analysis of spectral methods in the Bradley-Luce-Terry model [22, 20], we cannot control the entry-wise deviations of the left singular eigenvector of $Z^{(3)} - \overline{Z}^{(3)}$ because the matrix $\Theta^{(3)} - \overline{\Theta}^{(3)}$ is non-parametric and does not necessarily exhibit any spectral gap. To handle this, Liu and Moitra [56] suggest to compute a right singular vector of $Z^{(3)} - \overline{Z}^{(3)}$ and, using another independent sample, to compare the experts based on the corresponding weighted average of the experts. Unfortunately, while their

analysis provides near optimal results for $n = d$, this would not work for $n \leq d$. In **DoubleTrisection**, we apply a more involved workaround (i) to handle possible heteroskedastic noise and (ii) to improve the convergence rates in comparison to Liu and Moitra [56]. Indeed, we first compute in (3.17) a debiased version $\hat{v}$ of the left-singular vector of $Z^{(3)} - \overline{Z}^{(3)}$. Then, we compute the image $[Z^{(5)} - \overline{Z}^{(5)}]^T \hat{v}$, threshold it, and take its absolute value to obtain our estimated weights $\hat{w}^+$. In principle, $\hat{w}^+$ aims at being close to the right first singular vector of $\Theta^{(3)} - \overline{\Theta}^{(3)}$. In comparison to Liu and Moitra [56], $\hat{w}^+$ better handles the situation where the matrix $\Theta^{(3)} - \overline{\Theta}^{(3)}$ is highly rectangular (with many columns) and where its corresponding right singular vector is nearly sparse.

3.3.3.3 On the tree information and the definition of $\widehat{Q}_{WM}$

The oblivious estimator $\hat{\pi}_{HT}$ based on $\widehat{Q}_{\mathrm{cp}}$ is only proved to achieve the suboptimal error of Theorem 3.2.2. In this section, we explain how $\widehat{Q}_{WM}$ improves the performances of the procedure by relying on the neighboring experts to fix one possible weakness of $\widehat{Q}_{\mathrm{cp}}$ and so, improve the dimension reduction step.

Indeed, $\widehat{Q}_{\mathrm{cp}}$ selects spurious blocks of questions. In the previous toy example (Figure 3.4), some of the blocks corresponding to high-variation values of the expected mean expert $\overline{m}(\overline{P})$ do not bring any suitable information for ordering the experts in $\overline{P}$ because, in these blocks, all the experts are close to each other. In other words, the width of $\overline{P}$, that is the difference between the best and worst experts in $\overline{P}$, is small. It is not possible to easily estimate this width from the observations in $\overline{P}$ since this would require to have sorted the experts in $\overline{P}$ in the first place. Still, we can estimate this width by comparing the average of experts that are above $\overline{P}$ with average of experts that are below $\overline{P}$. A first idea would therefore be to consider a large enough number of experts above and below $\overline{P}$ in order to estimate the width with a small variance and to exclude regions such that the estimated width on a window of size r is small compared to h. This is exactly the purpose of the statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r}$. The selected blocks $\widehat{Q}_{WM}$ only contain regions such that the estimated width $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r}$ is large enough compared to h –see the left panel in Figure 3.5. Unfortunately, the statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r}$ may suffer from a large positive bias if the experts above or below $\overline{P}$ are away from $\overline{P}$. Moreover, considering only the scale r is not sufficient because we are forced to average over many experts above and below $\overline{P}$ in order to have a small variance at this small scale, leading to a large bias. For this reason, we consider all possible scales r_{cp} (in a dyadic grid) between r and $\tilde{r}$.

Another important idea in the dimension reduction scheme is the following: If there is a region of questions in which, not only the mean experts of the group $\overline{P}$ but also the mean experts in neighboring groups of $\overline{P}$ have a high variation, it is interesting to detect this high-variation region by relying on all these neighboring groups in order to decrease the variance of the CUSUM statistic. With this idea, we are able to consider the CUSUM statistic at a smaller scale $r_{\mathrm{cp}} \leq \tilde{r}$ –see the right panel in Figure 3.5. This is exactly the purpose of the statistic $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$.

In our procedure, we build a collection $\widehat{D}_{WM}$ that selects a question k if there exists a scale r_{cp} in $[4r, \tilde{r}]$ such that both the CUSUM statistic $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ at scale $2r_{\mathrm{cp}}$ is large and the empirical width $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ is large. This combines the two ideas described in the previous paragraphs which, in turn, allows us to further reduce the dimension in comparison to $\widehat{D}_{\mathrm{cp}}$ while ensuring that the selected questions in $\widehat{D}_{WM}$ contains all the relevant regions to trisect $\overline{P}$, namely regions of size r, on which $\overline{P}$ has a variation at least of the order of h and the width of $\overline{P}$ is at least of the order of h.

Interestingly, in the square case where $n = d$ considered in [56] or more generally when $n \geq d$, this dimension reduction variant is not necessary to achieve the minimax risk as the oblivious estimator $\hat{\pi}_{HT}$ is already optimal. The dimension reduction scheme $\widehat{Q}_{WM}$ allows us to improve the risk bound from that Theorem 3.2.2 to that of Theorem 3.2.3. In the specific case where the noise level ζ is equal to one, the term $n^{2/3}d^{1/3}$ in the risk bound is improved to the optimal one $n^{3/4}d^{1/4}$. Hence, building upon the neighboring groups in $\widehat{Q}_{WM}$ turns out to be the key ingredient to recover the minimax risk in the large d regime where $n \in [d^{1/3}, d]$.

3.4 Partial observations

We now come back to the partial observation setting. Given $\lambda > 0$, we are given $Poi(\lambda nd)$ independent observations (x_t, y_t) where x_t is sampled uniformly in $[n] \times [d]$ and, conditionally to x_t, $y_t = M_{x_t} + E_{x_t}$ is an observation of the full model (3.1) at position x_t. As noted above, λ stands for the sampling effort and the larger λ, the more samples on average.

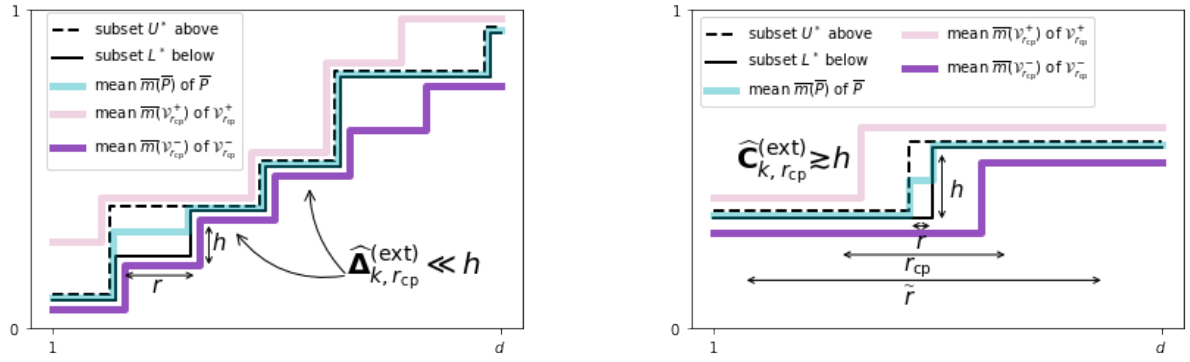


Figure 3.5: In these two panels, the group $\overline{P}$ is only made of two types of experts, those in U^* and those in L^*. The curves $\overline{m}(\mathcal{V}^+_{r_{cp}})$ and $\overline{m}(\mathcal{V}^-_{r_{cp}})$ respectively correspond to the expected average experts in the neighboring groups $\mathcal{V}^+_{r_{cp}}$ and $\mathcal{V}^-_{r_{cp}}$ defined in (3.21). In the left panel, the third and fourth blocks are not selected because the corresponding statistic $\widehat{\Delta}^{(ext)}_{k,r_{cp}}$ is small. In the right panel, both the statistic $\widehat{C}^{(ext)}_{k,r_{cp}}$ and $\widehat{\Delta}^{(ext)}_{k,r_{cp}}$ are large compared to h. If $\overline{P}$ is small, this block is selected using a scale $r_{cp} \lesssim \tilde{r}$.

3.4.1 Minimax lower bound

As in Section 3.2.1, we first state a minimax lower bound in the case where the noise matrix E is made of independent Gaussian random variables with variance ζ^2. Note that the following minimax lower bound also handle the noise case where $\zeta = 0$, i.e. the noiseless case.

Theorem 3.4.1. *There exist universal constants c and c' such that the following holds for any $n \geq 2$, any $d \geq 1$, $\lambda > 0$, and $\zeta \geq 0$:*

$$\inf_{\hat{\pi}} \sup_{\substack{\pi^* \in \Pi_n \\ M : M_{\pi^*-1} \in \mathbb{C}_{BISO}}} \mathbb{E}_{(\pi^*, M)}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2] \geq c\left[\left(\mathcal{R}_F[n, d, \zeta/\sqrt{\lambda}] + \frac{n}{\lambda}e^{-2\lambda}\right) \bigwedge nd\right] . \tag{3.27}$$

As in the previous minimax lower bound, the quantity nd simply appears because the entries of M lie in $[0, 1]$. In (3.27), we recognize two terms. First, $\mathcal{R}_F[n, d, \zeta/\sqrt{\lambda}]$ corresponds to the minimax risk for recovering π^* in a full observation model with noise $\zeta/\sqrt{\lambda}$. The second term $\frac{n}{\lambda}e^{-2\lambda}$ does not depend on ζ and is also present in the noiseless setting. It simply quantifies the fact that, for $\lambda < 1$, observations are lacking so that it is impossible to correctly rank experts if there are no observations on the questions on which they are distinct. As the minimax lower bound in (3.27) turns out to be nearly tight in light of Theorem 3.4.3, we refer to (3.27) as the minimax risk in the following. For the purpose of the discussion, we will first focus on the case where $\zeta = 1$ and $\lambda < 1$, which corresponds to the case where we really have partial observations on the matrix. We will then turn to $\zeta = 1$ and $\lambda > 1$, which corresponds to the case where we observe several times each entry of the matrix. Finally, we discuss the noiseless case where $\zeta = 0$.

Low-sample size. We first focus on the case where $\zeta = 1$ and $\lambda < 1$, which corresponds to the case where we really have partial observations on the matrix. If $\lambda \leq 1/d$, then the minimax risk is of the order of nd and it is impossible to perform significantly better than a random guess. This is not surprising as there are, in expectation, less than one observation on each row. For $\lambda \in [1/d, 1]$, the minimax risk is of the order of

$$\frac{nd^{1/6}}{\lambda^{5/6}} \bigwedge \frac{n^{3/4}d^{1/4}}{\lambda^{3/4}} + \frac{n}{\lambda}$$

In the rectangular case where $n \geq d$, the minimax risk is then of the order of n/λ for $\lambda \in [1/d, 1]$. When $n \in [d^{1/3}, d]$, the minimax risk is of the order of n/λ for $\lambda \in [1/d, n/d]$, and of the order of $\frac{n^{3/4}d^{1/4}}{\lambda^{3/4}}$ for $\lambda \in [n/d, 1]$. For even smaller $n \leq d^{1/3}$, there is one more regime since

$$\mathcal{R}_F[n, d, 1/\sqrt{\lambda}] \asymp \begin{cases} \frac{n}{\lambda} & \text{if } \lambda \in [\frac{1}{d}, \frac{n}{d}] ; \\ \frac{n^{3/4}d^{1/4}}{\lambda^{3/4}} & \text{if } \lambda \in [\frac{n}{d}, \frac{n^3}{d}] ; \\ \frac{nd^{1/6}}{\lambda^{5/6}} & \text{if } \lambda \in [\frac{n^3}{d}, 1] . \end{cases}$$

Large-sample size. In the setting where $\lambda > 1$ and $\zeta = 1$, there are several observations per entries. In this case, there are many regimes in (3.27) that depend on n, d, ζ, and λ. To simplify the discussion, we focus here

on the case $n = d$ and $\zeta = 1$. Then, the minimax risk is of the order of $\frac{n^{3/4}d^{1/4}}{\lambda^{3/4}}$ for $\lambda \in [1, n^2]$ and is of the order of $n\sqrt{d}/\lambda$ for $\lambda \geq n^2$. This 'easy rate' $n\sqrt{d}/\lambda$ is achieved by the simple procedure that ranks the experts according to the row sums [83, 60]. This simple method turns out to be optimal in the regime where there are more than n^2 observations per entry.

Noiseless case. In the extreme case where $\zeta = 0$ and $\lambda \geq 1/d$, the minimax risk is of the order of $(n/\lambda) \times e^{-2\lambda}$, which, for some small λ is of the order of n/λ. This minimax lower bound is quite simple to prove. Without loss of generality, suppose that $1/\lambda$ is an integer. Consider a matrix M such that all its columns, except its $1/\lambda$ first ones are constant and equal to one, so that it boils down to considering a ranking problem of size $n \times (1/\lambda)$. In this reduced model, there are two types of experts: (a) experts that are constant and equal to zero and (b) experts that are constant and equal to one. Obviously, if one is given at least one noiseless observation on a row, then it is possible to assign it to a group. However, on each row there is a probability e^{-1} of having no observations. Hence, on expectations there are n/e experts that are impossible to classify. For this reason, any estimator must suffer from a risk at least of the order n/λ.

3.4.2 Reduction to the full observation model

We now describe a scheme to adapt the estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$ that we developed in the full observation setting of Section 3.2, to this more general Poissonian setting (3.2), which encompasses the partial observation setting as well as the over-complete observation setting where each entry is sampled several times. Roughly, if λ is small, we simply decrease the number of columns of the matrix M in order to obtain a reduced matrix with full observations. Conversely, if λ is really large, which corresponds to the case of multiple observations per entry, we simply average the multiple observations per entry to reduce the noise levels.

As in Section 3.2, we fix $\delta \in (0, 1)$ that will correspond to a small probability. Given this $\delta > 0$, we denote $\Upsilon^* = \Upsilon^*(n, d, \zeta/\sqrt{\lambda \vee 1})$ the number of independent samples required in Section 3.2 for the estimation through $\hat{\pi}_{HT}$ or $\hat{\pi}_{WM}$ of the $n \times d$ matrix M with a noise level equal to $\zeta/\sqrt{\lambda \vee 1}$. Recall that Υ^* is of the order of $\log^8(nd(\lambda \vee 1)/(\delta\zeta_-))$.

Define $\lambda_- = \lambda/[4\Upsilon^*]$. For any $i \in [n]$ and any $S \subset [d]$, we write $n_{i,S}$ the number of observations in the sample falling in $\{i\} \times S$, that is $n_{i,S} = |\{t : \ x_t \in \{i\} \times S\}|$. The following lemma is a simple consequence of Chernoff inequality for Poisson random variables.

Lemma 3.4.2. *Assume that $\lambda_- \in [2/d, 1]$, we fix $l(\lambda) = \lfloor 1/\lambda_- \rfloor$. With probability higher than $1 - \delta$, we have*

$$\min_{i \in [n]} \min_{j \in [\lfloor d/l(\lambda) \rfloor]} n_{i,[(j-1)l(\lambda)+1, jl(\lambda)]} \geq \Upsilon^* \ .$$

Now assume that $\lambda_- > 1$. With probability higher than $1 - \delta$, we have

$$\min_{i \in [n]} \min_{j \in [d]} n_{i,\{j\}} \geq \lambda_- \Upsilon^* \ .$$

Henceforth, we work under the event introduced in the previous lemma. If this event does not hold, we choose $\hat{\pi}_{WMP}$ arbitrarily. To build $\hat{\pi}_{WMP}$, we consider three subcases that depend on the value of λ:

1. **Very small sample size**. If $\lambda_- \leq 2/d$, then we simply choose $\hat{\pi}_{WMP}$ uniformly at random over the set of all possible permutations. While this choice does not depend on the data and could therefore seem suboptimal, it is not the case, as the minimax lower bound states that it is impossible to perform better than random guess in this setting.

2. **Small sample size**. If $\lambda_- \in [2/d, 1]$, then we build Υ^* matrices $\mathcal{Y}^{\downarrow} = (Y^{\downarrow(0)}, Y^{\downarrow(1)}, Y^{\downarrow(\Upsilon^*-1)})$ of size $n \times \lfloor d/l(\lambda) \rfloor$ in the following way. For any $i \in [n]$, $j \in [\lfloor d/l(\lambda) \rfloor]$ and $s \in [0, \Upsilon^* - 1]$, $Y^{\downarrow(s)}_{i,j} = y_t$ where t is the $(s+1)$-th observation such that $x_t \in \{i\} \times [l(\lambda)j + 1, l(\lambda)(j+1)]$. On the event of Lemma 3.4.2, this definition is valid as we observe enough samples for any i, j. Then, we compute $\hat{\pi}_{WMP}$ as the variant $\hat{\pi}_{WM-SR}$, introduced in Section 3.5.7, applied to this sample of reduced matrices.

3. **Large sample size**. If $\lambda_- \geq 1$, then we build Υ^* matrices $\mathcal{Y}^{\downarrow} = (Y^{\downarrow(0)}, Y^{\downarrow(1)}, Y^{\downarrow(\Upsilon^*-1)})$ of size $n \times d$ in the following way. For any $i \in [n]$, $j \in [d]$, $l \in [\lfloor \lambda_- \rfloor]$, and $s \in [0, \Upsilon^* - 1]$, define $Y^{\downarrow(s)}_{i,j} = \frac{1}{\lfloor \lambda^- \rfloor} \sum_t y_t$ where the y_t's are the z-th observations such that $x_t = (i, j)$ with $z \in [1 + (s-1)\lfloor \lambda_- \rfloor, s\lfloor \lambda_- \rfloor]$. In other words, we build the samples $\mathcal{Y}^{\downarrow}$ be averaging $\lfloor \lambda_- \rfloor$ observations on each entries. Again, on the event of Lemma 3.4.2, this definition is valid as we observed enough samples for any i, j. Then, we define $\hat{\pi}_{WMP}$ as $\hat{\pi}_{WM}$ applied to this sample of averaged matrices. By averaging the independent observations, we reduce the noise level of each entry from ζ to $\zeta/\sqrt{\lambda_-}$.

For $\lambda_- \le 2/d$, there are very few observations on each row so that it is very difficult to compare the experts. For $\lambda_- \in [2/d, 1]$, we have access to less than Υ^* noisy observations of the matrix M. The rationale of our procedure is to group together $l(\lambda)$ consecutive questions together in such a way that there are enough observations on each of these groups. The resulting matrices of observations $Y^{\downarrow(s)}$ have around $\lambda d/\Upsilon^*$ columns. We could have applied the procedure $\hat{\pi}_{WM}$ defined in the previous section to $\mathcal{Y}^{\downarrow}$, but the corresponding subGaussian norm of the noise would be $1+\zeta$ (instead of ζ) because there is additional variability coming from the fact that any entry in the reduced matrices has been sampled uniformly among $l(\lambda)$ entries in the original matrices. This would lead us to a procedure achieving the minimax rate with respect to n, d, and λ but with a suboptimal dependency with respect to ζ since ζ would be replaced by $\zeta+1$. This is the reason why, for $\lambda_- \in [2/d, 1]$, we rely on a slight variant $\hat{\pi}_{WM-SR}$ (see Section 3.5.7) of $\hat{\pi}_{WM}$ that builds upon the fact that the variations that are due to the aggregation of M are very specific.

Theorem 3.4.3. *There exist four numerical constants c_1–c_4 such that the following holds. Fix $\delta = \zeta_-^2[(\lambda \vee 1)nd]^{-2}$. For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{BISO}$, the sorting tree estimator $\hat{\pi}_{WMP}$ defined above satisfies*

$$\mathbb{E}\Big[\|M_{\hat{\pi}_{WMP}^{-1}} - M_{\pi^{*-1}}\|_F^2\Big] \le c_1 \log^{c_2}\Big(\tfrac{nd(\lambda\vee 1)}{\zeta_-}\Big)\Big[\mathcal{R}_F(n,d,\zeta\lambda^{-1/2}) + \frac{n}{\lambda}e^{-c_3\lambda\log^{-c_4}(\frac{nd(\lambda\vee 1)}{\zeta_-})}\Big] \,. \tag{3.28}$$

Up to logarithmic terms and up to the logarithmic term inside the exponential term (3.28), both the minimax upper bound (3.28) and lower bound (3.27) match for all values of n, d, λ and ζ. As a consequence, this problem of estimating a single permutation π^* does not exhibit any significant computational gap.

Let us further discuss and compare the exponential term $n\lambda^{-1}e^{-c_3\lambda\log^{-c_4}(\frac{nd(\lambda\vee 1)}{\zeta_-})}$ in (3.27) and $n\lambda^{-1}e^{-2\lambda}$ in (3.28). First, observe that these two terms are larger than $R_F(n,d,\zeta\lambda^{-1/2})$ only when the noise level ζ is small, so that it is relevant to discuss them only when $\zeta \ll 1$. Second, note that there is a significant mismatch between these exponential terms only when λ is close to one, up to a polylogarithmic factor, since otherwise, either the exponential is close to one (for $\lambda \le 1$) or the exponential is so small that it becomes negligible in comparison to $R_F(n,d,\zeta\lambda^{-1/2})$. One may object that the logarithmic term $\log^{-c_4}(\frac{nd(\lambda\vee 1)}{\zeta_-})$ may be large in case ζ is really small –think e.g. of $\zeta = e^{-nd}$. Let us consider this extremely low noise setting where, say $\zeta \le 1/(nd)^2$. If one applies the procedure $\hat{\pi}_{WMP}$ with $\zeta_0 = 1/(nd)^2 \ge \zeta$, then the logarithmic terms become bounded inside the exponential. Since $\mathcal{R}_F(n,d,\zeta_0\lambda^{-1/2})$ is always smaller than $nd \wedge \frac{n}{\lambda}e^{-c_3\lambda}$ provided that $\lambda \le 1$, this estimator achieves the risk bound $\frac{n}{\lambda} \wedge nd$, which is optimal for all $\zeta \in [0, 1/(nd)^2]$ and all $\lambda \le 1$. To sum up, there is gap between our minimax lower and upper bounds only either (i) in the low-noise level with large but mild sampling effort, that is $\zeta = o(\log^{-c}(nd))$, $\zeta \ge (nd)^{-2}$, and $\lambda \in [\log^c\log(nd), \log^{c'}(nd)]$ for some c and $c' > 0$ or (ii) in the extremely low noise level with large sampling effort, that is $\zeta \le (nd)^{-2}$ and $\lambda \ge 1$.

In $\hat{\pi}_{WMP}$, we have plugged in the hierarchical sorting tree estimator with memory $\hat{\pi}_{WM}$. If we had plugged in the oblivious hierarchical sorting tree estimator $\hat{\pi}_{HT}$, then the resulting estimator would satisfy a similar rate similar to (3.28) except that the term $n^{3/4}d^{1/4}/\lambda^{3/4}$ would be replaced by the slower rate $n^{2/3}d^{1/3}/\lambda^{2/3}$.

3.4.3 Reconstruction of the matrix M

In this subsection, we assume again that the noise level $\zeta = 1$ to simplify the exposition. As alluded in Section 3.2, it is quite straightforward to estimate the matrix M and control the corresponding loss $\|\widehat{M} - M\|_F^2$ by a simple subsampling step explained e.g. in [60] that we recall here. First, we split the sample into two part by assigning independently each observation to the first subsample with probability 1/2 and the second subsample with probability 1/2. Then, we use the first subsample to estimate the permutation $\hat{\pi}$ of the experts. As for the second subsample $(x_t^{(2)}, y_t^{(2)})$, we define the empirical observed matrix $Y^{(2)}$ by

$$Y^{(2)}_{i,j} = \frac{1}{\lambda}\sum_t y_t^{(2)}\mathbf{1}_{x_t^{(2)}=(i,j)}.$$

Then, we compute the least-square estimator $\widehat{M}_{\hat{\pi}}$ of $M_{\hat{\pi}}$ in the class of bi-isotonic matrix $\widehat{M}_{\hat{\pi}} = \arg\min_{B\in\mathbb{C}_{\mathrm{BISO}}}\|B - Y^{(2)}_{\hat{\pi}}\|_F^2$. This estimator can be computed in near linear-time [53]. Then, Proposition 3.3 in [60] states, that with high probability, the loss $\|\widehat{M} - M\|_F^2$ is, up to logarithmic terms, smaller than the sum of the minimax risk for estimating a bi-isotonic matrix B and the loss $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$. Plugging this proposition with our estimator $\hat{\pi}_{WMP}$ with $\delta = (\lambda\vee 1)/(np)$, we readily arrive to the following risk bound for the corresponding estimator $\widehat{M}_{WMP}$.

Define $\mathcal{R}_1(n,d,\lambda) = \sqrt{\frac{nd}{\lambda}} \bigwedge \frac{nd}{\lambda^{2/3}(n\vee d)^{2/3}} \bigwedge \frac{nd}{\lambda}$. Mao et al. [60] have proved that, up to polylogarithmic factor and up to a possible additive term $(n \wedge d)/\lambda$, the minimax risk in square Frobenius norm for estimating a bi-isotonic matrix with partial observations is $\mathcal{R}_1(n,d,\lambda)$.

Corollary 3.4.4. *There exist two numerical constants c and c' such that the following holds. For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^*} \in \mathbb{C}_{BISO}$, we have*

$$\begin{aligned}\mathbb{E}\left[\|\widehat{M}_{WMP} - M\|_F^2\right] &\leq (nd) \bigwedge \left[c \log^{c'}((\lambda \vee 1)nd)\left(\mathcal{R}_1(n,d,\lambda) + \mathcal{R}_F(n,d,\lambda^{-1/2})\right)\right] \\ &\leq (nd) \bigwedge \left[c \log^{c'}((\lambda \vee 1)nd)\left(\mathcal{R}_1(n,d,\lambda) + \frac{n}{\lambda}\right)\right] . \end{aligned} \tag{3.29}$$

The proof is a straightforward consequence of Proposition 3.3 in [60] and Theorem 3.4.3 and is therefore omitted. It turns out that $\mathcal{R}_F(n,d,\lambda^{-1/2})$ is always smaller than $\mathcal{R}_1(n,d,\lambda)+\frac{n}{\lambda}$, so that the cost of reconstruction for not knowing π^* is n/λ.

This risk bound (3.29) is minimax optimal, up to polylogarithms, and this for all possible values of $n \geq 2$, d, and $\lambda > 0$. Indeed, in their Theorem 3.1, Mao et al. [60] provide a matching minimax lower bound in $\mathcal{R}_1(n,d,\lambda)$ in the specific case where $n \geq d$, but their proof easily extends to the case where $n \leq d$. Besides, our proof of the minimax lower bound $\frac{n}{\lambda}$ in Theorem 3.4.1 for the problem of estimating π^* straightforwardly extends to the problem of matrix estimation (recall that we consider $\zeta = 1$ here).

The least-square estimator $\hat{\pi}_{LS}$ of Mao et al. has also been proved to achieve the minimax risk for $n \geq d$ –see their theorem 3.1 in [60]. However, no efficient algorithm is known for computing this estimator in $\hat{\pi}_{LS}$, so that our estimator $\widehat{M}_{WMP}$ is, to the best of our knowledge, the first efficient minimax-optimal estimator for estimating M in this context, for any values of n, d, λ.

3.4.4 Bounds for the max loss of Mao et al. [60]

In [60], Mao et al. control, for an estimator $\hat{\pi}$ of the permutation, a different loss from ours. Up to normalization factors, they indeed focus on the maximum l_2 norm of the rows of $(M_{\hat{\pi}^{-1}})_{i,\cdot} - (M_{\pi^{*-1}})_{i,\cdot}$, that is

$$l_\infty(\hat{\pi}, \pi^*) = \sup_{i \in [n]} \|(M_{\hat{\pi}^{-1}})_{i,\cdot} - (M_{\pi^{*-1}})_{i,\cdot}\|_2^2 . \tag{3.30}$$

This loss also considered in [84, 19] corresponds to some maximum error of the estimated permutation so that $l_\infty(\hat{\pi}, \pi^*) \geq \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2/n$. Alternatively, we can define the loss l_{err}

$$l_{err}(\hat{\pi}, \pi^*) = \max_{\substack{i,j \in [n] : \\ \hat{\pi}(i) < \hat{\pi}(j) \text{ and } \pi^*(i) > \pi^*(j)}} \|M_{i,\cdot} - M_{j,\cdot}\|_2^2 ,$$

which quantifies the maximum distance between two experts that have not been ranked consistently. The loss l_∞ and l_{err} turn out to be equivalent as stated in the following lemma.

Lemma 3.4.5. *For any permutation $\hat{\pi}$, we have*

$$l_\infty(\hat{\pi}, \pi^*) \leq l_{err}(\hat{\pi}, \pi^*) \leq 4 l_\infty(\hat{\pi}, \pi^*) \tag{3.31}$$

To simplify the discussion in this section, we assume again that the noise level ζ equals one. Mao et al. [60] provide a simple polynomial time $\hat{\pi}_{\text{ref}}$ achieving

$$\mathbb{E}[l_\infty(\hat{\pi}_{\text{ref}}, \pi^*)] \lesssim d \bigwedge \frac{d^{1/4}}{\lambda^{3/4}} \log^{3/4}(n) . \tag{3.32}$$

Conversely, they prove in their Theorem 3.7 that any estimator $\hat{\pi}$ that only ranks the experts i and j according to the differences of the observations on the rows i and j must incur this risk bound– see [60] for further details. Besides, they conjecture that the risk bound (3.32) cannot be improved. In [56], Liu and Moitra already pointed out that the max loss $l_\infty(\hat{\pi}, \pi^*)$ is less suited than the loss $\|M_{\hat{\pi}} - M_{\pi^*}\|_F^2$ for the purpose of estimating the matrix M –see the discussion in the previous subsection. Still, controlling the max loss $l_\infty(\hat{\pi}, \pi^*)$ may be an objective per se, and the study of its minimax value and of the existence of related minimax estimators is relevant. In the following proposition, which is mainly a consequence of our results and proof techniques, we disprove Mao et al.'s conjecture by introducing an estimator $\hat{\pi}_{PC}$ achieving a faster rate than (3.32). Besides, this rate turns out to be minimax-optimal.

Proposition 3.4.6. *There exist numerical constants c, c', and c'' such that the following result holds. There exists a polynomial-time estimator $\hat{\pi}_{PC}$ that performs pair-wise comparisons between the experts and that achieves the risk bound*

$$\mathbb{E}[l_\infty(\hat{\pi}_{PC}, \pi^*)] \le c \log^{c'}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} \right] \bigwedge d \ . \tag{3.33}$$

Conversely, for any $n \ge 2$, any $d \ge 1$, and $\lambda > 0$, we have

$$\inf_{\hat{\pi}} \sup_{\pi^* \in \Pi_n} \sup_{M : M_{\pi^{*-1}} \in \mathbb{C}_{BISO}} \mathbb{E}_{(\pi^*, M)}[l_\infty(\hat{\pi}, \pi^*)] \ge c'' \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} \bigwedge d \right] . \tag{3.34}$$

For $\lambda \le 1/d$, it is not possible to perform significantly better than random guess. Then, in the interesting regime $\lambda \in [1/d, d^2]$, the risk is of the order of $\frac{d^{1/6}}{\lambda^{5/6}}$. It turns out that this rate corresponds, up to polylogarithmic terms, to the minimal distance between two experts so that one is able to consistently compare them. For very large sample size $\lambda \ge d^2$, we arrive at the easy regime which is of the order of $\frac{\sqrt{d}}{\lambda}$.

The estimator $\hat{\pi}_{PC}$ is based on pairwise comparisons. For any two experts i and j, we apply the procedure $\hat{\pi}_{WMP}$ to i and j with $\delta = [(\lambda \vee 1)(n^2 d)]^{-2}$. If the trisection (O, P, I) is of the form $(\varnothing, \{i\}, \{j\})$, we return $i \prec j$. If the trisection (O, P, I) is of the form $(\varnothing, \{j\}, \{i\})$, we return $j \prec i$. Otherwise, we return nothing. Applying this comparison algorithm to all (i, j), we recover a set of pairwise comparisons $\mathcal{PC} = \{(i, j) : i \prec j\}$. With high probability –see the proof for more details–, it turns that $\mathcal{PC}$ satisfies two properties:

(i) $\mathcal{PC}$ is consistent. For any $(i, j) \in \mathcal{PC}$, we have $\pi^*(i) < \pi^*(j)$.

(ii) $\mathcal{PC}$ contains all 2-tuple of experts that are far apart. More precisely, $\mathcal{PC}$ contains all (i, j) such that $\pi^*(i) < \pi^*(j)$, and

$$\|M_{i,\cdot} - M_{j,\cdot}\|_2^2 \ge c \log^{c'}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} \right] \bigwedge d \ , \tag{3.35}$$

for suitable constants c and c'.

Then, define the function $\phi : [n] \mapsto \mathbb{N}$ by $\phi(j) = |\{(i, j) : (i, j) \in \mathcal{PC}\}|$ which simply counts the number of experts i that are detected to be lower than j. Finally, we build $\hat{\pi}_{PC}$ as any permutation that ranks the experts consistently with ϕ.

In fact, the procedure for computing $\hat{\pi}_{PC}$ could be greatly simplified. Indeed, as we only perform pairwise comparisons, some parts of **BlockSort** turn out to be irrelevant. For instance, the PCA steps are not required. Besides, the sample splits could be avoided, and it could even be possible to work with a single observation. As the problem of optimal permutation recovery with respect to the l_∞ loss is not the main scope of this chapter, we do not provide a simplified and dedicated algorithm. Besides, we conjecture that our original estimator $\hat{\pi}_{WMP}$ also achieves the minimax risk (3.33) with respect to the l_∞ loss.

3.4.5 Full description of the procedures

In this section, we provide a fuller description of the estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$ as a collection of algorithms. We will rely on this description in the analysis of these estimators. To ease its understanding, we make this section completely self-contained. As a consequence, the material presented here is partly redundant with Section 3.3.

3.4.6 Sorting a group of experts

Some of the notation have already been introduced in Section 3.3. Still we define them again here for the sake of completeness. We write $\mathcal{D}$ for the set of all dyadic numbers, that is $\mathcal{D} = \{2^k : k \in \mathbb{Z}\}$. Equipped with $\mathcal{D}$, let

$$\mathcal{R} = \mathcal{D} \cap [1, d] \quad \text{and} \quad \mathcal{H} = \mathcal{D} \cap \left[\frac{\zeta^2}{nd}, 1 \right] \ , \tag{3.36}$$

respectively denote the dyadic collection of numbers beween 1 and d and the dyadic collection of numbers between $1/nd$ and 1.

Besides for an integer $r \in \mathcal{R}$, we write $\mathcal{Q}_r$ for the regular grid of $[d]$ of width r:

$$\mathcal{Q}_r = \left\{ 1, r+1, 2r+1, \dots \left\lfloor \frac{d}{r} \right\rfloor r + 1 \right\} \ .$$

In contrast to Section 3.3, we start by describing the simple comparison routine before moving to the dimension reduction techniques and to the general architecture of the procedures.

Given a collection $\overline{P}$ of experts, some data $Z \in \mathbb{R}^{\overline{P}\times Q}$ and a direction $w \in (\mathbb{R}^+)^Q$ and a pivot $\gamma \in [1 : |\overline{P}|]$, the following pivoting algorithm sorts the experts in $\overline{P}$ according to the projection of the data onto the vector w. More precisely, it returns four subsets $\overline{L} \subset L$ and $\overline{U} \subset U$ of experts such that the γ-th best expert according to the (Z, w)-order - as defined above Equation (3.15) - is significantly above all experts in L and below all experts in U. The subsets L and $\overline{L}$ (resp. U and $\overline{U}$) differ in the level of significance we require. We define the tuning parameters β_{tris} and $\overline{\beta}_{\text{tris}}$ for **Pivot**

$$\beta_{\text{tris}} = 4\sqrt{2}\zeta \ , \qquad \overline{\beta}_{\text{tris}} = 8\sqrt{2}\zeta. \tag{3.37}$$

Algorithm 4 Pivot(Z, w, γ)

Require: A matrix $Z \in \mathbb{R}^{\overline{P}\times Q}$ with $\overline{P} \subset [n]$ a set of experts and $Q \subset [d]$ a set of blocks, a direction $w \in \mathbb{R}_+^Q$, $w \neq 0$ and a pivot index γ
Ensure: Two couples of subsets (L, U) and $(\overline{L}, \overline{U})$ of $\overline{P}$

1: **for** $i \in \overline{P}$ **do**
2: Compute the statistic $\psi(i, w) = \langle Z_{i,\cdot}, \frac{w}{\|w\|_2} \rangle$
3: **end for**
4: Sort the statistics $\psi(i, w) : \psi(i_1, w) \leq \cdots \leq \psi(i_{|\overline{P}|}, w)$
5: $U = \{i \in \overline{P} \ : \ \psi(i, w) > \psi(i_\gamma, w) + \beta_{\text{tris}}\sqrt{\log\left(2\frac{|\overline{P}|}{\delta}\right)}\}$
6: $\overline{U} = \{i \in \overline{P} \ : \ \psi(i, w) > \psi(i_\gamma, w) + \overline{\beta}_{\text{tris}}\sqrt{\log\left(2\frac{|\overline{P}|}{\delta}\right)}\}$
7: $L = \{i \in \overline{P} \ : \ \psi(i, w) < \psi(i_\gamma, w) - \beta_{\text{tris}}\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)}\}$
8: $\overline{L} = \{i \in \overline{P} \ : \ \psi(i, w) < \psi(i_\gamma, w) - \overline{\beta}_{\text{tris}}\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)}\}$
9: **return** (L, U), $(\overline{L}, \overline{U})$

When the vector w is equal to $\mathbf{1}_Q$, we simply write **Pivot**(Z, γ) instead of **Pivot**$(Z, \mathbf{1}_Q, \gamma)$ for the sake of simplicity.

$$\textbf{Pivot}(Z, \gamma) = \textbf{Pivot}(Z, w = \mathbf{1}_Q, \gamma) \ . \tag{3.38}$$

In fact, **Pivot**(Z, γ) simply amounts to comparing the row sums of Z for each of the experts in P.

In the next two pages, we redefine in more detail the Double Trisection algorithm of Section 3.3. First, **DoubleTrisection – PCA** relies on a PCA-type argument to find a suitable direction $\hat{w}^+$ and then provides two trisections of the subset $\overline{P}$ of experts using the **Pivot** sub-routine.

Algorithm 5 DoubleTrisection – PCA$(\mathcal{Z}, \gamma)$

Require: 4 reduced samples $\mathcal{Z} = (Z^{(1)}, Z^{(2)}, Z^{(3)}, Z^{(4)})$ where $Z^{(1)}$, $Z^{(2)}$, $Z^{(3)} \in \mathbb{R}^{\widetilde{P}\times Q}$ and $Z^{(4)} \in \mathbb{R}^{\overline{P}\times Q}$ with some $\widetilde{P} \subset \overline{P}$, and a pivot index γ
Ensure: Four subsets $(L_{\text{pca}}, U_{\text{pca}})$ and $(\overline{L}_{\text{pca}}, \overline{U}_{\text{pca}})$ of $\overline{P}$

1: Compute the following vector with coefficients in $\widetilde{P}$:

$$\hat{v} = \underset{\|v\|\leq 1}{\arg\max}\Big[\|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2\Big] \in \mathbb{R}^{\widetilde{P}}$$

2: $\hat{z} = v^T Z^{(3)} \in \mathbb{R}^Q$
3: Define $\hat{w}^+$ by $(\hat{w}^+)_l = |\hat{z}_l|\mathbf{1}_{|\hat{z}_l|\geq 2\zeta\sqrt{2\log(2|Q|/\delta)}}$
4: $(L_{\text{pca}}, U_{\text{pca}}), (\overline{L}_{\text{pca}}, \overline{U}_{\text{pca}}) = \textbf{Pivot}(Z^{(4)}, \hat{w}^+, \gamma)$
5: **return** $(L_{\text{pca}}, U_{\text{pca}})$, $(\overline{L}_{\text{pca}}, \overline{U}_{\text{pca}})$

Next, **DoubleTrisection – Local**$(\mathcal{Z}, \gamma)$ builds two trisections of $\overline{P}$ based on the reduced samples. First, it builds these trisections by simply using the row sums on the data and then it improves them thanks to **DoubleTrisection – PCA**.

Algorithm 6 DoubleTrisection – Local$(\mathcal{Z}, \gamma)$

Require: 5 reduced samples $\mathcal{Z} = (Z^{(1)}, Z^{(2)}, Z^{(3)}, Z^{(4)}, Z^{(5)})$ in $\mathbb{R}^{\overline{P}\times Q}$, a pivot index γ and a threshold β_{tris}
Ensure: Two couples of subsets (L, U) and $(\overline{L}, \overline{U})$ of $\overline{P}$

1: $(L_{\text{cp}}, U_{\text{cp}}), (\overline{L}_{\text{cp}}, \overline{U}_{\text{cp}}) = \textbf{Pivot}(Z^{(1)}, \mathbf{1}_Q, \gamma)$
2: Set $\widetilde{P} = \overline{P} \setminus (\overline{L}_{\text{cp}} \cup \overline{U}_{\text{cp}})$
3: Set $\mathcal{Z}' = (Z^{(2)}(\widetilde{P}), Z^{(3)}(\widetilde{P}), Z^{(4)}(\widetilde{P}), Z^{(5)}(\overline{P}))$ be the sequence of reduced samples where the three first samples are restricted to $\widetilde{P}$.
4: $(L_{\text{pca}}, U_{\text{pca}}), (\overline{L}_{\text{pca}}, \overline{U}_{\text{pca}}) = \textbf{DoubleTrisection – PCA}(\mathcal{Z}', \gamma)$
5: Set $L = L_{\text{cp}} \cup L_{\text{pca}}$ and $\overline{L} = \overline{L}_{\text{cp}} \cup \overline{L}_{\text{pca}}$ and $U = U_{\text{cp}} \cup U_{\text{pca}}$ and $\overline{U} = \overline{U}_{\text{cp}} \cup \overline{U}_{\text{pca}}$
6: **return** $(L, U), (\overline{L}, \overline{U})$

To finish defining **DoubleTrisection**, we simply need to plug a dimension reduction procedure to select a subset of questions $Q \subset [d]$ and then to sum the data on these questions.

The two following algorithms are mainly definitions. For some data $Y \in \mathbb{R}^{[n]\times[d]}$, a set of experts $\overline{P}$ and a set of blocks $Q \subset \mathcal{Q}_r$, and a scale r, the $|\overline{P}| \times |Q|$ matrix $\textbf{Encode – Matrix}(Y, \overline{P}, Q, r)$ is simply a reduced matrix where we consider the normalized row sums of Y around the questions of Q at scale r.

Algorithm 7 Encode – Matrix$(Y, \overline{P}, Q, r)$

Require: A matrix $Y \in \mathbb{R}^{[n]\times[d]}$, a set of experts $\overline{P}$ and a set of blocks $Q \subset \mathcal{Q}_r$, a scale r
Ensure: A reduced matrix $Z \in \mathbb{R}^{\overline{P}\times Q}$

1: **for** $i \in \overline{P}$ and $l \in Q$ **do**
2: Define $Z_{i,l} = \frac{1}{\sqrt{r}} \sum_{k\in[l,l+r)} Y_{i,k}$ ▷ $Y_{i,k} = 1$ for $k \geq d+1$
3: **end for**
4: **return** $Z \in \mathbb{R}^{\overline{P}\times Q}$ ▷ the restriction of Z to $\overline{P}$ and Q

Second, $\textbf{Encode – Set}(D, r)$ transforms a subset $[d]$ of questions into a subset $Q \subset \mathcal{Q}_r$ of blocks of questions at scale r.

Algorithm 8 Encode – Set(D, r)

Require: A set of questions $D \subset [d]$, a scale $r \in \mathcal{R}$
Ensure: A set of blocks $Q \subset \mathcal{Q}_r$
return $Q = \{l \in \mathcal{Q}_r \ : \ [l, l+r) \cap D \neq \varnothing\}$

Then, we are in position to redefine this version of Algorithm 3. As in the original definition in Section 3.3, there are two variations of this procedure depending on whether we are building the estimator $\hat{\pi}_{HT}$ or the estimator $\hat{\pi}_{WM}$ that uses the memory of the tree. Algorithm $\textbf{DoubleTrisection}(\mathcal{Y}, \mathcal{T}, \overline{P}, \gamma)$ takes some original data and then reduces the dimension of the problem to build two trisections of the set $\overline{P}$ of experts.

Algorithm 9 DoubleTrisection$(\mathcal{Y}, \mathcal{T}, \overline{P}, \gamma)$

Require: 6 samples $\mathcal{Y} = (Y^{(1)}, \dots, Y^{(6)})$, a tree $\mathcal{T}$, a set of expert $\overline{P}$ included in a leaf G of $\mathcal{T}$ at maximal depth and a pivot index γ
Ensure: Two couples of subsets (L, U), $(\overline{L}, \overline{U})$ of $\overline{P}$

Initialize $L, U, \overline{L}, \overline{U} = \varnothing$
for $h \in \mathcal{H}, r \in \mathcal{R}$ **do**
  if Not using the memory of the tree **then**
    Set $\widehat{Q} := \widehat{Q}_{\mathrm{cp}}(h, r) =$ **DimensionReduction**$(Y^{(1)}, \overline{P}, h, r)$ - see Algorithm 12 or (3.20)
  else if Using the memory of the tree **then**
    Set $\widehat{Q} := \widehat{Q}_{WM}(h, r) =$ **DimensionReduction – WM**$(Y^{(1)}, \mathcal{T}, \overline{P}, h, r)$ - See Algorithm 13 or (3.26)
  end if
  Consider the five samples $\mathcal{Y}' = (Y^{(2)}, Y^{(3)}, Y^{(4)}, Y^{(5)}, Y^{(6)})$
  Consider the five reduced samples $\mathcal{Z} =$ **Encode – Matrix**$(\mathcal{Y}', \overline{P}, \widehat{Q}_{WM}, r)$
  Compute $(L_{\mathrm{loc}}, U_{\mathrm{loc}}), (\overline{L}_{\mathrm{loc}}, \overline{U}_{\mathrm{loc}}) =$ **DoubleTrisection – Local**$(\mathcal{Z}, \gamma)$
  Update $L = L \cup L_{\mathrm{loc}}$ and $\overline{L} = \overline{L} \cup \overline{L}_{\mathrm{loc}}$ and $U = U \cup U_{\mathrm{loc}}$ and $\overline{U} = \overline{U} \cup \overline{U}_{\mathrm{loc}}$
end for
return $(L, U), (\overline{L}, \overline{U})$

Finally, we reproduce **BlockSort** here that was originally defined in Algorithm 2. We recall that **BlockSort** iteratively applies a logarithmic number of times the procedure **DoubleTrisection** to build two suitable trisections of a set G of experts. Although implicit in this description, there are two different versions of the corresponding procedure whether we use the memory of the tree - estimator $\hat{\pi}_{WM}$ - or not - estimator $\hat{\pi}_{HT}$ in **DoubleTrisection**. In the following, $\tau_\infty = \lceil 4 \cdot 10^7 \log^7(\frac{nd}{\delta(\zeta_-)^2}) \rceil$ stands for the number of iterations in **BlockSort**.

Algorithm 10 BlockSort$(\mathcal{Y}, \mathcal{T}, G)$

Require: $6\tau_\infty$ samples $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(6\tau_\infty - 1)})$, the tree $\mathcal{T}$, a leaf $G \subset [n]$ in $\mathcal{T}$ at maximal depth
Ensure: A partition of G into three groups (O, P, I)

Set $\gamma = \lfloor |G|/2 \rfloor$ and O_0, I_0, $\overline{O}_0$, $\overline{I}_0 = \varnothing$
for $\tau = 0, \dots, \tau_\infty - 1$ **do**
  Consider 6 fresh samples $\mathcal{Y}_\tau = (Y^{(6\tau)}, \dots, Y^{(6\tau+5)})$
  set $\gamma = \lfloor |G|/2 \rfloor - |\overline{O}_\tau|$
  (L_τ, U_τ), $(\overline{L}_\tau, \overline{U}_\tau) =$ **DoubleTrisection**$(\mathcal{Y}_\tau, \mathcal{T}, G \setminus (\overline{O}_\tau \cup \overline{I}_\tau), \gamma)$ as in Algorithm 9
  Update $O_{\tau+1} = O_\tau \cup L_\tau$, $I_{\tau+1} = I_\tau \cup U_\tau$, $\overline{O}_{\tau+1} = \overline{O}_\tau \cup \overline{L}_\tau$, $\overline{I}_{\tau+1} = \overline{I}_\tau \cup \overline{U}_\tau$
end for
if $O_{\tau_\infty} \cap I_{\tau_\infty} \neq \varnothing$ **then**
  Set $O_{\tau_\infty} := O_{\tau_\infty} \setminus I_{\tau_\infty}$ and $I_{\tau_\infty} := I_{\tau_\infty} \setminus O_{\tau_\infty}$
end if
return $(O_{\tau_\infty}, G \setminus (O_{\tau_\infty} \cup I_{\tau_\infty}), I_{\tau_\infty})$

Under an event of high probability (to be later discussed), we have $O_{\tau_\infty} \cap I_{\tau_\infty} = \varnothing$. The correction at the end of the algorithm simply forces the algorithm to return a partition of G.

3.4.7 Hierarchical sorting trees and TreeSort algorithm

In this subsection, we formally describe how we build and navigate into a hierarchical tree. In the following, a node $G \in$ **Nodes** is a labelled subset of $[n]$. Its label belongs to $\{\mathbf{0}, \mathbf{p}, \mathbf{1}\}$. For a node G, we write **Type**(G) for the label (also called type) of G.

Definition 1. (Hierarchical sorting Trees) A hierarchical sorting tree $\mathcal{T}$ is a rooted tree that satisfies the three following properties:

- The root G of $\mathcal{T}$ corresponds to the set $[n]$ and its label is $\mathbf{0}$.
- Any node G of type $\mathbf{p}$ is a leaf.
- Any node G of type in $\{\mathbf{0}, \mathbf{1}\}$ is either a leaf or has three children (O, P, I) with type $\mathbf{0}$, $\mathbf{p}$, $\mathbf{1}$ respectively. Besides, (O, P, I) correspond to a partition of G.

We write $\mathcal{T}_0$ for the tree of depth 0. The procedure **TreeSort** iteratively builds a hierarchical sorting tree. Hence, we need to define the operation of adding children to a leaf in a tree $\mathcal{T}$. For a specific leaf G of type **0** or **1**, we consider three labelled subsets O, P, I of type **0**, **p**, **1**, respectively. Besides, those subsets satisfy the third condition in Definition 1. Then, $\mathcal{T}' = \mathbf{AddChild}(\mathcal{T}, G, (O, P, I))$ is the supertree of $\mathcal{T}$ where we have added the nodes (O, P, I) as children of G. Finally, we observe that for any $t > 0$, all the nodes at depth t of a hierarchical sorting tree $\mathcal{T}$ are disjoint.

In fact, we shall prove in Proposition 3.5.1 and in Corollary 3.5.4 that, with high probability, the final tree $\mathcal{T}_{t_\infty}$ turns out to be a valid hierarchical sorting tree as defined below.

Definition 2. (Valid hierarchical sorting Tree) A hierarchical sorting tree $\mathcal{T}$ is valid if non-terminal nodes G of $\mathcal{T}$ satisfy the two following additional properties: if we denote (O, P, I) their children of type **0**, **p**, **1** respectively, then

- All the experts in O are below those of I. In other words, for any $i \in O$ and any $j \in I$, we have $\pi^*(i) < \pi^*(j)$.
- $|O| < |G|$ and $|I| < |G|$.

The second property ($|O| < |G|$ and $|I| < |G|$) forces the tree to be finite.

For a node G in a such valid hierarchical sorting tree $\mathcal{T}$, $\mathbf{Depth}(\mathcal{T}, G)$ stands for the depth of G in $\mathcal{T}$. In light of this definition of valid hierarchical sorting trees, a labelled subset G cannot appear twice in a tree $\mathcal{T}$, so that $\mathbf{Depth}(\mathcal{T}, G)$ is well-defined.

We are now equipped to provide a more formal definition of **TreeSort**, although the procedure is in fact the same as the one described in Algorithm 1. Let $t_\infty = \lceil \log(n)/\log(2) \rceil$.

Algorithm 11 $\mathbf{TreeSort}(\mathcal{Y})$

Require: $6\tau_\infty t_\infty$ samples $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(6\tau_\infty t_\infty - 1)})$
Ensure: A final tree $\mathcal{T}$

$\mathcal{T} = \mathcal{T}_0$ ▷ The root is at depth 0
for $t = 0, \dots, t_\infty - 1$ **do**
  Consider $6\tau_\infty$ fresh samples $\mathcal{Y} = (Y^{(6t\tau_\infty)}, \dots, Y^{(6(t+1)\tau_\infty - 1)})$
  for $G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ **do** ▷ See (3.39) for the definition of $\mathcal{L}^{(\mathbf{0},\mathbf{1})}$
    $(O_G, P_G, I_G) = \mathbf{BlockSort}(\mathcal{Y}, \mathcal{T}, G)$
    Set $\mathbf{Type}(O_G) = \mathbf{0}$ and $\mathbf{Type}(P_G) = \mathbf{p}$ and $\mathbf{Type}(I_G) = \mathbf{1}$
  end for
  for $G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ **do**
    $\mathbf{AddChild}(\mathcal{T}, G, (O_G, P_G, I_G))$
  end for
end for
return $\mathcal{T}$

As explained in Section 3.3, the final estimators $\hat{\pi}_{HT}$ or $\hat{\pi}_{WM}$ are computed from their corresponding hierarchical sorting tree $\mathcal{T}$.

In order to define the **DimensionReduction – WM** algorithm in the next subsection, we need to introduce a few more notation. First, we define

$$\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T}) = \{G \in \mathbf{Leaves}(\mathcal{T}) \ : \ \mathbf{Type}(G) \in \{\mathbf{0}, \mathbf{1}\}\} \ , \tag{3.39}$$

as the collection of leaves of $\mathcal{T}$ that are either of type **0** or of type **1**. In the algorithm **TreeSort**, these are the leaves to be partitionned. In particular at step t of **TreeSort**, $\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ is only made of leaves at depth t.

For a subset $P \subset [n]$, $\mathbf{Leaf}(\mathcal{T}, P)$ is defined as the leaf $G \in \mathbf{Leaves}(\mathcal{T})$ containing P (if it exists). Finally, the groups $G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ inherit from a natural order provided that $\mathcal{T}$ is a valid hierarchical sorting tree. We can enumerate the groups $G_1, G_2, \dots, G_{|\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})|}$ in such a way that all the experts in G_s are below those of $G_{s'}$ for $s < s'$. To ease the presentation, we also introduce, for any positive integer s the groups $G_{|\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})|+s} = \{n + s\}$. The corresponding data and signal for the $n + s$-th expert satisfies $Y_{n+s,j} = 1 = M_{n+s,j}$ almost-surely for any $j \in [d]$. Also, for any positive integer s we introduce the groups $G_{1-s} = \{1 - s\}$. The corresponding data and signal for this synthetic expert satisfy $Y_{1-s,j} = 0 = M_{1-s,j} = 0$ almost-surely for any $j \in [d]$.

Then, for a specific leaf $G_s \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$, $\mathbf{Order}(\mathcal{T}, G)$ stands for the collection $(G^{(a)})$, $a \in \mathbb{Z}$ of leaves where $G^{(a)} = G_{a+s}$. In other words, we have $G^{(0)} = G_s$ and $G^{(1)}$ is the following group, and so on.

3.4.8 Dimension reduction algorithms

To finish the description of the two procedures, we fully describe the two dimension reduction algorithms both for the oblivious estimator $\hat{\pi}_{HT}$ and for the estimator $\hat{\pi}_{WM}$ with memory. These procedures were already introduced in Section 3.3. First, **DimensionReduction**$(Y, \overline{P}, h, r)$ considers the columns-wise mean of the restriction of Y to the group $\overline{P}$ and detects high-variation regions of this vector.

Algorithm 12 DimensionReduction$(Y, \overline{P}, h, r)$

Require: A sample $Y \in \mathbb{R}^{\overline{P} \times [d]}$, a set of experts $\overline{P}$, $h \in \mathcal{H}$ and r in $\mathcal{R}$
Ensure: An encoded set including the high-variation regions $\widehat{Q}_{\mathrm{cp}} := \widehat{Q}_{\mathrm{cp}}(Y, \overline{P}, h, r) \subset \mathcal{Q}_r$

1: $\overline{y}(\overline{P}) = \frac{1}{|\overline{P}|} \sum_{i \in \overline{P}} Y_{i,\cdot}$
2: $\tilde{r} = 8\left\lceil \left\lceil \left(\frac{32\zeta^2}{|\overline{P}|h^2} \log\left(\frac{2d}{\delta}\right)\right)\right\rceil \vee r \right\rceil$
3: Initialize $\widehat{D}_{\mathrm{cp}} = \varnothing$
4: **for** $k \in [d]$ **do**
5: Compute

$$\widehat{\mathbf{C}}_k(\overline{y}(\overline{P})) = \frac{1}{\tilde{r}} \left(\sum_{k'=k}^{k+\tilde{r}-1} \overline{y}_{k'}(\overline{P}) - \sum_{k'=k-\tilde{r}}^{k-1} \overline{y}_{k'}(\overline{P}) \right) \; ; \tag{3.40}$$

6: **end for**
7: $\widehat{D}_{\mathrm{cp}} = \{k \in [d] \; : \; \widehat{\mathbf{C}}_k(\overline{y}(\overline{P})) \geq h/4\}$
8: $\widehat{Q}_{\mathrm{cp}} =$ **Encode – Set**$(\widehat{D}_{\mathrm{cp}}, r)$
9: **return** $\widehat{Q}_{\mathrm{cp}}$

For the more involved dimension reduction procedure with memory **DimensionReduction – WM**, we compute the CUSUM statistic in larger groups $\mathcal{V} \supset \overline{P}$ to reduce its variance and we also require that the estimated "width" of the group of experts is high enough. More precisely, given three sets of expert $\mathcal{V}$, $\mathcal{V}^+$ and $\mathcal{V}^-$ and a sample Y, we consider the two following statistics, for any $k = 1, \ldots, d$ and $r' \in \mathcal{R}$:

$$\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r'}(\mathcal{V}^+, \mathcal{V}^-) = \frac{1}{2r'} \sum_{k'=k-r'}^{k+r'-1} \overline{y}_{k'}(\mathcal{V}^+) - \overline{y}_{k'}(\mathcal{V}^-) \; ; \qquad \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,r'}(\mathcal{V}) = \frac{1}{r'} \left(\sum_{k'=k}^{k+r'-1} \overline{y}_{k'}(\mathcal{V}) - \sum_{k'=k-r'}^{k-1} \overline{y}_{k'}(\mathcal{V}) \right) \; . \tag{3.41}$$

Here, $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r'}(\mathcal{V}^+, \mathcal{V}^-)$ computes the width - i.e. the difference - between the mean of experts in $\mathcal{V}^+$ and the mean of experts in $\mathcal{V}^-$. Since $\mathcal{V}^+$ and $\mathcal{V}^-$ are built in the algorithm below in such a way that experts in $\overline{P}$ are below those of $\mathcal{V}^+$ and above those of $\mathcal{V}^-$, $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r'}(\mathcal{V}^+, \mathcal{V}^-)$ provides an upper bound of the width between the best expert in $\overline{P}$ and the worst expert in $\overline{P}$.

The algorithm **DimensionReduction – WM** described below builds a collection of sets $\mathcal{V}^+$, $\mathcal{V}^-$, and $\mathcal{V}$ and detects questions such that both the CUSUM $\mathbf{C}^{(\mathrm{ext})}_{k,r'}(\mathcal{V})$ and the width $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r'}(\mathcal{V}^+, \mathcal{V}^-)$ are large enough. Further explanations are postponed to the analysis of the algorithm in Section 3.5.5. Below, we write $\lceil x \rceil^{dya}$ for $2^{\lceil \log_2(x) \rceil}$.

Algorithm 13 DimensionReduction – WM$(Y, \mathcal{T}, \overline{P}, h, r)$

Require: A sample $Y \in \mathbb{R}^{n\times d}$, a tree $\mathcal{T}$, a set $\overline{P}$ included in a leaf G of $\mathcal{T}$ of type **0** or **1**, $h \in \mathcal{H}$ and $r \in \mathcal{R}$
Ensure: A set of blocks $\widehat{Q}_{WM} := \widehat{Q}_{WM}(Y, \mathcal{T}, \overline{P}, h, r) \subset \mathcal{Q}_r$

$r_0 = 2^9 \log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{|\overline{P}|h^2}$ and $\tilde{r} = 4(\lceil r_0 \rceil^{dya} \vee r)$
$(G^{(a)})_{a\in\mathbb{Z}} = \mathbf{Order}(\mathcal{T}, G)$
for $r_{\mathrm{cp}} \in [4r, 2\tilde{r}] \cap \mathcal{R}$ **do**
  Set $a^+_{WM} = \min\{a \ : \ |G^{(1)}| + \cdots + |G^{(a)}| \geq 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\mathrm{cp}}h^2}\}$
  Set $a^-_{WM} = \min\{a \ : \ |G^{(-1)}| + \cdots + |G^{(-a)}| \geq 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\mathrm{cp}}h^2}\}$
  Set

$$\mathcal{V}^+_{r_{\mathrm{cp}}} := \mathcal{V}^+_{r_{\mathrm{cp}}}(\mathcal{T}, G, h) = \bigcup_{a=1}^{a^+_{WM}} G^{(a)} \quad \text{and} \quad \mathcal{V}^-_{r_{\mathrm{cp}}} := \mathcal{V}^-_{r_{\mathrm{cp}}}(\mathcal{T}, G, h) = \bigcup_{a\in -a^-_{WM}}^{-1} G^{(a)} \tag{3.42}$$

  if $r_{\mathrm{cp}} > \tilde{r}$ **then**
    Set $\mathcal{V}_{r_{\mathrm{cp}}} := \mathcal{V}_{r_{\mathrm{cp}}}(\mathcal{T}, \overline{P}, h) = \overline{P}$
  else if $r_{\mathrm{cp}} \leq \tilde{r}$ **then**
    Set $\mathcal{V}_{r_{\mathrm{cp}}} := \mathcal{V}_{r_{\mathrm{cp}}}(\mathcal{T}, \overline{P}, h) = \mathcal{V}^-_{r_{\mathrm{cp}}} \cup \mathcal{V}^+_{r_{\mathrm{cp}}}$
  end if
end for
$\widehat{Q}_{WM} = \varnothing$
for $r_{\mathrm{cp}} \in [4r, \tilde{r}] \cap \mathcal{R}$ **do**
  $\widehat{D}_{WM} = \varnothing$
  for $k = 1, \ldots, d$ **do**
    Compute $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} := \widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}}, \mathcal{V}^-_{r_{\mathrm{cp}}})$
    Compute $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} := \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}(\mathcal{V}_{2r_{\mathrm{cp}}})$
  end for
  Update $\widehat{D}_{WM} = \{k \in [d] \ : \ \widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} \geq h/16 \text{ and } \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} \geq h/16\}$
  Update $\widehat{Q}_{WM} = \widehat{Q}_{WM} \cup \mathbf{Encode-Set}(\widehat{D}_{WM}, r)$
end for
return $\widehat{Q}_{WM}$ ▷ The same set is defined in (3.26)

3.5 Proofs

3.5.1 Overview and organization of the proofs of Theorems 3.2.2 and 3.2.3

In this section, we divide the analysis of the procedures into several properties that will be proved to hold with high probability in the next sections.

3.5.1.1 Definitions

Since we build our estimator using a hierarchical tree, we need to quantify the error that we suffer at each depth of the tree. For $i \in [n]$, we write $M_i = M_{i,\cdot}$ for the expert i. By definition of π^*, we recall that

$$M_{\pi^{*-1}(1)} \leq M_{\pi^{*-1}(2)} \leq \ldots \leq M_{\pi^{*-1}(n)} \ .$$

For a given group of experts G, we write $\pi^*_{\{G\}}$ for the oracle ordering in $[1, |G|]$ of the group G according to π^*, that is for all $i, j \in G$, $\pi^*_{\{G\}}(i)$ and $\pi^*_{\{G\}}(j)$ belong to $[1, |G|]$ and

$$\pi^*_{\{G\}}(i) < \pi^*_{\{G\}}(j) \quad \text{iff} \quad \pi^*(i) < \pi^*(j) \ .$$

We say that a sequence of sets $\mathcal{G} = (G_1, \ldots, G_\alpha)$ is an ordered partition of a set S if $\{G_1, \ldots, G_\alpha\}$ is a partition of S. For a given ordered partition $\{G_1, \ldots, G_\alpha\}$ and $a \in [1, \alpha]$ and any $i \in G_a$ we write

$$\pi^-_{\mathcal{G}}(G_a) = \pi^-_{\mathcal{G}}(i) := \sum_{a'<a} |G_{a'}| \quad \text{and} \quad \pi^+_{\mathcal{G}}(G_a) = \pi^+_{\mathcal{G}}(i) := \sum_{a'\leq a} |G_{a'}| \ . \tag{3.43}$$

If we are to build a permutation π which is consistent with this ordered partition, then one easily checks that $\pi(i) \in [\pi^-_{\mathcal{G}}(i) + 1, \pi^+_{\mathcal{G}}(i)]$. For simplicity, we write $G(i)$ for the group G_a such that $i \in G_a$. For a given ordered

partition $\mathcal{G} = (G_1, \dots, G_\alpha)$, we define the oracle permutation associated to $\mathcal{G}$ by

$$\pi^*_{\mathcal{G}}(i) = \pi^-_{\mathcal{G}}(i) + \pi^*_{\{G(i)\}}(i) \ . \tag{3.44}$$

For example, $\pi^*_{\{[n]\}} = \pi^*$ is simply the true permutation. By definition, we have $\pi^*_{\mathcal{G}}(G(i)) = [\pi^-_{\mathcal{G}}(i) + 1, \pi^+_{\mathcal{G}}(i)]$. Given an ordered partition, $\pi^*_{\mathcal{G}}$ is the best permutation we could hope for after any statistical treatment.

Given an ordered partition $\mathcal{G} = (G_1, \dots, G_\alpha)$, we define the random estimation of π^* given $\mathcal{G}$ as $\hat{\pi}_{\mathcal{G}}(i)$ which is uniformly distributed in $[\pi^-_{\mathcal{G}}(i) + 1, \pi^+_{\mathcal{G}}(i)]$:

$$\hat{\pi}_{\mathcal{G}}(i) \in [\pi^-_{\mathcal{G}}(i) + 1, \pi^+_{\mathcal{G}}(i)] \ .$$

Note that $\hat{\pi}_{\mathcal{G}}$ is not necessarily bijective.

3.5.1.2 Deterministic analysis

In this subsection, we analyze **TreeSort** (Algorithm 11) and we characterize the loss of the estimator $\hat{\pi}$ in terms of that of the trisections that are computed inside the subroutine **BlockSort**$(\mathcal{Y}, \mathcal{T}, G)$. This algorithm takes a subset G of experts and computes two trisections of G. The first one

$$(O, P, I) = (O_{\tau_\infty}, G \setminus (O_{\tau_\infty} \cup I_{\tau_\infty}), I_{\tau_\infty})$$

is returned by the algorithm. The second one

$$(\overline{O}, \overline{P}, \overline{I}) = (\overline{O}_{\tau_\infty}, G \setminus (\overline{O}_{\tau_\infty} \cup \overline{I}_{\tau_\infty}), \overline{I}_{\tau_\infty})$$

is important for our analysis. From the definitions of the different procedures, one readily checks that $\overline{O} \subset O$ and $\overline{I} \subset I$. In fact, we shall prove later that, with high probability, the subsets (O, P, I) and $(\overline{O}, \overline{P}, \overline{I})$ satisfy the following stronger property.

Property 1.

1. *$\{O, P, I\}$ and $\{\overline{O}, \overline{P}, \overline{I}\}$ are partitions of the leaf G with $\overline{O} \subset O$, $\overline{I} \subset I$, and $P \subset \overline{P}$,*
2. *For $\omega = \pi^{*-1}_{\{O,P,I\}} \pi^*_{\{G\}}$, we have $\omega(i) = i$ for any $i \in \overline{O} \cup \overline{I}$.*
3. *For any $i \in O$ and $j \in I$, we have $\pi^*(i) < \pi^*(j)$.*
4. *We have $|O| \le |G|/2$ and $|I| \le |G|/2$.*

The last claim states that all experts in O are below all experts of I. The second claim can be understood as the fact that, if an expert i belongs to $\overline{O}$, then all experts below i belong to O.

Let $\mathcal{Y} = (Y^{(0)}, \dots, Y^{(6\tau_\infty - 1)})$ be a sequence of $6\tau_\infty$ matrices in $\mathbb{R}^{n \times d}$. We say that **BlockSort** satisfies Property 1 on $(\mathcal{Y}, \mathcal{T}, G)$ if the two partitions (O, P, I) and $(\overline{O}, \overline{P}, \overline{I})$ worked out in Algorithm 11 satisfy Property 1. We recall that by definition $\overline{O} \subset O$, $\overline{I} \subset I$, $P \subset \overline{P}$ so that $\overline{P}$ corresponds to the collection of experts that are either not sorted by **TreeSort** or are sorted with a small confidence.

For each $t = 0, \dots, t_\infty$, we write $\mathcal{T}_t$ for the hierarchical sorting tree at the beginning of step t of **TreeSort**. Besides, we write $\mathcal{G}_t$ for the corresponding ordered partition obtained by taking the leaves of the tree $\mathcal{T}_t$ in increasing order in the ternary base $\{\mathbf{0}, \mathbf{p}, \mathbf{1}\}$. We define the tree $\overline{\mathcal{T}}_{t_\infty}$ as the tree $\mathcal{T}_{t_\infty}$ where we replaced all the leaves P of type $\mathbf{p}$ - at any depth - by $\overline{P}$, where $(\overline{O}, \overline{P}, \overline{I})$ has been worked out by **TreeSort** at the same time as (O, P, I).

We also define

$$\begin{aligned} \mathcal{L}_t(\mathcal{T}_{t_\infty}) &= \{P \in \mathcal{G}_{t_\infty} \ : \ P \text{ is a nonempty leaf at depth } t \text{ of } \mathcal{T}_{t_\infty}\} \ ; \\ \overline{\mathcal{L}}_t(\overline{\mathcal{T}}_{t_\infty}) &= \{\overline{P} \in \mathcal{G}_{t_\infty} \ : \ \overline{P} \text{ is a nonempty leaf at depth } t \text{ of } \overline{\mathcal{T}}_{t_\infty}\} \ . \end{aligned} \tag{3.45}$$

For simplicity, we sometimes write $\mathcal{L}_t$ for $\mathcal{L}_t(\mathcal{T}_{t_\infty})$ and $\overline{\mathcal{L}}_t$ for $\overline{\mathcal{L}}_t(\overline{\mathcal{T}}_{t_\infty})$. $\mathcal{L}_t$ stands for the collection of experts that have not been sorted at the t-th iteration **TreeSort**. The sets in the collection $\overline{\mathcal{L}}_t$ are strictly larger and correspond to the collections of experts in $\overline{P}$ that are either not sorted by **TreeSort** or are sorted with less confidence. Let $M(\overline{P})$ be defined as the restriction of M to the experts in $\overline{P}$, and $\overline{M}(\overline{P})$ the $|\overline{P}| \times d$ matrix with constant columns which correspond to the mean row of $M(\overline{P})$. The following proposition characterizes the loss of the final estimator estimator $\hat{\pi}_{\mathcal{G}_{t_\infty}}$ which is obtained from a hierarchical sorting tree in terms of the variance of the experts M within the groups $\overline{P}$ in $\overline{\mathcal{L}}_t(\overline{\mathcal{T}}_{t_\infty})$.

Proposition 3.5.1 (Deterministic Analysis of **TreeSort**). *Assume that at each step of* **TreeSort**, *the routine* **BlockSort** *applied to the data satisfies Property 1. Then, the error of* $\hat{\pi} = \hat{\pi}_{\mathcal{G}_{t_\infty}}^{-1}$ *is controlled as follows*

$$\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq 10 t_\infty \sum_{t=1}^{t_\infty} \sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \ . \tag{3.46}$$

Besides, the hierarchical tree $\mathcal{T}_{t_\infty}$ *is valid (as in Definition 2) and all its non-empty leaves are of type* **p**.

Up to a normalization, $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$ corresponds to the variance of M within the group $\overline{P}$. The bound (3.46) expresses that the loss of a hierarchical sorting tree is controlled by the variance of the set $\overline{P}$ that are not sorted with confidence at each step of the algorithm. Also, we recall that $t_\infty = \lceil \log(n)/\log(2) \rceil$. This proposition only relies on Property 1 and on the construction of the tree. Hence, it applies both to the estimators $\hat{\pi}_{HT}$ and $\hat{\pi}_{WM}$.

The sets $(O, \overline{O}, I, \overline{I})$ built in **BlockSort** arise as unions of set (L, U) and $(\overline{L}, \overline{U})$ that are computed by **DoubleTrisection** for a set $\overline{P}$ and a pivot $\gamma \in [1, |\overline{P}|]$. For this reason, we now state a desired property of the result of the algorithm that will enforce Property 1.

Property 2 (Property on (L, U) and $(\overline{L}, \overline{U})$). *For* $\overline{P}' = \overline{P} \setminus (\overline{L} \cup \overline{U})$ *and* $P' = \overline{P} \setminus (L \cup U)$, *we have*

1. $\overline{L} \subset L$, *and* $\overline{U} \subset U$,
2. *if* $\omega = \pi_{\{L, P', U\}}^{*-1} \pi_{\{\overline{P}\}}^*$ *then for any* $i \in \overline{L} \cup \overline{U}$ *it holds that* $\omega(i) = i$,
3. *For any* $i \in L$ *and* $j \in U$ *we have* $\pi_{\{\overline{P}\}}^*(i) < \gamma < \pi_{\{\overline{P}\}}^*(j)$.

We say that **DoubleTrisection** with $(\mathcal{Y}, \mathcal{T}, \overline{P}, \gamma)$ satisfies Property 2 if the corresponding subsets (L, U) and $(\overline{L}, \overline{U})$ satisfy Property 2.

Proposition 3.5.2 (Deterministic Analysis of **BlockSort**). **BlockSort** *satisfies Property 1 on* $(\mathcal{Y}, \mathcal{T}, G)$ *if, at each step of Algorithm 10, each call of* **DoubleTrisection** *satisfies Property 2.*

In light of Propositions 3.5.1 and 3.5.2, it suffices to show that, with high probability, all applications of **DoubleTrisection** in the construction of the hierarchical sorting tree satisfy Property 2, and then to control the sum of within-group variances in (3.46).

3.5.1.3 High probability control of property 2

We write in this part of the proof (this subsection), for simplicity, $\mathcal{Y} = (Y^{(1)}, \dots, Y^{(6)})$ for 6 independent matrices that are identically distributed as $Y = M + E$ in (3.1), where we recall that the entries of E are centered, independent and ζ-subgaussian.

Fix a hierarchical sorting tree $\mathcal{T}$ (recall Definition 1), a leaf G of $\mathcal{T}$, a set $\overline{P} \subset G$, a pivot $\gamma \in \{1, \dots, |\overline{P}|\}$. Let $\mathcal{P}_2 := \mathcal{P}_2(\mathcal{T}, \overline{P}, \gamma, \beta_{\text{tris}}, \overline{\beta}_{\text{tris}})$ be the event holding true if **DoubleTrisection** satisfies Property 2 on $\mathcal{Y}$ for $(\mathcal{T}, \overline{P}, \gamma, \beta_{\text{tris}}, \overline{\beta}_{\text{tris}})$. The following proposition states that $\mathcal{P}_2$ holds with uniformly high probability.

Proposition 3.5.3. *For any* $\mathcal{T}$, *any leaf* G, *any* $\overline{P} \subset G$, *any pivot* $\gamma \in [|\overline{P}|]$, *we have* $\mathbb{P}(\mathcal{P}_2) \geq 1 - 3|\mathcal{H}||\mathcal{R}|\delta$.

This result is valid for both versions of **DoubleTrisection** where we use the memory of the tree (estimator $\hat{\pi}_{WM}$) or not (estimator $\hat{\pi}_{HT}$). Recall that in **BlockSort** there are at most τ_∞ calls of **DoubleTrisection**. Since the construction of the hierarchical tree requires at most $2^{t_\infty+1}$ applications of **BlockSort**, we arrive at the following straightforward corollary of Propositions 3.5.2, 3.5.1 and 3.5.3.

Corollary 3.5.4. *There exists an event* ξ *of probability higher than* $1 - 2^{t_\infty+1} 3\tau_\infty |\mathcal{H}||\mathcal{R}|\delta$ *such that all results of* **BlockSort** *within* **TreeSort** *satisfy Property 1. In particular, the tree* $\mathcal{T}_{t_\infty}$ *is a valid hierarchical sorting tree (as in Definition 2) whose non-empty leaves are all of type* **p**. *Besides, on this event we also have*

$$\|M_{\hat{\pi}_{\mathcal{G}_{t_\infty}}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq 10 t_\infty \sum_{t=1}^{t_\infty} \sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \ . \tag{3.47}$$

Again, this results applies to both variants of our procedure - with or without memory.

3.5.1.4 Control of the loss function

In contrast to the previous subsection, we now need to specify the dimension reduction scheme **DimensionReduction** (which corresponds to $\hat{\pi}_{HT}$) or **DimensionReduction – WM** (which corresponds to $\hat{\pi}_{WM}$) inside **DoubleTrisection** as the convergence rates depend on these quantities.

First we state the results for the method without memory: $\hat{\pi}_{HT}$.

Proposition 3.5.5. *Consider the oblivious hierarchical sorting tree estimator $\hat{\pi}_{HT}$. On the intersection of event ξ (defined in Corollary 3.5.4) and an event of probability higher than $1 - 5 \cdot 2^t \tau_\infty \delta$, it holds that*

$$\sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|^2 \lesssim \zeta^2 \log^{8.5}\left(\frac{6nd}{\delta\zeta_-}\right)\left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] .$$

Then we state the results for the method with memory: $\hat{\pi}_{WM}$.

Proposition 3.5.6. *Consider the hierarchical sorting tree estimator $\hat{\pi}_{WM}$. On the intersection of event ξ (defined in Corollary 3.5.4) and an event of probability higher $1 - 5 \cdot 2^t \tau_\infty \delta$, it holds that*

$$\sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|^2 \lesssim \zeta^2 \log^{9}\left(\frac{6nd}{\delta\zeta_-}\right)\left[\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}}\right) + n\right] .$$

Now, we are in position to easily conclude the proof of Theorems 3.2.2 and 3.2.3.

Proof of Theorem 3.2.2. Let $\hat{\pi}_{HT} := \hat{\pi}_{\mathcal{G}_{t_\infty}}$ denote the oblivious hierarchical sorting tree estimator. Combining Corollary 3.5.4 with Proposition 3.5.5 and a union bound over all $t = 0, \dots, t_\infty - 1$, it holds with probability higher than $1 - 8 \cdot 2^{t_\infty + 1} \tau_\infty |\mathcal{H}||\mathcal{R}|\delta$ that

$$\begin{aligned}
\|M_{\hat{\pi}_{HT}^{-1}} - M_{\pi^{\star -1}}\|_F^2 &\lesssim t_\infty^2 \, \zeta^2 \log^{8.5}\left(\frac{2nd}{\delta\zeta_-}\right)\left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] \\
&\lesssim \zeta^2 \log^{10.5}\left(\frac{2nd}{\delta\zeta_-}\right)\left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] .
\end{aligned}$$

□

Proof of Theorem 3.2.3. Let $\hat{\pi}_{\mathrm{WM}} := \hat{\pi}_{\mathcal{G}_{t_\infty}}$ denote the hierarchical sorting tree where we use the memory to reduce the dimension (Algorithm **DimensionReduction – WM**). Combining Corollary 3.5.4 with Proposition 3.5.6 and a union bound on $t = 0, \dots, t_\infty - 1$, it holds with probability higher than $1 - 8 \cdot 2^{t_\infty + 1} \tau_\infty |\mathcal{H}||\mathcal{R}|\delta$ that

$$\|M_{\hat{\pi}_{WM}^{-1}} - M_{\pi^{\star -1}}\|_F^2 \lesssim \zeta^2 \log^{11}\left(\frac{6nd}{\delta\zeta_-}\right)\left[\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}}\right) + n\right] .$$

□

In the next four sections, we prove the intermediary results. Propositions 3.5.1–3.5.3 are relatively simple. The main difficulty and the key arguments lie in the proofs of Proposition 3.5.5 and 3.5.6 which are respectively in Sections 3.5.3 and 3.5.5.

3.5.2 Proofs of Propositions 3.5.1, 3.5.2, and 3.5.3

Proof of Proposition 3.5.1. First, we prove by induction that $\mathcal{T}_{t_\infty}$ is a valid hierarchical sorting tree. Besides, the last part of Property 1 enforces that the cardinality of any non-terminal node G of $\mathcal{T}_{t_\infty}$ of depth t is at most $n/2^t$. As a consequence, the cardinality of any non-terminal node at depth $t_{\infty-1}$ is at most 1 and its children O and I are therefore empty.

We control the error using a telescopic sum. Recall that, by convention, $\pi^\star_{\mathcal{G}_0} = \pi^\star$. We start with the following inequality:

$$\|M_{\hat{\pi}_{\mathcal{G}_{t_\infty}}^{-1}} - M_{\pi^{\star -1}}\|_F^2 \le 2\|M_{\hat{\pi}_{\mathcal{G}_{t_\infty}}} - M_{\pi^{\star -1}_{\mathcal{G}_{t_\infty}}}\|_F^2 + 2t_\infty \sum_{t=1}^{t_\infty} \|M_{\pi^{\star -1}_{\mathcal{G}_t}} - M_{\pi^{\star -1}_{\mathcal{G}_{t-1}}}\|_F^2 . \tag{3.48}$$

Since, for any group P in $\mathcal{G}_{t_\infty}$, $\hat{\pi}_{\mathcal{G}_{t_\infty}}$ sorts the elements of P uniformly at random and $\pi^{*-1}_{\mathcal{G}_{t_\infty}}$ acts as another permutation of P, we deduce from the triangular inequality that

$$\begin{aligned}\|M_{\hat{\pi}^{-1}_{\mathcal{G}_{t_\infty}}} - M_{\pi^{*-1}_{\mathcal{G}_{t_\infty}}}\|_F^2 &\le \sum_{P\in\mathcal{G}_{t_\infty}} 2\|M(P)-\overline{M}(P)\|_F^2 = 2\sum_{t=1}^{t_\infty}\sum_{P\in\mathcal{L}_t}\|M(P)-\overline{M}(P)\|_F^2 \\ &\le 2\sum_{t=1}^{t_\infty}\sum_{\overline{P}\in\overline{\mathcal{L}}_t}\|M(\overline{P})-\overline{M}(\overline{P})\|_F^2 \ , \end{aligned} \tag{3.49}$$

where we used in the last line that $P\subset\overline{P}$. For the second term in (3.48), remark that $\pi^*_{\mathcal{G}_{t-1}}(P)=\pi^*_{\mathcal{G}_t}(P)$ for any $P\in\mathcal{L}_{t-1}$ so that the error at step t in the telescopic sum can be restricted to the groups G that are trisected at step $t-1$:

$$\sum_{t=1}^{t_\infty}\|M_{\pi^{*-1}_{\mathcal{G}_t}} - M_{\pi^{*-1}_{\mathcal{G}_{t-1}}}\|_F^2 = \sum_{t=1}^{t_\infty}\sum_{G\in\mathcal{G}_{t-1}\setminus(\cup_{s\ge1}\mathcal{L}_{t-s})}\sum_{i\in G}\|M_{\pi^{*-1}_{\mathcal{G}_t}(\pi^*_{\mathcal{G}_{t-1}}(i))} - M_i\|_2^2 \ .$$

Let (O,P,I) be the trisection obtained at the t-th iteration when we apply **BlockSort** to a group $G\in\mathcal{G}_{t-1}\setminus(\cup_{s\ge1}\mathcal{L}_{t-s})$. We also write $(\overline{O},\overline{P},\overline{I})$ for the more conservative trisection obtained at the end of **BlockSort**. For short, we write $\omega=\pi^{*-1}_{\mathcal{G}_t}\pi^*_{\mathcal{G}_{t-1}}$. We decompose the sum over $i\in G$:

$$\sum_{i\in G}\|M_{\omega(i)}-M_i\|^2 = \sum_{i\in\overline{O}}\|M_{\omega(i)}-M_i\|^2 + \sum_{i\in\overline{I}}\|M_{\omega(i)}-M_i\|^2 + \sum_{i\in\overline{P}}\|M_{\omega(i)}-M_i\|^2 \ ,$$

By Property 1, all the experts in $\overline{O}$ and in $\overline{I}$ are perfectly sorted within G by $\pi^{*-1}_{\mathcal{G}(t)}$. As a consequence, the two first sums in the right-hand side term of the above equality are equal to zero. To handle the last term, we introduce the row vector $m(\overline{P})$ as the mean of the experts of M over $\overline{P}$:

$$\begin{aligned}\sum_{i\in G}\|M_{\omega(i)}-M_i\|_2^2 = \sum_{i\in\overline{P}}\|M_{\omega(i)}-M_i\|_2^2 &\le 2\sum_{i\in\overline{P}}(\|M_i-m(\overline{P})\|_2^2 + \|m(\overline{P})-M_{\omega(i)}\|_2^2) \\ &= 4\|M(\overline{P})-\overline{M}(\overline{P})\|_F^2 \ ,\end{aligned}$$

where we used in the last line that ω acts as a permutation of $\overline{P}$. Since $\overline{P}\in\overline{\mathcal{L}}_t$, we obtain

$$\sum_{t=1}^{t_\infty}\|M_{\pi^{*-1}_{\mathcal{G}_t}} - M_{\pi^{*-1}_{\mathcal{G}_{t-1}}}\|^2 \le 4\sum_{t=1}^{t_\infty}\sum_{\overline{P}\in\overline{\mathcal{L}}_t}\|M(\overline{P})-\overline{M}(\overline{P})\|_F^2 \ .$$

Together with (3.48) and (3.49), this concludes the proof since $t_\infty\ge1$. □

Proof of Proposition 3.5.2. Consider any data $\mathcal{Y}$, any tree $\mathcal{T}$ and any leaf G of $\mathcal{T}$. Let (O,P,I) and $(\overline{O},\overline{P},\overline{I})$ denote the trisections built in **BlockSort**$(\mathcal{Y},\mathcal{T},G)$. For any $\tau<\tau_\infty$, let (L_τ,U_τ), $(\overline{L}_\tau,\overline{U}_\tau)$, (O_τ,I_τ) and $(\overline{O}_\tau,\overline{I}_\tau)$ be defined as in Algorithm 10. We also write $P_\tau=G\setminus(O_\tau\cup I_\tau)$ and $\overline{P}_\tau=G\setminus(\overline{O}_\tau\cup\overline{I}_\tau)$. We only need to prove that, for all τ, (O_τ,P_τ,I_τ), and $(\overline{O}_\tau,\overline{P}_\tau,\overline{I}_\tau)$ satisfy Property 1. Since

$$O_\tau=\bigcup_{\tau'<\tau}L_{\tau'} \quad\text{and}\quad I_\tau=\bigcup_{\tau'<\tau}U_{\tau'} \quad\text{and}\quad \overline{O}_\tau=\bigcup_{\tau'<\tau}\overline{L}_{\tau'} \quad\text{and}\quad \overline{I}_\tau=\bigcup_{\tau'<\tau}\overline{U}_{\tau'} \ ,$$

we easily deduce from Property 2 for (L_τ,U_τ) and $(\overline{L}_\tau,\overline{U}_\tau)$ that the first part of Property 1 is satisfied for (O_τ,P_τ,I_τ), and $(\overline{O}_\tau,\overline{P}_\tau,\overline{I}_\tau)$.

Let us turn to the third and fourth parts of Property 1. Let us call i_m the expert such that $\pi^*_{\{G\}}(i_m)=\lfloor|G|/2\rfloor$. In fact, we only need to prove that $\max_{i\in O_\tau}\pi^*_{\{G\}}(i)\le|G|/2$ and $\min_{i\in I_\tau}\pi^*_{\{G\}}(i)\ge|G|/2$. For this purpose, we prove by induction on τ that the pivot always satisfies $\pi^{*-1}_{\{\overline{P}_\tau\}}(\gamma)=i_m$ and that all the experts of O_τ (resp. I_τ) are below (resp. above) i_m, where γ depends on τ and is defined in Algorithm 10. Assume that this property holds at step τ. Since $\overline{O}_\tau$ only contains experts that are below the median expert and since $\gamma=\lfloor|G|/2\rfloor-|\overline{O}_\tau|$, it follows that $\pi^{*-1}_{\{\overline{P}\}}(\gamma)=i_m$. Consider any $i\in O_{\tau+1}$. If $i\in O_\tau$, then $\pi^*_{\{G\}}(i)\le|G|/2$ by induction. If $i\in L_\tau$, then it follows from Property 2 that i is below i_m, which in turn implies that $\pi^*_{\{G\}}(i)\le|G|/2$. By symmetry, the property also holds for I_τ. We have proved the third and the fourth parts of Property 1.

Finally, we consider the second part of Property 1. Assume that the property holds at step τ. This implies that, for any $i\in\overline{O}_\tau$, all experts below i belong to O_τ. Consider any expert $i\in\overline{O}_{\tau+1}$. If $i\in\overline{O}_\tau$, then, by induction, we have $\pi^*_{\{G\}}(i)=\pi^*_{\{O_\tau\}}(i)=\pi^*_{\{O_{\tau+1}\}}(i)$. Then, we turn to the case where i belongs to $\overline{L}_\tau\subset\overline{P}_\tau$. Consider any $j\in G$ such that $\pi^*_{\{G\}}(j)\le\pi^*_{\{G\}}(i)$. If $j\in\overline{O}_\tau$, then we obviously have $j\in\overline{O}_{\tau+1}$. If $j\in\overline{P}_\tau$, then

the second part of property 2 enforces that $j \in L_\tau$ and therefore $j \in O_{\tau+1}$. Finally, it is not possible that $j \in \overline{I}_\tau$ since this enforces that $\pi^*_{\{G\}}(j) > |G|/2 > \pi^*_{\{G\}}(i)$ and contradicts the hypothesis. We prove similarly that, for any expert $i \in \overline{I}_{\tau+1}$, all experts j above i belong to $I_{\tau+1}$.

□

Proof of Proposition 3.5.3. As **DoubleTrisection** is based on multiple applications of the **Pivot** algorithm, we start by considering the latter procedure.

Consider two sets $|\overline{P}| \subset [n]$ and $Q \subset [d]$ and a matrix $\Theta \in \mathbb{R}^{|\overline{P}|\times|Q|}$ which, up to the permutation $\pi^*_{\{\overline{P}\}}$, is bi-isotonic. Let Z be a noiy observation of Θ,

$$Z = \Theta + N \ , \tag{3.50}$$

where the noise matrix N is made of independent, centered, ζ-subGaussian random variables. Let $w \in \mathbb{R}^Q_+$ be a non-zero vector with nonegative coordinates. we write (L, U) and $(\overline{L}, \overline{U})$ for the result of **Pivot**(Z, w, γ). We define the event $\mathcal{P}_3 := \mathcal{P}_3(\overline{P}, Q, w, \gamma)$ as the event on Z such that (L, U) and $(\overline{L}, \overline{U})$ satisfy Property 2.

We remind that $P' = \overline{P} \setminus (L \cup U)$, and $\overline{P'} = \overline{P} \setminus (\overline{L} \cup \overline{U})$.

Lemma 3.5.7. *For any non-zero vector $w \in \mathbb{R}^Q_+$, any pivot $\gamma \in \{1, \dots, |\overline{P}|\}$, we have $\mathbb{P}[\mathcal{P}_3] \geq 1 - \delta$. Besides, on the same event of probability at least $1 - \delta$, we have*

$$\left|\langle \Theta_{i,\cdot} - \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq (2\zeta\sqrt{2} + \overline{\beta}_{\text{tris}})\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} \quad \text{if } i \in P' \ . \tag{3.51}$$

Before proving the lemma, let us explain why Proposition 3.5.3 is easily deduced from it. The procedure **DoubleTrisection** calls at most $3|\mathcal{H}||\mathcal{R}|$ times **Pivot**. Note that, each time, we rely on an independent sample to choose the direction w and to apply **Pivot**. Then, applying the Lemma, we derive that, with probability higher than $1 - 3|\mathcal{H}||\mathcal{R}|\delta$, each of these $3|\mathcal{H}||\mathcal{R}|$ sets (L, U) and $(\overline{L}, \overline{U})$ satisfy Property 2. Hence, we only need to check that Property 2 is stable by union. If, both $(L^{(1)}, U^{(1)})$ and $(\overline{L}^{(1)}, \overline{U}^{(1)})$ and $(L^{(2)}, U^{(2)})$ and $(\overline{L}^{(2)}, \overline{U}^{(2)})$ satisfy Property 2, then one easily checks that the first and third part of Property 2 are also true for $(L, U) = (L^{(1)} \cup L^{(2)}, U^{(1)} \cup U^{(2)})$ and $(\overline{L}, \overline{U}) = (\overline{L}^{(1)} \cup \overline{L}^{(2)}, \overline{U}^{(1)} \cup \overline{U}^{(2)})$. Consider any expert i in $\overline{L}$. Without loss of generality, we may assume that $i \in \overline{L}^{(1)}$ so that all experts below i in $\overline{P}$ belong to $L^{(1)}$ by the second part of Property 2. As a consequence, all these experts below i belong to L and we deduce that the second part of Property 2 holds. Similarly, we deal with experts $i \in \overline{U}$. This concludes the proof of Proposition 3.5.3.

□

Proof of Lemma 3.5.7. Since the noise matrix in (3.50) is made of independent ζ-subGaussian random variables, it follows from a union bound, that with probability higher than $1 - \delta$, we have

$$\left|\langle Z_{i,\cdot}, \frac{w}{\|w\|_2}\rangle - \langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq \zeta\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)} \ .$$

simultaneously for all i in $\overline{P}$. Since the entries of w are non-negative, the quantities $\langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2}\rangle$ are ordered according the permutation $\pi^*_{\{\overline{P}\}}$. Denote $i_\gamma = \pi^{*-1}_{\{\overline{P}\}}(\gamma)$ and $\widehat{i}_\gamma$ the index of γ-th value of $\langle Z_{i,\cdot}, \frac{w}{\|w\|_2}\rangle$ for $i \in \overline{P}$. Since at least γ experts satisfy $\langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2}\rangle \leq \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle$, we deduce from the above uniform deviation inequality that

$$\langle Z_{\widehat{i}_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle \leq \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle + \zeta\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)} \ .$$

By symmetry, we deduce that

$$\left|\langle Z_{\widehat{i}_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle - \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq \zeta\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)} \ .$$

As a consequence, we have

$$\begin{aligned}
\langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2} \rangle &< \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2} \rangle - (\beta_{\mathrm{tris}} - 2\zeta\sqrt{2}) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} \quad \text{if } i \in L\,; \\
\langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2} \rangle &> \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2} \rangle + (\beta_{\mathrm{tris}} - 2\zeta\sqrt{2}) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} \quad \text{if } i \in U\ ; \\
\left| \langle \Theta_{i,\cdot} - \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2} \rangle \right| &\le (2\zeta\sqrt{2} + \beta_{\mathrm{tris}}) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} \quad \text{if } i \in P'\ . \qquad (3.52)
\end{aligned}$$

The same inequalities hold for $\overline{L}$, $\overline{U}$, and $\overline{P}'$ provided that we replace β_{tris} by $\overline{\beta}_{\mathrm{tris}}$. It remains to show that (L,U) and $(\overline{L},\overline{U})$ satisfy Property 2. The first part of the property is obvious. Since $\beta_{\mathrm{tris}} \ge 2\sqrt{2}\zeta$, one observes that $\pi^*_{\{\overline{P}\}}(i) < \pi^*_{\{\overline{P}\}}(i_\gamma) = \gamma$ if $i \in L$. Similarly, $\pi^*_{\{\overline{P}\}}(i) > \gamma$ if $i \in U$ and the third part of Property 2 follows. Turning to the second part of the property, we consider without loss of generality some $i \in \overline{L}$ and we need to show that all j satisfying $\pi^*_{\{\overline{P}\}}(j) \le \pi^*_{\{\overline{P}\}}(i)$ belong to L. First, such a j does not belong to U since $\pi^*_{\{\overline{P}\}}(i) < \gamma$. Since $i \in \overline{L}$, we deduce that

$$\langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2} \rangle < \langle \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2} \rangle - (\overline{\beta}_{\mathrm{tris}} - 2\sqrt{2}\zeta) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)}\,,$$

which implies that

$$\left| \langle \Theta_{i_\gamma,\cdot} - \Theta_{j,\cdot}, \frac{w}{\|w\|_2} \rangle \right| > (\overline{\beta}_{\mathrm{tris}} - 2\sqrt{2}\zeta) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} \ge (2\zeta\sqrt{2} + \beta_{\mathrm{tris}}) \sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)}\,.$$

which in light of (3.52) implies that $j \notin P'$. We have proved that j belongs to L. Hence, Property 2 holds, which concludes the proof. □

3.5.3 Proof of Proposition 3.5.5

In this section, we prove Proposition 3.5.5 and thereby control the loss of the estimator $\hat{\pi}$ with simple dimension reduction. For this purpose, we analyze each step of the algorithm. In Section 3.5.3.2, we first prove that, by detecting the high-variation regions of M, we are able to aggregate M at some scale r without decreasing much the variation of M. This allows us to drastically reduce the dimension of the problem. Then, in Sections 3.5.3.3 and 3.5.3.3, we show that, unless this aggregated matrix Θ has small variations, **DoubleTrisection – PCA** and **Pivot** will remove some experts so that the corresponding new aggregated matrix Θ' exhibit significantly smaller variations. As a consequence, after a polylogarithmic number of iterations of the procedure, the variations of the matrix M restricted to the remaining experts of $\overline{P}$ is small enough.

3.5.3.1 Notation

As the arguments rely on considering aggregation of the matrix at different scales, we recall some notation. Let $Y = M + E$ denote a sample of the original matrix. For a set $P \subset [n]$ of experts and a set Q of blocks of questions and a scale $r \in R$, we respectively denote

$$\begin{aligned}
Z(P,Q,r) &= \textbf{Encode}-\textbf{Matrix}(Y,P,Q,r) \in \mathbb{R}^{P\times Q} \\
\Theta(P,Q,r) &= \textbf{Encode}-\textbf{Matrix}(M,P,Q,r) \\
N(P,Q,r) &= \textbf{Encode}-\textbf{Matrix}(E,P,Q,r) \quad ,
\end{aligned}$$

the aggregations of Y, M, and E at scale r so that

$$Z(P,Q,r) = \Theta(P,Q,r) + N(P,Q,r)\ .$$

By definition of **Encode – Matrix**, all the entries of N are independent and ζ-subGaussian. For any $p \times q$ matrix A, we define $\overline{a}$ as the row vector corresponding to the column-wise mean of A, that is $\overline{a}_j = \frac{1}{q}\sum_{i=1}^{q} A_{i,j}$. Besides, we write $\overline{A}$ for $p \times q$ matrix whose experts are all equal to $\overline{a}$.

3.5.3.2 Analysis of DimensionReduction

In this subsection, we mainly state, that for any $h \in \mathcal{H}$, $r \in \mathcal{R}$, the set $\widehat{Q}_{\mathrm{cp}}$ (which depends on h, r) detects the high-variation regions of M with high probability. Then, we show in Lemma 3.5.9, that for, for some $(h,r) \in \mathcal{H}\times\mathcal{R}$, the aggregation of M at scale r and at these high-variation regions contains most of the variance of M. This motivates us to work with this aggregated matrix henceforth. Consider any set $\overline{P}$ of experts and a sample

$$Y(\overline{P}) = M(\overline{P}) + E(\overline{P}) \ .$$

Fix any scale $r \in \mathcal{R}$ and any height $h \in \mathcal{H}$. Recall the two quantities r_0 and $\tilde{r}$ defined in **DimensionReduction** by

$$r_0 = 32\zeta^2 \log\left(\frac{2d}{\delta}\right)\frac{1}{|\overline{P}|h^2} \quad \text{and} \quad \tilde{r} = 8(\lceil r_0\rceil \vee r) \ . \tag{3.53}$$

In a nustshell, r_0 stands for the minimal scale at which a variation of order h in the mean $\overline{m}(\overline{P}) = \mathbb{E}[\overline{y}(\overline{P})]$ can be statistically detected. This is why we consider empirical variations of $\overline{y}(\overline{P})$ at scale $\tilde{r} \geq (r_0 \vee r)$ in Algorithm **DimensionReduction** to possibly detect variations at scale r.

The purpose of this subsection is to prove, that with high probability, the collections $\widehat{Q}_{\mathrm{cp}}(h,r) = \mathbf{DimensionReduction}(Y,\overline{P},h,r)$ of selected blocks of length r is not too large and that there exists at least one $(h,r) \in \mathcal{H}\times\mathcal{R}$ such that the aggregation of $M(\overline{P})$ at scale r restricted to the blocks $\widehat{Q}_{\mathrm{cp}}(h,r)$ captures most of the variance of $M(\overline{P})$.

For this purpose, we recall the CUSUM statistics introduced in **DimensionReduction** and we introduce its population counterpart. Given positive integers $k \in [d]$ and $r > 0$, consider

$$\widehat{\mathbf{C}}_{k,r} = \frac{1}{r}\left(\sum_{k'=k}^{k+r-1} \overline{y}_{k'}(\overline{P}) - \sum_{k'=k-r}^{k-1} \overline{y}_{k'}(\overline{P})\right) \quad \text{and} \quad \mathbf{C}^*_{k,r} = \frac{1}{r}\left(\sum_{k'=k}^{k+r-1} \overline{m}_{k'}(\overline{P}) - \sum_{k'=k-r}^{k-1} \overline{m}_{k'}(\overline{P})\right) \ .$$

Equipped with this notation, we define $\widehat{D}_{\mathrm{cp}}$ as in the algorithm as the set of positions d such that the association CUSUM statistic is above the threshold, and D^*_{cp} and $\overline{D}^*_{\mathrm{cp}}$ as some population versions of $\widehat{D}_{\mathrm{cp}}$, but with different tuning parameters:

$$\widehat{D}_{\mathrm{cp}}(h,r) = \left\{k \in [d] \ : \ \widehat{\mathbf{C}}_{k,\tilde{r}} \geq \frac{1}{4}h\right\} \ , \tag{3.54}$$

$$D^*_{\mathrm{cp}}(h,r) = \left\{k \in [d] \ : \ \mathbf{C}^*_{k,8r} \geq \frac{1}{2}h\right\} \ ; \quad \overline{D}^*_{\mathrm{cp}}(h,r) = \left\{k \in [d] \ : \ \mathbf{C}^*_{k,\tilde{r}} \geq \frac{1}{8}h\right\} \ . \tag{3.55}$$

Then, we consider the collection of blocks $\widehat{Q}_{\mathrm{cp}}(h,r)$, $Q^*_{\mathrm{cp}}(h,r)$, and $\overline{Q}^*_{\mathrm{cp}}(h,r)$ of size r that are associated with these positions. In terms of our algorithms, this means that $Q^*_{\mathrm{cp}}(h,r) = \mathbf{Encode-Set}(D^*_{\mathrm{cp}},r)$, $\overline{Q}^*_{\mathrm{cp}}(h,r) = \mathbf{Encode-Set}(\overline{D}^*_{\mathrm{cp}},r)$, and $\widehat{Q}_{\mathrm{cp}}(h,r) = \mathbf{Encode-Set}(\widehat{D}_{\mathrm{cp}},r)$. The first proposition states that, with high probability, $\widehat{Q}_{\mathrm{cp}}(h,r)$ is sandwidched between $Q^*_{\mathrm{cp}}(h,r)$ and $\overline{Q}^*_{\mathrm{cp}}(h,r)$, so that, on the corresponding event, it is sufficient to study these two quantities.

Lemma 3.5.8. *For all h,r, the event $\xi_{\mathrm{cp}} := \xi_{\mathrm{cp}}(\overline{P},h,r)$ defined by*

$$Q^*_{\mathrm{cp}} \subset \widehat{Q}_{\mathrm{cp}} \subset \overline{Q}^*_{\mathrm{cp}} \ , \tag{3.56}$$

holds true with probability at least $1-\delta$.

Then, we show that there are not too many significant blocks in $\overline{Q}^*_{\mathrm{cp}}$. The proof is based on the fact that the row vector $\overline{m}(\overline{P})$ is isotonic and lies in $[0,1]$. As a consequence, there cannot exist two many regions where the variations of $\overline{m}(\overline{P})$ is large.

Lemma 3.5.9. *For all $h \in \mathcal{H}$ and all $r \in \mathcal{R}$, we have*

$$|\overline{Q}^*_{\mathrm{cp}}| \leq \frac{64\tilde{r}}{rh} \ . \tag{3.57}$$

The next lemma states that, at least for a height $h \in \mathcal{H}$ and a scale $r \in \mathcal{R}$, the aggregation of M at scale r and restricted to the regions $Q^*_{\mathrm{cp}}(h,r)$ of significant variations contains almost all the variance of the signal.

For any number θ and any $\eta > 0$, we define $[\theta]_\eta = (-1)^{\mathrm{sgn}(\theta)}\eta\mathbf{1}_{|\theta|\geq\eta}$. For any matrix Θ, we write $[\Theta]_\eta$ for the thresholded matrix with coefficients $[\Theta_{i,j}]_\eta$.

Lemma 3.5.10. *For any set $\overline{P} \subset [n]$ and any bi-isotonic matrix $M \in [0,1]^{n\times d}$, there exist $r \in \mathcal{R}$ and $h \in \mathcal{H}$ such that*

$$\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \leq 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}|\left\|\left[\Theta(\overline{P},Q^*_{\mathrm{cp}}) - \overline{\Theta}(\overline{P},Q^*_{\mathrm{cp}})\right]_{\sqrt{r}h}\right\|_F^2 \ , \tag{3.58}$$

The proof of the above lemmas is postponed to Section 3.5.4.1.

3.5.3.3 Analysis of Pivot based on the row sums

We consider a specific subset $\overline{P}$ of experts, a subset Q of blocks of questions, the corresponding aggregated model

$$Z(\overline{P},Q) = \Theta(\overline{P},Q) + N(\overline{P},Q) \in \mathbb{R}^{\overline{P}\times Q} \quad , \tag{3.59}$$

and a pivot $\gamma \in [1,|\overline{P}|]$. Let $(\overline{L},\overline{U}) = \mathbf{Pivot}(Z,\mathbf{1}_Q,\gamma)$ be the conservative result of **Pivot** based on the row sums of $Z(\overline{P},Q)$ and let $\overline{P}' = \overline{P} \setminus (\overline{L}\cup\overline{U})$ be the subgroup of experts which have not been classified by **Pivot**. The following proposition states that, provided that for some η the norm $\|\left[\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\right]_\eta\|_F^2$ is large enough compared to $|\overline{P}|\sqrt{|Q|}$, the resulting matrix $\Theta(\overline{P}',Q) - \overline{\Theta}(\overline{P}',Q)$ after **Pivot** has a significantly smaller norm. We shall often use the following quantity.

$$\phi_{l_1} = 2(2\zeta\sqrt{2} + \overline{\beta}_{\mathrm{tris}}) \le 29\zeta \quad . \tag{3.60}$$

Proposition 3.5.11. *Consider any $\overline{P}\subset[n]$, any $r\in\mathcal{R}$, and any subset $Q\subset Q_r$. Also, fix any $\eta>0$ and any $\phi>0$. If*

$$\|\left[\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\right]_\eta\|_F^2 \ge \frac{1}{\phi}\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \ge 8\phi_{l_1}\eta\sqrt{\log(\tfrac{2|\overline{P}|}{\delta})}|\overline{P}|\sqrt{|Q|} \quad ,$$

then, with probability higher than $1-\delta$, we have

$$\|\Theta(\overline{P}',Q) - \overline{\Theta}(\overline{P}',Q)\|_F^2 \le \left(1 - \frac{1}{16\phi}\right)\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \quad .$$

3.5.3.4 Analysis of DoubleTrisection – PCA

In this subsection, we state the main result regarding the trisection of a set $\overline{P}$ based on the first singular vector of a suitable matrix. We start from a subset $\overline{P}$ of experts. In **DoubleTrisection – Local**, we start applying **Pivot** and define $\widetilde{P} = \overline{P} \setminus (\overline{L}_{\mathrm{cp}}\cup\overline{U}_{\mathrm{cp}})$ as the set of experts that have not been classified by **Pivot**. We are given four independent samples $\mathcal{Z} = (Z^{(1)},Z^{(2)},Z^{(3)},Z^{(4)})$ according to the aggregated model (3.59). The first three samples are restricted to $\widetilde{P}$, whereas the last one concerns $\overline{P}$. Fix $\gamma\in[1,|\overline{P}|]$. We consider $(\overline{L},\overline{U}) = \mathbf{DoubleTrisection-PCA}(\mathcal{Z},\gamma)$ and $\overline{P}' = \widetilde{P}\setminus(\overline{L}_{\mathrm{pca}}\cup\overline{U}_{\mathrm{pca}})$ the set of experts that have not been classified by **DoubleTrisection – PCA**.

Recall the definition (3.60) of ϕ_{l_1}. Henceforth, the matrix $\Theta(\widetilde{P},Q)$ is said to be undistinguishable in l_1-norm if it satisfies

$$\max_{i,j\in\overline{P}}\|\Theta_{i,\cdot}(\widetilde{P},Q) - \Theta_{j,\cdot}(\widetilde{P},Q)\|_1 \le \phi_{l_1}\sqrt{|Q|\log\left(\tfrac{2|\overline{P}|}{\delta}\right)} \quad . \tag{3.61}$$

Since, up to permutation of its experts, the matrix $\Theta(\widetilde{P},Q)$ is bi-isotonic, the l_1 norm $\|\Theta_{i,\cdot}(\widetilde{P},Q) - \Theta_{j,\cdot}(\widetilde{P},Q)\|_1$ is simply the difference of the row sums of $\Theta(\widetilde{P},Q)$. Since $\widetilde{P}$ has been deduced from $\overline{P}$ by applying $\mathbf{Pivot}(Z,\gamma)$, we can safely assume that $\Theta(\widetilde{P},Q)$ is undistinguishable in l_1-norm with high probability – see the next subsection for a proper justification.

The next result states that, if $\Theta(\widetilde{P},Q)$ is undistinguishable in l_1-norm and if the Frobenius norm of $\Theta(\widetilde{P},Q) - \overline{\Theta}(\widetilde{P},Q)$ is large enough, then the corresponding matrix $\Theta(\overline{P}',Q)$ obtained after trisection has a significantly smaller Frobenius norm.

Proposition 3.5.12. *Let $\overline{P}\subset[n]$ and $Q\subset[d]$. If $\Theta(\widetilde{P},Q)$ is undistinguishable in l_1-norm and if*

$$\|\Theta(\widetilde{P},Q) - \overline{\Theta}(\widetilde{P},Q)\|_F^2 \ge 2\cdot 10^5\zeta^2\log^3\left(\frac{6nd}{\delta\zeta_-}\right)\left(\sqrt{|\widetilde{P}||Q|} + |\widetilde{P}|\right) \quad , \tag{3.62}$$

then, with probability higher than $1-3\delta$, we have

$$\|\Theta(\overline{P}',Q) - \overline{\Theta}(\overline{P}',Q)\|_F^2 \le \left(1 - \frac{1}{200\log^2(nd/\zeta_-)}\right)\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \quad .$$

Then, we gather the two previous results to analyze the routine **DoubleTrisection – Local**. Fix any $\overline{P}\subset[n]$ and $Q\subset[p]$. Let $\mathcal{Z} = (Z^{(1)}(\overline{P},Q),Z^{(2)}(\overline{P},Q),Z^{(3)}(\overline{P},Q),Z^{(4)}(\overline{P},Q),Z^{(5)}(\overline{P},Q))$ be five independent samples of the model (3.50). Fix any $\gamma\in[1,|\overline{P}|]$. Let $(\overline{L},\overline{U})$ be the conservative result of $\mathbf{DoubleTrisection-Local}(\mathcal{Z},\gamma)$ and $\overline{P}' = \overline{P}\setminus(\overline{L}\cup\overline{U})$. In the following, we write $\Theta(\overline{P},\mathcal{Q}_r)$ for the aggregation of $M(\overline{P})$ at all blocks of size r.

Corollary 3.5.13. *Fix any $r \in \mathcal{R}$. If, for some $\overline{P} \subset [n]$, $Q \subset \mathcal{Q}_r$, and $\eta > 0$, $\Theta(\overline{P}, Q)$ satisfies*

$$\begin{cases} \|\left[\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\right]_\eta\|_F^2 \geq \frac{1}{120\log^2\left(\frac{nd}{\delta\zeta_-}\right)}\|\Theta(\overline{P},\mathcal{Q}_r) - \overline{\Theta}(\overline{P},\mathcal{Q}_r)\|_F^2 \\ \|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \geq 4\cdot 10^5\zeta^2\log^3\left(\frac{6nd}{\delta\zeta_-}\right)\left[\frac{\eta}{\zeta}|\overline{P}|\sqrt{|Q|} \wedge \left(\sqrt{|\overline{P}||Q|} + |\overline{P}|\right)\right] \end{cases} , \tag{3.63}$$

then, with probability higher than $1-4\delta$, we have

$$\|\Theta(\overline{P}',\mathcal{Q}_r) - \overline{\Theta}(\overline{P}',\mathcal{Q}_r)\|_F^2 \leq \left(1 - \frac{1}{3\cdot 10^5\log^4\left(\frac{nd}{\delta\zeta_-}\right)}\right)\|\Theta(\overline{P},\mathcal{Q}_r) - \overline{\Theta}(\overline{P},\mathcal{Q}_r)\|_F^2 .$$

3.5.3.5 Analysis of BlockSort

Next, we combine the results of the previous sections to control the error of **BlockSort**. We are given a collection $\mathcal{Y}$ of $6\tau_\infty$ samples of the model (3.1) and a valid hierarchical sorting tree $\mathcal{T}$ of depth t that we consider as fixed. Then, we take a leaf G of $\mathcal{T}$ with maximal depth and we consider $(O,P,I) = \textbf{BlockSort}(\mathcal{Y},\mathcal{T},G)$ the trisection of G, as well as $\overline{P} \supseteq P$ the more conservative intermediary set. In this section, we provide a high-probability control of $\overline{P}$.

For any height $h \in \mathcal{H}$ and scale $r \in \mathcal{R}$, recall that Q^*_{cp} is the subset(defined in Section 3.5.3.2) of block of questions at scale r such that the mean $\overline{m}(\overline{P})$ increases by at least $h/2$. Also recall the superset $\overline{Q}^*_{\mathrm{cp}} \supset Q^*_{\mathrm{cp}}$.

At a high level, the next proposition states that, after τ_∞ iterations of the **DoubleTrisection – Local** routines at all scales $r \in \mathcal{R}$ and all heights $h \in \mathcal{H}$, the size $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$ is quite small. This is mainly due to the fact that, by Lemma 3.5.10, at each step τ, there exists some $(r,h) \in \mathcal{R}\times\mathcal{H}$ such that the norm of the thresholded aggregated matrix $\left[\Theta(\overline{P}_\tau, Q^*_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}_\tau, Q^*_{\mathrm{cp}})\right]_{\sqrt{r}h}$ is of the same order as $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$. By Lemma 3.5.8, the estimated blocks $\widehat{Q}^*_{\mathrm{cp}}$ contain Q^*_{cp} with high probability. Hence, unless the norm of the thresholded aggregated matrix is small, we derive from corollary 3.5.13 that the norm of this aggregated matrix has contracted at step $\tau+1$. Hence, after τ_∞ steps, one could expect that the norm of $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$ is small. In fact, both the statement and the proof of this proposition are slightly more involved because we need to keep track of the scales and heights of interest. Define the function $\Psi(p,r,h,q)$ by

$$\Psi(p,r,h,q) = \frac{hp\sqrt{rq}}{\zeta} \wedge (\sqrt{pq}) + p . \tag{3.64}$$

Proposition 3.5.14. *With probability higher than $1 - 5\tau_\infty\delta$, there exists a subset $\overline{P}^\dagger$ such that $\overline{P} \subseteq \overline{P}^\dagger \subseteq G$ and the following property holds. For some $r^\dagger \in \mathcal{R}$ and some $h^\dagger \in \mathcal{H}$, upon writing $Q^\dagger_{\mathrm{cp}} = Q^*_{\mathrm{cp}}(\overline{P}^\dagger, h^\dagger, r^\dagger)$ and $\overline{Q}^\dagger_{\mathrm{cp}} = \overline{Q}^*_{\mathrm{cp}}(\overline{P}^\dagger, h^\dagger, r^\dagger)$, we have simultaneously*

$$\|\left[\Theta(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}})\right]_{\sqrt{r^\dagger}h^\dagger}\|_F^2 \leq 4\cdot 10^5\zeta^2\log^3\left(\frac{6nd}{\delta\zeta_-}\right)\Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{\mathrm{cp}}|) ; \tag{3.65}$$

$$\|M(\overline{P}^\dagger) - \overline{M}(\overline{P}^\dagger)\|_F^2 \leq 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}|\|[\Theta(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}})]_{\sqrt{r^\dagger}h^\dagger}\|_F^2 . \tag{3.66}$$

In other words, there exists a superset $\overline{P}^\dagger$ of $\overline{P}$ such that, for a suitable height and scale, at the high-variation regions, both the original matrix $M(\overline{P}^\dagger)$ and the thresholded aggregated matrix are controlled at the level $\Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{\mathrm{cp}}|)$. The virtue of the above result is that it easily adapts to the block sorting variant with memory. Unfortunately, the rate $\Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{\mathrm{cp}}|)$ is a bit difficult to handle. In the next corollary, we replace it by a simpler but cruder bound that only depends on $|G|$, $h^\dagger$ and d.

Corollary 3.5.15. *Under the same event of probability higher than $1-5\tau_\infty\delta$ as in the previous proposition, the set $\overline{P}^\dagger$, the scale $r^\dagger$, and the height $h^\dagger$ also satisfy*

$$\|\left[\Theta(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}})\right]_{\sqrt{r^\dagger}h^\dagger}\|_F^2 \lesssim \zeta^2\log^{3.5}\left(\frac{6nd}{\delta\zeta_-}\right)\left[\frac{h^\dagger|G|\sqrt{d}}{\zeta} \wedge \sqrt{|G|d} \wedge \sqrt{\frac{|G|}{h^\dagger}} + |G|\right] , \tag{3.67}$$

where we recall that G is the initial group.

3.5.3.6 Analysis of the complete procedure TreeSort

We are now equipped to prove Proposition 3.5.5.

Proof of Proposition 3.5.5. Let us fix an integer $t \le t_\infty$ and let us consider the collection $\overline{\mathcal{L}}_t$ of the 2^t groups $\overline{P}$ that are not sorted with confidence. Let us apply Proposition 3.5.14 to each of these sets $\overline{P}$. In view of this proposition, we define $(\overline{P}^\dagger, h^\dagger, r^\dagger)$ as well as $Q^\dagger_{\mathrm{cp}}$. We also define

$$s^\dagger = \Big|\Big\{l \ : \ \exists i,j \in \overline{P}^\dagger \text{ s.t. } \Big[\Theta_{i,l}(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) - \Theta_{j,l}(\overline{P}^\dagger, Q^\dagger_{\mathrm{cp}})\Big]_{\sqrt{r^\dagger}h^\dagger} \neq 0\Big\}\Big| \ .$$

In a nustshell, $s^\dagger$ is the number of columns of the thresholded aggregated matrix which are not equal to zero.

Definition 3. Define the dyadic collection $\mathcal{S} = \{1, 2, 4, \dots, 2^{\lceil \log_2(d) \rceil}\}$. For any $s \in \mathcal{S}$, $r \in \mathcal{R}$, and $h \in \mathcal{H}$, we consider the collection $\mathcal{P}^*(h,r,s) \subset \overline{\mathcal{L}}_t$ satisfying $r^\dagger = r$, $h^\dagger = h$, and $s^\dagger \in [s, 2s)$.

The following lemma controls the cardinality of $\mathcal{P}^*(h,r,s)$. This bound mainly relies on the facts that the matrix M is, up to a row permutation, bi-isotonic and that its entries lie in $[0,1]$.

Lemma 3.5.16. *Assume that there exists an ordering σ of $\overline{\mathcal{L}}_t$ that orders all groups $\overline{P}$'s. In other words, for any $r \le s$, any expert $i \in \overline{P}_{\sigma(r)}$ is below any expert $j \in \overline{P}_{\sigma(s)}$. Then, upon this assumption,*

$$|\mathcal{P}^*(h,r,s)| \le \frac{2d}{hrs} \wedge 2^t \le \sqrt{\frac{d2^{t+1}}{hrs}} \ ,$$

for any $h \in \mathcal{H}$, $r \in \mathcal{R}$, and $s \in \mathcal{S}$.

In fact, all the collections $\overline{\mathcal{L}}_t$ with $t = 0, \dots, t_\infty$ satisfy the assumption in the above under the event ξ defined in Corollary 3.5.4 –see the proof of Proposition 3.5.3.

Putting everything together and summing over the groups $\mathcal{P}^*(h,r,s)$, we derive from Proposition 3.5.14 and Corollary 3.5.15 that, with probability higher than $1 - 5 \cdot 2^t \tau_\infty \delta$, we have

$$\begin{aligned}
\sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 &\le 16\zeta^2 |\overline{\mathcal{L}}_t| + 96|\mathcal{R}||\mathcal{H}| \sum_{h,r,s} \sum_{\overline{P} \in \mathcal{P}^*(h,r,s)} \| \left[\Theta(\overline{P}, Q^\dagger_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}, Q^\dagger_{\mathrm{cp}})\right]_{\sqrt{r}h} \|_F^2 \\
&\overset{(a)}{\lesssim} \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{h,r,s} \sum_{\overline{P} \in \mathcal{P}^*(h,r,s)} \left[\left(\frac{n}{2^t\zeta^2} s^\dagger r h^2\right) \wedge \left(\sqrt{\frac{n}{2^t h}}\right) + \frac{n}{2^t}\right] \\
&\overset{(b)}{\lesssim} \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{h,r,s} \left[\frac{nsrh^2}{\zeta^2} \wedge \sqrt{\frac{dn}{srh^2}} + n\right] \\
&\overset{(c)}{\lesssim} \zeta^2 \log^{8.5}\left(\frac{6nd}{\delta\zeta_-}\right) \left[\frac{n^{2/3}d^{1/3}}{\zeta^{2/3}} + n\right] \ ,
\end{aligned}$$

where in (a), we combined Corollary 3.5.15 with the fact that the size of each group is at most $n/2^t$, the crude bound $\|[A]_\eta\|_F^2 \le \eta^2 d_1 d_2$ for any $d_1 \times d_2$ matrix and that $n/2^t \ge 1$. In (b), we relied on Lemma 3.5.16, whereas in (c) we used that $x \wedge y \le x^{1/3} y^{2/3}$.

We have proved the desired $n^{2/3}d^{1/3}/\zeta^{2/3} + n$ upper bound. The rate $nd^{1/6}/\zeta^{1/3}$ is proved using the same scheme except that we apply Corollary 3.5.15 differently in (a). More precisely, we have

$$\begin{aligned}
\sum_{\overline{P} \in \overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 &\le 16\zeta^2 |\overline{\mathcal{L}}_t| + 96|\mathcal{R}||\mathcal{H}| \sum_{h,r,s} \sum_{\overline{P} \in \mathcal{P}^*(h,r,s)} \| \left[\Theta(\overline{P}, Q^\dagger_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}, Q^\dagger_{\mathrm{cp}})\right]_{\sqrt{r}h} \|_F^2 \\
&\le \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{h,r,s} \sum_{\overline{P} \in \mathcal{P}^*(h,r,s)} \left[\left(\frac{n}{2^t\zeta} h\sqrt{d}\right) \wedge \left(\sqrt{\frac{n}{2^t h}}\right) \wedge \sqrt{\frac{n}{2^t}d} + \frac{n}{2^t}\right] \\
&\lesssim \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_h \sum_{r,s} |\mathcal{P}^*(h,r,s)| \left[\left(\frac{n}{2^t\zeta} h\sqrt{d}\right) \wedge \left(\sqrt{\frac{n}{2^t h}}\right) \wedge \sqrt{\frac{n}{2^t}d} + \frac{n}{2^t}\right] \\
&\overset{(a')}{\lesssim} \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_h \left[\frac{nh\sqrt{d}}{\zeta} \wedge \sqrt{\frac{n^2}{h}} \wedge \sqrt{n2^t d} + n\right] \\
&\overset{(b')}{\lesssim} \zeta^2 \log^{6.5}\left(\frac{6nd}{\delta\zeta_-}\right) \left[\frac{nd^{1/6}}{\zeta^{1/3}} \wedge n\sqrt{d} + n\right] \ ,
\end{aligned}$$

where in (a'), we used that $\sum_{r,s} |\mathcal{P}^*(h,r,s)| \le 2^t \le 2n$ and in (b') that $xy \le x^{1/3} y^{2/3}$. □

Proof of Lemma 3.5.16. To ease the notation, we write $\mathcal{P}^* = \mathcal{P}^*(h,r,s)$ in this proof. Since $\mathcal{P}^* \subset \overline{\mathcal{L}}_t$, we straightforwardly derive that $|\mathcal{P}^*| \leq |\overline{\mathcal{L}}_t| \leq 2^t$. Let us introduce the width of a matrix $\Theta \in \mathbb{R}^{[n]\times\mathcal{Q}_r}$ on the set $P \subset [n]$ and $Q \subset \mathcal{Q}_r$:

$$\mathrm{W}_{\infty,1}(\Theta, P, Q) = \max_{i,j\in P} \sum_{l\in Q} |\Theta_{i,l} - \Theta_{j,l}| \ .$$

Consider any set $\overline{P}$ and the corresponding quantities $\overline{P}^\dagger$, $s^\dagger$, $r^\dagger$, and $h^\dagger$. By definition of $s^\dagger$, we have $\mathrm{W}_{\infty,1}(\Theta, \overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) \geq s^\dagger\sqrt{r^\dagger}h^\dagger$. Recall that the matrix Θ is, up to a permutation of its rows, bi-isotonic. Besides, all the groups $\overline{P}$ in $\overline{\mathcal{L}}_t$ are perfectly ordered by assumption. As a consequence, the width of Θ on $[n]$ is larger or equal to the sum of the width on each set $\overline{P}$. Since $\mathcal{P}^*$ is an ordered sub-partition, it holds that

$$\mathrm{W}_{\infty,1}(\Theta, [n], \mathcal{Q}_r) \geq \sum_{\overline{P}\in\mathcal{P}^*} \mathrm{W}_{\infty,1}(\Theta, \overline{P}^\dagger, \mathcal{Q}_r) \geq \sum_{\overline{P}\in\mathcal{P}^*} \mathrm{W}_{\infty,1}(\Theta, \overline{P}^\dagger, Q^\dagger_{\mathrm{cp}}) \geq |\mathcal{P}^*| s^\dagger\sqrt{r}h \geq |\mathcal{P}^*| s\sqrt{r}h \ . \tag{3.68}$$

By definition of $\mathcal{Q}_r$, we have $|\mathcal{Q}_r| \leq 2d/r$. Since the values of Θ lie in $[0,\sqrt{r}]$, we deduce that $\mathrm{W}_{\infty,1}(\Theta,[n],\mathcal{Q}_r) \leq 2d/\sqrt{r}$. Together with (3.68), this yields

$$|\mathcal{P}^*| \leq \frac{2d}{rsh} \ ,$$

which concludes the proof.

□

3.5.4 Remaining proofs for Proposition 3.5.5

3.5.4.1 Proofs of the results on DimensionReduction (Section 3.5.3.2)

Proof of Lemma 3.5.8. It is sufficient to prove that $D^*_{\mathrm{cp}}(h,r) \subset \widehat{D}_{\mathrm{cp}}(h,r) \subset \overline{D}^*_{\mathrm{cp}}(h,r)$. Recall that we use the convention that $\overline{y}_i = \overline{m}_i = 0$ if $i \leq 0$ and $\overline{y}_i = \overline{m}_i = 1$ if $i > d$. Since the CUSUM statistic is linear, we have the decomposition

$$\widehat{\mathbf{C}}_{k,\tilde{r}} = \mathbf{C}^*_{k,\tilde{r}} + \frac{1}{\tilde{r}}\left(\sum_{k'=k}^{k+\tilde{r}-1} \overline{e}_{k'}(\overline{P}) - \sum_{k'=k-\tilde{r}}^{k-1} \overline{e}_{k'}(\overline{P})\right) \ ,$$

where the latter random variable is centered and $\zeta(|\overline{P}|\tilde{r}/2)^{-1/2}$-subGaussian. By a union bound, we derive that, with probability higher than $1-\delta$, we have

$$\max_{k\in[d]} \left|\widehat{\mathbf{C}}_{k,\tilde{r}} - \mathbf{C}^*_{k,\tilde{r}}\right| \leq \zeta\sqrt{\frac{2\cdot 2}{|\overline{P}|\tilde{r}}\log\left(\frac{2d}{\delta}\right)} \ .$$

Since $\tilde{r}$ is defined in such a way that

$$\zeta\sqrt{\frac{4}{|\overline{P}|}\log\left(\frac{2d}{\delta}\right)} \leq \frac{1}{8}\sqrt{\tilde{r}}h \ ,$$

we deduce that $\widehat{D}_{\mathrm{cp}}(h,r) \subset \overline{D}^*_{\mathrm{cp}}(h,r)$. Conversely, if k belongs to $D^*_{\mathrm{cp}}(h,r)$, we have $\mathbf{C}^*_{k,8r} \geq h/2$. Since $\overline{m}(\overline{P})$ is an isotonic vector and $\tilde{r} \geq 8r$, it follows that $\mathbf{C}^*_{k,\tilde{r}} \geq \mathbf{C}^*_{k,8r} \geq h/2$. We deduce that

$$\widehat{\mathbf{C}}_{k,\tilde{r}} \geq h[\tfrac{1}{2} - \tfrac{1}{8}] \geq \frac{h}{4} \ ,$$

which implies that $D^*_{\mathrm{cp}}(h,r) \subset \widehat{D}_{\mathrm{cp}}(h,r)$. □

Proof of Lemma 3.5.9. If an index k belongs to $\overline{D}^*(h,r)$, this implies that $\overline{m}_{k+\tilde{r}} - \overline{m}_{k-\tilde{r}} \geq h/8$, since the vector $\overline{m}$ is isotonic. Define $\kappa = 1+\lceil \tilde{r}/r \rceil$. Since $\overline{m}$ is an isotonic vector, for $l \in \overline{Q}^*_{\mathrm{cp}}(h,r)$, we deduce that $\overline{m}_{l+\kappa r} - \overline{m}_{l-\kappa r} \geq h/8$. Consider the regular grid $\mathcal{Q}_{\kappa r}$ of width κr and define $\overline{Q}^*(h,r,\kappa) = \{l \in \mathcal{Q}_{\kappa r} : \overline{Q}^*_{\mathrm{cp}}(h,r) \cap [l, l+\kappa r) \neq \varnothing\}$. Since, for $l \in \overline{Q}^*(h,r,\kappa)$, we have $\overline{m}_{l+2\kappa r} - \overline{m}_{l-2\kappa r} \geq h/8$ and since the total variation of $\overline{m}$ is at most one, this implies

$$\frac{h}{8}|\overline{Q}^*(h,r,\kappa)| \leq \sum_{l\in D(\kappa,r,h)} \overline{m}_{l+2\kappa r} - \overline{m}_{l-2\kappa r} \leq \sum_{l\in\mathcal{Q}_{\kappa r}} \overline{m}_{l+2\kappa r} - \overline{m}_{l-2\kappa r} \leq 4 \ .$$

Since $|\overline{Q}^*(h,r)| \leq \kappa|\overline{Q}^*(h,r,\kappa)|$ and since $\kappa \leq 2\tilde{r}/r$, we obtain the desired result.

□

Proof of Lemma 3.5.10. For any height $h \in \mathcal{H}$ –recall the definition of the dyadic class $\mathcal{H}$ in (3.36)– and any expert $i \in \overline{P}$, we consider the h-level set $M_{i,.} - \overline{m}$, that is

$$F(i,h) = \{k \in [d] \ : \ M_{i,k} - \overline{m}_k \geq h\} \ ; \qquad F(i,-h) = \{k \in [d] \ : \ M_{i,k} - \overline{m}_k \leq -h\} \ . \tag{3.69}$$

Since $F(i,h)$ and $F(i,-h)$ are subsets of $[d]$, we can decompose them into unions of disjoint intervals. For any positive integer $r \in \mathcal{R}$, we write $F(i,h,r)$ as the union of intervals of $F(i,h)$ whose size belongs $[2r-1, 4r-1)$. Finally, we consider the subset $F(i,h,r;2h) \subset F(i,h,r)$ of all intervals of $F(i,h,r)$ that intersect $F(i,2h)$. In other words, any maximal interval I in $F(i,h,r;2h)$ is a h-level set whose size belongs to $[2r-1, 4r-1)$ and such that $M_{i,.} - \overline{m}$ crosses the level $2h$ in I. We define similarly $F(i,h,r)$ and $F(i,h,r;2h)$ when h is negative and $-h \in \mathcal{H}$. It follows from these definitions that, for any h such that either $h \in \mathcal{H}$ or $-h \in \mathcal{H}$, we have

$$F(i,2h) \subset \bigcup_{r \in \mathcal{R}} F(i,h,r;2h) \ . \tag{3.70}$$

We define $F^*(h,r,2h)$ as the union of those intervals for $i \in \overline{P}$.

$$F^*(h,r,2h) = \bigcup_{i \in \overline{P}} F(i,h,r;2h) \ .$$

First, we claim that this collection of intervals $F^*(h,r,2h)$ is contained in the significant regions of variation of $\overline{m}$. This result heavily relies on the monotonicity assumptions.

Lemma 3.5.17. *For any $h \in \mathcal{H}$ and any $r \in \mathcal{R}$.*

$$[F^*(h,r,2h) \bigcup F^*(-h,r,-2h)] \subset D^*_{\mathrm{cp}}(h,r) \ .$$

Next, we quantify $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$ using regions of large variation of $M_{i,.} - \overline{m}$.

Lemma 3.5.18. *For any $\overline{P}$, it holds that*

$$\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \leq 16 \left[\zeta^2 + \sum_{i \in \overline{P}} \sum_{r \in \mathcal{R}, h \in \mathcal{H}} h^2 \left(|F(i,h,r;2h)| + |F(i,-h,r;-2h)| \right) \right] \ .$$

The last lemma connects these sets $|F(i,h,r;2h)|$ to the norm of the thresholded aggregated matrix.

Lemma 3.5.19. *For any $r \in \mathcal{R}$ and $h \in \mathcal{H}$, we consider $\Theta(\overline{P}, Q^*_{\mathrm{cp}}(h,r))$ the aggregation of M at scale r and at $Q^*_{\mathrm{cp}}(h,r)$. We have*

$$h^2 \sum_{i \in \overline{P}} [|F(i,h,r;2h)| + |F(i,-h,r;-2h)|] \leq 3 \| \left[\Theta(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) - \overline{\Theta}(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) \right]_{\sqrt{r}h} \|_F^2 \ .$$

Combining Lemmas 3.5.18 and 3.5.19, we conclude that

$$\begin{aligned}
\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 &\leq 16 \left[\zeta^2 + \sum_{i \in \overline{P}} \sum_{r \in \mathcal{R}, \ h \in \mathcal{H}} h^2 (|F(i,h,r;2h)| + |F(i,-h,r;-2h)|) \right] \\
&\leq 16 \left[\zeta^2 + 3 \sum_{r \in \mathcal{R}, \ h \in \mathcal{H}} \| \left[\Theta(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) - \overline{\Theta}(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) \right]_{\sqrt{r}h} \|_F^2 \right] \\
&\leq 16 \left[\zeta^2 + 6|\mathcal{R}||\mathcal{H}| \max_{r \in \mathcal{R}, \ h \in \mathcal{H}} \left\| \left[\Theta(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) - \overline{\Theta}(\overline{P}, Q^*_{\mathrm{cp}}(h,r)) \right]_{\sqrt{r}h} \right\|_F^2 \right] \ ,
\end{aligned}$$

which concludes the proof of Lemma 3.5.10.

□

Proof of Lemma 3.5.17. Consider any $i \in \overline{P}$, any height $h \in \mathcal{H}$, and any scale $r \in \mathcal{R}$. Without loss of generality, we only focus on $F(i,h,r;2h)$; the case of $F(i,-h,r;-2h)$ being analogous. Let I be an interval of $F(i,h,r;2h)$. Fix any question $k \in I$ such that $|M_{i,k} - \overline{m}_k| \geq 2h$. Since $k \in F(i,h,r)$, it follows that there exists $l < 4r$ such that $M_{i,k+l} - \overline{m}_{k+l} \leq h$. Since both the vectors $M_{i,.}$ and $\overline{m}$ are isotonic, it follows that $\overline{m}_{k+l} - \overline{m}_k \geq h$. Now consider any $k_0 \in I$. Using again the monotonicity of $\overline{m}$, we deduce that,

$$\mathbf{C}^*_{k_0, 8r} \geq \frac{4rh}{8r} \geq \frac{1}{2} h \ ,$$

and k_0 therefore belongs to $D^*_{\mathrm{cp}}(h,r)$. We have proved the desired result. □

Proof of Lemma 3.5.18. Consider any expert $i \in \overline{P}$. We decompose the norm of $[M_{i,\cdot} - \overline{m}(\overline{P})]$ using the level sets of this vector. We recall that $\mathcal{H}$ is of the form $\{h_{\min}, 2h_{\min}, 4h_{\min}, \dots\}$ where $h_{\min} \in [\zeta^2/nd, 2\zeta^2/nd]$.

$$\begin{aligned}
\|M_{i,\cdot} - \overline{m}(\overline{P})\|_2^2 &\le \sum_{h\in\mathcal{H}} \sum_{k=1}^{d} (M_{i,k} - \overline{m}_k(\overline{P}))^2 \mathbf{1}\{2h \le |M_{i,k} - \overline{m}_k(\overline{P})| < 4h\} + 4dh_{\min}^2 \\
&\le 16 \sum_{h\in\mathcal{H}} \sum_{k=1}^{d} h^2 \mathbf{1}\{2h \le |M_{i,k} - \overline{m}_k(\overline{P})| < 4h\} + \frac{16\zeta^2}{n^2 d} \\
&\le 16 \sum_{h\in\mathcal{H}} h^2 |F(i,2h)| + \frac{16\zeta^2}{n^2 d} \\
&\le 16 \sum_{h\in\mathcal{H}} \sum_{r\in\mathcal{R}} h^2 |F(i,h,r;2h)| + \frac{16\zeta^2}{n^2 d} \ ,
\end{aligned}$$

where in the last line, we used (3.70). Then, we sum over $i \in \overline{P}$ to conclude. □

Proof of Lemma 3.5.19. Consider any $i \in \overline{P}$, any height $h \in \mathcal{H}$, and any scale $r \in \mathcal{R}$. Without loss of generality, we only consider $F(i,h,r;2h)$ the case of $F(i,-h,r;-2h)$ being analogous. Let I be a maximal interval of $F(i,h,r;2h)$. We deduce from Lemma 3.5.17 that I is included in $D^*_{\rm cp}(h,r)$. Let I_0 be the largest sub-interval of I of the form $[qr, q'r)$ where q and $q' \in \mathcal{Q}_r$. Since $|I| \ge 2r-1$, it follows that $|I| \le 3|I_0|$. We write L_0 the subset of columns of the aggregated matrix $\Theta(\overline{P}, Q^*_{\rm cp}(h,r))$ corresponding to I_0 so that $|L_0| = |I_0|/r$. On each column l of L_0, we have $\Theta_{i,l}(\overline{P}, Q^*_{\rm cp}(h,r)) - \overline{\theta}_l(\overline{P}, Q^*_{\rm cp}(h,r)) \ge \sqrt{r}h$. Putting everything together, we get

$$h^2|I| \quad \le \quad 3h^2|I_0| = 3h^2 r|L_0| \le 3 \sum_{l\in L_0} \Big(\big[\Theta(\overline{P}, Q^*_{\rm cp}(h,r))_{i,l} - \overline{\theta}(\overline{P}, Q^*_{\rm cp}(h,r))_l\big]_{\sqrt{r}h} \Big)^2 \ .$$

Summing over all intervals I and over all experts $i \in \overline{P}$ and also accounting for the $F[i,-h,r;-2h]$ concludes the proof. □

3.5.4.2 Proof of Proposition 3.5.11

To simplify the notation, we define $\Phi_{l_1} = 2(2\zeta\sqrt{2} + \overline{\beta}_{\rm tris})\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} = \phi_{l_1}\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)}$.

For simplicity, we respectivly write $\Theta(\overline{P}) = \Theta(\overline{P}, Q)$ and $\Theta(\overline{P}') = \Theta(\overline{P}', Q)$ in this proof. Recall that $\overline{\theta}(\overline{P})$ stands the mean row of $\Theta(\overline{P})$ whereas $\overline{\theta}(\overline{P}')$ stands for the mean row of $\Theta(\overline{P}')$.

Invoking Lemma 3.5.7 with $w = \mathbf{1}_Q$ and since the matrix Θ is isotonic, we deduce that outside an event of probability smaller than δ, we have

$$\max_{i,j\in\overline{P}'} \|\Theta(\overline{P}')_{i,\cdot} - \Theta(\overline{P}')_{j,\cdot}\|_1 \le \Phi_{l_1}\sqrt{|Q|} \ . \tag{3.71}$$

since the matrix Θ is isotonic. We shall deduce from this inequality the desired bound. We consider two cases depending on the difference between $\overline{\theta}(\overline{P})$ and $\overline{\theta}(\overline{P}')$ the mean rows in $\overline{P}$ and $\overline{P}'$.

Case 1: $|\overline{P}'| \cdot \|\overline{\theta}(\overline{P}) - \overline{\theta}(\overline{P}')\|_2^2 > \frac{1}{16\phi}\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2$. Since $\overline{P}' \subset \overline{P}$, we deduce that

$$\begin{aligned}
\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 - \|\Theta(\overline{P}') - \overline{\Theta}(\overline{P}')\|_F^2 &\ge \sum_{i\in\overline{P}'} \|\Theta(\overline{P})_{i,\cdot} - \overline{\theta}(\overline{P})\|_2^2 - \|\Theta(\overline{P})_{i,\cdot} - \overline{\theta}(\overline{P}')\|_2^2 \\
&= |\overline{P}'| \cdot \|\overline{\theta}(\overline{P}) - \overline{\theta}(\overline{P}')\|_2^2 \\
&\ge \frac{1}{16\phi}\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 \ ,
\end{aligned}$$

where we used the condition in the last line. We have proved the desired result.
Case 2: $|\overline{P}'| \cdot \|\overline{\theta}(\overline{P}) - \overline{\theta}(\overline{P}')\|_2^2 \le \frac{1}{16\phi}\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2$. We start with the decomposition

$$\|\Theta(\overline{P}') - \overline{\Theta}(\overline{P}')\|_F^2 \le \|\Theta(\overline{P}') - \overline{\Theta}(\overline{P})\|_F^2 = \|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 - \|\Theta(\overline{P} \setminus \overline{P}') - \overline{\Theta}(\overline{P})\|_F^2 \ , \tag{3.72}$$

so that we only have to control $\|\Theta(\overline{P} \setminus \overline{P}') - \overline{\Theta}(\overline{P})\|_F^2$ from below. By definition of the operator $[\cdot]_\eta$, we have

$$\|\Theta(\overline{P} \setminus \overline{P}') - \overline{\Theta}(\overline{P})\|_F^2 \ge \|[\Theta(\overline{P} \setminus \overline{P}') - \overline{\Theta}(\overline{P})]_\eta\|_F^2 = \|[\Theta(\overline{P}) - \overline{\Theta}(\overline{P})]_\eta\|_F^2 - \|[\Theta(\overline{P}') - \overline{\Theta}(\overline{P})]_\eta\|_F^2.$$

By assumption, we have $\|[\Theta(\overline{P}) - \overline{\Theta}(\overline{P})]_\eta\|_F^2 \geq \phi^{-1}\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2$. Hence, as long as we prove that

$$\|[\Theta(\overline{P}') - \overline{\Theta}(\overline{P})]_\eta\|_F^2 \leq (2\phi)^{-1}\|[\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 \ , \tag{3.73}$$

we can safely conclude from (3.72) that

$$\|\Theta(\overline{P}') - \overline{\Theta}(\overline{P}')\|_F^2 \leq (1 - (2\phi)^{-1})\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 \ .$$

Thus, we only have to prove (3.73). Again, by definition of the thresholding operator, we have

$$\begin{aligned} \|[\Theta(\overline{P}') - \overline{\Theta}(\overline{P})]_\eta\|_F^2 &= \eta^2 \sum_{i\in\overline{P}',\ l\in Q} \mathbf{1}_{|\Theta_{i,l} - \overline{\theta}(\overline{P})_l| \geq \eta} \\ &\leq \eta^2 \sum_{i\in\overline{P}',\ l\in Q} \mathbf{1}_{|\Theta_{i,l} - \overline{\theta}(\overline{P}')_l| \geq \eta/2} + \mathbf{1}_{|\theta(\overline{P}')_l - \overline{\theta}(\overline{P})_l| \geq \eta/2} \ . \end{aligned} \tag{3.74}$$

By Markov inequality, the condition that defines Case 2 above implies that

$$\sum_{l\in Q} \mathbf{1}\{|\overline{\theta}(\overline{P}')_l - \overline{\theta}(\overline{P})_l| \geq \eta/2\} \leq \frac{1}{4\phi}\frac{\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2}{|\overline{P}'|\eta^2} \ . \tag{3.75}$$

From (3.71) and a convexity argument, we deduce that, for any $i \in \overline{P}'$, $\|\Theta(\overline{P}')_{i,\cdot} - \overline{\theta}(\overline{P}')\|_1 \leq \Phi_{l_1}\sqrt{|Q|}$. Then, applying again Markov inequality, we deduce that, for any expert i in $\overline{P}'$ and any $\eta > 0$, we have

$$\sum_{l\in Q} \mathbf{1}\{|\Theta(\overline{P}')_{i,l} - \overline{\theta}(\overline{P}')_l| \geq \eta/2\} \leq 2\Phi_{l_1}\frac{\sqrt{|Q|}}{\eta} \ .$$

Since we assume that $\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 \geq 8\phi\Phi_{l_1}\eta|\overline{P}|\sqrt{|Q|} \geq 8\phi\Phi_{l_1}\eta|\overline{P}'|\sqrt{|Q|}$, we deduce that

$$\sum_{l\in Q} \mathbf{1}\{|\Theta(\overline{P}')_{i,l} - \overline{\theta}(\overline{P}')_l| \geq \eta/2\} \leq \frac{1}{4\phi}\frac{\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2}{|\overline{P}'|\eta^2} \ . \tag{3.76}$$

So that, combining (3.74), (3.75) and (3.76), we arrive at

$$\|[\Theta(\overline{P}') - \overline{\Theta}(\overline{P})]_\eta\|_F^2 \leq \frac{1}{2\phi}\|\Theta(\overline{P}) - \overline{\Theta}(\overline{P})\|_F^2 \ .$$

We have proved (3.73).

3.5.4.3 Proof of Proposition 3.5.12

For simplicity, we write in this proof $\Theta := \Theta(\widetilde{P}, Q)$ and $\Theta(\overline{P}') := \Theta(\overline{P}', Q)$. Without loss of generality, we assume that the rows of Θ are already ordered according to the oracle order so that Θ is bi-isotonic.

First, the following lemma states that, the first singular value of $(\Theta - \overline{\Theta})$ is, up to polylogarithmic terms, of the same order as its Frobenius norm. This is mainly due to the fact that the entries of Θ lies in $[0, \sqrt{r}]$ and that Θ is a bi-isotonic matrix.

Lemma 3.5.20. *Assume that $\|\Theta - \overline{\Theta}\|_F \geq 2\zeta$. For any sets $\widetilde{P}$ and Q, we have*

$$\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \geq \frac{1}{16\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ .$$

Now, write $\hat{v} = \arg\max_{\|v\|_2\leq 1}\left[\|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2\right]$.

Lemma 3.5.21. *Fix any $\delta \in (0,1)$. If*

$$\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \geq 1600\zeta^2\left[\sqrt{|Q|(5|\widetilde{P}| + \log(6/\delta))} + 7|\widetilde{P}| + 2\log(6/\delta)\right] \ , \tag{3.77}$$

then, with probability higher than $1 - \delta$, we have

$$\|\hat{v}^T\left(\Theta - \overline{\Theta}\right)\|_2^2 \geq \frac{1}{2}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \ .$$

In light of Lemma 3.5.20 and Condition (3.62), the Condition (3.77) in Lemma 3.5.21 is valid. Consequently, there exists an event of probability higher than $1-\delta$ such that

$$\|\hat{v}^T\left(\Theta-\overline{\Theta}\right)\|_2^2 \geq \frac{1}{32\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 \ . \tag{3.78}$$

Next, we show that a thresholded version of $\hat{z}=(Z^{(3)}-\overline{Z}^{(3)})^T\hat{v}$ is almost aligned with $z^*=(\Theta-\overline{\Theta})^T\hat{v}$. We define the sets $S^*\subset Q$ and $\hat{S}\subset Q$ of blocks of questions by

$$S^* \ = \ \left\{l\in Q \ : \ |z_l^*|\geq 3\zeta\sqrt{2\log(2|Q|/\delta)}\right\} \ ; \quad \hat{S}=\left\{l\in Q \ : \ |\hat{z}_l|\geq 2\zeta\sqrt{2\log(2|Q|/\delta)}\right\} \ .$$

S^* stands for the collection of blocks of questions l such that z_l^* is large whereas $\hat{S}$ is the collection of blocks l with large $\hat{z}_l$. Finally, we consider the vectors w^* and $\hat{w}$ defined as theresholded versions of z^* and $\hat{z}$ respectively, that is $w_i^*=z_i^*1_{i\in S^*}$ and $\hat{w}_i=\hat{z}_i\mathbf{1}_{i\in\hat{S}}$. Note that, up to the sign, $\hat{w}$ stands for the active coordinates computed in **DoubleTrisection – PCA**.

We write v for any unit vector in $\mathbb{R}^{|\widetilde{P}|}$. Since the noise matrix $N^{(2)}$ is made of independent ζ-subGaussian random variables, it follows that $(v^T(N^{(3)}-\overline{N}^{(3)}))_l$ is a ζ-subGaussian random variables. Hence, we deduce that, for any fixed matrix Θ, subsets $\overline{P}$ and Q, and any unit vector v, we have

$$\mathbb{P}\left[\max_{l\in Q}\left|(v^T(N^{(3)}-\overline{N}^{(3)}))_l\right|\leq\zeta\sqrt{2\log(2|Q|/\delta)}\right]\geq 1-\delta \ .$$

Observe that $\hat{z}=z^*+\hat{v}^T(N^{(3)}-\overline{N}^{(3)})$. Conditioning on $\hat{v}$, we deduce that, on an event of probability higher than $1-\delta$, we have

$$\|\hat{z}-z^*\|_\infty\leq\zeta\sqrt{2\log(2|Q|/\delta)} \ . \tag{3.79}$$

Under this event, we have $S^*\subset\hat{S}$ and for $l\in\hat{S}$, we have $z_l^*/\hat{z}_l\in[1/2,2]$. Next, we shall prove that, under this event, $\hat{v}^T(\Theta-\overline{\Theta})\hat{w}/\|\hat{w}\|_2$ is large (in absolute value):

$$\left|\hat{v}^T(\Theta-\overline{\Theta})\hat{w}\right|=\left|(z^*)^T\hat{w}\right|=\sum_{l\in\hat{S}}z_l^*\hat{z}_l\geq\frac{2}{5}\sum_{l\in\hat{S}}(z_l^*)^2+(\hat{z}_l)^2\geq\frac{2}{5}[\|w^*\|_2^2+\|\hat{w}\|_2^2]\geq\frac{4}{5}\|\hat{w}\|_2\|w^*\|_2 \ ,$$

where we used in the first inequality that $z_l^*/\hat{z}_l\in[1/2,2]$ and in the second inequality that $S^*\subset\hat{S}$. Thus, it holds that

$$\left|\hat{v}^T(\Theta-\overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2\geq\frac{16}{25}\|w^*\|_2^2 \ . \tag{3.80}$$

It remains to prove that $\|w^*\|_2$ is large enough. Writing S^{*c} for the complementary of S^* in Q, it holds that

$$\|w^*\|_2^2=\|z^*\|_2^2-\sum_{l\in S^{*c}}(z_l^*)^2 \ , \tag{3.81}$$

so that we need to upper bound the latter quantity. Write $z_{S^{*c}}^*=z^*-w^*$. Coming back to the definition of z^*,

$$\begin{aligned}
\left[\sum_{l\in S^{*c}}(z_l^*)^2\right]^2 &= \left[\sum_{l\in S^{*c}}[\hat{v}^T(\Theta-\overline{\Theta})]_l z_l^*\right]^2 \\
&\leq \|\left(\Theta-\overline{\Theta}\right)z_{S^{*c}}^*\|_2^2=\sum_{i\in\widetilde{P}}\left(\sum_{l\in S^{*c}}(\Theta_{i,l}-\overline{\theta}_l)z_l^*\right)^2 \\
&\leq \frac{18\zeta^2}{|\widetilde{P}|^2}\log\left(\frac{2|Q|}{\delta}\right)\sum_{i\in\widetilde{P}}\left(\sum_{l\in S^{*c}}\sum_{j\in\widetilde{P}}|\Theta_{i,l}-\Theta_{j,l}|\right)^2 \\
&\leq \frac{18\zeta^2}{|\widetilde{P}|^2}\log\left(\frac{2|Q|}{\delta}\right)\sum_{i\in\widetilde{P}}\left(\sum_{j\in\widetilde{P}}\|\Theta_{i,\cdot}-\Theta_{j,\cdot}\|_1\right)^2 \\
&\leq 18\zeta^2\phi_{l_1}^2\log\left(\frac{2|Q|}{\delta}\right)\log\left(\frac{2|\widetilde{P}|}{\delta}\right)|\widetilde{P}||Q| \\
&\leq \left[145\zeta^2\log\left(\frac{2|Q||\widetilde{P}|}{\delta}\right)\sqrt{|\widetilde{P}||Q|}\right]^2 \ ,
\end{aligned}$$

where we used the definition of S^* in the third line as well as the Condition (3.71) in the fifth line. We recall that $\phi_{l_1} = 2(2\zeta\sqrt{2} + \overline{\beta}_{\text{tris}}) \le 29\zeta$ is defined in (3.60). Recall that $z^* = \hat{v}^T(\Theta - \overline{\Theta})$. Combining (3.78), (3.81), and Condition (3.62), we deduce that

$$\|w^*\|_2^2 \ge \frac{1}{64\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ ,$$

which, together with (3.80), yields

$$\left\|(\Theta - \overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \left|\hat{v}^T(\Theta - \overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \ge \frac{1}{100\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ .$$

Write $\hat{w}^{(1)}$ and $\hat{w}^{(2)}$ the positive and negative parts of $\hat{w}$ respectively so that $\hat{w} = \hat{w}^{(1)} - \hat{w}^{(2)}$ and $\hat{w}^+ = \hat{w}^{(1)} + \hat{w}^{(2)}$. We obviously have $\|\hat{w}\|_2 = \|\hat{w}^+\|_2$. Besides, if the rows of Θ are ordered according to the oracle permutation, then $(\Theta - \overline{\Theta})\hat{w}^{(1)}$ and $(\Theta - \overline{\Theta})\hat{w}^{(2)}$ are increasing vectors with mean zero. It then follows from Harris' inequality that these two vectors have a nonegative inner product. We have proved that

$$\left\|(\Theta - \overline{\Theta})\frac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \ge \left\|(\Theta - \overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \frac{1}{100\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ . \tag{3.82}$$

Equipped with this bound, we are now in position to show that the set $\overline{P}'$ of experts obtained from $\widetilde{P}$ when applying the pivoting algorithm with $\hat{w}^+/\|\hat{w}^+\|_2$ has a much smaller variance.

By Lemma 3.5.7, there exists an event of probability higher than $1 - \delta$ such that

$$\max_{i,j\in\overline{P}'}\left|\langle\Theta(\overline{P}')_{i,\cdot} - \Theta(\overline{P}')_{j,\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2}\rangle\right| \le \phi_{l_1}\sqrt{\log\left(\tfrac{2|\overline{P}|}{\delta}\right)} \ ,$$

where we recall that $\phi_{l_1} = 2(2\zeta\sqrt{2} + \overline{\beta}_{\text{tris}})$. By convexity, it follows that

$$\left\|(\Theta(\overline{P}') - \overline{\Theta}(\overline{P}'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \le \phi_{l_1}^2\log\left(\tfrac{2|\overline{P}|}{\delta}\right)|\overline{P}'| \le \phi_{l_1}^2\log\left(\tfrac{2|\overline{P}|}{\delta}\right)|\widetilde{P}| \ .$$

In light of Condition (3.62), this quantity is small compared to $\|\Theta - \overline{\Theta}\|_F^2$:

$$\|(\Theta(\overline{P}') - \overline{\Theta}(\overline{P}'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \le \frac{1}{200\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ , \tag{3.83}$$

which together with (3.82) leads to

$$\|(\Theta - \overline{\Theta})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(\Theta(\overline{P}') - \overline{\Theta}(\overline{P}'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \ge \frac{1}{200\log^2(nd/\zeta_-)}\|\Theta - \overline{\Theta}\|_F^2 \ . \tag{3.84}$$

Since $\overline{P}' \subset \widetilde{P}$, we deduce that, for any vector $w' \in \mathbb{R}^q$, we have $\|(\Theta - \overline{\Theta})w'\|_2^2 \ge \|(\Theta(\overline{P}') - \overline{\Theta}(\overline{P}'))w'\|^2$. It then follows from the Pythagorean theorem that

$$\|\Theta - \overline{\Theta}\|_F^2 - \|\Theta(\overline{P}') - \overline{\Theta}(\overline{P}')\|_F^2 \ge \|(\Theta - \overline{\Theta})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(\Theta(\overline{P}') - \overline{\Theta}(\overline{P}'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \ .$$

Then, together with (3.84), we arrive at

$$\|\Theta(\overline{P}') - \overline{\Theta}(\overline{P}')\|_F^2 \le \left(1 - \frac{1}{200\log^2(nd/\zeta)}\right)\|\Theta - \overline{\Theta}\|_F^2 \ .$$

Proof of Lemma 3.5.20. The proof mainly relies on a discretisation argument. Given any $a \in \mathbb{R}$ and any matrix U, we define the matrix $[U]_a^{\text{thres}}$ by $([U]_a^{\text{thres}})_{i,j} = U_{i,j}\mathbf{1}\{U_{i,j} \in (a, 2a]\}$. If a is negative, then the interval should be understood as $[2a, a)$. Recall that all the entries of $\Theta - \overline{\Theta}$ lie in $[-\sqrt{r}, \sqrt{r}]$. This allows us to decompose this matrix as follows

$$r^{-1/2}(\Theta - \overline{\Theta}) = \sum_{i\in\mathbb{N}^*}[r^{-1/2}(\Theta - \overline{\Theta})]_{2^{-i}}^{\text{thres}} + [r^{-1/2}(\Theta - \overline{\Theta})]_{-2^{-i}}^{\text{thres}} \ .$$

All the matrices in this decomposition have disjoint support. For all $i > \log_2(nd/\zeta)$, all the entries of the discretised matrices in the decomposition are smaller than $\zeta(nd)^{-1}$. Since $|\widetilde{P}| = p \le n$ and $|Q| = q \le d/r$, this implies that

$$\left\|\sum_{i\in\mathbb{N}^*, i>\log_2(nd/\zeta)}[r^{-1/2}(\Theta - \overline{\Theta})]_{2^{-i}}^{\text{thres}} + [r^{-1/2}(\Theta - \overline{\Theta})]_{-2^{-i}}^{\text{thres}}\right\|_F^2 \le \zeta^2\frac{nd}{(nd)^2 r} \le \frac{\zeta^2}{r} \ .$$

Coming back to the previous bound, we arrive at

$$\|r^{-1/2}(\Theta-\overline{\Theta})\|_F^2 \le \sum_{i\in\mathbb{N}^*, i\le \log_2(nd/\zeta)} \|[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i}}^{\text{thres}}\|_F^2 + \|[r^{-1/2}(\Theta-\overline{\Theta})]_{-2^{-i}}^{\text{thres}}\|_F^2 + \frac{\zeta^2}{r} \ .$$

As we assume that $\|\Theta-\overline{\Theta}\|_F \ge 2\zeta \ge 2\zeta/\sqrt{r}$,

$$\|r^{-1/2}(\Theta-\overline{\Theta})\|_F^2 \le \frac{4}{3} \sum_{i\in\mathbb{N}^*, i\le \log_2(nd/\zeta)} \|[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i}}^{\text{thres}}\|_F^2 + \|[r^{-1/2}(\Theta-\overline{\Theta})]_{-2^{-i}}^{\text{thres}}\|_F^2 \ .$$

Hence, there exists an integer $i_0 \in [1, \log_2(nd/\zeta)]$ such that

$$\frac{3\|\Theta-\overline{\Theta}\|_F^2}{8r\log_2(nd/\zeta)} \le \|[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}^{\text{thres}}\|_F^2 \vee \|[r^{-1/2}(\Theta-\overline{\Theta})]_{-2^{-i_0}}^{\text{thres}}\|_F^2 \ .$$

Assume w.l.o.g. that, for this $i_0 \le \log_2(nd/\zeta)$, we have

$$\frac{3\|\Theta-\overline{\Theta}\|_F^2}{8r\log_2(nd/\zeta)} \le \|[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}^{\text{thres}}\|_F^2 \ .$$

Now, we define a different discretised version. For a matrix U and some $a \in \mathbb{R}^+$, let $[U]_a$ be defined by $([U]_a)_{ij} = (a\mathbf{1}\{U_{i,j} \ge a\})_{i,j}$. We readily deduce that

$$\|\Theta-\overline{\Theta}\|_F^2 \le \frac{32}{3} r\log_2(nd/\zeta)\|[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}\|_F^2. \tag{3.85}$$

The entries of the matrix $[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}$ lie in $\{0, 2^{-i_0}\}$. Up to a permutation of the rows of Θ, we can assume that each column of $[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}$ is isotonic. One can easily check that a matrix that only takes two values and such that each column is isotonic can be transformed into a bi-isotonic matrix by applying a suitable permutation π_0 to its columns. We denote B the corresponding permuted matrix. Recall that we denote p and q the dimensions of B. Then, define the function $\phi : [p] \to \{0, \dots, q\}$ such that $\phi(i)$ is the number of non-zero entries in the $(p-i+1)$-th row of B. Since B is bi-isotonic, the function ϕ is non-increasing. Besides, we have

$$\sum_{i=1}^{p} \phi(i) = 2^{i_0} \sum_{i,j} B_{i,j} = 2^{2i_0} \sum_{i,j} B_{i,j}^2 = 2^{2i_0}\|B\|_F^2 \ .$$

Lemma 3.5.22. *Let d_1 and d_2 be two positive integers and consider a non-increasing function $f : [d_1] \to \mathbb{R}_+$. Then, there exists $m \in [d_1]$ such that $\sum_{i=1}^{d_1} f(i) \le \log(ed_1) m f(m)$.*

Applying this lemma to ϕ, we deduce that, for some $m \in [p]$, we have

$$2^{2i_0}\|B\|_F^2 \le \log(ep) m\phi(m) \ . \tag{3.86}$$

Since $\phi(m)$ is the number of non-zero entries on the $p+1-m$-th row of B, since B is bi-isotonic and since B only takes two values, this implies that B contains in the lower right a rectangle of size $m \times \phi(m)$ with value 2^{-i_0}. Define the vector $u \in \mathbb{R}^p$ such that $u_i = m^{-1/2}$ if $i \ge p-m+1$ and $u_i = 0$, otherwise. Define also the vector $v \in \mathbb{R}^q$ such $v_j = 1/\sqrt{\phi(m)}$ if $j \ge q - \phi(m) + 1$, and $v_j = 0$ otherwise. It follows from these definitions that $u^T B v = 2^{-i_0}\sqrt{m\phi(m)}$. Recall that $[r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}}$ corresponds to a row and column permutation of B. Hence, there exist two permutations π_1 and π_2 such that

$$u_{\pi_1}^T [r^{-1/2}(\Theta-\overline{\Theta})]_{2^{-i_0}} v_{\pi_2} = u^T B v \ .$$

By construction, the entries of $\Theta - \overline{\Theta}$ are higher than 2^{-i_0} for all entries such that $(u_{\pi_1})_i \ne 0$ and $(v_{\pi_1})_j \ne 0$. We deduce that

$$\left\|\Theta-\overline{\Theta}\right\|_{\text{op}} \ge \sqrt{r} u_{\pi_1}^T r^{-1/2}(\Theta-\overline{\Theta}) v = \sqrt{r} 2^{-i_0}\sqrt{m\phi(m)} \ge \frac{\sqrt{r}}{\sqrt{\log(ep)}}\|B\|_F \ .$$

Finally, we come back to (3.85) to conclude that $\|\Theta-\overline{\Theta}\|_{\text{op}} \ge [32\log(ep)\log_2(nd/\zeta)/3]^{-1/2}\|\Theta-\overline{\Theta}\|_F$, where we recall that $\zeta_- < \zeta$. □

Proof of Lemma 3.5.22. Define $a = \sup_{m=1}^{d_1} m f(m)$. As a consequence, we have $f(m) \le a/m$. This implies that

$$\sum_{i=1}^{d_1} f(i) \le \sum_{i=1}^{d_1} \frac{a}{i} \le a\log(ed_1) \ .$$

We have proved that $\log(ed_1)\sup_{m=1}^{d_1} m f(m) \ge \sum_{i=1}^{d_1} f(i)$. □

Proof of Lemma 3.5.21. We start with the two following lemmas. For short, we write $p = |\overline{P}|$ and $q = |Q|$ in this proof.

Lemma 3.5.23. *Let N' denote a random $d_1 \times d_2$ matrix whose entries follow independent, centered and ζ-subGaussian distributions. Let $\Omega \subset \mathbb{R}^{d_2}$ be a subspace of dimension d_2'. With probability larger than $1-\delta$, one has*

$$\sup_{u\in\mathbb{R}^{d_1},\ v\in\Omega:\ \|u\|_2\le 1,\ \|v\|_2\le 1} |u^T(N'-\overline{N'})v| \le 10\zeta\sqrt{d_1+d_2'+\log(2/\delta)}\ ,$$

where $\overline{N'} = d_1^{-1}\mathbf{1}_{d_1}\mathbf{1}_{d_1}^T N'$ is made of the mean row of N'.

Lemma 3.5.24. *Let N' be a random $d_1\times d_2$ matrix whose entries follow independent, centered and ζ-subGaussian distributions. It holds with probability larger than $1-\delta$ that*

$$\sup_{u\in\mathbb{R}^{d_1}:\|u\|_2\le 1} |\|u^T(N'-\overline{N'})\|_2^2 - \mathbb{E}\|u^T(N'-\overline{N'})\|_2^2| \le 64\zeta^2\left[\sqrt{d_2(5d_1+\log(2/\delta))} + (5d_1+\log(2/\delta))\right]\ .$$

We have

$$Z^{(1)} - \overline{Z}^{(1)} = \Theta - \overline{\Theta} + N^{(1)} - \overline{N}^{(1)}\ ,$$

so that, for any $v\in\mathbb{R}^p$,

$$\|v^T(Z^{(1)}-\overline{Z}^{(1)})\|_2^2 = \|v^T(\Theta-\overline{\Theta})\|_2^2 + \|v^TN^{(1)} - v^T\overline{N}^{(1)}\|_2^2 + 2\langle v^TN^{(1)} - v^T\overline{N}^{(1)}, v^T(\Theta-\overline{\Theta})\rangle\ ,$$

which, in turn, implies that

$$\begin{aligned}\Big|\|v^T(Z^{(1)}-\overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta-\overline{\Theta})\|_2^2 - \mathbb{E}\left[\|v^TN^{(1)} - v^T\overline{N}^{(1)}\|_2^2\right]\Big| \le \\ \Big|\|v^TN^{(1)} - v^T\overline{N}^{(1)}\|_2^2 - \mathbb{E}\left[\|v^TN^{(1)} - v^T\overline{N}^{(1)}\|_2^2\right]\Big| + 2|\langle v^TN^{(1)} - v^T\overline{N}^{(1)}, v^T(\Theta-\overline{\Theta})\rangle|\ .\end{aligned} \tag{3.87}$$

Write $W\subset\mathbb{R}^q$ for the image of $(\Theta-\overline{\Theta})^T$. Then, we apply Lemma 3.5.23 to derive that

$$\begin{aligned}\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1} |\langle v^T(N^{(1)}-\overline{N}^{(1)}), v^T(\Theta-\overline{\Theta})\rangle| &\le \|\Theta-\overline{\Theta}\|_{\mathrm{op}} \sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1,\ u\in W:\ \|u\|_2\le 1} |v^T(N^{(1)}-\overline{N}^{(1)})u| \\ &\le 10\zeta\|\Theta-\overline{\Theta}\|_{\mathrm{op}}\sqrt{2p+\log(6/\delta)}\ ,\end{aligned} \tag{3.88}$$

with probability higher than $1-\delta/3$ since the dimension of W is no larger than p. We deduce from Lemma 3.5.24 that, with probability higher than $1-\delta/3$, we have

$$\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\|v^T(N^{(1)}-\overline{N}^{(1)})\|_2^2 - \mathbb{E}\|v^T(N^{(1)}-\overline{N}^{(1)})\|_2^2\right| \le 64\zeta^2\left[\sqrt{q(5p+\log(6/\delta))} + 5p + \log(6/\delta)\right]\ .$$

Together with (3.87) and (3.88), we have that with probability larger than $1-2\delta/3$,

$$\begin{aligned}&\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\|v^T(Z^{(1)}-\overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta-\overline{\Theta})\|_2^2 - \mathbb{E}\|v^TN^{(1)} - v^T\overline{N}^{(1)}\|_2^2\right| \\ &\qquad\le 10\zeta\|\Theta-\overline{\Theta}\|_{\mathrm{op}}\sqrt{2p+\log(6/\delta)} + 64\zeta^2\left[\sqrt{q(5p+\log(6/\delta))} + (5p+\log(6/\delta))\right]\ .\end{aligned}$$

In the same way, we have that, with probability larger than $1-\delta/3$,

$$\begin{aligned}&\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\frac{1}{2}\|v^T(Z^{(1)}-\overline{Z}^{(1)} - Z^{(2)}+\overline{Z}^{(2)})\|_2^2 - \mathbb{E}\|v^T(N^{(1)}-\overline{N}^{(1)})\|_2^2\right| \\ &\quad= \frac{1}{2}\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\|v^T(Z^{(1)}-\overline{Z}^{(1)} - Z^{(2)}+\overline{Z^{(2)}})\|_2^2 - \mathbb{E}\|v^T(Z^{(1)}-\overline{Z}^{(1)} - Z^{(2)}+\overline{Z}^{(2)})\|_2^2\right| \\ &\quad\le 128\zeta^2\left[\sqrt{q(5p+\log(6/\delta)} + (5p+\log(6/\delta))\right]\ .\end{aligned}$$

Putting everything together we conclude that, on an event of probability higher than $1-3\delta$, we have simultaneously for all $v\in\mathbb{R}^p$ with $\|v\|_2\le 1$ that

$$\begin{aligned}&\left|\|v^T(Z^{(1)}-\overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta-\overline{\Theta})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)}-\overline{Z}^{(1)} - Z^{(2)}+\overline{Z}^{(2)})\|_2^2\right| \\ &\le 10\zeta\|\Theta-\overline{\Theta}\|_{\mathrm{op}}\sqrt{2p+\log(6/\delta)} + 192\zeta^2\left[\sqrt{q(3p+\log(6/\delta))} + (3p+\log(6/\delta))\right]\ .\end{aligned}$$

Since $\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \geq 1600\zeta^2[\sqrt{q(5p + \log(6/\delta))} + 7p + 2\log(6/\delta)]$, we deduce that, on the same event, we have

$$\sup_{v\in\mathbb{R}^p:\ \|v\|_2\leq 1} \left| \|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta - \overline{\Theta})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2 \right| \leq \frac{1}{4}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \ .$$

Writing $\psi(v) = \left| \|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2 \right|$, we deduce that, for v such that $\|v^T(\Theta - \overline{\Theta})\|_2^2 = \|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2$, we have $\Psi(v) \geq \frac{3}{4}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2$, whereas, for v such that $\|v^T(\Theta - \overline{\Theta})\|_2^2 < \frac{1}{2}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2$, we have $\Psi(v) < \frac{3}{4}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2$. We conclude that $\hat{v}$ satisfies $\|\hat{v}^T(\Theta - \overline{\Theta})\|_2^2 > \frac{1}{2}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2$.

□

Proof of Lemma 3.5.23. We start with a classical result. Variants of it can be found in random matrix textbooks (see e.g [91]). Still, we provide a simple dedicated proof below for the sake of completeness.

Lemma 3.5.25. *Let N' be a $d_1 \times d_2$ matrix whose entries follow independent, centered, and ζ-subGaussian distributions. Consider any vector subspace $\Omega \subset \mathbb{R}^{d_2}$ with dimension d_2'. With probability higher than $1 - \delta$, one has*

$$\sup_{u\in\mathbb{R}^{d_1},\ v\in\Omega:\ \|u\|_2\leq 1,\ \|v\|_2\leq 1} |u^T N' v| \leq 5\zeta\sqrt{d_1 + d_2' + \log(1/\delta)} \ .$$

We have the following decomposition

$$\sup_{\substack{u\in\mathbb{R}^{d_1},v\in\Omega \\ \|u\|_2\leq 1,\|v\|_2\leq 1}} |u^T(N' - \overline{N'})v| \quad \leq \quad \sup_{\substack{u\in\mathbb{R}^{d_1},v\in\Omega \\ \|u\|_2\leq 1,\|v\|_2\leq 1}} |u^T N' v| + \sup_{\substack{u\in\mathbb{R}^{d_1},v\in\Omega \\ \|u\|_2\leq 1,\|v\|_2\leq 1}} |u^T \overline{N'} v|.$$

The first expression in the right-hand side is handled with Lemma 3.5.25. Regarding the second one, we observe that $\overline{N'}v$ is a constant vector. As a consequence,

$$\sup_{\substack{u\in\mathbb{R}^{d_1},v\in\Omega \\ \|u\|_2\leq 1,\|v\|_2\leq 1}} |u^T \overline{N'} v| \leq \sqrt{d_1} \sup_{v\in\Omega} \|\overline{n}' v\|_2 \ ,$$

where $\overline{n}'$ is a $\zeta/\sqrt{d_1}$-subGaussian random vector. Then, we control this expression applying Lemma 3.5.25 to a $1 \times d_2'$ matrix. All in all, we have proved that, with probability higher than $1 - \delta$, we have

$$\sup_{\substack{u\in\mathbb{R}^{d_1},v\in\Omega \\ \|u\|_2\leq 1,\|v\|_2\leq 1}} |u^T(N' - \overline{N'})v| \leq 10\zeta\sqrt{(d_1 + d_2') + \log\left(\frac{2}{\delta}\right)} \ .$$

□

Proof of Lemma 3.5.25. Let $\mathcal{U}_d(\epsilon)$ denote the ϵ-covering number of the d-dimensional unit ball and let $\mathcal{U}_d(\epsilon)$ denote a corresponding minimal covering set. For Ω a d_2'-dimensional subspace of $\mathbb{R}^{d_2'}$, we also write with a slight abuse of notation $\mathcal{U}_{d_2'}(\epsilon)$ for a corresponding minimal covering set of its unit ball. Consider any $d_1 \times d_2$ matrix W. Write $w^* = \sup_{u:\|u\|_2\leq 1} \sup_{v\in\Omega,\ \|v\|_2\leq 1} |u^T W v|$ and $w = \sup_{u\in\mathcal{U}_{d_1}(1/4)} \sup_{v\in\mathcal{U}_{d_2'}(1/4)} |u^T W v|$. Given $u \in \mathbb{R}^{d_1}$, let $\pi(u)$ denote any closest point of u in $\mathcal{U}_{d_1}(1/4)$. Similarly, for $v \in \Omega$, $\pi'(v)$ stands for a closest point of v in $\mathcal{U}_{d_2'}(1/4)$. By triangular inequality, we have

$$\begin{aligned} w^* &\leq w + \sup_{u:\|u\|_2\leq 1} \sup_{v\in\Omega,\ \|v\|_2\leq 1} |u^T W v| - |\pi(u) W \pi'(v)| \\ &\leq w + \sup_{u:\|u\|_2\leq 1} \sup_{v\in\Omega,\ \|v\|_2\leq 1} |(u^T - \pi(u)^T) W v| + |\pi(u)^T W (v - \pi'(v))| \\ &\leq w + w^*/2 \ . \end{aligned}$$

We have proven that

$$\sup_{u:\|u\|_2\leq 1} \sup_{v\in\Omega,\ \|v\|_2\leq 1} |u^T W v| \leq 2 \sup_{u\in\mathcal{U}_{d_1}(1/4)} \sup_{v\in\mathcal{U}_{d_2'}(1/4)} |u^T W v| \ . \tag{3.89}$$

Since $\log(\mathcal{U}_d(\epsilon)) \leq d\log(3/\epsilon)$ (see e.g. [104]), we deduce from triangular inequality that, with probability higher than $1 - \delta$, we have

$$\sup_{u\in\mathbb{R}^{d_1},\ v\in\Omega:\ \|u\|_2\leq 1,\ \|v\|_2\leq 1} |u^T N' v| \leq 2\zeta\sqrt{2(d_1 + d_2')\log(12) + 2\log(1/\delta)} \ .$$

□

Proof of Lemma 3.5.24. Relying on (3.89) with $W = (N'-\overline{N}')(N'-\overline{N}')^T - \mathbb{E}\left[(N'-\overline{N}')(N'-\overline{N}')^T\right]$, we derive that $\sup_{u:\|u\|_2\le 1} |\|u^T(N'-\overline{N}')\|_2^2 - \mathbb{E}\|u^T(N'-\overline{N}')\|_2^2|$ is less than or equal to

$$2 \sup_{u\in\mathcal{U}_{d_1}(1/4)} \sup_{v\in\mathcal{U}_{d_1}(1/4)} u^T(N'-\overline{N}')(N'-\overline{N}')^T v - \mathbb{E}\left[u^T(N'-\overline{N}')(N'-\overline{N}')^T v\right]$$

As a consequence, it amounts to simultaneously control $|\mathcal{U}_{d_1}(1/4)|^2$ quadratic forms of subGaussian random variables. For this purpose, we use the Hanson-Wright inequality [91]. Below we provide a version of this inequality with explicit numerical constants.

Lemma 3.5.26. *Let x be d-dimensional ζ-subGaussian centered random vector with independent components. For any $d\times d$ matrix A and any $t>0$, we have*

$$\mathbb{P}\left[x^TAx - \mathbb{E}[x^TAx] \ge 32\zeta^2\left(\|A\|_F\sqrt{t} + \|A\|_{\mathrm{op}}t\right)\right] \le 2e^{-t}$$

For any fixed u and v, we interpret $u^T(N'-\overline{N}')(N'-\overline{N}')^Tv$ as a quadratic form of d_1d_2 independent random variables where the corresponding matrix B of the quadratic form satisfies $\|B\|_{\mathrm{op}}\le 1$ and $\|B\|_F \le \sqrt{d_2}$. Putting everything together we deduce that, with probability higher than $1-\delta$, we have

$$\begin{aligned}
&\sup_{u:\|u\|_2\le 1} \left|\|u^T(N'-\overline{N}')\|_2^2 - \mathbb{E}\|u^T(N'-\overline{N}'))\|_2^2\right| \\
&\quad\le 64\zeta^2\left[\sqrt{d_2(2d_1\log(12)+\log(2/\delta)} + 2d_1\log(12) + \log(2/\delta)\right] \\
&\quad\le 64\zeta^2\left[\sqrt{d_2(5d_1+\log(2/\delta)} + (5d_1+\log(2/\delta))\right] \ .
\end{aligned}$$

□

Proof of Lemma 3.5.26. We consider separately the diagonal terms of A and the non-diagonal terms. Write A^- for the matrix such that $A^-_{ij} = A_{ij}\mathbf{1}_{i\neq j}$. First, we use Section 2.8 in [75] to handle x^TA^-x. We know that

$$\mathbb{P}\left[x^TA^-x \ge 8\zeta^2\left(\|A^-\|_F\sqrt{t} + \sqrt{2}\|A^-\|_{\mathrm{op}}t\right)\right] \le e^{-t} \ ,$$

for any $t>0$. Regarding the diagonal part, we know from Rudelson and Vershynin [78] (Step 1 of the main proof) that $\|x_i^2 - \mathbb{E}[x_i^2]\|_{\psi_1} \le 4\zeta^2$ (see [91] for a definition of $\|.\|_{\psi_1}$). Then, we are in position to apply Bernstein's inequality [10] (Theorem 2.10) to $\sum_i a_{ii}x_i^2$ with $v = (16\zeta^2)^2\sum_i a_{ii}^2$ and $c = 16\zeta^2\max_i|a_{ii}|$. For any $t>0$, we have

$$\mathbb{P}\left[\sum_{i=1}^d a_{ii}(x_i^2 - \mathbb{E}[x_i^2]) \ge 16\zeta^2\left(\sqrt{2\sum_i a_{ii}^2 t} + \max_i|a_{ii}|t\right)\right] \le e^{-t} \ ,$$

which implies that

$$\mathbb{P}\left[\sum_{i=1}^d a_{ii}(x_i^2 - \mathbb{E}[x_i^2]) \ge 16\zeta^2\left(\|A\|_F\sqrt{2t} + \|A\|_{\mathrm{op}}t\right)\right] \le e^{-t} \ .$$

We combine the two deviation inequalities and use $\|A^-\|_{\mathrm{op}} \le 2\|A\|_{\mathrm{op}}$ to conclude that

$$\mathbb{P}\left[x^TAx - \mathbb{E}[x^TAx] \ge 32\zeta^2\left[\|A\|_F\sqrt{t} + \|A\|_{\mathrm{op}}t\right]\right] \le 2e^{-t} \ .$$

□

3.5.4.4 Proof of Corollary 3.5.13

Let $(\overline{L}_{\mathrm{cp}}, \overline{U}_{\mathrm{cp}})$ denote the conservative result of $\mathbf{Pivot}(Z^{(1)}, \mathbf{1}_Q, \gamma)$ and $\widetilde{P} = \overline{P}\setminus(\overline{L}_{\mathrm{cp}}\cup\overline{U}_{\mathrm{cp}})$.
Let $(\overline{L}_{\mathrm{pca}}, \overline{U}_{\mathrm{pca}}) = \mathbf{DoubleTrisection-PCA}(\mathcal{Z}, \gamma)$ with $\mathcal{Z} = (Z^{(2)}, Z^{(3)}, Z^{(4)}, Z^{(5)})$. Here, $(Z^{(2)}, Z^{(3)}, Z^{(4)})$ restricted to the experts in $\widetilde{P}$, whereas $Z^{(5)}$ is restricted to experts in P. Finally, we write $\overline{P}' = \widetilde{P}\setminus(\overline{L}_{\mathrm{pca}}\cup\overline{U}_{\mathrm{pca}})$. We first prove the following intermediary result

$$\|\Theta(\overline{P}', Q) - \overline{\Theta}(\overline{P}', Q)\|_F^2 \le \left(1 - \frac{1}{1920\log^2(\frac{nd}{\delta\zeta_-})}\right)\|\Theta(\overline{P}, Q) - \overline{\Theta}(\overline{P}, Q)\|_F^2 \ . \tag{3.90}$$

We consider two cases. First, we assume that $\frac{\eta}{\zeta}|\overline{P}|\sqrt{|Q|} \le \sqrt{|\overline{P}||Q|} + |\overline{P}|$. Then, it follows from Equation (3.63) that we are in position to apply Proposition 3.5.11 with $\phi = 120\log^2(\frac{nd}{\delta\zeta_-})$. Since $\overline{P}' \subset \widetilde{P}$, it follows that $\|\Theta(\overline{P}', Q) - \overline{\Theta}(\overline{P}', Q)\|_F^2 \le \|\Theta(\widetilde{P}, Q) - \overline{\Theta}(\widetilde{P}, Q)\|_F^2$ and (3.90) follows from Proposition 3.5.11.

Now, we assume that $\sqrt{|\overline{P}||Q|} + |\overline{P}| \le \frac{\eta}{\zeta}|\overline{P}|\sqrt{|Q|}$. If $\|\Theta(\widetilde{P},Q) - \overline{\Theta}(\widetilde{P},Q)\|_F^2 \le 0.5\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2$, then the result obviously holds. Otherwise, it follows from (3.63) that

$$\|\Theta(\widetilde{P},Q) - \overline{\Theta}(\widetilde{P},Q)\|_F^2 \ge 2\cdot 10^5 \log^3\left(\frac{6nd}{\delta\zeta_-}\right)\left[\eta|\overline{P}|\sqrt{|Q|} \wedge \left(\sqrt{|\overline{P}||Q|} + |\overline{P}|\right)\right] ,$$

Besides, with probability higher than $1-\delta$, $\Theta(\widetilde{P})$ is undistinguishable in l_1-norm by (3.71). Hence, we are in position to apply Proposition 3.5.12 and it follows that

$$\begin{aligned}\|\Theta(\overline{P}',Q) - \overline{\Theta}(\overline{P}',Q)\|_F^2 &\le \left(1 - \frac{1}{200\log^2(\frac{nd}{\delta\zeta_-})}\right)\|\Theta(\widetilde{P},Q) - \overline{\Theta}(\widetilde{P},Q)\|_F^2 \\ &\le \left(1 - \frac{1}{1920\log^2(\frac{nd}{\delta\zeta_-})}\right)\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 ,\end{aligned}$$

which is exactly Equation (3.90).

It remains to conclude from Equation (3.90). We start from

$$\begin{aligned}\|\Theta(\overline{P}',\mathcal{Q}_r) - \overline{\Theta}(\overline{P}',\mathcal{Q}_r)\|_F^2 &= \|\Theta(\overline{P}',Q) - \overline{\Theta}(\overline{P}',Q)\|_F^2 + \|\Theta(\overline{P}',\mathcal{Q}_r\setminus Q) - \overline{\Theta}(\overline{P}',\mathcal{Q}_r\setminus Q)\|_F^2 \\ &\le \left(1 - \frac{1}{1920\log^2(\frac{nd}{\delta\zeta_-})}\right)\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \\ &\quad + \|\Theta(\overline{P},\mathcal{Q}_r\setminus Q) - \overline{\Theta}(\overline{P},\mathcal{Q}_r\setminus Q)\|_F^2 \\ &\le \left(1 - \frac{1}{3\cdot 10^5\log^4\left(\frac{nd}{\delta\zeta_-}\right)}\right)\|\Theta(\overline{P},\mathcal{Q}_r) - \overline{\Theta}(\overline{P},\mathcal{Q}_r)\|_F^2 ,\end{aligned}$$

where we used in the last line that $\|\Theta(\overline{P},Q) - \overline{\Theta}(\overline{P},Q)\|_F^2 \ge 1/(120\log^2(\frac{nd}{\delta\zeta_-}))\|\Theta(\overline{P},\mathcal{Q}_r) - \overline{\Theta}(\overline{P},\mathcal{Q}_r)\|_F^2$.

3.5.4.5 Proof of Proposition 3.5.14

For all $\tau = 0,\dots,\tau_\infty$, let $(\overline{O}_\tau,\overline{P}_\tau,\overline{I}_\tau)$ be the sets defined in **BlockSort**.

Let also $(h^{\dagger\tau},r^{\dagger\tau}) = \arg\max_{h,r}\|\left[\Theta(\overline{P}_\tau,Q^*_{\rm cp}) - \overline{\Theta}(\overline{P}_\tau,Q^*_{\rm cp})\right]_{\sqrt{r}h}\|_F^2$, where we recall that $Q^*_{\rm cp}$ depends on h and r. For simplicity, we write $Q^{\dagger\tau}_{\rm cp} = Q^*_{\rm cp}(h^{\dagger\tau},r^{\dagger\tau})$. Equipped with this notation, we readily deduce from Lemma 3.5.10 that

$$\|M(\overline{P}_\tau) - \overline{M}(\overline{P}_\tau)\|_F^2 \le 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}|\left\|\left[\Theta(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp}) - \overline{\Theta}(P,Q^{\dagger\tau}_{\rm cp})\right]_{\sqrt{r^{\dagger\tau}}h^{\dagger\tau}}\right\|_F^2 . \tag{3.91}$$

If, for some $\tau < \tau_\infty$, we have

$$\left\|\left[\Theta(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp}) - \overline{\Theta}(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp})\right]_{\sqrt{r^{\dagger\tau}}h^{\dagger\tau}}\right\|_F^2 \le 4\cdot 10^5\zeta^2\log^3\left(\frac{6nd}{\delta\zeta_-}\right)\Psi(|\overline{P}_\tau|,r^{\dagger\tau},h^{\dagger\tau},|\overline{Q}^{\dagger\tau}_{\rm cp}|) , \tag{3.92}$$

then we can fix, for any such τ, $\overline{P}^\dagger = \overline{P}_\tau$, $h^\dagger = h^{\dagger\tau}$, $Q^\dagger = Q^{\dagger\tau}_{\rm cp}$, and $r^\dagger = r^{\dagger\tau}$ so that both the properties (3.65) and (3.66) hold.

Hence, we assume henceforth that, for all τ, Equation (3.92) does not hold and we shall arrive at a contradiction. In particular, this implies that $\|[\Theta(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp}) - \overline{\Theta}(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp})]_{\sqrt{r^{\dagger\tau}}h^{\dagger\tau}}\|_F^2 \ge \zeta^2 \ge \zeta^2(|\mathcal{R}||\mathcal{H}|)^{-1}$. We have $112|\mathcal{R}||\mathcal{H}| \le 120\log^2(\frac{nd}{\delta\zeta_-})$ provided that n is a large enough constant. In light of (3.91), this implies that, for all τ,

$$\begin{aligned}\left\|\left[\Theta(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp}) - \overline{\Theta}(\overline{P}_\tau,Q^{\dagger\tau}_{\rm cp})\right]_{\sqrt{r^{\dagger\tau}}h^{\dagger\tau}}\right\|_F^2 &\ge \frac{1}{120\log^2(\frac{nd}{\delta\zeta_-})}\|M(\overline{P}_\tau) - \overline{M}(\overline{P}_\tau)\|_F^2 &(3.93)\\ &\ge \frac{1}{120\log^2(\frac{nd}{\delta\zeta_-})}\|\Theta(\overline{P}_\tau,\mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_\tau,\mathcal{Q}_{r^{\dagger\tau}})\|_F^2 , &(3.94)\end{aligned}$$

where we recall that $\mathcal{Q}_{r^{\dagger\tau}}$ is the collection of all blocks at scale $r^{\dagger\tau}$ and we use the Pythagorean equality in the second line. Applying Lemma 3.5.8 at the scale $r^{\dagger\tau}\in\mathcal{R}$, at the height $h^{\dagger\tau}\in\mathcal{H}$, and at all steps $\tau = 0,\dots,\tau_\infty - 1$, we deduce that the event $\overline{\xi}_{\rm cp} := \bigcap_{\tau=0}^{\tau_\infty-1}\xi_{\rm cp}(\overline{P}_\tau,h^{\dagger\tau},r^{\dagger\tau})$ holds with probability at least $1-\tau_\infty\delta$. Under this event,

we write $\widehat{Q}^{\tau}_{\mathrm{cp}}$ for the estimated set defined at step τ and scales $(h^{\dagger\tau}, r^{\dagger\tau})$ in **BlockSort**. Then, it holds that $Q^{\dagger\tau}_{\mathrm{cp}} \subset \widehat{Q}^{\tau}_{\mathrm{cp}} \subset \overline{Q}^{\dagger\tau}_{\mathrm{cp}}$ and we deduce from (3.94) that

$$\left\| \left[\Theta(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}}) \right]_{\sqrt{r^{\dagger\tau}} h^{\dagger\tau}} \right\|_F^2 \geq \frac{1}{120 \log^2(\frac{nd}{\delta\tau^-})} \|\Theta(\overline{P}_\tau, \mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_\tau, \mathcal{Q}_{r^{\dagger\tau}})\|_F^2 \ .$$

Since (3.92) is not satisfied, we also have

$$\begin{aligned} \|\Theta(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}})\|_F^2 &\geq \left\| \left[\Theta(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}_\tau, \widehat{Q}^{\tau}_{\mathrm{cp}}) \right]_{\sqrt{r^{\dagger\tau}} h^{\dagger\tau}} \right\|_F^2 \\ &\geq \left\| \left[\Theta(\overline{P}_\tau, Q^{\dagger\tau}_{\mathrm{cp}}) - \overline{\Theta}(\overline{P}_\tau, Q^{\dagger\tau}_{\mathrm{cp}}) \right]_{\sqrt{r^{\dagger\tau}} h^{\dagger\tau}} \right\|_F^2 \\ &\geq 4 \cdot 10^5 \zeta^2 \log^3\left(\frac{6nd}{\delta\zeta_-}\right) \Psi(|\overline{P}_\tau|, r^{\dagger\tau}, h^{\dagger\tau}, |\widehat{Q}^{\tau}_{\mathrm{cp}}|) \ , \end{aligned}$$

since $\widehat{Q}^{\tau}_{\mathrm{cp}} \subset \overline{Q}^{\dagger\tau}_{\mathrm{cp}}$. Hence, we are in position to apply Corollary 3.5.13 at all steps τ with $r^{\dagger\tau}$, $\overline{P}_\tau$, $\widehat{Q}^{\tau}_{\mathrm{cp}}$, and $\eta = \sqrt{r^{\dagger\tau}} h^{\dagger\tau}$. There exists an event of probability higher than $1 - 4\tau_\infty\delta$ such that, at all steps τ, we have

$$\|\Theta(\overline{P}_{\tau+1}, \mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_{\tau+1}, \mathcal{Q}_{r^{\dagger\tau}})\|_F^2 \leq \left(1 - \frac{1}{3 \cdot 10^5 \log^4\left(\frac{nd}{\delta\zeta_-}\right)}\right) \|\Theta(\overline{P}_\tau; \mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_\tau; \mathcal{Q}_{r^{\dagger\tau}})\|_F^2 \ .$$

Together with Equation (3.93), we deduce that

$$\begin{aligned} &\|M(\overline{P}_\tau) - \overline{M}(\overline{P}_\tau)\|_F^2 - \|M(\overline{P}_{\tau+1}) - \overline{M}(\overline{P}_{\tau+1})\|_F^2 \\ &\quad \geq \|\Theta(\overline{P}_\tau, \mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_{\tau+1}, \mathcal{Q}_{r^{\dagger\tau}})\|_F^2 - \|\Theta(\overline{P}_{\tau+1}, \mathcal{Q}_{r^{\dagger\tau}}) - \overline{\Theta}(\overline{P}_\tau, \mathcal{Q}_{r^{\dagger\tau}})\|_F^2 \\ &\quad \geq \frac{1}{4 \cdot 10^7 \log^6(\frac{nd}{\delta\zeta_-})} \|M(\overline{P}_\tau) - \overline{M}(\overline{P}_\tau)\|_F^2 \ . \end{aligned}$$

Hence,

$$\begin{aligned} \|M(\overline{P}_{\tau_0})\|_F^2 &\geq \|M(\overline{P}_{\tau_0}) - \overline{M}(\overline{P}_{\tau_0})\|_F^2 \\ &\geq \|M(\overline{P}_{\tau_\infty}) - \overline{M}(\overline{P}_{\tau_\infty})\|_F^2 \left(1 - \frac{1}{4 \cdot 10^7 \log^6(\frac{nd}{\delta\zeta_-})}\right)^{-\tau_\infty} \ . \end{aligned}$$

Since (3.92) does not hold at $\tau = \tau_\infty$, this implies that the Frobenius norm in the right-hand side of the above inequality is larger than $2\zeta^2$ and, in light of the definition of $\tau_\infty = 4 \cdot 10^7 \log^7(\frac{nd}{\delta(\zeta_-)^2})$, the right-hand side is larger than $2nd$. This contradicts the fact that $\|M(\overline{P}_{\tau_0})\|_F^2 \leq nd$ since the entries of M lie in $[0,1]$.

Proof of Corollary 3.5.15. To ease the notation in this proof, we simply write P for $\overline{P}^{\dagger}$, r for $r^{\dagger}$, $\overline{Q}^{*}_{\mathrm{cp}}$ for $\overline{Q}^{\dagger}_{\mathrm{cp}}$, and h for $h^{\dagger}$. Since $\overline{Q}^{*}_{\mathrm{cp}}$ corresponds to a set of blocks of questions of size r, it follows that $|\overline{Q}^{*}_{\mathrm{cp}}| \leq d/r$. This, in turn, implies that $\sqrt{r}h|P|\sqrt{|\overline{Q}^{*}_{\mathrm{cp}}|} \leq |P|h\sqrt{d}$ and $\sqrt{|P||\overline{Q}^{*}_{\mathrm{cp}}|} \leq \sqrt{|P|d}$. We have proven that

$$\frac{\sqrt{r}h|P|\sqrt{|\overline{Q}^{*}_{\mathrm{cp}}|}}{\zeta} \wedge \left(\sqrt{|P||\overline{Q}^{*}_{\mathrm{cp}}|} + |P|\right) \leq \frac{|G|h\sqrt{d}}{\zeta} \wedge \left(\sqrt{|G|d} + |G|\right) \ . \tag{3.95}$$

Second, we know from Lemma 3.5.9 that $|\overline{Q}^{*}_{\mathrm{cp}}| \leq 64\frac{\tilde{r}}{rh}$ so that

$$\frac{\sqrt{r}h|P|\sqrt{|\overline{Q}^{*}_{\mathrm{cp}}|}}{\zeta} \wedge \left(\sqrt{|P||\overline{Q}^{*}_{\mathrm{cp}}|} + |P|\right) \lesssim \frac{|P|\sqrt{\tilde{r}h}}{\zeta} \wedge \left(\sqrt{|P|\frac{\tilde{r}}{rh}} + |P|\right) \ . \tag{3.96}$$

If $r > 32\zeta^2 \log(\frac{2d}{\delta})\frac{1}{|P|h^2}$ then, it follows from the definition (3.53) of $\tilde{r}$ that $\tilde{r} = 8r$ so that the right-hand side of (3.96) is at most of the order of $\sqrt{|P|/h} + |P|$. For a smaller r, we know from (3.53) that $\tilde{r} \lesssim \zeta^2 \log(\frac{2d}{\delta})/(|P|h^2)$, which in turn implies that

$$\frac{|P|\sqrt{\tilde{r}h}}{\zeta} \lesssim \sqrt{\log(\tfrac{2d}{\delta})}\sqrt{\frac{|P|}{h}} \ .$$

Hence, we deduce from (3.96) that

$$\frac{\sqrt{r}h|P|\sqrt{|\overline{Q}^*_{\mathrm{cp}}|}}{\zeta} \wedge \left(\sqrt{|P||\overline{Q}^*_{\mathrm{cp}}|} + |P|\right) \lesssim \sqrt{\log(\tfrac{2d}{\delta})}\left(\sqrt{\frac{|P|}{h}} + |P|\right) \ .$$

Together with (3.95), this leads us to

$$\frac{\sqrt{r}h|P|\sqrt{|\overline{Q}^*_{\mathrm{cp}}|}}{\zeta} \wedge \left(\sqrt{|P||\overline{Q}^*_{\mathrm{cp}}|} + |P|\right) \lesssim \sqrt{\log(\tfrac{2d}{\delta})}\left[\frac{|G|h\sqrt{d}}{\zeta} \wedge |G|\sqrt{d} \wedge \sqrt{\frac{|G|}{h}} + |G|\right] \ ,$$

which, together with (3.65) concludes the proof.

□

3.5.5 Proof of Proposition 3.5.6

In this section, we prove Proposition 3.5.6 which states a tighter bound than Proposition 3.5.5 on **TreeSort** when we use the variant **DimensionReduction – WM** to compute $\hat{\pi}_{WM}$. Recall that, for any $t = 1, \dots, t_\infty$, $\mathcal{T}_t$ stands for the hierarchical sorting tree built by **TreeSort** at the beginning of step t. Thus, $\mathcal{T}_t$ has depth t.

The main difference with the analysis of Proposition 3.5.5 lies in the analysis of the algorithm **DimensionReduction – WM**, which is the purpose of the next subsection. Then, we combine it with the general scheme of the proof of Proposition 3.5.6 to get the desired bound.

3.5.5.1 Analysis of DimensionReduction – WM

The key idea of **DimensionReduction – WM** is to examine the high-variation regions of the observations not only in a set of experts $\overline{P}$ but also in the neighboring sets of experts. For this reason, we remind the reader of the notation of **DimensionReduction – WM**. Through this subsection, we fix the step $t \geq 0$ of **TreeSort**. For simplicity, we write $\mathcal{T} := \mathcal{T}_t$. Recall that $\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ stands for the set of leaves of $\mathcal{T}$ of type **0** or **1**. By definition, those leaves are all at depth t. Let us focus on a specific leaf $G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$, and we consider a subset $\overline{P}$ of G.

Finally, we recall that we consider an ordering of the leaves $\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ at depth t and centered on G as:

$$(G^{(a)})_{a\in\mathbb{Z}} = \mathbf{Order}(\mathcal{T}, G) \ ,$$

where $G^{(0)} = G$.

Also, we fix any $h \in \mathcal{H}$ and $r \in \mathcal{R}$. As in **DimensionReduction – WM**, define

$$r_0 = 2^9 \log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{|\overline{P}|h^2} \quad \text{and} \quad \tilde{r} = 4(\lceil r_0 \rceil^{dya} \vee r) \ , \tag{3.97}$$

where $\lceil r_0 \rceil^{dya} = 2^{\lceil \log_2(r_0) \rceil}$ is the smaller power of 2 which is larger than r_0. Up to numerical constants, $\tilde{r}$ is defined as for the original procedure **DimensionReduction**. If $r \geq r_0$, then we can simply rely on CUSUM statistics at the scale $8r$ and on the set $\overline{P}$ to detect high variation regions in $\overline{P}$. If h (or $|\overline{P}|$) is so small that $r_0 > r$, we applied the CUSUM statistic at a larger scale $\tilde{r}$ in **DimensionReduction**. In this version, we compute the CUSUM statistics at a scale smaller than $\tilde{r}$ to the price of considering more experts than those in $\overline{P}$.

If $r < \lceil r_0 \rceil^{dya}$, let us consider any $r_{\mathrm{cp}} \in [8r, 2\tilde{r}] \cap \mathcal{R}$. We respectively define

$$a^+_{WM} := a^+_{WM}(\mathcal{T}, G, h, r_{\mathrm{cp}}) = \min\left\{a \ : \ |G^{(1)}| + \dots + |G^{(a)}| \geq 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\mathrm{cp}}h^2}\right\} \ ;$$
$$a^-_{WM} := a^-_{WM}(\mathcal{T}, G, h, r_{\mathrm{cp}}) = \min\left\{a \ : \ |G^{(-1)}| + \dots + |G^{(-a)}| \geq 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\mathrm{cp}}h^2}\right\} \ ,$$

as the minimum number of groups above and below G in such a way that there are enough experts to detect a h-variation in the mean at the scale r_{cp}. Then, $\mathcal{V}^+_{r_{\mathrm{cp}}}$ and $\mathcal{V}^-_{r_{\mathrm{cp}}}$ stand for the collection of experts in the corresponding groups:

$$\mathcal{V}^+_{r_{\mathrm{cp}}} := \mathcal{V}^+_{r_{\mathrm{cp}}}(\mathcal{T}, \overline{P}, h) = \bigcup_{a=1}^{a^+_{WM}} G^{(a)} \quad \text{and} \quad \mathcal{V}^-_{r_{\mathrm{cp}}} := \mathcal{V}^-_{r_{\mathrm{cp}}}(\mathcal{T}, \overline{P}, h) = \bigcup_{a=-a^-_{WM}}^{-1} G^{(a)} \ , \tag{3.98}$$

Finally, we define

$$\mathcal{V}_{r_{\mathrm{cp}}} := \mathcal{V}_{r_{\mathrm{cp}}}(\mathcal{T}, \overline{P}, h) = \begin{cases} \mathcal{V}^+_{r_{\mathrm{cp}}} \cup \mathcal{V}^-_{r_{\mathrm{cp}}} & \text{if } r_{\mathrm{cp}} \le \tilde{r} \\ \overline{P} & \text{if } r_{\mathrm{cp}} = 2\tilde{r} \end{cases} \tag{3.99}$$

which exactly corresponds to the definition at Line 4 and Line 6 of **DimensionReduction – WM**. For any $k \in [d]$, we recall here the definition of the statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ its deterministic counterpart:

$$\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} = \frac{1}{2r_{\mathrm{cp}}} \sum_{k'=k-r_{\mathrm{cp}}}^{k+r_{\mathrm{cp}}-1} \overline{y}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{y}_{k'}(\mathcal{V}^-_{r_{\mathrm{cp}}}) \quad \text{and} \quad \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}} = \frac{1}{2r_{\mathrm{cp}}} \sum_{k'=k-r_{\mathrm{cp}}}^{k+r_{\mathrm{cp}}-1} \overline{m}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{m}_{k'}(\mathcal{V}^-_{r_{\mathrm{cp}}}) \ .$$

In the notation of $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$, we remove the dependency on $\mathcal{V}^+_{r_{\mathrm{cp}}}$ and $\mathcal{V}^-_{r_{\mathrm{cp}}}$ to simplify the notation. Here, $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ stands for the width between the empirical means of the groups above $\overline{P}$ and below $\overline{P}$. Recall also the definition of the statistic $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ and introduce its deterministic counterpart:

$$\begin{aligned} \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} &= \frac{1}{2r_{\mathrm{cp}}} \left(\sum_{k'=k}^{k+2r_{\mathrm{cp}}-1} \overline{y}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) - \sum_{k'=k-2r_{\mathrm{cp}}}^{k-1} \overline{y}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) \right) \\ \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} &= \frac{1}{2r_{\mathrm{cp}}} \left(\sum_{k'=k}^{k+2r_{\mathrm{cp}}-1} \overline{m}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) - \sum_{k'=k-2r_{\mathrm{cp}}}^{k-1} \overline{m}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) \right) \ . \end{aligned}$$

Here, $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ stands for the mean CUSUM statistic over the experts in $\mathcal{V}_{2r_{\mathrm{cp}}}$. Consider any $r_{\mathrm{cp}} \in [4r, \tilde{r}] \cap \mathcal{R}$. Then, as in the algorithm **DimensionReduction – WM**, we define the collection of positions where both the width and the CUSUM statistic are large:

$$\widehat{D}_{WM} := \widehat{D}_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) = \left\{ k \ : \ \widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} \ge \frac{h}{16} \text{ and } \widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} \ge \frac{h}{16} \right\} \ .$$

See Figure 3.5 for illustrations. Then, we define D^*_{WM} and $\overline{D}^*_{WM}$ as the population counterparts of $\widehat{D}_{WM}$ with different constants

$$\begin{aligned} D^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) &= \left\{ k \in 1, \dots, d \ : \ \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} \ge \frac{h}{8} \text{ and } \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}} \ge \frac{h}{8} \right\} \ ; \\ \overline{D}^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) &= \left\{ k \in 1, \dots, d \ : \ \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} \ge \frac{h}{32} \text{ and } \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}} \ge \frac{h}{32} \right\} \ . \end{aligned}$$

Then, we consider the collections of blocks $\widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}})$, $Q^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}})$, and $\widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}})$ of size r. With our notation, this means that $Q^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) = \mathbf{Encode-Set}(D^*_{WM}, r)$, $\overline{Q}^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) = \mathbf{Encode-Set}(\overline{D}^*_{WM}, r)$, and $\widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) = \mathbf{Encode-Set}(\widehat{D}_{WM}, r)$. Finally, we consider the unions over all possible $r_{\mathrm{cp}} \in \mathcal{R}$ with $4r \le r_{\mathrm{cp}} \le \tilde{r}$:

$$\begin{aligned} \widehat{Q}_{WM} := \widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r) &= \bigcup_{r_{\mathrm{cp}}=4r}^{\tilde{r}} \widehat{Q}_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) \ ; \\ Q^*_{WM} := Q^*_{WM}(\mathcal{T}, \overline{P}, h, r) &= \bigcup_{r_{\mathrm{cp}}=4r}^{\tilde{r}} Q^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) \ ; \\ \overline{Q}^*_{WM} := \overline{Q}^*_{WM}(\mathcal{T}, \overline{P}, h, r) &= \bigcup_{r_{\mathrm{cp}}=4r}^{\tilde{r}} \overline{Q}^*_{WM}(\mathcal{T}, \overline{P}, h, r, r_{\mathrm{cp}}) \ . \end{aligned}$$

The following lemma states that, with high probability, $\widehat{Q}_{WM}$ is sandwiched between Q^*_{WM} and $\overline{Q}^*_{WM}$, so that, on the corresponding event, it is sufficient to study these two quantities.

Lemma 3.5.27. *Consider any valid hierarchical sorting tree $\mathcal{T}$, any subset $\overline{P}$ of a leaf G of $\mathcal{T}$, any $h \in \mathcal{H}$, and any $r \in \mathcal{R}$. With probability at least $1 - \delta$, it holds that*

$$Q^*_{WM} \subset \widehat{Q}_{WM} \subset \overline{Q}^*_{WM} \ . \tag{3.100}$$

Next, we show that the aggregation of $M(\overline{P})$ at Q^*_{WM} captures most of the variance of $M(\overline{P})$.

Lemma 3.5.28. *Assume that $\mathcal{T}$ is a valid hierarchical sorting tree. Then, there exist $h \in \mathcal{H}$ and $r \in \mathcal{R}$ such that*

$$\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \le 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}| \left\| \left[\Theta(\overline{P}, Q^*_{WM}) - \overline{\Theta}(\overline{P}, Q^*_{WM}) \right]_{\sqrt{r}h} \right\|_F^2 \ . \tag{3.101}$$

Recall that $\mathcal{T}_{t_\infty}$ (and in particular also $\mathcal{T} = \mathcal{T}_t$) is a valid hierarchical sorting tree under the event ξ of high probability defined in Corollary 3.5.4. This lemma is the counterpart of Lemma 3.5.10 for the oblivious **DimensionReduction** algorithm.

3.5.5.2 Analysis of the variant BlockSort with DoubleTrisection – WM

Recall the definition (3.64) of the function Ψ by $\Psi(p,r,h,q) = \frac{hp\sqrt{rq}}{\zeta} \wedge \sqrt{pq} + p$. In Proposition 3.5.14, we stated a high probability control for the result of **BlockSort** when fed with **DimensionReduction**. In particular, this proposition only used the properties of **DimensionReduction** stated in Lemmas 3.5.8 and 3.5.10. As we have proven in Lemmas 3.5.27 and 3.5.28 (their counterparts for **DimensionReduction – WM**), we readily obtain the following result whose proof is omitted.

Proposition 3.5.29. *Assume that $\mathcal{T}_t$ is a valid hierarchical sorting tree. Consider a leaf G of $\mathcal{T}$ of type* **0** *or* **1** *at depth t. With probability higher than $1-5\tau_\infty\delta$, there exists a subset $\overline{P}^\dagger$ such that $\overline{P} \subseteq \overline{P}^\dagger \subseteq G$ and the following property holds. For some $r^\dagger_{\mathrm{cp}} \geq r^\dagger \in \mathcal{R}$ and some $h^\dagger \in \mathcal{H}$, upon writing $Q^\dagger_{WM} = Q^*_{WM}$ and $\overline{Q}^\dagger_{WM} = \overline{Q}^*_{WM}$, we have simultaneously*

$$\left\| \left[\Theta(\overline{P}^\dagger, Q^\dagger_{WM}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{WM})\right]_{\sqrt{r^\dagger}h^\dagger} \right\|_F^2 \leq 4\cdot 10^5 \zeta^2 \log^3\left(\frac{6nd}{\delta\zeta_-}\right) \Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{WM}|) \ ; \tag{3.102}$$

$$\|M(\overline{P}^\dagger) - \overline{M}(\overline{P}^\dagger)\|_F^2 \leq 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}| \|[\Theta(\overline{P}^\dagger, Q^\dagger_{WM}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{WM})]_{\sqrt{r^\dagger}h^\dagger}\|_F^2 \ . \tag{3.103}$$

Since $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \leq \|M(\overline{P}^\dagger) - \overline{M}(\overline{P}^\dagger)\|_F^2$, the above proposition controls $\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2$ in terms of $\Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{WM}|)$.

3.5.5.3 Analysis of the complete procedure TreeSort with DoubleTrisection – WM

In light of Proposition 3.5.29, we need to control the cardinality of $|\overline{Q}^\dagger_{WM}|$. In comparison to the oblivious procedure analyzed in the previous section, the main improvement here is that the typical cardinalities $|\overline{Q}^\dagger_{WM}|$ are smaller than $|\overline{Q}^\dagger_{\mathrm{cp}}|$ thanks to the refined dimension reduction procedure **DimensionReduction – WM**.

Unfortunately, it is not possible to get a tight control of the cardinality of each $\overline{Q}^\dagger_{WM}$ individually. Still, we are able to show that among all groups $G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T}_t)$ that are refined in the t-th iteration of **TreeSort**, many of them will correspond to small $|\overline{Q}^\dagger_{WM}|$. To formalize this argument, we need to be careful about the dependencies of the quantities under consideration.

We start from the ordered collection $\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T}_t)$ of $v \leq 2^t$ leaves of types **0** or **1**. We write $G_1, \dots, G_v$ for these groups and we are given a collection $\overline{P}_1, \dots, \overline{P}_v$ of subgroups such that $\overline{P}_i \subset G_i$ for $i = 1, \dots, v$. Later, we will specify $\overline{P}_v = \overline{P}^\dagger_v$, but those sets can be considered arbitrarily.

For a specific group $\overline{P}_v \subset G_v$, we write $\overline{Q}^\dagger_{WM}(\overline{P}_v, h, r)$ instead of $\overline{Q}^\dagger_{WM}$ to emphasize its dependency on $\overline{P}_v$, r and h. Given a positive integer $p > 0$, we define $\mathcal{P}^*(p) = \{\overline{P}_v : \ |\overline{P}_v| \in [p, 2p)\}$ the collection of groups $\overline{P}_v$ of size in $[p, 2p)$.

Lemma 3.5.30. *Assume that $\mathcal{T}_t$ is a valid hierarchical sorting tree. For any $h \in \mathcal{H}$, $r \in R$, any integer p, any sequence $\overline{P}_v$ of subsets of G_v, it holds that*

$$\sum_{\overline{P} \in \mathcal{P}^*(p)} |\overline{Q}^*_{WM}(\overline{P}, h, r)| \lesssim \log(d) \left[\frac{\sqrt{nd(r_0 \vee r)}}{rh\sqrt{p}} \wedge \frac{d(r_0 \vee r)}{r^2 h} \wedge \frac{nd}{pr} \wedge \frac{n(r_0 \vee r)}{prh} \right] . \tag{3.104}$$

We are now equipped to prove Proposition 3.5.6.

Proof of Proposition 3.5.6. We work under the event ξ (Corollary 3.5.4) ensuring $\mathcal{T}_{t_\infty}$ and in particular $\mathcal{T}_t$ is a valid hierarchical sorting tree. For each group $G_s \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T}_t)$ we apply Proposition 3.5.29 and define a corresponding subgroup $\overline{P}^\dagger$, with $r^\dagger \in \mathcal{R}$, $h^\dagger \in \mathcal{H}$ and a corresponding collection of blocks $\overline{Q}^\dagger_{WM}$. Define the collection $\mathcal{D}_n = \{1, 2, 4, \dots, 2^{\lceil \log_2(n) \rceil}\}$. For $p \in \mathcal{D}_n$, we define $\mathcal{P}^*(p,h,r)$ as the collection of groups $\overline{P}^\dagger$ satisfying $|\overline{P}^\dagger| \in [p, 2p)$, $h^\dagger = h$, and $r^\dagger = r$.

Then, we derive from Proposition 3.5.29 that, on an additional event of probability higher than $1 - 5\cdot 2^t \tau_\infty \delta$,

we have

$$
\begin{aligned}
&\sum_{\overline{P}\in\overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \\
&\le 16\zeta^2|\overline{\mathcal{L}}_t| + 96|\mathcal{R}||\mathcal{H}| \sum_{p,h,r} \sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)} \left\| \left[\Theta(\overline{P}^\dagger, \overline{Q}^\dagger_{WM}) - \overline{\Theta}(\overline{P}^\dagger, \overline{Q}^\dagger_{WM})\right]_{\sqrt{r}h} \right\|_F^2 \\
&\overset{(a)}{\lesssim} \zeta^2 \log^5\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h,r} \sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)} \left[\sqrt{[\frac{h^2pr}{\zeta^2}\wedge 1]p|\overline{Q}^*_{WM}(\overline{P}^\dagger,h,r)|} + p\right] \\
&\overset{(b)}{\lesssim} \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h,r} \left[\sqrt{\frac{r}{r\vee r_0}p|\mathcal{P}^*(p,h,r)| \sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)} |\overline{Q}^*_{WM}(\overline{P}^\dagger,h,r)|} + p|\mathcal{P}^*(p,h,r)|\right] \\
&\overset{(c)}{\lesssim} \zeta^2 \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h,r} \left[\sqrt{\frac{nr}{r\vee r_0} \sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)} |\overline{Q}^*_{WM}(\overline{P}^\dagger,h,r)|} + n\right] \\
&\overset{(d)}{\lesssim} \zeta^2 \log^{6}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h,r} \left[\left(n^{3/4}d^{1/4}\left(\frac{1}{(r_0\vee r)ph^2}\right)^{1/4} \wedge n\sqrt{\frac{d}{p(r_0\vee r)}} \wedge \frac{n}{\sqrt{ph}} \wedge \sqrt{\frac{nd}{rh}}\right) + n\right] \\
&\lesssim \zeta^2 \log^{7}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h} \left[\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{nh}{\zeta}\sqrt{d} \wedge \frac{n}{\sqrt{h}} \wedge \frac{\sqrt{nd}}{\sqrt{h}}\right) + n\right] \\
&\overset{(e)}{\lesssim} \zeta^2 \log^{7}\left(\frac{6nd}{\delta\zeta_-}\right) \sum_{p,h} \left[\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}}\right) + n\right] \\
&\lesssim \zeta^2 \log^{9}\left(\frac{6nd}{\delta\zeta_-}\right) \left[\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}}\right) + n\right] ,
\end{aligned}
$$

where we applied Proposition 3.5.29 in (a), Jensen inequality and the definition of r_0 in (b), as well as the bound $|\mathcal{P}^*(p,h,r)| \le n/p$ in (c), Lemma 3.5.30 in (d), and $x\wedge y \le x^{1/3}y^{2/3}$ in (e).

□

3.5.5.4 Remaining proofs

Proof of Lemma 3.5.27. It is sufficient to prove that with high probability, $D^*_{WM}(\mathcal{T},P,h,r_{\mathrm{cp}}) \subset \widehat{D}_{WM}(\mathcal{T},P,h,r_{\mathrm{cp}}) \subset \overline{D}^*_{WM}(\mathcal{T},P,h,r_{\mathrm{cp}})$ for all $r_{\mathrm{cp}} \in [4r,\tilde{r}]\cap\mathcal{R}$. Recall that we use the convention that $\overline{y}_i = \overline{m}_i = 0$ if $i \le 0$ and $\overline{y}_i = \overline{m}_i = 1$ if $i > d$. Since the CUSUM and the envelope statistics are linear, we have the decompositions

$$
\begin{aligned}
\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}(\mathcal{V}_{2r_{\mathrm{cp}}}) &= \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}(\mathcal{V}_{2r_{\mathrm{cp}}}) + \frac{1}{2r_{\mathrm{cp}}}\left(\sum_{k'=k}^{k+2r_{\mathrm{cp}}-1} \overline{e}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) - \sum_{k'=k-2r_{\mathrm{cp}}}^{k-1} \overline{e}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}})\right) \\
\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}},\mathcal{V}^-_{r_{\mathrm{cp}}}) &= \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}},\mathcal{V}^-_{r_{\mathrm{cp}}}) + \frac{1}{2r_{\mathrm{cp}}}\sum_{k'=k-r_{\mathrm{cp}}}^{k+r_{\mathrm{cp}}-1} \left(\overline{e}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{e}_{k'}(\mathcal{V}^-_{r_{\mathrm{cp}}})\right) ,
\end{aligned}
$$

where the two latter random variables are centered and respectively $\zeta(r_{\mathrm{cp}}|\mathcal{V}_{2r_{\mathrm{cp}}}|)^{-1/2}$-subGaussian and $\zeta[r_{\mathrm{cp}}(|\mathcal{V}^+_{r_{\mathrm{cp}}}|\wedge|\mathcal{V}^-_{r_{\mathrm{cp}}}|)]^{-1/2}$-subGaussian. By a union bound, we deduce that, with probability higher than $1-\delta$, we have simultaneously

$$
\max_{r_{\mathrm{cp}}\in[4r,\tilde{r}]\cap\mathcal{R}} \max_{k\in[d]} \left|\widehat{\mathbf{C}}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} - \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}\right| \le \zeta\sqrt{\frac{2}{r_{\mathrm{cp}}|\mathcal{V}_{2r_{\mathrm{cp}}}|}\log\left(\frac{4d|\mathcal{R}|}{\delta}\right)} ; \tag{3.105}
$$

$$
\max_{r_{\mathrm{cp}}\in[4r,\tilde{r}]\cap\mathcal{R}} \max_{k\in[d]} \left|\widehat{\mathbf{\Delta}}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}} - \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}}\right| \le \zeta\sqrt{\frac{2}{(|\mathcal{V}^+_{r_{\mathrm{cp}}}|\wedge|\mathcal{V}^-_{r_{\mathrm{cp}}}|)r_{\mathrm{cp}}}\log\left(\frac{4d|\mathcal{R}|}{\delta}\right)} . \tag{3.106}
$$

To conclude, it suffices to check that $|\mathcal{V}^+_{r_{\mathrm{cp}}}|$, $|\mathcal{V}^-_{r_{\mathrm{cp}}}|$, and $|\mathcal{V}_{r_{\mathrm{cp}}}|$ have been chosen large enough so that the right-hand side of the two above equations is at most $h/32$.

By definition of $\mathcal{V}^+_{r_{\mathrm{cp}}}$ and $\mathcal{V}^-_{r_{\mathrm{cp}}}$, we know that $|\mathcal{V}^+_{r_{\mathrm{cp}}}|\wedge|\mathcal{V}^-_{r_{\mathrm{cp}}}| \ge 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\mathrm{cp}}h^2}$ which implies that (3.106) is at most $h/32$.

If $r_{\rm cp} \le \tilde{r}/2$, then $\mathcal{V}_{2r_{\rm cp}} = |\mathcal{V}^+_{2r_{\rm cp}}| + |\mathcal{V}^-_{2r_{\rm cp}}| \ge 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\rm cp}h^2}$, which implies that (3.105) is at most $h/32$. Finally, for $r_{\rm cp} = \tilde{r}$, we use

$$r_{\rm cp} \ge 4r_0 \ge 2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{|P|h^2}$$

and that $|\mathcal{V}| = |P|$ to conclude that (3.105) is at most $h/32$. □

Proof of Lemma 3.5.28. In the analysis of **DimensionReduction**, we introduced in (3.55) the sets $D^*_{\rm cp}(P,h,r)$ of questions such that the corresponding CUSUM of the mean expert in P is above $h/2$ at scale $8r$. Recall the set $Q^*_{\rm cp} := Q^*_{\rm cp}(P,h,r) = \textbf{Encode}-\textbf{Set}(D^*_{\rm cp}(P,h,r),r)$. In Lemma 3.5.10, we stated that, for some $h \in \mathcal{H}$ and $r \in \mathcal{R}$, we have

$$\|M(P) - \overline{M}(P)\|_F^2 \le 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}|\left\|\left[\Theta(P,Q^*_{\rm cp}) - \overline{\Theta}(P,Q^*_{\rm cp})\right]_{\sqrt{r}h}\right\|_F^2 \ . \tag{3.107}$$

Define $D^*_{\rm env} := D^*_{\rm env}(\mathcal{T},P,h,r) = \{k \in [d] \ : \ \mathbf{\Delta}^{*({\rm ext})}_{k,r}(\mathcal{V}^+_r,\mathcal{V}^-_r) \ge h/2\}$ for the questions where the population width between $\mathcal{V}^+_r$ and $\mathcal{V}^-_r$ at scale r is at least $h/2$. Besides, we define $Q^*_{\rm env} := Q^*_{\rm env}(\mathcal{T},P,h,r) = \textbf{Encode}-\textbf{Set}(D^*_{\rm env},r)$. If $l \in Q^*_{\rm cp} \setminus Q^*_{\rm env}$, then for any $i,j \in P$, we have

$$|\Theta_{i,l} - \Theta_{j,l}| \le \frac{1}{\sqrt{r}}\sum_{k'=lr}^{(l+1)r-1} \overline{m}_{k'}(\mathcal{V}^+_r) - \overline{m}_{k'}(\mathcal{V}^-_r) \le 2\sqrt{r}\mathbf{\Delta}^{(*{\rm ext})}_{lr,r} < \sqrt{r}h \ .$$

Hence, it follows that

$$\left\|\left[\Theta(P,Q^*_{\rm cp}) - \overline{\Theta}(P,Q^*_{\rm cp})\right]_{\sqrt{r}h}\right\|_F^2 = \left\|\left[\Theta(P,Q^*_{\rm cp}\cap Q^*_{\rm env}) - \overline{\Theta}(P,Q^*_{\rm cp}\cap Q^*_{\rm env})\right]_{\sqrt{r}h}\right\|_F^2 \ . \tag{3.108}$$

In light of (3.107) and (3.108), we only have to prove that, for any fixed $\mathcal{T}$, P, h, and r, we have

$$D^*_{\rm cp}(P,h,r) \cap D^*_{\rm env}(\mathcal{T},P,h,r) \subset \bigcup_{r_{\rm cp}\in[4r,\tilde{r}]\cap\mathcal{R}} D^*_{WM}(\mathcal{T},P,h,r,r_{\rm cp}) \ . \tag{3.109}$$

Since the remainder of the proof heavily relies on the comparisons between CUSUM statistics for different subsets of experts, we respectively write $\mathbf{C}^{*({\rm ext})}_{k,r}(\mathcal{V}_r)$ and $\Delta^{*({\rm ext})}_{k,r}(\mathcal{V}^+_r,\mathcal{V}^-_r)$ instead of $\mathbf{C}^{*({\rm ext})}_{k,r}$ and $\Delta^{*({\rm ext})}_{k,r}$ to better keep track of the dependencies. Fix any question $k \in D^*_{\rm cp}(P,h,r) \cap D^*_{\rm env}(\mathcal{T},P,h,r)$ and define

$$r_{\min} = \max\{r' \in \mathcal{R} \ : \ \mathbf{C}^{*({\rm ext})}_{k,r'}(\mathcal{V}_{r'}) < h/8\} \ ,$$

with the convention that $\max(\varnothing) = 1$. $r_{\min}$ can be interpreted as the largest scale r' in $\mathcal{R}$ such that the population CUSUM at scale r' applied to $\mathcal{V}_{r'}$ is smaller than $h/8$. By definition, we have $\mathcal{V}_{2\tilde{r}} = P$. As a consequence, for any $r' \ge 2\tilde{r}$, we have $\mathbf{C}^{*({\rm ext})}_{k,r'}(\mathcal{V}_{r'}) = \mathbf{C}^*_{k,r'}(P) \ge \mathbf{C}^*_{k,2\tilde{r}}(P) \ge \mathbf{C}^*_{k,8r}(P) \ge h/8$ since $k \in D^*_{\rm cp}(P,h,r)$ and since $\tilde{r} \ge 8r$ (see (3.97)). This implies that $r_{\min} \le \tilde{r}$. We consider two distinct cases.

Case 1: $r_{\min} \le 4r$**.** Then, we simply choose $r_{\rm cp} = 4r$. By definition of $r_{\min}$, we have $\mathbf{C}^{*({\rm ext})}_{k,2r_{\rm cp}}(\mathcal{V}_{2r_{\rm cp}}) \ge h/8$. Since $k \in D^*_{\rm env}(\mathcal{T},P,h,r)$, we can lower bound the envelope statistic as

$$\mathbf{\Delta}^{*({\rm ext})}_{k,r_{\rm cp}}(\mathcal{V}^+_{r_{\rm cp}},\mathcal{V}^-_{r_{\rm cp}}) \ge \frac{1}{4}\mathbf{\Delta}^{*({\rm ext})}_{k,r}(\mathcal{V}^+_r,\mathcal{V}^-_r) \ge h/8 \ .$$

We have proved that $k \in D^*_{WM}(\mathcal{T},P,h,r,r_{\rm cp})$.

Case 2: $r_{\min} \in (4r,\tilde{r}]$**.** In that case, we choose $r_{\rm cp} = r_{\min} \ge 8r$ (since $r_{\min}$ is a power of 2). By definition of $r_{\min}$, we have both $\mathbf{C}^{*({\rm ext})}_{k,2r_{\rm cp}} \ge h/8$ and $\mathbf{C}^{*({\rm ext})}_{k,r_{\rm cp}} < h/8$. Since $k \in D^*_{\rm cp}(P,h,r)$ and $r_{\rm cp} \ge 8r$, we also deduce by monotonocity that the CUSUM of the mean expert in P at scale $r_{\rm cp}$ is higher than $h/2$, this is $\mathbf{C}^*_{k,r_{\rm cp}} \ge \mathbf{C}^*_{k,8r} \ge h/2$ since $k \in D^*_{\rm cp}(P,h,r)$ – see (3.55).

Remark that, since $r_{\rm cp} \le \tilde{r}$, we have $\mathcal{V}_{r_{\rm cp}} = \mathcal{V}^+_{r_{\rm cp}} \cup \mathcal{V}^-_{r_{\rm cp}}$. Without loss of generality, we can assume that $|\mathcal{V}^+_{r_{\rm cp}}| \ge |\mathcal{V}^-_{r_{\rm cp}}|$. This implies in particular that

$$\begin{aligned} r_{\rm cp}\mathbf{C}^{*({\rm ext})}_{k,r_{\rm cp}}(\mathcal{V}_{r_{\rm cp}}) &\ge \frac{|\mathcal{V}^+_{r_{\rm cp}}|}{|\mathcal{V}_{r_{\rm cp}}|}\sum_{k'=k}^{k+r_{\rm cp}-1}\overline{m}_k(\mathcal{V}^+) - \frac{|\mathcal{V}^+_{r_{\rm cp}}|}{|\mathcal{V}_{r_{\rm cp}}|}\sum_{k'=k-r_{\rm cp}}^{k-1}\overline{m}_k(\mathcal{V}^+) \\ &\ge \frac{1}{2}\left(\sum_{k'=k}^{k+r_{\rm cp}-1}\overline{m}_k(\mathcal{V}^+_{r_{\rm cp}}) - \sum_{k'=k-r_{\rm cp}}^{k-1}\overline{m}_k(\mathcal{V}^+_{r_{\rm cp}})\right) \ . \end{aligned}$$

Since $\mathbf{C}^{*}_{k,r_{\rm cp}}(P) \geq h/2$ and $\mathbf{C}^{*(\rm ext)}_{k,r_{\rm cp}}(\mathcal{V}_{r_{\rm cp}}) \leq h/8$, this implies that

$$\begin{aligned} h/4 \leq \mathbf{C}^{*}_{k,r_{\rm cp}}(P) - 2\mathbf{C}^{*(\rm ext)}_{k,r_{\rm cp}}(\mathcal{V}_{r_{\rm cp}}) &\leq \frac{1}{r_{\rm cp}} \left(\sum_{k'=k}^{k+r_{\rm cp}-1} \overline{m}_{k'}(P) - \overline{m}_{k'}(\mathcal{V}^{+}_{r_{\rm cp}}) \right) + \frac{1}{r_{\rm cp}} \left(\sum_{k'=k-r_{\rm cp}}^{k-1} \overline{m}_{k'}(\mathcal{V}^{+}_{r_{\rm cp}}) - \overline{m}_{k'}(P) \right) \\ &\leq \frac{1}{r_{\rm cp}} \left(\sum_{k'=k-r_{\rm cp}}^{k-1} \overline{m}_{k'}(\mathcal{V}^{+}_{r_{\rm cp}}) - \overline{m}_{k'}(P) \right) \\ &\leq \frac{1}{r_{\rm cp}} \left(\sum_{k'=k-r_{\rm cp}}^{k+r_{\rm cp}-1} \overline{m}_{k'}(\mathcal{V}^{+}_{r_{\rm cp}}) - \overline{m}_{k'}(\mathcal{V}^{-}_{r_{\rm cp}}) \right) \\ &= 2\mathbf{\Delta}^{*(\rm ext)}_{k,r_{\rm cp}}(\mathcal{V}^{+}_{r_{\rm cp}}, \mathcal{V}^{-}_{r_{\rm cp}}) \ . \end{aligned}$$

Hence, we have proved that $\mathbf{\Delta}^{*(\rm ext)}_{k,r_{\rm cp}}(\mathcal{V}^{+}_{r_{\rm cp}}, \mathcal{V}^{-}_{r_{\rm cp}}) \geq h/8$ and $\mathbf{C}^{*(\rm ext)}_{k,2r_{\rm cp}}(\mathcal{V}_{r_{\rm cp}}) \geq h/8$. Thus, $k \in D_{WM}(\mathcal{T}, P, h, r_{\rm cp})$. We have shown (3.109) and the proof is finished.

□

Proof of Lemma 3.5.30. We fix $h \in \mathcal{H}$ and $r \in \mathcal{R}$. Let us consider a subgroup $\overline{P} \subset G \in \mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ Recall that the blocks $\overline{Q}^{*}_{WM}(\overline{P}, h, r) = \cup_{r_{\rm cp}=4r}^{\tilde{r}} \overline{Q}^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})$ – see the definitions in Section 3.5.5.1. Again, we remove the dependency on $\mathcal{T}$ in $Q^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})$ for the ease of exposition. First, we bound $\sum_{\overline{P}\in\mathcal{P}^{*}(p)} |\overline{Q}^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})|$ before summing over the range over all possible $r_{\rm cp}$.

Let us consider some $l \in \overline{Q}^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})$. By definition, there exists at least one question $k(l) \in [lr, (l+1)r)$ such that we have simultaneously $\overline{\mathbf{C}}^{*(\rm ext)}_{k(l),2r_{\rm cp}} \geq h/32$ and $\overline{\mathbf{\Delta}}^{*(\rm ext)}_{k(l),r_{\rm cp}} \geq h/32$. For $l \in \mathcal{Q}_r \setminus \overline{Q}^{*}_{WM}(r_{\rm cp})$, we simply define $k(l) = l$. We deduce from this definition that

$$|\overline{Q}^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})| \leq \sum_{l\in\mathcal{Q}_r} \mathbf{1}\{\overline{\mathbf{C}}^{*(\rm ext)}_{k(l),2r_{\rm cp}} \geq h/32\} \mathbf{1}\{\overline{\mathbf{\Delta}}^{*(\rm ext)}_{k(l),r_{\rm cp}} \geq h/32\} \ . \tag{3.110}$$

This implies that

$$|\overline{Q}^{*}_{WM}(\overline{P}, h, r, r_{\rm cp})| \leq \frac{32}{h} \sum_{l\in\mathcal{Q}_r} \overline{\mathbf{C}}^{*(\rm ext)}_{k(l),2r_{\rm cp}} \leq 2^{8} \frac{r_{\rm cp}}{rh} \tag{3.111}$$

where the last inequality comes from the fact that the total variation of $\overline{m}(\mathcal{V}_{2r_{\rm cp}})$ is at most 1 and that, for any $l \in \mathcal{Q}_r$, the interval $[k(l)-2r_{\rm cp}, k(l)+2r_{\rm cp})$ intersects at most $8r_{\rm cp}/r$ intervals of the form $[k(l')-2r_{\rm cp}, k(l')+2r_{\rm cp})$ with $l' \in \mathcal{Q}_r$.

Let p be an integer and assume that $|\overline{P}| \in [p, 2p)$. Let us introduce $\Gamma := \Gamma(p, h, r_{\rm cp}) = \frac{\tilde{r}}{r_{\rm cp}} \geq 1$, where we recall that $\tilde{r} \geq 4r_0$ is defined by $r_0 = 2^{9}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{ph^2}$ in (3.97). Intuitively, Γ would correspond to the number a^{+}_{WM} and a^{-}_{WM} of sets of experts above $\overline{P}$ or below $\overline{P}$ that would be considered if those sets were of size p. More generally, $\mathcal{V}^{+}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h) = \cup_{a=1}^{a^{+}_{WM}} G^{(a)}$ contains at most Γ groups of size at least p among $G^{(1)}, \dots, G^{(a_{WM}-1)}$ since the total size of the groups $G^{(a)}$ with $a \leq a_{WM} - 1$ must be less than $2^{11}\log(4d|\mathcal{R}|/\delta)\frac{\zeta^2}{r_{\rm cp}h^2}$. Thus, we deduce that $\mathcal{V}^{-}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h) \cup \overline{P} \cup \mathcal{V}^{+}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h)$ contains at most $2\Gamma + 3$ groups of size at least p.

The following lemma states that the neighbourhoods $\mathcal{V}^{-}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h) \cup \overline{P} \cup \mathcal{V}^{+}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h)$ of groups $\overline{P}$ in $\mathcal{P}^{*}(p)$ only intersect on a few groups.

Lemma 3.5.31. *Consider any group $\overline{P} \in \mathcal{P}^{*}(p)$. There exists at most $4\Gamma + 3$ groups $\overline{P}' \in \mathcal{P}^{*}(p)$ such that*

$$\left(\mathcal{V}^{-}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h) \cup \overline{P} \cup \mathcal{V}^{+}_{r_{\rm cp}}(\mathcal{T}, \overline{P}, h)\right) \cap \left(\mathcal{V}^{-}_{r_{\rm cp}}(\mathcal{T}, \overline{P}', h) \cup \overline{P}' \cup \mathcal{V}^{+}_{r_{\rm cp}}(\mathcal{T}, \overline{P}', h)\right) \neq \varnothing \ . \tag{3.112}$$

As in the proof of Lemma 3.5.16, we introduce the width of the matrix M on a set A of experts and an interval of questions $[k_1, k_2]$ by

$$W_{\infty,1}(M, A, [k_1, k_2]) := \max_{i,j\in A} \sum_{k=k_1}^{k_2} |M_{i,k} - M_{j,k}| \ .$$

From (3.110) again, we deduce that

$$\begin{aligned}\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})| &\le \frac{32}{h}\sum_{\overline{P}\in\mathcal{P}^*(p)}\sum_{l\in\mathcal{Q}_r}\overline{\boldsymbol{\Delta}}^{*(\mathrm{ext})}_{k(l),r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P},h),\mathcal{V}^-_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P},h))\\ &\le \frac{32}{h}\sum_{l\in\mathcal{Q}_r}\sum_{\overline{P}\in\mathcal{P}^*(p)}\frac{1}{2r_{\mathrm{cp}}}W_{\infty,1}(\mathcal{V}^+_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P},h)\cup\overline{P}\cup\mathcal{V}^-_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P},h),[k-r_{\mathrm{cp}},k+r_{\mathrm{cp}}])\\ &\le \frac{32}{rh}(4\Gamma+3)d\ ,\end{aligned}$$

where the last inequality comes Lemma 3.5.31 and the fact that the sum over disjoints sets $\mathcal{V}^-_{r_{\mathrm{cp}}}(\overline{P})\cup\overline{P}\cup\mathcal{V}^+_{r_{\mathrm{cp}}}(\overline{P})$ of $W_{\infty,1}(\mathcal{V}^+_{r_{\mathrm{cp}}}(\overline{P})\cup\overline{P}\cup\mathcal{V}^-_{r_{\mathrm{cp}}}(\overline{P}),[k-r_{\mathrm{cp}},k+r_{\mathrm{cp}}))$ is upper bounded by $2r_{\mathrm{cp}}$ since the total variation of any column of M is at most 1.

Combining (3.111) with the latter upper bound together with $|\mathcal{P}^*(p)|\le n/p$ we deduce that

$$\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\frac{nr_{\mathrm{cp}}}{prh}\wedge\frac{\Gamma d}{rh}\ . \tag{3.113}$$

If $r_0>r$, then we have $\Gamma\le\frac{8r_0}{r_{\mathrm{cp}}}$. This implies that

$$\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\frac{nr_{\mathrm{cp}}}{prh}\wedge\frac{r_0 d}{r_{\mathrm{cp}}rh}\lesssim\frac{\sqrt{ndr_0}}{rh\sqrt{p}}\ .$$

Since $r_{\mathrm{cp}}\le[4r,\tilde{r}]\cap\mathcal{R}$, there are at most $c\log(d)$ possible values for r_{cp}, we conclude that

$$\sum_{r_{\mathrm{cp}}}\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\frac{nr_0}{prh}\wedge\frac{r_0 d}{r^2h}\wedge\frac{\sqrt{ndr_0}}{rh\sqrt{p}}\ .$$

Otherwise, if $r_0\le r$, then $\Gamma\le 8$ and $r_{\mathrm{cp}}\in[4r,8r]$. We deduce from (3.113) that

$$\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\frac{n}{ph}\wedge\frac{d}{rh}\le\frac{\sqrt{nd}}{\sqrt{r}h\sqrt{p}}\ .$$

We have proved that, in any case,

$$\sum_{r_{\mathrm{cp}}}\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\log(d)\left[\frac{d(r_0\vee r)}{r^2h}\wedge\frac{\sqrt{nd(r_0\vee r)}}{rh\sqrt{p}}\right]. \tag{3.114}$$

To establish the remaining bound for the sum of $|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|$, we control each $|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|$ individually in a similar fashion to what we did for the analysis of the oblivious hierarchical sorting estimator $\hat{\pi}_{HT}$. First, we have $\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})\subset\mathcal{Q}_r$ so that $|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\le d/r$. Besides, arguing as in the proof of Lemma 3.5.9, $|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim r_{\mathrm{cp}}/(rh)\lesssim(r_0\vee r)/(rh)$.

$$\sum_{r_{\mathrm{cp}}}\sum_{\overline{P}\in\mathcal{P}^*(p)}|\overline{Q}^*_{WM}(\overline{P},h,r,r_{\mathrm{cp}})|\lesssim\log(d)\left[\frac{nd}{pr}\wedge\frac{n(r_0\vee r)}{prh}\right]. \tag{3.115}$$

Combining (3.114) and (3.115) concludes the proof. □

Proof of Lemma 3.5.31. Consider two distinct groups $\overline{P}$ and $\overline{P}'$ in $\mathcal{P}^*(p)$. Let $(G^{(a)}(\overline{P}))_{a\in\mathbb{Z}}=\mathbf{Order}(\mathcal{T},\overline{P})$ be the ordering of $\mathcal{L}^{(\mathbf{0},\mathbf{1})}(\mathcal{T})$ centered on $\overline{P}$ and $a'\in\mathbb{Z}$ the index of the leaf $G^{(a')}(\overline{P})$ containing $\overline{P}'$. Obviously, $|G^{(a')}|\ge|\overline{P}'|\ge p$.

Without loss of generality, we assume that $a'>0$. In that case, if (3.112) is satisfied then necessarily

$$(\mathcal{V}^+_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P},h)\cup\overline{P})\cap\mathcal{V}^-_{r_{\mathrm{cp}}}(\mathcal{T},\overline{P}',h)\neq\varnothing\ .$$

This can only happen if the number of leaves $G^{(a)}(\overline{P})$ for $0<a<a'$ that are of size at least p is less than or equal to 2Γ. The same holds if $a'<0$ and this proves the lemma. □

3.5.6 Proof of Lemma 3.4.2 and Theorem 3.4.3

3.5.6.1 Proof of Lemma 3.4.2

We start with the case $\lambda_- \in [2/d, 1]$. The random variable $n_{i,[(j-1)l(\lambda)+1,jl(\lambda)]}$ is distributed as a Poisson random variable with parameter $\lambda l(\lambda)$. Let us apply Chernoff's inequality for Poisson random variable (e.g. [10], section 2.2). We have

$$\mathbb{P}\left[n_{i,[(j-1)l(\lambda)+1,jl(\lambda)]} \le \lambda l(\lambda)/2\right] \le \exp\left[-\frac{3}{28}\lambda l(\lambda)\right] \le \frac{\delta}{nd} ,$$

provided that $\lambda l(\lambda) \ge \frac{28}{3}\log(nd/\delta)$. Since $\lambda_- \le 1$, we have $\lambda l(\lambda)/2 \ge \Upsilon^*$. In view of the definition of Υ^*, the condition $\lambda l(\lambda) \ge \frac{28}{3}\log(nd/\delta)$ is therefore valid and we conclude that

$$\mathbb{P}\left[n_{i,[(j-1)l(\lambda)+1,jl(\lambda)]} \le \Upsilon^*\right] \le \frac{\delta}{nd} ,$$

and the first result follows. Turning to the second result, we observe that $n_{i,\{j\}}$ is distributed as a Poisson random variable. We apply again Chernoff's inequality to derive that

$$\mathbb{P}\left[n_{i,\{j\}} \le \frac{\lambda}{2}\right] \le \exp\left[-\frac{3}{28}\lambda\right] \le \frac{\delta}{nd} ,$$

since $\lambda \ge \frac{28}{3}\log(nd/\delta)$. Since $\lambda \ge 2\lambda_-\Upsilon^*$, the result follows.

3.5.6.2 Proof of Theorem 3.4.3

If $\lambda_- \le 2/d$, we use the trivial bound $\|M_{\hat{\pi}^{-1}_{WMP}} - M_{\pi^{*-1}}\|_F^2 \le nd$, which ensures that

$$\mathbb{E}\left[\|M_{\hat{\pi}^{-1}_{WMP}} - M_{\pi^{*-1}}\|_F^2\right] \le \frac{n}{\lambda_-} \le c\log^{c'}\left(\frac{nd\lambda^{1/2}}{\zeta_-}\right)\frac{n}{\lambda} .$$

If $\lambda_- \ge 1$, then Lemma 3.4.2 ensures that, with probability higher than $1-\delta$, we are able to build the Υ^* subsamples and we are in position to apply Theorem 3.2.3 with subGaussian norm $\zeta/\lfloor\lambda_-\rfloor^{1/2}$. Hence, with probability higher than $1 - c'\log^9(nd\lambda_-^{1/2}/(\delta\zeta_-))\delta$, we have

$$\begin{aligned}\|M_{\hat{\pi}^{-1}_{WMP}} - M_{\pi^{*-1}}\|_F^2 &\le c\log^{11}\left(\frac{nd\lfloor\lambda_-\rfloor^{1/2}}{\delta\zeta_-}\right)\mathcal{R}_F(n,d,\zeta\lfloor\lambda_-\rfloor^{-1/2}) \\ &\le c'\log^{c''}\left(\frac{nd\lambda^{1/2}}{\zeta_-}\right)\mathcal{R}_F(n,d,\zeta\lfloor\lambda\rfloor^{-1/2}) ,\end{aligned}$$

where we use the definition of δ and λ_- in the last line. On the complementary event, we simply use that $\|M_{\hat{\pi}^{-1}_{WMP}} - M_{\pi^{*-1}}\|_F^2 \le nd$. Since δ has been chosen small enough, we can conclude that

$$\mathbb{E}[\|M_{\hat{\pi}^{-1}_{WMP}} - M_{\pi^{*-1}}\|_F^2] \le c'\log^{c''}\left(\frac{nd\lambda^{1/2}}{\zeta_-}\right)\mathcal{R}_F(n,d,\zeta\lambda^{-1/2}) .$$

It remains to consider the case where $\lambda_- \in [2/d, 1]$. Working under the event of probability higher than $1-\delta$ ensured by Lemma 3.4.2, we have Υ^* independent samples $Y^{\downarrow(0)}, \dots, Y^{\downarrow(\Upsilon^*-1)}$ of size $n \times \lfloor d/l(\lambda)\rfloor$. Define the matrix $M^{\downarrow}$ of size $n \times \lfloor d/l(\lambda)\rfloor$ by $M^{\downarrow}_{i,j} = M_{i,l(\lambda)(j-1)+1}$. Obviously, $M^{\downarrow}_{\pi^{*-1}}$ is a bi-isotonic matrix. Besides, for $s = 0, \dots, \Upsilon^* - 1$, $(i,j) \in [n] \times \lfloor d/l(\lambda)\rfloor$, we have the decomposition

$$Y^{\downarrow(s)}_{ij} = M^{\downarrow(s)}_{ij} + E^{\downarrow(s)}_{ij} ,$$

where $M^{\downarrow(s)}_{ij}$ belongs to $[M^{\downarrow}_{ij}, M^{\downarrow}_{ij+1}]$ with the convention $M^{\downarrow(s)}_{i,\lfloor d/l(\lambda)\rfloor+1} = 1$ and the $E^{\downarrow(s)}_{ij}$'s are independent and, for fixed i and j, are i.i.d. distributed and ζ-subGaussian. In fact, the $M^{\downarrow(s)}_{ij}$ are random since $M^{\downarrow(s)}_{ij}$ has been sampled uniformly in $\{M_{i,l(\lambda)(j-1)+1}, M_{i,l(\lambda)(j-1)+2}, \dots, M_{i,l(\lambda)(j-1)+l(\lambda)}\}$. Besides, those are correlated with the noise $E^{\downarrow(s)}_{ij}$. For the sake of the analysis, it is in fact easier to consider that $M^{\downarrow(s)}_{ij}$ has been set by an adversary. Hence, we fall into the semi-random model of Section 3.5.7 and we are in position to apply Theorem 3.5.32 to $\hat{\pi}_{WM-SR}$. With probability at least $1 - c'n\log^9(\frac{nd}{\delta\zeta_-})\delta$, we have

$$\|M^{\downarrow}_{\hat{\pi}^{-1}_{\mathrm{WM-SR}}} - M^{\downarrow}_{\pi^{*-1}}\|_F^2 \le c\log^{11}\left(\frac{2nd}{\delta\zeta_-}\right)[\mathcal{R}_F(n,\lfloor d/l(\lambda)\rfloor,\zeta) + n] ,$$

Define the matrix $M^{\downarrow\uparrow}$ of size $n\times d$ such that each column is duplicated $l(\lambda)$ times, except the last one which has been duplicated $l(\lambda)-l(\lambda)\lfloor d/l(\lambda)\rfloor$. We readily deduce that

$$\|M^{\downarrow\uparrow}_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M^{\downarrow\uparrow}_{\pi^{*-1}}\|_F^2 \le c' l(\lambda)\log^{11}\left(\frac{2nd}{\delta\zeta_-}\right)[\mathcal{R}_F(n,\lfloor d/l(\lambda)\rfloor,\zeta)+n] \quad , \tag{3.116}$$

By triangular inequality, we have

$$\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M_{\pi^{*-1}}\|_F^2 \le 2\|M^{\downarrow\uparrow}_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M^{\downarrow\uparrow}_{\pi^{*-1}}\|_F^2 + 8\|M-M^{\downarrow\uparrow}\|_F^2 \ .$$

Thus it remains to upper bound the square Euclidean norm of each row of $M-M^{\downarrow\uparrow}$:

$$\begin{aligned}
\sum_{j=1}^d [M-M^{\downarrow\uparrow}]_{i,j}^2 &= \sum_{k=1}^{\lfloor d/l(\lambda)\rfloor}\sum_{r=1}^{l(\lambda)}[M_{i,(k-1)l(\lambda)+r} - M_{i,(k-1)l(\lambda)+1}]^2 \\
&\quad + \sum_{r=1}^{d-l(\lambda)\lfloor d/l(\lambda)\rfloor}\left[M_{i,(\lfloor d/l(\lambda)\rfloor-1)l(\lambda)+r} - M_{i,(\lfloor d/l(\lambda)\rfloor-1)l(\lambda)+1}\right]^2 \\
&\le \sum_{k=1}^{\lfloor d/l(\lambda)\rfloor} l(\lambda)[M_{i,kl(\lambda)} - M_{i,(k-1)l(\lambda)+1}]^2 + l(\lambda)\left[M_{i,d} - M_{i,(\lfloor d/l(\lambda)\rfloor-1)l(\lambda)+1}\right]^2 \\
&\le 2l(\lambda) \ ,
\end{aligned}$$

since the total variation of the i-th row of M is at most one. Hence, $\|M-M^{\downarrow\uparrow}\|_F^2 \le 2nl(\lambda)$. Together with (3.116), we conclude that

$$\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M_{\pi^{*-1}}\|_F^2 \le c' l(\lambda)\log^{11}\left(\frac{2nd}{\delta\zeta_-}\right)[\mathcal{R}_F(n,\lfloor d/l(\lambda)\rfloor,\zeta)+n] \ .$$

with probability at least $1-c'n\log^9(\frac{nd}{\delta\zeta_-})\delta$. Since δ has been chosen small enough and since $\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}}-M_{\pi^{*-1}}\|_F^2 \le nd$, we conclude that

$$\mathbb{E}\left[\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M_{\pi^{*-1}}\|_F^2\right] \le c' l(\lambda)\log^{11}\left(\frac{2nd}{\zeta_-}\right)[\mathcal{R}_F(n,\lfloor d/l(\lambda)\rfloor,\zeta)+n] \ .$$

Since $l(\lambda)\le c'\log^{c''}(nd(\lambda\vee 1)/\zeta_-)/\lambda$, we deduce from this bound that

$$\begin{aligned}
&\mathbb{E}\left[\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M_{\pi^{*-1}}\|_F^2\right] \\
&\le c'\log^{c''}\left(\frac{2nd}{\zeta_-}\right)\left[\left(\frac{\zeta}{\sqrt{\lambda}}\right)^2\left\{\frac{nd^{1/6}}{(\frac{\zeta}{\sqrt{\lambda}})^{1/3}}\wedge\frac{n^{3/4}d^{1/4}}{(\frac{\zeta}{\sqrt{\lambda}})^{1/2}}\wedge n\sqrt{d\lambda}\wedge\frac{n^{2/3}\sqrt{d}\lambda^{1/3}}{(\frac{\zeta}{\sqrt{\lambda}})^{1/3}}+n\right\}+\frac{n}{\lambda}\right] \\
&\le c'\log^{c''}\left(\frac{2nd}{\zeta_-}\right)\left[\mathcal{R}_F[n,d,\zeta/\sqrt{\lambda}]+\frac{n}{\lambda}\right] \ ,
\end{aligned}$$

since $\lambda_-\le 1$. Again, since $\lambda_-\le 1$, we have $\lambda\le c_3\log^{c_4}(nd(\lambda\vee 1)/\zeta)$ for some numerical constant c_3 and c_4. We conclude that

$$\mathbb{E}\left[\|M_{\hat\pi^{-1}_{\mathrm{WM}-SR}} - M_{\pi^{*-1}}\|_F^2\right] \le c'\log^{c''}\left(\frac{2nd}{\zeta_-}\right)\left[\mathcal{R}_F[n,d,\zeta/\sqrt{\lambda}]+\frac{n}{\lambda}e^{-\frac{\lambda}{c_3\log^{c_4}(nd(\lambda\vee1)/\zeta)}}\right] \ ,$$

which concludes the proof.

3.5.7 Permutation estimation in the semi-random model

3.5.7.1 Model and algorithm

We now consider a slightly different model with Υ^* samples $Y^{(1)},\dots,Y^{(\Upsilon^*-1)}$. The noise matrices $E^{(1)},\dots,E^{(\Upsilon^*-1)}$ are sampled independently (as previously) and $Y_{ij}^{(t)} = E_{ij}^{(t)} + M_{ij}^{(t)}$ where $M_{ij}^{(t)}$ is chosen by an adversary in $[M_{ij}, M_{i,j+1}]$. This slightly different model is mainly motivated by the analysis of the partial observation scheme in Section 3.4. In particular, building upon this model and relying on the corresponding modifications in the algorithm allows us to recover the right dependency with respect to ζ in Section 3.4.

We consider a slight variant $\hat\pi_{WM-SR}$ of the estimator $\hat\pi_{WM}$ to handle the adversarial differences. The procedure $\hat\pi_{WM-SR}$ is computed exactly as $\hat\pi_{WM}$ except that

- In **Pivot** (Algorithm 4), the threshold $\beta_{\rm tris}\sqrt{\log(\frac{2|P|}{\delta})}$ is replaced by $\beta_{\rm tris}\sqrt{\log(\frac{2|P|}{\delta})} + 4\|\omega\|_\infty/\|\omega\|_2$ and $\overline{\beta}_{\rm tris}\sqrt{\log(\frac{2|P|}{\delta})}$ is replaced by $\overline{\beta}_{\rm tris}\sqrt{\log(\frac{2|P|}{\delta})} + 8\|\omega\|_\infty/\|\omega\|_2$
- In **DimensionReduction – WM** (Algorithm 13), we respectively replace the definitions of the CUSUM and empirical width by

$$\widehat{\mathbf{\Delta}}^{(\rm ext)}_{k,r'}(\mathcal{V}^+,\mathcal{V}^-) = \sum_{k'=k-r'}^{k+r'-1} \overline{y}_{k'}(\mathcal{V}^+) - \overline{y}_{k'-1}(\mathcal{V}^-) \; ; \tag{3.117}$$

$$\widehat{\mathbf{C}}^{(\rm ext)}_{k,r'}(\mathcal{V}) = \sum_{k'=k}^{k+r'-1} \overline{y}_{k'}(\mathcal{V}) - \sum_{k'=k-r'-1}^{k-2} \overline{y}_{k'}(\mathcal{V}) \; . \tag{3.118}$$

Theorem 3.5.32. *There exist three numerical constants c, c', and c_0 such that the following holds. Fix $\delta > 0$ and assume that $\Upsilon \geq c_0 \log^8(nd/(\delta\zeta_-))$. For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{BISO}$, the hierarchical sorting tree estimator with memory $\hat{\pi}_{WM-SR}$ satisfies*

$$\|M_{\hat{\pi}^{-1}_{\rm WM-SR}} - M_{\pi^{*-1}}\|_F^2 \leq c\log^{11}\left(\frac{2nd}{\delta\zeta_-}\right)[\mathcal{R}_F(n,d,\zeta)+n] \; , \tag{3.119}$$

with probability at least $1 - c'n\log^9(\frac{nd}{\delta\zeta_-})\delta$.

3.5.7.2 Proof of Theorem 3.5.32

The proof follows the main steps as that of Theorem 3.2.3 and we mainly emphasize here the differences. In the proof of Theorem 3.2.3, we often work with the aggregated model (3.50) $Z = \Theta + N$ which is restricted to a subset P of experts and a subset $Q \subset \mathcal{Q}_r$ of questions aggregated at scale r – see **Encode – Matrix** for details. For $t = 0, \dots, \Upsilon - 1$, the counterpart of (3.50) is the following

$$Z^{(t)} = \Theta^{(t)} + N^{(t)} \; , \tag{3.120}$$

where the entries of $N^{(t)}$ are independent and ζ-subGaussian and $\Theta^{(t)}$ stands for the corresponding aggregation of the matrix $M^{(t)}$. Since the total variation of each row of M is at most one, one readily checks that

$$\sum_{j\in Q} |\Theta^{(t)}_{ij} - \Theta_{ij}| \leq 1 \; . \tag{3.121}$$

Since $\hat{\pi}_{WM-SR}$ is a hierarchical sorting tree estimator, we are in position to control its loss using Proposition 3.5.1. For this purpose, we need to prove that Proposition 3.5.3 still holds in the semi-random model which, in turn, would imply that Corollary 3.5.4 is true. In fact, the proof of Proposition 3.5.3 is verbatim the same except that Lemma 3.5.7 is replaced by the following lemma.

We remind that $P' = \overline{P} \setminus (L \cup U)$, and $\overline{P'} = \overline{P} \setminus (\overline{L} \cup \overline{U})$.

Lemma 3.5.33. *For any non-zero vector $w \in \mathbb{R}^Q_+$, any pivot $\gamma \in \{1, \dots, |\overline{P}|\}$, we have $\mathbb{P}[\mathcal{P}_3] \geq 1 - \delta$. Besides, on the same event of probability at least $1 - \delta$, we have*

$$\left|\langle \Theta_{i,\cdot} - \Theta_{i_\gamma,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq (2\zeta\sqrt{2} + \overline{\beta}_{\rm tris})\sqrt{\log\left(\frac{2|\overline{P}|}{\delta}\right)} + 10\frac{\|w\|_\infty}{\|w\|_2} \quad \text{if } i \in P' \; . \tag{3.122}$$

Proof of Lemma 3.5.33. Consider any sample $t \in [0, \Upsilon - 1]$, any vector $w \in \mathbb{R}^q$, and any $i \in \overline{P}$. As a straightforward consequence of (3.121), we deduce that

$$\left|\langle \Theta^{(t)}_{i,\cdot} - \Theta_{i,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq \frac{\|w\|_\infty}{\|w\|_2} \; . \tag{3.123}$$

We then deduce from a union bound, that with probability higher than $1 - \delta$, we have

$$\left|\langle Z^{(t)}_{i,\cdot}, \frac{w}{\|w\|_2}\rangle - \langle \Theta_{i,\cdot}, \frac{w}{\|w\|_2}\rangle\right| \leq \zeta\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)} + \frac{\|w\|_\infty}{\|w\|_2} \; .$$

simultaneously for all i in $\overline{P}$. The rest of the proof of Lemma 3.5.33 is left unchanged provided that we replace $\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)}$ by $\sqrt{2\log\left(\frac{2|\overline{P}|}{\delta}\right)} + \frac{\|w\|_\infty}{\|w\|_2}$. □

Then, being in position to apply Corollary 3.5.4, we state the counterpart of Proposition 3.5.6.

Proposition 3.5.34. *On the intersection of event ξ (defined in Corollary 3.5.4) and an event of probability higher than $1-5\cdot 2^t\tau_\infty\delta$, it holds that*

$$\sum_{\overline{P}\in\overline{\mathcal{L}}_t}\|M(\overline{P})-\overline{M}(\overline{P})\|^2 \lesssim \log^9\left(\frac{6nd}{\delta\zeta_-}\right)[\mathcal{R}_F(n,d,\zeta)+n] \ .$$

We conclude the proof of Theorem 3.5.32 by combining Proposition 3.5.34 with Corollary 3.5.4. Hence, we only need to prove the last proposition.

3.5.7.3 Proof of Proposition 3.5.34

Again, we only emphasize the differences with the proof of Proposition 3.5.6. We start with the analysis of **DimensionReduction – WM**. Recall that we slightly changed the definition of the CUSUM statistics

$$\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}} = \frac{1}{2r_{\mathrm{cp}}}\left(\sum_{k'=k}^{k+2r_{\mathrm{cp}}-1}\overline{y}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}}) - \sum_{k'=k-2r_{\mathrm{cp}}-1}^{k-2}\overline{y}_{k'}(\mathcal{V}_{2r_{\mathrm{cp}}})\right)$$

by shifting the second sum by one index. The definition of the population CUSUM statistic $\mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ is left unchanged. Similarly, we slightly changed the definition of $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ to

$$\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}} = \frac{1}{2r_{\mathrm{cp}}}\sum_{k'=k-r_{\mathrm{cp}}}^{k+r_{\mathrm{cp}}-1}\overline{y}_{k'}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{y}_{k'-1}(\mathcal{V}^-_{r_{\mathrm{cp}}}) \ ,$$

by shifting again the right hand-side observation by one. With these simple shifts, $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ and $\widehat{\mathbf{C}}^{(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}$ both overestimates $\mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}}$ and $\mathbf{C}^{*(ext)}_{k,2r_{\mathrm{cp}}}$ and arguing as in the proof of Lemma 3.5.27, we will prove that $Q^*_{WM}\subset\widehat{Q}_{WM}$ with probability at least $1-\delta$ –see Lemma 3.5.35 below. However, we need to adapt the definition of $\overline{D}^*_{WM}(\mathcal{T},P,h,r,r_{\mathrm{cp}})$ to cope with this possible bias. Define

$$\overline{D}^*_{WM\text{-}SR\text{-}1}(\mathcal{T},P,h,r,r_{\mathrm{cp}}) = \left\{k\in 1,\dots,d \ : \ \mathbf{C}^{*(\mathrm{ext})}_{k,2r_{\mathrm{cp}}}\geq\frac{h}{128} \text{ and } \mathbf{\Delta}^{*(\mathrm{ext})}_{k,r_{\mathrm{cp}}}\geq\frac{h}{128}\right\} \ ; \tag{3.124}$$

$$\overline{D}^*_{WM\text{-}SR\text{-}2}(\mathcal{T},P,h,r,r_{\mathrm{cp}}) = \left\{k\in 1,\dots,d \ : \ \overline{m}_{k+r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}}) - \overline{m}_{k-r_{\mathrm{cp}}}(\mathcal{V}^+_{r_{\mathrm{cp}}})\geq\frac{hr_{\mathrm{cp}}}{128}\right\} \ ; \tag{3.125}$$

$$\overline{D}^*_{WM\text{-}SR\text{-}3}(\mathcal{T},P,h,r,r_{\mathrm{cp}}) = \left\{k\in 1,\dots,d \ : \ \overline{m}_{k+r_{\mathrm{cp}}-1}(\mathcal{V}^-_{r_{\mathrm{cp}}}) - \overline{m}_{k-r_{\mathrm{cp}}-1}(\mathcal{V}^-_{r_{\mathrm{cp}}})\geq\frac{hr_{\mathrm{cp}}}{128}\right\} \ ; \tag{3.126}$$

$$\overline{D}^*_{WM\text{-}SR\text{-}4}(\mathcal{T},P,h,r,r_{\mathrm{cp}}) = \left\{k\in 1,\dots,d \ : \ \overline{m}_{k+2r_{\mathrm{cp}}}(\mathcal{V}_{2r_{\mathrm{cp}}}) - \overline{m}_{k-2r_{\mathrm{cp}}-1}(\mathcal{V}_{2r_{\mathrm{cp}}})\geq\frac{hr_{\mathrm{cp}}}{128}\right\} \ . \tag{3.127}$$

Then, we define the corresponding subsets $\overline{Q}^*_{WM\text{-}SR\text{-}1}$, $\overline{Q}^*_{WM\text{-}SR\text{-}2}$, $\overline{Q}^*_{WM\text{-}SR\text{-}3}$, and $\overline{Q}^*_{WM\text{-}SR\text{-}4}$ of $\mathcal{Q}_r$. For short, we write $\overline{Q}^*_{WM\text{-}SR}=\overline{Q}^*_{WM\text{-}SR\text{-}1}\cup\overline{Q}^*_{WM\text{-}SR\text{-}2}\cup\overline{Q}^*_{WM\text{-}SR\text{-}3}\cup\overline{Q}^*_{WM\text{-}SR\text{-}4}$. We have the following counterpart of Lemma 3.5.27.

Lemma 3.5.35. *Consider any valid hierarchical sorting tree $\mathcal{T}$, any subset P of a leaf G of $\mathcal{T}$, any $h\in\mathcal{H}$, and any $r\in\mathcal{R}$. With probability at least $1-\delta$, it holds that*

$$Q^*_{WM}\subset\widehat{Q}_{WM}\subset\overline{Q}^*_{WM\text{-}SR} \ . \tag{3.128}$$

Obviously, Lemma 3.5.28 is still true since it does not depend on the data generating process. Then, we adapt Propositions 3.5.11 and 3.5.12 to this adversarial setting.

Proposition 3.5.36. *Consider any $\overline{P}\subset[n]$, any $r\in\mathcal{R}$, and any subset $Q\subset Q_r$. Also, fix any $\eta>0$ and any $\phi>0$. Provided that*

$$\|\left[\Theta(\overline{P},Q)-\overline{\Theta}(\overline{P},Q)\right]_\eta\|_F^2\geq\frac{1}{\phi}\|\Theta(\overline{P},Q)-\overline{\Theta}(\overline{P},Q)\|_F^2\geq 8\eta|\overline{P}|\left[\phi_{l_1}\sqrt{\log(\tfrac{2|\overline{P}|}{\delta})}\sqrt{|Q|}+20\right] \ ,$$

then, with probability higher than $1-\delta$, we have

$$\|\Theta(\overline{P}',Q)-\overline{\Theta}(\overline{P}',Q)\|_F^2\leq\left(1-\frac{1}{16\phi}\right)\|\Theta(\overline{P},Q)-\overline{\Theta}(\overline{P},Q)\|_F^2 \ .$$

Recall the definition (3.60) of ϕ_{l_1}. Henceforth, the matrix $\Theta(\widetilde{P}, Q)$ is said to be indistinguishable in l_1-norm if it satisfies

$$\max_{i,j\in\overline{P}} \|\Theta_{i,\cdot}(\widetilde{P}, Q) - \Theta_{j,\cdot}(\widetilde{P}, Q)\|_1 \le \phi_{l_1}\sqrt{|Q|\log\left(\tfrac{2|\overline{P}|}{\delta}\right)} + 20 \ . \tag{3.129}$$

Proposition 3.5.37. *Let $\overline{P} \subset [n]$ and $Q \subset [d]$. If $\Theta(\widetilde{P}, Q)$ is indistinguishable in l_1-norm and if*

$$\|\Theta(\widetilde{P}, Q) - \overline{\Theta}(\widetilde{P}, Q)\|_F^2 \ge 10^6 \log^3\left(\frac{6nd}{\delta\zeta_-}\right)\left[\zeta^2\left(\sqrt{|\widetilde{P}||Q|} + |\widetilde{P}|\right) + |\widetilde{P}|\right] \ , \tag{3.130}$$

then, with probability higher than $1-3\delta$, we have

$$\|\Theta(\overline{P}', Q) - \overline{\Theta}(\overline{P}', Q)\|_F^2 \le \left(1 - \frac{1}{200\log^2(nd/\zeta_-)}\right)\|\Theta(\overline{P}, Q) - \overline{\Theta}(\overline{P}, Q)\|_F^2 \ .$$

Equipped with these two propositions, we arrive at the counterpart of Propositions 3.5.14 and 3.5.29. Recall Definition (3.64) of the function Ψ by $\Psi(p, r, h, q) = \frac{hp\sqrt{rq}}{\zeta} \wedge \sqrt{pq} + p$.

Proposition 3.5.38. *Assume that $\mathcal{T}_t$ is a valid hierarchical sorting tree. Consider a leaf G of $\mathcal{T}$ of type **0** or **1** at depth t. With probability higher than $1-5\tau_\infty\delta$, there exists a subset $\overline{P}^\dagger$ such that $\overline{P} \subseteq \overline{P}^\dagger \subseteq G$ and the following property holds. For some $r^\dagger_{\rm cp} \ge r^\dagger \in \mathcal{R}$ and some $h^\dagger \in \mathcal{H}$, upon writing $Q^\dagger_{WM} = Q^*_{WM}$ and $\overline{Q}^\dagger_{WM-SR} = \overline{Q}^*_{WM-SR}$, we have simultaneously*

$$\left\|\left[\Theta(\overline{P}^\dagger, Q^\dagger_{WM-SR}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{WM})\right]_{\sqrt{r^\dagger}h^\dagger}\right\|_F^2 \le 2\cdot 10^6 \log^3\left(\frac{6nd}{\delta\zeta_-}\right)\left[\zeta^2\Psi(|\overline{P}^\dagger|, r^\dagger, h^\dagger, |\overline{Q}^\dagger_{WM-SR}|) + |\overline{P}^\dagger|\right] \ ; \tag{3.131}$$

$$\|M(\overline{P}^\dagger) - \overline{M}(\overline{P}^\dagger)\|_F^2 \le 16\zeta^2 + 96|\mathcal{R}||\mathcal{H}|\|[\Theta(\overline{P}^\dagger, Q^\dagger_{WM}) - \overline{\Theta}(\overline{P}^\dagger, Q^\dagger_{WM})]_{\sqrt{r^\dagger}h^\dagger}\|_F^2 \ . \tag{3.132}$$

The proof is analogous to that of Proposition 3.5.29, up to some numerical constants, and is omitted.

Then, we state the counterpart of Lemma 3.5.30 to control $|\overline{Q}^\dagger_{WM-SR}|$. In comparison to this lemma, we have an additional term $n/(prh)$.

Lemma 3.5.39. *Assume that $\mathcal{T}_t$ is a valid hierarchical sorting tree. For any $h \in \mathcal{H}$, $r \in R$, any integer p, any sequence $\overline{P}_v$ of subsets of G_v, it holds that*

$$\sum_{\overline{P}\in\mathcal{P}^*(p)} |\overline{Q}^*_{WM-SR}(\overline{P}, h, r)| \lesssim \log(d)\left[\left\{\frac{\sqrt{nd(r_0\vee r)}}{rh\sqrt{p}} \wedge \frac{d(r_0\vee r)}{r^2h} \wedge \frac{nd}{pr} \wedge \frac{n(r_0\vee r)}{prh}\right\} + \frac{n}{prh}\right] \ . \tag{3.133}$$

Then, we apply Proposition 3.5.38 to control the loss on an additional event of probability higher than $1 - 5\cdot 2^t\tau_\infty\delta$.

$$\begin{aligned}
&\sum_{\overline{P}\in\overline{\mathcal{L}}_t} \|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \\
&\le 16\zeta^2|\overline{\mathcal{L}}_t| + 96|\mathcal{R}||\mathcal{H}| \sum_{p,h,r}\sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)} \left\|\left[\Theta(\overline{P}^\dagger, \overline{Q}^\dagger_{WM}) - \overline{\Theta}(\overline{P}^\dagger, \overline{Q}^\dagger_{WM})\right]_{\sqrt{r}h}\right\|_F^2 \\
&\lesssim \log^5\left(\frac{6nd}{\delta\zeta_-}\right)\sum_{p,h,r}\sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)}\left[\zeta^2\sqrt{[\tfrac{h^2pr}{\zeta^2}\wedge 1]p|\overline{Q}^*_{WM}(\overline{P}^\dagger, h, r)|} + (\zeta^2\vee 1)p\right] \\
&\lesssim \log^{5.5}\left(\frac{6nd}{\delta\zeta_-}\right)\sum_{p,h,r}\left[\zeta^2\sqrt{\frac{nr}{r\vee r_0}\sum_{\overline{P}^\dagger\in\mathcal{P}^*(p,h,r)}|\overline{Q}^*_{WM}(\overline{P}^\dagger, h, r)|} + (\zeta^2\vee 1)n\right] \\
&\lesssim \log^6\left(\frac{6nd}{\delta\zeta_-}\right)\sum_{p,h,r}\left[\zeta^2\left(n^{3/4}d^{1/4}\left(\frac{1}{(r_0\vee r)ph^2}\right)^{1/4} \wedge n\sqrt{\frac{d}{p(r_0\vee r)}} \wedge \frac{n}{\sqrt{ph}} \wedge \sqrt{\frac{nd}{rh}}\right) + \zeta^2 n\sqrt{\frac{1}{pr_0h}} + (\zeta^2\vee 1)n\right] \\
&\overset{(a)}{\lesssim} \log^7\left(\frac{6nd}{\delta\zeta_-}\right)\sum_{p,h}\left[\zeta^2\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{nh}{\zeta}\sqrt{d} \wedge \frac{n}{\sqrt{h}} \wedge \frac{\sqrt{nd}}{\sqrt{h}}\right) + (\zeta^2\vee 1)n\right] \\
&\lesssim \log^9\left(\frac{6nd}{\delta\zeta_-}\right)\left[\zeta^2\left(\frac{n^{3/4}d^{1/4}}{\zeta^{1/2}} \wedge n\sqrt{d} \wedge \frac{n^{2/3}\sqrt{d}}{\zeta^{1/3}} \wedge \frac{nd^{1/6}}{\zeta^{1/3}}\right) + (\zeta^2\vee 1)n\right] \ ,
\end{aligned}$$

where, in (a), we use that $pr_0h \ge pr_0h^2 \gtrsim 1$, the rest of the bounds being analogous to the proof of Proposition 3.5.34. This concludes the proof.

3.5.7.4 Proofs of the lemmas

Proof of Lemma 3.5.35. By a union bound and arguing as in the proof of Lemma 3.5.27, we deduce that, with probability higher than $1-\delta$, we have simultaneously

$$\max_{r_{\rm cp}\in[4r,\tilde{r}]\cap\mathcal{R}}\max_{k\in[d]}\left|\widehat{\mathbf{C}}^{*(\rm ext)}_{k,2r_{\rm cp}}-\mathbb{E}\left[\widehat{\mathbf{C}}^{*(\rm ext)}_{k,2r_{\rm cp}}\right]\right|\le h/32\ ; \tag{3.134}$$

$$\max_{r_{\rm cp}\in[4r,\tilde{r}]\cap\mathcal{R}}\max_{k\in[d]}\left|\widehat{\mathbf{\Delta}}^{*(\rm ext)}_{k,r_{\rm cp}}-\mathbb{E}\left[\widehat{\mathbf{\Delta}}^{(\rm ext)}_{k,r_{\rm cp}}\right]\right|\le h/32\ . \tag{3.135}$$

Because of the adversarial observations, we now have

$$\mathbf{C}^{*(\rm ext)}_{k,2r_{\rm cp}}\le\mathbb{E}\left[\widehat{\mathbf{C}}^{*(\rm ext)}_{k,2r_{\rm cp}}\right]\le\mathbf{C}^{*(\rm ext)}_{k,2r_{\rm cp}}+\frac{\overline{m}_{k+2r_{\rm cp}}(\mathcal{V})-\overline{m}_{k-2r_{\rm cp}-1}(\mathcal{V})}{2r_{\rm cp}}\ ,$$

$$\mathbf{\Delta}^{*(\rm ext)}_{k,r_{\rm cp}}\le\mathbb{E}\left[\widehat{\mathbf{\Delta}}^{*(\rm ext)}_{k,r_{\rm cp}}\right]\le\mathbf{\Delta}^{*(\rm ext)}_{k,r_{\rm cp}}+\frac{\overline{m}_{k+r_{\rm cp}}(\mathcal{V}^+)-\overline{m}_{k-r_{\rm cp}-1}(\mathcal{V}^-)}{2r_{\rm cp}}+\frac{\overline{m}_{k+r_{\rm cp}-1}(\mathcal{V}^+)-\overline{m}_{k-r_{\rm cp}}(\mathcal{V}^-)}{2r_{\rm cp}}\ .$$

Combining the above bounds with (3.134) and (3.135) allows us to conclude. □

Proof of Proposition 3.5.36. With the notation of the proof of Proposition 3.5.11, the condition (3.71) is now replaced by

$$\max_{i,j\in\overline{P}'}\|\Theta(\overline{P}')_{i,\cdot}-\Theta(\overline{P}')_{j,\cdot}\|_1\le\Phi_{l_1}\sqrt{|Q|}+20\ , \tag{3.136}$$

where we used Lemma 3.5.33 with $w=\mathbf{1}_Q$. The rest of the proof is left unchanged except that we replace $\Phi_{l_1}\sqrt{|Q|}$ by $\Phi_{l_1}\sqrt{|Q|}+20$. □

Proof of Proposition 3.5.37. Lemma 3.5.20 is still true. However, Lemma 3.5.21 needs to be updated to

Lemma 3.5.40. *Fix any $\delta\in(0,1)$. If*

$$\|\Theta-\overline{\Theta}\|^2_{\rm op}\ge 6400\left[|\widetilde{P}|+\zeta^2\left[\sqrt{|Q|(5|\widetilde{P}|+\log(6/\delta))}+7|\widetilde{P}|+2\log(6/\delta)\right]\right]\ , \tag{3.137}$$

then, with probability higher than $1-\delta$, we have

$$\|\hat{v}^T\left(\Theta-\overline{\Theta}\right)\|_2^2\ge\frac{1}{2}\|\Theta-\overline{\Theta}\|^2_{\rm op}\ .$$

In light of Condition (3.130), this assumption is valid. Together with Lemma 3.5.20, we deduce that there exists an event of probability higher than $1-\delta$ such that

$$\|\hat{v}^T\left(\Theta-\overline{\Theta}\right)\|_2^2\ge\frac{1}{2}\|\Theta-\overline{\Theta}\|^2_{\rm op}\ge\frac{1}{32\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2\ .$$

As the vectors $\hat{z}$ and $\hat{w}$ are defined though $Z^{(3)}$, we rather focus on $\Theta^{(3)}$. By (3.121), we have $\|\Theta^{(3)}-\Theta\|_{\rm op}\le\sqrt{|\widetilde{P}|}$.

$$\begin{aligned}\|\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})\|_2^2 &\ge \|\hat{v}^T(\Theta-\overline{\Theta})\|_2^2-4\|\Theta-\Theta^{(3)}\|_{\rm op}\|\Theta-\overline{\Theta}\|_{\rm op}\ge\|\hat{v}^T(\Theta-\overline{\Theta})\|_2^2-4\sqrt{|\widetilde{P}|}\|\Theta-\overline{\Theta}\|_{\rm op}\\ &\ge \frac{9}{20}\|\Theta-\overline{\Theta}\|^2_{\rm op}\ge\frac{1}{36\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2\ .\end{aligned} \tag{3.138}$$

Then, the analysis of $\hat{z}$ and z^* follows the same steps as in the original proofs, - see Section 3.5.4.3 - the main difference being that we invoke (3.136) instead of (3.71). More precisely, we still have

$$\left|\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2\ge\frac{16}{25}\|w^*\|_2^2\ . \tag{3.139}$$

and

$$\|w^*\|_2^2=\|z^*\|_2^2-\sum_{l\in S^{*c}}(z^*_l)^2\ . \tag{3.140}$$

The control of $\sum_{l\in S^{*c}}(z_l^*)^2$ is slightly different.

$$\begin{aligned}
\left[\sum_{l\in S^{*c}}(z_l^*)^2\right]^2 &= \left[\sum_{l\in S^{*c}}[\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})]_l z_l^*\right]^2 \\
&\le \|\left(\Theta^{(3)}-\overline{\Theta}^{(3)}\right)z^*_{S^{*c}}\|_2^2 = \sum_{i\in\widetilde{P}}\left(\sum_{l\in S^{*c}}(\Theta^{(3)}_{i,l}-\overline{\theta}^{(3)}_l)z_l^*\right)^2 \\
&\le \frac{18\zeta^2}{|\widetilde{P}|^2}\log\left(\frac{2|Q|}{\delta}\right)\sum_{i\in\widetilde{P}}\left(\sum_{l\in S^{*c}}\sum_{j\in\widetilde{P}}|\Theta^{(3)}_{i,l}-\Theta^{(3)}_{j,l}|\right)^2 \\
&\le \frac{18\zeta^2}{|\widetilde{P}|^2}\log\left(\frac{2|Q|}{\delta}\right)\sum_{i\in\widetilde{P}}\left(\sum_{j\in\widetilde{P}}\|\Theta^{(3)}_{i,\cdot}-\Theta^{(3)}_{j,\cdot}\|_1\right)^2 \\
&\le 18\zeta^2\log\left(\frac{2|Q|}{\delta}\right)|\widetilde{P}|\left[\phi_{l_1}\log^{1/2}\left(\frac{2|\widetilde{P}|}{\delta}\right)\sqrt{Q}+22\right]^2 \\
&\le \left[250\zeta^2\log\left(\frac{2|Q||\widetilde{P}|}{\delta}\right)(\sqrt{|\widetilde{P}||Q|}+1)+400|\widetilde{P}|\right]^2 ,
\end{aligned}$$

where we used (3.136) as well as the fact $\|\Theta^{(3)}_{i,\cdot}-\Theta_{i,\cdot}\|_1\le 1$. Recall that $z^*=\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})$. Combining Section 3.5.7.4, (3.140), and Condition (3.130), we deduce that

$$\|w^*\|_2^2 \ge \frac{1}{72\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 ,$$

which, together with (3.139), yields

$$\left\|(\Theta^{(3)}-\overline{\Theta}^{(3)})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \left|\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \ge \frac{1}{120\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 .$$

Then, we come back to the matrix $\Theta-\overline{\Theta}$ using again (3.121).

$$\left\|(\Theta-\overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \frac{9}{10}\left|\hat{v}^T(\Theta^{(3)}-\overline{\Theta}^{(3)})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 - 9|\widetilde{P}| .$$

Then, we apply Harris' inequality as in the original proof of the lemma to conclude that

$$\left\|(\Theta-\overline{\Theta})\frac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \ge \left\|(\Theta-\overline{\Theta})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \frac{9}{1200\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 - 9p \ge \frac{1}{150\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 . \quad (3.141)$$

Applying the pivot algorithm to $\hat{w}^+$, we deduce from Lemma 3.5.33 that there exists an event of probability higher than $1-\delta$ such that

$$\max_{i,j\in\overline{P}'}\left|\langle\Theta(\overline{P}')_{i,\cdot}-\Theta(\overline{P}')_{j,\cdot},\frac{\hat{w}^+}{\|\hat{w}^+\|_2}\rangle\right| \le \phi_{l_1}\sqrt{\log\left(\tfrac{2|\overline{P}|}{\delta}\right)}+20 .$$

By convexity, it follows that

$$\left\|[\Theta(\overline{P}')-\overline{\Theta}(\overline{P}')]\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \le 2\phi_{l_1}^2\log\left(\tfrac{2|\overline{P}|}{\delta}\right)|\overline{P}'|+800|\overline{P}'| \le |\widetilde{P}|\left[2\phi_{l_1}^2\log\left(\tfrac{2|\overline{P}|}{\delta}\right)+800\right] .$$

In light of Condition (3.62), this quantity is small compared to $\|\Theta-\overline{\Theta}\|_F^2$.

$$\|(\Theta(\overline{P}')-\overline{\Theta}(\overline{P}'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \le \frac{1}{200\log^2(nd/\zeta_-)}\|\Theta-\overline{\Theta}\|_F^2 . \quad (3.142)$$

Then, we conclude from (3.142) as we did from (3.83) in the original proof.

□

Proof of Lemma 3.5.40. For short, we write $p=|\widetilde{P}|$. Since $Z^{(t)}=\Theta^{(t)}+N^{(t)}$ for $t=1,2$, the difference wih Lemma 3.5.21 is that $\Theta^{(1)}$ and $\Theta^{(2)}$ are involved in the terms $v^T(Z^{(1)}-\overline{Z}^{(1)})$ and $v^T(Z^{(2)}-\overline{Z}^{(2)})$. Hence,

arguing as in the proof of Lemma 3.5.21, we derive that, on an event of probability higher than $1-3\delta$, we have simultaneously for all $v \in \mathbb{R}^p$ with $\|v\|_2 \le 1$ that

$$\begin{aligned}
&\Big| \|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta^{(1)} - \overline{\Theta}^{(1)})\|_2^2 + \\
&\frac{1}{2}\|v^T(\Theta^{(1)} - \overline{\Theta}^{(1)} - \Theta^{(2)} + \overline{\Theta}^{(2)})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2 \Big| \\
&\le 10\zeta \left[\|\Theta^{(1)} - \overline{\Theta}^{(1)}\|_{\mathrm{op}} + \frac{1}{2}\|\Theta^{(1)} - \overline{\Theta}^{(1)} - \Theta^{(2)} + \overline{\Theta}^{(2)}\|_{\mathrm{op}} \right] \sqrt{2p + \log(6/\delta)} \\
&\quad + 192\zeta^2 \left[\sqrt{q(3p + \log(6/\delta))} + (3p + \log(6/\delta)) \right] .
\end{aligned}$$

By (3.121), we have $\|\Theta^{(t)} - \Theta\|_{\mathrm{op}} \le \sqrt{p}$ for $t = 1, 2$. Hence, the above bound simplifies in

$$\begin{aligned}
&\Big| \|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta - \overline{\Theta})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2 \Big| \\
&\le 10\zeta \left[\|\Theta - \overline{\Theta}\|_{\mathrm{op}} + 4\sqrt{p} \right] \sqrt{2p + \log(6/\delta)} + 192\zeta^2 \left[\sqrt{q(3p + \log(6/\delta))} + (3p + \log(6/\delta)) \right] \\
&\quad + 12p + 4\sqrt{p}\|\Theta - \overline{\Theta}\|_{\mathrm{op}} .
\end{aligned}$$

Since we assume that $\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 \ge 6400\Big[p + \zeta^2[\sqrt{q(5p + \log(6/\delta))} + 7p + 2\log(6/\delta)]\Big]$, we deduce that, on the same event, we have

$$\sup_{v \in \mathbb{R}^p:\ \|v\|_2 \le 1} \Big| \|v^T(Z^{(1)} - \overline{Z}^{(1)})\|_2^2 - \|v^T(\Theta - \overline{\Theta})\|_2^2 - \frac{1}{2}\|v^T(Z^{(1)} - \overline{Z}^{(1)} - Z^{(2)} + \overline{Z}^{(2)})\|_2^2 \Big| \le \frac{1}{4}\|\Theta - \overline{\Theta}\|_{\mathrm{op}}^2 .$$

The rest of the proof is left unchanged.

□

Proof of Lemma 3.5.39. Recall that $\overline{Q}^*_{WM-SR}(\overline{P}, h, r)$ decomposes as the union of $\overline{Q}^*_{WM-SR-1}$, $\overline{Q}^*_{WM-SR-2}$, and $\overline{Q}^*_{WM-SR-3}$, $\overline{Q}^*_{WM-SR-4}$. Since $\overline{Q}^*_{WM-SR-1}$ is defined analogously to $\overline{Q}^*_{WM}$ –but with a different numerical constant–, we can argue as in the proof of Lemma 3.5.39, which yields

$$\sum_{\overline{P} \in \mathcal{P}^*(p)} |\overline{Q}^*_{WM-SR-1}(\overline{P}, h, r)| \lesssim \log(d) \left[\frac{\sqrt{nd(r_0 \vee r)}}{rh\sqrt{p}} \wedge \frac{d(r_0 \vee r)}{r^2 h} \wedge \frac{nd}{pr} \wedge \frac{n(r_0 \vee r)}{prh} \right] .$$

It remains to consider the three last sets. We only focus on $\overline{Q}^*_{WM-SR-2}(\overline{P}, h, r)$, the last ones being analogous. We first focus on a single set $\overline{Q}^*_{WM-SR-2}(\overline{P}, h, r, r_{\mathrm{cp}})$. If k belongs to $\overline{D}^*_{WM-SR-2}(\overline{P}, h, r, r_{\mathrm{cp}})$, this implies that the total variation of $\overline{m}(\mathcal{V}^+_{r_{\mathrm{cp}}})$ between $k - r_{\mathrm{cp}}$ and $k + r_{\mathrm{cp}}$ is at least $hr_{\mathrm{cp}}/128$. Since the total variation of $\overline{m}(\mathcal{V}^+_{r_{\mathrm{cp}}})$ is at most one, there are at most $c/(hr_{\mathrm{cp}})$ regions of $\mathcal{Q}_{r_{\mathrm{cp}}}$ that contain at least a point $\overline{D}^*_{WM-SR-2}(\overline{P}, h, r, r_{\mathrm{cp}})$, which entails that there are at most $c' \frac{r_{\mathrm{cp}}}{r} \cdot \frac{1}{hr_{\mathrm{cp}}} = c'/(hr)$ regions of $\mathcal{Q}_r$ that contain at least a point $\overline{D}^*_{WM-SR-2}(\overline{P}, h, r, r_{\mathrm{cp}})$. Since r_{cp} takes at most a logarithmic number of values and since $|\mathcal{P}^*(p)| \le n/p$, we obtain

$$\sum_{r_{\mathrm{cp}}} \sum_{\overline{P} \in \mathcal{P}^*(p)} |\overline{Q}^*_{WM-SR-2}(\overline{P}, h, r, r_{\mathrm{cp}})| \lesssim \log(d) \frac{n}{prh} ,$$

which concludes the proof. □

3.5.8 Proofs for the l_∞ loss

Proof of Lemma 3.4.5. Without loss of generality, we assume that π^* is the identity. Fix any $i \in [n]$ and assume that $\hat{\pi}^{-1}(i) \neq i$. Consider for instance the case where $m = \hat{\pi}^{-1}(i) > i$. As a consequence, there are at least $l = m - i$ experts that are below m in the oracle order and above m in the estimated order . Denote j the smallest of those experts. Hence, we have $j \le i < m$ and $\hat{\pi}(j) \ge \hat{\pi}(m)$. Besides, since $j \le i \le m$, we deduce from the bi-isotonic assumption that

$$\|M_{i,.} - M_{\hat{\pi}^{-1}(i),.}\|_2^2 \le \|M_{j,.} - M_{\hat{\pi}^{-1}(i),.}\|_2^2 = \|M_{j,.} - M_{m,.}\|_2^2 .$$

Taking the supremum over all i implies that $l_\infty(\hat{\pi}, \pi^*) \le l_{err}(\hat{\pi}, \pi^*)$. Let us turn to the second inequality. Consider any $i < j$ such that $\hat{\pi}(i) > \hat{\pi}(j)$. We consider three subcases.

(i) If $\hat{\pi}(i) \ge j$, then we have $\|M_{i,.} - M_{j,.}\|_2^2 \le \|M_{i,.} - M_{\hat{\pi}(i),.}\|_2^2$.

(ii) If $\hat{\pi}(j) \leq i$, then $\|M_{i,\cdot} - M_{j,\cdot}\|_2^2 \leq \|M_{\hat{\pi}(j),\cdot} - M_{j,\cdot}\|_2^2$.

(iii) It remains to consider the case where we have $i < \hat{\pi}(j) < \hat{\pi}(i) < j$. As a consequence, for each $k \in [d]$, we have $M_{j,k} - M_{i,k} \leq M_{j,k} - M_{\hat{\pi}(j),k} + M_{\hat{\pi}(i),k} - M_{i,k}$, which in turn implies that

$$\|M_{j,\cdot} - M_{i,\cdot}\|_2^2 \leq 4 l_\infty(\hat{\pi}, \pi^*) \ .$$

Taking the supremum over all i and reminding the definition of j concludes the proof. □

Proof of Proposition 3.4.6. For $n = 2$, all the losses are equal. Hence, the minimax lower bound (3.34) is a straightforward consequence of the general minimax lower bound of Theorem 3.4.1 by a reduction to the case where $n = 2$ (recall that $\zeta = 1$ here) - This reduction is achieved by putting to 0 the signal corresponding to all $n - 2$ experts that do not corresponds to the 2 experts of interest that will be most difficult to distinguish so that estimating the permutation amounts to deciphering between these two experts. Hence, we derive that

$$\inf_{\hat{\pi}} \sup_{\pi^* \in \Pi_n} \sup_{M : M_{\pi^*-1} \in \mathbb{C}_{\mathrm{BISO}}} \mathbb{E}_{(\pi^*, M)}[l_\infty(\hat{\pi}, \pi^*)] \geq c'' \left[\left(\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} + \frac{1}{\lambda} \right) \bigwedge d \right] .$$

It turns out that the term $1/\lambda$ is higher than d if $\lambda \leq 1/d$ and is smaller than $\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda}$ for larger λ's. Hence, we can conclude that

$$\inf_{\hat{\pi}} \sup_{\pi^* \in \Pi_n} \sup_{M : M_{\pi^*-1} \in \mathbb{C}_{\mathrm{BISO}}} \mathbb{E}_{(\pi^*, M)}[l_\infty(\hat{\pi}, \pi^*)] \geq c'' \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} \bigwedge d \right] .$$

Regarding the upper bound, we build upon the analysis of $\hat{\pi}_{WMP}$ in the specific case of $n = 2$. Consider any fixed i and j. With probability higher than $1 - c\delta \log^{c'}[nd(\lambda \vee 1)]$, it follows from the proof of Theorems 3.2.3 and 3.4.3 that (i) $\hat{\pi}_{WMP}$ builds a valid hierarchical sorting tree and (ii) the set $\overline{P} \subset \{i, j\}$ built at the end of **BlockSort** satisfies

$$\|M(\overline{P}) - \overline{M}(\overline{P})\|_F^2 \leq c_1 \log^{c_2}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} + \frac{1}{\lambda} \right] . \tag{3.143}$$

It follows from (i) that (i, j) (resp. (j, i)) is added to $\mathcal{PC}$ only if $\pi^*(i) < \pi^*(j)$ (resp. $\pi^*(i) < \pi^*(j)$). Besides, if

$$\|M_{i,\cdot} - M_{j,\cdot}\|_2 > 2c_1 \log^{c_2}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} + \frac{1}{\lambda} \right] , \tag{3.144}$$

then, this implies that $|\overline{P}| \leq 1$, otherwise this would contradict Equation (3.143).

Then, taking a union bound over all possible (i, j), we deduce that there exists an event of probability higher than $1 - cn^2\delta \log^{c'}(nd(\lambda \vee 1))$, such that $\mathcal{PC}$ is consistent and contains all 2-tuples of experts that satisfy (3.144).

Turning to the estimated permutation $\hat{\pi}_{PC}$, we consider any two experts such that $\pi^*(i) < \pi^*(j)$ and $\hat{\pi}_{PC}(i) > \hat{\pi}_{PC}(j)$. The latter condition implies that $\phi(i) \geq \phi(j)$. Since $\mathcal{PC}$ is consistent, we have $\pi^*(i) \geq 1 + \phi(i)$. Define $\pi^*_-(j)$ as the number of experts k that are below j and are far apart from j in the sense of Equation (3.144). We know that, under the above event, we have that $\phi(j) \geq \pi^*_-(j)$. This implies that $\pi^*(i) > \pi^*_-(j)$. As a consequence, i and j are not far apart in the sense of Equation (3.144). This implies that

$$l_{err}(\hat{\pi}_{PC}, \pi^*) \leq 2c_1 \log^{c_2}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} + \frac{1}{\lambda} \right] .$$

Since $l_{err}(\hat{\pi}_{PC}, \pi^*)$ is equivalent to $l_\infty(\hat{\pi}_{PC}, \pi^*)$, this bound also holds (with a larger constant) for the latter loss. Since δ has been chosen small enough and since the loss is always smaller than d, we arrive at the following risk bound

$$\mathbb{E}\left[l_\infty(\hat{\pi}_{PC}, \pi^*)\right] \leq c_1' \log^{c_2'}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} + \frac{1}{\lambda} \right] ,$$

which, in turn, implies that

$$\mathbb{E}[l_\infty(\hat{\pi}_{PC}, \pi^*)] \leq c \log^{c'}(nd(\lambda \vee 1)) \left[\frac{d^{1/6}}{\lambda^{5/6}} \bigwedge \frac{\sqrt{d}}{\lambda} \right] \bigwedge d \ .$$

□

3.5.9 Proof of the minimax lower bounds

3.5.9.1 Proof of Theorem 3.4.1

3.5.9.2 Noiseless minimax lower bound

Here, we shall prove the following minimax lower bound holding in the noiseless case $\zeta = 0$.

$$\mathcal{R}^*[n,d,\lambda,0] \geq c\left[\frac{n}{\lambda}e^{-2\lambda} \wedge nd\right] \tag{3.145}$$

Obviously, the bound remains valid for general $\zeta \geq 0$. Define the positive integer $d_- = 1 \vee \lfloor 1/\lambda \rfloor \wedge d] \leq d$. We build a prior distribution ν of M as follows. For each row $i = 1,\dots,n$, we sample $W_i \sim \mathcal{B}(1/2)$. If $\zeta_i = 1$, the i-th row of M is constant and equal to 1. if $W_i = 0$, then the i-th row of M has its d_- first entries equal to 0, while the remaining entries are equal to 1.

We write $\mathbf{P}$ and $\mathbf{E}$ for the corresponding marginal probability and expectations of the data (x_t, y_t).

$$\mathcal{R}^*[n,d,\lambda,0] \geq \inf_{\hat{\pi}} \mathbf{E}\left[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_2^2\right] .$$

For each entry $i = 1,\dots,n$, we write $N_i = \sum_t \mathbf{1}_{x_t \in \{i\}\times[d_-]}$ the number of observations on the d_- first columns of the i-th row. If $N_i \geq 1$, then the statistician knows the value of W_i. Conversely, if $N_i = 0$, then she has no information on the value of W_i. Given an estimator $\hat{\pi}$, it is always possible to reduce its loss by ranking at the top the experts such that $N_i \geq 1$ and $W_i = 1$, ranking below the experts such that $N_i \geq 1$ and $W_i = 0$, and putting in between the experts such that $N_i = 0$. Conditionally to the observations (x_t, y_t), the values of W_i such that $N_i = 0$ are still distributed according to a Bernoulli distribution. As a consequence, for any $\hat{\pi}$ which has been rearranged as explained above, the conditional risk satisfies

$$\mathbf{E}\left[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_2^2 \middle| (x_t, y_t)\right] \geq d_- \times g(\sum_{i=1}^{n} \mathbf{1}_{N_i=0}) ,$$

where $g(k)$ corresponds to the expected number of error of $\hat{\pi}$ when there are exactly k rows without any observations. Since conditionally to $\hat{\pi}$, the corresponding values of W have been sampled independently as Bernoulli random variables with parameter $1/2$, we arrive at the following expression for $g(k)$:

$$g(k) = \sum_{i=1}^{k} \mathbb{P}[\{W_i = 1\} \cap \{\sum_{j=1}^{k} W_j \leq k - i\}] + \mathbb{P}[\{W_i = 0\} \cap \{\sum_{j=1}^{k} W_j > k - i\}] .$$

We have $g(1) = 0$, $g(2) = 1/2$, $g(3) = 1$. For $k \geq 4$, we focus on the $\lfloor k/4 \rfloor$ first and $\lfloor k/4 \rfloor$ last entries to deduce that

$$\begin{aligned} g(k) &\geq \mathbb{E}\left[\sum_{i=1}^{\lfloor k/4 \rfloor} W_i\right] P\left[\sum_{i=\lfloor k/4 \rfloor+1}^{k} W_i \leq k/2\right] + \mathbb{E}\left[\sum_{i=k-\lfloor k/4 \rfloor+1}^{k} (1 - W_i)\right] P\left[\sum_{i=1}^{k-\lfloor k/4 \rfloor} (1 - W_i) \leq k/2\right] \\ &\geq 0.5\lfloor \frac{k}{4} \rfloor . \end{aligned}$$

Hence, there exists a universal constant $c > 0$ such that we have $g(k) \geq c(k-1)$ for any $k \geq 1$. Since N_i follows a Poisson distribution with parameter λd_-, $V = \sum_{i=1}^{n} \mathbf{1}_{N_i=0}$ follows a binomial distribution with parameters $(e^{-\lambda d_-}, n)$. We obtain $\mathcal{R}^*[n,d,\lambda,0] \geq c d_- \mathbb{E}[(V-1)_+]$. If $\mathbb{E}[V] \geq 2$, then we simply use $\mathbb{E}[(V-1)_+] \geq \mathbb{E}[V]/2$. If $\mathbb{E}[V] < 2$, we use $\mathbb{E}[(V-1)_+] \geq \mathbb{P}[V = 2] = \frac{n(n-1)}{2} e^{-2\lambda d_-}(1 - e^{-2\lambda d_-})^{n-2} \geq c' n^2 e^{-2\lambda d_-}$. In any case, we conclude that

$$\mathcal{R}^*[n,d,\lambda,0] \geq c'' d_- n e^{-2\lambda d_-} .$$

If $\lambda \leq 1/d$, then $d_- = d$, and the right hand-side is higher than $c'' n d e^{-2}$. If $\lambda \in [1/d, 1]$, then we have $d_- \in [1/(2\lambda), 1/\lambda]$ and the right hand-side risk is higher than cn/λ. Finally, if $\lambda \geq 1$, we take $d_- = 1$ and the right hand-side is higher than $c' n e^{-2\lambda}$. We have proved Equation (3.145).

3.5.9.3 Proof of the remaining regimes

Since the minimax risk is increasing with n and d, we can assume without loss of generality that both n and d express as a power of 2.

We shall first build a collection of prior distributions $\nu_{\mathbf{G}}$ indexed by $\mathbf{G} \in \mathcal{G}$ on M. We denote $\mathbf{P}_{\mathbf{G}}^{(\mathbf{full})}$ and $\mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}$ the corresponding marginal probability distributions and expectations on the data (x_t, y_t). Since we aim

at proving the lower bound in the Gaussian setting, we assume that the data y_t is a normal random variable with mean M_{x_t} and variance ζ^2 conditionally on M and x_t. The minimax risk (3.4) is higher than the worst Bayesian risk.

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq \inf_{\hat{\pi}} \sup_{\mathbf{G}\in\mathcal{G}} \mathbf{E}_{\mathbf{G}}^{\mathbf{full}}\left[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2\right] . \tag{3.146}$$

We first spend some time defining the corresponding prior distributions before applying a sequence of reduction arguments.

3.5.9.4 Construction of the Prior distribution on M

Let $\tilde{n} \in [n]$ be an a power of 2 so that $n/\tilde{n}$ is an integer. From a broad perspective, the general purpose of this prior construction is to break down the permutation estimation problem into $n/\tilde{n}$ independent bisection problems of size $\tilde{n}$. We will fix the value of $\tilde{n}$ at the end of the proof. The permuted matrix $M_{\pi^{*-1}}$ will turn out to be block constant and we introduce $\tilde{d} \in [d]$ the number of blocks of questions, each of them being of size $d/\tilde{d}$. Here we assume that $\tilde{d}$ is a power of 2 so that $d/\tilde{d}$ is an integer. $\tilde{d}$ will be also fixed at the end of the proof.

We introduce the staircase matrix C of dimension $(n/\tilde{n}) \times \tilde{d}$ such that $C_{\iota,\kappa} = \iota\tilde{n}/(4n) + \kappa/(4\tilde{d})$. Also write U for the constant $\tilde{n} \times d/\tilde{d}$ matrix whose entries are all equal to one. With this notation, the Kronecker product matrix $C \otimes U$ of size $n \times d$ is a bi-isotonic staircase matrix with blocks of size $\tilde{n} \times (d/\tilde{d})$.

Then, we shall perturb the matrix $C \otimes U$ in order to simultaneously craft $n/\tilde{n}$ independent clustering problems of size $\tilde{n}$ each. Set $\tilde{\lambda} = \lambda\frac{d}{\tilde{d}}$ and $\lambda_0 = \tilde{n}\tilde{\lambda}$. Let υ be a positive number and let also q be an integer smaller than or equal to $\tilde{d}$ and

$$M = C \otimes U + \upsilon \frac{\zeta}{\sqrt{\lambda_0}} B^{(\mathbf{full})} , \tag{3.147}$$

where the random matrix $B^{(\mathbf{full})} \in \{0,1\}^{n\times d}$ is defined below.

For this purpose, we consider a collection $\mathcal{G}$ of subsets of $[\tilde{n}]$ with size $\tilde{n}/2$ that are well-separated in symmetric difference as defined by the following lemma.

Lemma 3.5.41. *There exists a numerical constant c_0 such that the following holds for any even integer $\tilde{n}$. There exists a collection $\mathcal{G}$ of subsets of $[\tilde{n}]$ with size $\tilde{n}/2$ whose satisfies $\log(|\mathcal{G}|) \geq c_0|\tilde{n}|$ and whose elements are $\tilde{n}/4$-separated, that is $|G_1 \Delta G_2| \geq \tilde{n}/4$ for any $G_1 \neq G_2$.*

The above result is a straightforward consequence of Varshamov-Gilbert's lemma – see e.g. [90].

For each block $\iota \in [n/\tilde{n}]$, we fix a subset $G^{(\iota)}$ from $\mathcal{G}$. Then, we consider its 'translation' $G^{t(\iota)} = \{x + (\iota-1)\tilde{n} : x \in G^{\iota}\}$. The experts of $G^{t(\iota)}$ will correspond to the subgroup of 'higher' experts in the group ι. We write $\mathbf{G} = (G^{t(1)}, \dots, G^{t(n/\tilde{n})})$ and $\boldsymbol{\mathcal{G}}$ the corresponding collection of all possible $\mathbf{G}$. Given any such $\mathbf{G}$, we shall define a prior distribution $\nu_{\mathbf{G}}$ on M.

For $\iota \in [n/\tilde{n}]$, we sample uniformly a subset $Q^{(\iota)}$ of q block of questions among the $\tilde{d}$ blocks. In each of these q blocks, the corresponding rows of $B^{(\mathbf{full})}$ are equal to one. More formally, upon writing $\mathbf{1}_{d/\tilde{d}}$ for the constant vector of size $d/\tilde{d}$, we have

$$B^{(\mathbf{full})} = \sum_{\iota=1}^{n/\tilde{n}} \mathbf{1}_{G^{t(\iota)}} (Q^{(\iota)} \otimes \mathbf{1}_{d/\tilde{d}})^T . \tag{3.148}$$

To sum up, we define a prior distribution $\nu_{\mathbf{G}}$ on $B^{(\mathbf{full})}$ (and equivalently on M) such that, under $\nu_{\mathbf{G}}$, all the rows of $B^{(\mathbf{full})}$ that do not belong to any $G^{t(\iota)}$ are zero. All the rows belonging to the same set $G^{t(\iota)}$ are equal and block constants with $\tilde{d}$ blocks of size $d/\tilde{d}$, among which q blocks are exactly equal to one.

Coming back to the matrix M defined in (3.148), we see that as soon as

$$2\upsilon\zeta/\sqrt{\lambda_0} \leq \tilde{n}/(4n) \wedge 1/(4\tilde{d}) , \tag{3.149}$$

then, almost surely, the matrix M, is up to a (non-unique) permutation, bi-isotonic and its coefficients are in $[0,1]$. Defining the subset $\overline{G}^{(\iota)} = \{i + (\iota-1)\tilde{n} : i \in [\tilde{n}]\}$, we see that, under $\nu_{\mathbf{G}}$, recovering a suitable permutation π^* is exactly equivalent to estimating the subgroup $G^{t(\iota)} \subset \overline{G}^{(\iota)}$ for each $\iota = 1, \dots, n/\tilde{n}$. This construction of M is illustrated in Figure 3.6. To sum up, the prior distribution distribution $\nu_{\mathbf{G}}$ on M requires the choice of the parameters $\tilde{n} \in [n]$, $\tilde{d} \in [d]$, the sparsity $q \in [\tilde{d}]$, and some signal level $\upsilon > 0$ satisfying (3.149).

As we shall use several reduction arguments, we need to introduce some new notation. First, we respectively denote $\mathbf{P}_{\mathbf{G}}^{(\mathbf{full})}$ and $\mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}$ for the marginal probability and expectation with respect to the data when M is sampled according to $\nu_{\mathbf{G}}$.

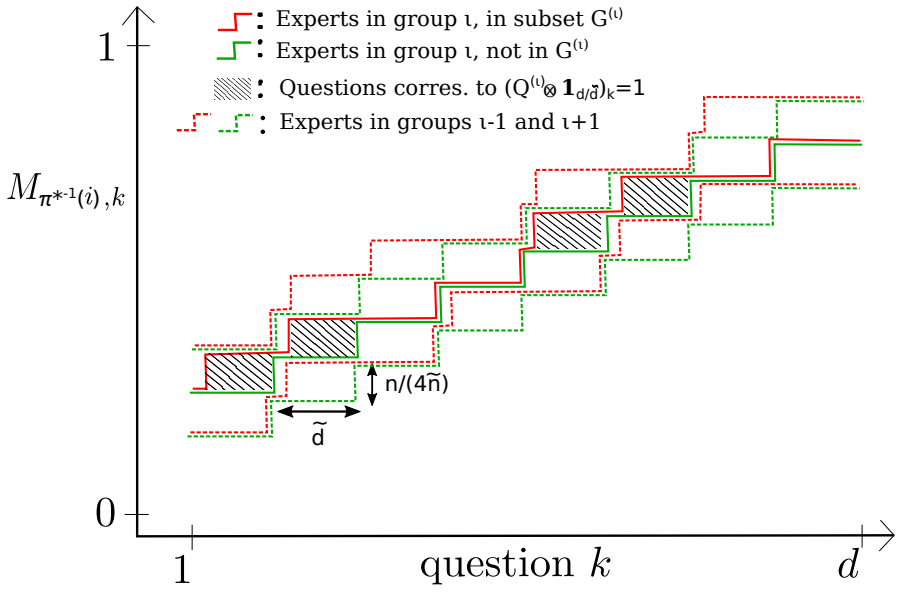


Figure 3.6: Example of a matrix M sampled from $\nu_{\mathbf{G}}$.

The distribution of the rows $\overline{G}^{t(\iota)}$ in M under $\nu_{\mathbf{G}}$ only depends on $G^{t(\iota)}$. In what follows, we write $\nu_{G^{t(\iota)}}$ for this distribution. Similarly, we write $\mathbf{P}^{(\mathbf{full})}_{G^{t(\iota)}}$ for the corresponding marginal distribution of the observations (x_t, y_t) such that $(x_t)_1 \in \overline{G}^{t(\iota)}$. By the poissonization trick, the distribution $\mathbf{P}^{(\mathbf{full})}_{\mathbf{G}}$ is a product measure of $\mathbf{P}^{(\mathbf{full})}_{G^{t(\iota)}}$ for $\iota = 1, \dots, n/\tilde{n}$. We write $\mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}}$ for the corresponding expectation.

3.5.9.4.1 Step 2: Problem Reduction We start with prior distributions $\nu_{\mathbf{G}}$.

$$\boldsymbol{\mathcal{R}}^*[n,d,\lambda,\zeta] \geq \inf_{\hat{\pi}} \sup_{\mathbf{G}\in\boldsymbol{\mathcal{G}}} \mathbb{E}^{(\mathbf{full})}_{\mathbf{G}} \|M_{\pi^{*-1}} - M_{\hat{\pi}^{-1}}\|_2^2$$

For each of these matrices M sampled from a distribution $\nu_{\mathbf{G}}$, it turns out that $\pi^*(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$. Hence, to estimate π^*, we only need to estimate each $G^{t(\iota)} \subset \overline{G}^{(\iota)}$ from the data. Intuitively, we therefore can restrict ourselves to estimators $\hat{\pi}$ satisfying $\hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$. More precisely, if an estimator $\tilde{\pi}$ does not satisfy this condition, then we can modify $\tilde{\pi}$ in $\hat{\pi}$ in order to enforce the $\overline{G}^{(\iota)}$'s to be stable. Since, by Condition (3.149) experts in different $\overline{G}^{(\iota)}$ are far from each other, it turns out that the loss of $\hat{\pi}$ is smaller than that of $\tilde{\pi}$.

$$\begin{aligned}
\mathcal{R}^*[n,d,\lambda,\zeta] &\geq \inf_{\hat{\pi}:\ \hat{\pi}(\overline{G}^{(\iota)})=\overline{G}^{(\iota)}} \sup_{\mathbf{G}\in\boldsymbol{\mathcal{G}}} \sum_{\iota=1}^{n/\tilde{n}} \mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}} \left[\left\| \left(M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}} \right\|_F^2 \right] \\
&\geq \inf_{\hat{\pi}:\ \hat{\pi}(\overline{G}^{(\iota)})=\overline{G}^{(\iota)}} \sum_{\iota=1}^{n/\tilde{n}} \sup_{G^{t(\iota)}} \mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}} \left[\left\| \left(M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}} \right\|_F^2 \right] \\
&\geq \sum_{\iota=1}^{n/\tilde{n}} \inf_{\hat{\pi}^{(\iota)}} \sup_{G^{t(\iota)}} \mathbf{E}^{(\mathbf{full})}_{G^{t(\iota)}} \left[\left\| \left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}} \right\|_F^2 \right] ,
\end{aligned}$$

where, in the last line, $\hat{\pi}^{(\iota)}$ stands for any estimator of the restriction π^* to $\overline{G}^{(\iota)}$. By symmetry, we arrive at

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq \frac{n}{\tilde{n}} \inf_{\hat{\pi}^{(1)}} \sup_{G^{t(1)}} \mathbf{E}^{(\mathbf{full})}_{G^{t(1)}} \left[\left\| \left(M_{\hat{\pi}^{(1)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(1)}} \right\|_F^2 \right] \tag{3.150}$$

In summary, we have reduced the problem of estimating π^* into the sum of $n/\tilde{n}$ problems of size $\tilde{n}$. Under $\nu_{G^{t(\iota)}}$, the restriction of M to $\overline{G}^{(\iota)}$ contains $\tilde{n}/2$ good experts (those in $G^{t(\iota)}$) and $\tilde{n}/2$ bad experts. The square Euclidean distance between these two types of experts is $\frac{q\upsilon^2 d\zeta^2}{\lambda_0 \tilde{d}}$. If we denote $\hat{G}^{t(\iota)}$ the set of the $\tilde{n}/2$ best experts according to $\hat{\pi}^{(\iota)}$, then the loss writes as

$$\left\| \left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}} \right\|_F^2 = \frac{q\upsilon^2 d\zeta^2}{\lambda_0 \tilde{d}} \left| \hat{G}^{(\iota)} \Delta G^{t(\iota)} \right| .$$

Coming back to (3.150), we obtain

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq \frac{nq\upsilon^2 d\zeta^2}{\tilde{n}\lambda_0\tilde{d}} \inf_{\hat{G}^{(1)}} \sup_{G^{t(1)}} \mathbf{E}^{(\mathbf{full})}_{G^{t(1)}}\left[\left|\hat{G}^{(1)}\Delta G^{t(1)}\right|\right] \ .$$

Since all possible values of $G^{t(1)}$ are $\tilde{n}/4$-apart by definition of the collection $\mathcal{G}$, we deduce that

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq \frac{nq\upsilon^2 d\zeta^2}{8\lambda_0\tilde{d}} \inf_{\hat{G}^{(1)}} \sup_{G^{t(1)}} \mathbf{P}^{(\mathbf{full})}_{G^{t(1)}}\left[\hat{G}^{(1)} \neq G^{t(1)}\right] \ .$$

For any group $G^{t(1)}$, under $\nu_{G^{t(1)}}$, the rows of the restrictions of M to $\overline{G}^{t(1)}$ are block-constant with $\tilde{d}$ blocks of $d/\tilde{d}$ questions. Consider the $\tilde{n}\times\tilde{d}$ matrices N and $Y^{\downarrow}$ defined by

$$N_{i,j} = \sum_t \mathbf{1}_{x_t\in\{i\}\times[(j-1)(d/\tilde{d})+1,j(d/\tilde{d})+1]} \ ; \qquad Y^{\downarrow}_{i,j} = \sum_t \mathbf{1}_{x_t\in\{i\}\times[(j-1)(d/\tilde{d})+1,j(d/\tilde{d})+1]}\left(y_t - \frac{\tilde{n}}{4n} - \frac{j}{4\tilde{d}}\right) \ .$$

To simplify the notation, we write henceforth G and $\hat{G}$ for $G^{t(1)}$ and $\hat{G}^{(1)}$ respectively. We also write $\mathbf{P}_G$ for the corresponding marginal distribution of N and $Y^{\downarrow}$. By a sufficiency argument, it turns out that

$$\inf_{\hat{G}}\sup_{G} \mathbf{P}^{(\mathbf{full})}_{G}\left[\hat{G}\neq G\right] = \inf_{\hat{G}}\sup_{G} \mathbf{P}_G\left[\hat{G}\neq G\right] \ .$$

Hence, we arrive at the following conclusion

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq \frac{nq\upsilon^2 d\zeta^2}{8\lambda_0\tilde{d}} \inf_{\hat{G}}\sup_{G} \mathbf{P}_G\left[\hat{G}\neq G\right] \ . \tag{3.151}$$

Let us introduce a third-part distribution $\mathbf{P}_0$ on N and $Y^{\downarrow}$ corresponding to the case $\upsilon = 0$. Each of the entry of N therefore follows an independent Poisson distribution with parameter $\tilde{\lambda}$ and, given $N_{i,j}$, we have $Y^{\downarrow}\sim\mathcal{N}(0,N_{i,j}\zeta^2)$. We then deduce from Fano's inequality [90] that

$$\inf_{\hat{G}}\sup_{G\in\mathcal{G}} \mathbf{P}_G(\hat{G}\neq G) \geq 1 - \frac{1+\max_{G\in\mathcal{G}}\mathrm{KL}(\mathbf{P}_G\|\mathbf{P}_0)}{\log(|\mathcal{G}|)} \ , \tag{3.152}$$

where KL(.||.) stands for the Kullback-Leibler divergence. Then, the following lemma bounds these Kullback-Leibler divergences.

Lemma 3.5.42. *Assume that $\lambda_0 = \tilde{n}\lambda d/\tilde{d}\geq 1$ and that $8\upsilon^2\leq 1$. For any $G\in\mathcal{G}$, we have*

$$\mathrm{KL}(\mathbf{P}_G\|\mathbf{P}_0) \leq \frac{4\upsilon^2 q^2}{\tilde{d}} \ .$$

In the specific case where $\tilde{d} = q = 1$, we have $\mathrm{KL}(\mathbf{P}_G\|\mathbf{P}_0) = \upsilon^2/2$ for any $G\in\mathcal{G}$, any $\lambda_0>0$, and any $\upsilon>0$.

Let us summarize our findings by combining (3.151), (3.152), with Lemma 3.5.42 and the different constraints on the parameters (3.149).

Proposition 3.5.43. *Provided that $\tilde{n}$, $\tilde{d}$, q, and υ satisfy the two following conditions*

$$\lambda \geq \frac{\tilde{d}}{\tilde{n}d} \ ; \tag{3.153}$$

$$\upsilon \leq 2^{-3/2}\bigwedge\left[c_0\frac{\sqrt{\tilde{d}\tilde{n}}}{q}\bigwedge c_1\frac{\sqrt{\lambda}}{\zeta}\left[\frac{\tilde{n}^{3/2}d^{1/2}}{n\tilde{d}^{1/2}}\wedge\frac{\sqrt{\tilde{n}d}}{\tilde{d}^{3/2}}\right]\right] \ , \tag{3.154}$$

then, we have

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq c''\frac{nq\upsilon^2\zeta^2}{\tilde{n}\lambda} \ . \tag{3.155}$$

In the specific case where we fix $\tilde{d} = q = 1$ and $\tilde{n} = n$, we can deduce from combining (3.151), (3.152), and the second part of Lemma 3.5.42 that

$$\mathcal{R}^*[n,d,\lambda,\zeta] \geq c''\frac{n\upsilon^2\zeta^2}{\tilde{n}\lambda} \ ,$$

provided that $\upsilon^2\leq c'\frac{\lambda nd}{\zeta^2}\wedge n$. By choosing υ^2 of the order of the right-hand side, we then deduce that

$$\mathcal{R}^*[n,d,\zeta] \geq c\left[\frac{n\zeta^2}{\lambda}\wedge nd\right] \ . \tag{3.156}$$

3.5.9.5 Step 3. Choice of the parameters and conclusion

Writing $\lambda' = \lambda/\zeta^2$, recall that we aim at proving that

$$R[n,d,\lambda,\zeta] \geq c\left[\left[\frac{nd^{1/6}}{\lambda'^{5/6}} \bigwedge \frac{n^{3/4}d^{1/4}}{\lambda'^{3/4}} \bigwedge \frac{n^{2/3}\sqrt{d}}{\lambda'^{5/6}} \bigwedge \frac{n\sqrt{d}}{\lambda'}\right] + \frac{n}{\lambda'} + \frac{n}{\lambda}e^{-2\lambda}\right] \bigwedge nd \ . \tag{3.157}$$

Since we have proved the lower bound (3.145) and (3.156), we only have to prove the corresponding minimax lower bound for the remaining four rates. For this purpose, we shall fix the values of $\tilde{n}$, $\tilde{d}$, q, and υ and apply from Proposition 3.5.43. In the sequel we write $\lfloor x \rfloor_{dya}$ for $2^{\lfloor \log_2(x) \rfloor}$.

Case 1: Rate $\frac{nd^{1/6}}{\lambda'^{5/6}}$. This rate can only occur if $n \leq d$, $\lambda' \in [n^3/d, d^2]$ and $\lambda \geq 1 \wedge [\lambda'^{5/6}/d^{1/6}]$. In this case, we take $\tilde{n} = 2$, $\tilde{d} = \lfloor (\lambda' d)^{1/3} \rfloor_{dya}$, and $q = \lfloor \sqrt{\tilde{d}} \rfloor$. One readily checks that the conditions (3.153) and (3.154) are satisfied for a universal numerical value of υ. Then, Proposition 3.5.43 leads to the desired rate.

Case 2: Rate $\frac{n^{3/4}d^{1/4}}{\lambda'^{3/4}}$. This rate can only occur if $\lambda \geq [1 \wedge (n\lambda'^3/d)^{1/4}]$ and (a) either $n \leq d$ and $\lambda' \in [\frac{n}{d}, \frac{n^3}{d}]$ or (b) $n \in [d; d^2]$ and $\lambda' \in [\frac{n}{d}, \frac{d^3}{n}]$. In this case, we take $\tilde{d} = \lfloor (\lambda' n d)^{1/4} \rfloor_{dya}$, $\tilde{n} = \lfloor n/\tilde{d} \rfloor_{dya}$, and $q = \lfloor \sqrt{\tilde{n}\tilde{d}} \rfloor$. One readily checks that the conditions (3.153) and (3.154) are satisfied for an universal numerical value of υ. Then, Proposition 3.5.43 leads to the desired rate.

Case 3: Rate $\frac{n^{2/3}\sqrt{d}}{\lambda'^{5/6}}$. This rate can only occur if $\lambda \geq [1 \wedge \frac{\lambda'^{5/6} n^{1/3}}{\sqrt{d}}]$ and (a) either $n \in [d, d^2]$ and $\lambda' \in [\frac{d^3}{n}, n^2]$ or (b) $n \geq d^2$ and $\lambda' \in [\frac{n^2}{d^3}, n^2]$. In this case, we take $\tilde{n} = \lfloor (n^2/\lambda')^{1/3} \rfloor_{dya}$, $\tilde{d} = d$, and $q = \lfloor \sqrt{\tilde{n}\tilde{d}} \rfloor$. One readily checks that the conditions (3.153) and (3.154) are satisfied for an universal numerical value of υ. Then, Proposition 3.5.43 leads to the desired rate.

Case 4: Rate $\frac{n\sqrt{d}}{\lambda'}$. This rate can only occur if $\lambda \geq 1$ and $\lambda' \geq (n \vee d)^2$. In this case, we take $\tilde{n} = 2$, $\tilde{d} = d$, and $q = \lfloor \sqrt{d} \rfloor$. One readily checks that the conditions (3.153) and (3.154) are satisfied for a universal numerical value of υ. Then, Proposition 3.5.43 leads to the desired rate. This concludes the proof.

3.5.9.6 Proof of Lemma 3.5.42

Proof of Lemma 3.5.42. In order to bound the Kullback-Leibler discrepancy $\mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0)$, we first observe that the rows of N and $Y^{\downarrow}$ outside G have the same distribution on $\mathbf{P}_G$ and $\mathbf{P}_0$. Besides, all the rows of N and $Y^{\downarrow}$ in G are identically distributed on $\mathbf{P}_G$ and on $\mathbf{P}_0$. Define the vectors $\overline{N}$ and $\overline{Y}^{\downarrow}$ by $\overline{N}_j = \zeta^{-1} \sum_{i \in G} N_{i,j}$ and $Y^{\downarrow}_j = \zeta^{-1} \sum_{i \in G} Y^{\downarrow}_{i,j}$ are a sufficient statistic for deciphering $\mathbf{P}_G$ and $\mathbf{P}_0$, we have $\mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0) = \mathrm{KL}(\mathbf{P}' \| \mathbf{P})$ where $\mathbf{P}'$ and $\mathbf{P}$ stand for the corresponding marginal distributions of $\overline{N}$ and $\overline{Y}^{\downarrow}$.

Set $u = \upsilon/\sqrt{\lambda_0}$. Under $\mathbf{P}$, given $\overline{N}$, the $\overline{Y}^{\downarrow}_j$'s are independent and satisfy $\overline{Y}^{\downarrow}_j \sim \mathcal{N}(0, \overline{N}_j)$. Under $\mathbf{P}'$, conditionally to the subset Q of size q and conditionally to $\overline{N}$, the $\overline{Y}^{\downarrow}_j$'s are independent and satisfy $\overline{Y}^{\downarrow}_j \sim \mathcal{N}(u\overline{N}_j \mathbf{1}\{j \in Q\}, \overline{N}_j)$.

In the specific case of $q = \tilde{d} = 1$, we can explicitely compute the Kullback Leibler divergence. Conditionally to $\overline{N}_1 = x$, $\overline{Y}^{\downarrow}$ is either distributed $\mathcal{N}(0, x)$ under $\mathbf{P}$ and $\mathcal{N}(ux, x)$ under $\mathbf{P}'$. Hence, their conditional Kullback-divergence is $u^2 x/2$. Integrating with respect to x, we conclude that

$$\mathrm{KL}(\mathbf{P}' \| \mathbf{P}) = \mathbf{E}\left[\frac{u^2}{2}\overline{N}\right] = \frac{u^2\lambda_0}{2} = \frac{\upsilon^2}{2} \ .$$

We have shown the second result.

Let us come back to the general case. For $z = 1, 0$, define

$$\alpha_z(x,y) = \frac{\lambda_0^x e^{-\lambda_0}}{x!} \frac{1}{\sqrt{2\alpha x}} \exp\Big(-\frac{(y - uxz)^2}{2x}\Big) \ .$$

Then, the density of $\mathbf{P}$ with respect to $\mu \otimes \lambda$ where μ is the discrete measure and λ is the Lebesgues measure is $\prod_j \alpha_0(\overline{N}_j, \overline{Y}^{\downarrow}_j)$. Besides, the density of $\mathbf{P}'$ is

$$\int \Big[\prod_j \alpha_{\mathbf{1}_{j \in Q}}(\overline{N}_j, \overline{Y}^{\downarrow}_j)\Big] d\eta(Q) \ ,$$

where η stands for the uniform distribution over $\{Q \in \{0,1\}^{\tilde{d}} : \|Q\|_0 = q\}$. It is more convenient to first control the χ^2 distance $\mathbf{P}$ and $\mathbf{P}'$. Since this distance is, up to an additive term of order 1, the second moment of the likelihood ratio between $\mathbf{P}$ and $\mathbf{P}'$, we arrive at the following

$$\begin{aligned}
&\chi^2(\mathbf{P}', \mathbf{P}) + 1 \\
&= \int \Big[\prod_{j \in Q \cap Q'} \frac{[\alpha_1(x_j, y_j)]^2}{\alpha_0(x_j, y_j)} d\mu(x_j) dy_j \Big] \Big[\prod_{j \in Q \Delta Q'} \alpha_1(x_j, y_j) d\mu(x_j) dy_j \Big] d\eta(Q) d\eta(Q') \\
&= \int \Big[\prod_{j \in Q \cap Q'} \frac{[\alpha_1(x_j, y_j)]^2}{\alpha_0(x_j, y_j)} d\mu(x_j) dy_j \Big] d\eta(Q) d\eta(Q') \ ,
\end{aligned}$$

since α_1 is a density. Let us work out each of these ratios.

$$\begin{aligned}
\int \frac{\alpha_1^2(x,y)}{\alpha_0(x,y)} dx dy &= \int \alpha_0(x,y) \exp\Big[\frac{2yux - u^2x^2}{x} \Big] d\mu(x) dy \\
&= \sum_{x=0}^{\infty} \frac{\lambda_0^x e^{-\lambda_0}}{x!} e^{u^2 x} = \exp(\lambda_0 (e^{u^2} - 1)) := \exp(\mathcal{I}) \ .
\end{aligned}$$

Coming back to the χ^2 distance, we arrive at the following equality

$$\chi^2(\mathbf{P}', \mathbf{P}) = \int \exp\left(\mathcal{I} |Q \cap Q'|\right) d\eta(Q) d\eta(Q') - 1 \ .$$

Here, $|Q \cap Q'|$ is distributed as an Hypergeometric distribution with parameters $\tilde{d}$, q, and $q/\tilde{d}$. We know from Aldous (p.173) [2] that $|Q \cap Q'|$ follows the same distribution as the random variable $\mathbb{E}(W|\mathcal{B})$ where W is a binomial random variable of parameters q, $q/\tilde{d}$ and $\mathcal{B}$ is some suitable σ-algebra. By Jensen's inequality, we deduce that

$$\chi^2(\mathbf{P}', \mathbf{P}) \le \mathbb{E}[\exp(\mathcal{I} W)] - 1 = \Big[1 + \frac{q}{\tilde{d}} (\exp(\mathcal{I}) - 1) \Big]^q - 1 \ .$$

Recall that $\lambda_0 u^2 = \upsilon^2 \le 1/8$. Hence, provided that $\lambda_0 = \tilde{n}\lambda d/\tilde{d} \ge 1$, we have $\mathcal{I} \le 2\lambda_0 u^2 = 2\upsilon^2$. It then follows that

$$\chi^2(\mathbf{P}', \mathbf{P}) \le \exp\left(q^2/\tilde{d} (\exp(\mathcal{I}) - 1) \right) - 1 \le \exp\left(4\upsilon^2 q^2/\tilde{d} \right) - 1.$$

To conclude, we use the classical bound $\mathrm{KL}(\mathbf{P}' \| \mathbf{P}) \le \log\left(1 + \chi^2(\mathbf{P}', \mathbf{P})\right)$ –see e.g. [90]. This leads us to

$$\mathrm{KL}(\mathbf{P}' \| \mathbf{P}) \le \frac{4\upsilon^2 q^2}{\tilde{d}} \ .$$

□

3.5.9.7 Proof of Theorem 3.2.1

Fix n, d, ζ, and $\kappa \ge 2$, and assume that, for some c', there exists an estimator $\hat{\pi}$ satisfying

$$\sup_{\pi^* \in \Pi_n} \sup_{M : M_{\pi^{*-1}} \in \mathbb{C}_{\mathrm{BISO}}} \mathbb{E}_{(\pi^*, M)} \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \le c' \left[\log^{-\kappa}(nd/\zeta_-) \mathcal{R}_F[n, d, \zeta] \bigwedge nd \right] \ , \tag{3.158}$$

with $\Upsilon = \lfloor \log^{\kappa}(nd/\zeta_-) \rfloor$ samples.

Let us show that this bound would contradict the minimax lower bound in the Poisson setting. Fix $\lambda = \frac{112}{3} \log^{\kappa}(nd/\zeta_-)$ and consider the model (3.2). Define the estimator $\tilde{\pi}$ such that $\tilde{\pi} = \hat{\pi}$ under the event $\mathcal{A}$ such that there are at least Υ observations on each entry and $\tilde{\pi}$ is computed arbitrarily otherwise. By (3.158), $\tilde{\pi}$ satisfies

$$\mathbb{E}_{(\pi^*, M)} \|M_{\tilde{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \le nd \mathbb{P}[\mathcal{A}^c] + c' \left[\log^{-\kappa}(nd/\zeta_-) \mathcal{R}_F[n, d, \zeta] \bigwedge nd \right] \ . \tag{3.159}$$

By Chernoff'inequality for Poisson random variable, we deduce that

$$\mathbb{P}[\mathcal{A}^c] \quad \le \quad nd \exp\left[-\frac{3}{28} \lambda \right] \le nd e^{-4 \log^{\kappa}(nd/\zeta_-)} \le \frac{\zeta_-^2}{nd} e^{-4 \log^{\kappa}(nd/\zeta_-) + 2 \log(nd/\zeta_-)}$$

There exists a constant c_0 such that for any $\kappa \ge 2$, $e^{-4x^{\kappa} + 2x} \le \frac{c_0}{x^{2\kappa}}$. We deduce that

$$\mathbb{P}[\mathcal{A}^c] \quad \le \quad \frac{\zeta_-^2}{nd} \frac{c_0}{\log^{2\kappa}(nd/\zeta_-)} \ ,$$

where we used that $e^x \geq 1 + x^\beta/\beta$ for any $x \geq 0$ and any $\beta > 0$ and that $\kappa \geq 2$. We then deduce from (3.159) that

$$\mathbb{E}_{(\pi^*,M)} \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq \left(c' + \frac{c_0}{\log^\kappa(nd/\zeta_-)}\right) \left[\log^{-\kappa}(nd/\zeta_-)\mathcal{R}_F[n,d,\zeta] \bigwedge nd\right] . \tag{3.160}$$

For $\lambda \geq 1$, $\mathcal{R}_F[n,d,\zeta/\sqrt{\lambda}] \geq \frac{\mathcal{R}_F[n,d,\zeta]}{\lambda}$. We deduce that

$$\mathbb{E}_{(\pi^*,M)} \|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq \frac{112}{3}\left(c' + \frac{c_0}{\log^\kappa(nd/\zeta_-)}\right) \left[\mathcal{R}_F[n,d,\zeta/\sqrt{\lambda}] \bigwedge nd\right] .$$

Taking c' small enough compared to the numerical constant c in Theorem 3.4.1 contradicts this last theorem provided that nd/ζ_- is larger than some some numerical constant. Hence, no estimator can achieve (3.158) for this constant c' when (nd/ζ_-) is large enough.

It remains to consider the case where nd/ζ_- is smaller than some constant $c'' \geq 2$. We only need to prove that the minimax risk is lower bounded by $\frac{c_0}{\Upsilon}$ where Υ is the sample size. Since the minimax risk is non-decreasing with respect to n, d, and ζ, we only have to consider the case $n = 2$, $d = 1$, $\zeta = 2/c''$. Define $a = \zeta/\sqrt{\Upsilon}$. Consider a problem where either $M = (a,0)^T$ or $M = (0,a)^T$. Then, with positive probability, no test is able to distinguish both hypotheses and the risk of any estimator is at most of the order $a^2 = \zeta^2/\Upsilon$. The result follows.

Chapter 4

Ranking a permuted matrix under the isotonic model

We consider a ranking problem where we have noisy observations from a matrix with isotonic columns whose rows have been permuted by some permutation π^*. This encompasses many models, including crowd-labeling and ranking in tournaments by pair-wise comparisons. In this work, we provide an optimal and polynomial-time procedure for recovering π^*, settling an open problem in [33]. As a byproduct, our procedure is used to improve the state-of-the art for ranking problems in the stochastically transitive model (SST). Our approach is based on iterative pairwise comparisons by suitable data-driven weighted means of the columns. These weights are built using a combination of spectral methods with new dimension-reduction techniques. In order to deal with the important case of missing data, we establish a new concentration inequality for sparse and centered rectangular Wishart-type matrices.

4.1 Introduction

Ranking problems have recently spurred a lot of interest in the statistical and computer science literature. This includes a variety of problems ranging from ranking experts/workers in crowd-sourced data, ranking players in a tournament or equivalently sorting objects based on pairwise comparisons.

To fix ideas, let us consider a problem where we have noisy partial observations from an unknown matrix $M \in [0,1]^{n\times d}$. In crowdsourcing problems, n stands for the number of experts (or workers), d stands for the number of questions (or tasks) and $M_{i,k}$ for the probability that expert i answers question k correctly. For tournament problems, we have $n = d$ players (or objects) and $M_{i,k}$ stands for the probability that player i wins against player k. Based on these noisy data, the general goal is to provide a full ranking of the experts or of the players.

Originally, these problems were tackled using parametric model for the matrix M. Notably, this includes the noisy sorting model [12] or Bradley-Luce-Terry model [11]. Still, it has been observed that these simple models are often unrealistic and do not tend to fit well.

This has spurred a recent line of literature where strong parametric assumptions are replaced by non-parametric assumptions [81, 83, 84, 85, 60, 59, 56, 33, 8, 79]. In particular, for tournament problems, the strong stochastically transitive (SST) model presumes that the square matrix M is, up to a common permutation π^* of the rows and of the columns, bi-isotonic and satisfies the skew symmetry condition $M_{i,k} + M_{k,i} = 1$. Although optimal rates for estimation of the permutation π^* have been pinpointed in the earlier paper of Shah et al. [83], there remains a large gap between these optimal rates and the best known performances of polynomial-time algorithms. This has led to conjecture the existence of a statistical-computational gap [60, 56].

For crowdsourcing data, the counterpart of the SST model is the so-called bi-isotonic model, where the rectangular matrix M is bi-isotonic, up to an unknown permutation π^* of its rows and an unknown permutation η^* of its columns. This model turns out to be really similar to the SST model and the existence of a statistical-computational gap has also been conjectured [60].

In this chapter, we tackle a slightly different route and we consider the arguably more general isotonic model [33]. The only assumption is that all the columns of M are nondecreasing up to an unknown permutation of the rows, making the isotonic model more flexible than the bi-isotonic and SST models. It is in fact the most general model under which an unambiguous ranking of the experts is well-defined. In this model as well, there is a gap between the (statistical) optimal rates, and the rate obtained by the (polynomial-time) algorithm in [33].

Our main contributions are as follows. For the isotonic model, we establish the optimal rate for recovering the permutation, and we introduce a polynomial-time procedure achieving this rate, thereby settling the absence of any computational gap in this model. Besides, our procedure and results have important consequences when applied to the SST and bi-isotonic model. More specifically, we achieve the best known guarantees in these two models [56, 59] and even improve them in some regimes.

4.1.1 Problem formulation

Let us further introduce our model. A bounded matrix $A \in [0,1]^{n\times d}$ is said to be isotonic if its columns are nondecreasing, that is $A_{i,k} \le A_{i+1,k}$ for any $i \in [n-1]$ and $k \in [d]$. Henceforth, we write $\mathbb{C}_{\text{iso}}$ for the collection of all $n \times d$ isotonic matrices taking values in $[0,1]$. In our model, we recall that we assume that the signal matrix M is isotonic up to an unknown permutation of its rows. In other words, there exists a permutation π^* of $[n]$ such that the matrix $M_{\pi^{*-1}}$ defined by $(M_{\pi^{*-1}})_{i,k} = (M_{\pi^{*-1}(i),k})$ has nondecreasing columns, that is

$$M_{\pi^{*-1}(i),k} \le M_{\pi^{*-1}(i+1),k} \ , \tag{4.1}$$

for any $i \in \{1,\dots,n-1\}$ and $k \in \{1,\dots,d\}$, or equivalently $M_{\pi^{*-1}} \in \mathbb{C}_{\text{iso}}$. Henceforth, π^* is called an oracle permutation. Using the terminology of crowdsourcing, we refer to i^{th} row of M as <u>expert</u> i and to k^{th} column as <u>question</u> k.

In this work, we have N partial and noisy observations of the matrix M of the form (x_t, y_t) where

$$y_t = M_{x_t} + \varepsilon_t \quad t = 1,\dots,N \ . \tag{4.2}$$

For each t, the position $x_t \in [n]\times[d]$ is sampled uniformly. The noise variables ε_t's are independent and their distributions only depend on the position x_t. We only assume that all these distributions are centered and are subGaussian with a subGaussian norm of at most 1 – see e.g. [94]. In particular, this encompasses the typical case where the y_t's follow Bernoulli distributions with parameters M_{x_t}.

As usual in the literature e.g. [74, 56, 60], we use, for technical convenience, the Poissonization trick which amounts to assuming that the number N of observations has been sampled according to a Poisson distribution with parameter λnd. We refer to $\lambda > 0$ as the sampling effort. When $\lambda > 1$, we have, in expectation, several independent observations per entry (i,j) - and $\lambda = 1$ means that there is on average one observation per entry. In this chaper, we are especially interested in the sparse case where λ is much smaller than one, i.e. the case where we have missing observations for some entries. We refer to $\lambda = 1$ as the full observation regime at it bears some similarity to the case often considered in the literature –e.g. [83, 33], where we have a full observation of the matrix,

$$Y = M + E' \in \mathbb{R}^{n\times d} \ . \tag{4.3}$$

The entries of the noise matrix E' are independent, centered, and 1-subGaussian.

In this work, we are primarily interested in estimating the permutation π^*. Given an estimator $\hat\pi$, we use the square Frobenius norm $\|M_{\hat\pi^{-1}} - M_{\pi^{*-1}}\|_F^2$ as the loss. This loss quantifies the distance between the matrix M reordered according to the estimator $\hat\pi$ and the matrix M sorted according to the oracle permutation π^*. This loss is explicitly used in [56, 74] and is implicit in earlier works –see e.g. [83].

We define the associated optimal risk of permutation recovery as a function of the number n of experts, the number d of question and the sampling effort λ,

$$\mathcal{R}^*_{\text{perm}}(n,d,\lambda) = \inf_{\hat\pi} \sup_{\substack{\pi^*\in\Pi_n \\ M:\, M_{\pi^{*-1}}\in\mathbb{C}_{\text{iso}}}} \mathbb{E}_{(\pi^*,M)}\big[\|M_{\hat\pi^{-1}} - M_{\pi^{*-1}}\|_F^2\big] \ , \tag{4.4}$$

where the infimum is taken over all estimators. Here, Π_n stands for the collection of all permutations of $[n]$. If the main focus is not only to estimate π^*, but also to reconstruct the unknown matrix M, we also consider the optimal reconstruction rate

$$\mathcal{R}^*_{\text{reco}}(n,d,\lambda) = \inf_{\hat M} \sup_{\substack{\pi^*\in\Pi_n \\ M:\, M_{\pi^{*-1}}\in\mathbb{C}_{\text{iso}}}} \mathbb{E}\big[\|\hat M - M\|_F^2\big] \ . \tag{4.5}$$

It turns out that reconstructing the matrix M is more challenging than estimating the permutation π^*. Considering both risks allows to disentangle the reconstruction of the matrix M: looking at both enables to distinguish the error that is due to estimating the permutation, from the error that comes from estimating an isotonic matrix.

4.1.2 Past results on the isotonic model and our contributions

In the specific case where $d = 1$ (a single column), our model is closely related to uncoupled isotonic regression and is motivated by optimal transport. Rigollet and Niles-Weed [76] have established that the estimation error $\inf_{\hat{M}^{\text{sorted}}} \sup_{\pi^*, M} \|\hat{M}^{\text{sorted}} - M_{\pi^{*-1}}\|_F^2$ of the sorted vector $M_{\pi^{*-1}}$ is of the order of $n(\frac{\log\log(n)}{\log(n)})^2$.

For the general case $d \geq 1$, Flammarion et al. [33] have shown[1] that the optimal reconstruction error in the full observation model (4.3) is of the order of $n^{1/3}d + n$. However, the corresponding procedure is not efficient. They also introduce an efficient procedure that first estimates π^* using a score based on row comparisons on Y. Unfortunately, this method only achieves a reconstruction error of the order of $n^{1/3}d + n\sqrt{d}$ which is significantly slower than the optimal one. Whether or not there is a statistical-computational gap was therefore an open problem.

We prove in this work that there is no computational statistical gap in this model. More precisely, we introduce estimators that are both polynomial-time and minimax optimal up to some polylog factors. To that end, we characterize the optimal risks $\mathcal{R}^*_{\text{perm}}(n, d, \lambda)$ and $\mathcal{R}^*_{\text{reco}}(n, d, \lambda)$ of permutation estimation and matrix reconstruction, for all possible number of experts $n \geq 1$, number of questions $d \geq 1$ and all sampling efforts λ, up to some polylog factors in nd. Table 4.1 summarizes our findings in the arguably most interesting cases[2] $\lambda \in [1/(n \wedge d), 1]$.

	$n \leq d^{3/2}\sqrt{\lambda}$	$d^{3/2}\sqrt{\lambda} \leq n$
$\mathcal{R}^*_{\text{perm}}$	$n^{2/3}\sqrt{d}\lambda^{-5/6}$	n/λ
$\mathcal{R}^*_{\text{reco}}$	$n^{1/3}d\lambda^{-2/3}$	n/λ

,

Table 4.1: Optimal rates in our model, for all possible values of n, d and $\lambda \in [1/(n \wedge d), 1]$, up to a polylogarithmic factor in nd. These rates are achieved by polynomial-time estimators.

4.1.3 Implication for other models and connection to the literature

As discussed earlier, the isotonic model is quite general and encompasses both the bi-isotonic model for crowd-sourcing problems as well the SST model for tournament problems.

Let us first focus on the SST model which corresponds to the case where $n = d$ together with a bi-isotonicity and a skew-symmetry assumption. In the full observation scheme (related to the case $\lambda = 1$) where one observes the noisy matrix $n \times n$, Shah et al. [83] have established that the optimal rates for estimating π^* and reconstructing the matrix M are of the order of n. In contrast, their efficient procedure which estimates π^* according to the row sums of Y only achieves the rate of $n^{3/2}$. In more recent years, there has been a lot of effort dedicated to improving this $\sqrt{n}$ statistical-computational gap. The SST model was also generalized to partial observations by [19], which corresponds to $\lambda \leq 1$. They introduced an efficient procedure that targets a specific sub-class of the SST model, and that achieves a rate of order $n^{3/2}\lambda^{-1/2}$ in the worst case for matrix reconstruction.

Recently, a few important contributions tackling both the bi-isotonic model and the SST model made important steps towards better understanding the statistical-computational gap. We first explain how their results translate in the SST model. Mao et al. [60, 59] introduced a polynomial-time procedure handling partial observation, achieving a rate of order $n^{5/4}\lambda^{-3/4}$ for matrix reconstruction. Nonetheless, [60] failed to exploit global information shared between the players/experts – as they only compare players/experts two by two – as pointed out by [56]. Building upon this remark, [56] managed to get the better rate $n^{7/6+o(1)}$ with a polynomial-time method in the case $\lambda = n^{o(1)}$.

Let us turn to the more general bi-isotonic model. Here, the rectangular matrix $M \in \mathbb{R}^{n \times d}$ is bi-isotonic up an unknown permutation π^* of the rows and an unknown permutation η^* of the columns. Since M is not necessarily square, this model can be used in more general crowd-sourcing problems. The optimal rate for reconstruction in this model with partial observation has been established in [60] to be of order $\nu(n, d, \lambda) := (n \vee d)/\lambda + \sqrt{nd/\lambda} \wedge n^{1/3}d\lambda^{-2/3} \wedge d^{1/3}n\lambda^{-2/3}$ up to polylog factors, in the non-trivial regime where $\lambda \in [1/(n \wedge d), 1]$. However, the polynomial-time estimator provided by Mao et al. [60] only achieves the rate $n^{5/4}\lambda^{-3/4} + \nu(\lambda, n, d)$. In a nutshell, Mao et al. first compute column sums to give a first estimator of the permutation of the

[1] The authors consider the isotonic model as a subcase of a seriation model, where each columns of $M_{\pi^{*-1}}$ is only assumed to be unimodal.

[2] We are indeed mostly interested in the more realistic sparse observation regime (meaning $\lambda \leq 1$). The case $\lambda \leq 1/d$ leads to the trivial minimax bound of order nd for both reconstruction and estimation, as in this case we have less than one observations per expert on average. As for the case $\lambda > 1/d$ but $\lambda \leq 1/n$, we have less than one observation per question on average, and this leads to a minimax risk of order $n\sqrt{d/\lambda}$ for permutation estimation and of order nd for matrix recontruction.

questions. Then, they compare the experts on aggregated blocks of questions, and finally compare the questions on aggregated blocks of experts. As explained in the previous paragraph for SST models, Liu and Moitra [56] improved this rate to $n^{7/6+o(1)}$ in the square case ($n = d$), with a subpolynomial number of observations per entry ($\lambda = n^{o(1)}$). Their estimators of the permutations π^*, η^* were based on hierarchical clustering and on local aggregation of high variation areas. Both [56, 60] made heavily use of the bi-isotonicity structure of M by alternatively sorting the columns and rows. As mentioned for the SST model, the order of magnitude $n^{7/6+o(1)}$ remains nevertheless suboptimal, and whether there exists an efficient algorithm achieving the optimal rate in this bi-isotonic model remains an open problem.

We now discuss the implications of our work concerning the bi-isotonic model and SST model. First, in the full observation setting ($\lambda = 1$) and square case for the bi-isotonic model ($n = d$), we reach in polynomial-time the upper bound $n^{7/6}$ up to polylog factors, for both permutation estimation and matrix reconstruction. In particular, we improve the rate in [56] by a subpolynomial factor in n, and we do not need a subpolynomial number of observations per entry. Moreover, our procedure being primarily designed for the isotonic model, it does not require any shape constraint on the rows in contrast to [56, 60]. Beyond the full observation regimes, we provide guarantees on our estimator of π^* for different values of λ. In particular, in Corollary 4.2.5, we derive an estimator of the matrix M that achieves a maximum reconstruction risk $\sup_{\pi^*,\eta^*,M} \mathbb{E}\big[\|\hat{M} - M_{\pi^{*-1}\eta^{*-1}}\|_F^2\big]$ of order less than $n^{7/6}\lambda^{-5/6}$ up to polylogs, thereby improving the state-of-the-art polynomial-time methods in partial observation [60]. Lastly, we perform our analysis in the general rectangular case, giving guarantees for general values of d.

The optimal risks and the known polynomial-time upper bounds for the isotonic, bi-isotonic with two permutations and SST models are summarized in Table 4.2. For the sake of simplicity, we focus in the table to the specific case case $n = d$ and $\lambda \in [1/n, 1]$.

Different models, with $M \in \mathbb{R}^{n\times n}$		**Isotonic** $M_{\pi^{*-1}}$ **has nondecreasing columns**	Bi-isotonic(π^*, η^*) $M_{\pi^{*-1}\eta^{*-1}}$ has nondecreasing columns and rows	SST $M_{\pi^{*-1}\pi^{*-1}}$ has nondecreasing columns and rows, and $M_{ik} + M_{ki} = 1$
Permutation estimation	Poly. Time	$n^{7/6}\lambda^{-5/6}$ **[Th 4.2.2]**	$n^{7/6+o(1)}$ [56]$(\lambda = n^{o(1)})$ $n^{7/6}\lambda^{-5/6}$ **[Th 4.2.2]**	$n^{7/6+o(1)}$ [56]$(\lambda = n^{o(1)})$ $n^{7/6}\lambda^{-5/6}$ **[Th 4.2.2]**
	optimal rate	$n^{7/6}\lambda^{-5/6}$ **[Th 4.2.1]**	n/λ [60]	n/λ [60]
Matrix reconstruction	Poly. Time	$n^{3/2}$ $(\lambda = 1)$[33] $n^{4/3}\lambda^{-2/3}$ **[Cor 4.2.5]**	$n^{7/6+o(1)}$ [56]$(\lambda = n^{o(1)})$ $n^{5/4}\lambda^{-3/4}$ [60] $n^{7/6}\lambda^{-5/6}$ **[Cor 4.2.5]**	$n^{7/6+o(1)}$ [56]$(\lambda = n^{o(1)})$ $n^{5/4}\lambda^{-3/4}$ [60] $n^{7/6}\lambda^{-5/6}$ **[Cor 4.2.5]**
	optimal rate	$n^{4/3}\lambda^{-2/3}$ [33] (also [Prop 4.2.3])	n/λ [60]	n/λ [60]

Table 4.2: For the isotonic model, the optimal rate for permutation estimation (resp. matrix reconstruction) corresponds to $\mathcal{R}^*_{\text{perm}}$ (resp. $\mathcal{R}^*_{\text{reco}}$). For the two other columns, the optimal rates are similarly defined as minimax risk over the corresponding models. The Poly. Time rows correspond to state-of-the art rates achieved by polynomial-time methods. All the rates are given up to polylogarithmic factors in n.

Finally, we mention the even more specific model where the matrix M is bi-isotonic up to a single permutation π^* acting on the rows. This corresponds to the case where η^* is known in the previous paragraph [60, 74, 56]. Equivalently, this also corresponds to our isotonic model (4.2) with the additional assumption that all the rows are nondecreasing, that is $M_{i,k} \le M_{i,k+1}$. For this model, it is possible to leverage the shape constrains on the rows to build efficient and optimal estimators, this for all n, d, and λ – see [74].

4.1.4 Overview of our techniques

In this work, we introduce the iterative soft ranking (**ISR**) procedure, which gives an estimator $\hat{\pi}$ based on the observations. Informally, this method iteratively updates a weighted directed graph between experts, where the weight between any two experts quantifies the significance of their comparison. The procedure increases the weights at each step. After it stops, the final estimator is an arbitrary permutation $\hat{\pi}$ that agrees as well as possible with the final weighted directed graph.

As mentioned in [56], it is hopeless to use only local information between pairs of experts to obtain a rate of order $n^{7/6}$ up to polylogs, and we must exploit global information. Still, we do it in a completely different

way than Liu and Moitra [56], who were building upon the bi-isotonicity of the matrix.

One first main ingredient of our procedure is a new dimension reduction technique. At a high level, suppose that we have partially ranked the rows in such a way that, for a given triplet (P, O, I) of subsets of $[n]$, we are already quite confident that experts in P are below those in I and above those in O. Relying on the shape constraint of the matrix M, it is therefore possible to build a high-probability confidence regions for rows in P based on the rows in O and the rows in I. If, for a question j, the confidence region is really narrow, this implies that all experts in P take almost the same value on this column. As a consequence, this question is almost irrelevant for further comparing the experts in P. In summary, our dimension reduction technique selects the set of questions for which the confidence region of P is wide enough, and in that way reduces the dimension of the problem while keeping most of the relevant information.

The second main ingredient, once the dimension is reduced, is to use a spectral method to capture some global information shared between experts. That is why our procedure makes significant use of spectral methods to compute the updates of the weighted graph. Although this spectral scheme already appears in recent works [74, 56], those are used here for updating the weight of the comparison graph rather than performing a clustering as in [56]. Moreover, the analysis of the spectral step in the partial observation regime ($\lambda \ll 1$) leads to technical difficulties – see the discussion in Section 4.3.5.

Related to the latter problem, we need to establish a new tail bound on sparse rectangular matrices. More specifically, for a rectangular matrix X with centered independent entries that satisfy a Bernstein type condition, we provide a high-probability control of the operator norm of $XX^T - \mathbb{E}[XX^T]$. This result, based on non-commutative matrix Bernstein concentration inequality, may be of independent interest e.g. for controlling the spectral properties of a sparse bipartite random graph. We state it in Section 4.4, independently of the rest of this chapter.

4.1.5 Notation

Given a vector u and $p \in [1, \infty]$, we write $\|u\|_p$ for its l_p norm. For a matrix A, $\|A\|_F$ and $\|A\|_{\mathrm{op}}$ stand for its Frobenius and its operator norm. We write $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) for the largest (resp. smallest) integer smaller than (resp. larger than) or equal to x. Although M stands for an $n \times d$ matrix, we extend it sometimes in an infinite matrix defined for all $i \in \mathbb{N}, k \in \{1, \dots, d\}$ by setting $M_{ik} = 0$ when $i \le 0$ and $M_{ik} = 1$ when $i \ge n + 1$. The corresponding infinite matrix $M_{\pi^{*(-1)}}$ which is obtained by permuting the n original rows is still isotonic and takes values in $[0, 1]$. We shall often work with submatrices $M(P, Q)$ of M that are restricted to a subset $P \subset [n]$ and $Q \subset [d]$ of rows and columns. If A is any matrix in $\mathbb{R}^{P \times Q}$, we write $\overline{A}$ for the matrix whose rows are all equal to the average row of A, namely $\overline{A}_{ik} = \frac{1}{|P|} \sum_{j \in P} A_{jk}$.

4.2 Results

In this section, we first establish the statistical limit with a lower bound on $\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda)$. Then, we state the existence of a polynomial-time estimator that is minimax optimal up to polylog factors. More precisely, we prove that for all integers n, d and $\lambda \in [1/d, 8n^2]$, the optimal rate of permutation estimation $\mathcal{R}^*_{\mathrm{perm}}$ is of the order of

$$\rho_{\mathrm{perm}}(n, d, \lambda) := \frac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} \bigwedge n\sqrt{\frac{d}{\lambda}} + \frac{n}{\lambda} \ , \tag{4.6}$$

up to some polylog factors. As a corollary, we then establish that the optimal rate of matrix reconstruction $\mathcal{R}^*_{\mathrm{reco}}$ is of order

$$\rho_{\mathrm{reco}}(n, d, \lambda) := \frac{n^{1/3} d}{\lambda^{2/3}} + \frac{n}{\lambda} \ , \tag{4.7}$$

up to polylog factors. We therefore establish that these two problems do not exhibit a computational-statistical gap.

4.2.1 Minimax lower bound for permutation estimation

Assume that $\lambda \in [1/d, 8n^2]$ is fixed and that we are given $N = Poi(\lambda nd)$ independent observations under model (4.2). Namely, we observe $(x_t, y_t)_{t=1,\dots,N}$ where x_t is sampled uniformly in $[n] \times [d]$ and $y_t = M_{x_t} + \varepsilon_t$ conditionally to x_t. The following theorem states that ρ_{perm} is a lower bound on the maximum risk of permutation estimation for all $n, d, \lambda \in [1/d, 8n^2]$, up to some numerical constant.

Theorem 4.2.1. *There exists a universal constant $c > 0$ such that, for any $n \ge 2$, $d \ge 1$, and $\lambda \in [1/d, 8n^2]$, we have*

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c\rho_{\mathrm{perm}}(n, d, \lambda) \ . \tag{4.8}$$

In the proof, we show a slightly stronger result that also covers the cases $\lambda < 1/d$ and $\lambda > 8n^2$, where $\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda)$ is in fact respectively lower bounded by a quantity of order nd and $n\sqrt{d}/\lambda$. For the sake of readability, we chose to omit these arguably less interesting cases in the statement of Theorem 4.2.1 and of Theorem 4.2.2.

4.2.2 Optimal permutation estimation

Let us fix a quantity $\delta \in (0,1)$ that will correspond to a small probability. We need to introduce some notation. We write

$$\phi_{\mathrm{L}_1} = 10^4 \log\left(\frac{10^2 nd}{\delta}\right) \ . \tag{4.9}$$

Our procedure depends on a sequence of tuning parameters. For this reason, we introduce a subset $\Gamma \subset \mathbb{R}^+$, henceforth called a grid. The grid Γ is said to be valid if it contains a sequence $\gamma_0 \geq \cdots \geq \gamma_{2\lfloor \log_2(n)\rfloor+2}$ of length $2\lfloor \log_2(n)\rfloor + 3$ such that for all u,

$$\gamma_u - \gamma_{u+1} \geq \gamma_{2\lfloor \log_2(n)\rfloor+2} + \phi_{\mathrm{L}_1} \quad \text{and} \quad \gamma_{2\lfloor \log_2(n)\rfloor+2} \geq \phi_{\mathrm{L}_1} \ . \tag{4.10}$$

In light of this definition, we could simply choose the valid sequence $\Gamma = \{\phi_{\mathrm{L}_1}, 2\phi_{\mathrm{L}_1}, \dots, (2\lfloor \log_2(n)\rfloor + 3)\phi_{\mathrm{L}_1}\}$ with a corresponding γ_0 that is polylogarithmic. Still, for practical purpose, we consider general grids; examples of such grids are discussed in more details in Section 4.3.6.

For any valid subset Γ, we define $\bar{\gamma}$ as the smallest possible value of γ_0 over all sequences that satisfy (4.10).

$$\bar{\gamma} = \min\{\gamma \ : \ \exists(\gamma_u) \text{ satisfying (4.10) s.t. } \gamma_0 = \gamma\} \ . \tag{4.11}$$

Our main procedure **ISR**, for iterative soft ranking, will be described in detail in Section 4.3. The only tuning parameters are the number of steps T and the valid grid Γ.

Theorem 4.2.2. *There exists $C > 0$ such that the following holds. Let $\lambda \in [1/d, 8n^2]$ and $\delta > 0$. Assume that Γ is a valid grid and that $T \geq 4\bar{\gamma}^6$ with $\bar{\gamma}$ defined in (4.11). For any permutation $\pi^* \in \Pi_n$ and any matrix M such that $M_{\pi^{*-1}} \in \mathbb{C}_{\mathrm{iso}}$, the estimator $\hat{\pi}$ from Algorithm* **ISR**(T,Γ) *defined in the next section satisfies*

$$\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq CT\bar{\gamma}^6 \rho_{\mathrm{perm}}(n,d,\lambda) \ ,$$

with probability at least $1 - 10T\delta$.

In particular, if we suitably choose Γ (as discussed above) and $T = 4\lceil \bar{\gamma}^6 \rceil$ and $\delta = 1/(nd)^2$, we deduce from Theorem 4.2.2 that

$$\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda) \leq C' \log^{C'}(nd)\rho_{\mathrm{perm}}(n,d,\lambda) \ ,$$

for some numerical constant $C' > 0$. In the case where $\lambda = n^{o(1)}$ and $n = d$, this bound achieves the order of magnitude $n^{7/6}$, which aligns with the result presented in Theorem 2 of Liu and Moitra [56]. However, it is important to note that the analysis made in [56] focuses on the statistically easier bi-isotonic model, and their procedure heavily relies on the isotonicity structure imposed on the questions.

4.2.3 Optimal reconstruction of the matrix M

We now turn to the problem of estimating the signal matrix M. Obviously, the reconstruction of the matrix M from the observation of model in(4.2) is at least as hard as if we knew the permutation π^*. In this favorable situation, estimating M amounts to estimating d isotonic vectors from partial and noisy observations $Y_{ik} = \frac{1}{\lambda}\sum_t y_t \mathbf{1}_{x_t=(ik)}$. The isotonic regression problem is already well understood, and we state the following lower bound without proof since it directly follows from [60] (see in particular Theorem 3.1 therein). We recall that $\rho_{\mathrm{reco}}(n,d,\lambda)$ is defined in (4.7).

Proposition 4.2.3. *There exists a universal constant $c > 0$ such that, for any $n \geq 2$, any $d \geq 1$, and any $\lambda > 0$, we have*

$$\mathcal{R}^*_{\mathrm{reco}}(n,d,\lambda) \geq c\rho_{\mathrm{reco}}(n,d,\lambda) \ . \tag{4.12}$$

In particular, since $\rho_{\mathrm{perm}}(n,d,\lambda) \ll \rho_{\mathrm{reco}}(n,d,\lambda)$ in many regimes in n, d, λ, this proposition implies that the reconstruction of a permuted isotonic matrix is harder than the estimation of the permutation, namely that $\mathcal{R}^*_{\mathrm{perm}} \ll \mathcal{R}^*_{\mathrm{reco}}$.

To build an optimal estimator of M, we compute the estimated permutation $\hat{\pi}$ of Theorem 4.2.2 and estimate an isotonic matrix based on this ordering. This approach is similar to what is done in [60, 74], for

related problems where a bi-isotonic assumption is done. For simplicity, set the tuning parameters T, Γ for Algorithm 14 so that $T = 4\lceil\bar{\gamma}^6\rceil$ and $\bar{\gamma}^6 \le C'\log^{C'}(nd/\delta)$. We split the samples y_t defined in (4.2) into two independent sequences of samples $(y_t^{(1)})$, $(y_t^{(2)})$. First, we compute the estimator $\hat{\pi}$ of π^* with the first subsamples $(y_t^{(1)})$. Then, we define $\hat{M}_{\mathrm{iso}}$ as the projection of $Y^{(2)}_{\hat{\pi}}$ onto the convex set of isotonic matrices, where $Y^{(2)}$ is the matrix defined by $Y^{(2)}_{ik} = \frac{1}{\lambda}\sum_t y_t^{(2)}\mathbf{1}_{x_t^{(2)}=(i,k)}$. More precisely, set

$$\hat{M}_{\mathrm{iso}} = \operatorname*{arg\,min}_{\tilde{M}\in\mathbb{C}_{\mathrm{iso}}} \|\tilde{M} - Y^{(2)}_{\hat{\pi}^{-1}}\|_2^2 \ .$$

The following corollary controls the risk of $\hat{M}_{\mathrm{iso}}$.

Corollary 4.2.4. *Assume that $\lambda \in [1/d, 8n^2]$. There exists a universal constant C'' such that the following holds for any permutation $\pi^* \in \Pi_n$ and any matrix $M \in \mathbb{C}_{\mathrm{iso}}$.*

$$\mathbb{E}[\|(\hat{M}_{\mathrm{iso}})_{\hat{\pi}} - M\|_F^2] \le C''\log^{C''}(nd)\rho_{\mathrm{reco}}(n,d,\lambda) \ .$$

As a consequence, the polynomial-time estimator $\hat{M}_{\mathrm{iso}}$ achieves the optimal risk for all values of n and d. For $\lambda = 1$, the optimal risk $\rho_{\mathrm{reco}}(n,d,1)$ is of the order of $n^{1/3}d + n$. In particular, our risk bound strictly improves over the one of Flammarion et al. [33] - e.g. their procedure achieves the estimation error $n\sqrt{d}$ for $n \ge d^{1/3}$. Their slower convergence rates are mainly due to the fact that their estimator of the permutation π^* is suboptimal in this regime.

4.2.4 Polynomial-time reconstruction in the bi-isotonic model

We now turn our attention to the problem of estimating the matrix M when M satisfies the additional assumption of being bi-isotonic up to unknown permutations π^* and η^* of its rows and columns respectively. In other words, the matrix $M_{\pi^{*-1}\eta^{*-1}}$ has non-decreasing entries. As explained in the introduction, this model has attracted a lot of attention in the last decade and encompasses the SST model for tournament problems.

To simplify the exposition, we focus in this section on the case $n = d$ and $\lambda \in [\frac{1}{n}, 1]$. Since the bi-isotonic model is a specific case of the isotonic model, we could rely on the estimator $\widehat{M}_{\mathrm{iso}}$ introduced in the previous subsection. In fact, we can improve this estimation rate by relying on the bi-isotonicity of the matrix $M_{\pi^{*-1}\eta^{*-1}}$.

As previously, we choose the tuning parameters of Algorithm 14 in such a way that $T = 4\lceil\bar{\gamma}^6\rceil$ and $\bar{\gamma}^6 \le C'\log^{C'}(nd/\delta)$. Then, we use the following procedure:

1. Subsample the data into 3 independent samples $(y_t^{(1)})$, $(y_t^{(2)})$, $(y_t^{(3)})$.
2. Run our procedure Algorithm 14 to obtain an estimator $\hat{\pi}$ of the permutation π^* of the rows, using the first sample.
3. Run again Algorithm 14 to obtain an estimator $\hat{\eta}$ of the permutation η^* of the columns, using the second sample.
4. Compute the least-square estimator $\hat{M}_{\mathrm{biso}} = \arg\min_{\tilde{M}\in\mathbb{C}_{\mathrm{biso}}} \|\tilde{M} - Y^{(3)}_{\hat{\pi}^{-1}\hat{\eta}^{-1}}\|_2^2$, where $\mathbb{C}_{\mathrm{biso}}$ is the set of all bi-isotonic matrices with entries in $[0,1]$ and $Y^{(3)}_{ik} = \frac{1}{\lambda}\sum_t y_t^{(3)}\mathbf{1}_{x_t^{(3)}=(i,k)}$.

The following corollary states that $\hat{M}_{\mathrm{biso}}$ achieves a reconstruction rate of order $n^{7/6}\lambda^{-5/6}$ in the bi-isotonic model.

Corollary 4.2.5. *Assume that $\lambda \in [1/n, 8n^2]$. There exists a universal constant C'' such that*

$$\sup_{\substack{\pi^*,\eta^*\in\Pi_n \\ M:\, M_{\pi^{*-1}\eta^{*-1}}\in\mathbb{C}_{\mathrm{biso}}}} \mathbb{E}\left[\|(\hat{M}_{\mathrm{biso}})_{\hat{\pi}\hat{\eta}} - M\|_F^2\right] \le C''\log^{C''}(n)n^{7/6}\lambda^{-5/6} \ .$$

Here, we have fixed $n = d$ to simplify the exposition, but we could extend the analysis to general n and d. Our risk bound improves over the rate $n^{5/4}\lambda^{-3/4}$ of Mao et al. [60]. In [56], Liu and Moitra have introduced a procedure achieving the rate $n^{7/6}$ in the specific case where $\lambda = n^{o(1)}$. In some way, our procedure generalizes their results for general λ, while being applicable to the more general isotonic models.

Still, we recall that the optimal risk (without computational constraints) for estimating the matrix M is of the order n/λ – see e.g. [83, 60]. This remains an open problem to establish the existence of a computational-statistical gap or to construct a polynomial-time procedure achieving this risk on SST and bi-isotonic models.

4.3 Description of the ISR procedure

4.3.1 Weighted directed graph $\mathcal{W}$ and estimator $\hat{\pi}$

Our approach involves the iterative construction of a weighted directed graph $\mathcal{W}$, represented by an antisymmetric matrix in $\mathbb{R}^{n\times n}$. More formally, for any experts i, j in $[n]$, we have $\mathcal{W}(i,j) = -\mathcal{W}(j,i)$. In a nutshell, $\mathcal{W}(i,j)$ quantifies our evidence of the comparisons between expert i and expert j. If $\mathcal{W}(i,j)$ is large and positive (resp. negative), we are confident that the expert i is above (below) the expert j. Most of the procedure is dedicated to the construction of $\mathcal{W}$. Before this, let us explain how we deduce our estimator $\hat{\pi}$ from $\mathcal{W}$.

For a given weighted directed graph $\mathcal{W}$, we define its corresponding directed graph at threshold $\gamma > 0$ as

$$\mathcal{G}(\mathcal{W},\gamma) = \{(i,j) \in [n]^2 \ : \ \mathcal{W}(i,j) > \gamma\} \ . \tag{4.13}$$

For any thresholds $\gamma < \gamma'$, it holds that $\mathcal{G}(\mathcal{W},\gamma) \subset \mathcal{G}(\mathcal{W},\gamma')$. In other words, the function $\gamma \to \mathcal{G}(\mathcal{W},\gamma)$ is nondecreasing. When $\gamma \geq \max_{i,j} |\mathcal{W}(i,j)|$, $\mathcal{G}(\mathcal{W},\gamma) = \varnothing$ is the trivial graph with no edges. Let $\hat{\gamma}$ be the smallest threshold γ such that $\mathcal{G}(\mathcal{W},\gamma)$ is a directed acyclic graph (**DAG**). By monotonicity, $\mathcal{G}(\mathcal{W},\hat{\gamma})$ is also the largest **DAG** among $\{\mathcal{G}(\mathcal{W},\gamma), \gamma \geq \hat{\gamma}\}$. We then build the estimator $\hat{\pi}$ by picking any permutation that is consistent with the graph $\widehat{\mathcal{G}} := \mathcal{G}(\mathcal{W},\hat{\gamma})$, that is if $(i,j) \in \widehat{\mathcal{G}} \cap [n]^2$ then $\hat{\pi}(i) \geq \hat{\pi}(j)$. To put it another way, the general idea of our procedure can be summarized into these three components:

1. Construct a weighted directed graph $\mathcal{W}$ between the experts.
2. Compute the largest directed acyclic graph $\widehat{\mathcal{G}}$ of $\mathcal{W}$.
3. Take any arbitrary permutation $\hat{\pi}$ that is consistent with $\widehat{\mathcal{G}}$.

The construction of $\mathcal{W}$ is at the core of this chapter, and the computation of $\widehat{\mathcal{G}}$ and $\hat{\pi}$ will be discussed in Section 4.3.7. Still, we already point out that the third point can be dealt in polynomial time using Mirsky's algorithm [64].

4.3.2 Construction of $\mathcal{W}$ with ISR

4.3.2.1 Description of the subsampling

Let us now describe the construction of the weighted directed graph $\mathcal{W}$. Let $T \geq 1$ be an arbitrary integer, representing the number of steps of our procedure. In what follows, we explain how we subsample the data from (4.2) into $5T$ independent matrices $(Y^{(s)})_{s=1\dots 5T}$. Recall that we are given N observations (x_t, y_t), where N follows a Poisson distribution $\mathcal{P}(\lambda nd)$. Let us divide the observations into $5T$ batches $(N^{(s)})_{s=0,\dots,5T-1}$, aggregated into matrices of averaged observations $Y^{(s)}$. To that end, we let S_u be i.i.d. uniform random variables in $\{0,\dots,5T-1\}$ representing a random batch for observation u, and we define

$$N^{(s)} = \{u \in \{1\dots,N\} \ : \ S_u = s\} \quad \text{and} \quad Y^{(s)}_{ik} = \sum_{t\in N^{(s)}} \frac{y_t}{\mathbf{r}^{(s)}_{ik} \vee 1} \mathbf{1}\{x_t = (i,k)\} \ , \tag{4.14}$$

where, for any $(i,k) \in [n]\times[d]$, $\mathbf{r}^{(s)}_{ik} = \sum_{t\in N^{(s)}} \mathbf{1}\{x_t = (i,k)\}$ is the number of times the coefficient position (i,k) is observed in batch s. $Y^{(s)}_{ik}$ is equal to 0 if (i,k) is not observed in batch s and it is equal to the average of the observations y_t for which $x_t = (i,k)$ otherwise. We also define the mask matrix $B^{(s)}$ as being equal to 0 at location (i,k) if the value is missing from batch s, and to 1 otherwise.

$$B^{(s)}_{ik} = \mathbf{1}\{\mathbf{r}^{(s)}_{ik} \geq 1\} \ . \tag{4.15}$$

Define $\lambda_0 = \lambda/5T$. In our sampling scheme, where the data is divided into $5T$ samples, each coefficient $B^{(s)}_{ik}$ has a probability of $1 - e^{-\lambda_0}$ of being equal to one. It is worth mentioning that a different subsampling scheme was performed in [74], consisting in aggregating consecutive columns. However, such a scheme is not applicable in our case as we do not assume the rows of M to be nondecreasing, unlike in [74].

4.3.2.2 Neighborhoods in comparison graphs

At each step $t = 0,\dots,T-1$ of the procedure, we aim to enrich our knowledge of the order of the experts, which we formally do by nondecreasing the weights of $\mathcal{W}$ in absolute value. At $T = 0$, we start with the weights $\mathcal{W}_{ij}$ all being equal to zero. A meaningful update of $\mathcal{W}$ around a reference expert i can be done when we restrict ourselves to experts that are in a neighborhood of i. Broadly speaking, a neighborhood of i is a set made of all the experts j that are not comparable to i with respect to a given partial order.

More precisely, for any directed graph $\mathcal{G}$ and any experts $i, j \in \{1, \dots, n\}$, we say that i and j are $\mathcal{G}$-comparable if there is a path from i to j or from j to i in $\mathcal{G}$. The neighborhood $\mathcal{N}(\mathcal{G}, i)$ of i in $\mathcal{G}$ can then naturally be defined as the set of experts j that are not $\mathcal{G}$-comparable with i. Equipped with the concept of neighborhood, our overall strategy involves iterating over all possible thresholds $\gamma \in \Gamma$ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic, as well as all possible experts i. At each iteration, we apply the soft local ranking procedure Algorithm 15 described in the next subsection. Algorithm 15 updates the weights between i and any expert j in the neighborhood $\mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma), i)$ of i. Our approach can be summarized as follows:

1. Subsample the data - see Section 4.3.2.1.
2. Initialize $\mathcal{W}$ to be the directed graph with all weights set to 0.
3. For all $t = 0, \dots, T-1$ and $\gamma \in \Gamma$ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic and all $i \in [n]$, update $\mathcal{W}$ with the soft local ranking procedure Algorithm 15.

Algorithm 14 **ISR**(T, Γ)

Require: N and observations $(x_t, y_t)_{t=1,\dots,N}$ according to (4.2), a number of steps T and a valid grid Γ as in (4.10)
Ensure: A weighted graph $\mathcal{W}$ and an estimator $\hat{\pi}$
1: Aggregate the observation into $5T$ matrices of observation $(Y^{(s)})$ as in (4.14)
2: Initialize $\mathcal{W}(i, j) = 0$ for all $(i, j) \in [n]^2$, and $\hat{\gamma} = 0$
3: **for** $t = 0, \dots, T-1$ **do**
4: **for** $\gamma \in \Gamma \cap [\hat{\gamma}, +\infty)$ **do**
5: Compute $\mathcal{G} = \mathcal{G}(\mathcal{W}, \gamma)$ the directed graph at threshold γ of $\mathcal{W}$ as in (4.13) and set $P = \mathcal{N}(\mathcal{G}, i)$.
6: Take 5 samples $\mathcal{Y} = (Y^{(5t)}, \dots, Y^{(5t+4)})$
7: **for** $i \in [n]$ **do**
8: Apply **SLR**$(\mathcal{Y}, \mathcal{W}, \gamma, i, \mathcal{G}, P)$ to update $\mathcal{W}$
9: **end for**
10: **end for**
11: Set $\hat{\gamma}$ as the smallest γ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic
12: **end for**
13: Set $\widehat{\mathcal{G}} = \mathcal{G}(\mathcal{W}, \hat{\gamma})$ be the largest acyclic **DAG** (see (4.13))
14: Set $\hat{\pi}$ to be any arbitrary permutation that is consistent with $\widehat{\mathcal{G}}$
15: **return** $\mathcal{W}$ and $\hat{\pi}$

The main Line 8 of Algorithm 14 aims to provide a soft ranking of the neighborhood P of i by setting positive (resp. negative) weights $\mathcal{W}_{ij}$ to experts $j \in P$ that are significantly below (resp. above) i. Line 11 together with restricting $\gamma \geq \hat{\gamma}$ simply guarantees that all the considered graph $\mathcal{G}$ are acyclic. Finally, Lines 13 and 14 simply correspond to the construction of the final permutation, described in the second and third points of Section 4.3.1.

4.3.3 Description of the updating procedure

4.3.3.1 Local weighted sums

Let us describe the process of updating a given weighted graph $\mathcal{W}$, which will be used twice at each call of the soft local ranking Algorithm 15. Let us fix a weighted graph $\mathcal{W}$, an element $s \in \{0, \dots, 5T-1\}$ and $Y := Y^{(s)}$ the matrix defined in (4.14). We also let $i \in [n]$ be an arbitrary expert corresponding to Line 7 of Algorithm 14, and γ be any threshold in the grid Γ. We write $P := \mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma), i) \subset [n]$ for the neighborhood of i in $\mathcal{G}(\mathcal{W}, \gamma)$, echoing the notation of the sets that are trisected in [74].

Since the matrix M is, up to a row-permutation, a column-wise isotonic matrix, it follows that, if the expert i is above j, then for any vector $w \in \mathbb{R}^d_+$, we have $\sum_{k=1}^d w_{ik} M_{ik} \geq \sum_{k=1}^d w_{jk} M_{jk}$. As a consequence, the crux of the algorithm is to find suitable data-driven weights w that allow to discriminate the experts. As explained in the introduction, earlier works focused on uniform weights $w = \mathbf{1}_{[d]}$ [83] which, unfortunately leads to suboptimal results. Before discussing the choice of the weights w in the following subsections, let us first formalize how we leverage on w to compare the experts and update the graph $\mathcal{W}$.

Given a subset $Q \subset [d]$ of columns and a non-zero vector $w \in \mathbb{R}^Q_+$, we first check whether the following condition is satisfied:

$$\lambda_0 \|w\|_2^2 \geq \|w\|_\infty^2 \ , \tag{4.16}$$

where we recall that $\lambda_0 = \lambda/5T$. This condition is always verified when $\lambda_0 \geq 1$, and it is equivalent to $\lambda_0|Q| \geq 1$ when $w = \mathbf{1}_Q$. Condition (4.16) ensures that w is not too sparse which could be harmful when many observations are lacking (λ_0 small).

If this condition is not satisfied, then we leave the weights of $\mathcal{W}$ unchanged. Otherwise, we define the (Y, P, w)-updating weights $\mathcal{U} := \mathcal{U}(Y, P, w)$ around i as

$$\mathcal{U}_{ij} = \frac{1}{\sqrt{\frac{1}{\lambda_0} \wedge \lambda_0}} \cdot \langle Y_{i\cdot} - Y_{j\cdot}, \frac{w}{\|w\|_2} \rangle \ , \tag{4.17}$$

where, for all $w' \in \mathbb{R}^Q$ and $a \in \mathbb{R}^d$, we write $\langle a, w' \rangle = \sum_{k \in Q} a_k w'_k$. We can then update the weighted directed graph around i by setting, for all $i \in P$ such that $|\mathcal{U}_{ij}| \geq |\mathcal{W}_{ij}|$,

$$\mathcal{W}_{ij} = \mathcal{U}_{ij} \quad \text{and} \quad \mathcal{W}_{ji} = -\mathcal{U}_{ij} \ . \tag{4.18}$$

As explained above, if we replace $Y_{i\cdot}$ and $Y_{j\cdot}$ by $M_{i\cdot}$ and $M_{j\cdot}$ respectively in (4.17), then the corresponding value of the statistic is non-negative if expert i is above j. Hence, a large value for $\mathcal{U}_{ij}$ provides evidence that i is above j.

Computing $\mathcal{U}(Y, P, w)$ for suitable directions w is the basic brick or our procedure, since it is through the update (4.18) that we iteratively increase the weights of $\mathcal{W}$. This update shares some similarities to the pivoting algorithm introduced in [56] and also used in [74], in the sense that while we are fixing an arbitrary reference expert i to compute pairwise comparisons, they fix a set P and compute a pivot expert i_0 that would correspond to a quantile of the set $\{\langle Y_{j\cdot}, \frac{w}{\|w\|_2} \rangle, j \in P\}$ in the case $\lambda_0 = 1$.

Note that the orientation of a given weighted edge (i, j) can change during the procedure if it turns out that $|\mathcal{U}_{ij}| \geq |\mathcal{W}_{ij}|$ and that $\mathcal{U}_{ij}\mathcal{W}_{ij} \leq 0$. This simply means that if the direction w leads to a more significant weight between some experts i and j, then we are more confident to use the vector w and to revise the order between i and j.

For $Q \subset [d]$, choosing $w = \mathbf{1}_Q$ in (4.17) amounts to compute the average of the observations over all questions in Q. We now explain in the main sections how we iteratively build adaptive weights w that allow to improve over the naive global average given by $w = \mathbf{1}_{[d]}$.

4.3.3.2 Definitions of a rank in a DAG

We first introduce a few definitions on directed acyclic graphs $\mathcal{G}$, which we formally define as a set of directed edges $(i, j) \in [n]^2$ for which there is no cycle. We denote $\mathbf{path}(i, j) = \{(k_1, \dots, k_L) \ : \ L > 0 \text{ and } (i, k_1), \dots, (k_L, j) \in \mathcal{G}\}$ as the set of all possible paths from i to j, and we write $|s|$ for the length of any path s. We say that i and j are $\mathcal{G}$-comparable if $\mathbf{path}(i, j) \cup \mathbf{path}(j, i) \neq \varnothing$, and we write $\mathcal{N}(i, \mathcal{G})$ for the set of all experts that are not $\mathcal{G}$-comparable with i. If i, j are $\mathcal{G}$-comparable, it either holds that $\mathbf{path}(i, j) = \varnothing$ or $\mathbf{path}(j, i) = \varnothing$. We say in the first case that i is $\mathcal{G}$-below j and that i is $\mathcal{G}$-above j in the second case. We also define the relative rank from i according to $\mathcal{G}$ as the length of the longest path in $\mathcal{G}$ from i to j, or minus the longest past from j to i depending on whether i is $\mathcal{G}$-above or $\mathcal{G}$-below j:

$$\mathbf{rk}_{\mathcal{G},i}(j) = \max\{|s| \ : \ s \in \mathbf{path}(i, j)\} - \max\{|s| \ : \ s \in \mathbf{path}(j, i)\} \ . \tag{4.19}$$

Here, we use the convention $\max \varnothing = 0$. With this definition, the neighborhood of a given expert i is equal to the set of experts whose relative rank is equal to 0, that is $\mathcal{N}(\mathcal{G}, i) = \mathbf{rk}^{-1}_{\mathcal{G},i}(0)$. Moreover, an expert $j \in [n]$ is $\mathcal{G}$-above (resp. $\mathcal{G}$-below) i if and only if $\mathbf{rk}_{\mathcal{G},i}(j) \geq 1$ (resp. $\mathbf{rk}_{\mathcal{G},i}(j) \leq -1$). Although $\mathcal{G}$ stands for a finite set of edges with endpoints in $[n]$, we extend it to a set of edges with endpoints in $\mathbb{Z}^2$ by putting in $\mathcal{G}$ every $(i, j) \in \mathbb{Z}^2$ such that $i > j$ and $j \leq 0$ or $i \geq n + 1$.

4.3.3.3 Description of the soft local ranking algorithm

To update the weighted directed graph $\mathcal{W}$ in Line 8 of Algorithm 14, we apply the soft local ranking procedure **SLR** to all experts $i \in [n]$ and all thresholds γ. To define our soft local ranking procedure, let us fix $\mathcal{W}$, an expert i and a threshold γ such that $\mathcal{G}(\mathcal{W}, \gamma)$ is acyclic. As a shorthand, we write $\mathcal{G}$ and P respectively for the thresholded graph $\mathcal{G}(\mathcal{W}, \gamma)$ and the neighborhood $\mathcal{N}(\mathcal{G}, i)$ of i in $\mathcal{G}$.

We write $\mathcal{D}$ for the set of all dyadic numbers: $\mathcal{D} = \{2^k \ : \ k \in \mathbb{Z}\}$ and we define the set $\mathcal{H} = \mathcal{D} \cap \left[\frac{1}{nd}, 1\right]$. We denote $\overline{y}(P)$ as the mean of the vectors $Y_{j\cdot}$ over all $j \in P$, that is $\overline{y}_k(P) = \frac{1}{|P|} \sum_{j \in P} Y_{jk}$, for any $k \in [d]$. **SLR** relies on the following steps repeated over all height $h \in \mathcal{H}$. It is also described in Algorithm 15.

1. **Dimension reduction.** Using the first sample $Y^{(1)}$, we first reduce the dimension by selecting a subset $\widehat{Q}^h \subset [d]$ corresponding to wide confidence regions. Recall that $\mathbf{rk}_{\mathcal{G},i}$ is the relative rank to i defined in

(4.19). For any $a > 0$, define the sets $\mathcal{N}_a := \mathcal{N}_a(\mathcal{G}, i)$ (resp. $\mathcal{N}_{-a} := \mathcal{N}_{-a}(\mathcal{G}, i)$) of experts j which are $\mathcal{G}$-above (resp. $\mathcal{G}$-below) all the experts of P and whose relative rank to any $i' \in P$ is at most a in absolute value:

$$\mathcal{N}_a = \bigcap_{i' \in P} \mathbf{rk}^{-1}_{\mathcal{G},i'}([1,a]) \quad \text{and} \quad \mathcal{N}_{-a} = \bigcap_{i' \in P} \mathbf{rk}^{-1}_{\mathcal{G},i'}([-1,-a]) \ . \tag{4.20}$$

Secondly, we define for any question $k \in [d]$ and $a \geq 1$ the width statistic $\widehat{\mathbf{\Delta}}_k$ as the difference between the mean of the experts in $\mathcal{N}_a$ and the mean of the experts in $\mathcal{N}_{-a}$. Then, $\hat{a}_k$ is set to be the first value of $a \geq 1$ such that any $a' \geq a$ has a corresponding width statistic of at least $(\lambda_0 \wedge 1)h$:

$$\widehat{\mathbf{\Delta}}_k(a) = \overline{y}_k(\mathcal{N}_a) - \overline{y}_k(\mathcal{N}_{-a}) \quad \text{and} \quad \hat{a}_k(h) = \max\left\{a \geq 1 \ : \ \frac{1}{\lambda_0 \wedge 1}\widehat{\mathbf{\Delta}}_k(a) < h\right\} + 1 \ . \tag{4.21}$$

Finally, we define $\widehat{Q}^h := \widehat{Q}^h(\mathcal{G}, i)$ as the set of indices k such that $\hat{a}_k(h)$ is relatively small.

$$\widehat{Q}^h = \{k \in [d] \ : \ |\mathcal{N}_{\hat{a}_k(h)}| \wedge |\mathcal{N}_{-\hat{a}_k(h)}| \leq \frac{1}{\lambda_0 h^2}\} \ . \tag{4.22}$$

Intuitively, if the experts above and below i vary by more than h on a specific question k, then this question should belong to $\widehat{Q}^h$. Conversely, if the experts below and above i are nearly equal on the question k, than $\hat{a}_k(h)$ will be large and k will not be selected in $\widehat{Q}^h$.

2. **Average-based weighted sums.** Still using the first sample $Y^{(1)}$, we examine the corresponding submatrix $Y^{(1)}(P, \widehat{Q})$ restricted to questions in $\widehat{Q}$. If the row sums of Y are larger than the current edges, we update the weighted edges. More formally, we compute the $(Y^{(1)}, P, \mathbf{1}_{\widehat{Q}})$-updating weighted edges $(\mathcal{U}_{\widehat{Q}})$ around i as defined in (4.17) and update $\mathcal{W}$ as in (4.18). We then also update $\mathcal{G} = \mathcal{G}(\mathcal{W}, \gamma)$ and $P = \mathcal{N}(\mathcal{G}, i)$.

3. **PCA-based weighted sums.** Relying on the samples $Y^{(2)}$, $Y^{(3)}$, $Y^{(4)}$, $Y^{(5)}$, we do a slight abuse of notation and write $Y^{(s)}$ for the restriction of $Y^{(s)}$ to the subset $P, \widehat{Q}^h$ for $s = 2, 3, 4, 5$. Ideally, we would get an informative direction w from the largest right singular vector of $\mathbb{E}[Y^{(2)} - \overline{Y}^{(2)}] \in \mathbb{R}^{P \times \widehat{Q}^h}$. Indeed, it is known (see the proofs for more details) that the entries of the first right singular vector of an isotonic matrix all share the same sign and are most informative to compare the experts. However, computing directly the empirical right-singular vector of $Y^{(2)} - \overline{Y}^{(2)}$ does not lead to the desired bounds because (i) this matrix is perhaps highly rectangular (ii) the noise is possibly heteroskedastic and (iii) this matrix is perhaps sparse because of the many missing observations when λ_0 is small. Here, we use a workaround which is reminiscent of that of [74] and discussed later. First, we compute $\hat{v}$ as a proxy for the first left singular vector of $\mathbb{E}[Y^{(2)} - \overline{Y}^{(2)}]$.

$$\hat{v} := \hat{v}(P, \widehat{Q}^h) = \operatorname*{arg\,max}_{v \in \mathbb{R}^P : \ \|v\|_2 \leq 1} \left[\|v^T(Y^{(2)} - \overline{Y}^{(2)})\|_2^2 - \frac{1}{2}\|v^T(Y^{(2)} - \overline{Y}^{(2)} - Y^{(3)} + \overline{Y}^{(3)})\|_2^2\right] \ . \tag{4.23}$$

The right-hand side term in (4.23) deals with the heteroskedasticity of the noise matrix E in (4.3). $\hat{v}$ in (4.23) can be computed efficiently since it corresponds to the leading eigenvector of a symmetric matrix. For technical reasons occurring in the sparse observation regime (i.e. when λ_0 is small), we then threshold the largest absolute values of the coefficients of $\hat{v}$ at $\sqrt{\lambda_0}$ and define $(\hat{v}_-)_i = \hat{v}_i \mathbf{1}\{|\hat{v}_i| \leq \sqrt{\lambda_0}\}$. After having calculated $\hat{v}_-$, we consider as in [74] the image $\hat{z} = \hat{v}_-^T(Y^{(4)} - \overline{Y}^{(4)}) \in \mathbb{R}^{\widehat{Q}}$ of $\hat{v}_-$. We then threshold the smallest values of $\hat{z}$ and take the absolute values of the components. Thus, we get $\hat{w}^+ \in \mathbb{R}^{\widehat{Q}}$ defined by $(\hat{w}^+)_l = |\hat{z}_l| \mathbf{1}\{|\hat{z}_l| \geq \gamma\sqrt{\lambda_0 \wedge \frac{1}{\lambda_0}}\}$ for any $l \in \widehat{Q}$.

Finally, we consider the last submatrix $Y^{(5)} = Y^{(5)}(P, \widehat{Q})$. We apply these weights $\hat{w}^+$ to compute the row-wise weighted sums of $Y^{(5)}$ and update the weighted edges. More formally, we compute the $(Y^{(5)}, P, \hat{w}^+)$-updating weighted edges $\mathcal{U}(Y^{(5)}, P, \hat{w})$ around i as defined in (4.17). We finally update the weighted directed graph $\mathcal{W}$ with $\mathcal{U}(Y^{(5)}, P, \hat{w}^+)$ as in (4.18).

Algorithm 15 SLR$((Y^{(s)})_{s=1,\dots,5}, \mathcal{W}, \gamma, i, \mathcal{G}, P)$

Require: 6 samples $(Y^{(s)})_{s=1,\dots,5}$, a weighted directed graph $\mathcal{W}$, a threshold γ such that $\mathcal{G}(\mathcal{W},\gamma)$ is acyclic and an expert $i \in [n]$. $\mathcal{G}$ and P are shorthands for the thresholded graph $\mathcal{G}(\mathcal{W},\gamma)$ and the neighborhood $\mathcal{N}(\mathcal{G},i)$.
Ensure: An update of $\mathcal{W}$

for $h \in \mathcal{H}$ **do**
  Compute $\widehat{Q}^h := \widehat{Q}(\mathcal{G},i)$ as in (4.22) using sample $Y^{(1)}$
  Let $\mathcal{U}_{\widehat{Q}^h}$ be the $(Y^{(1)}, P, \mathbf{1}_{\widehat{Q}^h})$-updating weighted edges around i as in (4.17), using again sample $Y^{(1)}$
  Update $\mathcal{W}$ with $\mathcal{U}(\widehat{Q}^h)$ as in (4.18) and update $\mathcal{G} = \mathcal{G}(\mathcal{W},\gamma)$, $P = \mathcal{N}(\mathcal{G},i)$
  Restrict the samples $(Y^{(s)})_{s=2,\dots,5}$ to $P, \widehat{Q}^h$ in the following remaining steps
  Compute the PCA-like direction $\hat{v} := \hat{v}(P,\widehat{Q}^h)$ as in (4.23) and define $(\hat{v}_-)_i = \hat{v}_i \mathbf{1}\{|\hat{v}_i| \le \sqrt{\lambda_0}\}$
  Compute $\hat{z} = \hat{v}_-^T(Y^{(4)} - \overline{Y}^{(4)})$ and define $\hat{w}^+$ by $(\hat{w}^+)_l = |\hat{z}_l| \mathbf{1}\{|\hat{z}_l| \ge \gamma\sqrt{\lambda_0 \wedge \frac{1}{\lambda_0}}\}$ for any $l \in \widehat{Q}^h$
  Let $\mathcal{U}(Y^{(5)}, \hat{w}^+)$ be the $(Y^{(5)}, P, \hat{w}^+)$-updating weighted edges around i as in (4.17)
  Update $\mathcal{W}$ with $\mathcal{U}(Y^{(5)}, \hat{w}^+)$ as in (4.18)
end for

4.3.4 Toy example illustrating Algorithm 15

To understand why the steps described in Algorithm 15 are relevant, assume that $\pi^* = \mathrm{id}$ and consider the following simple example where $n = 204$, $d = 10$, and where the isotonic matrix $M_{\pi^{*-1}}$ can be decomposed into three blocks of rows as

$$M_{\pi^{*-1}} = \alpha + \frac{h}{2}\begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & 0 & -1 & -1 \\ \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} \end{pmatrix}.$$

In the above matrix, α is any number in $(h, 1-h)$, and $\mathbf{0}, \mathbf{1}$ are the columns in $\mathbb{R}^{100}$ whose coefficients are respectively all equal to 0 and 1. Assume that the statistician already knows that the first and the third blocks are made of experts that are respectively above and below the second block. If $\mathcal{W}, P, \gamma$ are the parameters fixed in Algorithm 15, the three blocks correspond respectively to the subsets $\mathcal{N}_1 \cup \mathcal{N}_2$, P and $\mathcal{N}_{-1} \cup \mathcal{N}_{-2}$ in our example. Provided that $\mathcal{N}_{-2}$ and $\mathcal{N}_2$ are large enough, the set $\widehat{Q}^h$ only keeps columns corresponding to indices k where $\widehat{\mathbf{\Delta}}_k(1)$ is large – those are highlighted in blue.

Then, we can work on the reduced subset $\widehat{Q}^h$ of columns highlighted in blue. As one may check, $\widehat{Q}^h$ contains all the relevant columns to decipher the experts in the block P. Besides, the expected matrix of observations restricted to the block P and to $\widehat{Q}^h$ is of rank one:

$$\mathbb{E}[Y - \overline{Y}] = \frac{h}{2}\begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & -1 & -1 & 0 & -1 & -1 \\ 0 & -1 & -1 & 0 & -1 & -1 \end{pmatrix}.$$

In particular, the right singular vector of this matrix is of the form $(0,1,1,0,1,1)$ and provides suitable weights to decipher the two largest experts from the two lowest experts in the above matrix. The PCA-based weighted sums steps above precisely aims at estimating these weights.

4.3.5 Comments on the procedure and relation to the literature

Finding confidence regions $\widehat{Q}$ before computing weighted sums on the corresponding columns is at the core of our procedure. This idea generalizes the RankScore procedure of [33] which rather computes averages on the subsets $[d]$ or on the singletons $\{1\}, \dots, \{d\}$. As mentioned in the introduction, only using the subsets of the RankScore method in [33] does not allow to reach the optimal rate for permutation estimation or matrix reconstruction.

In Algorithm 15, the computation of subsets $\widehat{Q}^h$ is reminiscent of some aspects of the non oblivious trisection procedure used in [74] for the bi-isotonic model. In fact, the statistic $\widehat{\mathbf{\Delta}}_k$ corresponds to the statistic $\widehat{\mathbf{\Delta}}^{(\mathrm{ext})}_{k,1}$ in [74]. Apart from that, the selection of subsets of questions was quite different in [74] as it mostly involved

change-point detection ideas as introduced in [56]. However, those ideas are irrelevant in our setting because the rows do not exhibit any specific structure in the isotonic model.

The high-level sorting method in [74] is based on a hierarchical sorting tree with memory. In contrast, our new algorithm is based on an iterative refinement of a weighted comparison graph. This new algorithm is more natural and benefits from the fact that it is almost free of any tuning parameter. Indeed, at the end of Algorithm 14, we simply use the threshold $\hat{\gamma}$ corresponding to the largest acyclic $\hat{\mathcal{G}}$ graph in $\mathcal{W}$. No significant threshold needs to be chosen, since any permutation that is consistent with $\hat{\mathcal{G}}$ is also necessarily consistent with $\mathcal{W}$ thresholded at values larger than $\hat{\gamma}$.

The spectral step in [74] is quite similar to the third step of our procedure described in section 4.3.3.3, except for the first thresholding of $\hat{v}$ to obtain $\hat{v}_-$. In [74], this workaround was not needed mainly because in the bi-isotonic model, it is possible to aggregate sparse observations by merging consecutive columns – see [74] for further details. This is however not possible here.

As mentioned in the introduction, Liu and Moitra [56] obtain an upper bound of the permutation loss of the order of $n^{7/6}$ for the estimation of two unknown permutations in the case where $M \in \mathbb{R}^{n\times n}$ is bi-isotonic. Broadly speaking, their method involves iterating a clustering method called block-sorting over groups of rows or columns that are close with each other. Using this sorting method based on block-sorting, their whole approach alternates between row sorting and column sorting for a subpolynomial number of time. Besides, their procedure makes heavily use of bi-isotonicity of the matrix. It turns out that Algorithm 15 reaches the same rate in this bi-isotonic model by running only once on the rows, and once on the columns, as described in Section 4.2.4. In other words, if the problem is to estimate only π^* in the bi-isotonic model, we proved that only the isotonicity of the columns is necessary to achieve the state-of-the-art polynomial-time upper bound of order $n^{7/6}$.

4.3.6 Examples of valid grids Γ

Remark that the simple set $\{(u+1)\cdot\phi_{\mathrm{L}_1}, u \in \{0,\dots,2\lfloor\log_2(n)\rfloor+2\}\}$ is a valid grid of logarithmic size with $\bar{\gamma} \le (2\log_2(n)+3)\phi_{\mathrm{L}_1}$. This set is the smallest valid grid achieving the smallest possible value of $\bar{\gamma}$. However, it depends on the quantity ϕ_{L_1} which is perhaps a bit pessimistic in practice.

An other choice can be to take $\mathbb{R}^+$ itself, albeit infinite. Indeed, the set $\{\mathcal{G}(\mathcal{W},\gamma), \gamma \ge 0\}$ is made of at most n^2 possible directed graphs for any $\mathcal{W}$ during the whole procedure. Choosing $\mathbb{R}^+$ is convenient since it does not depend on the constants in ϕ_{L_1} that are likely to be overestimated. The drawback of choosing $\mathbb{R}^+$ though is that the number of tested γ in Algorithm 15 becomes quadratic in n.

Finally, a good compromise is to take the set $\{(1+\frac{1}{\log_2(n)})^{u'},\ u' \in \mathbb{Z}\}$. It is easy to check that it contains a sequence satisfying (4.10) whose length is at least $2\lfloor\log_2(n)\rfloor+3$ and whose maximum $\bar{\gamma}$ is a polylogarithmic function in nd/δ.

4.3.7 Discussion on the computation of $\widehat{\mathcal{G}}$ and $\hat{\pi}$

Once we have suitable weighted graph $\mathcal{W}$, it remains to construct the permutation $\hat{\pi}$, as in the second and third point of Section 4.3.1.

For the second point, checking that a given directed graph is acyclic can be done through depth first search with a computational complexity less than n, so that computing $\hat{\gamma}$ can be done with less than $|\Gamma|n$ operations. As discussed in Section 4.3.3, it is possible to choose Γ to be of size of order less than $\log(n)$. If Γ is bounded and is such that any different thresholds γ,γ' in Γ satisfy $|\gamma-\gamma'| \ge \eta$ for some $\eta > 0$, the computation of $\hat{\gamma}$ can always be done with complexity of order less than $n\log(\max(\Gamma)/\eta)$.

Regarding the third point, a permutation $\hat{\pi}$ can be computed in polynomial time from the directed acyclic graph $\hat{\mathcal{G}}$ using Mirsky's algorithm [64] – see also [72]. It simply consists in finding the minimal experts i in $\hat{\mathcal{G}}$, removing them and repeat this process. This construction is in fact equivalent to ranking the experts according to the index $\mathbf{rk}_{\hat{\mathcal{G}},0}$ as defined in (4.19).

4.4 Concentration inequality for rectangular matrices

In this section, we state a concentration inequality for rectangular random matrices with independent entries satisfying a Bernstein-type condition. This section can be read independently of the rest of this chapter. Let p and q be two positive integers and $X \in \mathbb{R}^{p\times q}$ be a random matrix with independent and mean zero coefficients. Assume that there exists $\sigma > 0$ and $K \ge 1$ such that for any $i = 1,\dots,p$ and $k = 1,\dots,q$,

$$\forall u \ge 1, \quad \mathbb{E}[(X_{ik})^{2u}] \le \frac{1}{2}u!\sigma^2 K^{2(u-1)} \ . \tag{4.24}$$

This Bernstein-type condition (4.24) is exactly the same as Assumption 1 in [7] – see [7] for a discussion. Let $\Lambda \in \mathbb{R}^{p\times p}$ be any orthogonal projection matrix, i.e. $\Lambda = \Lambda^T$ and $\Lambda^2 = \Lambda$. We write r_Λ for the rank of Λ.

Proposition 4.4.1. *There exists a positive numerical constant κ such that the following holds for any $\delta > 0$.*

$$\|\Lambda(XX^T - \mathbb{E}[XX^T])\Lambda\|_{\mathrm{op}} \le \kappa\left[\sqrt{(\sigma^4 pq + \sigma^2 q)\log(p/\delta)} + (\sigma^2 r_\Lambda + K^2\log(q))\log(p/\delta)\right] . \tag{4.25}$$

For the sake of the discussion, consider the particular case where $X_{ik} = B_{ik}E_{ik}$, with B_{ik} and E_{ik} being respectively independent Bernoulli random variable of parameter σ^2 and centered Gaussian random variable with variance 1. By a simple computation done e.g. in (4.77), X_{ik} satisfies condition (4.24) with K being of the order of a constant. Hence, if $K^2\log(q) \le \sigma^2 p$, applying Proposition 4.4.1 with the identity matrix Λ gives

$$\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}} \le 2\kappa\sigma^2\left[\sqrt{pq\log(p/\delta)} + p\log(p/\delta)\right] , \tag{4.26}$$

with probability at least $1-\delta$.

Up to our knowledge, the inequality (4.26) is tighter than state-of-the-art result random rectangular sparse matrices in the regime where $q \gg p$ and $\sigma^2 \ll 1$. In fact, most of the results in the literature concerning random matrices state concentration inequalities for the non-centered operator norm $\|XX^T\|_{\mathrm{op}}$ – see the survey of Tropp [88].

More specifically, Bandeira and Van Handel [5] provide tight non-asymptotic bounds for the spectral norm of a square symmetric random matrices with independent Gaussian entries, and derive tail bounds for the operator norm of XX^T. For instance, Corollary 3.11 in [5], implies that, for some numerical constant c, $\mathbb{E}[\|XX^T\|_{\mathrm{op}}^2] \le c(\sigma^2(p\vee q) + \log(p\vee q))$. Together with a triangular inequality, Bandeira and Van Handel imply $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2 \le c\sigma^2((p\vee q) + \log(\frac{p\vee q}{\delta}))$ with probability higher than $1-\delta$.

While the order of magnitude $\sigma^2(p\vee q)$ is tight for controlling the operator norm $\|XX^T\|_{\mathrm{op}}^2$ of the non-centered Gram matrix with high probability, (4.26) implies that the right bound for $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2$ is rather $\sigma^2\sqrt{pq}$ which is significantly smaller in the regime $p \ll q$ and $\sigma^2 \ll 1$.

In the proof of Theorem 4.2.2, we could have used those previous results for controlling the matrices of the form $\|XX^T - \mathbb{E}[XX^T]\|_{\mathrm{op}}^2$. However, we would have then achieved a suboptimal risk upper bound. Indeed, Proposition 4.4.1 plays critical role in the proof of Theorem 4.2.2, when we need to handle matrices with partial observations that are possibly highly rectangular in the spectal step of the procedure (4.23).

The proof of Proposition 4.4.1 relies on the observation that the matrix $XX^T - \mathbb{E}[XX^T]$ is the sum of q centered rank 1 random matrices. This allows us to apply Matrix Bernstein-type concentration inequalities for controlling the operator norm of this sum – see [88] or Section 6 of [94].

4.5 Proofs

4.5.1 Proof of Theorem 4.2.2

4.5.1.1 Notation and signal-noise decomposition

We first introduce some notation, and in particular the noise matrices on which we will apply concentration inequalities. In what follows, we define for any matrix $A \in \mathbb{R}^{n\times d}$, and any vector $w \in \mathbb{R}^d$:

$$\langle A_{i\cdot}, w\rangle = \sum_{k=1}^{d} A_{ik}w_k \ . \tag{4.27}$$

If w belongs to $\mathbb{R}^Q$ where Q is some subset of $[d]$, we also write $< A_{i\cdot}, w >= \sum_{k\in Q}^d A_{ik}w_k$. The same notation stands for the scalar product on matrices, namely $\langle A, A'\rangle = \mathrm{Tr}(A^TA')$ if $A' \in \mathbb{R}^{n\times d}$. If A and A' are two matrices in $\mathbb{R}^{n\times d}$, then we write the coordinate-wise product $(A\odot A')_{ik} = A_{ik}A'_{ik}$. In what follows, we assume that $\pi^* = \mathrm{id}$. We make this assumption without loss of generality since we can reindex each expert i with $i' = \pi^{*-1}(i)$. Recalling that B is defined in (4.15) we define

$$\lambda_1 := \mathbb{P}(B_{ik}^{(s)} = 1) = 1 - e^{-\lambda_0} \ . \tag{4.28}$$

If $\lambda_0 \le 1$, we have $\lambda_0 \ge \lambda_1 \ge (1-\frac{1}{e})\lambda_0$. We assume in what follows that $\lambda_0 \le 1$, which corresponds to the case where there are potentially many unobserved coefficients. The case $\lambda_0 \ge 1$ will be treated in Section 4.5.6. For an observation matrix $Y^{(s)}$ defined in (4.14), we make the difference between $\mathbb{E}[Y^{(s)}] = \lambda_1 M$, which is the unconditional expectation of $Y^{(s)}$, and $\mathbb{E}[Y^{(s)}|B^{(s)}] = B^{(s)}\odot M$, which is the expectation of $Y^{(s)}$ conditionally to the matrix B. We write the noise matrix

$$E^{(s)} = Y^{(s)} - \lambda_1 M \quad \text{and} \quad \widetilde{E}^{(s)} = Y^{(s)} - B^{(s)} \odot M \ . \tag{4.29}$$

Recall that $\varepsilon_t = y_t - M_{x_t}$ is the subGaussian noise part in model (4.2), and that N_s is defined in (4.14). Each coefficient $\widetilde{E}^{(s)}_{ik}$ can be rewritten as the average of the noise ε_t. that are present in $N^{(s)}$ and that correspond to coefficient $x_t = (i,k)$.

$$\widetilde{E}^{(s)}_{ik} = \sum_{t \in N^{(s)}} \frac{\varepsilon_t}{\mathbf{r}^{(s)}_{ik} \vee 1} \mathbf{1}\{x_t = (i,k)\} \ . \tag{4.30}$$

From now on, we often omit the dependence in s. We will extensively use the decomposition $Y = \lambda_1 M + E$, where λ_1 is defined in (4.28) and E in (4.29). Recalling that $B_{ik} = \mathbf{1}\{r_{ik} \geq 1\}$, we often rewrite E as the sum of two centered random variables:

$$E_{ik} = (B_{ik} - \lambda_1) M + B_{ik} \widetilde{E}_{ik} \ .$$

Handling the concentration of the noise is more challenging in the case $\lambda_0 \leq 1$ than in the full observation regime $\lambda_0 \geq 1$ discussed in Section 4.5.6. While subGaussian concentration inequalities are effective in the full observation regime $\lambda_0 \geq 1$, they lead to slower estimation rate in the case $\lambda_0 \leq 1$, for instance in Lemma 4.5.1. Indeed, it turns out that the variance of a coefficient ε_{ik} is of order $\lambda_0 \leq 1$, while the Hoeffding inequality only implies that $B_{ik} - \lambda_1$, and in particular ε_{ik} are c-subGaussian for some numerical constant c. To overcome this issue, one of the main ideas is to use Bernstein-type bounds on the coefficients of E and on the random matrix $EE^T - \mathbb{E}[EE^T]$- see Lemma 4.5.7 and Proposition 4.4.1.

4.5.1.2 General property on $\mathcal{W}$

Recall that we assume that $\lambda_0 \leq 1$, so that $\frac{1}{\lambda_0} \wedge \lambda_0 = \lambda_0$ in (4.18), and that ϕ_{L_1} is defined in (4.9) by $\phi_{\mathrm{L}_1} := 10^4 \log(10^2 nd/\delta)$. In the following, we let ξ be the event on which the noise concentrates well for all the pairs (Q,w) considered during the whole procedure. More precisely, we say that we are under event ξ, if for any $s = 0, \dots, 5T-1$ and for any pair (Q,w) that is used to compute a refinement as in (4.17) we have

$$\left| \langle E^{(s)}_{i\cdot} - E^{(s)}_{j\cdot}, w \rangle \right| \leq \tfrac{1}{3} \phi_{\mathrm{L}_1} \sqrt{\lambda_0} \quad \text{for any } (i,j) \in [n]^2 \ . \tag{4.31}$$

Lemma 4.5.1. *The event ξ holds true with probability at least $1 - 2T\delta$.*

The idea of Lemma 4.5.1 is to apply a bernstein-type inequality and a union bound on all the possible dot products $\langle E^{(s)}_{i\cdot}, w \rangle$, for all the $5T$ possible s and the at most $2T$ possible w. The upper bound is of the order of the square of the variance of E_{ik} up to the polylogarithm factor ϕ_{L_1}. The crucial point is that if $\langle E^{(s)}_{i\cdot}, w \rangle$ is not λ_0-subGaussian, it satisfies the Bernstein's Condition [2.15 of [62]] with variance $\nu = \lambda_0$ and scaling factor $b = \|w\|_\infty$. We then obtain an upper bound of order $\sqrt{\lambda_0}$ since any w considered in the update step (4.18) must satisfy (4.16). Recall that $\bar{\gamma}$ is defined in (4.11). We fix in what follows a sequence $\overline{\gamma} = \gamma_0 > \gamma_1 > \gamma_2 > \cdots > \gamma_{\lfloor 2\log_2(n) \rfloor} = \gamma_{\min}$ in Γ satisfying property (4.10). We say that u is the level of the corresponding threshold γ_u. We say $\mathcal{W}$ and (γ_u) satisfies the property $\mathcal{C}(\mathcal{W}, (\gamma_u))$ if the following holds

1. **consistency:** For any $(i,j) \in \mathcal{G}(\mathcal{W}, \gamma_{\min})$ it holds that $\pi^*(i) > \pi^*(j)$.
2. **weak-transitivity:** Fix any $u \in \{0, \dots, \lfloor 2\log_2(n) \rfloor - 1\}$. For any experts i, j, k, if i is $\mathcal{G}(\mathcal{W}, \gamma_u)$-above j and $k \in \mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma_{u+1}), j)$, then any $i' \geq i$ is also $\mathcal{G}(\mathcal{W}, \gamma_{\min})$-above k.

The first point of the above property means that at threshold $\gamma_{\min}$, there is no mistake in the directed graph $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, meaning that if there is an edge from i to j in $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, then i is truly above j. Moreover, we only state the consistency property of the graph $\mathcal{G}(\mathcal{W}, \gamma_{\min})$, but this property also implies that, for any more conservative threshold $\gamma \geq \gamma_{\min}$, any $(i,j) \in \mathcal{G}(\mathcal{W}, \gamma)$ satisfies $\pi^*(i) > \pi^*(j)$. This is due to the fact that $\mathcal{G}(\mathcal{W}, \gamma) \subset \mathcal{G}(\mathcal{W}, \gamma_{\min})$. The weak transitivity property states in particular that if there is a path from i to j in the more conservative graph $\mathcal{G}(\mathcal{W}, \gamma_u)$, then there is a path from i to any k in the neighborhood of j at the less conservative threshold $\gamma_{\min}$. The following lemma states that the above property remains true for the weighted graph $\mathcal{W}'$, after any update (4.18) of the whole procedure.

Lemma 4.5.2. *Under ξ, the property $\mathcal{C}(\mathcal{W}', (\gamma_u))$ holds true for any directed weighted graph $\mathcal{W}'$ obtained at any stage of Algorithm 14 and Algorithm 15.*

We denote in the following $\mathcal{W}_t$ for the directed weighted graph at the beginning of step t. For any $u \in [0, \lfloor 2\log_2(n) \rfloor]$, we also write as a shorthand $\mathcal{G}_{t,u} = \mathcal{G}(\mathcal{W}_t, \gamma_u)$ for the directed graph at beginning of step t and level u and $P_{t,u}(i) = \mathcal{N}(\mathcal{G}_{t,u}, i)$ for the set of experts that are not comparable with i according to $\mathcal{G}_{t,u}$. For any sequence of experts I, we write $\mathcal{P}_{t,u}(I)$ for the sequence of subsets $(P_{t,u}(i))_{i \in I}$. Let us now divide the T steps of the algorithm into $\tau_{\max} = \lfloor \log_2(n) \rfloor + 1$ epochs of $K = \lfloor T/\tau_{\max} \rfloor$ steps. For any $\tau \in [0, \tau_{\max}]$, we also write

$\mathcal{G}^K_{\tau,u} = \mathcal{G}_{\tau K,u}$, $P^K_{\tau,u}(i) = P_{\tau K,u}(i)$ and $\mathcal{P}^K_{\tau,u}(I) = \mathcal{P}^K_{\tau,u}(I)$. Now we consider for each epoch τ a sequence of experts $I(\tau) = (i_1(\tau), \dots, i_{L_\tau}(\tau))$ defined by induction:

- $I(0)$ is the empty sequence
- For $\tau \geq 0$, let $(i_1, \dots, i_L)$ be the sequence ordered according to π^* and corresponding to the union of the already constructed sequences $\bigcup_{\tau' \leq \tau} I(\tau')$, and $i = 0$, $i_{L+1} = n+1$. For any $l \in [0, L]$, let A_l be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_{l+1} but $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_l. For all l such that A_l is not empty, we define i'_l as the expert of A_l which is any expert closest to the median $\lfloor (i_l + i_{l+1})/2 \rfloor$, and the new sequence $I(\tau+1) \coloneqq (i'_l)$.

By definition, remark that $I(1)$ is equal to $(\lfloor (n+1)/2 \rfloor)$. The induction step aims at building a sequence $I(\tau+1)$ that is disjoint from $\cup_{\tau' \leq \tau} I(\tau')$, and that cuts each set A_l of experts that are above i_l and below i_{l+1} according to the graph at epoch $\tau+1$ and level $2\tau+1$. Given the already constructed collections of perfectly ordered experts $I(\tau')$ for $\tau' \leq \tau$, the idea of $I(\tau+1)$ is that it tends to fill the gaps between the neighborhoods in $\mathcal{G}_{\tau+1,2\tau+1}$ of any two successive experts in $\cup_{\tau' \leq \tau} I(\tau')$.

By monotonicity, it holds that for any expert i, epoch τ and level u that $P^K_{\tau+1,u+1}(i) \subset P^K_{\tau+1,u}(i) \subset P^K_{\tau,u}(i)$. We say that the sets $P^K_{\tau,2\tau}(i)$ and $P^K_{\tau,2\tau+1}(i)$ are the neighborhoods of i at the beginning of epoch τ and that the sets $P^K_{\tau+1,2\tau}(i), P^K_{\tau+1,2\tau+1}(i)$ are the neighborhood of i at the end of epoch τ. The neighborhoods at the end of a given epoch τ are obtained from the neighborhoods at the beginning the of epoch τ after K steps of the Algorithm 14. On the other hand, we say that the sets $P^K_{\tau,2\tau}, P^K_{\tau+1,2\tau}$ are the conservative subsets at epoch τ, since they correspond to a more conservative directed graph with threshold $\gamma_{2\tau} \geq \gamma_{2\tau+1}$. The following lemma states that, at any epoch τ, the conservative subsets at the beginning of epoch τ are well separated according to the true order $\pi^* = \mathrm{id}$:

Lemma 4.5.3. *Under event ξ, for any $\tau \in [0, \tau_{\max}]$, letting $(i_1, \dots, i_L) = I(\tau)$, we have*

$$P^K_{\tau,2\tau}(i_1) < \cdots < P^K_{\tau,2\tau}(i_L).$$

In other words, Lemma 4.5.3 implies that, for any $l < l'$, any expert in $P^K_{\tau,2\tau}(i_l)$ is π^*-below any expert in $P^K_{\tau,2\tau}(i_{l'})$. As a consequence, it holds that for any $l \in [1, L_\tau - 2]$,

$$P^K_{\tau,2\tau}(i_l) \overset{\mathcal{G}^K_{\tau,2\tau}}{<} P^K_{\tau,2\tau}(i_{l+2}) \ . \tag{4.32}$$

Namely, any expert in $P^K_{\tau,2\tau}(i_l)$ is $\mathcal{G}^K_{\tau,2\tau}$-below any expert in $P^K_{\tau,2\tau}(i_{l+2})$. Indeed, Lemma 4.5.3 and first point of event ξ imply that any expert j in $P^K_{\tau,2\tau}(i_l)$ is $\mathcal{G}^K_{\tau,2\tau}$-below i_{l+1}, since j cannot be in $P^K_{\tau,2\tau}(i_l)$. On the other hand, i_{l+1} is itself $\mathcal{G}^K_{\tau,2\tau}$-below any expert of $P^K_{\tau,2\tau}(i_{l+2})$ for the same reason. The following lemma states that the ending less conservative subsets are covering the set of all experts.

Lemma 4.5.4. *Under event ξ, it holds that*

$$[n] = \bigcup_{\tau=0}^{\tau_{\max}-1} \bigcup_{i \in I(\tau)} P^K_{\tau+1,2\tau+1}(i) \ .$$

Let $\hat{\pi}$ be the estimator obtained from the final weighted directed graph $\mathcal{W}$ at the end of the procedure, that is any permutation on $[n]$ that is consistent with the largest acyclic graph of the form $\mathcal{G}(\mathcal{W}, \gamma)$ for all $\gamma > 0$. For any sequence of subsets $\mathcal{P} = (P_1, \dots, P_L)$ we define

$$\mathrm{SN}(\mathcal{P}) = \sum_{P \in \mathcal{P}} \|M(P) - \overline{M}(P)\|_F^2 \ . \tag{4.33}$$

The following proposition that we can control the L_2 error of $\hat{\pi}$ by the maximum over all epoch τ of the sum over τ of the square norms of the groups in $\mathcal{P}^K_{\tau+1,2\tau+1}$.

Proposition 4.5.5. *Under event ξ, it holds that*

$$\|M_{\hat{\pi}^{-1}} - M\|_F^2 \leq 4 \sum_{\tau=0}^{\tau_{\max}-1} \mathrm{SN}(\mathcal{P}^K_{\tau+1,2(\tau+1)}) \ . \tag{4.34}$$

Recall that $\bar{\gamma}$ is defined in (4.11), and that Γ can be taken to be a valid grid with $\bar{\gamma}$ smaller than a polylogarithm in n, d, δ. The final proposition states that at any level u and any step t, any sequence of subset that can be ordered according to the already constructed graph $\mathcal{G}_{t,u}$ as in (4.32) will either have a square norm smaller than the minimax rate ρ_{perm}, defined in (4.6) or almost exponentially decrease its square norm with high probability.

Proposition 4.5.6. *Fix any $u \in [0, 2\tau_{\max}]$ and step $t < T$, and assume that $I = (i_1, \dots, i_L)$ is a sequence of experts that satisfies $P_{t,u}(i_1) \overset{\mathcal{G}_{t,u}}{<} \dots \overset{\mathcal{G}_{t,u}}{<} P_{t,u}(i_L)$. Then on the intersection of the event ξ (defined in (4.31)) and an event of probability higher than $1 - 5\delta$, it holds that*

$$\mathrm{SN}(\mathcal{P}_{t+1,u}(I)) \le \left[C\bar{\gamma}^6 \rho_{\mathrm{perm}}(n,d,\lambda_0)\right] \vee \left[\left(1 - \frac{1}{4\bar{\gamma}^2}\right)\mathrm{SN}(\mathcal{P}_{t,u}(I))\right] \ ,$$

for some numerical constant C.

Let us fix $\tau \in \{0, \dots, \tau_{\max} - 1\}$. Applying Proposition 4.5.6 for each $t = K\tau, \dots, K\tau + K - 1$ and $u = 2(\tau+1)$ -the hypothesis of Proposition 4.5.6 being satisfied by (4.32), we obtain with probability $1 - 5(K+T)\delta$ that

$$\begin{aligned}\mathrm{SN}(\mathcal{P}_{\tau+1,2(\tau+1)}) &\le \left[C\bar{\gamma}^6 \rho_{\mathrm{perm}}(n,d,\lambda_0)\right] \vee e^{-\frac{T}{4\tau_{\max}\bar{\gamma}^4}} nd \\ &\le CT\bar{\gamma}^6 \rho_{\mathrm{perm}}(n,d,\lambda) \ ,\end{aligned}$$

if T is larger than $4\bar{\gamma}^6 \ge 4\log^2(nd)\bar{\gamma}^4$. We conclude the proof of Theorem 4.2.2 with Proposition 4.5.5, using that $4\tau_{\max} \le \bar{\gamma}$:

$$\|M_{\hat{\pi}^{-1}} - M\|_F^2 \le 4 \sum_{\tau=0}^{\tau_{\max}-1} \mathrm{SN}(\mathcal{P}^K_{\tau+1,2(\tau+1)}) \le CT\bar{\gamma}^7 \rho_{\mathrm{perm}}(n,d,\lambda) \ .$$

4.5.2 Proofs of the lemmas of Section 4.5.1 and of Proposition 4.5.5

4.5.2.1 Proof of Proposition 4.5.5

Let $\hat{\pi}$ be any arbitrary permutation that is consistent with the largest **DAG** $\mathcal{G}(\mathcal{W}, \bar{\gamma})$, as defined in Section 4.3.1. Recall that we assume in this proof that $\pi^* = \mathrm{id}$. By Lemma 4.5.4, for any $i \in [n]$ there exists $\tau \in [0, \tau_{\max} - 1]$ and $i_0 \in I(\tau)$ such that $i \in P^K_{\tau+1,2\tau+1}(i_0)$.

Let us define the interval $[a,b]$ as the maximal interval containing i_0 and that is included in the more conservative set $P^K_{\tau+1,2\tau}$. Now, if $j > b$, then by definition there exists j' such that $j \ge j' > b$ and $j' \notin P^K_{\tau+1,2\tau}$. Summarizing the properties, we have $j \ge j' \overset{\mathcal{G}_{\tau+1,2\tau}}{>} i_0$, and that i is in the neighborhood of i_0 in the graph $\mathcal{G}_{\tau+1,2\tau+1}$. Hence, applying the weak-transitivity property (first in $\mathcal{C}$), holding true on event ξ - see Lemma 4.5.2, we obtain that j is $\mathcal{G}(\mathcal{W}_{K(\tau+1)}, \gamma_{\min})$-above i. By the consistency property (second point in $\mathcal{C}$), j is also necessarily $\mathcal{G}(\mathcal{W}, \gamma_{\min})$-above i, and this proves that all the $n-b$ experts j satisfying $j > b$ are $\mathcal{G}(\mathcal{W}, \gamma_{\min})$ above i. Hence, it holds that $\hat{\pi}(i) \le b$. By symmetry, we also prove that $\hat{\pi}(i) \ge a$, so that

$$\hat{\pi}(i) \in [a,b] \subset P^K_{\tau+1,2\tau}(i_0). \tag{4.35}$$

Finally, we have

$$\begin{aligned}\|M_{\hat{\pi}^{-1}} - M\|_F^2 &= \sum_{i=1}^n \|M_{\hat{\pi}(i)\cdot} - M_{i\cdot}\|_F^2 \\ &\le \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau+1}(i_0)} \|M_{\hat{\pi}(i)\cdot} - M_{i\cdot}\|^2 \\ &\le 2 \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau+1}(i_0)} \|M_{i\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 + \|M_{\hat{\pi}(i)\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 \\ &\le 4 \sum_{\tau=0}^{\tau_{\max}} \sum_{i_0 \in I(\tau)} \sum_{i \in P^K_{\tau+1,2\tau}(i_0)} \|M_{i\cdot} - \overline{m}(P^K_{\tau+1,2\tau}(i_0))\|^2 \ ,\end{aligned}$$

where we used Lemma 4.5.4 for the first inequality and (4.35) for the last inequality.

4.5.2.2 Proof of the lemmas of Section 4.5.1

We postpone the proof of Lemma 4.5.1 to the next subsection.

Proof of Lemma 4.5.2. Recall that we consider the case $\lambda_0 \le 1$, so that $\lambda_0 \wedge 1/\lambda_0 = \lambda_0$ in (4.18).

Consider any substep of the whole procedure where the current directed weighted graph is $\mathcal{W}'$. For the first point, remark that i is $\mathcal{G}(\mathcal{W}', \gamma_{\min})$-above j only if there exists a previous substep during which we find out

that $\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle \geq \gamma_{\min}$ on some direction $w \in \mathbb{R}^Q$, where Y is the sample used to refine the edges (4.17). Since $\gamma_{\min} > \phi_{\mathrm{L}_1}$, then decomposing $Y = \lambda_1 M + E$ as in (4.29), we have

$$\lambda_1 \langle M_{i\cdot} - M_{j\cdot}, w\rangle \geq \langle Y_{i\cdot} - Y_{j\cdot}, w\rangle - \langle E_{i\cdot} - E_{j\cdot}, w\rangle > 0 \ , \tag{4.36}$$

where the last inequality comes from (4.31), using the notation (4.27). Since the coefficients of w are nonegative, we have proven that i is above j. For the second point, assume that i is $\mathcal{G}(\mathcal{W}, \gamma_u)$-above j, and take $i' \geq i$. As before, there exists a direction w used during the procedure such that $\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle \geq \gamma_u$. Now consider any $k \in \mathcal{N}(\mathcal{G}(\mathcal{W}, \gamma_{u+1}), j)$. On the direction w, we have under the event ξ defined in (4.31) that

$$\begin{aligned}
\langle Y_{i'\cdot} - Y_{k\cdot}, w\rangle &\geq \lambda_1 \langle M_{i'\cdot} - M_{k\cdot}, w\rangle - \tfrac{1}{3}\phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\geq \lambda_1 \langle M_{i\cdot} - M_{k\cdot}, w\rangle - \tfrac{1}{3}\phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\geq \langle Y_{i\cdot} - Y_{j\cdot}, w\rangle - \langle Y_{k\cdot} - Y_{j\cdot}, w\rangle - \phi_{\mathrm{L}_1}\sqrt{\lambda_0} \\
&\geq (\gamma_u - \gamma_{u+1} - \phi_{\mathrm{L}_1})\sqrt{\lambda_0} \geq \gamma_{\min}\sqrt{\lambda_0} \ ,
\end{aligned}$$

where the last inequality comes from the assumption (4.10). We conclude that i' is $\mathcal{G}(\mathcal{W}', \gamma_{\min})$-above k. □

Proof of Lemma 4.5.3. We proceed by induction over $\tau \geq 0$. The lemma is trivial for $\tau = 0, 1$ since $I(0)$ is empty and $I(1) = (\lfloor (n+1)/2 \rfloor)$. Let $\tau \geq 1$ and i_1, i_2, i_3 be three experts in $I(\tau) \cup \{0, n+1\}$ such that $i_1 < i_2 < i_3$. Let A be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_1 and $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_2, and A' be the set of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_2 and $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_3. Assume that both sets A and A' are nonempty, and let $j \in A$ and $j' \in A'$. Let us apply the weak-transitivity of $\mathcal{W}, (\gamma_u)$ in Property $\mathcal{C}$ - which holds true under ξ from Lemma 4.5.2 - with $u = 2\tau + 1$. Since j is $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_2, any $k \in P^K_{\tau+1,2(\tau+1)}(j)$ is π^*-below i_2. We also prove that any $k' \in P^K_{\tau+1,2(\tau+1)}(j')$ is π^*-above i_2. We conclude that $P^K_{\tau+1,2(\tau+1)}(j) < P^K_{\tau+1,2(\tau+1)}(j')$, and the proof of the lemma follows. □

Proof of Lemma 4.5.4. We prove that, by construction, any expert $i \in [n]$ is at distance less than $(n+1)/2^{\tau+1}$ of $\bigcup_{\tau'=0}^{\tau} \bigcup_{i \in I(\tau)} P^K_{\tau+1,2\tau+1}(i) \cup \{0, n+1\}$. This is obvious for $\tau = 0$ since any expert is at distance less than $(n+1)/2$ of 0 or $n+1$. Let $(i_1, \dots, i_L) = \bigcup_{\tau' \leq \tau} I(\tau')$ be the collection of experts in the union of all possible $I(\tau')$ that is ordered according to π^*. If j is any expert in $[n]$, then we let $l \in [0, L]$ be such that $i_l \leq j \leq i_{l+1}$. We can assume that $j \notin P^K_{\tau+1,2\tau+1}(i_l)$ and $j \notin P^K_{\tau+1,2\tau+1}(i_{l+1})$ because otherwise the distance of j to $\bigcup_{\tau'=0}^{\tau} \bigcup_{i \in I(\tau)}^{L} P^K_{\tau+1,2\tau+1}(i)$ is 0. Using property $\mathcal{C}$ holding true from Lemma 4.5.2, it holds that the set A of experts that are $\mathcal{G}^K_{\tau+1,2\tau+1}$-above i_l but $\mathcal{G}^K_{\tau+1,2\tau+1}$-below i_{l+1} contains j and therefore is nonempty. Now, let $m = \lfloor (i_l + i_{l+1})/2 \rfloor$ and i' be any expert closest to m in A, as defined in the construction of $I(\tau+1)$, and assume without loss of generality that $m \leq i'$. We consider the following cases:

- $m \leq i' \leq j$: In that case, j is at distance less than $(i_{l+1} - m)/2$ of i' or i_{l+1}.
- $m \leq j < i'$: This case is not possible since i' is the closest expert to m in A.
- $j < m < i'$: In that case, since i' minimizes the distance to m, we necessarily have that $m \in P^K_{\tau+1,2\tau+1}(i_l) \cup P^K_{\tau+1,2\tau+1}(i_{l+1})$. Hence, j is at distance less than $(m - i_l)/2$ of m or i_l.

We have proved that the distance of any j to $\bigcup_{\tau'=0}^{\tau+1} \bigcup_{i \in I(\tau)}^{L} P^K_{\tau+1,2\tau+1}(i) \cup \{0, n+1\}$ is at most $(m - i_l)/2$ or $(i_{l+1} - m)/2$. Using the induction hypothesis, we have that $m - i_l$ and $i_{l+1} - m$ are both less than $n/2^{\tau+1}$, which concludes the induction.

Finally, applying this property with $\tau_{\max} - 1 = \lfloor \log_2(n) \rfloor$ gives a distance strictly smaller than 1, which proves the result. □

4.5.2.3 Proof of Lemma 4.5.1

Let us start with the following lemma, which gives a concentration bound when $\lambda_0 \leq 1$:

Lemma 4.5.7. *For any $\delta' > 0$ and for any matrix $W \in \mathbb{R}^{n \times d}$, the following inequality holds with probability at least $1 - \delta'$:*

$$|\langle E, W\rangle| \leq \sqrt{4e^2 \|W\|_F^2 \lambda_0 \log\left(\frac{2}{\delta'}\right)} + \|W\|_\infty \log\left(\frac{2}{\delta'}\right) \ . \tag{4.37}$$

Now we apply Lemma 4.5.7 with the matrix W with 0 coefficients except at line i where it is equal to the vector $\frac{w}{\|w\|_2}$ as defined in (4.17), and we deduce that

$$|\langle E_{i,\cdot}, \frac{w}{\|w\|_2}\rangle| \le \sqrt{4e^2\lambda_0 \log\left(\frac{2}{\delta'}\right)} + \frac{\|w\|_\infty}{\|w\|_2}\log\left(\frac{2}{\delta'}\right) \le 11\sqrt{\lambda_0}\log(2/\delta') \ , \tag{4.38}$$

where the last inequality comes from Condition (4.16) on w. Now choosing $\delta' = \delta/(4Tn^6)$, a union bound over the at most $2n^2T|\mathcal{H}|(|\Gamma| \wedge n^2)$ pairs (Q, w) considered during the procedure, we deduce the bound of Lemma 4.5.1 for all $\lambda_0 \le 1$.

Proof of Lemma 4.5.7. Recall that $E, \widetilde{E}$ are defined in (4.29) and that we have in particular

$$E_{ik} = (B_{ik} - \lambda_1)M_{ik} + \widetilde{E}_{ik} \ .$$

Let $x > 0$. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}[e^{xE_{ik}}] \le \sqrt{\mathbb{E}[e^{2x(B_{ik}-\lambda_1)M_{ik}}]}\sqrt{\mathbb{E}[e^{2x\widetilde{E}_{ik}}]} \ ,$$

where we recall that $\lambda_1 = 1 - e^{-\lambda_0} \le \lambda_0$. We have

$$\mathbb{E}[e^{2x(B_{ik}-\lambda_1)M_{ik}}] \le e^{-2\lambda_1 x M_{ik}}(\lambda_1(e^{2xM_{ik}} - 1) + 1) \le e^{\lambda_1 e^2 x^2} \ ,$$

and

$$\mathbb{E}[e^{2x\widetilde{E}_{ik}}] \le \lambda_1(e^{2x^2} - 1) + 1 \le e^{\lambda_1 e^2 x^2} \ ,$$

where we used the inequalities $e^{2x^2} - 1 \le e^2x^2$ and $e^{2x} - 1 - 2x \le e^2x^2$ for any $x \in [-1, 1]$.

In particular, if $t > 0$, a Chernoff bound with $x = \frac{t}{2\|W\|_F^2\lambda_0 e^2} \wedge 1$ gives

$$\mathbb{P}(\langle W, E\rangle \ge t) \le \exp(-(\tfrac{t^2}{4\|W\|_F^2\lambda_0 e^2} \wedge t)) \ ,$$

so that with probability at least $1 - \delta'$:

$$|\langle W, E\rangle| \le \sqrt{4e^2\|W\|_F^2\lambda_0 \log\left(\frac{2}{\delta'}\right)} + \log\left(\frac{2}{\delta'}\right) \ .$$

□

4.5.3 Proof of Proposition 4.5.6

Step 0 : general definitions

In this proof, we fix $u \in \{0, \dots, 2\lfloor \log_2(n)\rfloor + 2\}$ and a corresponding threshold γ_u in the sequence in Γ satisfying $\gamma_u \ge \phi_{\mathrm{L}_1}$ - see (4.10) - and a step $t < T$. We assume that $I = (i_1, \dots, i_L)$ is a fixed sequence of experts that satisfies $P_{t,u}(i_1) \overset{\mathcal{G}_{t,u}}{\prec} \dots \overset{\mathcal{G}_{t,u}}{\prec} P_{t,u}(i_L)$.

From now on, we ease the notation by omitting the dependence in t, u, γ_u and we write $\mathcal{G} = \mathcal{G}_{t,u}$, $\mathcal{G}' = \mathcal{G}_{t+1,u}$, $\mathcal{P} = (P_1, \dots, P_L)$ for $\mathcal{P}_{t,u}$ and $\mathcal{P}'$ for $\mathcal{P}_{t+1,u}$. We denote $\widetilde{\mathcal{G}}^h$ for the directed graph at threshold γ_u of the directed weighted graph $\widetilde{\mathcal{W}}^h$ obtained at the end the first update Line 3 of Algorithm 15. We also write $\widetilde{P}_l^h = \mathcal{N}(\widetilde{\mathcal{G}}^h, i_l)$ and $\widetilde{\mathcal{P}}^h = (\widetilde{P}_1^h, \dots, \widetilde{P}_L^h)$ for the corresponding sequence of subsets at height $h \in \mathcal{H}$. By monotonicity, it holds for any $h \in \mathcal{H}$ that

$$P'_l \subset \widetilde{P}_l^h \subset P_l \ .$$

4.5.3.1 Step 1: Analysis of the selected set $\widehat{Q}$

Recall the definition of the neighborhoods (4.20) of the set P_l in the graph $\mathcal{G}$:

$$\mathcal{N}_a(l) = \bigcap_{i \in P_l} \mathbf{rk}_{\mathcal{G},i_l}^{-1}([1, a]) \quad \text{and} \quad \mathcal{N}_{-a}(l) = \bigcap_{i \in P_l} \mathbf{rk}_{\mathcal{G},i_l}^{-1}([-1, -a]) \ ,$$

Define for $\kappa > 0$ and $l \in [1, L]$ the population version $\mathbf{\Delta}_k^*$ of the width statistic $\widehat{\mathbf{\Delta}}_k$ - see (4.21) - as the the difference of the best and worst expert of $P(i_l)$ if $a = 0$ and as the difference of the average of the experts in $\mathcal{N}_a(l)$ and the average of the expert in $\mathcal{N}_{-a}(l)$:

$$\mathbf{\Delta}_k^*(0, l) = \max_{i,j \in P(i_l)} |M_{i,k} - M_{j,k}| \quad \text{and} \quad \mathbf{\Delta}_k^*(a, l) = \overline{m}_k(\mathcal{N}_a(l)) - \overline{m}_k(\mathcal{N}_{-a}(l)) \text{ if a} \ge 1. \tag{4.39}$$

We also define $a^*(h,l)$ as the minimum $a \geq 1$ such that there are at least $\frac{1}{\lambda_0 h^2}$ experts in $\mathcal{N}_a(l)$ and in $\mathcal{N}_{-a}(l)$:

$$a^*(h,l) = \min\{a \geq 1 \ : \ |\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)| \geq \frac{1}{\lambda_0 h^2}\} \ . \tag{4.40}$$

Now, define for $\phi \geq 1$:

$$\begin{aligned} Q_l^{*h}(\phi) &:= \{k \in [d] \ : \ \mathbf{\Delta}_k^*(0,l) \in [\phi h, 2\phi h]\} \\ \overline{Q}_l^{*h}(\phi) &:= \{k \in [d] \ : \ \mathbf{\Delta}_k^*(a^*(\phi^{-1}h,l),l) \geq h/2\} \ . \end{aligned} \tag{4.41}$$

The following lemma states that, for ϕ of order $\log(nd/\delta)$, we can sandwich $\widehat{Q}_l^h$ between the two fixed sets Q_l^{*h} and $\overline{Q}_l^{*h}$:

Lemma 4.5.8. *Let l be a fixed index in $\{1,\dots,L\}$ and h a fixed height in $\mathcal{H}$. There exists a numerical constant $\kappa_0 > 0$ such that, with probability at least $1 - \delta/(L|\mathcal{H}|)$, we have*

$$Q_l^{*h}(\kappa_0 \log(nd/\delta)) \subset \widehat{Q}_l^h \subset \overline{Q}_l^{*h}(\kappa_0 \log(nd/\delta)) \ . \tag{4.42}$$

4.5.3.2 Step 2 : l1-control of the intermediary sets $\widetilde{\mathcal{P}}^h$

Recall that γ_u is a threshold corresponding to a sequence in Γ as defined in (4.10). For any sets $P \subset [n], Q \subset [d]$, we say that $M(P,Q)$ is indistinguishable in L_1-norm if it satisfies

$$\max_{i,j \in P} \|M_{i\cdot}(P,Q) - M_{j\cdot}(P,Q)\|_1 \leq 3\gamma_u \sqrt{\frac{|Q|}{\lambda_0}} \ . \tag{4.43}$$

For a fixed $l \in \{1,\dots,L\}$, let $\xi_{\mathrm{L}_1}(l,h)$ be the event under which $M(\widetilde{P}_l^h, \widehat{Q}_l^h)$ is indistinguishable in L_1-norm.

Lemma 4.5.9. *Let l be a fixed index in $\{1,\dots,L\}$ and $h \in \mathcal{H}$ such that $\lambda_0|Q_l^{*h}| \geq 1$. The event $\xi_{\mathrm{L}_1}(l,h)$ holds true with probability at least $1 - \delta/(L|\mathcal{H}|)$.*

Let κ_0 be a numerical constant given by Lemma 4.5.8 and let $\phi_0 = \kappa_0 \log(nd/\delta)$. In what follows, we write for simplicity $(Q_l^{*h}, \widehat{Q}_l^h, \overline{Q}_l^h) = (Q_l^{*h}(\phi_0), \widehat{Q}_l^h(\phi_0), \overline{Q}_l^h(\phi_0))$. Lemma 4.5.9 provides an upper bound only on the L_1 distance between rows of M restricted to the subsets $\widetilde{P}_l^h$ and $\widehat{Q}_l^h$, while the square norm of a group (4.33) is defined with the L_2 distance. with (4.43). The idea is that for any k in Q^{*h}, and for any $i \in \widetilde{P}^h$, we have that $|M_{ik} - \overline{m}_k|^2 \leq 2\phi_0 h |M_{ik} - \overline{m}_k|$. In particular, $\|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_2^2 \leq 2\phi_0 h \|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_1$. Hence, it holds from Lemma 4.5.8, Lemma 4.5.9 and a union bound over all $l \in \{1,\dots,L\}$ and all $h \in \mathcal{H}$ satisfying $\lambda_0|Q_l^{*h}| \geq 1$ that with probability at least $1 - 2\delta$,

$$\sum_{i \in \widetilde{P}_l^h} \|M_{i\cdot}(\widetilde{P}_l^h, Q_l^{*h}) - \overline{m}_\cdot(\widetilde{P}_l^h, Q_l^{*h})\|_2^2 \leq 6\phi_0\gamma_u \left[h|\widetilde{P}_l^h| \sqrt{\frac{|\overline{Q}_l^{*h}|}{\lambda_0}} \right] \ , \tag{4.44}$$

simultaneously for all $l \in \{1,\dots,L\}$ and $h \in \mathcal{H}$ satisfying $\lambda_0|Q_l^{*h}| \geq 1$.

Proof of Lemma 4.5.9. Let l be a fixed index in $\{1,\dots,L\}$ and h be a fixed height in $\mathcal{H}$. If $a \geq 1$, the subset P_l is disjoint from the sets $\mathcal{N}_a(l) \cup \mathcal{N}_{-a}(l)$ so that $\widehat{Q}_l^h$ is independent of $Y^{(1)}(P_l)$. Remark also that condition (4.16) is satisfied since $\lambda_0|Q_l^{*h}| \geq 1$ and $Q_l^{*h} \subset \widehat{Q}_l^h$.

Recall that we assume that $\lambda_0 \leq 1$. We write $w = \mathbf{1}_{\widehat{Q}_l^h}$, and we recall that $B = (B_{ik})$ is the matrix defined in (4.15). Let $i,j \in \widetilde{P}_l^h$ so that, by definition, we have that $|\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle| \leq \gamma_u \sqrt{\lambda_0 |\widehat{Q}_l^h|}$. With probability at least $1 - \delta/L$, for all i,j in P_l we have that

$$\lambda_1 \, |\langle M_{i\cdot} - M_{j\cdot}, w\rangle| \leq |\langle Y_{i\cdot} - Y_{j\cdot}, w\rangle| + |\langle E_{i\cdot} - E_{j\cdot}, w\rangle| \leq (\gamma_u + \phi_{\mathrm{L}_1}/2)\sqrt{\lambda_0 |\widehat{Q}_l^h|} \ . \tag{4.45}$$

where the last inequality comes from Lemma 4.5.7 applied with $\delta' = \delta/n^3$ and from the definition of ϕ_{L_1} (4.9). Recalling the two inequalities $\lambda_1 = 1 - e^{-\lambda_0} \geq \lambda_0/2$ and $\phi_{\mathrm{L}_1} \leq \gamma_u$, we obtain the result. □

4.5.3.3 Step 3 : Local square norm reduction

Henceforth we condition to the sample $Y^{(1)}$ of Algorithm 15 which allows us to assume that, for any $h \in \mathcal{H}$, the two sequences of sets $\widetilde{\mathcal{P}}^h$ and $\widehat{\mathcal{Q}}^h$ are fixed.

For $\kappa_1 > 0$, let $\xi_{\text{loc}}(l, h, \kappa_1)$ be the event holding true if the local square norm of $M(P_l, \widehat{Q}_l^h)$ has decreased at the end of Algorithm 15, that is

$$\begin{aligned}\|M(P_l', \widehat{Q}_l^h) - \overline{M}(P_l', \widehat{Q}_l^h)\|_F^2 \le \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}\right] \\ \vee \left(1 - \frac{1}{4\gamma_u^2}\right)\|M(P_l, \widehat{Q}_l^h) - \overline{M}(P_l, \widehat{Q}_l^h)\|_F^2 \ .\end{aligned} \tag{4.46}$$

The following proposition states that given the fact that the experts in $\widetilde{P}_l^h$ are indistinguishable in L_1-norm and $\lambda_0(|\widetilde{P}_l^h| \wedge |Q_l^{*h}|) \ge 1$, the event ξ_{loc} holds true simultaneously for all l and h with high probability.

Proposition 4.5.10. *There exists a numerical constant κ_1 such that the following holds, for any fixed index l in $\{1, \dots, L\}$, and fixed height h in $\mathcal{H}$. Conditionally to $Y^{(1)}$, the event $\xi_{\text{L}_1}(l)$ and $\lambda_0(|\widetilde{P}_l^h| \wedge |Q_l^{*h}|) \ge 1$, the event $\xi_{\text{loc}}(l, h, \kappa_1)$ holds true with probability at least $1 - 3\delta/(L|\mathcal{H}|)$.*

Proposition 4.5.10 is at the core of the analysis, and its proof contains a significant part of the arguments. This proposition and its proof are similar to Proposition D.5 in [74], but the main difficulty with respect to [74] is that we do not achieve the optimal rate in $\lambda_0 \le 1$ using only the subgaussianity of the coefficients of the noise E. A key step in the proof of Proposition 4.5.10 is Proposition 4.4.1, which implies Lemma 4.5.13 and gives a concentration inequality of the operator norm of $EE^T - \mathbb{E}[EE^T]$. Proposition 4.4.1 is effective in that case since the coefficients of E will be proven to satisfy (4.24).

Then, the idea is that when a group P_l' has a square norm of order at least $\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}$, the PCA-based procedure defined as in (4.23) will output a vector $\hat{v}$ that is well aligned with the first left singular vector of $M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h)$. Moreover, the isotonic structure of $M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h)$ implies in fact that its operator norm is greater than a polylogarithmic fraction of its Frobenius norm (see Lemma 4.5.12 or Lemma E.4 in [74]], so that $\|\hat{v}^T(M(\widetilde{P}_l^h, \widehat{Q}_l^h) - \overline{M}(\widetilde{P}_l^h, \widehat{Q}_l^h))\|_2^2$ is of the same order as the square Frobenius norm. Hence after updating the edges, we can prove that the experts in $\widetilde{P}_l^h \setminus P_l'$ were contributing significantly to the Frobenius norm, which enforces the contraction part in the second term of the maximum in (4.46). All the details of the proof can be found in Section 4.5.5.

4.5.3.4 Step 4 : Control of the size of the sets $\overline{Q}_l^{*h}$

For any $p \in [n] \cap \{2^k \ : \ k \in \mathbb{Z}^+\}$, let $\mathcal{L}(p)$ be the sets of indices $l = 1, \dots, L$ whose corresponding group size $|P_l|$ belongs to $[p, 2p)$. The two upper bounds implied by (4.44) and (4.46) both depend on the selected subset of columns, which is included in $\overline{Q}_l^{*h}$ under the event of Lemma 4.5.8. The following lemma provides an upper bound on the sum over $l \in \mathcal{L}(p)$ of the size of the sets $\overline{Q}_l^{*h}(\phi)$ defined in (4.41), for any $\phi > 0$.

Lemma 4.5.11. *For any $\phi \ge 1$ and any $h \in \mathcal{H}$, it holds that*

$$\sum_{l \in \mathcal{L}(p)} |\overline{Q}^{*h}(\phi)| \le 12\phi^2 \left(\frac{1}{p\lambda_0 h^2} \vee 1\right)\frac{d}{h} \ .$$

The proof of Lemma 4.5.11 is mainly implied by the fact that the coefficients of M are bounded by 1. Then, the idea is that in the case where all the sets P_l are of size p, it is enough to take a number of group a of order at most $\frac{1}{p\lambda_0 h^2} \vee 1$ above and below each P_l to ensure that the corresponding neighborhood of P_l has size $|\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)| \ge \frac{1}{\lambda_0 h^2}$.

4.5.3.5 Step 5 : Conclusion of the previous steps

We first decompose the square norm $\text{SN}(\mathcal{P})$ as defined in (4.33) into two terms. Assume that the event of Lemma 4.5.8, $\xi_{\text{L}_1}(l)$ and $\xi_{\text{loc}}(l, h, \kappa_1)$ - see Lemma 4.5.9 and Proposition 4.5.10 - hold true. Define $\mathcal{L}_-$ as the sequence of indices l such that the corresponding reduced subsets P_l' have low local square norm for all $h \in \mathcal{H}$. More precisely, we say that $l \in \mathcal{L}_-$ if for all $h \in \mathcal{H}$ we have

$$\begin{aligned}\|M(P_l', \widehat{Q}_l^h) - \overline{M}(P_l', \widehat{Q}_l^h)\|_F^2 \le \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0}\sqrt{|P_l||\widehat{Q}_l^h|} + \frac{|P_l|}{\lambda_0}\right] \\ \vee \frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2 \ .\end{aligned} \tag{4.47}$$

We also define the complementary $\mathcal{L}_+ = [1, L] \setminus \mathcal{L}_-$ and their corresponding subsets $\mathcal{P}'_+, \mathcal{P}'_-$ in $\mathcal{P}'$. We have the following decomposition:

$$\mathrm{SN}(\mathcal{P}') = \mathrm{SN}(\mathcal{P}'_+) + \mathrm{SN}(\mathcal{P}'_-) \ . \tag{4.48}$$

Let us now give an upper bound of $\mathrm{SN}(\mathcal{P}'_+)$. For any $l \in \mathcal{L}_+$, there exists by definition an element $h_l \in \mathcal{H}$ such that $\|M(P'_l, \widehat{Q}^{h_l}_l) - \overline{M}(P'_l, \widehat{Q}^{h}_l)\|_F^2 > \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0} \sqrt{|P_l| \|\widehat{Q}^h_l|} + \frac{|P_l|}{\lambda_0} \right] \vee \frac{1}{2|\mathcal{H}|} \|M(P_l, \widehat{Q}^h_l) - \overline{M}(P_l, \widehat{Q}^h_l)\|_F^2$. Hence, applying (4.46) with $h = h_l$, we have that, for any $l \in \mathcal{L}_+$,

$$\begin{aligned}
\|M(P'_l) - \overline{M}(P'_l)\|_F^2 &= \|M(P'_l, \widehat{Q}^{h_l}_l) - \overline{M}(P'_l, \widehat{Q}^{h_l}_l)\|_F^2 + \|M(P'_l, [d] \setminus \widehat{Q}^{h_l}_l) - \overline{M}(P'_l, [d] \setminus \widehat{Q}^{h_l}_l)\|_F^2 \\
&\le \|M(P_l) - \overline{M}(P_l)\|_F^2 - \frac{1}{4\gamma_u^2} \|M(P_l, \widehat{Q}^{h_l}_l) - \overline{M}(P_l, \widehat{Q}^{h_l}_l)\|_F^2 \\
&\le \left(1 - \frac{1}{\gamma_u^3}\right) \|M(P_l) - \overline{M}(P_l)\|_F^2 \ ,
\end{aligned}$$

where the third inequality comes from the second term of (4.47) together with $P'_l \subset P_l$ and $\gamma_u \ge \phi_{L_1} \ge 8|\mathcal{H}|$, with ϕ_{L_1} defined in (4.9). Hence we obtain that

$$\mathrm{SN}(\mathcal{P}'_+) \le \left(1 - \frac{1}{\gamma_u^3}\right) \mathrm{SN}(\mathcal{P}_+) \ . \tag{4.49}$$

Finally, we give an upper bound of $\mathrm{SN}(\mathcal{P}'_-)$. Let us write $\mathcal{D}_n = \{2^k \ : \ k \in \mathbb{Z}^+\} \cap [n]$ for the set of dyadic integer smaller than n. Given $p \in \mathcal{D}_n$, we write $\mathcal{L}_-(p) = \mathcal{L}(p) \cap \mathcal{L}_-$ for the set of indices in $\mathcal{L}_-$ such that $|P_l| \in [p, 2p)$, and $\mathcal{P}'_-(p)$ for the corresponding sequence of subsets in $\mathcal{P}'_-(p)$. Let $\phi_0 = \kappa_0 \log(nd/\delta)$, where κ_0 is a numerical constant given by Lemma 4.5.8. By definition of Q^{*h}_l, the square norm of a group P'_l restricted to questions that do not belong the set $\cup_{h \in \mathcal{H}} Q^{*h}_l$ is smaller than $\phi_0 nd \cdot \min(\mathcal{H}) \le \phi_0$. Hence, we have that

$$\mathrm{SN}(\mathcal{P}'_-) = \sum_{p \in \mathcal{D}_n} \mathrm{SN}(\mathcal{P}'_-(p)) \le \phi_0 + \sum_{(p,h) \in \mathcal{D}_n \times \mathcal{H}} \sum_{l \in \mathcal{L}_-(p)} \|M(P'_l, Q^{*h}_l) - \overline{M}(P'_l, Q^{*h}_l)\|_F^2 \ . \tag{4.50}$$

If $\lambda_0 |Q^{*h}_l| \le 1$ then we use the trivial inequality $\|M(P'_l, Q^{*h}_l) - \overline{M}(P'_l, Q^{*h}_l)\|_F^2 \le |P'_l| |Q^{*h}_l| \le |P^h_l|/\lambda_0$,since the entries of M are bounded by one.

If $\lambda_0 |Q^{*h}_l| \ge 1$ and $|\widetilde{P}^h_l| \lambda_0 \le 1$, we have that $h|\widetilde{P}^h_l| \sqrt{\frac{|\overline{Q}^{*h}_l|}{\lambda_0}} \le \sqrt{\frac{|\widetilde{P}^h_l| |\overline{Q}^{*h}_l|}{\lambda_0^2}}$, using the fact that $h \le 1$. Hence, since the experts in $P'_l \subset \widetilde{P}^h$ are indistinguishable in L_1 norm by Lemma 4.5.9, (4.44) holds true, and we have

$$\begin{aligned}
\|M(P'_l, Q^{*h}_l) - \overline{M}(P'_l, Q^{*h}_l)\|_F^2 &\le 6\phi_0 \gamma_u \left[h|\widetilde{P}^h_l| \sqrt{\frac{|\overline{Q}^{*h}_l|}{\lambda_0}} \right] \\
&\le 6\phi_0 \gamma_u \left[\sqrt{h^2 |\widetilde{P}^h_l|^2 \frac{|\overline{Q}^{*h}_l|}{\lambda_0}} \wedge \sqrt{\frac{|\widetilde{P}^h_l| |\overline{Q}^{*h}_l|}{\lambda_0^2}} \right] \\
&\le 12\phi_0 \gamma_u \left[\sqrt{(h^2 p \lambda_0 \wedge 1) \frac{p|\overline{Q}^{*h}_l|}{\lambda_0^2}} + \frac{p}{\lambda_0} \right] \ .
\end{aligned}$$

Finally, if $\lambda_0(|Q^{*h}_l| \vee |\widetilde{P}^h_l|) \ge 1$, we are in position to apply Proposition 4.5.10. For all $l \in \mathcal{L}_-(p)$ and $h \in \mathcal{H}$ that $\|M(P'_l, Q^{*h}_l) - \overline{M}(P'_l, Q^{*h}_l)\|_F^2$ is either smaller than $\frac{1}{2|\mathcal{H}|} \|M(P_l) - \overline{M}(P_l)\|_F^2$, or it is smaller than $\kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0} \sqrt{|P_l| \|\widehat{Q}^h_l|} + \frac{|P_l|}{\lambda_0} \right]$. From (4.44), it is also smaller than $6\phi_0 \gamma_u h |\widetilde{P}^h_l| \sqrt{\frac{|\overline{Q}^{*h}_l|}{\lambda_0}}$. As a consequence, we obtain the following upper bound:

$$\begin{aligned}
\|M(P'_l, Q^{*h}_l) - \overline{M}(P'_l, Q^{*h}_l)\|_F^2 \le & \kappa_2 \gamma_u^4 \left[\sqrt{(h^2 p \lambda_0 \wedge 1) \frac{p|\overline{Q}^{*h}_l|}{\lambda_0^2}} + \frac{p}{\lambda_0} \right] \\
& \vee \frac{1}{2|\mathcal{H}|} \|M(P_l) - \overline{M}(P_l)\|_F^2 \ ,
\end{aligned} \tag{4.51}$$

with $\kappa_2 = 12(\kappa_0 \vee \kappa_1)$, and using that $\phi_0 \le \kappa_0\gamma_u$ and $|\widetilde{P}_l^h| \le |P_l| \le 2p$.

By the two previous cases on l, the inequality (4.51) is valid for any $l \in \mathcal{L}_-(p)$. Now, we decompose (4.50) into two terms, corresponding to the maximum in (4.51). First, since each P_l is in at most one $\mathcal{P}_-(p)$ for $p \in \mathcal{D}_n$, we have

$$\sum_{(p,h)\in\mathcal{D}_n\times\mathcal{H}} \sum_{l\in\mathcal{L}_-(p)} \frac{1}{2|\mathcal{H}|}\|M(P_l) - \overline{M}(P_l)\|_F^2 \le \frac{1}{2}\,\mathrm{SN}(\mathcal{P}_-) \ . \tag{4.52}$$

Secondly, we have that

$$\begin{aligned}
&\kappa_2\gamma_u^4 \sum_{(p,h)\in\mathcal{D}_n\times\mathcal{H}} \sum_{l\in\mathcal{L}_-(p)} \left[\sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\overline{Q}_l^{*h}|}{\lambda_0^2}} + \frac{p}{\lambda_0}\right] \\
&\qquad \le \kappa_2\gamma_u^6 \left[\max_{p,h} \sum_{l\in\mathcal{L}_-(p)} \sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\overline{Q}_l^{h*}|}{\lambda_0^2}} + \frac{p}{\lambda_0}\right] \\
&\qquad \overset{(a)}{\le} 2\kappa_2\gamma_u^6 \max_{p,h}\left[\frac{n}{\lambda_0} + \sqrt{(h^2p\lambda_0\wedge 1)\frac{p|\mathcal{L}(p)|\sum_{l\in\mathcal{L}(p)}|\overline{Q}_l^{h*}|}{\lambda_0^2}}\right] \\
&\qquad \overset{(b)}{\le} 4\kappa_2^2\gamma_u^7 \max_{p,h}\left[\frac{n}{\lambda_0} + \sqrt{(h^2p\lambda_0\wedge 1)\left(\frac{n^2d}{\lambda_0^2p}\wedge\left(\frac{nd}{p\lambda_0^3h^3}\vee\frac{nd}{\lambda_0^2h}\right)\right)}\right] \\
&\qquad \le 4\kappa_2^2\gamma_u^7 \max_{p,h}\left[\frac{n}{\lambda_0} + nh\sqrt{\frac{d}{\lambda_0}}\wedge\sqrt{\frac{n^2dh^2}{\lambda_0}}\wedge\frac{nd}{\lambda_0^2h}\right] \\
&\qquad \overset{(c)}{\le} 4\kappa_2^2\gamma_u^7\left[\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right] \ ,
\end{aligned}$$

where in (a) we used the Jensen inequality, in (b) we used Lemma 4.5.11 with $\phi = \phi_0$ together with the trivial inequality $\sum_{l\in\mathcal{L}(p)}|\overline{Q}_l^{h*}| \le nd/p$ and in (c) the fact that $x\wedge y \le x^{2/3}y^{1/3}$ and $h \le 1$.

Finally, combining this last inequality with (4.48), (4.49) and (4.52),
we obtain

$$\begin{aligned}
\mathrm{SN}(\mathcal{P}') &= \mathrm{SN}(\mathcal{P}'_+) + \mathrm{SN}(\mathcal{P}'_-) \\
&\le \left(1 - \frac{1}{\gamma_u^3}\right)\mathrm{SN}(\mathcal{P}_+) + 4\kappa_2^2\gamma_u^7\left[\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right]\vee\left[\frac{1}{2}\,\mathrm{SN}(\mathcal{P}_-)\right] \\
&\le \left[C\bar{\gamma}^7\left(\frac{n}{\lambda_0} + n\sqrt{\frac{d}{\lambda_0}}\wedge\frac{n^{2/3}\sqrt{d}}{\lambda_0^{5/6}}\right)\right]\vee\left[\left(1-\frac{1}{\bar{\gamma}^3}\right)\mathrm{SN}(\mathcal{P})\right] \ ,
\end{aligned}$$

where we recall that $\bar{\gamma}$ is defined in (4.11) and satisfies $\bar{\gamma} \ge \gamma_u$. This concludes the proof of Proposition 4.5.6.

4.5.4 Proof of the lemmas of Section 4.5.3

Recall that we can write

$$E = (B - \mathbb{E}[B]) \odot M + B \odot \widetilde{E} \ , \tag{4.53}$$

where $\widetilde{E} = Y - \mathbb{E}[Y|B]$ and that B is a matrix of Bernoulli random variables with parameter λ_1.

Proof of Lemma 4.5.8. Assume first that $\lambda_0 \le 1$. Let us fix $l \in \{1,\dots,L\}$ and $h \in \mathcal{H}$. We omit the dependence in l in this proof to ease the notation, and we write P for P_l. Let us define

$$E'_k(a) := \frac{1}{|\mathcal{N}_a|}\sum_{i\in\mathcal{N}_a} E_{ik} - \frac{1}{|\mathcal{N}_{-a}|}\sum_{i\in\mathcal{N}_{-a}} E_{ik} \quad \text{and} \quad \nu(a) := |\mathcal{N}_a| \wedge |\mathcal{N}_{-a}| \ . \tag{4.54}$$

Using Lemma 4.5.7 with a column matrix W with coefficient in $\{0, \frac{1}{|\mathcal{N}_a|}, -\frac{1}{|\mathcal{N}_{-a}|}\}$ and a union bound over all $k \in [d]$ and $a \in [n]$, we have with probability at least $1 - \delta/L$ that:

$$\frac{1}{\lambda_0}|E'_k(a)| \le \kappa'_0\log(nd/\delta)\left[\sqrt{\frac{1}{\lambda_0\nu(a)}} + \frac{1}{\lambda_0\nu(a)}\right] \ , \tag{4.55}$$

for some numerical constant κ_0'. In what follows, we work under that (4.55) holds true for all $a \in [n]$ and $k \in [d]$.

First inclusion. Let $k \in Q^*(\kappa_0 \log(nd/\delta)h)$ with numerical constant κ_0 to be fixed later. Let $a' \geq 1$ be any integer such that $\nu(a') \geq 1/(\lambda_0 h^2)$. We have

$$\frac{1}{\lambda_0}|E_k'(a')| \leq 2\kappa_0' \log(nd/\delta)h \ , \tag{4.56}$$

since we work under the event defined by (4.55) and since $h^2 \leq h$. Then by consistency of the already constructed graph $\mathcal{G}_{t,u}$ at the beginning of step t, $\mathcal{N}_{a'}$ (resp. $\mathcal{N}_{-a'}$) contains by definition (4.20) only experts that are π^*-above (resp. below) all the experts of P. Since by assumption k is in Q^{*h}, it holds that $\mathbf{\Delta}_k^*(a') \geq \mathbf{\Delta}_k^*(0) \geq \kappa_0 \log(nd/\delta)h$ - see the definition (4.41) of Q^{*h}. Hence, recalling the signal-noise decomposition (4.53), we have that

$$\frac{1}{\lambda_0}\widehat{\mathbf{\Delta}}_k(a') = \frac{\lambda_1}{\lambda_0}\mathbf{\Delta}_k^*(a') + \frac{1}{\lambda_0}E_k'(a') \geq \log(nd/\delta)((1-1/e)\kappa_0 - 2\kappa_0')h \ . \tag{4.57}$$

Choosing $\kappa_0 \geq 10\kappa_0' + 1$, we obtain by definition (4.21) that $\nu(\hat{a}_k(h)) \leq \frac{1}{\lambda_0 h^2}$ so that $k \in \widehat{Q}^h$.

Second inclusion. Let $k \in \widehat{Q}^h$, and $a' = a^*((\kappa_0 \log(nd/\delta))^{-1}h)$ be as defined in (4.40). By definition, it holds that $\nu(a') \geq \kappa_0 \log(nd/\delta)/(\lambda_0 h^2) \geq \frac{1}{\lambda_0 h^2}$. Hence, since $k \in \widehat{Q}^h$, we have by definition (4.22) that $\nu(\hat{a}_k(h)) \leq \frac{1}{\lambda_0 h^2} \leq \nu(a')$, which implies in particular that $\hat{a}_k(h) \leq a'$. Then, by definition (4.21) of $\hat{a}_k(h)$ we have that $\frac{1}{\lambda_0}\widehat{\mathbf{\Delta}}_k(a') \geq h$. Using the concentration inequality (4.55) with $h' = (\kappa_0 \log(nd/\delta))^{-1}h$ and the fact that $\lambda_0 \geq \lambda_1$ we obtain

$$\mathbf{\Delta}_k^*(a') \geq h - \frac{2\kappa_0'}{\kappa_0}h \ , \tag{4.58}$$

and we get the second inclusion by also choosing $\kappa_0 \geq 4(\kappa_0' + 1)$. □

Proof of Lemma 4.5.11. For simplicity, we renumber $\mathcal{L}(p) = (1, 2, \ldots, L' := |\mathcal{L}(p)|)$. Let us write $\nu(a,l) = |\mathcal{N}_a(l)| \wedge |\mathcal{N}_{-a}(l)|$ and $\Lambda = \left\lfloor \frac{\phi^2}{p\lambda_0 h^2} \right\rfloor + 1$. We let $a^* := a^*(\phi^{-1}h, l)$ be as defined in (4.40) so that for any l, $\nu(a^*, l) \geq \frac{\phi^2}{\lambda_0 h^2}$.

By assumption of Proposition 4.5.6, it holds that $P_1 \overset{\mathcal{G}}{\prec} P_2 \overset{\mathcal{G}}{\prec} \ldots \overset{\mathcal{G}}{\prec} P_{|\mathcal{L}(p)|}$ where we recall $\mathcal{G} = \mathcal{G}_{t,u}$ is the already constructed graph - see Section 4.5.3.1. Hence, it holds that $\mathbf{rk}_{\mathcal{G},i}(j) \geq \Lambda$ for any $i \in P_l$ and $j \in P_{l+\Lambda}$ - see (4.19) for the definition of $\mathbf{rk}$. Since there are at least $p\Lambda \geq \frac{\phi^2}{\lambda_0 h^2}$ experts in the union $P_{l+1} \cup \cdots \cup P_{l+\Lambda}$, we conclude that $a^* \leq \Lambda$, and that any expert in $\mathcal{N}_{a^*}$ (resp. $\mathcal{N}_{-a^*}$) is below the maximal expert of $P_{l+\Gamma}$ (resp. above) the minimal expert of $P_{l-\Lambda}$. This implies that, upon writing $\overline{\mathbf{\Delta}}_k^*(l)$ for the difference of these maximal and minimal experts, we have by definition (4.41) of $\overline{Q}^{*h}$ that $\overline{\mathbf{\Delta}}_k^*(l) > h/2$ for all k in $\overline{Q}^{*h}$. This implies in particular that

$$\sum_{l \in \mathcal{L}(p)} |\overline{Q}_l^{*h}(h,\phi)| \leq \sum_{k=1}^d \sum_{l \in \mathcal{L}(p)} \mathbf{1}\{\overline{\mathbf{\Delta}}_k^*(l) \geq h/2\} \leq \frac{2}{h}\sum_{k=1}^d \sum_{l \in \mathcal{L}(p)} \overline{\mathbf{\Delta}}_k^*(l) \leq (2\Lambda+1)\frac{2d}{h} \leq 6\frac{\Lambda d}{h} \ , \tag{4.59}$$

where in the last inequality we used the fact that $M_{i,k} \in [0,1]$ and that the sequence $P_{l-\Lambda}, \ldots, P_{l+\Lambda}$ is of length $2\Lambda + 1$, for any $l \in \mathcal{L}(p)$. □

4.5.5 Proof of Proposition 4.5.10

Let us fix any $l \in \{1, \ldots, L\}$ and $h \in \mathcal{H}$. Since l, h and $\widehat{Q}_l^h$ are fixed in this proof, we simplify the notation and we write $(P', \widetilde{P}, Q) = (P_l', \widetilde{P}_l^h, \widehat{Q}_l^h)$ and $M := M(\widetilde{P}, Q)$ and $M(P') := M(P', Q)$. We also fix $\delta' = \delta/(L|\mathcal{H}|)$, where we recall that $L \leq n$ is the number of groups.

Let us assume that

$$\|M(P') - \overline{M}(P')\|_F^2 \geq \ \kappa_1 \gamma_u^4 \left[\frac{1}{\lambda_0}\sqrt{|\widetilde{P}||Q|} + \frac{|\widetilde{P}|}{\lambda_0} \right] \ , \tag{4.60}$$

for some constant κ_1 to be fixed later. In what follows, we show that under assumption (4.60) for some large enough numerical constant κ_1, we necessarily have that the square norm of P' is a contraction of the square norm of P, that is

$$\|M(P') - \overline{M}(P')\|_F^2 \leq \left(1 - \frac{1}{4\gamma_u^2}\right)\|M - \overline{M}\|_F^2 \ . \tag{4.61}$$

Step 1: control of the vector $\hat{v}$

First, the following lemma states that the first singular value of $(M - \overline{M})$ is, up to polylogarithmic terms, of the same order as its Frobenius norm. This is mainly due to the fact that the entries of M lie in $[0,1]$ and that $M - \overline{M}$ is an isotonic matrix.

Lemma 4.5.12 (Lemma E.4 in [74]). *Assume that $\|M - \overline{M}\|_F \geq 2$. For any sets $\widetilde{P}$ and Q, we have*

$$\|M - \overline{M}\|_{\text{op}}^2 \geq \frac{4}{\gamma_u^2}\|M - \overline{M}\|_F^2 \ .$$

This lemma was already stated and proved as Lemma E.4 in [74], recalling that $\gamma_u > \phi_{\text{L}_1} \geq 8\log(nd)$ – see (4.9) and (4.10).

Now, write $\hat{v} = \arg\max_{\|v\|_2 \leq 1} \left[\|v^T(Y^{(2)} - \overline{Y}^{(2)})\|_2^2 - \frac{1}{2}\|v^T(Y^{(2)} - \overline{Y}^{(2)} - Y^{(3)} + \overline{Y}^{(3)})\|_2^2\right]$, where the argmax is taken over all v in $\widetilde{P}$.

Lemma 4.5.13. *Assume that $\lambda_0|\widetilde{P}| \geq 1$. There exists a numerical constant κ_0' such that if*

$$\|M - \overline{M}\|_{\text{op}}^2 \geq \kappa_0' \log^2(nd/\delta')\left(\frac{1}{\lambda_0}\sqrt{|Q||\widetilde{P}|} + \frac{|\widetilde{P}|}{\lambda_0}\right) \ , \tag{4.62}$$

then, with probability higher than $1 - \delta'$, we have

$$\|\hat{v}^T\left(M - \overline{M}\right)\|_2^2 \geq \frac{1}{2}\|M - \overline{M}\|_{\text{op}}^2 \ .$$

In light of Lemma 4.5.12 and Condition (4.60), the Condition (4.62) in Lemma 4.5.13 is valid if we choose κ_1 in Proposition 4.5.10 such that $\kappa_1 \geq 16\kappa_0'$. Consequently, there exists an event of probability higher than $1 - \delta'$ such that

$$\|\hat{v}^T\left(M - \overline{M}\right)\|_2^2 \geq \frac{2}{\gamma_u^2}\|M - \overline{M}\|_F^2 \ . \tag{4.63}$$

Step 2: control of the vector $\hat{v}_-$

Now remark that since $\|\hat{v}_i\|_2 = 1$, there are at most $\frac{1}{\lambda_0}$ of experts i such that $\hat{v}_i > \sqrt{\lambda_0}$. Hence, we have that

$$\begin{aligned}
\|\hat{v}_-^T\left(M - \overline{M}\right)\|_2^2 &\geq \frac{2}{\gamma_u^2}\|M - \overline{M}\|_F^2 - \sum_{i \in \widetilde{P}} \mathbf{1}_{\hat{v}_i > \sqrt{\lambda_0}}\|M_{i\cdot} - \overline{m}\|_2^2 \\
&\overset{(a)}{\geq} \frac{2}{\gamma_u^2}\|M - \overline{M}\|_F^2 - \frac{3\gamma_u}{\lambda_0}\sqrt{\frac{|\widehat{Q}|}{\lambda_0}} \\
&\overset{(b)}{\geq} \frac{1}{\gamma_u^2}\|M - \overline{M}\|_F^2 \ .
\end{aligned}$$

(a) comes from the fact that any expert in $\widetilde{P}$ satisfies (4.43) under the event of Lemma 4.5.9. (b) comes from Condition (4.60) and the assumption that $\lambda_0|\widetilde{P}| \geq 1$.

Step 3: control of the vector $\hat{w}$

Next, we show that a thresholded version of $\hat{z} = (Y^{(4)} - \overline{Y}^{(4)})^T\hat{v}_-$ is almost aligned with $z^* = \lambda_1(M - \overline{M})^T\hat{v}_-$. We define the sets $S^* \subset Q$ and $\hat{S} \subset Q$ of questions by

$$S^* = \left\{k \in Q \ : \ |z_k^*| \geq 2\gamma_u\sqrt{\lambda_0}\right\} \ ; \quad \hat{S} = \left\{k \in Q \ : \ |\hat{z}_k| \geq \gamma_u\sqrt{\lambda_0}\right\} \ . \tag{4.64}$$

S^* stands for the collection of questions k such that z_k^* is large whereas $\hat{S}$ is the collection questions k with large $\hat{z}_k$. Finally, we consider the vectors w^* and $\hat{w}$ defined as theresholded versions of z^* and $\hat{z}$ respectively, that is $w_k^* = z_k^* 1_{k \in S^*}$ and $\hat{w}_k = \hat{z}_k \mathbf{1}_{k \in \hat{S}}$. Note that, up to the sign, $\hat{w}$ stands for the active coordinates computed in **SLR**, Line 7 of Algorithm 15.

Recall that we assume that $\lambda_0 \leq 1$. We write v for any unit vector in $\mathbb{R}^{|\widetilde{P}|}$. Let us apply Lemma 4.5.7 for each column $k \in Q$ of the noise matrix E with the matrix W equal to $v - (\frac{1}{|\widetilde{P}|}\sum_{i \in \widetilde{P}} v_i)\mathbf{1}_{\widetilde{P}}$ at column k and 0 elsewhere. We deduce that, for any fixed matrix M, any subsets $\widetilde{P}$ and Q, and any unit vector $v \in \mathbb{R}^{\widetilde{P}}$ such that $\|v\|_\infty \leq 2\sqrt{\lambda_0}$, we have

$$\mathbb{P}\left[\max_{k \in Q}\left|(v^T(E^{(3)} - \overline{E}^{(3)}))_k\right| \leq 100\log(2|Q|/\delta')\sqrt{\lambda_0}\right] \geq 1 - \delta' \ . \tag{4.65}$$

Observe that $\hat{z} = z^* + (E^{(3)} - \overline{E}^{(3)})^T \hat{v}_-$. Conditioning on $\hat{v}_-$, we deduce that, on an event of probability higher than $1 - \delta'$, we have

$$\|\hat{z} - z^*\|_\infty \le 100 \log(2|Q|/\delta')\sqrt{\lambda_0} \le \frac{\gamma_u}{2}\sqrt{\lambda_0} \ , \tag{4.66}$$

where the last inequality comes from $\gamma_u > \phi_{\mathrm{L}_1}$. Hence it holds that $S^* \subset \hat{S}$ and for $k \in \hat{S}$, we have $z^*_k/\hat{z}_k \in [1/2, 2]$. Next, we shall prove that, under this event, $\lambda_1 \hat{v}_-^T(M - \overline{M})\hat{w}/\|\hat{w}\|_2$ is large (in absolute value):

$$\lambda_1 \left|\hat{v}_-^T(M - \overline{M})\hat{w}\right| = \left|(z^*)^T\hat{w}\right| = \sum_{k\in\hat{S}} z^*_k \hat{z}_l \ge \frac{2}{5}\sum_{k\in\hat{S}}(z^*_k)^2 + (\hat{z}_l)^2 \ge \frac{2}{5}[\|w^*\|_2^2 + \|\hat{w}\|_2^2] \ge \frac{4}{5}\|\hat{w}\|_2\|w^*\|_2 \ ,$$

where we used in the first inequality that $z^*_k/\hat{z}_k \in [1/2, 2]$ and in the second inequality that $S^* \subset \hat{S}$. Thus, it holds that

$$\lambda_1^2 \left|\hat{v}_-^T(M - \overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \ge \frac{16}{25}\|w^*\|_2^2 \ . \tag{4.67}$$

It remains to prove that $\|w^*\|_2$ is large enough. Writing S^{*c} for the complementary of S^* in Q, it holds that

$$\|w^*\|_2^2 = \|z^*\|_2^2 - \sum_{k\in S^{*c}} (z^*_k)^2 \ , \tag{4.68}$$

so that we need to upper bound the latter quantity. Write $z^*_{S^{*c}} = z^* - w^*$. Coming back to the definition of z^*,

$$\begin{aligned}
\left[\sum_{k\in S^{*c}} (z^*_k)^2\right]^2 &= \left[\sum_{k\in S^{*c}} \lambda_1 [\hat{v}_-^T(M - \overline{M})]_k z^*_k\right]^2 \\
&\le \|\lambda_1 \left(M - \overline{M}\right) z^*_{S^{*c}}\|_2^2 = \sum_{i\in\widetilde{P}} \left(\sum_{k\in S^{*c}} \lambda_1 (M_{ik} - \overline{m}_k) z^*_k\right)^2 \\
&\overset{(a)}{\le} \frac{4\gamma_u^2}{|\widetilde{P}|^2}\lambda_0 \sum_{i\in\widetilde{P}} \left(\sum_{k\in S^{*c}}\sum_{j\in\widetilde{P}} \lambda_1 |M_{ik} - M_{jk}|\right)^2 \\
&\le \frac{4\gamma_u^2}{|\widetilde{P}|^2}\lambda_0 \sum_{i\in\widetilde{P}} \left(\sum_{j\in\widetilde{P}} \lambda_1 \|M_{i\cdot} - M_{j\cdot}\|_1\right)^2 \\
&\overset{(b)}{\le} 40\gamma_u^4\lambda_0^2|\widetilde{P}||Q| \\
&\le \left[7\gamma_u^2\lambda_0\sqrt{|\widetilde{P}||Q|}\right]^2 \le \left[\frac{1}{2\gamma_u^2}\lambda_0^2\|M - \overline{M}\|_F^2\right]^2 \ .
\end{aligned}$$

In (a), we used the definition of S^*. In (b), we used (4.43) that holds true since we are under the event Lemma 4.5.8 and $\lambda_0|Q| \ge 1$. The last inequality comes from Condition (4.60), choosing $\kappa_1 \ge 14$.

Recall that $z^* = \hat{v}_-^T(M - \overline{M})$. Combining (4.63) and (4.68), we deduce that

$$\|w^*\|_2^2 \ge \frac{1}{2\gamma_u^2}\lambda_0^2\|M - \overline{M}\|_F^2 \ , \tag{4.69}$$

which, together with (4.67) and $\lambda_0 \ge \lambda_1$, yields

$$\left\|(M - \overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \left|\hat{v}_-^T(M - \overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right|^2 \ge \frac{1}{2\gamma_u^2}\|M - \overline{M}\|_F^2 \ . \tag{4.70}$$

Write $\hat{w}^{(1)}$ and $\hat{w}^{(2)}$ the positive and negative parts of $\hat{w}$ respectively so that $\hat{w} = \hat{w}^{(1)} - \hat{w}^{(2)}$ and $\hat{w}^+ = \hat{w}^{(1)} + \hat{w}^{(2)}$. We obviously have $\|\hat{w}\|_2 = \|\hat{w}^+\|_2$. Besides, if the rows of M are ordered according to the oracle permutation, then $(M - \overline{M})\hat{w}^{(1)}$ and $(M - \overline{M})\hat{w}^{(2)}$ are nondecreasing vectors with mean zero. It then follows from Harris' inequality that these two vectors have a nonegative inner product. We have proved that

$$\left\|(M - \overline{M})\frac{\hat{w}^+}{\|\hat{w}^+\|_2}\right\|_2^2 \ge \left\|(M - \overline{M})\frac{\hat{w}}{\|\hat{w}\|_2}\right\|_2^2 \ge \frac{1}{2\gamma_u^2}\|M - \overline{M}\|_F^2 \ . \tag{4.71}$$

Step 4: Showing that $\hat{w}$ satisfies Condition (4.16)

Recall that we assume for simplicity that $\lambda_0 \le 1$. First we upper bound $\|w\|_\infty^2$ by using (a) that $\hat{z}$ is close to z^* with (4.66), (b) that for any $k \in Q$, $v^T M_k \le \|v\|_1$ and (c) that $\lambda_0|\widetilde{P}| \ge 1$:

$$\|\hat{w}\|_\infty^2 \overset{(a)}{\le} 2\|z^*\|_\infty^2 + \gamma_u^2\lambda_0 \overset{(b)}{\le} 2\lambda_0^2\|\hat{v}\|_1^2 + \gamma_u^2\lambda_0 \overset{(c)}{\le} 3\gamma_u^2\lambda_0^2|\widetilde{P}| \ . \tag{4.72}$$

Secondly, we lower bound $\|w\|_2^2$ by using (a) that $S^* \subset \hat{S}$ and that $z_k^*/\hat{z}_k \in [1/2, 2]$, (b) that $\|w^*\|_2^2$ captures a significant part of the L_2 norm -see (4.69), and (c) the Condition (4.60) with $\kappa_1 \ge 24$:

$$\|\hat{w}\|_2^2 \overset{(a)}{\ge} \frac{1}{4}\|w^*\|_2^2 \overset{(b)}{\ge} \frac{1}{8\gamma_u^2}\lambda_0^2\|M - \overline{M}\|_F^2 \overset{(c)}{\ge} 3\gamma_u^2\lambda_0|\widetilde{P}| \ . \tag{4.73}$$

We deduce that $\|\hat{w}\|_\infty^2 \le \lambda_0\|\hat{w}\|_2^2$, which is exactly Condition (4.16). This shows that $\hat{w}^+$ is considered for the update (4.18) in the final step of the procedure Line 9 of Algorithm 15.

Step 5: upper bound of the Frobenius norm restricted to P'

Equipped with this bound, we are now in position to show that the set P' of experts obtained from $\widetilde{P}$ when applying the pivoting algorithm with $\hat{w}^+/\|\hat{w}^+\|_2$ has a much smaller square norm. By Lemma 4.5.7 used with the matrix W equal to 0 except at line i where it is equal to the vector $\hat{w}^+/\|\hat{w}^+\|_2$, there exists an event of probability higher than $1 - \delta'$ such that

$$\max_{i,j\in P'} \left| \langle E_{i\cdot} - E_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2} \rangle \right| \le \phi_{\mathrm{L}_1}\sqrt{\lambda_0} \le \gamma_u\sqrt{\lambda_0} \ ,$$

where we recall that ϕ_{l_1} is defined in (4.9). Hence, since the vector $\hat{w}$ is considered in the update (4.18), we have $\max_{i,j\in P'} \left| \langle Y_{i\cdot} - Y_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2} \rangle \right| \le \gamma_u\sqrt{\lambda_0}$ and

$$\max_{i,j\in P'} \left| \langle M_{i\cdot} - M_{j\cdot}, \frac{\hat{w}^+}{\|\hat{w}^+\|_2} \rangle \right| \le 2\gamma_u\sqrt{\frac{1}{\lambda_0}} \ . \tag{4.74}$$

By convexity, it follows that

$$\left\| (M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2} \right\|_2^2 \le 4\gamma_u^2\frac{1}{\lambda_0}|P'| \le 4\gamma_u^2\frac{1}{\lambda_0}|\widetilde{P}| \ .$$

In light of Condition (4.60), this quantity is small compared to $\|M - \overline{M}\|_F^2$:

$$\|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \le \frac{1}{4\gamma_u^2}\|M - \overline{M}\|_F^2 \ , \tag{4.75}$$

which together with (4.71) leads to

$$\|(M - \overline{M})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \ge \frac{1}{4\gamma_u^2}\|M - \overline{M}\|_F^2 \ . \tag{4.76}$$

Since $P' \subset \widetilde{P}$, we deduce that, for any vector $w' \in \mathbb{R}^q$, we have $\|(M - \overline{M})w'\|_2^2 \ge \|(M(P') - \overline{M}(P'))w'\|^2$. It then follows from the Pythagorean theorem that

$$\|M - \overline{M}\|_F^2 - \|M(P') - \overline{M}(P')\|_F^2 \ge \|(M - \overline{M})\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 - \|(M(P') - \overline{M}(P'))\tfrac{\hat{w}^+}{\|\hat{w}^+\|_2}\|_2^2 \ .$$

Then, together with (4.76), we arrive at

$$\|M(P') - \overline{M}(P')\|_F^2 \le \left(1 - \frac{1}{4\gamma_u^2}\right)\|M - \overline{M}\|_F^2 \ .$$

We have shown that if (4.60) is satisfied, then there is a contraction in the sense of (4.61). This in turn gives the upper bound (4.46), and it concludes the proof of Proposition 4.5.10.

Proof of Lemma 4.5.13. Recall that we consider the case $\lambda_0 \le 1$ and that the case $\lambda_0 \ge 1$ is discussed in Section 4.5.6. We start with the two following lemmas. To ease the notation, we assume in this proof that $\widetilde{P} = \{1, \dots, p\}$, that $Q = \{1, \dots q\}$. We only consider the matrices restricted to the sets $\widetilde{P}, Q$ and we write $E := E(\widetilde{P}, Q)$. Let us define $J = \mathbf{1}\mathbf{1}^T \in \mathbb{R}^{p\times p}$ the matrix with constant coefficients equals to 1 and $A = (\mathbf{I}_p - \frac{1}{p}J)$ be the projector on the orthogonal of $\mathbf{1}$, so that $E - \overline{E} = AE \in \mathbb{R}^{p\times q}$. The two following lemmas are direct consequences of Proposition 4.4.1, and a discussion of the corresponding concenration inequality on random rectangular matrices can be found in Section 4.4. We state weaker concentration inequalities than what is proven in Proposition 4.4.1 in order to factorize the polylogarithmic factors and to ease the reading of the proof.

Lemma 4.5.14. *Assume that $\lambda_0 \le 1$ and that $\lambda_0(p \vee q) \ge 1$. It holds with probability larger than $1 - \delta'/4$ that*

$$\|EE^T - \mathbb{E}[EE^T]\|_{\mathrm{op}} \le \kappa_0'' \log^2(pq/\delta')\,[\lambda_0\sqrt{pq} + \lambda_0 p] \quad .$$

Lemma 4.5.15. *Assume that $\lambda_0 \le 1$ and that $\lambda_0(p \vee q) \ge 1$. With probability larger than $1 - \delta'/4$, one has for any orthogonal projection $\Lambda \in \mathbb{R}^{q\times q}$ satisfying $\mathrm{rank}(\Lambda) \le p$ that*

$$\|\Lambda E^T E \Lambda\|_{\mathrm{op}} \le \kappa_1'' \log^2(pq/\delta')\,[\lambda_0\sqrt{pq} + \lambda_0 p] \quad ,$$

Proofs of Lemma 4.5.14 and Lemma 4.5.15. First, we recall that for any i, k, we have that $E_{ik} = (B_{ik} - \lambda_1)M_{ik} + \widetilde{E}_{ik}$, and that $\widetilde{E}$ is an average of 1-subGaussian random variables, as described in (4.30) For any $u \ge 0$ we have

$$\mathbb{E}[E_{ik}^{2u}] \le 3^u \mathbb{E}\left[B_{ik} + \lambda_0^{2u} + \widetilde{E}_{ik}^{2u}\right] \le 3^u \left(2\lambda_0 + u!\mathbb{E}[e^{\widetilde{E}_{ik}^2}]\right) \le \frac{1}{2}u!\lambda_0 1000^u \quad , \tag{4.77}$$

where for the last inequality we used the following inequalities:

$$\mathbb{E}[e^{\widetilde{E}_{ik}^2}] \le \sum_{u\ge 1} e^{-\lambda_0}\frac{\lambda_0^u}{u!}e^{1/u} \le \lambda_0 e \quad .$$

Hence, condition (4.24) is satisfied with $K = 1000$ and $\sigma^2 = \lambda_0$ for the coefficients of E. We just apply Proposition 4.4.1 with $X = E$ for Lemma 4.5.14. For Lemma 4.5.15, we apply Proposition 4.4.1 with $X = E^T$ and we remark that $\|\Lambda E^T E\Lambda\|_{\mathrm{op}}^2 \le 2\|\Lambda E^T E - \mathbb{E}[E^T E]\Lambda\|_{\mathrm{op}}^2 + 2\|\mathbb{E}[E^T E]\|_{\mathrm{op}}^2$ together with the fact that $\|\mathbb{E}[E^T E]\|_{\mathrm{op}}^2 \le c'\lambda_0 p$ for some numerical constant c'. □

Remark that since we assume in Lemma 4.5.13 that $\lambda_0 p \ge 1$, it holds that $\sqrt{\lambda_0 p} \le \lambda_0 p$ and $\sqrt{\lambda_0 q} \le \lambda_0^2\sqrt{pq}$, so that both upper bounds of Lemma 4.5.14 and Lemma 4.5.15 reduce - up to logarithmic factors - to $\lambda_0\sqrt{pq} + \lambda_0 p$. We write for short in the following

$$F := F(p, q, \lambda_0, \delta') = \log^2(pq/\delta')[\lambda_0\sqrt{pq} + \lambda_0 p] \quad , \tag{4.78}$$

and $\kappa_2'' = 8(\kappa_0'' \vee \kappa_1'')$.

Now let us write

$$AY = \lambda_1 AM + AE \quad ,$$

so that, for any $v \in \mathbb{R}^p$, recalling that $AY = Y - \overline{Y}$,

$$\|v^T AY\|_2^2 = \lambda_1^2\|v^T AM\|_2^2 + \|v^T AE\|_2^2 + 2\lambda_1\langle v^T AE, v^T AM\rangle \quad ,$$

which, in turn, implies that

$$\begin{aligned}\left|\|v^T AY\|_2^2 - \lambda_1^2\|v^T AM\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| &\le \left|\|v^T AE\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| + 2\lambda_1|v^T AME^T(Av)| \\ &\overset{(a)}{\le} \|A(EE^T - \mathbb{E}[EE^T])A\|_{\mathrm{op}} + 2\lambda_1\|AME^T E(AM)^T\|_{\mathrm{op}}^{1/2} \\ &\le \|EE^T - \mathbb{E}[EE^T]\|_{\mathrm{op}} + 2\lambda_1\|AM\|_{\mathrm{op}}\|\Lambda E^T E\Lambda\|_{\mathrm{op}}^{1/2} \quad ,\end{aligned}$$

Where we define $\Lambda \in \mathbb{R}^{d\times d}$ as the orthogonal projector on the image of $\ker(AM)^\perp$ which is of rank less than p. For (a), we used the fact that A is contracting the operator norm as an orthogonal projector so that $\|Av\|_2 \le 1$. We now apply Lemma 4.5.14 and Lemma 4.5.15 together with the fact that $\lambda_1 \le \lambda_0$, and we obtain with probability at least $1 - \delta'/2$ that

$$\sup_{v\in\mathbb{R}^p, \|v\|=1}\left|\|v^T AY\|_2^2 - \lambda_1^2\|v^T AM\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| \le \kappa_2'' F + \lambda_1\|AM\|_{\mathrm{op}}\sqrt{\kappa_2'' F} \quad . \tag{4.79}$$

where F is defined in (4.78). In the same way, we have that, with probability larger than $1 - \delta'/2$,

$$\begin{aligned}\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\frac{1}{2}\|v^T A(Y - Y')\|_2^2 - \mathbb{E}\left[\|v^T AE\|_2^2\right]\right| &= \frac{1}{2}\sup_{v\in\mathbb{R}^p:\ \|v\|_2\le 1}\left|\|v^T A(Y - Y')\|_2^2 - \mathbb{E}\|v^T A(Y - Y')\|_2^2\right| \\ &\le \kappa_3'' F \quad ,\end{aligned}$$

for some numerical constant κ_3''. Putting everything together we conclude that, on an event of probability higher than $1 - \delta'$, we have simultaneously for all $v \in \mathbb{R}^p$ with $\|v\|_2 \le 1$ that

$$\left|\|v^T AY\|_2^2 - \|v^T AM\|_2^2 - \frac{1}{2}\|v^T A(Y - Y')\|_2^2\right| \le \kappa_4'' F + \lambda_1\|AM\|_{\mathrm{op}}\sqrt{\kappa_4'' F} \quad ,$$

with $\kappa_4'' = \kappa_2'' \vee \kappa_3''$. Choosing the numerical constant κ_0' of Lemma 4.5.13 such that $\kappa_0' \geq 4 \cdot 16(1-1/e)^{-1}\kappa_4''$ we have

$$\lambda_1^2 \|AM\|_{\mathrm{op}}^2 \geq 4 \cdot 16 \kappa_4'' F \ ,$$

since it holds that $\lambda_1 \geq (1-1/e)\lambda_0$. We deduce that on the same event:

$$\sup_{v \in \mathbb{R}^p:\ \|v\|_2 \leq 1} \left| \|v^T AY\|_2^2 - \|v^T AM\|_2^2 - \frac{1}{2}\|v^T A(Y-Y')\|_2^2 \right| \leq \frac{1}{4}\|AM\|_{\mathrm{op}}^2 \ .$$

Writing $\psi(v) = \left| \|v^T(Y-\overline{Y})\|_2^2 - \frac{1}{2}\|v^T A(Y-Y')\|_2^2 \right|$, we deduce that, for v such that $\|v^T AM\|_2^2 = \|AM\|_{\mathrm{op}}^2$, we have $\Psi(v) \geq \frac{3}{4}\|AM\|_{\mathrm{op}}^2$, whereas, for v such that $\|v^T AM\|_2^2 < \frac{1}{2}\|AM\|_{\mathrm{op}}^2$, we have $\Psi(v) < \frac{3}{4}\|AM\|_{\mathrm{op}}^2$. We conclude that $\hat{v}$ satisfies $\|\hat{v}^T AM\|_2^2 > \frac{1}{2}\|AM\|_{\mathrm{op}}^2$ with probability at least $1-\delta'$.

□

4.5.6 Proof of Theorem 4.2.2 when $\lambda_0 \geq 1$

The aim of this section is to provide an extension of the proof of Theorem 4.2.2 to the case $\lambda_0 \geq 1$. Recall that we fix δ to be a small probability the proof of Theorem 4.2.2, and that E and $\widetilde{E}$ are the matrices defined in (4.29) and (4.30) by

$$\widetilde{E}_{ik}^{(s)} = \sum_{t \in N^{(s)}} \frac{\varepsilon_t}{\mathbf{r}_{ik}^{(s)} \vee 1} \mathbf{1}\{x_t = (i,k)\} \quad \text{and} \quad E_{ik}^{(s)} = (B_{ik}^{(s)} - \lambda_1)M + B_{ik}^{(s)} \widetilde{E}_{ik}^{(s)} \ .$$

In what follows, we consider the two subcases where $\lambda_0 > 16\log(5nd/\delta)$ or $\lambda_0 \leq 16\log(5nd/\delta)$, which essentially rely on the two following ideas:

- If $\lambda_0 \leq 16\log(5nd/\delta)$, we use the fact that the coefficients of E defined in (4.29) are 5-subGaussian together with the same signal-noise decomposition $Y = \lambda_1 M + E$ as in the proofs when $\lambda_0 \leq 1$. The difference from the case $\lambda_0 \leq 1$ lies in the application of subGaussian inequalities of E_{ik} instead of Bernstein inequalities as in (4.37).
- If $\lambda_0 > 16\log(5nd/\delta)$, we show that the event $\{\mathbf{r}_{ik}^{(s)} \geq \lambda_0/2\}$ holds true for all i,k,s with high probability. Working conditionally to this event, we use the decomposition $Y = M + \widetilde{E}$, and we show that the noise $\widetilde{E}$ has $\frac{2}{\lambda_0}$-subGaussian independent coefficients. The rationale behind using $\widetilde{E}$ when λ_0 is large is that $\widetilde{E}_{ik}$ takes advantage of the mean of $2/\lambda_0$ subGaussian variables with high probability.

Let $\mathbf{r}_{\min}^{(s)} = \min_{i,k} \mathbf{r}_{ik}^{(s)}$ be the minimum number of observation at positions (i,k) in N_s - see (4.14). In the case $\lambda_0 > 16\log(5nd/\delta)$, the following lemma states that with high probability, we observe all the coefficients for all sample s in the full observation regime.

Lemma 4.5.16. *Assume that $\lambda_0 \geq 16\log(5nd/\delta)$. The event $\{\mathbf{r}_{\min}^{(s)} \geq \lambda_0/2\}$ holds simultaneously for all sample s with probability at least $1-5T\delta$.*

Proof of Lemma 4.5.16. We apply Chernoff's inequality - see e.g. section 2.2 of [62] - to derive that for any i,k

$$\mathbb{P}(\mathbf{r}_{ik}^{(s)} \leq \lambda_0/2) \leq \exp(-\frac{1}{8}\lambda_0) \leq \delta/(nd) \ , \tag{4.80}$$

where we use the inequality $(1-\log(2))/2 \geq 1/8$. We conclude with a union bound over all coefficients in $[n]\times[d]$ and all $5T$ samples. □

Let us now omit the dependence of E and $\widetilde{E}$ in the sample s. In what follows, use that the coefficients of E are 5-subGaussian, which is a consequence of the fact that E_{ik} is the sum of a centered variable bounded by 1 and a 1-subgaussian random variable $\widetilde{E}_{ik}$, so that by Cauchy-Schwarz and the Hoeffding inequality we have

$$\mathbb{E}[\exp(xE_{ik})] \leq \sqrt{\exp(4x^2/8)}\sqrt{\exp(4x^2/2)} = \exp(5/4x^2) \ . \tag{4.81}$$

Under the event of Lemma 4.5.16, we use that $\widetilde{E}_{ik}$ is $\lambda_0/2$-subGaussian, as an average of at least $2/\lambda_0$ random variables that are 1-subGaussians:

$$\mathbb{E}[\exp(x\widetilde{E}_{ik})] \leq \exp(\tfrac{1}{\lambda_0}x^2) \ , \tag{4.82}$$

4.5.6.1 Adjustements for the general analysis

We first make the changes that should be done in Section 4.5.1 to have a proper proof in the case $\lambda_0 \geq 1$.

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, we simply replace λ_0 by $1/\lambda_0$ in the upper bound of (4.31) for the event ξ in Lemma 4.5.1. In the proof of the restated Lemma 4.5.1, we can replace the inequality (4.37) by

$$|\langle E, W\rangle| \leq \sqrt{10\|W\|_F^2 \log\left(\frac{2}{\delta'}\right)}\ , \tag{4.83}$$

for any matrix $W \in \mathbb{R}^{n\times d}$, with probability at least $1-\delta'$. We can then obtain $1/\lambda_0$ instead of λ_0 simply by using that $\phi_{\mathrm{L}_1}/\sqrt{\lambda_0} \geq \sqrt{\phi_{\mathrm{L}_1}}$, recalling that ϕ_{L_1} is defined in (4.9).

If $\lambda_0 > 16\log(5nd/\delta)$, we say that we are under event ξ if the event of Lemma 4.5.16 holds and (4.31) holds for all pairs (Q, w), replacing E by $\widetilde{E}$, and λ_0 by $1/\lambda_0$. The proof of the new version of Lemma 4.5.1 lies in the Hoeffding inequality applied to $\widetilde{E}$ under the event of Lemma 4.5.16, leading to the subsequent equation:

$$|\langle \widetilde{E}, W\rangle| \leq \sqrt{\frac{4\|W\|_F^2}{\lambda_0} \log\left(\frac{2}{\delta'}\right)}\ , \tag{4.84}$$

for any matrix $W \in \mathbb{R}^{n\times d}$, with probability at least $1-\delta'$. This equation then replaces (4.37).

4.5.6.2 Adjustments to the proofs of Proposition 4.5.6

We now adapt the proofs in Section 4.5.3 of Proposition 4.5.6 to the case $\lambda_0 \geq 1$.

All the lemmas of Section 4.5.3 can be stated as is for any $\lambda_0 \geq 1$, and the only adjustments concern the proofs of Lemma 4.5.8, Lemma 4.5.9 and Proposition 4.5.10.

4.5.6.3 Adjustments in the proofs of Lemma 4.5.8 and Lemma 4.5.9

Consider the proof of Lemma 4.5.8. First, if $\lambda_0 \geq 16\log(5nd/\delta)$, we place ourselves under the event Lemma 4.5.16 and replace λ_1 by 1 and all the E by $\widetilde{E}$. Instead of inequality (4.55), we use the fact that the coefficients of $\widetilde{E}$ are $2/\lambda_0$-subGaussian - see (4.82) - leading to the following inequality with probability at least $1-\delta$:

$$\left|\widetilde{E}_k(a)\right| := \left|\frac{1}{|\mathcal{N}_a|}\sum_{i\in\mathcal{N}_a}\widetilde{E}_{ik} - \frac{1}{|\mathcal{N}_{-a}|}\sum_{i\in\mathcal{N}_{-a}}\widetilde{E}_{ik}\right| \leq \kappa_0' \log(nd/\delta)\sqrt{\frac{1}{\lambda_0\nu(a)}}\ , \tag{4.85}$$

for some numerical constant κ_0'. The rest of the proof remains unchanged.

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, we use the fact that E has 5-subGaussians coefficients - see (4.81) and we do not divide by λ_0 in (4.57) - see the definition of $\widehat{\mathbf{\Delta}}$ (4.21).

Concerning Lemma 4.5.9, the adjustments are the same as for Lemma 4.5.1, namely working under the event of Lemma 4.5.16, replacing E by $\widetilde{E}$, λ_0 by $1/\lambda_0$ and λ_1 by 1 if $\lambda_0 \geq 16\log(5nd/\delta)$, and using the fact that the coefficient of E are 5-subGaussians - see (4.81) if $\lambda_0 \in [1, 16\log(5nd/\delta)]$.

4.5.6.4 Adjustments in the proof of Proposition 4.5.10

We now adapt the proofs in Section 4.5.5 of Proposition 4.5.10 to the case $\lambda_0 \geq 1$. First, Lemma 4.5.13 can be stated as is, and its proof when $\lambda_0 \geq 1$ is directly implied by Lemma E.5 in [74] with $\Theta := M$ either conditionally on Lemma 4.5.16 with noise $N := \widetilde{E}$ and $\zeta^2 := 2/\lambda_0$ when $\lambda_0 \geq 16\log(5nd/\delta)$ or with noise $N := E$ and $\zeta^2 := 5$ when $\lambda_0 \leq 16\log(5nd/\delta)$.

Secondly, remark that if $\lambda_0 \geq 1$, it holds that $\hat{v}_- = \hat{v}$ and that Condition (4.16) on $\hat{w}$ is automatically satisfied, so that step 2 and step 4 can be removed from the proof in that case. For Step 3 and 5, we do the following adjustments:

If $\lambda_0 \in [1, 16\log(5nd/\delta)]$, the proof remains unchanged except that we use that the coefficients of E are 5-subGaussian -see (4.81).

If $\lambda_0 \geq 16\log(5nd/\delta)$, we work conditionally on the event of Lemma 4.5.16 and we replace λ_1 by 1 and E by $\widetilde{E}$. The subgaussian concentration bound on $\widetilde{E}$ (4.84) allows us to replace λ_0 by $\frac{1}{\lambda_0}$ in the equations from (4.64) to (4.69).

4.5.7 Proof of Corollaries 4.2.4 and 4.2.5

Proof of Corollary 4.2.4. Assume that $\pi^* = \mathrm{id}$ for simplicity. Let P_{iso} be the projector on the set of isotonic matrices, and $E' = Y^{(2)}_{\hat{\pi}^{-1}} - M_{\hat{\pi}^{-1}}$ so that $\hat{M}_{\mathrm{iso}} = P_{\mathrm{iso}}(M_{\hat{\pi}^{-1}} + E')$. Remark that the loss can be decomposed as

$$\|(\hat{M}_{\mathrm{iso}})_{\hat{\pi}} - M\|_F^2 = \|P_{\mathrm{iso}} M_{\hat{\pi}^{-1}} - P_{\mathrm{iso}} M + P_{\mathrm{iso}}(M + E') - M + M - M_{\hat{\pi}^{-1}}\|_F^2 \ .$$

Using the non-expansiveness of P_{iso} and the triangular inequality as in the proof of proposition 3.3 of [60], we deduce that

$$\|\hat{M}_{\mathrm{iso}} - M\|_F^2 \le 4\|M_{\hat{\pi}^{-1}} - M\|_F^2 + 2\|P_{\mathrm{iso}}(M + E') - M\|_F^2 \ . \tag{4.86}$$

Since the projection of $M + E'$ on isotonic matrices is equal to the columnwise projection on isotonic vectors, it holds that $\sup_{M \in \mathbb{C}_{\mathrm{iso}}(n,d)} \mathbb{E}\|P_{\mathrm{iso}}(M + E') - M\|_F^2 = d \sup_{M \in \mathbb{C}(n,1)} \mathbb{E}\|P_{\mathrm{iso}}(M_{\cdot 1} + E'_{\cdot 1}) - M_{\cdot 1}\|_F^2$, where we also use the notation P_{iso} for the projector on isotonic vectors. The rate of estimation in L_2 norm of an isotonic vector with bounded total variation partial observation can be found in [107], with $V := 1$ and $\sigma^2 := 1/\lambda$. Hence, we obtain that $\sup_{M \in \mathbb{C}(n,1)} \mathbb{E}\|P_{\mathrm{iso}}(M_{\cdot 1} + E'_{\cdot 1}) - M_{\cdot 1}\|_F^2 \le C_1 n^{1/3}/\lambda^{2/3}$. Upper bounding the first term in (4.86) with a quantity of order $\rho_{\mathrm{perm}} \le 2\rho_{\mathrm{reco}}$ by Theorem 4.2.2 concludes the proof.

□

Proof of Corollary 4.2.5. We follow the same steps as in Corollary 4.2.4. Assume that $\pi^* = \eta^* = \mathrm{id}$, $E' = Y^{(3)}_{\hat{\pi}^{-1}\hat{\eta}^{-1}} - M$, and let P_{biso} be the projector on bi-isotonic matrices. We have that

$$\|(\hat{M}_{\mathrm{biso}})_{\hat{\pi}\hat{\eta}} - M\|_F^2 \le 4\|M_{\hat{\pi}^{-1}\hat{\eta}^{-1}} - M\|_F^2 + 2\|P_{\mathrm{biso}}(M + E') - M\|_F^2 \ . \tag{4.87}$$

M is isotonic in both directions so that we can apply Theorem 4.2.2 in rows and columns. After the first two steps of the above procedure, we obtain two estimator $\hat{\pi}, \hat{\eta}$ that satisfy

$$\sup_{\substack{\pi^*, \eta^* \in \Pi_n \\ M : M_{\pi^{*-1}\eta^{*-1}} \in \mathbb{C}_{\mathrm{biso}}}} \mathbb{E}\left[\|M_{\hat{\pi}^{-1}\hat{\eta}^{-1}} - M_{\pi^{*-1}\eta^{*-1}}\|_F^2\right] \le C'' \log^{C''}(n) n^{7/6} \lambda^{-5/6} \ . \tag{4.88}$$

The second term of (4.87) is the risk of a bi-isotonic regression by least square, and is smaller than $n/\lambda \le n^{7/6}\lambda^{-5/6}$ - see e.g. [60]. □

4.5.8 Proof of the minimax lower bound

Proof of Theorem 4.2.1. Since $\rho_{\mathrm{perm}}(n, d, \lambda)$ is nondecreasing with n and d, we can assume without loss of generality that both n and d express as a power of 2.

The following proof is strongly related to the proof of Theorem 4.1 in [74]. While a worst case distribution is defined on the set of matrices that have nondecreasing rows and nondecreasing columns in [74], we aim here at defining a worst case distribution on matrices only have nondecreasing columns. Since the isotonic model is less constrained than the bi-isotonic model studied in [74], the permutation estimation problem is statistically harder, and the lower bound has a greater order of magnitude.

As in [74], the general idea is first to build a collection of prior $\nu_{\mathbf{G}}$ indexed by some $\mathbf{G} \in \boldsymbol{\mathcal{G}}$ on M, then to reduce the problem to smaller problems and finally to specify the prior in function of the regime in n, d and λ. By assumption, the data y_t is distributed as a normal random variable with mean M_{x_t} and variance 1, conditionally on M and x_t. We write as in [74] $\mathbf{P}^{(\mathbf{full})}_{\mathbf{G}}$ and $\mathbf{E}^{(\mathbf{full})}_{\mathbf{G}}$ the corresponding marginal probability distributions and expectations on the data (x_t, y_t). Our starting point is the fact that the minimax risk (4.4) is higher than the worst Bayesian risk:

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge \inf_{\hat{\pi}} \sup_{\mathbf{G} \in \boldsymbol{\mathcal{G}}} \mathbf{E}^{\mathbf{full}}_{\mathbf{G}}\left[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2\right] \ . \tag{4.89}$$

Step 1: Construction of the prior distribution on M

Let $p \in \{2, \dots, n\}$ and $q \in [d]$ be two powers of 2 to be fixed later, and $\overline{G}^{(\iota)} := [(\iota - 1)p + 1, \iota p]$, for $\iota \in \{1, \dots, n/p\}$. The general idea is to build a simple prior distribution on isotonic matrices in $\mathbb{R}^{\overline{G}^{(\iota)} \times d}$, and to derive a prior distribution on isotonic matrices in $\mathbb{R}^{n \times d}$ by combining n/p independent simple prior distributions defined on each strip $\mathbb{R}^{\overline{G}^{(\iota)} \times d}$.

Let $w \in \mathbb{R}^n$ be a vector that is constant on each group $\overline{G}^{(\iota)} = [(\iota - 1)p + 1, \iota p]$ and that has linearly nondecreasing steps:

$$w_i = \left\lfloor \frac{i}{p} \right\rfloor \frac{p}{4n} \in [0, 1/4] \ . \tag{4.90}$$

Letting $\mathbf{1}_{[d]}$ be constant equal to 1 in $\mathbb{R}^d$, we define

$$M = w\mathbf{1}_{[d]}^T + \frac{\upsilon}{\sqrt{p\lambda}} B^{(\mathbf{full})} \ , \tag{4.91}$$

where the random matrix $B^{(\mathbf{full})} \in \{0,1\}^{n\times d}$ is defined as in [74]. We recall the definition of its distribution in what follows for the sake of completeness.

Consider a collection $\mathcal{G}$ of subsets of $[p]$ with size $p/2$ that are well-separated in symmetric difference as defined by the following lemma.

Lemma 4.5.17. *There exists a numerical constant c_0 such that the following holds for any even integer p. There exists a collection $\mathcal{G}$ of subsets of $[p]$ with size $p/2$ which satisfies $\log(|\mathcal{G}|) \geq c_0|p|$ and whose elements are $p/4$-separated, that is $|G_1 \Delta G_2| \geq p/4$ for any $G_1 \neq G_2$.*

The above result is stated as is in [74] and is a consequence of Varshamov-Gilbert's lemma - see e.g. [90].

For each $\iota \in [n/p]$, we fix a subset $G^{(\iota)}$ from $\mathcal{G}$, and its translation $G^{t(\iota)} = \{(\iota-1)p + x : x \in G^{(\iota)}\} \subset \overline{G}^{(\iota)}$. The experts of $G^{t(\iota)}$ will correspond the $p/2$ experts in $\overline{G}^{(\iota)}$ that are above the $p/2$ experts in $\overline{G}^{(\iota)} \setminus G^{t(\iota)}$. We write $\mathbf{G} = (G^{t(1)}, \dots, G^{t(n/p)})$ and $\boldsymbol{\mathcal{G}}$ the corresponding collection of all possible $\mathbf{G}$. Given any such $\mathbf{G}$, we shall define a distribution $\nu_{\mathbf{G}}$ of $B^{(\mathbf{full})}$, and equivalently of M by (4.91).

For $\iota \in [n/p]$, we sample uniformly a subset $Q^{(\iota)}$ of q questions among the d columns. In each of these q columns, the corresponding rows of $B^{(\mathbf{full})}$ are equal to one. More formally, we have

$$B^{(\mathbf{full})} = \sum_{\iota=1}^{n/p} \mathbf{1}_{G^{t(\iota)}} \mathbf{1}_{Q^{(\iota)}} \ . \tag{4.92}$$

As mentioned above, the definition of $B^{(\mathbf{full})}$ is the same as in [74], if $\tilde{d}$ is set to be equal to d. They define a block constant matrix when $\tilde{d} < d$ to get an appropriate prior distribution for bi-isotonic matrices, but we do not need to do that here since we do not put any constraint on the rows of M.

The matrix M defined in (4.92) is isotonic up to a permutation of its rows and has coefficients in $[0,1]$, if the following inequality is satisfied.

$$\frac{\upsilon}{\sqrt{p\lambda}} \leq \frac{p}{8n} \ . \tag{4.93}$$

This constraint is strictly weaker than its counterpart (149) in [74], and this is precisely what makes the lower bound in the isotonic setting larger than the lower bound in the bi-isotonic setting of [74]. Our purpose will be to wisely choose parameters p, q and $\upsilon > 0$ to maximize the Bayesian risk (4.89) with $\nu_{\mathbf{G}}$.

Step 2: Problem Reduction

In what follows, we use the same reduction arguments as in [74]. Using the notation of [74], we write $\mathbf{P}_{\mathbf{G}}^{(\mathbf{full})}$ and $\mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}$ for the probability distribution and corresponding expectation of the data (x_t, y_t), when M is sampled according to $\nu_{\mathbf{G}}$. Since the distribution of the rows of M in $\overline{G}^{t(\iota)}$ only depend on $G^{t(\iota)}$, we write $\nu_{G^{t(\iota)}}$ for the distribution of these rows. We also write $\mathbf{P}_{G^{t(\iota)}}^{(\mathbf{full})}$ and $\mathbf{E}_{G^{t(\iota)}}^{(\mathbf{full})}$ for the corresponding marginal distribution and corresponding expectation of the observations (x_t, y_t) such that $(x_t)_1 \in \overline{G}^{t(\iota)}$. By the Poissonization trick, the distribution $\mathbf{P}_{\mathbf{G}}^{(\mathbf{full})}$ is a product measure of $\mathbf{P}_{G^{t(\iota)}}^{(\mathbf{full})}$ for $\iota = 1, \dots, n/p$.

Let $\tilde{\pi}$ be any estimator of π^*. Let us provide more details than [74] to prove that $\tilde{\pi}$ can be modified into an estimator $\hat{\pi}$ satisfying $\hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$ for all $\iota = 1, \dots, n/p$, and reducing the loss $\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq \|M_{\tilde{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ almost surely, for all possible prior $\nu_{\mathbf{G}}$. For that purpose, we introduce

$$N(\pi) = \sum_{\iota=1}^{n/p} \sum_{i \in \overline{G}^{(\iota)}} \mathbf{1}\{i \notin \overline{G}^{(\iota)}\} \ .$$

If $N(\tilde{\pi}) > 0$, then there exists ι_0 and $i_0 \in \overline{G}^{(\iota_0)}$ such that $\tilde{\pi}(i_0) \in \overline{G}^{(\iota_1)}$ with $\iota_1 \neq \iota_0$. Then, $\tilde{\pi}$ being a permutation, we consider its associated cycle containing i_0, which we denote by $(i_1, \dots, i_K)$. Let $(i'_1, i'_2, \dots, i'_L)$ be the elements of this cycle such that $\tilde{\pi}(i'_l) \notin \overline{G}^{(\iota_l)}$, where ι_l satisfies $i'_l \in \overline{G}^{(\iota_l)}$. Then it holds that for any $l = 1, \dots, L-1$, $\tilde{\pi}(i'_l) \in \overline{G}^{(\iota_{l+1})}$, and $\tilde{\pi}(i'_L) \in \overline{G}^{(\iota_1)}$. We now define $\tilde{\pi}'(i) = \tilde{\pi}(i)$ for all i, except on the cycle $(i'_1, \dots, i'_L)$ where we set $\tilde{\pi}'(i'_l) = \tilde{\pi}(i'_{l-1})$. Then, we easily check that $N(\tilde{\pi}') = N(\tilde{\pi}) - L < N(\tilde{\pi})$, and that $\|M_{\tilde{\pi}'^{-1}} - M_{\pi^{*-1}}\|_F^2 \leq \|M_{\tilde{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2$ if condition (4.93) is satisfied.

We can therefore restrict ourselves to estimators $\hat{\pi}$ such that $\hat{\pi}(\overline{G}^{(\iota)}) = \overline{G}^{(\iota)}$ for all ι. There is however still another catch to obtain the same lines as in [74]. Indeed, the restriction $\hat{\pi}^{(\iota)}$ of $\hat{\pi}$ to $\overline{G}^{(\iota)}$ is measurable with respect to the observation Y, but not necessarily to $Y(\overline{G}^{(\iota)})$. Still, this restriction can be writen as $\hat{\pi}^{(\iota)} = \hat{\pi}^{(\iota)}(Y(\overline{G}^{(\iota)}), Y([n] \setminus \overline{G}^{(\iota)}))$, and, for any $\alpha > 0$, there exists $y^{*(\iota)}(\alpha)$ such that

$$\mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}\left[\|M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\|_F^2\right] \geq \mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}\left[\|M_{\bar{\pi}^{(\iota)-1}(\alpha)} - M_{\pi^{*-1}}\|_F^2\right] - \alpha \ ,$$

where $\bar{\pi}^{(\iota)} := \hat{\pi}^{(\iota)}(Y(\overline{G}^{(\iota)}), y^{*(\iota)}(\alpha))$ is measurable with respect to $Y(\overline{G}^{(\iota)})$. Since it is possible such a stable estimator for any $\alpha > 0$, we finally obtain the inequality

$$\begin{aligned}\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda) &\geq \inf_{\hat{\pi}:\ \hat{\pi}(\overline{G}^{(\iota)})=\overline{G}^{(\iota)}} \sup_{\mathbf{G}\in\mathcal{G}} \sum_{\iota=1}^{n/p} \mathbf{E}_{\mathbf{G}}^{(\mathbf{full})}\left[\|\left(M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2\right] \\ &\geq \sum_{\iota=1}^{n/p} \inf_{\hat{\pi}^{(\iota)}} \sup_{G^{t(\iota)}} \mathbf{E}_{G^{t(\iota)}}^{(\mathbf{full})}\left[\|\left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2\right] \ .\end{aligned}$$

The problem of estimating the permutation π^* is now broken down into the n/p smaller problems of estimating the subsets $G^{t(\iota)} \subset \overline{G}^{(\iota)}$. The square Euclidean distance between to experts in $\overline{G}^{(\iota)}$ of experts is 0 is they are both either in or not in $G^{t(\iota)}$ and it is equal to $\frac{q\upsilon^2}{p\lambda}$ otherwise. Let us focus on the easier problem of estimating the subsets $G^{t(\iota)}$ and define $\hat{G}^{t(\iota)}$ the set of the $p/2$ experts that are ranked above according $\hat{\pi}^{(\iota)}$. Then, we have that

$$\|\left(M_{\hat{\pi}^{(\iota)-1}} - M_{\pi^{*-1}}\right)_{\overline{G}^{(\iota)}}\|_F^2 = \frac{q\upsilon^2}{p\lambda}|\hat{G}^{(\iota)} \Delta G^{t(\iota)}| \geq \frac{q\upsilon^2}{4\lambda}\mathbf{1}\{\hat{G}^{(\iota)} \neq G^{t(\iota)}\} \ ,$$

where the last inequality comes from the construction of the sets $G^{t(\iota)}$ by Lemma 4.5.17. Hence, we deduce that

$$\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda) \geq \frac{q\upsilon^2}{4\lambda} \sum_{\iota=1}^{n/p} \inf_{\hat{\pi}^{(\iota)}} \sup_{G^{t(\iota)}} \mathbf{P}_{G^{t(\iota)}}^{(\mathbf{full})}\left[\hat{G}^{(\iota)} \neq G^{t(\iota)}\right] \ , \tag{4.94}$$

so that by symmetry,

$$\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda) \geq \frac{nq\upsilon^2}{4p\lambda} \inf_{\hat{G}^{(1)}} \sup_{G^{t(1)}} \mathbf{P}_{G^{t(1)}}^{(\mathbf{full})}\left[\hat{G}^{(1)} \neq G^{t(1)}\right] \ .$$

Consider the $p \times d$ matrices N and $Y^{\downarrow}$ defined by

$$N_{ik} = \sum_t \mathbf{1}_{x_t=(i,k)} \ ; \qquad Y^{\downarrow}_{ik} = \sum_t \mathbf{1}_{x_t=(i,k)}(y_t - w_i) \ ,$$

where w is defined in (4.90). To simplify the notation, we write henceforth G and $\hat{G}$ for $G^{t(1)}$ and $\hat{G}^{(1)}$ respectively. Letting $\mathbf{P}_G$ for the corresponding marginal distribution of N and $Y^{\downarrow}$, the same sufficiency argument as in [74] gives that

$$\inf_{\hat{G}} \sup_{G} \mathbf{P}_G^{(\mathbf{full})}\left[\hat{G} \neq G\right] = \inf_{\hat{G}} \sup_{G} \mathbf{P}_G\left[\hat{G} \neq G\right] \ .$$

We finally obtain the following inequality:

$$\mathcal{R}^*_{\mathrm{perm}}(n,d,\lambda) \geq \frac{nq\upsilon^2}{4p\lambda} \inf_{\hat{G}} \sup_{G} \mathbf{P}_G\left[\hat{G} \neq G\right] \ . \tag{4.95}$$

Let $\mathbf{P}_0$ be the distribution on N and $Y^{\downarrow}$ corresponding to the case $\upsilon = 0$. The entries of N of are independent and follow a Poisson distribution of parameter λ. Conditionally to N_{ik}, we have $Y^{\downarrow}_{ik}$ is a Gaussian variable with mean 0 and variance N_{ik}. Then, we deduce from Fano's inequality [90] that

$$\inf_{\hat{G}} \sup_{G\in\mathcal{G}} \mathbf{P}_G(\hat{G} \neq G) \geq 1 - \frac{1 + \max_{G\in\mathcal{G}} \mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0)}{\log(|\mathcal{G}|)} \ , \tag{4.96}$$

where $\mathrm{KL}(.\|.)$ stands for the Kullback-Leibler divergence. The following lemma gives an upper bound of these Kullback-Leibler divergences. It can be found in [74], with the slightly stronger assumption that $p\lambda \geq 1$.

Lemma 4.5.18 (Lemma J.2 of [74]). *There exists a numerical constant c_1 such that the following holds true. If $v^2 \le 1 \wedge p\lambda$, then for any $G \in \mathcal{G}$, we have*

$$\mathrm{KL}(\mathbf{P}_G \| \mathbf{P}_0) \le c_1 \frac{v^2 q^2}{d} \quad .$$

The proof of Lemma 4.5.18 can be found in [74], with $\tilde{n} := p$ and $\tilde{d} = d$. The slightly stronger assumption that $p\lambda \ge 1$ made in Lemma J.2 in [74] is in fact not necessary. Indeed, it is only used to prove that $\mathcal{I} := \lambda p(e^{v^2/(\lambda p)} - 1) \le c_1' v^2$ in the proof of Lemma J.2 in [74], and this inequality remains valid under the assumption of Lemma 4.5.18, that is $u^2 := v^2/(\lambda p) \le 1$.

Step 3: Choice of suitable parameters p, q and v

By combining (4.95), (4.96), with Lemma 4.5.18 and the different constraints on the parameters (4.93), we directly obtain the following proposition.

Proposition 4.5.19. *There exists a numerical constant c such that if $p \in \{2, \dots, n\}$, $q \in \{1, \dots, d\}$ are dyadic integers, and v satisfy the following condition:*

$$v \quad \le \quad c\left[1 \wedge \sqrt{p\lambda} \wedge \frac{\sqrt{pd}}{q} \wedge \sqrt{\lambda}\frac{p^{3/2}}{n}\right] \quad , \tag{4.97}$$

then we have

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c\frac{nqv^2}{p\lambda} \quad . \tag{4.98}$$

The above proposition being a direct consequence of what preceeds it, we consider that it does not require a proof. Let us now apply Proposition 4.5.19 for different parameters p, q and v to conclude the proof of Theorem 4.2.1.

First, using the lower bound in the bi-isotonic case – see Theorem 4.1 of [74], we have for some constant c' that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c'(n/\lambda \wedge nd) \quad . \tag{4.99}$$

In what follows, we write $\lfloor x \rfloor_{\mathrm{dya}}$ for the greatest integer that is a power of two and smaller than x. Let us consider the following inequality:

$$\lambda \ge 1/d \vee n^2/d^3 \quad . \tag{4.100}$$

In the case where (4.100) is not satisfied, then $n\sqrt{d/\lambda} \wedge n^{2/3}\sqrt{d}\lambda^{-5/6} \le n/\lambda \wedge nd$ and the lower bound of Theorem 4.2.1 is proven by (4.99).

We subsequently assume that (4.100) is satisfied.

Case 1: $\lambda n \le 1$. In this case, we choose $q = \left\lfloor \sqrt{\frac{d}{\lambda}} \right\rfloor_{\mathrm{dya}}$ and $p = n/2$. We have that $q \in \{1, \dots, d\}$ since $\lambda \le 1$ in that case and by assumption (4.100), $\lambda \ge 1/d$. We deduce from Proposition 4.5.19 applied with $v/c = \sqrt{p\lambda} = \sqrt{pd}/q$ that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c'' n\sqrt{\tfrac{d}{\lambda}} \quad .$$

Case 2: $\lambda \in [\frac{1}{n}, 8n^2]$. In this case, we choose $q = \left\lfloor \frac{n^{1/3}\sqrt{d}}{\lambda^{1/6}} \right\rfloor_{\mathrm{dya}}$ and $p = \left\lfloor \frac{n^{2/3}}{\lambda^{1/3}} \right\rfloor_{\mathrm{dya}}$. We deduce from (4.100) that $q \le d$. Since $\lambda \in [\frac{1}{n}, 8n^2]$, we also necessarily have that $q \ge 1, p \ge 2$ and $p \le n$. Applying the above proposition with $v/c = 1 = \sqrt{pd}/q = \sqrt{\lambda}p^{3/2}/n$, we deduce that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c'' \tfrac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} \quad .$$

Case 3: $\lambda \ge 8n^2$. When λ satisfies this condition that is out of the scope of Theorem 4.2.1 but discussed below Theorem 4.2.1, we choose $q = \left\lfloor \sqrt{d} \right\rfloor_{\mathrm{dya}}$ and $p = 2$. Applying the above proposition with $v/c = 1$, we deduce that

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c'' \tfrac{n\sqrt{d}}{\lambda} \quad .$$

We have proved that for any n, d and λ, we have the lower bound

$$\mathcal{R}^*_{\mathrm{perm}}(n, d, \lambda) \ge c''\left[n\sqrt{\tfrac{d}{\lambda}} \wedge \tfrac{n^{2/3}\sqrt{d}}{\lambda^{5/6}} \wedge \tfrac{n\sqrt{d}}{\lambda} + n/\lambda\right] \wedge nd \quad .$$

This concludes in particular the proof of Theorem 4.2.1, stated for $\lambda \in [1/d, 8n^2]$.

□

4.5.9 Proof of Proposition 4.4.1

Let us introduce $\mathbb{P}_k = \Lambda(X_{.k}X_{.k}^T - \mathbb{E}[X_{.k}X_{.k}^T])\Lambda \in \mathbb{R}^{p\times p}$, so that

$$\Lambda(XX^T - \mathbb{E}[XX^T])\Lambda = \sum_{k=1}^{q} \mathbb{P}_k \quad . \tag{4.101}$$

Lemma 4.5.20. *There exists a numerical constant κ_3''' such that for any $x \in [0, (\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$, we have*

$$\|\mathbb{E}[e^{x\,\mathbb{P}_k}]\|_{\mathrm{op}} \le \exp(\kappa_3''' x^2(\sigma^2 + \sigma^4 p)) + \frac{1}{q} \quad .$$

Moreover, applying the Matrix Chernoff techniques for the independent matrices $\mathbb{P}_k$ (see lemma 6.12 and 6.13 of [94]), we have for any $t > 0$ that

$$\begin{aligned}
\log(\mathbb{P}(\|\sum_{k=1}^{q}\mathbb{P}_k\|_{\mathrm{op}} \ge t)) &\le \log(\mathrm{tr}\left[\mathbb{E}[e^{x\sum_{k=1}^{q}\mathbb{P}_k}]\right]) - xt \\
&\le \log\left(\mathrm{tr}\left[\exp\left(\sum_{k=1}^{q}\log(\mathbb{E}[e^{x\,\mathbb{P}_k}])\right)\right]\right) - xt \\
&\le \log(p) + \sum_{k=1}^{q}\|\log(\mathbb{E}[e^{x\,\mathbb{P}_k}])\|_{\mathrm{op}} - xt \\
&= \log(p) + \sum_{k=1}^{q}\log(\|\mathbb{E}[e^{x\,\mathbb{P}_k}]\|_{\mathrm{op}}) - xt \quad .
\end{aligned}$$

Applying Lemma 4.5.20, it holds for any $x \in [0, (\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$ that

$$\begin{aligned}
\sum_{k=1}^{q}\log(\|\mathbb{E}[e^{x\,\mathbb{P}_k}]\|_{\mathrm{op}}) &\le q\log\left(\exp\left(\kappa_3''' x^2(\sigma^2+\sigma^4 p)\right) + \frac{1}{q}\right) \\
&\le \kappa_3''' x^2(\sigma^2 q + \sigma^4 pq) + 1 \quad ,
\end{aligned}$$

where in the last inequality we used the fact that for any $a \ge 1$ and $u > 0$, $\log(a+u) \le \log(a) + u/a$.

Hence, we obtain

$$\log(\mathbb{P}(\|\sum_{k=1}^{q}\mathbb{P}_k\|_{\mathrm{op}} \ge t)) \le \log(ep) + \kappa_3''' x^2(\sigma^2 q + \sigma^4 pq) - xt \quad .$$

Hence, if $t \le 2\frac{\sigma^2 q+\sigma^4 pq}{\sigma^2 r_\Lambda + K^2\log(q)}$, we choose $x = \frac{t}{2\kappa_3'''(\sigma^2 q+\sigma^4 pq)}$ and if $t > 2\frac{\sigma^2 q+\sigma^4 pq}{\sigma^2 r_\Lambda + K^2\log(q)}$ we choose $x = \frac{1}{\kappa_3'''(\sigma^2 r_\Lambda + K^2\log(q))}$, which gives

$$\mathbb{P}(\|\sum_{k=1}^{q}\mathbb{P}_k\|_{\mathrm{op}} \ge t) \le ep\max\left[\exp\left(-\frac{1}{\kappa_3}\left(\frac{t^2}{4(\sigma^2 q+\sigma^4 pq)} \vee \frac{t}{2(\sigma^2 r_\Lambda + K^2\log(q))}\right)\right)\right] \quad .$$

We deduce that with probability at least $1-\delta$, it holds that

$$\|\sum_{k=1}^{q}\mathbb{P}_k\|_{\mathrm{op}} \le \kappa\left[\sqrt{(\sigma^4 pq + \sigma^2 q)\log(p/\delta'')} + (\sigma^2 r_\Lambda + K^2\log(q))\log(p/\delta'')\right] \quad ,$$

for some numerical constant κ.

Proof of Lemma 4.5.20. Since $\|\Lambda X_{.k}X_{.k}^T\Lambda\|_{\mathrm{op}} = \|\Lambda X_{.k}\|_2^2$, we state the following lemma controlling the moment generating function of the L_2 norm of the projection $\Lambda X_{.k}$:

Lemma 4.5.21. *There exists a numerical constant κ_0''' such that for any $x \le \frac{1}{\kappa_0''' K^2}$ we have*

$$\mathbb{E}[e^{x\|\Lambda X_{.k}\|_2^2}] \le e^{\kappa_0''' \sigma^2 r_\Lambda x} \quad .$$

Now we define the event $\xi_{\text{op}} := \{\max_{k=1,\dots,d} \|\Lambda X_{\cdot k}\|_2^2 \le \kappa_0'''(\sigma^2 r_\Lambda + K^2 \log(q^3))\}$, where κ_0''' is the numerical constant given by Lemma 4.5.21. Applying the same lemma together with the Chernoff bound, a union bound over all $k = 1 \dots d$ gives

$$\mathbb{P}(\xi_{\text{op}}^c) \le \tfrac{1}{q^2} \ .$$

We consider in what follows the relation order $\preceq$ induced by the cone of nonegative symmetric matrices $\mathbb{S}_n^+$, namely $X' \preceq X''$ if and only if $X'' - X' \in \mathbb{S}_n^+$. Under the event ξ_{op}, it holds that for any $u \ge 2$,

$$\begin{aligned}
\mathbb{P}_k^u &\preceq \|\,\mathbb{P}_k\,\|_{\text{op}}^{u-2}\,\mathbb{P}_k^2 \\
&\preceq \|\Lambda(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])\Lambda\|_{\text{op}}^{u-2}\,\mathbb{P}_k^2 \\
&\preceq (\kappa_1'''(\sigma^2 r_\Lambda + K^2 \log(q)))^{u-2}\,\mathbb{P}_k^2 \ ,
\end{aligned}$$

for some numerical constant κ_1''' (depending on κ_0'''). In the third inequality we used the definition of ξ_{op} the fact that $\mathbb{E}[\|\Lambda X_{\cdot k}\|_2^2] \le \kappa_0'''\sigma^2 r_\Lambda$.

We now give an upper bound of $\|\mathbb{E}[\mathbb{P}_k^2]\|_{\text{op}}$, which is the operator norm of the variance of $\mathbb{P}_k$ as defined in section 6 in [94]. Remark that since any matrix $U \in \mathbb{R}^{q\times q}$ satisfies $U\Lambda U^T \preceq UU^T$, we have that $\mathbb{P}_k \preceq \Lambda(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2\Lambda$.

Let us compute the expectation of $(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2$:

$$\mathbb{E}[(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2]_{ij} = \sum_{l\in P} \mathbb{E}[(X_{ik}X_{lk} - \mathbb{E}[X_{ik}X_{lk}])(X_{lk}X_{jk} - \mathbb{E}[X_{lk}X_{jk}])] \ .$$

The off diagonal terms are zero, and the i^{th} diagonal element satisfies:

$$\mathbb{E}[(X_{\cdot k}^T X_{\cdot k} - \mathbb{E}[X_{\cdot k}^T X_{\cdot k}])^2]_{ii} = \mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2] + \sum_{j\neq i} \mathbb{E}[X_{ik}^2]\mathbb{E}[X_{jk}^2] \ . \tag{4.102}$$

By assumption (4.24), the first term of (4.102) satisfies

$$\mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2] \le 4\mathbb{E}[(X_{ik}^4)] \le 48\sigma^2 K^2 \ .$$

The second term of (4.102) is smaller than $\sigma^4 p$, still by assumption (4.24). Hence we have some numerical constant κ_2''' that

$$\|\mathbb{E}[\mathbb{P}_k^2]\|_{\text{op}} \le \|\mathbb{E}[(X_{ik}^2 - \mathbb{E}[X_{ik}^2])^2]\|_{\text{op}} \le \kappa_2'''(\sigma^2 + \sigma^4 p) \ .$$

Now, by the definition of the exponential of matrices, the triangular inequality and the fact that $\mathbb{P}_k$ is centered, we have

$$\|\mathbb{E}[\exp(x\,\mathbb{P}_k)]\|_{\text{op}} = 1 + \sum_{u\ge 2} \frac{x^u}{u!}\|\mathbb{E}[\mathbb{P}_k^u\,\mathbf{1}_{\xi_{\text{op}}}]\|_{\text{op}} + \sum_{u\ge 2} \frac{x^u}{u!}\|\mathbb{E}[\mathbb{P}_k^u\,\mathbf{1}_{\xi_{\text{op}}^c}]\|_{\text{op}} \ . \tag{4.103}$$

By definition of ξ_{op} together with the upper bound of the variance of $\mathbb{P}_k^2\,\mathbf{1}_{\xi_{\text{op}}} \preceq \mathbb{P}_k^2$, it holds for any $x \in [0, (\kappa_1'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$ that

$$\begin{aligned}
\sum_{u\ge 2} x^u \|\mathbb{E}[\mathbb{P}_k^u\,\mathbf{1}_{\xi_{\text{op}}}]\|_{\text{op}} &\le x^2\|\mathbb{E}[\mathbb{P}_k^2]\|_{\text{op}} \sum_{u\ge 2} \frac{x^{u-2}}{u!}(\kappa_1'''(\sigma^2 r_\Lambda + K^2\log(q)))^{u-2} \\
&\le x^2\kappa_2'''(\sigma^2 + \sigma^4 p) \sum_{u\ge 0} \frac{x^u}{(u+2)!}(\kappa_1'''(\sigma^2 r_\Lambda + K^2\log(q)))^u \\
&\le \exp(\kappa_3''' x^2(\sigma^2 + \sigma^4 p)) - 1 \ ,
\end{aligned}$$

for some numerical constant κ_3'''. We now control the second term of (4.103) under the complementary event ξ_{op}, for any $x \in [0, (2\kappa_0'''(\sigma^2 r_\Lambda + K^2\log(q)))^{-1}]$:

$$\begin{aligned}
\sum_{u\ge 2} \frac{x^u}{u!}\|\mathbb{E}[\mathbb{P}_k^u\,\mathbf{1}_{\xi_{\text{op}}^c}] &\le \mathbb{E}[\exp(x\|\,\mathbb{P}_k\,\|_{\text{op}}\mathbf{1}_{\xi_{\text{op}}^c}]] \\
&\overset{(a)}{\le} \sqrt{\frac{1}{q^2}}\sqrt{\mathbb{E}[\exp(2x\|\,\mathbb{P}_k\,\|_{\text{op}})]} \\
&\overset{(b)}{\le} \frac{1}{q}\exp(x\kappa_0'''\sigma^2 r_\Lambda) \\
&\le \frac{1}{q} \ ,
\end{aligned}$$

where in (a) we used the Cauchy-Schwarz inequality for real random variables and in (b) we applied Lemma 4.5.21.

□

Proof of Lemma 4.5.21. We use the result of [7] which is a generalization of the Hanson-Wright inequality to random variables with coefficients with Bernstein's moments.

[Assumption 1 of [7]] is satisfied with parameters σ^2 and K, and we have the following upper bound on the moment generating function of the quadratic form $\|\Lambda X_{.k}^T\|_2^2 = |X_{.k}\Lambda X_{.k}^T|$:

$$\mathbb{E}\big[e^{x\|\Lambda X_{.k}^T\|_2^2}\big] \leq e^{x\mathbb{E}[\|\Lambda X_{.k}^T\|_2^2]} e^{\kappa_0''' x^2 K^2 \sigma^2 \|\Lambda\|_F^2} \leq e^{\kappa_1''' x \sigma^2 r_\Lambda} \quad , \tag{4.104}$$

for any x satisfying condition (6) of [7], that is $128x\|\Lambda\|_{\mathrm{op}} K^2 \leq 1$. For the last inequality, we used the fact that $\|\Lambda\|_F^2 = \mathrm{rank}(\Lambda)$. We obtain the result by choosing $\kappa_2''' = \kappa_1''' \vee 128$. □

Chapter 5

Multiple change-point detection for high-dimensional data

This chapter makes two contributions to the field of change-point detection. In a general change-point setting, we provide a generic algorithm for aggregating local homogeneity tests into an estimator of change-points in a time series. Interestingly, we establish that the error rates of the collection of tests directly translate into detection properties of the change-point estimator. This generic scheme is then applied to various problems including covariance change-point detection, nonparametric change-point detection and sparse multivariate mean change-point detection. For the latter, we derive minimax optimal rates that are adaptive to the unknown sparsity and to the distance between change-points when the noise is Gaussian. For sub-Gaussian noise, we introduce a variant that is optimal in almost all sparsity regimes.

This chapter is based on [73].

5.1 Introduction

Change-point detection has a long history since the seminal work of Wald [95] that lead to flourishing lines (see [68, 89] for recent surveys). Earlier contributions focused on the problems of detecting and localizing change-points in a univariate time series. Spurred by applications in genomics [69] and finance, there has been a recent trend in the literature towards the analysis of more complex time series for instance in a high-dimensional linear space [48] or even belonging to a non-Euclidean space [24].

In this work, we study high-dimensional time series whose mean may change possibly on a few number of coordinates. See the introduction of [103] for an account of possible applications and practical motivations. In particular, we build a procedure which is able to detect and localize change-points under minimal assumptions on the height of these change-points. Along the way towards this optimal procedure, we define and analyze a scheme for general change-point problems that aggregates a collection of local tests into an estimator change-points. This generic scheme is of independent interest and easily allows to derive optimal change-point procedure in other complex settings such as covariance change-points problems or nonparametric change-point problems. In this introduction, we first describe this generic scheme before turning to our results in high-dimensional sparse change-point detection and finally discussing other applications.

5.1.1 General change-point setting

In the most general form of a change-point problem, we consider a random sequence $Y = (y_1, y_2, \dots, y_n)$ in some measured space $\mathcal{Y}^n$ and, for $t = 1, \dots, n$, we write $\mathbb{P}_t$ for the marginal distribution of y_t. We are also given a functional Γ mapping the probability distribution $\mathbb{P}_t$ to some space $\mathcal{V}$. Then, the purpose of change-point detection is to detect changes in the sequence $(\Gamma(\mathbb{P}_1), \Gamma(\mathbb{P}_2), \dots, \Gamma(\mathbb{P}_n))$ in $\mathcal{V}^n$ and to estimate the positions of these changes. This setting is really general and does not require that the random variables (y_t) are independent.

Let us shortly explain how this general framework encompasses most offline change-point detection problems. In the Gaussian mean univariate change-point setting, we have $\mathcal{Y} = \mathbb{R}$, the distribution $\mathbb{P}_t$ corresponds to the normal distribution with mean $\theta_t \in \mathbb{R}$ and variance σ^2 and $\Gamma(\mathbb{P}_t) = \theta_t$. In the (heteroscedastic) mean univariate change-point problem, the distribution $\mathbb{P}_t$ is not necessarily Gaussian and, in particular, the variance of y_t is allowed to vary with t. Still, one is only interested in detecting variations of $\Gamma(\mathbb{P}_t) = \int x d\mathbb{P}_t = \mathbb{E}[y_t]$. By contrast, in the *variance* univariate change-point problems, one focuses on changes in the variance of y_t. This can be

done by taking $\Gamma(\mathbb{P}_t) = \int x^2 d\mathbb{P}_t - [\int x d\mathbb{P}_t]^2 = \mathrm{Var}(y_t)$. If one is interested in possibly nonparametric changes in the distributions, then the functional Γ is simply taken to be the identity map. In semi-parametric quantile change-point detection [49], the univariate distributions $\mathbb{P}_t$ can be arbitrary whereas $\Gamma(\mathbb{P}_t)$ is a quantile of $\mathbb{P}_t$.

To further formalize the change-point detection problem in the sequence $(\Gamma(\mathbb{P}_1), \Gamma(\mathbb{P}_2), \ldots, \Gamma(\mathbb{P}_n))$, we define an integer $0 \leq K \leq n-1$ and a vector of integers $\tau = (\tau_1, \ldots, \tau_K)$ satisfying $1 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n+1$ such that $\Gamma(\mathbb{P}_t)$ is constant over each interval $[\tau_k, \tau_{k+1}-1]$ and $\Gamma(\mathbb{P}_{\tau_k-1}) \neq \Gamma(\mathbb{P}_{\tau_k})$. Hence, τ_k corresponds to the position of the k^{th} change-point. We shall often refer to τ_k as a change-point. Equipped with this notation, we are interested in building an estimator $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}})$ of τ from the time series Y. Here, $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$ correspond to the estimated change-points of τ and $\hat{K}$ to the number of the estimated change-points.

5.1.1.1 Desirable Guarantees of an estimator.

Before describing the generic scheme for estimating τ, let us first formalize the desired properties of a good change-point procedure. Informally, the primary objectives are to detect most if not all change-points while estimating no (or at least very few) spurious change-points.

Regarding the latter objective, it is usually required that the number of change-points K is not overestimated by $\hat{\tau}$. Here, we require a slightly stronger local property introduced in [92]. An estimator $\hat{\tau}$ of size $\hat{K}$ is said to detect no spurious change-points **(NoSp)** if

$$\begin{cases} \left|\{\hat{\tau}_{k'}, 1 \leq k' \leq \hat{K}\} \cap \left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2}\right]\right| \leq 1 \ , & \text{for all } 1 \leq k \leq K \ ; \\ \{\hat{\tau}_{k'}, 1 \leq k' \leq \hat{K}\} \subset \left[\tau_1 - \frac{\tau_1 - 1}{2}, \tau_K + \frac{n+1-\tau_K}{2}\right] \ . \end{cases} \tag{5.1}$$

The second condition simply ensures that no change-point is estimated near the boundaries of the time series. The first condition entails that, for each change-point τ_k there is at most one estimated change-point $\hat{\tau}_k$ in the interval $[\tau_k - (\tau_k - \tau_{k-1})/2, \tau_k + (\tau_{k+1} - \tau_k)/2]$. In other words, (**NoSp**) requires that, on each sub-interval, the number of change-points is not overestimated.

Let us now formalize the objective of detecting the change-points. In this work, we consider as in [92] realistic settings where some change-points are so close or their heights are so small that they are impossible to detect. As a consequence, we can only hope to detect the subset of significant change-points. In what follows, we define a subset $\mathcal{K}^* \subset [K]$ of change-point indices that correspond to significant change-points. Obviously, the significance of a particular change-point is relative to the problem under consideration - data distribution, nature of change-points - and the definition is problem dependent. As an example, we define in the next subsection the suitable notion of energy and significance of a change-point in the mean multivariate change-point setting. In Section 5.6, we formalize this notion for covariance and univariate nonparametric change-point problems. In light of this discussion, the second guarantee we aim for is to **detect** all significant change-points. A change-point τ_k is said to be detected if there is at least one estimated change-point $\hat{\tau}_l$ in the interval $[\tau_k - (\tau_k - \tau_{k-1})/2, \tau_k + (\tau_{k+1} - \tau_k)/2]$. Equivalently, this means that at least one of the estimated change-points is closer to τ_k than to any other true change-point.

Aside from (**NoSp**) and (**detect**) properties, one may additionally aim at localizing the change-points as well as possible – see the discussions in [97]. Given a specific change-point τ_k detected by an estimator $\widehat{\tau}$, its localization error $d_{H,1}(\widehat{\tau}, \tau_k)$ is defined by

$$d_{H,1}(\widehat{\tau}, \tau_k) = \min_{l=1,\ldots,|\widehat{\tau}|} |\widehat{\tau}_l - \tau_k| \ ,$$

which is the smallest distance between τ_k and one of the estimated change-points. While this work mainly focused on the detection problem, we shall also provide localization bounds along the way.

5.1.1.2 A generic roadmap for change-point detection.

In this chapter, our first contribution is a generic procedure for aggregating a collection of tests into an estimator $\widehat{\tau}$ of τ. For two positive integers (l, r), we consider the time interval $[l-r, l+r)$. Suppose we are given a collection $\mathcal{G}$ of such (l, r). For each $(l, r) \in \mathcal{G}$, we are also given a homogeneity test $T_{l,r}$ of the null hypothesis $\mathcal{H}_0$: $\{(\Gamma(\mathbb{P}_t))$ is constant over the segment $[l-r, l+r)\}$. This hypothesis is equivalent to the absence of any change-point on the interval $(l-r, l+r)$. Given such a collection of homogeneity tests $(T_{l,r})$, $(l, r) \in \mathcal{G}$, we build in this chapter an estimator $\widehat{\tau}$ that satisfies the following properties. If the multiple testing procedure does not reject any true null hypothesis (no false positives), then $\widehat{\tau}$ does not estimate any spurious change-point, that is, it satisfies (**No Sp**). Furthermore, any change-point τ_k that is detected by some test $T_{\bar{\tau}_k, \bar{r}_k}$, where $\bar{\tau}_k$ is close enough to τ_k and $\bar{r}_k$ is small enough is **detected** by the estimator $\widehat{\tau}$. In other words, we establish a completely generic result that translates properties of the multiple testing procedure into **detection** properties. Thus, the construction of a change-point procedure boils down to building a suitable multiple testing procedure $(T_{l,r})$,

$(l, r) \in \mathcal{G}$ whose family-wise error rate (FWER) is controlled, while being able to detect all the significant change-points. In turn, this allows us to reduce the problem of change-point detection under minimal distance between the change-points to the well-established field of minimax testing.

5.1.1.3 Related Work and possible applications.

In the last years, there has been a growing interest into the extension of univariate mean change-point procedures such as wild binary segmentation (WBS) [36] to other problems such as covariance change-point [96], network change-point [97], or nonparametric change-point [70]. For each of these problems (and for others), it turns out that the general ideas of WBS can be instantiated. However, for each setting, the proofs need to be fully adapted in a case by case manner. Besides, the resulting procedures are only optimal up to logarithmic terms.

Recently, Chan and Chen [47] and Kovács et al. [51] have introduced bottom-up aggregation procedures for mean change-point segmentation (see also [52] for localization improvements). Moreover, Kovács et al. [51, 52] illustrate the numerical performances to other change-point models, such as graphical models or multivariate mean-change point models. In fact, one may extend their procedures to generic problems, but the theoretical guarantees are only provided for univariate models, and it remains unclear whether one can extend them beyond very specific cases.

In contrast, it is quite straightforward to adapt our generic procedure to any new setting once suitable homogeneity multiple tests have been crafted. As the most prominent example, we consider the sparse high-dimensional mean change-point detection and establish the optimality of our procedure – see the next subsection for details. In Section 5.6, we also handle the covariance change-point detection and the univariate nonparametric change-point detection problems. In each case, we pinpoint the first tight minimal conditions for detection.

Besides, we could apply our strategy to other problems such changes in auto-regressive models [99], changes in the inverse covariance matrix of y_i [41, 51] or changes in a high-dimensional regression model [77]. All such change-point problems can be addressed through the construction and careful analysis of two-sample tests for auto-regressive models, inverse covariance matrices, and linear regression models respectively. Similarly, we can build Kernel change-point procedures [3, 40] from kernel two-sample tests [43].

5.1.2 Sparse multivariate change-point setting

As explained above, our primary application of our generic scheme is the multivariate mean change-point detection problem with sparse variations where one observes a time series $Y = (y_1, \dots, y_n) \in \mathbb{R}^{p \times n}$ with unknown means $\Theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^{p \times n}$ so that we have the decomposition

$$y_t = \theta_t + \varepsilon_t \qquad t = 1, \dots, n \ , \tag{5.2}$$

where the noise matrix $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ is made of independent and mean zero random vectors of size p. In this chapter, we make two distributional assumptions on the noise. Either we suppose that all random vectors ε_i follow independent normal distribution with variance $\sigma^2 \mathbf{I}_p$ (see section 5.3) or that the components of ε_i follow independent sub-Gaussian distributions with variance σ^2 (see section 5.4). In either case, we assume that σ^2 is known.

Here, we are interested in the variations of the *mean* vector θ_t so that, relying on the formalism of the previous subsection, we have $\Gamma(\mathbb{P}_t) = \theta_t$. Considering the vector of change-points $\tau = (\tau_1, \dots, \tau_K)$, we can define $K + 1$ vectors $\mu_0, \dots, \mu_K$ in $\mathbb{R}^p$ satisfying $\mu_k \neq \mu_{k+1}$ for all $k = 0, \dots, K - 1$ such that

$$\theta_t = \sum_{k=0}^{K} \mu_k \mathbf{1}_{\tau_k \leq t < \tau_{k+1}} \ .$$

Equivalently, μ_k is the constant mean of y over the interval $[\tau_k, \tau_{k+1} - 1]$. The difference $\mu_k - \mu_{k-1}$ in $\mathbb{R}^p$ measures the variation of Θ at the change-point τ_k and can possibly have many null coordinates. In this possibly sparse multi-dimensional setting, the significance of a change-point is measured through three quantities Δ_k, r_k, and s_k. First, the height Δ_k of the change-point τ_k is defined as the Euclidean norm of the signal difference. The length r_k of the change-point τ_k is the minimal distance from τ_k to another change-point, τ_{k-1} or τ_{k+1}. More precisely,

$$\Delta_k = \|\mu_k - \mu_{k-1}\| \ ; \qquad r_k = \min(\tau_{k+1} - \tau_k, \tau_k - \tau_{k-1}) \ . \tag{5.3}$$

As a simple example, Figure 5.1 depicts a one dimensional piece-wise constant sequence Θ with 3 change-points illustrating the setting presented above. In the univariate change-point literature (e.g. [36, 37, 23]) the height and the length of a change-point characterize the significance of a change-point. In the multivariate setting,

where the change-points can be sparse, meaning the number of non null coordinates of the vector $\mu_k - \mu_{k-1}$ is possibly small, one also considers the sparsity s_k of change-point τ_k, defined by

$$s_k = \|\mu_k - \mu_{k-1}\|_0 \quad , \tag{5.4}$$

where, for any $v \in \mathbb{R}^p$, $\|v\|_0 = \sum_{1 \le i \le p} \mathbf{1}\{v_i \neq 0\}$.

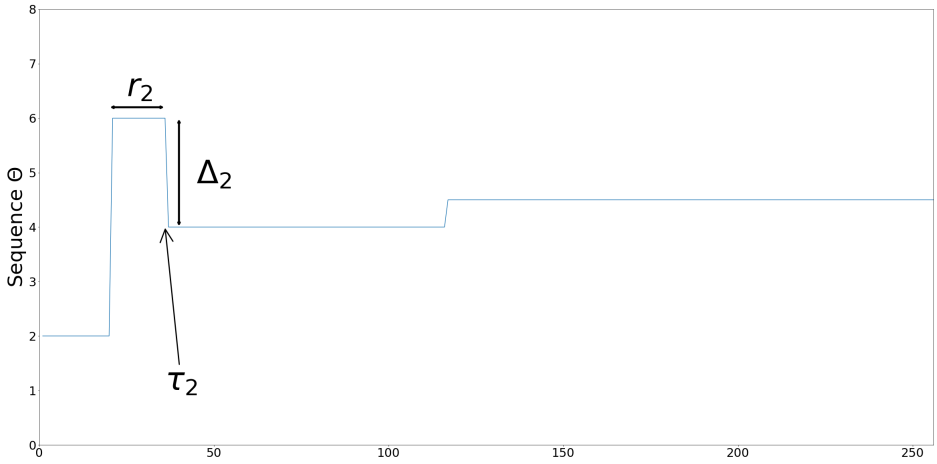


Figure 5.1: An example of a piece-wise constant sequence Θ with 3 change-points and $p = 1$.

5.1.2.1 Two-sample tests and CUSUM statistics

Our objective is to detect and recover positions $(\tau_k)_{k \le K}$ under minimal conditions on the change-point height Δ_k, change-point length r_k and sparsity s_k. In view of the generic change-point procedure discussed in the previous subsection, this mainly boils down to building suitable tests of the assumptions $\{\Theta$ is constant over $[l-r, l+r)\}$ versus $\{\Theta$ is not constant on this segment$\}$. Following the literature on binary and wild binary segmentation, we consider the CUSUM statistic

$$\mathbf{C}_{l,r}(Y) = \sqrt{\frac{r}{2\sigma^2}} \left(\frac{1}{r} \sum_{i=l}^{l+r-1} y_i - \frac{1}{r} \sum_{i=l-r}^{l-1} y_i \right) .$$

This statistic computes the normalized difference of empirical mean of y_i on $[l-r, l)$ and $[l, l+r)$. If the noise is Gaussian and if Θ is constant on $[l-r, l+r)$, then $\mathbf{C}_{l,r}(Y)$ simply follows a standard p-dimensional normal distribution. To simplify, consider a specific instance of our testing problem where we want to test whether $\{\Theta$ is constant over $[l-r, l+r)\}$ versus $\{\Theta$ contains exactly one change-point at l on the segment $[l-r, l+r)\}$. This corresponds to a two-sample mean testing problem, for which the CUSUM statistic $\mathbf{C}_{l,r}(Y)$ is a sufficient statistic if the noise is Gaussian. Then, given $\mathbf{C}_{l,r}(Y)$, one wants to test whether its expectation is 0 (no change-point on $[l-r, l+r)$) versus its expectation is non-zero but is s-sparse for some unknown s. This classical detection problem is well understood [30], and it is well known that a combination of a χ^2-type test with a higher-criticism-type test is optimal. Here, the challenge stems from the fact that we do not want to perform a single such test, but a large collection of tests over a collection of $(l, r) \in \mathcal{G}$.

5.1.2.2 Our contribution

As usual in the mean change-point literature, we consider the energy $r_k \Delta_k^2$ of the change-point τ_k. Up to a possible factor in $[1/2, 1]$, $r_k \Delta_k^2$ is the square distance between Θ and its projection on the space of vectors Θ' with change-point at $(\tau_1, \ldots, \tau_{k-1}, \tau_{k+1}, \ldots, \tau_K)$ –see e.g. [92] for a discussion in the univariate setting. In other words, the energy $r_k \Delta_k^2$ characterizes the significance of the change-point τ_k. In Section 5.3, we introduce a multi-scale change-point detection procedure detecting any change-point τ_k whose energy is higher, up to a numerical constant, than $\sigma^2 s_k \log(1 + \frac{\sqrt{p}}{s_k} \sqrt{\log(n/r_k)}) + \sigma^2 \log(n/r_k)$. This result is valid for arbitrary length r_k and sparsity s_k, and does not require the knowledge of these two quantities. In summary, our procedure does not estimate any spurious change-point (**NoSp**) and **detects** all the change-points whose energy are higher than

the latter threshold. In Section 5.5, we establish that, as soon as the unknown number K of the change-points is larger than 1, the condition $\sigma^2 s_k \log(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log(n/r_k)}) + \sigma^2 \log(n/r_k)$ on the energy is tight with respect to n, p, r_k and s_k, in the sense that no procedure achieving (**NoSp**) is able to detect with high probability a change-point whose energy is smaller (up to some constant) than the latter threshold. In Section 5.4, we consider the more general setting where the noise is L-sub-gaussian with known variance, and we establish a similar result to the Gaussian case up to a logarithmic loss in some regimes. Finally, we illustrate in Section 5.8 the behavior of our procedure on numerical experiments.

5.1.2.3 Related work

For dense change-points ($s_k = p$) but with unknown covariance for the noise, Wang et al. [102] (see also [101]) study the behavior of a procedure based on U-statistics of the CUSUM. Jirak [48] and Yu and Chen [105] introduce binary segmentation procedures based on the l_∞ norm of the CUSUMs. Although those work explicitly characterize the asymptotic distribution of the test statistics and, for some of them, allow temporal dependencies in the data, the corresponding energy requirements for change-point detection are either not studied or turn out to be suboptimal.

Closest to our work, Chan and Chen [47] study a bottom-up approach to detect change-points of a Gaussian multivariate time series in an asymptotic setting. More specifically, the authors consider an asymptotic regime where the size of the time series is exponential in the dimension: $n = e^{p^\zeta}$ with $\zeta \in (0,1)$. The authors also assume that the number K of change-points remains finite when $n, p \to \infty$ and that the minimal sparsity s of these change-points is polynomial is p. In this specific regime, their procedures provably recover change-points under a near-minimal (up to logarithmic factors with respect to n) condition on the energy. In contrast, our results provide non-asymptotic and tight results for all scaling with respect to n and p, allow for arbitrarily large number K of change-points and allow for the presence of non-significant change-points. In the same specific asymptotic setting, [46] introduce a so called score test statistic used in a change-point detection procedure which is shown to achieve the same performance as [47] in the gaussian model but also handle Poisson observations.

Recently, Liu et al. [57] have characterized the optimal detection rate of a possibly sparse change-point in the specific case where there is at most one change-point, but the optimal rates are significantly slower in the multiple change-point setting. See also [31] and [28] for earlier results. Wang and Samworth [103] have proposed the INSPECT method based on sparse projection to handle sparse change-points, but INSPECT provably detects the change-points under strong assumption on the energy; see Section 5.3 for a precise comparison.

In the univariate setting ($p = 1$), minimal energy requirements for change-point detection are well understood [34, 37, 98, 92] and are nearly achieved by a wide range of procedures including penalized least-square and multi-scale tests methods.

5.2 A generic algorithm for multiscale change-point detection on a grid

In this section, we study the problem of change-point detection in the general setting defined in Section 5.1.1. We introduce a bottom-up algorithm that aggregates a collection of homogeneity tests, performed at many positions, and for many scales, of our data. Then, we establish that, under some conditions on these tests, the procedure detects significant change-points.

5.2.1 Grid and multiscale statistics

Since our purpose is to translate a collection of local tests $T = (T_{l,r})_{(l,r)\in\mathcal{G}}$ indexed by a grid $\mathcal{G}$ into a change-point detection procedure, we first need to formalize what we mean by a grid. Henceforth, we call a grid $\mathcal{G}$ of $[n]$ a collection of locations and scales where a scale r is a positive integer smaller or equal to $\lfloor n/2 \rfloor$ and a location l is an integer between $r+1$ and $n-r$. This couple (l,r) refers to the segment $[l-r, l+r)$ centered at l and with radius r. Formally, $\mathcal{G}$ is therefore a subset of $J_n = \big\{(l,r) : \ r = 1, \dots, \lfloor \frac{n}{2} \rfloor \text{ and } l = r+1, \dots, n-r+1\big\}$. Given a grid $\mathcal{G}$, we call $\mathcal{R}$ its collection of scales, that is $\mathcal{R} = \{r : \ \exists l \text{ s.t. } (l,r) \in \mathcal{G}\}$. Finally, for a scale $r \in \mathcal{R}$, $\mathcal{D}_r$ stands for the corresponding collection of locations, that is $\mathcal{D}_r = \{l : \ (l,r) \in \mathcal{G}\}$. Although we do not make any assumption on the grid $\mathcal{G}$ for the time being, we will mainly consider two specific grids in this section: the **complete** grid $\mathcal{G}_F = J_n$ and the **dyadic** grid $\mathcal{G}_D$ defined by $\mathcal{R} = \{1, 2, 4, \dots, 2^{\lfloor \log_2(n) \rfloor - 1}\}$, $\mathcal{D}_1 = [2, n]$, and

$$\mathcal{D}_r = \left\{ r+1, 3\lfloor r/2 \rfloor + 1, 4\lfloor r/2 \rfloor + 1, \dots, \Big(\frac{n}{\lfloor r/2 \rfloor} - 2\Big)\lfloor \frac{r}{2} \rfloor + 1, n - r + 1 \right\} \quad \text{for } r \in \mathcal{R} \setminus \{1\} \ . \tag{5.5}$$

See Figure 5.2 for a visual representation of the dyadic grid. At some points, we shall also mention a-adic grids $\mathcal{G}_a$. For any $a \in (0,1)$, $\mathcal{G}_a$ is defined by $\mathcal{R} = \{1, \lfloor a^{-1} \rfloor, \lfloor a^{-2} \rfloor, \dots, \lfloor a^{1 - \lfloor \log(n)/\log(a) \rfloor} \rfloor\}$ and $\mathcal{D}_r$ as in (5.5).

Interestingly, the cardinality of the dyadic grid or more generally of the a-adic grid is order $O(n)$, whereas the complete grid $\mathcal{G}_D$ is quadratic.

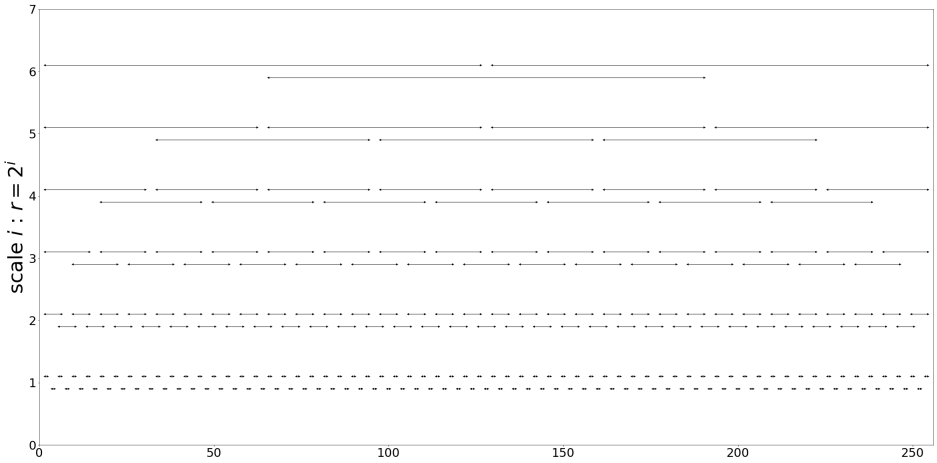


Figure 5.2: The dyadic grid is represented as follows : for each $r = 2^i$ and $l \in \mathcal{D}_r$, we draw the interval $[l - r + 1, l + r - 1]$ at position $(l, \log_2(r))$.

Grids are reminiscent of the c-normal systems of intervals introduced by Nemirovsky [67] (see also [55] for a definition) although our definition allows for non-necessarily normal intervals.

Given a fixed grid $\mathcal{G}$, a multiscale test is simply a collection of test $T = (T_{l,r})_{(l,r)\in\mathcal{G}}$ indexed by the elements of $\mathcal{G}$, which amounts to testing at all scales $r \in \mathcal{R}$ and all locations $l \in \mathcal{D}_r$ whether the functional $\Gamma(\mathbb{P}_t)$ is constant over the segment $[l - r, l + r)$. Equivalently, $T_{l,r}$ tests whether there exists a change-point in $[l - r + 1, l + r - 1]$.

5.2.2 From a multiscale test to a change-point detection procedure

Our purpose is to introduce a generic procedure to translate a multiscale procedure into a vector of change-points. Intuitively, if, for some $(l, r) \in \mathcal{G}$, we have $T_{l,r} = 1$, then the functional $\Gamma(\mathbb{P}_t)$ is certainly not constant over $[l - r, l + r)$ which entails that there is possibly at least one change-point in $[l - r + 1, l + r - 1]$. As a consequence, the multiscale test gives a collection $\mathcal{I}(T) = \{[l - r + 1, l + r - 1] \text{ s.t. } T_{l,r} = 1\}$ of intervals that tentatively contain at least one change-point.

If all these intervals were disjoint, then one simply would take $\widehat{\tau}$ as the sequence of centers of these intervals. Unfortunately, when two intervals $[l_1 - r_1 + 1, l_1 + r_1 - 1]$ and $[l_2 - r_2 + 1, l_2 + r_2 - 1]$ in $\mathcal{I}(T)$ have a non-empty intersection, one cannot necessarily decipher whether there is only one change-point in the intersection of both intervals or if each interval contains a specific change-point. Hence, our general objective is to transform the collection $\mathcal{I}(T)$ into a collection of non-intersecting intervals by either discarding or merging some of them.

We propose the following bottom-up iterative procedure for building a collection of non-intersecting intervals. Start with $\mathcal{T}_0 = \mathcal{S}_0 = \varnothing$. For any scale $r \in \mathcal{R}$, we compute the collections $\mathcal{S}_r$ of intervals of scale r and the collection $\mathcal{T}_r$ of locations based on the following

$$\mathcal{T}_r = \left\{ l \in \mathcal{D}_r, \quad T_{l,r} = 1 \quad \text{and} \quad [l - r + 1, l + r - 1] \bigcap \Big(\bigcup_{r' < r,\ r' \in \mathcal{R}} \mathcal{S}_{r'} \Big) = \varnothing \ ; \right\}$$
$$\mathcal{S}_r = \bigcup_{l \in \mathcal{T}_r} [l - r + 1, l + r - 1] \ .$$

The sets $\mathcal{T}_1$ and $\mathcal{S}_1$ are made of all positions l such that $T_{l,1} = 1$. More generally, $\mathcal{T}_r$ contains all locations l such that $T_{l,r} = 1$ and the corresponding interval $[l - r + 1, l + r - 1]$ does not intersect with any of the detected intervals at a smaller scale $r' < r$. The set $\mathcal{S}_r$ contains all intervals associated to $\mathcal{T}_r$.

One can easily check that $\mathcal{S} = \bigcup_r \mathcal{S}_r$ is a union of closed non-intersecting intervals. Denote $\mathcal{C} = \{\mathrm{C}_1, \ldots, \mathrm{C}_{\hat{K}}\}$ the partition of $\mathcal{S}$ into connected components such that, for all $1 \le i < j \le \hat{K}$, $\max \mathrm{C}_i < \min \mathrm{C}_j$. Finally, we estimate the vector of change-points $\widehat{\tau}$ by taking the center of each segment C_k. In other words, we take $\hat{\tau}_k := \frac{1}{2}(\min \mathrm{C}_k + \max \mathrm{C}_k)$ for any $1 \le k \le \widehat{K}$. This bottom-up aggregation procedure is summarized in Algorithm 16 and illustrated in Figure 5.3 below.

Remark: If, for some $r \in \mathcal{R}$ and some $l_1 < l_2 \in \mathcal{D}_r$, we have $T_{l_1,r} = 1$, $T_{l_2,r} = 1$, and $l_1 + r - 1 \geq l_2 - r + 1$, then $\mathcal{S}_r$ contains the segment $[l_1 - r + 1, l_2 + r - 1]$. In other words, our aggregation procedure merges two intervals if and only if they correspond to the same scales. In Section 5.9, we also introduce a variant of the algorithm where, instead of merging these two intersecting with identical scale, we discard one of them.

Algorithm 16 Bottom-up aggregation procedure of multiscale tests

Require: Observations $y_t, t = 1 \dots n$ and local test statistics $(T_{l,r})_{(l,r)\in\mathcal{G}}$
Ensure: Estimated change-points $(\hat{\tau}_k)_{k\leq\hat{K}}$

$\mathcal{T}_r, \mathcal{S}_r = \varnothing$ for all $r \in \mathcal{R}$ and $\mathcal{S} = \varnothing$
for Increasing $r \in \mathcal{R}$ **do**
  for $l \in \mathcal{D}_r$ s.t. $T_{l,r} = 1$ **do**
    if $[l - r + 1, l + r - 1] \bigcap \mathcal{S} = \varnothing$ **then**
      $\mathcal{T}_r \leftarrow \mathcal{T}_r \cup \{l\}$
      $\mathcal{S}_r \leftarrow \mathcal{S}_r \cup [l - r + 1, l + r - 1]$
    end if
  end for
  $\mathcal{S} = \mathcal{S} \bigcup \mathcal{S}_r$
end for
Let $(\mathrm{C}_k)_{k=1,\dots,\hat{K}}$ be the connected components of $\mathcal{S}$ sorted in increasing order
return $\left(\hat{\tau}_k = \frac{1}{2}(\min \mathrm{C}_k + \max \mathrm{C}_k)\right)_{k=1,\dots,\hat{K}}$

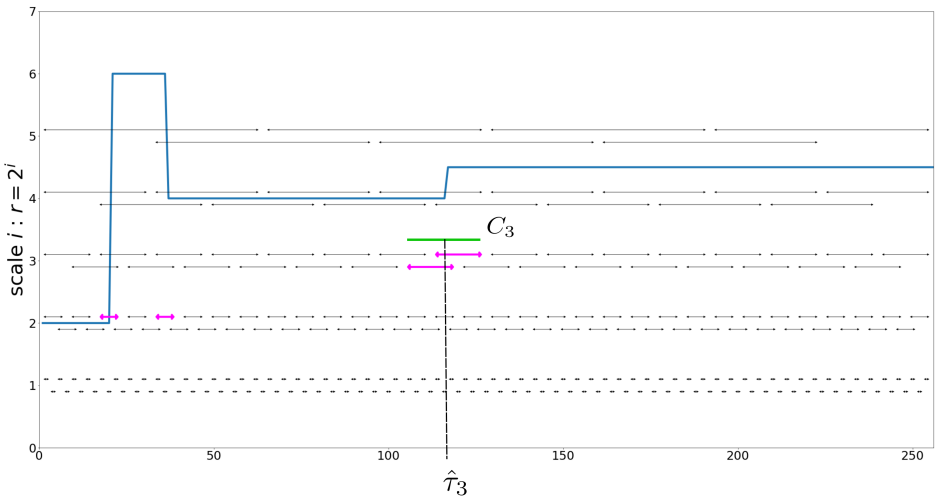


Figure 5.3: Example of our change-point detection procedure with three change-points. The first two change-points have large heights and are detected at a small scale r (in magenta) while the third one is detected at a larger scale r.

Computational Cost. A naive implementation of Algorithm 16 - and also of Algorithm 17 defined in Section 5.9 - requires to compute all tests $T_{l,r}$ on the grid, whereas the aggregation procedure only needs to compute a number of tests $T_{l,r}$ proportional to the size of the grid. More precisely, if the computational cost of $T_{l,r}$ is $\Lambda_{l,r}$ for each (l,r) in the grid $\mathcal{G}$, then the aggregation procedure requires $O(\sum_{(l,r)\in\mathcal{G}} \Lambda_{l,r})$ computations. If for all (l,r), the cost $\Lambda_{l,r}$ is proportional to r, that is $\Lambda_{l,r} = O(r\Lambda)$, then the overall computational cost is $O(\Lambda \sum_{(l,r)\in\mathcal{G}} r)$ which is $O(\Lambda n^3)$ for the complete grid and $O(\Lambda n \log(n))$ for the dyadic grid. One can speed up the full procedure by computing the statistics $T_{l,r}$ and aggregating on the fly by checking whether $[l-r+1, l+r-1]$ intersects $\mathcal{S}$ before evaluating $T_{l,r} = 1$. Indeed, the connected components C_k can be computed at each increasing scale r. Hence, at scale r, one only needs to compute the tests $T_{l,r}$ at locations l such that $[l - r + 1, l + r - 1]$ does not intersect the connected components detected at scales $r' < r$.

5.2.3 General analysis

In this subsection, we provide an abstract theorem translating error controls of the multiple test procedure T in terms of properties of $\widehat{\tau}$. As explained in the introduction, the time series (y_t) may contain change-points that are too small to be detected. Having this in mind, we define a subset $\mathcal{K}^* \subset [K]$ of indices corresponding to so-called significant change-points. As our purpose is to provide deterministic condition so that the change-points in $\mathcal{K}^*$, we need to introduce, for each $k \in \mathcal{K}^*$, an element of the grid $(\bar{\tau}_k, \bar{r}_k) \in \mathcal{G}$ at which the statistic T is expected to detect τ_k. One could think of $\bar{\tau}_k$ as some position close to τ_k and to $\bar{r}_k$ as some radius which is large enough to convey information on the change-point. Recall that the length r_k of the change-point τ_k is defined by $r_k = \min(\tau_{k+1} - \tau_k, \tau_k - \tau_{k-1})$. We assume that the scales $\bar{r}_k$ and the location $\bar{\tau}_k$ of detection satisfy the two following conditions:

$$4(\bar{r}_k - 1) < r_k \quad \text{and} \quad |\bar{\tau}_k - \tau_k| \leq \bar{r}_k - 1. \tag{5.6}$$

The first condition ensures that the scale $\bar{r}_k < r_k/4 + 1$ is small enough compared to the length r_k. The second condition is always satisfied if $\bar{\tau}_k$ is the best approximation of τ_k in $\mathcal{D}_{\bar{r}_k}$ and if the grid $\mathcal{G}$ satisfies the following approximation property

(App): For all $r \in \mathcal{R}$ and all $l \in [r+1, n-r+1]$, there exists $l' \in \mathcal{D}_r$ such that $|l' - l| \leq r - 1$.

This property entails that any point l can be approximated at distance $r-1$ by some location in $\mathcal{D}_r$. This also implies that each point $l \in [r+1, n-r]$ belongs to at least one segment $(l'-r, l'+r)$ where l_1 lies in $\mathcal{D}_r$. In practice, the a-adic grids $\mathcal{G}_a$ and the complete grid satisfy **(App)**.

Next, we introduce an event on the tests $(T_{l,r})$ under which the change-point estimator $\widehat{\tau}$ of Algorithm 16 performs well. In the following, we write $\mathcal{H}_0$, the collection of all possible $(l,r) \in J_n$ such that there is no change in $[l-r+1, l+r-1]$, i.e. $\Gamma(\mathbb{P}_t)$ is constant on $[l-r, l+r)$. Equivalently, we have

$$(l,r) \in \mathcal{H}_0 \quad \text{iff} \quad (l-r, l+r) \cap \{\tau_k, k = 1, \dots, K\} = \varnothing \ . \tag{5.7}$$

For a collection $\mathcal{K}^*$ and some elements of the grid $(\bar{\tau}_k, \bar{r}_k)$ satisfying (5.6), the Event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$ is defined as the conjunction of the two following properties: (i) **(No false positive)** $T_{l,r} = 0$ for all $(l,r) \in \mathcal{H}_0 \cap \mathcal{G}$ (ii) **(Detection of significant change-points)** for every $k \in \mathcal{K}^*$, we have $T_{\bar{\tau}_k, \bar{r}_k} = 1$.

The first property states that T performs no type I errors on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$, whereas the second property enforces that all the significant change-points are detected by the specific tests $T_{\bar{\tau}_k, \bar{r}_k}$.

Theorem 5.2.1. *The following holds for any grid $\mathcal{G}$, any local test statistic T, any non-negative integer K, any distribution with K change-points, any $\mathcal{K}^* \subset [K]$ and scales and locations $(\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*}$ in $\mathcal{G}$ satisfying Assumption (5.6). Under the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$, the estimated change-point vector $\hat{\tau}$ returned by Algorithm 16 satisfies the two following properties*

- **Significant change-points are detected**: *for all $k \in \mathcal{K}^*$, there exists $k' \leq \hat{K}$ such that $|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1 < \frac{r_k}{4}$.*
- *(***NoSp***): No Spurious change-point is detected (5.1).*

The first property states that so-called significant change-points $(\tau_k)_{k \in \mathcal{K}^*}$ are detected by the generic algorithm at the right scale. The no-spurious property (5.1) guarantees that, around any true change-point τ_k, the procedure estimates at most one single change-point $\widehat{\tau}_l$. Importantly, the theorem does not make any assumption on the non-significant change-points. In fact, change-points τ_k with $k \in [K] \setminus \mathcal{K}^*$ may or may not be detected. In general, we can only conclude from Theorem 5.2.1 that $|\mathcal{K}^*| \leq \widehat{K} \leq K$ on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$.

Theorem 5.2.1 is abstract, but its main virtue is to translate multiple testing properties into change-point detection properties. For a specific problem such as multivariate mean change-point detection considered in the next section, the construction of a near optimal procedure boils down to introducing a collection of local test statistics, such that (a) change-points τ_k belong to $\mathcal{K}^*$ under minimal conditions, (b) the scale $\bar{r}_k$ is the smallest possible, and (c) the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$ holds with high probability.

In the case where all the change-points are significant, the result of Theorem 5.2.1 can be reformulated as follows:

Corollary 5.2.2. *The following holds for any grid $\mathcal{G}$, any local test statistic T, any non-negative integer K, any distribution with K change-points, any $(\bar{\tau}_k, \bar{r}_k)_{k=1,\dots,K}$ in $\mathcal{G}$ satisfying Assumption (5.6). Under the event $\mathcal{A}(T, [K], (\bar{\tau}_k, \bar{r}_k)_{k=1,\dots,K}$, the estimated change-point vector $\hat{\tau}$ returned by Algorithm 16 satisfies $\widehat{K} = K$ and,*

$$|\hat{\tau}_k - \tau_k| < \bar{r}_k - 1 \leq \frac{r_k}{4} \qquad \text{for all } k = 1, \dots, K \ .$$

Let us respectively define the Hausdorff distance and the Wasserstein distance of two vectors $(u_1, \dots, u_K)$ and $(v_1, \dots, v_K)$ in $\mathbb{R}^K$ by $d_H(u,v) = \max_{k=1,\dots,K} |u_k - v_k|$ and $d_W(u,v) = \sum_{k=1,\dots,K} |u_k - v_k|$. Then, Corollary 5.2.2 straightforwardly implies that, if $\mathcal{K}^* = [K]$, then these two losses are bounded as follows

$$d_H(\hat{\tau}, \tau) \le \max_{k=1,\dots,K} (\bar{r}_k - 1) \quad \text{and} \quad d_W(\hat{\tau}, \tau) \le \sum_{k=1,\dots,K} (\bar{r}_k - 1) \ .$$

As an alternative of Algorithm 16, one could use other bottom-up aggregating procedures. For instance, Algorithm 17 defined in Section 5.9 also satisfies Theorem 5.2.1. Although these two algorithms are closely related, Algorithm 16 is slightly more conservative than Algorithm 17 since it merges all detection intervals at a given resolution while Algorithm 17 only keeps one interval at a given resolution when multiple intervals intersect - the one with smallest index t. While the minimax properties of both methods are comparable - at least up to a multiple constant - the choice of aggregation method will have an influence in practice on the outcome: Algorithm 16 will be slightly more stable, detect less change-points, and provide wider confidence interval around them, while Algorithm 17 will be slightly more sensitive to smaller changes, i.e. detect smaller change-points, will be more precise, and somewhat less stable.

Theorem 5.2.1 ensures that, if $T_{\bar{\tau}_k, \bar{r}_k} = 1$ with $(\bar{\tau}_k, \bar{r}_k)$ satisfying Assumption (5.6), then the change-point τ_k is detected. Inspecting the proof of Theorem 5.2.1, one easily checks that Assumption (5.6) is minimal for Algorithm 16 (and also for Algorithm 17). Still, one may wonder whether any generic algorithm has to require that $4(\bar{r}_k - 1) < r_k$ to detect the change-points or if there exists a generic algorithm where the constant 4 in the above condition can be improved.

Comparison with narrowest over threshold methods. As mentioned in the introduction, other aggregation procedures have been proposed in the literature. In particular, the narrowest over threshold scheme proposed by [6] and later used in [51] is also closely related to the local segmentation algorithm of Chan and Chen [47]. A simple extension of these procedures for generic change-point problems and for a general collection of tests $(T_{l,r})$ would amount to modifying algorithm 16 by selecting locations l in $\mathcal{D}_r$ such that $T_{l,r} = 1$ and $[l - r + 1, l + r - 1]$ does not intersect previously detected change-points, whereas we require in Algorithms 16 and 17, that $[l - r + 1, l + r - 1]$ does not intersect previously detected confidence intervals. In some way, the narrowest-over threshold scheme is therefore less conservative. Unfortunately, there is no generic result in the form of Theorem 5.2.1 for such procedures and, from informal arguments, we doubt that the corresponding procedure provably achieves (**NoSp**) under a control of the FWER of the tests. Inspecting the proof of Theorem 1 in [6] and Theorem 3 in [51] for univariate mean change-point problems, one observes that the chosen threshold is much larger than what is needed to control the FWER so that the theoretical threshold is certainly over-conservative – see step 5 of the proof of Theorem 1 in [6]. In contrast, Theorem 1 in [47] for univariate change-point problems is based on the minimal threshold, but the proof relies on the important assumption that the number K of change-point remains bounded while n goes to infinity. Besides, it is not clear how one could extend the arguments to more general settings.

5.3 Multivariate Gaussian change-point detection

We now turn to the multivariate change-point model introduced in Section 5.1.2. Throughout this section, we assume that the random vectors ε_t are independently and identically distributed with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. Since we shall apply the general aggregation procedures introduced in the previous section, our main job here is to introduce a near-optimal testing procedure.

Fix some quantity $\delta \in (0,1)$. At the end of the section, $1 - \delta$ will correspond to the probability of the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$ introduced in the previous section. Alternatively, one may interpret δ as an upper bound of the desired probability that the change-point detection procedure detects a spurious change-points. Recall that, for a change-point τ_k, s_k stands for the sparsity of the difference $\mu_{k+1} - \mu_k$. The energy of a given change-point τ_k is c_0-high if

$$r_k \Delta_k^2 \ge c_0 \sigma^2 \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k} \sqrt{\log\left(\frac{n}{r_k \delta} \right)} \right) + \log\left(\frac{n}{r_k \delta} \right) \right] \ , \tag{5.8}$$

for some universal constant c_0 to be defined later. We show in this section that when c_0 is large enough, all high-energy change-points can be detected. Conversely, it is established in Section 5.5 that Condition (5.8) is (up to a multiplicative constant) optimal for detecting change-points and cannot be weakened.

Let us now discuss the different regimes contained in Equation (5.8). In what follows, define

$$\psi^{(g)}_{n,r,s} := s \log\left(1 + \frac{\sqrt{p}}{s} \sqrt{\gamma_r} \right) + \gamma_r \ ; \qquad \gamma_r := \log\left(\frac{n}{r\delta} \right) \ ,$$

in order to alleviate notations. If $\gamma_r \geq p/2$, then $\psi^{(g)}_{n,r,s} \asymp \gamma_r$ where $u \asymp v$ means that for two positive numerical constants c_1 and c_2, one has $c_1 v \leq u \leq c_2 v$. This corresponds to the minimal energy condition for detection in the univariate case, i.e. when $p = 1$; see [92]. The condition $\gamma_r \geq p/2$ occurs when p is rather small and the scale r is much smaller than n. If $\gamma_r \leq p/2$, then

$$\psi^{(g)}_{n,r,s} \asymp \begin{cases} \gamma_r & \text{if } s \leq \frac{\gamma_r}{\log(p)-\log(\gamma_r)} \\ s\log\left(2\frac{p}{s^2}\gamma_r\right) & \text{if } \frac{\gamma_r}{\log(p)-\log(\gamma_r)} < s < \sqrt{p\gamma_r} \\ \sqrt{p\gamma_r} & \text{if } s \geq \sqrt{p\gamma_r} \end{cases} .$$

We define $\mathcal{K}^* \subset [K]$ as the subset of indices such that τ_k satisfies (5.8). For any $k \in \mathcal{K}^*$, we define r^*_k as the minimum radius r such that an inequality similar to (5.8) is satisfied for $r\Delta_k^2$, namely

$$r^*_k = \min\left\{ r \in \mathbb{R}^+ : \quad r\Delta_k^2 \geq c_0\sigma^2\left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r\delta}\right)}\right) + \log\left(\frac{n}{r\delta}\right)\right]\right\} . \tag{5.9}$$

In the following, we introduce multi-scale tests for respectively dense and sparse change-points. For simplicity, we restrict our attention to the dyadic grid $\mathcal{G}_D = (\mathcal{R}, \mathcal{D})$ introduced in the previous section (see Equation (5.5)), the complete grid being used in the next section.

To apply Theorem 5.2.1, we will consider an event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ in the proof of Corollary 5.3.3 where the scale $\bar{r}_k \in \mathcal{R}$ is of the same order as $r^*_k \in \mathbb{R}^+$.

5.3.1 Dense change-points

We focus here on dense change-points for which s_k is possibly as large as p. Given $\kappa > 0$, τ_k is a κ-dense high-energy change-point if

$$r_k\Delta_k^2 \geq \kappa\sigma^2\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} + \log\left(\frac{n}{r_k\delta}\right)\right) . \tag{5.10}$$

The requirement (5.10) is analogous to (5.8) when $s_k \geq [p\log(n/(r_k\delta))]^{1/2}$. For any κ-dense high-energy change-point, we define $\bar{r}^{(\mathrm{d})}_k \in \mathcal{R}$ as the minimum radius $r \in \mathcal{R}$ such that an inequality of the same type as (5.10) is satisfied for $r\Delta_k^2$,

$$\bar{r}^{(\mathrm{d})}_k = \min\left\{ r \in \mathcal{R} : \, 8r\Delta_k^2 \geq \kappa\sigma^2\left(\sqrt{p\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right)\right\} .$$

Intuitively, $\bar{r}^{(\mathrm{d})}_k$ corresponds to the smallest scale such that τ_k is guaranteed to be detected. By definition, we have $4(\bar{r}^{(\mathrm{d})}_k - 1) \leq r_k$. Let $\bar{\tau}^{(\mathrm{d})}_k$ be the best approximation of τ_k in the grid with scale $\bar{r}^{(\mathrm{d})}_k$. By definition of the dyadic grid, we have $|\bar{\tau}^{(\mathrm{d})}_k - \tau_k| \leq \bar{r}^{(\mathrm{d})}_k/4$.

For any positive integers $r \in [1; n]$ and $l \in [r+1, n+1-r]$, we define the statistic $\Psi^{(\mathrm{d})}_{l,r} := \|\mathbf{C}_{l,r}\|^2 - p$. If θ is constant over $[l-r, l+r)$, then the expectation of $\Psi^{(\mathrm{d})}_{l,r}$ is zero. Recall that the rescaled CUSUM statistic $\mathbf{C}_{l,r}$ depends on the noise level σ, and the statistic $\Psi^{(\mathrm{d})}_{l,r}$ therefore requires the knowledge of σ. To calibrate the corresponding test $T^{(\mathrm{d})}_{l,r}$ rejecting for large values of $\Psi^{(\mathrm{d})}_{l,r}$ we introduce

$$T^{(\mathrm{d})}_{l,r} := \mathbf{1}\left\{\Psi^{(\mathrm{d})}_{l,r} > x^{(\mathrm{d})}_r\right\} ; \qquad x^{(\mathrm{d})}_r := 4\left(\sqrt{p\log\left(\frac{2n}{r\delta}\right)} + \log\left(\frac{2n}{r\delta}\right)\right) .$$

Proposition 5.3.1. *There exists a universal constant $\kappa_{\mathrm{d}} > 0$ and an event $\xi^{(\mathrm{d})}$ of probability larger than $1 - 2\delta$ such that (i) $T^{(\mathrm{d})}_{l,r} = 0$ for all $(l, r) \in \mathcal{H}_0 \cap \mathcal{G}_D$ and (ii) $T^{(\mathrm{d})}_{\bar{\tau}^{(\mathrm{d})}_k, \bar{r}^{(\mathrm{d})}_k} = 1$ for all κ_{d}-dense high-energy change-point τ_k.*

The above proposition ensures that, on the event $\xi^{(\mathrm{d})}$, the collection of tests $T^{(\mathrm{d})}_{l,r}$ detects all dense high-energy change-points at the scale $\bar{r}^{(\mathrm{d})}_k$ and makes no false positives on the dyadic grid $\mathcal{G}_D$. If we plugged this collection of tests into the general multiple change-point procedure, then Theorem 5.2.1 would entail that all κ_{d}-dense high-energy change-points are discovered and localized and that $\widehat{\tau}$ does not detect any spurious change-point. In the next subsection, we introduce alternative tests that are tailored to sparse change-points and thereby allow to detect change-points that are not κ_{d}-dense high-energy but still satisfy the energy condition (5.8).

5.3.2 Sparse change-points

5.3.2.1 Energy condition

For a given $1 \le k \le K$, the change-point τ_k is a κ-sparse high-energy change-point if $s_k \le [p \log(n/(r_k\delta))]^{1/2}$ and

$$r_k \Delta_k^2 \ge \kappa\sigma^2 \left(s_k \log\left(\frac{p}{s_k^2} \log\left(\frac{n}{r_k\delta} \right) \right) + \log\left(\frac{n}{r_k\delta} \right) \right) \quad . \tag{5.11}$$

If τ_k is a κ-sparse high-energy change-point, we define $\bar{r}_k^{(s)}$ as the minimum scale such that an inequality similar to (5.11) is satisfied :

$$\bar{r}_k^{(s)} = \min\left\{ r \in \mathcal{R} : \quad 8r\Delta_k^2 \ge \kappa\sigma^2 \left(s_k \log\left(\frac{p}{s_k^2} \log\left(\frac{n}{r\delta} \right) \right) + \log\left(\frac{n}{r\delta} \right) \right) \right\} \quad .$$

As in the dense case, we have $4(\bar{r}_k^{(s)} - 1) \le r_k$. Set $\bar{\tau}_k^{(s)}$ as the best approximation of τ_k in the grid $\mathcal{D}_{\bar{r}_k^{(s)}}$ at scale τ_k. By definition of the dyadic grid, we have $|\bar{\tau}_k^{(s)} - \tau_k| \le \bar{r}_k^{(s)}/4$. We introduce below two statistics for handling this problem.

5.3.2.2 Berk-Jones Test

The Berk-Jones test [65] is a variation of the Higher-Criticism test originally introduced in [30] for signal detection. It has been previously studied in [15] for sparse segment detection. We decided to use the Berk-Jones test in this chapter because of its intrinsic formulation in terms of the quantiles of a Bernoulli distribution, but the Higher-Criticism test would reach the same rates of detection within a constant factor. We use the notation $\mathbb{N}^*$ to denote the set of positive itegers. Given (l, r) in the grid $\mathcal{G}_D$, we first introduce $N_{x,l,r}$ as the number of coordinates of $\mathbf{C}_{l,r}$ that are larger than x in absolute value.

$$N_{x,l,r} = \sum_{i=1}^{p} \mathbf{1}_{|\mathbf{C}_{l,r,i}| > x} \tag{5.12}$$

If $(l, r) \in \mathcal{H}_0$, then the rescaled CUSUM statistic follows a standard normal distribution and $N_{x,l,r}$ therefore follows a Binomial distribution with parameters p and $2\overline{\Phi}(x)$. The Berk-Jones test amounts to rejecting the null, when at least one of the statistics $N_{x,l,r}$, for $x \in \mathbb{N}^*$, is significantly large. Next, we formalize what we mean by 'large'.

For any $u > 0$, any $q_0 \in [0, 1]$, and positive integer p_0, denote $\overline{Q}(u, p_0, q_0) = \mathbb{P}[\mathcal{B}(p_0, q_0) > u]$ the tail distribution function of a Binomial distribution with parameters p_0 and q_0. Given $\delta \in [0, 1]$, we then write $\overline{Q}^{-1}(\delta, p_0, q_0)$ for the corresponding quantile function,

$$\overline{Q}^{-1}(\delta, p_0, q_0) = \inf_u \left[\mathbb{P}[\mathcal{B}(p_0, q_0) > u] \le \delta \right] \quad .$$

Given a scale $r \in \mathcal{R}$ and a positive integer x, we define the weights

$$\delta_{x,r}^{(BJ)} = \frac{6\delta r}{\pi^2 x^2 |\mathcal{D}_r| n} \quad . \tag{5.13}$$

This allows us to define the Berk-Jones statistic over $[l - r, l + r)$ as the test rejecting the null when at least one $N_{x,l,r}$ is large.

$$T_{l,r}^{(BJ)} = \max_{x \in \mathbb{N}^*} \mathbf{1}\left\{ N_{x,l,r} > \overline{Q}^{-1}(\delta_{x,r}^{(BJ)}, p, 2\overline{\Phi}(x)) \right\} \quad . \tag{5.14}$$

Equivalently, $T_{l,r}^{(BJ)}$ is an aggregated test based on the statistics $N_{x,l,r}$ with weights $\delta_{x,r}^{(BJ)}$. From the above remark and a union bound, we deduce that the probability that the collection of tests $\{T_{l,r}^{(BJ)}, (l, r) \in \mathcal{G}_D\}$ rejects a least one false positive is at most δ:

$$\mathbb{P}\left[\max_{(l,r) \in \mathcal{H}_0 \cap \mathcal{G}_D} T_{l,r}^{(BJ)} = 1 \right] \le \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{x \in \mathbb{N}^*} \delta_{x,r}^{(BJ)} \le \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \frac{\delta r}{|\mathcal{D}_r| n} \le \sum_{r \in \mathcal{R}} \frac{\delta r}{n} \le \delta \quad ,$$

where we recall that $(l, r) \in \mathcal{H}_0$ if and only if Θ is constant on $[l - r, l + r)$. Although one may think from the definition (5.14) that $T_{l,r}^{(BJ)}$ involves an infinite number of $N_{x,l,r}$, this is not the case. Indeed, $N_{x,l,r}$ is

a non-increasing function of x whereas for all x such that $2p\overline{\Phi}(x) \le \delta_{x,r}^{(\mathrm{BJ})}$, we have $\overline{Q}^{-1}(\delta_{x,r}^{(\mathrm{BJ})}, p, 2\overline{\Phi}(x)) = 0$. Writing $x_{0,r}$ the smallest x such that $2p\overline{\Phi}(x) \le \delta_{x,r}^{(\mathrm{BJ})}$ we derive

$$T_{l,r}^{(\mathrm{BJ})} = \max_{x=1,\dots,x_{0,r}} \mathbf{1}\left\{N_{x,l,r} > \overline{Q}^{-1}(\delta_{x,r}^{(\mathrm{BJ})}, p, 2\overline{\Phi}(x))\right\} .$$

Since, for any $x > 0$, we have $\overline{\Phi}(x) \le e^{-x^2/2}$, one can deduce that $x_{0,r} \le c[\log(np/(r\delta))]^{1/2}$, for some numerical constant $c > 0$.

5.3.2.3 Partial norm statistics

The Berk-Jones test is able to detect change-points τ_k for which there exists s such that the s largest squared coordinates of $\mu_k - \mu_{k-1}$ are larger than $C(\log(ep/s^2) + \log(n/r_k)/s)$ with a large enough constant C. However, it may happen that τ_k satisfies the energy condition (5.8) and that the s largest coordinates of $\mu_k - \mu_{k-1}$ are negligible compared to $\log(n/r_k)/s$, mainly because $s \mapsto 1/s$ is not summable. To solve this issue, we introduce a second sparse statistic based on the partial sums. Let

$$\mathcal{Z} = \left\{1, 2, 2^2, \dots, 2^{\lfloor \log_2(p) \rfloor}\right\}$$

denote the dyadic set. Only the sparsities $s \in \mathcal{Z}$ will be analysed by the partial norm statistic. For any (l,r) in the grid $\mathcal{G}_D$, we respectively write $\mathbf{C}_{l,r,(1)}, \mathbf{C}_{l,r,(2)}, \dots$ the reordered entries of $\mathbf{C}_{l,r}$ by decreasing absolute value, that is $|\mathbf{C}_{l,r,(1)}| \ge \dots \ge |\mathbf{C}_{l,r,(p)}|$. Then, for $s \in \mathcal{Z}$, we define the partial CUSUM norm by

$$\Psi_{l,r,s}^{(\mathrm{p})} = \sum_{i=1}^{s} \left(\mathbf{C}_{l,r,(i)}\right)^2 . \tag{5.15}$$

Then, we define the test $T_{l,r}^{(\mathrm{p})}$ rejecting the null when at least one of the partial norms is large

$$x_{r,s}^{(\mathrm{p})} := x_{r,s}^{(\mathrm{p})}(\delta) = 4s\log\left(\frac{2ep}{s}\right) + 4\log\left(\frac{n}{r\delta}\right); \qquad T_{l,r}^{(\mathrm{p})} = \max_{s\in\mathcal{Z}} \mathbf{1}\left\{\Psi_{l,r,s}^{(\mathrm{p})} > x_{r,s}^{(\mathrm{p})}\right\} .$$

Finally, we define the sparse test by aggregating both the Berk-Jones test and the partial norm test. For any $(l,r) \in \mathcal{G}_D$, let $T_{l,r}^{(\mathrm{s})} = T_{l,r}^{(\mathrm{p})} \vee T_{l,r}^{(\mathrm{BJ})}$. The next proposition controls the error of this collection of tests.

Proposition 5.3.2. *There exists a universal constant $\kappa_{\mathrm{s}} > 0$ and an event $\xi^{(\mathrm{s})}$ of probability larger than $1 - 4\delta$ such that (i) $T_{l,r}^{(\mathrm{s})} = 0$ for all $(l,r) \in \mathcal{H}_0 \cap \mathcal{G}_D$ and (ii) $T_{\bar{r}_k^{(\mathrm{s})}, \bar{r}_k^{(\mathrm{s})}}^{(\mathrm{s})} = 1$ for all κ_{s}-sparse high-energy change-point τ_k.*

Here we introduced two different statistics for the same sparse regime $s_k \le [p\log(n/(r_k\delta))]^{1/2}$ - the Berk-Jones statistic and the partial sums statistic - mainly to solve a problem of integrability. We made this choice for the sake of simplicity, but we could have used a single test, as presented in [57]

$$\Psi_{x,l,r}^{(\mathrm{LGS})} = \sum_{i=1}^{p} \left(\mathbf{C}_{l,r,i}^2 - \mathbb{E}\left[Z | Z \ge x\right]\right) \mathbf{1}\{\mathbf{C}_{l,r,i}^2 \ge x\} ,$$

where Z follows a standard normal distribution $\mathcal{N}(0,1)$. This statistic leads to the same type of result as the Berk-Jones statistic when enough coordinates $\mu_k - \mu_{k-1}$ are large in absolute value, and it is comparable to the partial sums statistic when its threshold x becomes low enough.

5.3.3 Consequences

To conclude this section, it suffices to observe that, for c_0 in (5.8), any c_0-high-energy change-point τ_k in the sense of (5.8) is either a $\frac{c_0}{2}$-dense or a $\frac{c_0}{2}$-sparse high-energy change-point. Hence, upon defining the test $T_{l,r} = T_{l,r}^{(\mathrm{d})} \vee T_{l,r}^{(\mathrm{s})}$ for $(l,r) \in \mathcal{G}_D$, we consider the change-point procedure $\widehat{\tau}$ defined in Algorithm 16. Gathering Theorem 5.2.1 with Proposition 5.3.1 and Proposition 5.3.2, we obtain the following.

Corollary 5.3.3. *There exists a universal constant $c_0 > 0$ such that, with probability higher than $1 - 6\delta$, the estimator $\widehat{\tau}$ satisfies (**NoSp**) and **detects** all c_0-high-energy change-points (as defined in (5.8)) τ_k in the sense*

$$d_{H,1}(\hat{\tau}, \tau_k) < \frac{r_k^*}{2} \le \frac{r_k}{2} ,$$

where r_k^* is defined in (5.9).

If the change-points are of high-energy, that is $\mathcal{K}^* = [K]$, then Corollary 5.3.3 can be reformulated as follows:

Corollary 5.3.4. *Assume that for all $k = 1, \dots, K$, τ_k is a c_0-high-energy change-point (see (5.8)) where c_0 is the same as in Corollary 5.3.3. Then, with probability higher than $1 - 6\delta$, the estimator $\widehat{\tau}$ satisfies $\hat{K} = K$ and*

$$|\hat{\tau}_k - \tau_k| < \frac{r_k^*}{2} \le \frac{r_k}{2} \ , \qquad \text{for all } k = 1, \dots, K \ .$$

In particular, one can respectively bound the Hausdorff and the Wasserstein losses, with probability higher than $1 - 6\delta$ by

$$d_H(\hat{\tau}, \tau) \le \max_{k=1,\dots,K} \frac{r_k^*}{2} \quad \text{and} \quad d_W(\hat{\tau}, \tau) \le \sum_{k=1,\dots,K} \frac{r_k^*}{2} \ . \tag{5.16}$$

In Section 5.5, we establish that the Condition (5.8) is (up to a multiplicative constant) unimprovable and corresponds to the detection threshold for multivariate change-points.

Corollary 5.3.4 can be compared to the result of [103] on multivariate change-point detection in the multiple change-point setting. Using a method based on the CUSUM statistic and assuming that there are only high-energy change-points, the authors also obtain an upper bound on the energy necessary to detect the change-points. However, this result does not adapt to r_k, Δ_k, s_k, and the detection rate is suboptimal in many regimes. Writing $r = \min_{k=1,\dots,K} r_k$, $\Delta = \min_{k=1,\dots,K} \Delta_k$ and $s = \max_{k=1,\dots,K} s_k$, Theorem 5 of [103] requires two conditions of the type $r\Delta^2 \ge c(\frac{n}{r})^4 \log(np)$ and $r\Delta^2 \ge cs\frac{n}{r}\log(np)$. This detection rate is therefore suboptimal by a polynomial factor in n/r when r is of smaller order than n, and by a logarithmic factor $\log(np)$ instead of $\log(1 + \sqrt{p}/s \log(n/r)) + \frac{1}{s}\log(n/r)$ when r is of order n. Closer to our results, [47] have introduced another bottom-up procedure in the very specific asymptotic setting $n = e^{p^\zeta}$ for $\zeta \in (0, 1)$ with a fixed K number of change-points. Assuming that, for each change-point, at least s coordinates of $\mu_{k+1} - \mu_{k+1}$ are larger than ζ in absolute value, [47] establish that their procedure provably detects the change-points as long as

$$rs\zeta^2 \ge c \begin{cases} \sqrt{p\log(n)} & \text{if } s \ge 0.5\sqrt{p\log(n)} \\ s\log\left(\frac{p}{s^2}\log(n)\right) & \text{if } s \le 0.5\sqrt{p\log(n)} \end{cases} .$$

In their specific asymptotic regime and when all non-zero coordinates are of the same order, and all the change-points have a similar length r_k, their result is similar to ours up to the logarithmic terms. Indeed, for equispaced change-points, our logarithmic term $\log(n/r_k) = \log(K)$ is much smaller than $\log(n)$. Besides, their result does not handle the presence of low-energy change-points and does not hold beyond the asymptotic regime $n = e^{p^\zeta}$. In contrast, our condition (5.8) for high-energy change-points entails that the detection conditions are qualitatively different for other scalings in n and p. On the technical side, our condition (5.8) is of l_2 type whereas that in [47] is of minimal non-zero type. Recovering the tight l_2 conditions turns out to be much more challenging as we need to handle situations where some coordinates have different orders of magnitude. This is the main reason why we need to resort to a combination of the Berk-Jones and the partial-norm statistics.

Comparison to one change-point problem. When one knows that $K \le 1$ (at most one change-point), then [57] proved that it is possible to detect τ_1 if and only if $r_1\Delta_1^2 \ge c\sigma^2\big[s_1\log(1 + \frac{1}{s_1}\sqrt{p\log\log 8n}) + \log\log 8n\big]$. As in the univariate setting, the problem with only one change-point is simpler than for general $K \ge 2$. As for our procedure, Liu et al. [57] rely on statistics based on the CUSUM - a chi square statistics in the dense case and a thresholded sum of squared coordinates in the sparse case - to detect and localize τ_1. It turns out that the detection procedure of [57] adapts to distance $r_1 = \max(\tau_1 - 1, n + 1 - \tau_1)$ the boundary, and one could refine their result by stating that τ_1 is detectable if and only if $r_1\Delta_1^2 \ge c\sigma^2[s_1\log(1 + \frac{1}{s_1}\sqrt{p\log\log(2n/r_1)}) + \log\log(2n/r_1)]$ which is more smaller when r_1 is of the order of n. This refined result is in the same spirit as our bounds for mutiple change-point, but the rate is faster because one obtains $\log\log(n/r_1)$ - instead of $\log(n/r_k)$ in our case. The reason for this faster rate is due to the relative simplicity of the problem with only one change-point. Indeed, in single change-point detection, there is no need to look for change-points at all positions and scale at the same time, since scale and positions are related. This implies that it is possible to attain faster rates than in multiple change-point detection. The comparison between single and multiple change-point detection is thoroughly done in [92] for univariate models.

Computational Cost. The cost of the tests $T_{l,r}^{(\mathrm{d})}$ in the dense regime is $O(rp)$. The computation of the partial norm statistic requires to sort the coordinates $\mathbf{C}_{l,r,i}$ of the CUSUM statistic, which takes $O(p(r + \log(p)))$ operations. Since only the thresholds $x \le c\log(np/(r\delta))^{1/2}$ are needed to compute the Berk-Jones statistic, it holds that, for $\delta \ge (np)^{-c}$ with a numerical constant $c > 0$, the computational cost of the Berk-Jones statistic is $O(p(r + \log(np)))$. Thus, for each (l, r), the overall computational cost of the test $T_{l,r} = T_{l,r}^{(\mathrm{d})} \vee T_{l,r}^{(\mathrm{s})}$ is $\Lambda = O(p(r + \log(np)))$, and the computational cost of the whole change-point detection procedure on the dyadic grid is $O(np\log(np))$.

5.4 Multi-scale change-point detection with sub-Gaussian noise

We now turn to the more general case of sub-Gaussian distributions [91]. Given a random variable Z, define its ψ_2-norm by $\|Z\|_{\psi_2} = \inf\{x > 0,\ \mathbb{E}[\exp(Z^2/x^2)] \le 2\}$. Given $L > 0$, a mean zero real random variable is said to be L-sub-Gaussian if $\|Z\|_{\psi_2} \le L$. This implies in particular that, for all $x \ge 0$, one has $\mathbb{P}(|Z| \ge x) \le 2\exp(-x^2/L^2)$. Throughout this section, we assume that, for $t = 1, \dots, n$, the random vectors ε_t are independent, have independent L-sub-Gaussian components $\varepsilon_{t,i}$, for $i = 1, \dots, p$ with variance σ^2. As in the previous section, we apply the general aggregation procedures introduced in Section 5.2. As a consequence, our main task boils down to introducing a near-optimal multiple testing procedure indexed by a grid for detecting the existence of a change-point. Here, we shall rely on the complete grid $\mathcal{G}_F = J_n = \left\{(l,r) \ :\ r = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor \text{ and } l = r+1, \dots, n-r\right\}$ whose size is quadratic with respect to n. All the results presented in this section are still valid (but with different numerical constants) if we keep the dyadic grid $\mathcal{G}_D$ as in the previous section. Here, we use the complete grid as a proof of concept that one can rely on the full collection of possible segments without deteriorating the rates. Still, controlling the behavior of the procedure on the complete grid is technically more involved and requires chaining arguments. A detailed comparison between the complete and dyadic grids is made in Section 5.7.

In order to emphasize the common points with the previous section, we use the same notation $\mathcal{K}^*$ for the collection of high-energy change-points[1], $\bar{r}_k$ for the scales associated to the k-th change-points[2], Ψ for the statistics, T for the test and x for the thresholds although these quantities are slightly changed to cope with the sub-Gaussian tail distribution. We follow the same scheme as for the Gaussian case and first introduce multi-scale tests for dense change-points before turning to sparse change-points. As in the previous section, we consider some $\delta \in (0,1)$ corresponding to the type I error probability.

5.4.1 Dense change-points with sub-Gaussian noise

Recall that, for a change-point τ_k, s_k stands for the sparsity of the difference $\mu_{k+1} - \mu_k$. We focus here on dense change-points for which s_k is possibly as large as p. Given $\kappa > 0$, τ_k is a κ-dense high-energy change-point if

$$r_k \Delta_k^2 \ge \kappa L^2 \left(\sqrt{p \log\left(\frac{n}{r_k \delta}\right)} + \log\left(\frac{n}{r_k \delta}\right) \right) \ . \tag{5.17}$$

This condition is very similar to its counterpart (5.10) for Gaussian noise. Still, we introduce it here for the sake of completeness. For $k \in [K]$ such that τ_k is a κ-dense high-energy change-point, we define $\bar{r}_k^{(\mathrm{d})}$ as the minimum length such that an inequality similar to (5.17) is satisfied :

$$\bar{r}_k^{(\mathrm{d})} = \min\left\{ r \in \mathbb{N}^* : \quad 4r\Delta_k^2 \ge \kappa L^2 \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \right\} \ .$$

As in the Gaussian case in Section 5.3, $\bar{r}_k^{(\mathrm{d})}$ corresponds to the smallest scale such that τ_k is guaranteed to be detected. For any κ-dense high-energy change-point, it holds that $4(\bar{r}_k^{(\mathrm{d})} - 1) < r_k$. For any positive integers $(l,r) \in \mathcal{G}_F$, we consider the same CUSUM-based statistic $\Psi_{l,r}^{(\mathrm{d})} := \|\mathbf{C}_{l,r}\|^2 - p$ as for Gaussian noise. Let $\bar{c}_{\mathrm{thresh}}^{(\mathrm{d})} > 0$ be a tuning parameter to be discussed later. To calibrate the corresponding multiple test procedures $(T_{l,r}^{(\mathrm{d})})$ with $(l,r) \in \mathcal{G}_F$ rejecting for large values of $\Psi_{l,r}^{(\mathrm{d})}$ we introduce

$$T_{l,r}^{(\mathrm{d})} := \mathbf{1}\left\{ \Psi_{l,r}^{(\mathrm{d})} > x_r^{(\mathrm{d})} \right\} \ ; \qquad x_r^{(\mathrm{d})} = \bar{c}_{\mathrm{thresh}}^{(\mathrm{d})} \frac{L^2}{\sigma^2} \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \ .$$

Proposition 5.4.1. *There exists a numerical constant $\bar{c}_{\mathrm{thresh}}^{(\mathrm{d})} > 0$ such that the following holds for any $\kappa_{\mathrm{d}} > 32\bar{c}_{\mathrm{thresh}}^{(\mathrm{d})}$. With probability higher than $1-\delta$, one has (i) $T_{l,r}^{(\mathrm{d})} = 0$ for all $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$ and (ii) $T_{\tau_k, \bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} = 1$ for all κ_{d}-dense high-energy change-points τ_k.*

In comparison to Proposition 5.3.1 in the previous section, there are two differences. First, we need to cope with sub-Gaussian distribution by applying the Hanson-Wright inequality. Most importantly, the grid $\mathcal{G}_F$ is much larger than $\mathcal{G}_D$ so that we cannot simply consider each test $T_{l,r}$ separately and simply apply a union bound as in the previous section. To handle the dependencies between the statistics $\Psi_{l,r}^{(\mathrm{d})}$, we have to apply a chaining argument. In fact, the thresholds $x_r^{(\mathrm{d})}$ are similar to their counterpart in the previous section, whereas

[1] See Equation (5.20) as the energy condition is slightly different in the sub-Gaussian setting.
[2] Re-defined in Equation (5.21).

the number $|\mathcal{G}_F|$ of tests is now proportional to n^2. In principle, the benefit of using the full grid $\mathcal{G}_F$ is that $(\tau_k, \bar{r}_k^{(\mathrm{d})})$ belongs to $\mathcal{G}_F$ so that we can consider the CUSUM statistic based on a segment $[\tau_k - \bar{r}_k^{(\mathrm{d})}, \tau_k + \bar{r}_k^{(\mathrm{d})}]$ centered around the change-point τ_k. In contrast, $(\tau_k, \bar{r}_k^{(\mathrm{d})})$ does not necessarily belong to the dyadic grid $\mathcal{G}_D$ and we needed to consider its best approximation $(\bar{\tau}_k^{(\mathrm{d})}, \bar{r}_k^{(\mathrm{d})})$. The segment $[\bar{\tau}_k^{(\mathrm{d})} - \bar{r}_k^{(\mathrm{d})}, \bar{\tau}_k^{(\mathrm{d})} + \bar{r}_k^{(\mathrm{d})}]$ is therefore not centered on τ_k and the corresponding statistic $\Psi^{(\mathrm{d})}_{\bar{\tau}_k^{(\mathrm{d})}, \bar{r}_k^{(\mathrm{d})}}$ is in expectation smaller than $\Psi^{(\mathrm{d})}_{\tau_k, \bar{r}_k^{(\mathrm{d})}}$. In summary, both the collections of dense tests $\Psi^{(\mathrm{d})}_{l,r}$ on $\mathcal{G}_D$ and $\mathcal{G}_F$ are able to detect change-points whose energy is, up to some multiplicative constants, higher than $L^2[[p\log(\frac{n}{r_k\delta})]^{1/2} + \log(\frac{n}{r_k\delta})]$.

5.4.2 Sparse change-points with sub-Gaussian noise

Unlike in the Gaussian case, we do not know the exact distribution of the noise. As a consequence, the Berk-Jones test and more generally higher-criticism type tests cannot be applied to this setting. This is why we only rely on the partial norm statistic. Recall that $\mathcal{Z} = \{1, 2, 2^2, \dots, 2^{[\log_2(p)]}\}$ stands for a dyadic set of sparsities. For $(l,r) \in \mathcal{G}_F$ and $s \in \mathcal{Z}$, we also recall that the partial CUSUM norm is defined as $\Psi^{(\mathrm{p})}_{l,r,s} = \sum_{i=1}^{s} \big(\mathbf{C}_{l,r,(i)}\big)^2$. Then, for any $(l,r) \in \mathcal{G}_F$, the test $T^{(\mathrm{p})}_{l,r}$ rejects the null when at least one of the partial norms is large

$$x^{(\mathrm{p})}_{r,s} = s + \bar{c}^{(\mathrm{p})}_{\mathrm{thresh}} \frac{L^2}{\sigma^2}\left[s\log\left(\frac{2ep}{s}\right) + \log\left(\frac{n}{r\delta}\right)\right]; \qquad T^{(\mathrm{p})}_{l,r} = \max_{s\in\mathcal{Z}} \mathbf{1}\left\{\Psi^{(\mathrm{p})}_{l,r,s} > x^{(\mathrm{p})}_{r,s}\right\} \ ,$$

where $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$ is a tuning parameter in Proposition 5.4.2 below. The partial norm test alone is not able to detect sparse high-energy change-points in the sense of (5.11) and we need to introduce a stronger condition on the energy. Given $\kappa > 0$, a change-point τ_k is a κ-sparse high-energy change-point in the sub-Gaussian setting if $s_k \le [p\log(\frac{n}{r_k\delta})]^{1/2}$ and

$$r_k\Delta_k^2 \ge \kappa L^2\left[s_k\log\left(\frac{ep}{s_k}\right) + \log\left(\frac{n}{r_k\delta}\right)\right] \ . \tag{5.18}$$

Both Conditions (5.11) and (5.18) are compared at the end of the subsection. For a κ-sparse high-energy change-point τ_k, we define its scale $\bar{r}_k^{(\mathrm{s})}$ by

$$\bar{r}_k^{(\mathrm{s})} = \min\left\{r \in \mathbb{N}^* : \quad 4r\Delta_k^2 \ge \kappa L^2\left[s_k\log\left(\frac{ep}{s_k}\right) + \log\left(\frac{n}{r\delta}\right)\right]\right\} \ . \tag{5.19}$$

For any κ-sparse high-energy change-point, it holds that $4(\bar{r}_k^{(\mathrm{s})} - 1) \le r_k$.

Proposition 5.4.2. *There exists a numerical constant $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}} > 0$ such that the following holds for any $\kappa_{\mathrm{s}} > 32\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$. With probability higher than $1-\delta$, one has (i) $T^{(\mathrm{p})}_{l,r} = 0$ for all $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$ and (ii) $T^{(\mathrm{p})}_{\tau_k, \bar{r}_k^{(\mathrm{s})}} = 1$ for all κ_{s}-sparse high-energy change-point τ_k in the sense of (5.18).*

As for Proposition 5.4.1, the proof relies on a careful analysis of the joint distributions of the statistics $\Psi^{(\mathrm{p})}_{l,r,s}$ to handle the multiplicity of $\mathcal{G}_F$.

5.4.3 Consequences

Let $c_0 > 0$ be some constant that we will discuss later. A change-point τ_k is then said to be a c_0-high-energy change-points –in the sub-Gaussian setting– if

$$r_k\Delta_k^2 \ge c_0L^2\left[\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} \wedge \left(s_k\log\left(\frac{ep}{s_k}\right)\right)\right) + \log\left(\frac{n}{r_k\delta}\right)\right] \ . \tag{5.20}$$

We here re-introduce $\mathcal{K}^* \subset [K]$ as the subset of indices such that τ_k satisfies (5.20).

We gather both tests by considering, for any $(l,r) \in \mathcal{G}_F$, the test $T_{l,r} = T^{(\mathrm{d})}_{l,r} \vee T^{(\mathrm{p})}_{l,r}$ with tuning parameters $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}}$ and $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$ as in Propositions 5.4.1 and 5.4.2. Consider any $c_0 > 32(\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}} \vee \bar{c}^{(\mathrm{p})}_{\mathrm{thresh}})$ and any c_0-high-energy change-point τ_k, which is either a c_0-sparse or a c_0-dense high-energy change-point. Defining

$$\bar{r}_k = \bar{r}_k^{(\mathrm{d})} \wedge \bar{r}_k^{(\mathrm{s})}, \tag{5.21}$$

we straightforwardly derive from Proposition 5.4.1 and Proposition 5.4.2 the following result.

Corollary 5.4.3. *There exists two numerical constants* $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}} > 0$ *and* $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}} > 0$ *such that the following holds. With probability higher than* $1-\delta$*, it holds that (i)* $T_{l,r} = 0$ *for all* $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$ *and (ii)* $T_{\tau_k,\bar{r}_k} = 1$ *for any* c_0*-high-energy change-point* τ_k *in the sense of* (5.20)*.*

Then, it suffices to combine this multiple testing procedure with Algorithm 16 to get the change-point procedure $\widehat{\tau}$. Since, for a high-energy change-point in the sense of (5.20), we have $4(\bar{r}_k - 1) < r_k$, we are in position to apply Theorem 5.2.1.

Corollary 5.4.4. *There exist two numerical constant* $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}} > 0$ *and* $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}} > 0$ *such that the following holds. With probability higher than* $1-\delta$*, the estimator* $\widehat{\tau}$ *satisfies (***NoSp***) and* **detects** c_0*-high-energy change-point* τ_k *(as defined in*(5.20)*), that is*

$$d_{H,1}(\widehat{\tau}, \tau_k) \le \bar{r}_k - 1 \le \frac{r_k}{4} \ ,$$

where $\bar{r}_k$ *is defined in* (5.21)*.*

In the case where all change-points are c_0-high-energy change-points in the sense of (5.20), all of them are detected, and a result similar to Corollary 5.3.4 holds here, replacing $r_k^*/2$ by $\bar{r}_k - 1$. Also, both the Hausdorff distance and the Wasserstein distance, can be bounded as in Equation (5.16) if we replace $r_k^*/2$ by $\bar{r}_k - 1$.

As already stated, we could have obtained a similar result (but with different constants) using the dyadic grid $\mathcal{G}_D$ instead of $\mathcal{G}_F$. To conclude this section, let us compare the conditions (5.20) and (5.8) for high-energy. Define

$$\psi^{(sg)}_{n,r,s} = \sqrt{p\gamma_r} \wedge \left(s \log\left(\frac{ep}{s}\right)\right) + \gamma_r \ ,$$

where we recall that $\gamma_r = \log\left(\frac{n}{r\delta}\right)$. If $\gamma_r \ge p/2$, then $\psi^{(sg)}_{n,r,s} \asymp \gamma_r$. In low dimension, the energy threshold for multivariate change-point detection is the same as in the univariate setting, see [92]. If $\gamma_r \le p/2$, then

$$\psi^{(sg)}_{n,r,s} \asymp \begin{cases} \gamma_r & \text{if } s \le \frac{\gamma_r}{\log(p)-\log(\gamma_r)} \\ s\log\left(e\frac{p}{s}\right) & \text{if } \frac{\gamma_r}{\log(p)-\log(\gamma_r)} < s < \frac{\sqrt{p\gamma_r}}{\log(p)-\log(\gamma_r)} \\ \sqrt{p\gamma_r} & \text{if } s \ge \frac{\sqrt{p\gamma_r}}{\log(p)-\log(\gamma_r)} \end{cases}$$

As a consequence, $\psi^{(sg)}_{n,r,s}$ and $\psi^{(g)}_{n,r,s}$ are of the same order of magnitude for all s when $\gamma_r \ge p/2$. When $\log(n/r\delta) < p$, they are also of the same order of magnitude except when s is close but smaller than $\sqrt{p\gamma_r}$, for which the ratio $\psi^{(sg)}_{n,r,s}/\psi^{(g)}_{n,r,s}$ between these two quantities can be as large as $\log(p) - \log(\gamma_r)$. This gap corresponds to the regime where the test based on the Berk-Jones statistic defined in Equation (5.14), used in the Gaussian case, outperforms the test based on the partial CUSUM norm statistic defined in Equation (5.15).

In the definitions of the tests, the tuning constants $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$ and $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}}$ are left implicit, although one can find suitable values by following the proofs of Propositions 5.4.1 and 5.4.2. In practice, the practitioner can calibrate them by a Monte-Carlo method by simulating a Gaussian multivariate times series without any change-points. Then, $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$ and $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}}$ are chosen so that the Family-wise error rate (FWER) of the two collections $(T^{(\mathrm{d})}_{l,r})$ and $T^{(\mathrm{p})}_{l,r}$ is equal to δ.

Computational Cost. The computational cost of the statistic $T_{l,r} = T^{(\mathrm{d})}_{l,r} \vee T^{(\mathrm{p})}_{l,r}$ is $O(p(r + \log(p)))$. Thus, a naive computation of all the tests $T_{l,r}$ for (l,r) in the complete grid $\mathcal{G}_F$ requires $O(p\log(p)\sum_{(l,r)\in\mathcal{G}_F} r) = O(pn(n^2 + \log(p)))$ operations. Nevertheless, using the fact that $\sum_{i=l+1}^{l+r} Y_i = (\sum_{i=l}^{l+r-1} Y_i) + Y_{l+r} - Y_l$, it is possible to compute all the tests at scale r with cost $O(np\log(p))$. Since there are n possible scales r on the complete grid, the whole procedure cost is $O(n^2 p\log(p))$. Using a grid $\mathcal{G} = \{(l,r) \in \mathcal{G}_F \ : \ r \in \mathcal{R}\}$ that contains dyadic scales and all possible locations l for each scale, the whole change-point detection would then require only $O(np\log(n)\log(p))$ computations, since there are only $\log(n)$ possible scales r for such grids.

5.5 Minimax lower bound

In this section, we write for any $\Theta \in \mathbb{R}^{p\times n}$, the distribution of the time series $Y = (y_1, \dots, y_n)$ in the model (5.2) with Gaussian noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. In Section 5.3, we have established that any change-point satisfying the condition (5.8), that is

$$r_k \Delta_k^2 \ge c_0 \sigma^2 \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right)\right] \ ,$$

is detected by our change-point procedure. We now show that this energy condition is unimprovable from a minimax point of view. More precisely, let us define, for any $u > 0$, the class $\bar{\mathcal{P}}(u)$ of mean parameters Θ with arbitrary $K \geq 0$ number of change points and such that any change-point τ_k for $1 \leq k \leq K$ satisfies

$$r_k \Delta_k^2 \geq \frac{1}{2}\sigma^2 \left[s_k \log\left(1 + u\frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k}\right)}\right) + u \log\left(\frac{n}{r_k}\right)\right] \ . \tag{5.22}$$

For u small enough, it turns out no change-point estimator is able to detect all change-points without estimating any spurious change-point with high probability on the full class $\bar{\mathcal{P}}(u)$. Still, using this large class provides somewhat pessimistic bounds. For instance, the most challenging distributions in $\bar{\mathcal{P}}(u)$ for the purpose of change-point detection satisfy $s_k = p$ and $r_k = 1$ (very close change-points). As a consequence, relying on the full collection $\bar{\mathcal{P}}(u)$ turns too pessimistic. To establish that our bounds are adaptive with respect to the sparsity s_k and the length r_k, we define, for any positive integers $1 \leq r \leq \lfloor n/2 \rfloor$ and any $1 \leq s \leq p$ the collection

$$\bar{\mathcal{P}}(u, r, s) = \{\Theta \in \bar{\mathcal{P}}(u) \ : \ \min_k r_k \geq r \text{ and } \max_k s_k \leq s\} \ .$$

By convention, constant means Θ with no change-points ($K = 0$) also belong to $\bar{\mathcal{P}}(u, r, s)$. In the class $\bar{\mathcal{P}}(u, r, s)$, all change-points have a sparsity at most s and a length at least r. Hence, $\bar{\mathcal{P}}(u, r, s)$ becomes larger when s increases or when r increases.

Theorem 5.5.1. *Fix any $u \in (0, 1/8)$. For any $\sigma > 0$, $n \geq 2$, $p \geq 1$, any length $1 \leq r \leq n/4$, and any sparsity $1 \leq s \leq p$, we have*

$$\inf_{\hat{\tau}} \sup_{\Theta \in \bar{\mathcal{P}}(u,r,s)} \mathbb{P}_\Theta(\hat{K} \neq K) \geq \frac{1}{4} \ ,$$

where the infimum is taken over all estimators $\hat{\tau}$ of the change-point vector τ and and $\hat{K} = |\hat{\tau}|$.

Thus, in the Gaussian setting, if all the change-points have a high-energy in the sense of (5.8) but with a smaller multiplicative constant factor, no change-point estimator can consistently estimate the true number of change-points. The next corollary restates this negative results in the same lines as Corollary 5.3.4.

Corollary 5.5.2. *Fix any $u \in (0, 1/8)$. For any $\sigma > 0$, $n \geq 2$, $p > 1$, any length $1 \leq r \leq n/4$, any sparsity $1 \leq s \leq p$, and any estimator $\hat{\tau}$, there exists some $\Theta \in \bar{\mathcal{P}}(u, r, s)$ such that with $\mathbb{P}_\Theta$-probability larger than $1/4$, at least one of the two following properties is satisfied*

- *$\hat{\tau}$ contains at least one* **spurious** *change-point*
- *at least a change-point τ_k with $1 \leq k \leq K$ is not* **detected***, i.e. there is no change-point estimated in the interval $[(\tau_{k-1} + \tau_k)/2, (\tau_k + \tau_{k+1})/2]$.*

This corollary is to be compared to Corollary 5.3.4 - indeed, the energy condition in Equation (5.22) differs from Equation (5.8) only by a numerical multiplicative constant. As a consequence, the energy condition (5.22) is minimal for detection by a change-point estimator that achieves (**NoSp**).

5.6 Application to other change-point problems

In this section, we apply the general methodology of Section 5.2 to two other problems, namely detection of covariance and nonparametric change-points. This allows us to obtain the first tight minimax detection conditions for these problems.

5.6.1 Covariance change-point detection

Following Wang et al. [96], we consider the covariance change-point model where the covariance matrices Σ_t of the centered random vectors $y_t \in \mathbb{R}^p$ are piece-wise constant. Then, the goal is to estimate the times $0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = n + 1$ such that Σ_t is varying. See [96] for motivations. As in that work, we assume that the random vectors y_t are independent and are sub-Gaussian with a uniformly bounded Orlicz norm, that is $\max_{t=1,\ldots,n} \|y_t\|_{\psi_2} \leq B$ for some known fixed B. The Orlicz norm of a random vector y is the supremum of the Orlicz norm of any uni-dimensional projection of y – see e.g. [91]. If the y_t's follow a normal distribution, this amounts to assuming that $\max_{t=1,\ldots,n} \|\Sigma_t\|_{op} \leq 2B^2$ where $\|.\|_{op}$ is for the operator norm. The purpose of Wang et al. was to detect small changes in operator norm, that is detecting instants τ_k such that $\Sigma_{\tau_k} \neq \Sigma_{\tau_k - 1}$ with $\|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}$ possibly small. Apart from the operator norm, other norms have also been considered e.g. in [29]. Here, we focus on the operator norm as in [96].

Recalling the generic procedure introduced in Section 5.2, we consider the dyadic grid $\mathcal{G}_D$ and some $\delta \in (0,1)$. For any $(l,r) \in \mathcal{G}$, we respectively write $\widehat{\Sigma}_{l,-r}$ and $\widehat{\Sigma}_{l,r}$ for the empirical covariance matrices

$$\widehat{\Sigma}_{l,-r} = r^{-1} \sum_{t=l-r}^{l-1} y_t y_t^T \; ; \qquad \widehat{\Sigma}_{l,r} = r^{-1} \sum_{t=l}^{l+r-1} y_t y_t^T \; .$$

Then, we consider the test $T_{l,r}$ rejecting for large values of $\|\widehat{\Sigma}_{l,r} - \widehat{\Sigma}_{l,-r}\|_{op}$.

$$T_{l,r} = \mathbf{1}\left\{ \|\widehat{\Sigma}_{l,r} - \widehat{\Sigma}_{l,-r}\|_{op} \geq c_0 B^2 \left[\sqrt{\frac{p}{r}} + \frac{p}{r} + \sqrt{\frac{\log(\frac{2n}{\delta r})}{r}} + \frac{\log(\frac{2n}{\delta r})}{r} \right] \right\} \; , \tag{5.23}$$

where the numerical tuning constant c_0 is set in the proof of the following proposition. Relying on concentration bounds [50] for the empirical covariance matrix of sub-Gaussian random vectors, we easily prove that the FWER of the multiple testing procedure $(T_{l,r})$ with $(l,r) \in \mathcal{G}_D$ is small. Then, we can analyze the type II error probability and plug it into the generic result (Theorem 5.2.1) to control the behavior of the change-point estimator $\widehat{\tau}$. This leads us to the following result. In the sequel, a change-point τ_k is said to have a high-energy if

$$r_k \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}^2 \geq c_1 B^4 \left[\left(p + \log\left(\frac{2n}{r_k \delta} \right) \right) \wedge r_k \right] \; , \tag{5.24}$$

where the numerical constant c_1 is introduced in the proof of the following proposition. We recall that, by definition of the model, we have $\|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} \leq 4B^2$.

Proposition 5.6.1. *There exist positive numerical constants c_0, c_1, and c_2 such that the following holds for any $B > 0$ and any sequence of independent centered random vectors (y_t) satisfying $\max_t \|y_t\|_{\psi_2} \leq B$. With probability higher than $1-\delta$, the change-point estimator $\widehat{\tau}$ satisfies (***NoSp***) and* **detects** *all high-energy change-points in the sense of (5.24). Besides, any such high-energy change-point τ_k satisfies*

$$d_{H,1}(\widehat{\tau}, \tau_k) \leq c_2 B^4 \frac{p + \log\left(2\delta^{-1} B^{-4} n \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}^2\right)}{\|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}^2} \leq \frac{r_k}{4} \; , \tag{5.25}$$

under the same event of probability than $1-\delta$.

Let us compare our condition (5.24) for detection with Theorem 2 in Wang et al. [96]. The authors assume that all the change-points satisfy

$$\min_k r_k \min_k \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}^2 \geq c_1' B^4 p \log(n) \; .$$

In addition to the fact that we allow some change-points to have an arbitrarily low energy, our requirement for detection scales like $\sqrt{p} + \sqrt{\log(n/r_k)}$ instead of $\sqrt{p \log(n)}$.

The next proposition establishes that the latter condition is minimal. By homogeneity, we can only consider the case where $B = 3/2$. We focus our attention on Gaussian distributions so that the distribution of the sequence $(y_1, \dots, y_n)$ is uniquely defined by the sequence $(\Sigma_1, \dots, \Sigma_n)$ of covariance matrices. Given an integer $1 \leq r \leq n/4$ and $\zeta \in (0, 1/\sqrt{2})$, we define $\bar{\mathcal{P}}(r,\zeta)$ the collection of sequences $\eta = (\Sigma_1, \dots, \Sigma_n)$ of covariance matrices that satisfy either $\Sigma_t = I_p$ or $\|\Sigma_t\|_{op} = 1 + \zeta$. Besides, the corresponding change-points $(\tau_1, \dots, \tau_K)$ of η must satisfy $\min_k r_k \geq r$ and $\min_k \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} \geq \zeta$. For $\eta \in \bar{\mathcal{P}}(r,\zeta)$, we write $\mathbb{P}_\eta$ for the corresponding distribution of $(y_1, \dots, y_n)$.

Proposition 5.6.2. *There exists a positive numerical constant c such that, for any n, p and any length $1 \leq r \leq n/4$ the following holds. Provided that $r\zeta^2 \leq c(p + \log(n/r)) \wedge \frac{r}{2}$, we have*

$$\inf_{\hat{\tau}} \sup_{\eta \in \bar{\mathcal{P}}(r,\zeta)} \mathbb{P}_\eta(\hat{K} \neq K) \geq \frac{1}{4}.$$

As a consequence, our procedure $\widehat{\tau}$ achieves the minimal separation condition (5.24) for change-point detection. In their work, [96] obtain faster localization errors than (5.25) to the price of stronger separation conditions. Our focus in this work is to provide optimal detection conditions and we did not try to optimize (5.24).

5.6.2 Univariate nonparametric change-point detection

We now turn to the univariate nonparametric change-point model considered in [70]. Let $m \geq 1$ be any positive integer. At each time $t = 1, \dots, n$, the random vector y_t is an m-sample of a univariate distribution with

cumulative distribution function F_t. Then, we aim at detecting a vector $\tau = (\tau_1, \dots, \tau_K)$ of change-points such that $F_{\tau_k} \neq F_{\tau_{k-1}}$. As in [70], we quantify the distance between two distributions by the Kolmogorov distance $\|F_1 - F_2\|_\infty = \sup_{z\in\mathbb{R}} |F_1(z) - F_2(z)|$.

As in the previous subsection, we build a procedure $\widehat{\tau}$ with our generic algorithm on the dyadic grid. Regarding the collection of tests $(T_{l,r})$, we consider two-sample Kolmogorov-Smirnov tests. More precisely, we denote $\widehat{F}_t$ the empirical distribution function associated with the sample y_t and we define the test

$$T_{l,r} = \mathbf{1}\left\{ \left\| r^{-1}\left(\sum_{t=l}^{l+r-1} \widehat{F}_t - \sum_{t=l-r}^{l-1} \widehat{F}_t \right) \right\|_\infty \geq \sqrt{2\frac{\log(4n/(\delta r))}{mr}} \right\} .$$

In the following, a change-point τ_k is said to have a high-energy if

$$r_k \|F_{\tau_k} - F_{\tau_{k-1}}\|_\infty^2 \geq \frac{c_1}{m} \log\left(\frac{n}{r_k \delta}\right) , \tag{5.26}$$

where the numerical constant c_1 is introduced in the proof of the next proposition. As in Subsection 5.6.1, it is straightforward to prove, based on Dvoretzky–Kiefer–Wolfowitz inequality, that the FWER of the multiple testing procedures $(T_{l,r})$ with $(l, r) \in \mathcal{G}_D$ is small. Then, we analyze the type II error probability of this test and plug it into the generic result (Theorem 5.2.1) to control the behavior of the change-point estimator $\widehat{\tau}$.

Proposition 5.6.3. *There exist positive numerical constants c_1 and c_2 such that the following holds. With probability higher than $1-\delta$, the change-point estimator $\widehat{\tau}$ satisfies (***NoSp***) and* **detects** *all high-energy change-points τ_k in the sense of (5.26). Besides, any such high-energy change-points τ_k satisfies*

$$d_{H,1}(\widehat{\tau}_{k'}, \tau_k) \leq c_2 \frac{\log\left(\delta^{-1} nm \|F_{\tau_k} - F_{\tau_{k-1}}\|_\infty^2\right)}{m\|F_{\tau_k} - F_{\tau_{k-1}}\|_\infty^2} \leq \frac{r_k}{4} , \tag{5.27}$$

under the same event of probability than $1-\delta$.

In [70], the authors introduce a procedure detecting all the change-points provided that

$$\min_k r_k \min_k \|F_{\tau_k} - F_{\tau_{k-1}}\|_\infty^2 \geq c_1 \frac{\log(n)}{m} .$$

Comparing this last condition with (5.26), we observe that our logarithmic term is tighter and that we allow arbitrarily low-energy change-points.

The next proposition establishes that the condition (5.26) is unimprovable. Given an integer $1 \leq r \leq n/4$ and $\zeta \in (0, 1/4)$, we focus our attention on the collection $\bar{\mathcal{P}}(r, \zeta)$ of sequences $(F_1, \dots, F_n)$ of distributions such that the corresponding change-points $(\tau_1, \dots, \tau_K)$ satisfy $\min_k r_k \geq r$ and $\min_k \|F_{\tau_k} - F_{\tau_{k-1}}\|_\infty \geq \zeta$. For $\eta \in \bar{\mathcal{P}}(r, \zeta)$, we write $\mathbb{P}_\eta$ for the corresponding distribution of the sequence $(y_1, \dots, y_n)$.

Proposition 5.6.4. *There exists a positive numerical constant c such that, for any n, p and any length $1 \leq r \leq n/4$ the following holds. Provided that $r\zeta^2 \leq c' \log(n/r)/m$, we have*

$$\inf_{\hat{\tau}} \sup_{\eta\in\bar{\mathcal{P}}(r,\zeta)} \mathbb{P}_\eta(\hat{K} \neq K) \geq \frac{1}{4}.$$

5.7 Discussion

5.7.1 Noise distribution for multivariate change-point detection

Comparison between Gaussian and sub-Gaussian rates. In this work, we have studied two types of noise distribution: Gaussian (Section 5.3) and general sub-Gaussian distributions (Section 5.4) without further knowledge on the distribution functions. Since the Gaussian setting is a specific instance of the sub-Gaussian setting, it is clear that the minimax lower bounds from Section 5.5 apply in both settings. As described in the previous subsection, the performances in the sub-Gaussian case almost match those in the Gaussian setting except for s_k slightly lower but close to $\sqrt{p\log(en/r_k)}$. Indeed, in that regime, Berk-Jones or Higher-Criticism type statistics heavily rely on the probability distribution function of the noise, which is not available in the general sub-Gaussian case. Still, we could slightly improve the sub-Gaussian rates if we further assume that the noise components are identically distributed with common CDF F.

- If F is known (know noise distribution), then one may adapt Berk-Jones test by replacing $\bar{\Phi}(x)$ in Equation (5.14) by $F(-x) + (1 - F(x))$. This would allow us to recover the exact same detection condition as in the Gaussian setting.
- If F is unknown and if there are not too many change-points, one could hope to estimate the quantiles of the CUSUM statistic at each scale r and plug them into a Berk-Jones statistics. This goes however beyond the scope of this chapter.

Unknown variance or more general variance matrix. We assumed in the sparse multivariate sections that the variance σ^2 is known. Whereas the partial norm test only requires the knowledge of an upper bound on σ, the dense statistic $\Psi^{(\mathrm{d})}_{l,r}$ requires the exact knowledge of the variance. As soon as there are not too many change-points, it is possible to roughly estimate σ and therefore accommodate the partial norm test with an unknown variance. In contrast, the dense statistics needs to be replaced by a U-statistics. Consider any even positive integer r and define

$$\widetilde{\mathbf{C}}_{l,r}(Y) = \frac{\sqrt{r}}{2}\left(\frac{2}{r}\sum_{t=1}^{r/2} Y_{l-2(t-1)-1} - \frac{2}{r}\sum_{t=1}^{r/2} Y_{l+2(t-1)}\right), \quad \widetilde{\mathbf{C}}'_{l,r}(Y) = \frac{\sqrt{r}}{2}\left(\frac{2}{r}\sum_{t=1}^{r/2} Y_{l-2t} - \frac{2}{r}\sum_{t=1}^{r/2} Y_{l+2(t-1)+1}\right),$$

where $\widetilde{\mathbf{C}}_{l,r}(Y)$ and $\widetilde{\mathbf{C}}'_{l,r}(Y)$ are independent. If there is one change-point at position l and no other change-points in $(l-r, l+r)$, then these statistics are identically distributed and we consider $\widetilde{\Psi}^{''(\mathrm{d})}_{l,r} = \langle \widetilde{\mathbf{C}}_{l,r}(Y), \widetilde{\mathbf{C}}'_{l,r}(Y)\rangle$ whose expectation is null when there are no change-points in the segment. As a consequence, $\widetilde{\Psi}^{''(\mathrm{d})}_{l,r}$ does not require the knowledge of σ; only an upper bound of σ is required to calibrate the corresponding test. Such a U-statistics has already been introduced in [102] and analyzed in an asymptotic setting. Unfortunately, since we can only consider even r, this precludes us to detecting change-points that are very close together with $r_k = 1$.

In the general case where there is spatial covariance in the noise, that is $\mathrm{var}(\epsilon_t) = \Sigma$ for an unknown but general Σ, we can still use the same U-statistic described in the previous paragraph for the dense case. For the sparse case, one could use the supremum norm of the CUSUM statistics as in Jirak [48] and Yu and Chen [105]. To calibrate those tests, we need to estimate both the Frobenius and the operator norm of Σ, which seems to be doable as soon as there are not too many change-points. If the spatial covariance matrix $\mathrm{var}(\epsilon_t)$ is unknown and even allowed to change with time, we suspect that the problem becomes intrinsically more involved.

5.7.2 Optimal localization rates

In this work, we mainly considered the problem of **detecting** change-points in the mean of a random vector. We provided tight conditions on the energy so that a change-point is detectable. When such a change-point τ_k is detected, Corollary 5.3.3 states that its position is estimated up to an error of r^*_k, which is also of the order of $\sigma^2 \Psi^{(g)}_{n,\overline{\tau}_k,s_k} \Delta_k^{-2}$– see the definition (5.9). It is not clear whether this error is optimal or not.

In the univariate setting ($p = 1$), [92] has established that, above the detection threshold, a specific change-point position τ_k can be localized at the rate $\sigma^2\Delta_k^{-2}$. In the multivariate setting, the situation is more tricky and there are certainly several localization regimes beyond the detection threshold. It is an interesting direction of research to pinpoint the exact localization rate between $\sigma^2\Delta_k^{-2}$ and $\sigma^2\Psi^{(g)}_{n,\overline{\tau}_k,s_k}\Delta_k^{-2}$. We leave this for future work.

5.7.3 On the choice of the grid in the generic algorithm

Our general procedure is defined for almost any arbitrary grid. Optimal procedures with the dyadic grid are introduced in Sections 5.3 and 5.6, whereas we use a near-optimal procedure on the complete grid in Section 5.4.

From a computational perspective, the procedure's worst-case complexity is proportional to the size $|\mathcal{G}|$ of the grid $\mathcal{G}$. In that respect, the dyadic grid and more generally the a-adic grids benefit from a linear size whereas the size of the complete grid is quadratic.

From a mathematical perspective, it is much easier to control the behaviour of the procedure for an a-adic grid by a simple Bonferroni correction on all the statistics as it turns out that this correction is sufficient for our purpose – see the proofs of Section 5.3. In constrast, controlling larger collections of tests turns out to be much more challenging as one needs to carefully take into account the dependences between the test statistics, which becomes all the more challenging for complex models. As an example, we introduced in Section 5.3 Berk-Jones statistics to achieve the tight minimax condition for change-point detection. Unfortunately, we did not manage to apply a suitable chaining argument to these statistics and were therefore unable to control the behavior of the corresponding change-point detection procedure on the complete grid.

From a purely statistical perspective, it is difficult to appreciate the respective benefits of denser or sparser grids. On the one hand, for denser grids, the approximation $\overline{\tau}_k$ of τ_k at scale r will be closer to τ_k so that the corresponding test $T_{\overline{\tau}_k,r}$ may be more powerful. On the other hand, for a denser grid, the tests possibly suffer from a higher price for multiplicity. This price can be mild if one takes into account the dependences between the tests. Still, except perhaps in the univariate Gaussian change-point model for which delicate controls of the CUSUM process exist, it is challenging to provide theoretical guidance towards the best choice of the grid.

5.7.4 Optimality of the generic algorithm in a broader context.

Algorithm 16 aggregates homogeneity tests and provides theoretical guarantees on the event $\mathcal{A}\left(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*}\right)$ - i.e. the event where the outcomes of the tests are consistent - as stated in Theorem 5.2.1. In the possibly sparse high-dimensional mean change-point model, we introduced a suitable multiple testing procedure which, when combined with Algorithm 16, leads to a minimax optimal change-point detection procedure.

We described in Section 5.2 how to adapt this approach to other change-point problems and this was already illustrated in Section 5.6 with covariance and nonparametric problems. One may then wonder whether this roadmap still leads to minimax optimal procedures for general problems. Consider the general setting from Section 5.1 where we are interested in detecting change-points in $(\Gamma(\mathbb{P}_t))_{t\in[n]}$. Upon endowing the space $\mathcal{V}$ with some distance d, we define, for any k,

$$\bar{\Delta}_k = d\left(\Gamma\left(\mathbb{P}_{\tau_k}\right), \Gamma\left(\mathbb{P}_{\tau_{k-1}}\right)\right) \quad ,$$

which corresponds to the change-point height. Then, one may wonder how large $\bar{\Delta}_k$ has to be - as a function of r_k - so that a change-point detection procedure achieving the no-spurious property (**NoSp**) with high probability is able to detect τ_k. In this discussion, we restrict our attention to independent observations, that is the random variables y_t are assumed to be independent and we consider the dyadic grid $\mathcal{G}_D$.

Fix $\delta \in (0,1)$. At each scale $r \in \{1, 2, \ldots, 2^{\lfloor \log_2(n)\rfloor - 1}\}$ and for each $l \in \mathcal{D}_r$, with $\mathcal{D}_r$ defined in (5.5), we consider the testing problem $H_{0,l,r} : \{\mathbb{P} :\ \Gamma(\mathbb{P}_{l-r}) = \ldots = \Gamma(\mathbb{P}_{l+r-1})\}$ versus

$$H_{\rho,l,r} : \left\{ \mathbb{P} : \begin{array}{c} \Gamma(\mathbb{P}_{l-r}) = \ldots = \Gamma(\mathbb{P}_{l-m-1}) \\ \Gamma(\mathbb{P}_{l-m}) = \ldots = \Gamma(\mathbb{P}_{l+r-1}) \\ d(\Gamma(\mathbb{P}_{l-m-1}), \Gamma(\mathbb{P}_{l-m}) \geq \rho) \end{array} \quad \text{for some integer } m \in [-r/2, r/2] \right\}$$

This amounts to testing whether there is a single change-point near l of height at least ρ in the segment $(l-r, l+r)$. Given $\delta \in (0,1)$ and a test T we define the δ-separation distance of T by

$$\rho^*_{l,r}(T, \delta) = \inf\left\{ \rho : \sup_{\mathbb{P}\in H_{0,l,r}} \mathbb{P}(T = 1) \vee \sup_{\mathbb{P}\in H_{\rho,l,r}} \mathbb{P}(T = 0) \leq \delta \right\} .$$

This corresponds to the minimal change-point height that is detected by the test T. Then, the *minimax* separation distance $\rho^*_{l,r}(\delta)$ is simply $\inf_T \rho_{l,r}(T, \delta)$, i.e. the infimum over all tests T of the separation distance. By translation invariance of the testing problem, note that $\rho^*_{l,r}(\delta)$ does not depend on l and is henceforth denoted $\rho^*_r(\delta)$.

For any (l, r), take any test $T_{l,r}$ (nearly)[3] achieving the minimax separation distance $\rho^*_r(\delta|\mathcal{D}_r|^{-1}\beta_r)$ with $\beta_r = 6\log_2^{-2}(n/r))\pi^{-2}$. Then, it follows from a simple union bound on the dyadic grid that, with probability higher than $1-\delta$, the collection of tests $T_{l,r}$, where (l, r) belongs to the dyadic grid, does not detect any false positive and detects any change-point τ_k such that $\overline{\Delta}_k$ is higher than $\rho^*_{\tilde{r}_k}(\delta|\mathcal{D}_{\tilde{r}_k}|^{-1}\beta_{\tilde{r}_k})$, where $\tilde{r}_k$ is the largest scale in $\mathcal{R}$ such that $4(\tilde{r}_k - 1) \leq r_k$. As a consequence of Theorem 5.2.1, the corresponding detection procedure achieves, with probability higher than $1-\delta$, the property (**NoSp**) and **detects** any change-point satisfying the energy condition $\overline{\Delta}_k \geq \rho^*_{\tilde{r}_k}(r\delta\beta_r/2n)$.

Conversely, we believe that this energy condition is almost tight. Indeed, fix any even range $r \geq 2$. To simplify the discussion suppose that $n/(2r)$ is an integer. We consider a specific instance of the problem where the statistician knows that there are $n/(2r)-1$ evenly-spaced change-points respectively at $2r+1, 4r+1, \ldots, n-2r+1$ that allow to reduce the change-point detection problem to $n/(2r)$ change-point detection problem in intervals $(l-r, l+r]$ for $l = r+1, 3r+1, 5r+1, \ldots$. Furthermore, it is known that, in each such segment, there exists at most one change-point that is situated in $[l-0.5r, l+0.5r]$, and if the change-point is present then its height is at least $\rho = \rho^*_r(\delta) - \zeta$ for ζ arbitrarily small. Since all $n/(2r)-1$ evenly-spaced change-points $2r+1, 4r+1, \ldots, n-2r+1$ are known to the statistician, detecting all remaining change-points is equivalent to building an $n/(2r)$ multiple test of the hypotheses $H_{0,l,r}$ versus $H_{\rho,l,r}$ for $l = r+1, 3r+1, 5r+1, \ldots$. If a change-point procedure achieves (**NoSp**) and **detects** all change-points with radius at least $r/2$ and height at least ρ with probability at least $1-\delta$, then one is able, with probability uniformly higher than $1-\delta$, to simultaneously perform without error $n/(2r)$ independent tests $H_{0,l,r}$ versus $H_{\rho,l,r}$. Since any single test must endure an error with probability at least δ in the worst case, no collection of independents tests is able to endure less than $1-(1-\delta)^{n/(2r)}$. When n/r is large and $\delta < 2r/n$, the latter is of the order of $\delta 2r/n$. Based on this, we conjecture that no change-point

[3] Since the minimax separation distance is defined as an infimum, it is not necessarily achieved by a test. Still, we can build a test whose separation distance is arbitrarily close to the optimal one. We neglect the additive error term for the purpose of the discussion.

procedure is able to achieve, with probability higher than $1-\delta$ the property (**NoSp**), and also to **detect** all change-points with radius at least $r/2$ and height at least $\rho_r^*(2r\delta/n) - \zeta$ for $\zeta > 0$ arbitrarily small.

Comparing the performances of our procedure with the negative arguments that we just outlined, we see that aggregating optimal tests on a dyadic grid allows to detect change-points with (almost) uniform height higher $\rho^*_{\tilde{r}_k}(r_k\delta\beta_{r_k}/(2n))$ whereas, as explained above, we conjecture that a change-point τ_k can be detected only if $\Delta_k \geq \rho^*_{r_k}(2r_k\delta/n)$. Since $\tilde{r}_k \geq (r_k/8) \vee 1$- as we considered the dyadic grid when constructing $\tilde{r}_k$ - the difference between these two bounds is mostly due to the term β_r which is of the order of $\log^2(n/r)$. Whereas it is possible to detect change-points at a given scale with a test of type I error probability $2r\delta/n$, our multi-scale procedure relies on a collection of single tests with type I error probability of the order of $r\delta/n/\log^2(n/r)$. This mild mismatch - that we introduce to deal with the multiplicity of scales - of order $\log^2(n/r)$ is harmless for the Gaussian mean-detection problem. Indeed, one may deduce from our analysis in Section 5.3 that $\rho^*_{r_k}(2r_k\delta/n)$ is of the same order as $\rho^*_{\tilde{r}_k}(\delta|\mathcal{D}_{\tilde{r}_k}|^{-1}\beta_{\tilde{r}_k})$.

In conclusion, one can build through Algorithm 16 an almost optimal change-point procedure in any model provided that we are given optimal homogeneity tests of the form $H_{0,l,r}$ versus $H_{\rho,l,r}$. This provides a universal reduction of the problem of change-point detection to the problem of homogeneity testing.

5.8 Numerical experiments

In this section, we illustrate the behavior of our procedure to detect change-points in a sparse high-dimensional setting (5.2).

Performance Measure. To assess the quality of change-point estimator $\hat{\tau}$, we first measure whether the estimated number of change-points $\widehat{K} = |\hat{\tau}|$ is equal to the true number K of change-points. We also define the **SAND** loss as the proportion of **S**purious estimated change-points **A**nd true change-points that are **N**ot **D**etected:

$$\mathbf{SAND}((\tau_k),(\hat{\tau}_{k'})) = \frac{1}{K}\sum_{k=1}^{K}\left|\left|[(\tau_k+\tau_{k-1})/2,(\tau_k+\tau_{k+1})/2]\cap\{\hat{\tau}_k, k\in[\hat{K}]\}\right| - 1\right| \ .$$

Change-point Detection Methods. In the experiments, we implemented the bottom-up aggregation procedure Algorithm 16 with partial norm tests $T^{(\mathrm{p})}$ and dense test $T^{(\mathrm{d})}$ corresponding to Section 5.4 on a semi-complete grid $\mathcal{G}_F = \{(l,r) \ : \ l \in \{r+1,\dots,n-r+1,\ r\in\mathcal{R}\}\}$ - we take scales r in the dyadic set for computational purposes. On a location l and a scale r, each test statistic can be seen as a partial norm test relying on the statistic $\Psi^{(\mathrm{p})}_{l,r,s}$ defined in Section 5.4.2 and a threshold $\mathrm{Thresh}(r,s)$ which is either equal to $x_r^{(\mathrm{d})}$ when $s = d$ - see Section 5.4.1 - or to $x^{(\mathrm{p})}_{r,s}$ when $s \in \mathcal{Z}_r := \{1,2,4,\dots,2^{\lfloor\log_2(s_{\max})\rfloor}\}$ with $s_{\max} := \frac{\sqrt{p\gamma_r}}{\log(p)-\log(\gamma_r)}$ - see Section 5.4.3 for the definition of the boundary between sparse and dense regimes $s_{\max}$. We actually do not use the definition of $x_r^{(\mathrm{d})}$ and $x^{(\mathrm{p})}_{r,s}$ for our thresholds $\mathrm{Thresh}(r,s)$ since they rely on constants that are not necessarily tight, but we rather calibrate them by a Monte-Carlo method using 10.000 independent samples. For each sample consisting in a time series made of n gaussian normal centered vector in $\mathbb{R}^p$, and for each $r\in\mathcal{R}, s\in\mathcal{Z}_r\cup\{p\}$, we compute the maximum over all l of the statistics $\Psi^{(\mathrm{p})}_{l,r,s}$. Considering the list of all the 10.000 maximums and taking $\delta = 5\%$, $\mathrm{Thresh}(r,s)$ is then defined as the $(1-\delta/(2|\mathcal{R}||\mathcal{Z}_r|))$-quantile if $s\in\mathcal{Z}_r$ and as the $(1-\delta/(2|\mathcal{R}|))$-quantile if $s = p$, so that, by a union bound, the total probability of finding a false positive is less than δ. Note that this calibration step only depends on n, p, and σ and only needs to be performed once and for all.

We compare our procedure with the inspect method of [103] which is available as an R package. The tuning parameters of inspect are computed with the automatic method defined in the same R package.

In all the following experiments, we fix the dimension $p = 100$ and the sample size $n = 200$. We generate a piecewise constant signal $(\eta_t)_{t=1}^n$ in $\mathbb{R}^p$ with possible change-points $(\tau_1,\dots,\tau_K)$ using one of the three following settings. We then add a scaling factor $\alpha > 0$ and apply our procedure to the data $y_t = \alpha\eta_t + \varepsilon_t$, which amounts to setting $\theta_t = \alpha\eta_t$ in model (5.2). We fix the variance of all the coordinates of ε_t to be equal to one. Increasing α on a grid with step 0.1 allows us to experimentally identify a transition between the regime where we do not detect precisely the change-points - in which case the two losses tend to be close to one - and the regime where we do detect the change-points - in which cases the losses are smaller. We consider three simulation settings:

1. **Segment**. We generate a signal η which is zero everywhere, except on $[80,100]$ where we set it equal to a random vector Δ with $\|\Delta\| = 1$ and $\|\Delta\|_0 = s$, for $s = 1, 20, 100$. In each one of these cases, we choose the location of the s non null coordinates of Δ uniformly at random and their value uniformly at random in the set $\{-1/\sqrt{s}, 1/\sqrt{s}\}$. Each time, η has 2 true change-points, and we generate the noise (ϵ_t) as independent centered and normalized gaussian vectors.

2. **Multiple Change points**. We generate 10 uniform random locations $\tau_1 < \tau_2 < \ldots < \tau_{10}$ on $[1, 200]$. For each location τ_i, we generate a uniform random integer $s_i \in [1, 100]$ and a vector Δ_i as in the segment setting with $\|\Delta_i\| = 1$ and $\|\Delta_i\|_0 = s_i$. We generate a uniform random real number $N_i \in [1, 5]$ and define the time series η_i by $(\eta_i)_t = N_i \Delta_i \mathbf{1}_{t \geq \tau_i}$. Finally, the signal $\eta = \sum_{i=1}^{10} \eta_i$ has exactly 10 change-points with random locations. As previously, the noise components (ε_t) follow independent centered and standard gaussian vectors.

3. **Time-dependencies**. We use the same signal as in the segment setting with $s = 20$ but we move away from our assumptions by considering time dependencies. More precisely, the (ε_t)'s are now defined according to an AR process such taht $\varepsilon_{t+1} = \rho\varepsilon_t + \sqrt{1-\rho^2}\varepsilon'_{t+1}$ for $t \geq 0$ where (ε'_t) are independent centered and normalized gaussian vectors, $\rho = 0.05$ for the simulation and by convention $\varepsilon_0 \sim \mathcal{N}(0, I_p)$.

Risk estimation with Monte-Carlo. In each setting, we generate 500 independent samples and compute the twpo losses $\mathbf{SAND}((\tau_k), (\hat{\tau}_{k'}))$ and $\mathbf{1}\{\hat{K} \neq K\}$. We estimate the risks $\mathbb{E}[\mathbf{SAND}((\tau_k), (\hat{\tau}_{k'}))]$ and $\mathbb{P}(K \neq \hat{K})$ by averaging the loss over the 500 trials. We also compute 95% confidence intervals.

Results. In the segment setting - see Figure 5.4, 5.5, 5.6, the risks tend to decrease as α increases since the higher α, the higher the energy of the generated change-points are. As s increases, we can see that both methods need a higher scaling factor to achieve the same risk, which translates the fact that the higher s, the more energy is needed to detect a change-point with vector Δ of sparsity s. In the segment settings, our bottom-up procedure tends to achieve significantly smaller loss than the inspect method on average. It is not the case in the multiple change-points setting - see Figure 5.7 - where the inspect method tends to perform slightly better. In the setting with time-dependencies - see Figure 5.8 - the risks are worse than the corresponding setting without time-dependencies - see Figure 5.5 - mainly because adding time-dependencies tends to create more spurious change-points (i.e. false positives).

Computation time Our code is implemented with python 3.9 and it mainly uses the convolution function conv1d from pytorch 1.12.1 to compute the Cusum statistics. Simulations are run on CPU (Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz) with 32Go of memory. Running our method on pure noise - i.e. $\theta_t = 0$ for all t - takes 101 ± 2 ms while the inspect method takes only 18 ± 2 ms to run on average, but optimizing our code is out of the scope of this chapter. All the experiments are described in the repository https://github.com/epilliat/multicpdetec.

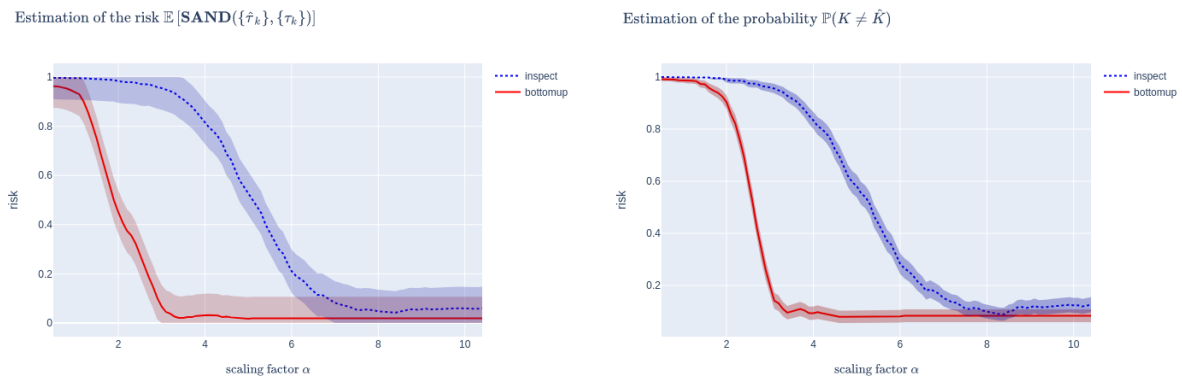


Figure 5.4: Estimation of $\mathbb{E}[\mathbf{SAND}((\tau_k), (\hat{\tau}_{k'}))]$ and $\mathbb{P}(\hat{K} \neq K)$ in the segment setting with $s = 1$.

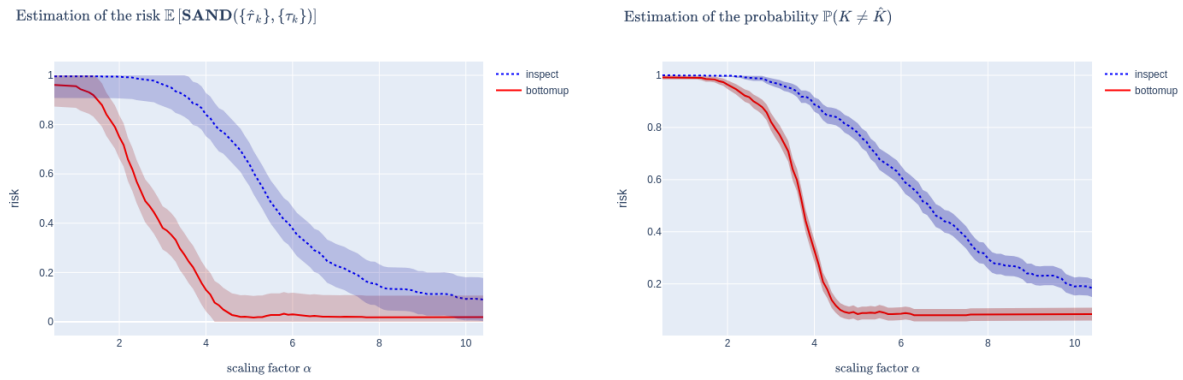


Figure 5.5: Estimation of $\mathbb{E}[\mathbf{SAND}((\tau_k),(\hat{\tau}_{k'}))]$ and $\mathbb{P}(\hat{K} \neq K)$ in the segment setting with $s = 20$.

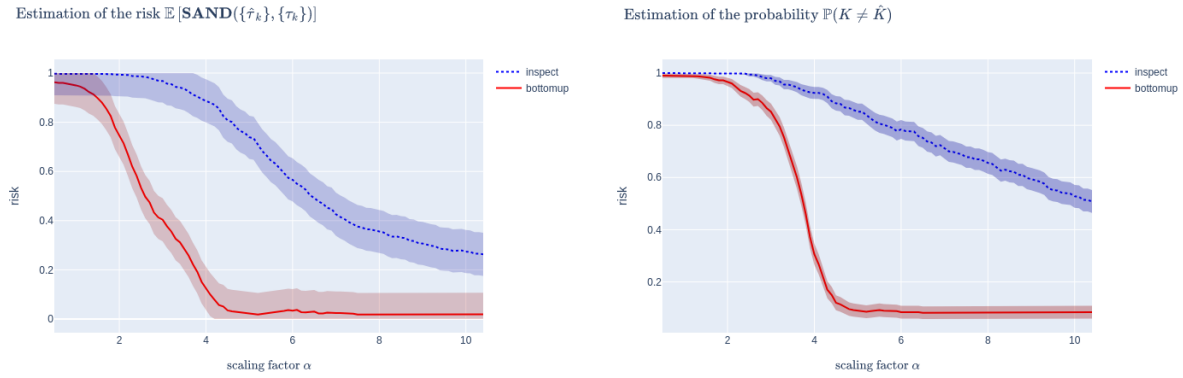


Figure 5.6: Estimation of $\mathbb{E}[\mathbf{SAND}((\tau_k),(\hat{\tau}_{k'}))]$ and $\mathbb{P}(\hat{K} \neq K)$ in the segment setting with $s = 100$.

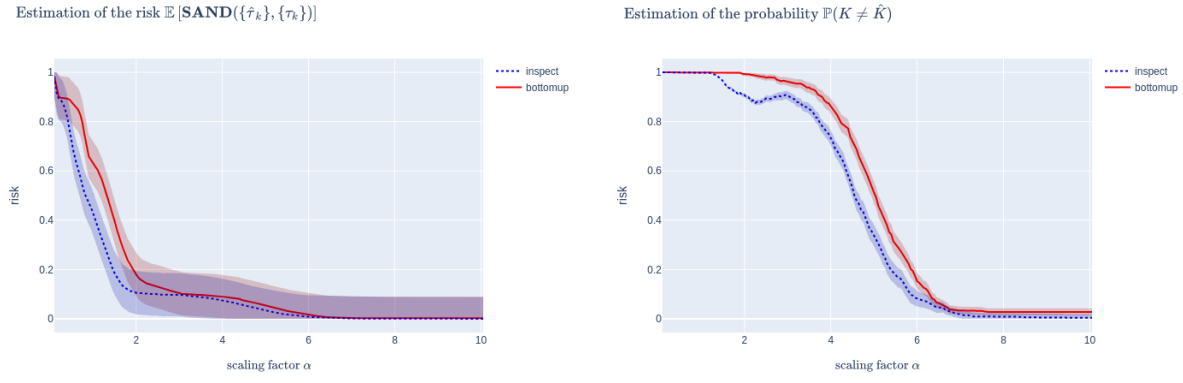


Figure 5.7: Estimation of $\mathbb{E}[\mathbf{SAND}((\tau_k),(\hat{\tau}_{k'}))]$ and $\mathbb{P}(\hat{K} \neq K)$ in a multiple change-point setting with $K = 10$ where change-points have random norms in $[1, 5]$ and random sparsities in $[1, p]$.

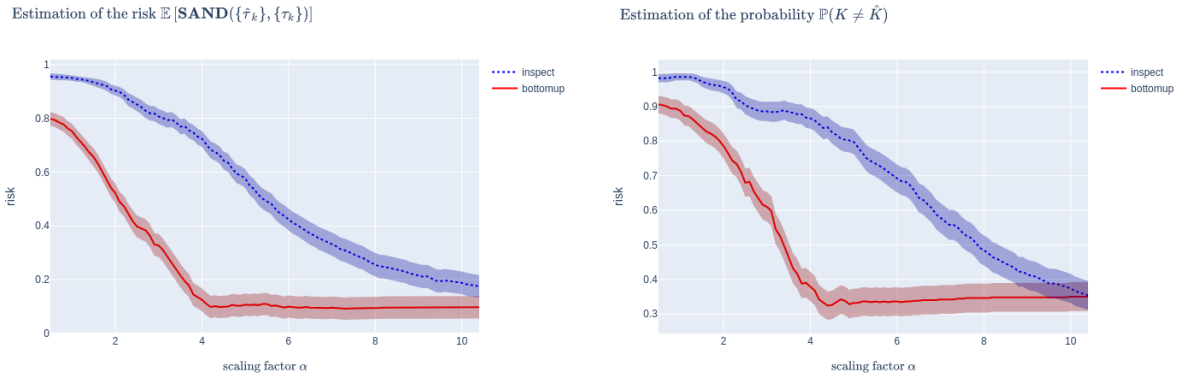


Figure 5.8: Estimation of $\mathbb{E}[\mathbf{SAND}((\tau_k),(\hat{\tau}_{k'}))]$ and $\mathbb{P}(\hat{K} \neq K)$ in the segment setting with $s = 20$ but with time-dependent noise that have an auto-correlation of $\rho = 5\%$.

5.9 An alternative algorithm

In Algorithm 17 below, we also introduce a variant of the procedure, where instead of merging relevant interesting intervals at the same scale, we only keep one of them. More precisely, we choose the convention of discarding the interval $[l-r+1, l+r-1]$ if there exists $l' < l$ such that $T_{l',r} = 1$ and $[l-r+1, l+r-1] \cap [l'-r+1, l'+r-1] \neq \varnothing$. Alternatively, we could have chosen to discard one of the intervals at random.

Algorithm 17 Variant bottom-up aggregation procedure of multiscale tests

Require: Observations $y_t, t = 1 \dots n$ and local test statistic $(T_{l,r})_{(l,r)\in\mathcal{G}}$
Ensure: Estimated change-points $(\hat{\tau}_k)_{k\leq\hat{K}}$

$\mathcal{S} = \varnothing$ $\mathcal{T} = \varnothing$
for Increasing $r \in \mathcal{R}$ **do**
  for $l \in \mathcal{D}_r$ s.t. $T_{l,r} = 1$ **do**
    if $[l-r+1, l+r-1] \cap \mathcal{S} = \varnothing$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup [l-r+1, l+r-1]$
      $\mathcal{T} \leftarrow \mathcal{T} \cup \{l\}$
    end if
  end for
end for
return $\mathcal{T}$

5.10 Proofs

5.10.1 Proof of Theorem 5.2.1

Let $\Theta \in \mathbb{R}^{n\times p}$, T be a local test statistic, $\mathcal{K}^*$ be a set of indices of significant change-points and $(\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*}$ be elements of the grid $\mathcal{G}$ that satisfy (5.6). We assume that $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ holds, that is:

1. **(No False Positive)** $T_{l,r} = 0$ for all $(l,r) \in \mathcal{H}_0 \cap \mathcal{G}$, where $\mathcal{H}_0$ is defined by (5.7)
2. **(Significant change-point detection)** for every $k \in \mathcal{K}^*$, we have $T_{\bar{\tau}_k, \bar{r}_k} = 1$.

For every $r \in \mathcal{R}$ define

$$\begin{aligned} \mathcal{T}_r^* &= \{l \in \mathcal{T}_r \ : \ \exists k \in \mathcal{K}^* \text{ s.t. } \tau_k \in [l-r+1, l+r-1]\}, \\ \mathcal{S}_r^* &= \bigcup_{l\in\mathcal{T}_r^*} [l-r+1, l+r-1]. \end{aligned}$$

In other words, for all $r \in \mathcal{R}$, $\mathcal{T}_r^*$ is the subset of $\mathcal{T}_r$ for which each interval of detection $[l-r+1, l+r-1]$ contains a significant change-point. The next proposition recursively analyzes the detection sets corresponding to significant change-points $(\mathcal{S}_r^*)_{r\geq 1}$. The first inclusion means that significant change-points which can be detected with a local statistic with radius smaller than r are detected before step r, while the second inclusion means that each connected component of $\bigcup_{r\in\mathcal{R}} \mathcal{S}_r^*$ is included in a close neighborhoods of some significant change-point τ_k, $k \in \mathcal{K}^*$.

Proposition 5.10.1. *For all $r \in \mathcal{R} \cup \{0\}$, we have the double inclusion*

$$\{\tau_k \ : \ k \in \mathcal{K}^* \text{ and } \bar{r}_k \leq r\} \subset \bigcup_{r'\leq r, r'\in\mathcal{R}} \mathcal{S}_{r'}^* \subset \bigcup_{k\in\mathcal{K}^*} [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ . \tag{5.28}$$

The next proposition shows that for each step $r \in \mathcal{R}$, the subset of detection corresponding to non significant change-point is disjoint from $\bigcup_{r'\in\mathcal{R}} \mathcal{S}_{r'}^*$.

Proposition 5.10.2. *For all $r \in \mathcal{R}$, we have*

$$\bigcup_{l\in\mathcal{T}_r\setminus\mathcal{T}_r^*} [l-r+1, l+r-1] \cap \left(\bigcup_{r'\in\mathcal{R}} \mathcal{S}_{r'}^* \right) = \varnothing \ .$$

Recall that $(\mathrm{C}_k)_{k=1,\dots,\hat{K}}$ are defined as the connected component of $\bigcup_{r\in\mathcal{R}} \mathcal{S}_r$. To ease the notation, re-index (C_k) so that τ_k is the closest true change-point to $\hat{\tau}_k = \frac{\min \mathrm{C}_k + \max \mathrm{C}_k}{2}$. Since there is no false positive, $\tau_k \in \mathrm{C}_k$. By Proposition 5.10.2, the two closed subset $\bigcup_{r\in\mathcal{R}} \bigcup_{l\in\mathcal{T}_r\setminus\mathcal{T}_r^*} [l-r+1, l+r-1]$ and $\bigcup_{r\in\mathcal{R}} \mathcal{S}_r^*$ are disjoint. For all $k \in \mathcal{K}^*$, it holds by Proposition 5.10.1 that $\tau_k \in \bigcup_{r\in\mathcal{R}} \mathcal{S}_r^*$, so that C_k is a connected component of $\bigcup_{r\in\mathcal{R}} \mathcal{S}_r^*$ containing the significant change-point τ_k. In particular, $\hat{K} \geq |\mathcal{K}^*|$. We have

- By Proposition 5.10.1, $\mathrm{C}_k \subset [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)]$ for every $k \in \mathcal{K}^*$. Thus

$$|\hat{\tau}_k - \tau_k| \leq (\bar{r}_k - 1) < \frac{r_k}{4}.$$

- For all $k \in [K] \setminus \mathcal{K}^*$, either τ_k does not belong to $\bigcup_{r \in \mathcal{R}} \mathcal{S}_r$ and it is simply not detected, or it is the closest true change-point to $\hat{\tau}_k = \frac{\min \mathrm{C}_k + \max \mathrm{C}_k}{2}$ so that

$$\hat{\tau}_k \in \left[\tau_k - \frac{\tau_k + \tau_{k-1}}{2}, \tau_k + \frac{\tau_k + \tau_{k+1}}{2} \right] .$$

  In particular,

$$\{\hat{\tau}_{k'}, k' \leq \hat{K}\} \subset \left[\tau_1 - \frac{\tau_1 - \tau_0}{2}, \tau_K + \frac{\tau_{K+1} - \tau_K}{2} \right] .$$

- Finally, if there exists two estimated change-points $\hat{\tau}_{k_1}, \hat{\tau}_{k_2}$ in $\left[\tau_k - \frac{\tau_k + \tau_{k-1}}{2}, \tau_k + \frac{\tau_k + \tau_{k+1}}{2}\right]$, then either C_{k_1} or C_{k_2} does not contain τ_k. Then Θ is constant on C_{k_1} or on C_{k_2} and we obtain a contradiction since there is no false positive.

This concludes the proof of Theorem 5.2.1.

Proof of Proposition 5.10.1. To prove the proposition, we do an induction on $r \in \mathcal{R} \cup \{0\}$. The case $r = 0$ is trivial since by definition, $\mathcal{S}_0 = \varnothing$. Let $r \in \mathcal{R}$ and assume that the double inclusion Proposition 5.10.1 holds for all $r' < r, r' \in \mathcal{R} \cup \{0\}$.

First inclusion: Let $k \in \mathcal{K}^*$ be such that $\bar{r}_k = r$ and assume that the corresponding significant change-point τ_k has not been detected before step r, that is $\tau_k \notin \bigcup_{r' < r} \mathcal{S}^*_{r'}$. Since $k \in \mathcal{K}^*$, this implies in particular that $\tau_k \notin \bigcup_{r' < r} \mathcal{S}_{r'}$. Let us show that $\tau_k \in \mathcal{S}_r$. To this end we prove that

$$[\bar{\tau}_k - r + 1, \bar{\tau}_k + r - 1] \cap \bigcup_{r' < r, r' \in \mathcal{R}} \mathcal{S}_{r'} = \varnothing \tag{5.29}$$

and

$$T_{\bar{\tau}_k, r} = 1, \tag{5.30}$$

which will be enough since $|\bar{\tau}_k - \tau_k| \leq \bar{r}_k - 1 = r - 1$.

- **Proof of (5.29):** Assume for the sake of contradiction that there exists an integer z which belongs to $[\bar{\tau}_k - r + 1, \bar{\tau}_k + r - 1] \cap \bigcup_{\substack{r' < r \\ r' \in \mathcal{R}}} \mathcal{S}_{r'}$. There exists $r' < r$ such that $z \in \mathcal{S}_{r'}$ and $l(z) \in \mathcal{T}_{r'}$ such that $z \in [l(z) - r' + 1, l(z) + r' - 1]$. Since $\tau_k \notin \bigcup_{r' < r} \mathcal{S}_{r'}$, we have $\tau_k \notin [l(z) - r' + 1, l(z) + r' - 1]$. Moreover,

$$\begin{aligned} |l(z) - \tau_k| &\leq |l(z) - z| + |z - \bar{\tau}_k| + |\bar{\tau}_k - \tau_k| \\ &\leq (r' - 1) + (r - 1) + |\bar{\tau}_k - \tau_k| \\ &< r_k - r' \ , \end{aligned}$$

  Where the last inequality comes from the hypothesis $3(\bar{r}_k - 1) + |\bar{\tau}_k - \tau_k| \leq r_k$ Consequently,

$$[l(z) - r', l(z) + r'] \subset [\tau_k - r_k, \tau_k + r_k) \setminus \{\tau_k\} \ ,$$

  so that θ is constant on $[l(z) - r', l(z) + r') \cap \mathbb{N}$. Thus, $(l(z), r') \in \mathcal{H}_0$ and $l(z) \notin \mathcal{T}_{r'}$ since there is no false positive. This gives a contradiction and concludes the proof of (5.29).

- **Proof of (5.30):** This is simply a consequence of the fact that significant change-point are detected on the grid (See Item 2 in the definition of $\mathcal{A}$).

We have just shown that $\tau_k \in \mathcal{S}_r$ and hence $\tau_k \in \mathcal{S}^*_r$ so that the first inclusion holds at step r.

Second inclusion : Let x be an element of $\mathcal{S}_r^*$. There exists $l(x) \in \mathcal{T}_r^*$ such that $x \in [l(x) - r + 1, l(x) + r - 1]$. By definition of $\mathcal{T}_r^*$, there exists a significant change-point τ_k (i.e. such that $k \in \mathcal{K}^*$) belonging to $[l(x) - r + 1, l(x) + r - 1]$.

We necessarily have $\bar{r}_k \geq r$. Indeed, if $\bar{r}_k < r$, then by the induction hypothesis, $\tau_k \in \mathcal{S}_{r'}^*$ for some $r' < r$, which contradicts the fact that $\mathcal{S}_{r'}^*$ is disjoint from $[l(x) - r + 1, l(x) + r - 1] \subset \mathcal{S}_r^*$. Consequently,

$$\begin{aligned} |l(x) - \tau_k| + r - 1 &\leq 2r - 2 \\ &\leq 2(\bar{r}_k - 1) \end{aligned}$$

Thus

$$x \in [l(x) - r + 1, l(x) + r - 1] \subset [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ .$$

We have just shown that $\mathcal{S}_r^* \subset \bigcup_{k \in \mathcal{K}^*} [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)]$.
Therefore, the proposition is verified at step r and the induction is proved. □

Proof of Proposition 5.10.2. Let $k \in \mathcal{K}^*$ and C_k be the detected connected component containing the significant change-point τ_k

$$\mathrm{C}_k = \bigcup_{r' \in \mathcal{R}} \mathcal{S}_{r'}^* \cap [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ .$$

We know from Proposition 5.10.1 that C_k is a connected component of $\bigcup_{r' \in \mathcal{R}} \mathcal{S}_{r'}^*$ and we want to prove now that C_k does not overlap with $\bigcup_{l \in \mathcal{T}_r \setminus \mathcal{T}_r^*} [l - r + 1, l + r - 1]$ for some $r \in \mathcal{R}$. Let r_0 be such that C_k is the connected component of $\mathcal{S}_{r_0}$,

$$C_k \subset \mathcal{S}_{r_0}^* \ .$$

Such an r_0 exists and is unique since the sets $(\mathcal{S}_{r'}^*)$ are disjoint. We have from Proposition 5.10.1 that $\tau_k \in \bigcup_{r' \in \mathcal{R}, r' \leq \bar{r}_k} \mathcal{S}_{r'}^*$ so that

$$r_0 \leq \bar{r}_k \ .$$

Let $r \in \mathcal{R}$ and $l \in \mathcal{T}_r \setminus \mathcal{T}_r^*$ and assume without loss of generality that $l + r - 1 < \tau_k$. Since there is no false positive, $(l, r) \notin \mathcal{H}_0$ and there exists at least one true change-point in the interval of detection $[l - r + 1, l + r - 1]$. Denote $\tau_a, \ldots, \tau_b$ with $a \leq b$ the true change-points belonging to $[l - r + 1, l + r - 1]$. By definition of $\mathcal{T}_r \setminus \mathcal{T}_r^*$, $\tau_a, \ldots, \tau_b$ are not significant change-points, i.e. $a, a + 1, \ldots, b \notin \mathcal{K}^*$. We consider the two cases $r > \bar{r}_k$ and $r \leq \bar{r}_k$

- $\underline{r > \bar{r}_k}$: In that case, since the sets $(\mathcal{S}_{r'})$ are disjoint and $\mathrm{C}_k \subset \mathcal{S}_{r_0}^*$, we have $\mathrm{C}_k \cap [l - r + 1, l + r - 1] = \varnothing$.
- $\underline{r \leq \bar{r}_k}$: In that case, we have

  $$l + r - 1 \leq \tau_b + 2(r - 1) \leq \tau_b + 2(\bar{r}_k - 1) < \tau_k - 2(\bar{r}_k - 1) \ ,$$

  where we used the fact that $4(\bar{r}_k - 1) < r_k \leq \tau_k - \tau_b$. Since by Proposition 5.10.1 we have $\mathrm{C}_k \subset [l - r + 1, l + r - 1]$, we also have in that case $\mathrm{C}_k \cap [l - r + 1, l + r - 1] = \varnothing$.

This concludes the proof of the proposition. □

5.10.2 Proofs for Gaussian multivariate change-point detection

From now on, we use the following notation for all $(l, r) \in J_n$.

- For any $(v_1, \ldots, v_n)$ with $v_t \in \mathbb{R}^p$, the left mean and right mean of v on $[l - r, l + r)$ are denoted by

  $$\bar{v}_{l,+r} = \frac{1}{r} \sum_{t=l}^{l+r-1} v_t \quad \bar{v}_{l,-r} = \frac{1}{r} \sum_{t=l-r}^{l-1} v_t \ .$$

- The population term of the CUSUM statistic $\mathbf{C}_{l,r}$ is written

  $$U_{l,r} = \sqrt{\frac{r}{2}} \left(\bar{\theta}_{l,+r} - \bar{\theta}_{l,-r} \right) \ .$$

- With these notation, we write $v_{l,+r,i}, v_{l,-r,i}, U_{l,r,i}$ for the i^{th} coordinate of the vector $v_{l,+r}, v_{l,-r}, U_{l,r}$.
- We define, for $1 \leq s \leq p$, the order statistics $U_{l,r,(s)}$ by $|U_{l,r,(1)}| \geq |U_{l,r,(2)}| \geq \ldots |U_{l,r,(p)}|$.

5.10.2.1 Proof of Proposition 5.3.1

Step 0: Consequence of Equation (5.10) on the grid. Let $k \in [K]$ and assume that τ_k is a κ_{d}-dense high-energy change-point (see Equation (5.10)). We have that

$$\begin{aligned}\left\|U_{\bar{\tau}_k^{(\mathrm{d})},\bar{r}_k^{(\mathrm{d})}}\right\|^2 &\geq \frac{9}{16}\left\|U_{\tau_k,\bar{r}_k^{(\mathrm{d})}}\right\|^2 \\ &\geq \frac{9}{16\times 12}\kappa_{\mathrm{d}}\left(\sqrt{p\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})},\delta}\right)}+\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})},\delta}\right)\right),\end{aligned} \tag{5.31}$$

since by definition $\|\tau_k-\bar{\tau}_k^{(\mathrm{d})}\|\leq \bar{r}_k^{(\mathrm{d})}/4$, so that $\|\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\|^2\geq\frac{9}{16}\|\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\|^2$.

Step 1: Introduction of useful high probability events. Remark that

$$\frac{r}{2}\left[\left\|\overline{y}_{l,+r}-\overline{y}_{l,-r}\right\|^2-\left\|\overline{\theta}_{l,-r}-\overline{\theta}_{l,+r}\right\|^2\right]-\sigma^2 p = r\langle\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r},\overline{\theta}_{l,+r}-\overline{\theta}_{l,-r}\rangle+\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p\ .$$

The first term, written as

$$r\langle\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r},\overline{\theta}_{l,+r}-\overline{\theta}_{l,-r}\rangle\ ,$$

is a crossed term between the noise and the mean vector θ. Lemma 5.10.3 states that near the change-points and on the grid defined by the sets $\mathcal{R},\mathcal{D}_r$, it is jointly controlled with high probability.

Lemma 5.10.3. *Let $1\geq\delta>0$. The event*

$$\begin{aligned}\xi_1^{(\mathrm{d})} = \bigcap_{k\in[K]}\Bigg\{&\bar{r}_k^{(\mathrm{d})}\left|\langle\overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}}-\overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}},\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \\ &\leq\frac{1}{8}\bar{r}_k^{(\mathrm{d})}\left\|\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\right\|^2+16\sigma^2\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)\Bigg\}\ .\end{aligned}$$

holds with probability larger than $1-\delta$.

The second term, written as

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p\ ,$$

is a term of pure noise. Lemma 5.10.4 states that it is controlled jointly with high probability on the grid defined by the sets $\mathcal{R},\mathcal{D}_r$.

Lemma 5.10.4. *Let $1\geq\delta>0$. The event*

$$\xi_2^{(\mathrm{d})}=\bigcap_{r\in\mathcal{R}}\bigcap_{l\in\mathcal{D}_r}\left\{\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p\right|\leq 4\sigma^2\left[\sqrt{p\log\left(2\frac{n}{r\delta}\right)}+\log\left(2\frac{n}{r\delta}\right)\right]\right\}\ ,$$

holds with probability larger than $1-\delta$.

Set now

$$\xi^{(\mathrm{d})}:=\xi^{(\mathrm{d})}=\xi_1^{(\mathrm{d})}\cap\xi_2^{(\mathrm{d})}\ .$$

Note that

$$\mathbb{P}(\xi^{(\mathrm{d})})\geq 1-2\delta\ .$$

Step 2: Study in the 'no change-point' situation. Consider $r\in\mathcal{R},l\in\mathcal{D}_r$ such that $\{\tau_k,k\in[K]\}\cap[l-r,l+r)=\varnothing$. Note that since $\{\tau_k,k\in[K]\}\cap[l-r,l+r)=\varnothing$, we have $\overline{\theta}_{l,-r}=\overline{\theta}_{l,+r}$ so that

$$\frac{r}{2}\left\|\overline{\theta}_{l,-r}-\overline{\theta}_{l,+r}\right\|^2=0\ ,$$

and

$$r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r} \rangle = 0.$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 5.10.4

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{r\delta}\right)} + \log\left(2\frac{n}{r\delta}\right) \right] = \sigma^2 x_r^{(\mathrm{d})} \ .$$

And so

$$\Psi_{l,r}^{(\mathrm{d})} \leq x_r^{(\mathrm{d})} \ ,$$

so that

$$T_{l,r}^{(\mathrm{d})} = 0 \ ,$$

on $\xi^{(\mathrm{d})}$. This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in [K]$ τ_k is a κ_{d}-dense high-energy change-point - that is Equation(5.10) holds. We have from (5.31) that for κ_{d} large enough,

$$\begin{aligned} \frac{\overline{r}_k^{(\mathrm{d})}}{2} \left\| \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} \right\|^2 &\geq \frac{9}{16\times 12} \kappa_{\mathrm{d}} \sigma^2 \left(\sqrt{p \log\left(\frac{n}{\overline{r}_k^{(\mathrm{d})}, \delta}\right)} + \log\left(\frac{n}{\overline{r}_k^{(\mathrm{d})}, \delta}\right) \right) \\ &> 4\sigma^2 x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \ . \end{aligned}$$

So on $\xi^{(\mathrm{d})}$ this implies that - see Lemma 5.10.3

$$\overline{r}_k^{(\mathrm{d})} \left| \langle \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \rangle \right| \leq \frac{\overline{r}_k^{(\mathrm{d})}}{4} \left\| \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \right\|^2 .$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 5.10.4

$$\left| \frac{\overline{r}^{(\mathrm{d})}}{2} \left\| \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{\overline{r}_k^{(\mathrm{d})}}\delta^{-1}\right)} + \log\left(2\frac{n}{\overline{r}_k^{(\mathrm{d})}}\delta^{-1}\right) \right] = \sigma^2 x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} .$$

And so on $\xi^{(\mathrm{d})}$, combining the three previous displayed equations implies

$$\Psi_{\overline{\tau}_k^{(\mathrm{d})},\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \geq \frac{\frac{\overline{r}_k^{(\mathrm{d})}}{2} \left\| \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \right\|^2}{2\sigma^2} - x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} > (2-1) x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} = x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \ ,$$

so that

$$T_{\overline{\tau}_k^{(\mathrm{d})},\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} = 1 \ .$$

This concludes the proof of the second part of the proposition.

Proof of Lemma 5.10.3. Let $k \in [K]$. Since the vectors ε_t are i.i.d. and distributed as $\mathcal{N}(0, \sigma^2 \mathbf{I}_p)$, it holds that

$$\overline{r}_k^{(\mathrm{d})} \langle \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \rangle \sim \mathcal{N}\left(0, 2\overline{r}_k^{(\mathrm{d})} \sigma^2 \left\| \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \right\|^2 \right) \ .$$

And so for $\delta_k > 0$, it holds with probability larger than $1 - \delta_k$ it holds that

$$\begin{aligned} &\overline{r}_k^{(\mathrm{d})} \left| \langle \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \rangle \right| \\ &\leq 2\sigma \left\| \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},+\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(\mathrm{d})},-\overline{r}_k^{(\mathrm{d})}} \right\| \sqrt{\overline{r}_k^{(\mathrm{d})} \log(2\delta_k^{-1})} \ . \end{aligned}$$

Let us set $\delta_k = \frac{(\bar{r}_k^{(\mathrm{d})})^2\delta}{2n^2}$. Note that

$$\sum_{k\in[K]}\delta_k = \sum_{r\in R}\sum_{k\in[K]:\bar{r}_k^{(\mathrm{d})}=r}\frac{(\bar{r}_k^{(\mathrm{d})})^2\delta}{2n^2} \leq \sum_{r\in R}\sum_{l\in D_r}\frac{r^2\delta}{2n^2} \leq \sum_{r\in\mathcal{R}}\frac{r\delta}{2n} \leq \delta \ ,$$

since $r_k \geq \bar{r}_k^{(\mathrm{d})}$ and $|\mathcal{D}_r| \leq 2n/r$, and also by definition of $\mathcal{R}$ which implies $\sum_{r\in\mathcal{R}}\frac{r}{n} \leq 1$. And so if $\delta \leq 1$, then with probability larger than $1-\delta$, for any $k \in [K]$, we have

$$\bar{r}_k^{(\mathrm{d})}\left|\langle\overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \leq 2\sigma\left\|\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\right\| \sqrt{2\bar{r}_k^{(\mathrm{d})}\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}}\delta^{-1}\right)} \ .$$

This implies in particular that with probability larger than $1-\delta$, for any $k \in [K]$, we have

$$\bar{r}_k^{(\mathrm{d})}\left|\langle\overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \leq \frac{\bar{r}_k^{(\mathrm{d})}}{2}\frac{\left\|\overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(\mathrm{d})},-\bar{r}_k^{(\mathrm{d})}}\right\|^2}{4} + 16\sigma^2\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}}\delta^{-1}\right) \ .$$

□

Proof of Lemma 5.10.4. Let $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$. Since the vectors ε_t are i.i.d. and distributed as $\mathcal{N}(0,\sigma^2\mathbf{I}_p)$, it holds that

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 \sim \sigma^2\chi_p^2,$$

which implies by properties of the χ_p^2 distribution - see e.g. Lemma 1 of [54] - that for any $\delta_r > 0$ we have with probability larger than $1-\delta_r$

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq 2\sigma^2\sqrt{p\log(2/\delta_r)} + 2\sigma^2\log(2/\delta_r) \ .$$

If we set, for $\delta > 0$, $\delta_r = \frac{r^2\delta}{2n^2}$, we have that with probability larger than $1-\frac{r\delta}{n}$, that $\forall l \in D_r$

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq 2\sigma^2\sqrt{p\log(2/\delta_r)} + 2\sigma^2\log(2/\delta_r) \ ,$$

since $|\mathcal{D}_r| \leq 2n/r$. And so with probability larger than $1-\delta$, for all $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq 2\sigma^2\sqrt{p\log(2/\delta_r)} + 2\sigma^2\log(2/\delta_r) \ ,$$

since $\sum_{r\in\mathcal{R}}\frac{r}{n} \leq 1$. And so finally for $\delta \leq 1$ and with probability larger than $1-\delta$, for all $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq 4\sigma^2\left[\sqrt{p\log\left(2\frac{n}{r}\delta^{-1}\right)} + \log\left(2\frac{n}{r}\delta^{-1}\right)\right] \ .$$

This concludes the proof. □

5.10.2.2 Proof of Proposition 5.3.2

Step 1 : Analysis of the Berk-Jones statistics We first define a threshold $x_{r,s}^{(\mathrm{BJ})}$ for the Berk-Jones statistics for all $r, s \geq 1$

$$x_{r,s}^{(\mathrm{BJ})} = \min\left\{x \geq 2 : \overline{\Phi}(x) \leq \frac{s^2}{28^2 p\log(2\delta_{x,r}^{-1})}\right\} \ , \tag{5.32}$$

where we recall that $\delta_{x,r}$ are the weights defined by (5.13):

$$\delta_{x,r} = \frac{6\delta r}{\pi^2 x^2|\mathcal{D}_r|n} \ .$$

Remark that $(x^{(\mathrm{BJ})}_{r,s})$ is nonincreasing with s and define for all $r \geq 1$

$$\bar{s}_r = \min\left\{ s \in \mathcal{Z} \ : \quad s \geq \frac{28}{3} \log\left(2\delta^{-1}_{x^{(\mathrm{BJ})}_{r,s},r} \right) \right\} \ . \tag{5.33}$$

The second point of the following proposition ensures that if there exists $s \in \mathcal{Z}$ such that $U_{l,r,(s)} \geq t_s$ for some $s \geq \bar{s}_r$, for $(l,r) = (\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k)$, then $T^{(\mathrm{BJ})}_{l,r} = 1$ with high probability. We recall that $|U_{l,r,(1)}| \geq \cdots \geq |U_{l,r,(p)}|$ are the sorted absolute values of the coordinate of $U_{l,r}$ and that $\mathcal{H}_0$ is defined by (5.7).

Proposition 5.10.5. *There exists an event $\xi^{(\mathrm{BJ})}$ of probability larger than $1-2\delta$ such that the following holds:*

- $T^{(\mathrm{BJ})}_{l,r} = 0$ *for any* $(l,r) \in \mathcal{H}_0 \cap G$.
- *For all $k \in [K]$, if there exists $s \in \mathcal{Z}$ such that $s \geq \bar{s}_{\bar{r}^{(\mathrm{s})}_k}$ and $U_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k, (s)} > x^{(\mathrm{BJ})}_{\bar{r}^{(\mathrm{s})}_k, s}$, then $T^{(\mathrm{BJ})}_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k} = 1$.*

Step 2 : Analysis of the partial norm statistics Since it may happen that τ_k is a sparse high-energy change-point but there is no $s \geq \bar{s}_{\bar{r}^{(\mathrm{s})}_k}$ such that $U_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k, (s)} \geq x^{(\mathrm{BJ})}_{\bar{r}^{(\mathrm{s})}_k, s}$, we use the following proposition on the partial norm test statistic $T^{(\mathrm{p})}_{l,r}$:

Proposition 5.10.6. *There exists an event $\xi^{(\mathrm{p})}$ of probability larger than $1-2\delta$ such that the following holds:*

- $T^{(\mathrm{p})}_{l,r} = 0$ *for any* $(l,r) \in \mathcal{H}_0 \cap G$.
- *for any $k \in [K]$, if there exists $s \in \mathcal{Z}$ such that*

$$\sum_{s'=1}^{s} \left| U_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k, (s')} \right|^2 > 4 x^{(\mathrm{p})}_{\bar{r}^{(\mathrm{s})}_k, s} \ , \tag{5.34}$$

then $T^{(\mathrm{p})}_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k} = 1$.

Step 3 : Combination of the two Statistics Let us return to the proof of Proposition 5.3.2. To conclude the proof, it suffices to show that if τ_k is a κ_{s}-sparse high-energy change-point - see (5.11) - for some large enough constant κ_{s}, then the result of one of the two preceding propositions holds. This is precisely what the following lemma shows.

Lemma 5.10.7. *There exists a constant κ_{s} such that if τ_k is a κ_{s}-sparse high-energy change-point, then one of the following propositions is true:*

- *There exists $s \in \mathcal{Z}$ such that $s > \bar{s}_{\bar{r}^{(\mathrm{s})}_k}$ and $\left| U_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k, (s)} \right| > x^{(\mathrm{BJ})}_{\bar{r}^{(\mathrm{s})}_k, s}$.*
- *There exists $s \in \mathcal{Z}$ such that $s \leq \bar{s}_{\bar{r}^{(\mathrm{s})}_k}$ and $\sum_{s'=1}^{s} \left| U_{\bar{\tau}^{(\mathrm{s})}_k, \bar{r}^{(\mathrm{s})}_k, (s')} \right|^2 > 4 x^{(\mathrm{p})}_{\bar{r}^{(\mathrm{s})}_k, s}$.*

Proof of Proposition 5.10.5. The first part of the proposition is a simple consequence of the definition together with an union bound.

$$\begin{aligned} \mathbb{P}\left[\max_{(l,r) \in \mathcal{H}_0} T^{(\mathrm{BJ})}_{l,r} = 1 \right] &\leq \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{x \in \mathbb{N}^*} \delta^{(\mathrm{BJ})}_{x,r} \\ &\leq \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \frac{\delta r}{|\mathcal{D}_r| n} \leq \sum_{r \in \mathcal{R}} \frac{\delta r}{n} \leq \delta. \end{aligned}$$

We focus on the second part of the proposition. To ease the reading, we introduce some notation

$$\begin{aligned} \gamma_{x,r} = \overline{Q}^{-1}[\delta_{x,r}, p, 2\overline{\Phi}(x)] \ ; \quad \eta_{x,r,s} = \overline{Q}^{-1}[1-\delta_{x,r}/2, p-s, 2\overline{\Phi}(x)] \ ; \\ \psi_{x,r,s}(u) = \overline{Q}^{-1}[1-\delta_{x,r}/2, s, \overline{\Phi}(x-u) + \overline{\Phi}(x+u)] \ , \end{aligned}$$

for $x \geq 0$. In fact, $\gamma_{x,r}$ is the threshold of the statistics $N_{x,l,r}$. As for $\eta_{x,r,s}$, it stands for the contribution to $N_{x,l,r}$ of the $(p-s)$ coordinates i such that $\theta_{\cdot,i}$ is constant over $[l-r, l+r)$. Finally, $\psi_{x,r,s}(u)$ stands for the contribution to $N_{x,l,r}$ of the s coordinates i whose population CUSUM statistics $U_{l,r,i}$ is equal to u.

Lemma 5.10.8. *Consider any $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$. If for some positive integers s and x we have*

$$\psi_{x,r,s}(|U_{l,r,(s)}|) > \gamma_{x,r} - \eta_{x,r,s} \ , \tag{5.35}$$

then $\mathbb{P}[T_{l,r}^{(\mathrm{BJ})} = 1] \geq 1 - \delta_{x,r}$.

Denote $\mathcal{H}[\theta]$ the collection of (l,r) with $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$ that satisfy Condition (5.35) for some s and some x. We easily deduce from the above Lemma together with an union bound that, with probability higher than $1-\delta$, $T_{l,r}^{(\mathrm{BJ})} = 1$ for all $(l,r) \in \mathcal{H}[\theta]$.

Let us now provide a more explicit characterisation of $\mathcal{H}[\theta]$ with the following Lemma.

Lemma 5.10.9. *For any $1 \leq s \leq p$ and $r \in \mathcal{R}$ define x_s by*

$$x_s := x_{r,s}^{(\mathrm{BJ})} = \min\left\{ x \geq 2 : \overline{\Phi}(x) \leq \frac{s^2}{28^2 p \log(2\alpha_{x,r}^{-1})} \right\} \ .$$

We have $\psi_{x_s,r,s}(t_s) > \gamma_{x_s,r} - \eta_{x_s,r,s}$ provided that

$$s \geq \frac{28}{3} \log(2\delta_{x_s,r}^{-1}) \ . \tag{5.36}$$

Combining Lemma 5.10.9 and Lemma 5.10.8, we conclude the proof of the proposition. □

Proof of Lemma 5.10.8. Denote S any subset of size s, such that for any $j \in S$, $|U_{l,r,j}| \geq |U_{l,r,(s)}|$. Define

$$N_{x,l,r}^{(1)} = \sum_{i=1}^{p} \mathbf{1}_{i \notin S} \mathbf{1}_{|\mathbf{C}_{l,r,i}| > x}, \qquad N_{x,l,r}^{(2)} = \sum_{i=1}^{p} \mathbf{1}_{i \in S} \mathbf{1}_{|\mathbf{C}_{l,r,i}| > x}$$

Since, for any $x > 0$, the function $u \mapsto \overline{\Phi}(x+u) + \overline{\Phi}(x-u)$ is non-decreasing. As a consequence, the random variable $N_{x,l,r}^{(1)}$ is stochastically dominated by a Binomial distribution with parameters $(p-s, 2\overline{\Phi}(x))$. Besides, $N_{x,l,r}^{(2)}$ is stochastically dominated by a Binomial distribution with parameters $(s, \overline{\Phi}(x + |U_{l,r,(s)}|) + \overline{\Phi}(x - |U_{l,r,(s)}|))$. We obtain

$$\begin{aligned}
\mathbb{P}[T_{l,r}^{(BJ)} = 0] &\leq \mathbb{P}[N_{x,l,r} \leq \gamma_{x,r}] \leq \mathbb{P}[N_{x,l,r}^{(1)} < \eta_{x,r,s}] + \mathbb{P}[N_{x,l,r}^{(2)} \leq \gamma_{x,r} - \eta_{x,r,s}] \\
&\leq \frac{\delta_{x,r}}{2} + 1 - \overline{Q}[\gamma_{x,r} - \eta_{x,r,s}, s, \overline{\Phi}(x - |U_{l,r,(s)}|) + \overline{\Phi}(x + |U_{l,r,(s)}|)] \\
&\leq \frac{\delta_{x,r}}{2} + \frac{\delta_{x,r}}{2} \leq \delta_{x,r} \ .
\end{aligned}$$

□

Proof of Lemma 5.10.9. From Bernstein inequality, we deduce that, for any positive integers s and x,

$$\begin{aligned}
\gamma_{x,s} &\leq 2p\overline{\Phi}(x) + 2\sqrt{p\overline{\Phi}(x)\log(\delta_{x,r}^{-1})} + \frac{2}{3}\log(\delta_{x,r}^{-1}) \ ; \\
\eta_{x,r,s} &\geq 2(p-s)\overline{\Phi}(x) - 2\sqrt{p\overline{\Phi}(x)\log(2\delta_{x,r}^{-1})} - \frac{2}{3}\log(2\delta_{x,r}^{-1}) \ .
\end{aligned}$$

Hence, it follows that

$$\gamma_{x,s} - \eta_{x,r,s} \leq 2s\overline{\Phi}(x) + 4\sqrt{p\overline{\Phi}(x)\log(2\delta_{x,r}^{-1})} + \frac{4}{3}\log(2\delta_{x,r}^{-1}) \ .$$

For $u = x$, we have $\overline{\Phi}(x-u) + \overline{\Phi}(x+u) \geq \overline{\Phi}(0) = 1/2$ and we derive from Bernstein inequality that

$$\psi_{x,r,s}(t) \geq \frac{s}{2} - \sqrt{s\log(2\delta_{x,r}^{-1})} - \frac{2}{3}\log(2\delta_{x,r}^{-1}) \ .$$

As a ce, $\psi_{x,r,s}(t) > \gamma_{x,s} - \eta_{x,r,s}$ as long as

$$s(1 - 4\overline{\Phi}(x)) > 12\sqrt{p\overline{\Phi}(x)\log(2\delta_{x,r}^{-1})} + \frac{12}{3}\log(2\delta_{x,r}^{-1}) \ .$$

Provided that we take $x \geq 2$, the latter holds if

$$s \geq 14\sqrt{p\overline{\Phi}(x)\log(2\delta_{x,r}^{-1})} + \frac{14}{3}\log(2\delta_{x,r}^{-1}) \tag{5.37}$$

In view of the definition (5.32) of x_s, we have $14\sqrt{p\overline{\Phi}(x_s)\log(2\delta_{x_s,r}^{-1})} \leq s/2$. Hence, under Condition (5.33), (5.37) holds and we conclude that $\psi_{x_s,r,s}(x_s) > \gamma_{x_s,s} - \eta_{x_s,r,s}$.

□

Proof of Proposition 5.10.6. The following lemma ensures that the partial norm test returns 0 with high probability jointly at all positions where there is no change-point. We write $\bar{C}_p^s$ for the set of all combinations of s indices taken from $[p]$.

Lemma 5.10.10 (concentration of the pure noise for the second sparse statistic)**.** *If $1 \geq \delta > 0$, then the event*

$$\xi_1^{(\mathrm{p})} = \left\{ \forall r \in \mathcal{R}, l \in \mathcal{D}_r, s \in \mathcal{Z} \quad \max_{S \in \bar{C}_p^s} \sum_{i \in S} \frac{r}{2\sigma^2} \left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)^2 \leq x_{r,s}^{(\mathrm{p})} \right\}$$

holds with probability higher than $1-\delta$.

We now state the following lemma, which ensures that the partial norm test returns 1 with high probability jointly at relevant positions which are close to a change-point.

Lemma 5.10.11 (concentration on the change-points for the second sparse statistic)**.** *We write $\bar{\mathcal{K}}^*$ for the set of $k \in [K]$ such that*

- $s_k \leq \sqrt{p\log\left(\frac{n}{r_k\delta}\right)}$
- $\sum_{s'=1}^{s} \left| U_{\bar{\tau}_k^{(\mathrm{s})}, \bar{r}_k^{(\mathrm{s})}, (s')} \right|^2 \geq 4x_{\bar{r}_k^{(\mathrm{s})},s}^{(\mathrm{p})}$

If $1 \geq \delta > 0$, the event

$$\xi_2^{(\mathrm{p})} = \left\{ \forall k \in \bar{\mathcal{K}}^* \ : \exists s \in \mathcal{Z} \quad s.t.\ \Psi^{(\mathrm{p})}_{\bar{\tau}_k^{(\mathrm{s})}, \bar{r}_k^{(\mathrm{s})}, s} > x^{(\mathrm{p})}_{\bar{r}_k^{(\mathrm{s})}, s} \right\} ,$$

holds with probability higher than $1-\delta$.

Lemmas 5.10.10 and 5.10.11 directly imply the result of the proposition. □

Proof of Lemma 5.10.10. Let $r \in \mathcal{R}, l \in \mathcal{D}_r, s \leq \bar{s}_r$ and $S \in \bar{C}_p^s$. Let $\delta > 0$, $\delta_{r,s} = \left(\frac{r}{n}\right)^2 \left(\frac{s}{2ep}\right)^s \delta$. Since $\sqrt{\frac{r}{2\sigma^2}}\left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)$ follows a $\mathcal{N}(0,1)$ distribution for all l, r, i, we have by Bernstein's inequality that with probability larger than $1-\delta_{r,s}$,

$$\begin{aligned}
\sum_{i \in S} \left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)^2 &\leq s + 2\sqrt{s\log\left(\frac{1}{\delta_{r,s}}\right)} + \log\left(\frac{1}{\delta_{r,s}}\right) \\
&\leq 2\left(s + \log\left(\frac{1}{\delta_{r,s}}\right)\right) \\
&= 2\left(s + s\log\left(\frac{2ep}{s}\right) + \log\left(\frac{n^2}{r^2\delta}\right)\right) \\
&\leq 4\left(s\log\left(\frac{2ep}{s}\right) + \log\left(\frac{n}{r\delta}\right)\right) .
\end{aligned}$$

Since the number of such S is smaller than $\left(\frac{ep}{s}\right)^s$, a union bound gives

$$\begin{aligned}
\mathbb{P}\left(\xi_1^{(\mathrm{p})}\right) &\geq 1 - \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{s \in \mathcal{Z}} |\bar{C}_p^s| \left(\frac{s}{2ep}\right)^s \left(\frac{r}{n}\right)^2 \delta \\
&\geq 1 - \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{s \in \mathcal{Z}} \left(\frac{1}{2}\right)^s \left(\frac{r}{n}\right)^2 \delta \\
&\geq 1 - \delta ,
\end{aligned}$$

which yields the result. □

Proof of Lemma 5.10.11. Let $k \in \bar{\mathcal{K}}^*$, and $s \in \mathcal{Z}$ such that

$$\sum_{i=1}^{s} U^2_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(i)} > 4x^{(\mathrm{p})}_{\bar r_k^{(\mathrm{s})},s} \quad . \tag{5.38}$$

To ease the reading, we write $(\tau, r) = (\bar\tau_k^{(\mathrm{s})}, \bar r_k^{(\mathrm{s})})$. Then on the event $\xi_1^{(\mathrm{p})}$ which holds with probability $1-\delta$, we have

$$\begin{aligned}
\Psi^{(\mathrm{p})}_{\tau,r,s} &= \max_{S\in\bar C_p^s} \sum_{i\in S} \frac{r}{2\sigma^2}\left(\bar\theta_{\tau,+r,i} + \bar\varepsilon_{\tau,+r,i} - \bar\theta_{\tau,-r,i} - \bar\varepsilon_{\tau,-r,i}\right)^2 \\
&\geq \max_{S\in\bar C_p^s} \sum_{i\in S} \frac{1}{2} U^2_{\tau,r,i} - \frac{r}{2\sigma^2}\left(\bar\varepsilon_{\tau,+r,i} - \bar\varepsilon_{\tau,-r,i}\right)^2 \\
&> 2x^{(\mathrm{p})}_{r,s} - x^{(\mathrm{p})}_{r,s} \\
&= x^{(\mathrm{p})}_{r,s} \quad ,
\end{aligned}$$

where in the second inequality, we used the fact that $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$ for all $a, b \in \mathbb{R}$. □

Proof of Lemma 5.10.7. First remark that there exists a large enough constant C such that for all $r, s \geq 1$,

$$\begin{aligned}
\left(x^{(\mathrm{BJ})}_{r,s}\right)^2 &\leq C \log\left(\frac{ep}{s^2}\log\left(\frac{n}{r\delta}\right)\right) \\
\bar s_r &\leq C \log\left(\log\left(\frac{ep}{\bar s_r^2}\right)\frac{n}{r\delta}\right) \quad ,
\end{aligned}$$

where we recall that $\bar s_r$ is defined by (5.33) and $x^{(\mathrm{BJ})}_{r,s}$ by (5.32). These two inequalies come from the fact that for all $t \geq 2$ and all $A > 0$, if $t \leq A + \log(t)$ then $t \leq 2A$. Assume that for all $s' = \bar s_{\bar r_k^{(\mathrm{s})}} + 1, \dots, s_k$ we have $|U_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(s')}| < x^{(\mathrm{BJ})}_{\bar r_k^{(\mathrm{s})},s'}$. To ease the notation, we write $\bar s = \bar s_{\bar r_k^{(\mathrm{s})}} \wedge s_k$ and in what follows we prove that $\sum_{s'=1}^{\bar s} |U_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(s')}|^2 > 4x^{(\mathrm{p})}_{\bar r_k^{(\mathrm{s})},\bar s}$ when κ_{s} is a large enough constant. We have

$$\begin{aligned}
\sum_{s'=\bar s_{\bar r_k^{(\mathrm{s})}}+1}^{s_k} U^2_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(s')} &\leq C_1 \sum_{i=0}^{\lfloor\log(s_k)\rfloor} 2^i \log\left(\frac{ep}{2^{2i}}\log\left(\frac{n}{\bar r_k^{(\mathrm{s})}\delta}\right)\right) \\
&\leq C_1 s_k \log\left(2e\log\left(\frac{n}{\bar r_k^{(\mathrm{s})}\delta}\right)\right) + C_1 \sum_{i=0}^{\lfloor\log(s_k)\rfloor} 2^i \log\left(\frac{p}{2^{2(i+1)}}\right) \quad ,
\end{aligned}$$

for some universal constant C_1. To handle the second term remark that since $x \mapsto \log\left(\frac{p}{x^2}\right)$ is decreasing, we have

$$\begin{aligned}
\sum_{i=0}^{\lfloor\log(s_k)\rfloor} 2^i \log\left(\frac{p}{2^{2(i+1)}}\right) &\leq \int_1^{2s_k} \log\left(\frac{p}{x^2}\right) dx \\
&= 2s_k \log\left(\frac{p}{(2s_k)^2}\right) + 2s_k - 1 \\
&\leq 2s_k \log\left(\frac{p}{s_k^2}\right) \quad ,
\end{aligned}$$

and thus

$$\sum_{s'=\bar s_{\bar r_k^{(\mathrm{s})}}+1}^{s_k} U^2_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(s')} \leq 2C_1 s_k \log\left(2e\frac{p}{s_k^2}\log\left(\frac{n}{\bar r_k^{(\mathrm{s})}\delta}\right)\right),$$

which finally gives

$$\begin{aligned}
\sum_{s'=1}^{\bar s} U^2_{\bar\tau_k^{(\mathrm{s})},\bar r_k^{(\mathrm{s})},(s')} &\geq \frac{9}{16}\bar r_k^{(\mathrm{s})}\Delta_k^2 - 2C_1 s_k \log\left(\frac{2ep}{s_k^2}\log\left(\frac{n}{\bar r_k^{(\mathrm{s})}\delta}\right)\right) \\
&\geq 4x^{(\mathrm{p})}_{\bar r_k^{(\mathrm{s})},\bar s}.
\end{aligned}$$

In the first inequality we used the fact that

$$\left|\bar{\tau}_k^{(\mathrm{s})} - \tau_k\right| \le \frac{1}{4}\bar{r}_k^{(\mathrm{s})},$$

so that for all i,

$$\begin{aligned}\left|\bar{\theta}_{\bar{\tau}_k^{(\mathrm{s})},+\bar{r}_k^{(\mathrm{s})},i} - \bar{\theta}_{\bar{\tau}_k^{(\mathrm{s})},-\bar{r}_k^{(\mathrm{s})},i}\right| &= \frac{1}{\bar{r}_k^{(\mathrm{s})}}\left|\left(\bar{r}_k^{(\mathrm{s})} + \bar{\tau}_k^{(\mathrm{s})} - \tau_k\right)\mu_{k,i} - \left(\bar{r}_k^{(\mathrm{s})} - \bar{\tau}_k^{(\mathrm{s})} + \tau_k\right)\mu_{k-1,i}\right| \\ &\ge \left(1 - \frac{\left|\bar{\tau}_k^{(\mathrm{s})} - \tau_k\right|}{\bar{r}_k^{(\mathrm{s})}}\right)\left|\mu_{k,i} - \mu_{k-1,i}\right| \\ &> \frac{3}{4}\left|\mu_{k,i} - \mu_{k-1,i}\right| = \frac{3}{4}U_{k,i}\ .\end{aligned}$$

In the second inequality, we used the fact that

- $8\bar{r}_k^{(\mathrm{s})}\Delta_k^2 \ge \kappa_{\mathrm{s}}\sigma^2\left(s_k \log\left(\frac{p}{s_k^2}\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right) + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right)$ for a large enough constant κ_{s} (see (5.11)),
- $x \mapsto x\log\left(\frac{ep}{x^2}\right)$ is increasing for $x \le p$, so that s_k can be replaced by $\bar{s}$,
- $\bar{s} \le C\log\left(\log\left(\frac{ep}{\bar{s}^2}\right)\frac{n}{r\delta}\right)$.

This concludes the proof of the lemma. □

5.10.2.3 Proof of Corollary 5.3.3

Let $\xi^{(\mathrm{d})}$ and $\xi^{(\mathrm{s})}$ be two events such that Proposition 5.3.1 and Proposition 5.3.2 hold respectively with constants $\kappa_{\mathrm{d}}, \kappa_{\mathrm{s}}$ and with probability $1-2\delta$ and $1-4\delta$, and write $\xi = \xi^{(\mathrm{d})} \cap \xi^{(\mathrm{s})}$. From now on, we work on the event ξ, which holds with probability $1-6\delta$. Let us choose $c_0 \ge 2(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}})$ in (5.8). For all k such that τ_k is a c_0-high-energy change-point, define

$$(\bar{\tau}_k, \bar{r}_k) = \begin{cases} (\bar{\tau}_k^{(\mathrm{d})}, \bar{r}_k^{(\mathrm{d})}) \text{ if } s_k > \sqrt{p\log\left(\frac{n}{r_k\delta}\right)} \\ (\bar{\tau}_k^{(\mathrm{s})}, \bar{r}_k^{(\mathrm{s})}) \text{ if } s_k \le \sqrt{p\log\left(\frac{n}{r_k\delta}\right)} \end{cases}.$$

$(\bar{r}_k, \bar{\tau}_k)$ is well defined. Indeed, If $s_k \le \sqrt{p\log\left(\frac{n}{r_k\delta}\right)}$ then

$$s_k\log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right) \ge \frac{1}{2}\left(s_k\log\left(\frac{p}{s_k^2}\log\left(\frac{n}{r_k\delta}\right)\right) + \log\left(\frac{n}{r_k\delta}\right)\right)\ .$$

Now if $s_k \ge \sqrt{p\log\left(\frac{n}{r_k\delta}\right)}$ then using $\log(1+x) \ge \frac{x}{2}$ for $x \in [0,1]$ we have

$$s_k\log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right) \ge \frac{1}{2}\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} + \log\left(\frac{n}{r_k\delta}\right)\right)\ .$$

According to Theorem 5.2.1, it is sufficient to prove that the event $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ defined in Section 5.2.3 holds on ξ:

1. **(No false positive):** for every $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$, if Θ is constant on $[l-r, l+r)$ then

$$T_{l,r} = T_{l,r}^{(\mathrm{d})} \vee T_{l,r}^{(\mathrm{s})} = 0,$$

by Proposition 5.3.1 and Proposition 5.3.2.

2. **(High-energy change-point detection):** for every k such that τ_k has c_0-high-energy, it holds by definition of $\bar{r}_k^{(\mathrm{d})}$ and $\bar{r}_k^{(\mathrm{s})}$ that

$$4(\bar{r}_k - 1) \le r_k.$$

Moreover, $T_{\bar{\tau}_k,\bar{r}_k}^{(\mathrm{s})} = 1$ if $(\bar{\tau}_k, \bar{r}_k) = (\bar{\tau}_k^{(\mathrm{d})}, \bar{r}_k^{(\mathrm{d})})$ by Proposition 5.3.2 and $T_{\bar{\tau}_k,\bar{r}_k}^{(\mathrm{d})} = 1$ if $(\bar{\tau}_k, \bar{r}_k) = (\bar{\tau}_k^{(\mathrm{s})}, \bar{r}_k^{(\mathrm{s})})$ by Proposition 5.3.1.

Theorem 5.2.1 ensures that for all $k \in [K]$ such that τ_k is a c_0-high-energy change-point, there exists $k' \in [\hat{K}]$ such that

$$|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1.$$

It remains to show that

$$\bar{r}_k - 1 \leq \frac{r_k^*}{2},$$

where r_k^* is define by (5.9). Using $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$ and $\log(1+x) \geq \log(x)$ for $x \geq 1$ we have

$$8\bar{r}_k \Delta_k^2 \leq 4(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}}) \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{\bar{r}_k \delta}\right)}\right) + \log\left(\frac{n}{\bar{r}_k \delta}\right)\right],$$

when $\bar{r}_k \geq 2$. Thus $2(\bar{r}_k - 1) \leq r_k^*$ for $c_0 \geq 2(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}})$. This concludes the proof of Corollary 5.3.3.

5.10.3 Proofs for sub-Gaussian multivariate change-point detection

We recall that in this section, we work on the complete grid $\mathcal{G}_F = J_n$ defined in Section 5.2.

5.10.3.1 Proof of Proposition 5.4.1

Step 1: Introduction of useful high probability events. We first introduce two events $\xi_1^{(\mathrm{d})}$ and $\xi_2^{(\mathrm{d})}$ on which the noise can be controlled. Remark that by a simple computation, the noise can be decomposed as follows :

$$\frac{r}{2}\left[\left\|\overline{y}_{l,+r} - \overline{y}_{l,-r}\right\|^2 - \left\|\overline{\theta}_{l,-r} - \overline{\theta}_{l,+r}\right\|^2\right] - \sigma^2 p = r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r}\rangle + \frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p \ .$$

The first term written as

$$r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r}\rangle$$

is a crossed term between the noise and the mean vector θ. Lemma 5.10.12 states that for l equal to a true change-point τ_k and r of order r_k^*, it is controlled on event $\xi_1^{(\mathrm{d})}$ with high probability.

Lemma 5.10.12 (concentration of the crossed terms)**.** *Assume that κ is a large enough universal constant. The event*

$$\xi_1^{(\mathrm{d})} = \left\{ \forall k \in [K] \text{ s.t. Equation (5.17) holds for } k, \right.$$
$$\left. \bar{r}_k^{(\mathrm{d})} \left|\langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}\rangle\right| \leq \frac{\bar{r}_k^{(\mathrm{d})}}{4}\left\|\overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}\right\|^2 \right\}$$

holds with probability higher than $1-\delta$.

The second term written as

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p \ ,$$

is a term of pure noise. Lemma 5.10.13 states that it is controlled on event $\xi_2^{(\mathrm{d})}$ with high probability.

Lemma 5.10.13 (concentration of the pure noise)**.** *There exists a constant $\bar{c}_{\mathrm{conc}} > 0$ such that the event*

$$\xi_2^{(\mathrm{d})} = \left\{ \forall (l,r) \in J_n, \ \left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right)\right\}$$

holds with probability higher than $1-2\delta$.

Set now

$$\xi^{(\mathrm{d})} := \xi_1^{(\mathrm{d})} \cap \xi_2^{(\mathrm{d})} \ .$$

Note that

$$\mathbb{P}(\xi^{(\mathrm{d})}) \geq 1 - 3\delta \ .$$

Step 2: Study in the 'no change-point' situation. We remind that $\mathcal{H}_0$ stands for elements (l,r) such that there is no change-point in $[l-r,l+r)$ and that it is defined in (5.7). Consider $(l,r) \in J_n \cap \mathcal{H}_0$. Note that since $\{\tau_k, k \in [K]\} \cap [l-r,l+r) = \varnothing$, we have $\overline{\theta}_{l,-r} = \overline{\theta}_{l,+r}$ so that

$$\frac{r}{2}\left\|\overline{\theta}_{l,-r} - \overline{\theta}_{l,+r}\right\|^2 = 0 \ ,$$

and

$$r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r}\rangle = 0 \ .$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 5.10.13

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \le \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{p\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right) \le \sigma^2 x_r^{(\mathrm{d})},$$

for $\bar{c}_{\mathrm{thresh}} \ge \bar{c}_{\mathrm{conc}}$ - note that $\bar{c}_{\mathrm{conc}} > 0$ is a universal constant. And so

$$\Psi_{l,r}^{(\mathrm{d})} \le x_r^{(\mathrm{d})} \ ,$$

so that

$$T_{l,r}^{(\mathrm{d})} = 0 \ ,$$

on $\xi^{(\mathrm{d})}$. This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in [K]$ such that τ_k is a κ-dense high-energy change-point - see Equation (5.17). We have

$$\frac{\bar{r}_k^{(\mathrm{d})}}{2}\left\|\overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}}\right\|^2 \ge \frac{\kappa}{8}L^2\left(\sqrt{p\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)\right).$$

So on $\xi^{(\mathrm{d})}$ choosing κ large enough implies that - see Lemma 5.10.12

$$\bar{r}_k^{(\mathrm{d})}\left|\langle \overline{\varepsilon}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \le \frac{\bar{r}_k^{(\mathrm{d})}}{4}\left\|\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\right\|^2 \ .$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 5.10.13

$$\left|\frac{\bar{r}_k^{(\mathrm{d})}}{2}\left\|\overline{\varepsilon}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\right\|^2 - \sigma^2 p\right| \le \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{p\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)\right) \le \sigma^2 x_{\bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \ ,$$

for $\bar{c}_{\mathrm{thresh}} \ge \bar{c}_{\mathrm{conc}}$ - note that $\bar{c}_{\mathrm{conc}} > 0$ is a universal constant. Thus on $\xi^{(\mathrm{d})}$, combining the three previous displayed equations implies

$$\begin{aligned}\Psi_{\tau_k,\bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} &\ge \frac{\bar{r}_k^{(\mathrm{d})}}{4\sigma^2}\left\|\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\right\|^2 - x_{\bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \\ &\ge \left(\frac{c_0}{16} - \bar{c}_{\mathrm{thresh}}\right)\frac{L^2}{\sigma^2}\left(\sqrt{p\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)\right) > x_{\bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \ ,\end{aligned}$$

since $\kappa > 32\bar{c}_{\mathrm{thresh}}$. And so on $\xi^{(\mathrm{d})}$:

$$T_{\tau_k,\bar{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} = 1 \ .$$

This concludes the proof of the second part of the proposition.

Proof of Lemma 5.10.12. Let k be in $[K]$ and such that Equation (5.17) is satisfied. Remark that θ is constant on $[\tau_k - \bar{r}_k^{(\mathrm{d})}, \tau_k)$ and is equal to μ_{k-1}, and is also constant on $[\tau_k, \tau_k + \bar{r}_k^{(\mathrm{d})})$ and is equal to μ_k. First, from the definition of the ψ_2-norm of a vector, there exists a universal constant $C > 0$ such that for all $k = 1 \dots K$,

$$\begin{aligned}
\left\| \bar{r}_k^{(\mathrm{d})} \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right\|_{\psi_2} &\leq \bar{r}_k^{(\mathrm{d})} \left\| \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \right\|_{\psi_2} |\mu_k - \mu_{k-1}| \\
&\leq C \sqrt{\bar{r}_k^{(\mathrm{d})}} \, \|\varepsilon_1\|_{\psi_2} \, |\mu_k - \mu_{k-1}| \\
&\leq C \sqrt{\bar{r}_k^{(\mathrm{d})}} L \, |\mu_k - \mu_{k-1}| \\
&\leq CL \sqrt{r_k \Delta_k^2} \ .
\end{aligned}$$

Thus by definition of sub-Gaussianity, for all $t > 0$,

$$\mathbb{P}\left(\bar{r}_k^{(\mathrm{d})} \left| \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right| \geq t \right) \leq \exp\left(-c \frac{t^2}{L^2 r_k \Delta_k^2} \right) \ ,$$

for some constant $c > 0$. Finally we apply the concentration inequality to $t = \frac{r_k \Delta_k^2}{4}$ - remembering that τ_k is a κ-dense high-energy change-point in the sense of Equation (5.17) - and sum over k to obtain a union bound over ξ_2^c :

$$\begin{aligned}
\mathbb{P}\left(\xi_2^c\right) &\leq \sum_{k=1}^{K} \mathbb{P}\left(r \left| \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right| \geq \frac{r_k \Delta_k^2}{4} \right) \\
&\leq \sum_{k=1}^{K} \exp\left(-c \frac{r_k \Delta_k^2}{16 L^2} \right) \\
&\leq \sum_{k=1}^{K} \exp\left(-c' \kappa \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}} \delta^{-1} \right) \right) \qquad (c' = c/16) \\
&\leq \sum_{k=1}^{K} \left(\frac{\bar{r}_k^{(\mathrm{d})}}{n} \right)^{c'\kappa} \delta^{c'\kappa} \\
&\leq \delta \ ,
\end{aligned}$$

where the last inequality comes from the fact that $\sum_{k=1}^{K} \bar{r}_k^{(\mathrm{d})} \leq n$ and the fact that κ is chosen large enough so that $c'\kappa \geq 1$. □

Proof of Lemma 5.10.13. Remark first that by homogeneity, we can assume without loss of generality that $L = 1$. To provide a proof, we will use the Hanson-Wright inequality in high dimension, which is a way to control quadratic forms of the noise.

Lemma 5.10.14 (Hanson-Wright inequality in high dimension)**.** *Let $A = (a_{ij})$ be a $m \times m$ matrix and $\varepsilon_1, \dots, \varepsilon_m$ be sub-Gaussian vectors of dimension p with norm smaller than 1. Then*

$$\mathbb{P}\left(\left| \sum_{1 \leq i,j \leq m} a_{i,j} \langle \varepsilon_i, \varepsilon_j \rangle - \mathbb{E}\left[\sum_{1 \leq i,j \leq m} a_{i,j} \langle \varepsilon_i, \varepsilon_j \rangle \right] \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{p \|A\|_F^2}, \frac{t}{\|A\|_{op}} \right) \right) \ ,$$

where c is an absolute constant, $\|A\|_F^2 = \sum_{i,j} a_{i,j}^2$ is the squared Frobenius norm of A and $\|A\|_{op}$ is the operator norm of A.

The proof of this lemma relies on the classical Hanson Wright inequality that is proved for example in [78]. To prove the proposition, we will use a chaining argument. To this end, we let $(N_u)_{u \geq 0}$ be the following covering sets of J_n :

$$N_u = J_n \cap \left\{ i 2^{\kappa_1 - u}, i \in \mathbb{N} \right\}^2 \ ,$$

where we define $\kappa_1 = \lfloor \log_2(n) \rfloor$, and more generally $\kappa_r = \lfloor \log_2(n/r) \rfloor$ for $r = 1, \dots n$. Remark that the higher u is, the finer the covering set N_u is, and $N_{\kappa_1} = J_n$. For all $u \geq 0$, we define the projection map π_u from J_n to N_u by

$$\pi_u(l, r) = \underset{(\hat{l}, \hat{r}) \in N_u}{\arg\min} \left(|\hat{l} - l| + |\hat{r} - r| \right) \ .$$

In the sequel, we will use the slight abuse of notation for (l,r) in J_n :

$$(l_u, r_u) = \pi_u(l,r) \ .$$

A useful lemma to control the distance between (l,r) and its projection (l_u, r_u) can be stated as follow.

Lemma 5.10.15. *For all $(l,r) \in J_n$ and $0 \le u \le \kappa_1$ such that $N_u \neq \varnothing$,*

$$|l_u - l| + |r_u - r| \le 2\frac{n}{2^u} \ .$$

Let $(l,r) \in J_n$. From know on, we write $\varepsilon_{l,+r} = r\bar{\varepsilon}_{l,+r} = \sum_{t=l}^{l+r-1} \varepsilon_t$ and $\varepsilon_{l,-r} = r\bar{\varepsilon}_{l,+r}$. The chaining relation can be written as

$$\begin{aligned}\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r} - \bar{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p = \frac{1}{2r}\left[\left\|\varepsilon_{l_{\kappa_r},+r_{\kappa_r}} - \varepsilon_{l_{\kappa_r},-r_{\kappa_r}}\right\|^2 - 2r_{\kappa_r}\sigma^2 p\right] \\ + \frac{1}{2r}\sum_{v=\kappa_r}^{\kappa_1}\left[\left\|\varepsilon_{l_{v+1},+r_{v+1}} - \varepsilon_{l_{v+1},-r_{v+1}}\right\|^2 - \left\|\varepsilon_{l_v,+r_v} - \varepsilon_{l_v,-r_v}\right\|^2 - 2(r_{v+1} - r_v)\sigma^2 p\right] \ .\end{aligned}$$

Remark that the chaining summation starts at scale $u = \kappa_r$ so that $\frac{n}{2^u} \asymp r$. The first term of the chaining is an approximation on the grid at level u of the term $\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r} - \bar{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p$. The second term can be viewed as an error term, and we will show that it is of the same order as the first term. Since both terms are quadratic forms of the noise, we will need an upper bound on the norm of their corresponding matrix to apply the Hanson Wright inequality - see Lemma 5.10.14.

Lemma 5.10.16 (Control of the Frobenius norm)**.** *Let (l,r) be a fixed element of J_n. Let A and B be the corresponding matrix of the two following quadratic form :*

$$\varepsilon^T A \varepsilon = \left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 \quad \text{and} \quad \varepsilon^T B \varepsilon = \left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - \left\|\varepsilon_{l',+r'} - \varepsilon_{l',-r'}\right\|^2 \ .$$

Then

$$\begin{aligned}\|A\|_F^2 &\le 16r^2 \\ \|B\|_F^2 &\le 24\left(|l-l'| + |r-r'|\right)\left(r + r' + |l-l'|\right) \ .\end{aligned}$$

The following lemma aims at upper bounding the first term of the chaining relation with high probability.

Lemma 5.10.17. *There exists a constant C_N such that for all n, the event*

$$\xi_N^{(\mathrm{d})} = \bigcap_{u \ge 0} \bigcap_{\substack{(l,r)\in N_u \\ r \le 3\frac{n}{2^u}}} \left\{\left|\left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2r\sigma^2 p\right| \le C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right)\right\} \ .$$

holds with probability higher than $1-\delta$.

For $u = \kappa_r$, $(l_u, r_u) \in N_u$ Lemma 5.10.15 gives $r_u \le r + 2\frac{n}{2^u} \le 3\frac{n}{2^u}$. Consequently, on the event $\xi_N^{(\mathrm{d})}$, we obtain

$$\left|\frac{1}{2r}\left\|\varepsilon_{l_{\kappa_r},+r_{\kappa_r}} - \varepsilon_{l_{\kappa_r},-r_{\kappa_r}}\right\|^2 - \frac{r_{\kappa_r}}{r}\sigma^2 p\right| \le C'_N\left(\sqrt{p\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right) \ ,$$

for C'_N a large absolute constant. To upper bound the second term, we use the following lemma :

Lemma 5.10.18. *For all (l,r) and (l',r') in J_n, set*

$$\xi_{\Delta,v}^{(\mathrm{d})}(l,r,l',r') = \left\{\left|\left\|\varepsilon_{l',+r'} - \varepsilon_{l',-r'}\right\|^2 - \left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2(r'-r)\sigma^2 p\right| \le C_\Delta\sqrt{\frac{rn}{2^v}}\left(\sqrt{p\log\left(2^v\delta^{-1}\right)} + \log\left(2^v\delta^{-1}\right)\right)\right\} \ .$$

There exists a constant C_Δ such that, for all n, the event

$$\xi_\Delta^{(\mathrm{d})} = \bigcap_{v\ge 0}\left\{\xi_{\Delta,v}^{(\mathrm{d})}(l,r,l',r') \ \text{ holds for all } ((l,r),(l',r')) \in N_v \times N_{v+1} \ \text{ s.t. } |l-l'| + |r-r'| \le 3\frac{n}{2^v}\right\} \ .$$

holds with probability higher than $1-\delta$.

For $v \geq \kappa_r$, $((l_v, r_v), (l_{v+1}, r_{v+1})) \in N_v \times N_{v+1}$ and by Lemma 5.10.15,

$$\begin{aligned}|r_v - r_{v+1}| + |l_v - l_{v+1}| &\leq |r_v - r| + |l_v - l| + |r - r_{v+1}| + |l - l_{v+1}| \\ &\leq 3\frac{n}{2^v}.\end{aligned}$$

Therefore, on the event $\xi_\Delta^{(\mathrm{d})}$,

$$\begin{aligned}&\left|\frac{1}{2r}\sum_{v=\kappa_r}^{\kappa_1-1}\left[\|\varepsilon_{l_{v+1},+r_{v+1}} - \varepsilon_{l_{v+1},-r_{v+1}}\|^2 - \|\varepsilon_{l_v,+r_v} - \varepsilon_{l_v,-r_v}\|^2 - 2(r_{v+1} - r_v)\sigma^2 p\right]\right| \\ &\leq C_\Delta \frac{1}{2r}\sum_{v=\kappa_r}^{\kappa_1-1}\sqrt{\frac{r_v n}{2^v}}\left(\sqrt{p\log\left(2^v\delta^{-1}\right)} + \log\left(2^v\delta^{-1}\right)\right) \\ &\leq C'_\Delta \sum_{v'\geq 0}\frac{1}{2^{v'}}\left(\sqrt{p\log\left(\frac{n2^{v'}}{r\delta}\right)} + \log\left(\frac{n2^{v'}}{r\delta}\right)\right) \\ &\leq C''_\Delta\left(\sqrt{p\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right),\end{aligned}$$

where C'_Δ, C''_Δ are large absolute constants. Hence, letting $\bar{c}_{\mathrm{conc}} = C'_N + C''_\Delta$ we obtain

$$\xi_N^{(\mathrm{d})} \cap \xi_\Delta^{(\mathrm{d})} \subset \xi_2^{(\mathrm{d})} \quad ,$$

which must be of probability higher than $1 - 2\delta$. □

Proof of Lemma 5.10.15. Since the mesh of the grid N_u is equal to $2^{\kappa_1 - u} \leq \frac{n}{2^u}$, there exists $(\tilde{l}, \tilde{r}) \in N_u$ such that

$$|l - \tilde{l}| \leq \frac{n}{2^u} \quad \text{and} \quad |r - \tilde{r}| \leq \frac{n}{2^u} \quad .$$

□

Proof of Lemma 5.10.16. Let us write

$$\varepsilon^T A\,\varepsilon = \sum_{l-r\leq i,j<l+r} a_{ij}\langle\varepsilon_i, \varepsilon_j\rangle \quad \text{and} \quad \varepsilon^T B\,\varepsilon = \sum_{m_1\leq i,j<m_2} b_{ij}\langle\varepsilon_i, \varepsilon_j\rangle,$$

where $m_1 = \min(l - r, l' - r')$, $m_2 = \max(l + r, l' + r')$. Remark that for all i, j in $[l - r, l + r)$, $a_{ij} \leq 2$. This gives the first inequality.
For the second inequality, assume without loss of generality that $l \leq l'$. As for the first inequality, $b_{ij} \leq 2$ for all $i, j \in [m_1, m_2)$. Remark that b_{ij} can be non zero only if (i, j) is in one of the following cases :

1. i or j is in $[\min(l + r, l' + r'), \max(l + r, l' + r'))$
2. i or j is in $[\min(l - r, l' - r'), \max(l - r, l' - r'))$
3. i or j is in $[l, l')$.

Hence there is at most $(4(|l - l'| + |r - r'|) + 2|l - l'|)(r + r' + |l - l'|)$ non zero b_{ij}, and we obtain the second inequality. □

Proof of Lemma 5.10.17. The probability of $(\xi_N^{(\mathrm{d})})^c$ can be written as :

$$\mathbb{P}\left((\xi_N^{(\mathrm{d})})^c\right) = \mathbb{P}\Bigg(\exists u \geq 0, \exists (l, r) \in N_u \quad \text{s.t.} \quad r \leq 3\frac{n}{2^u} \text{ and}$$

$$\left|\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\|^2 - 2r\sigma^2 p\right| \leq C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right)\Bigg) \quad .$$

First, fix $u \geq 0$ and $(l, r) \in N_u$ such that $r \leq 3\frac{n}{2^u}$.

Applying the first inequality of Lemma 5.10.16 and the Hanson-Wright inequality - see Lemma 5.10.14, we obtain for all $t \geq 0$

$$\mathbb{P}\left(\left|\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\|^2 - 2r\sigma^2 p\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{pr^2}, \frac{t}{r}\right)\right) \quad ,$$

where c is an absolute constant. Choosing

$$t = C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right) \ ,$$

we obtain

$$\mathbb{P}\left(\left|\left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2r\sigma^2 p\right| \geq C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right)\right) \leq C\left(\frac{1}{2^u}\right)^{cC_N}\delta^{cC_N} \ ,$$

where c, C are absolute constants. Since the cardinal of N_u is upper bounded by 2^{2u+2}, A union bound on each N_u for each $u \geq 0$ gives :

$$\begin{aligned}\mathbb{P}\left((\xi_N^{(\mathrm{d})})^c\right) &\leq \sum_{u\geq 0} C\left|N_u\right|\left(\frac{1}{2^u}\right)^{cC_N}\delta^{cC_N} \\ &\leq \sum_{u\geq 0} 4C\left(\frac{1}{2^u}\right)^{2-cC_N}\delta^{cC_N} \ ,\end{aligned}$$

which is convergent. For C_N large enough, we obtain $\mathbb{P}\left(\xi_N^c\right) \leq 1-\delta$. □

Proof of Lemma 5.10.18.

$$\mathbb{P}\left((\xi_\Delta^{(\mathrm{d})})^c\right) = \mathbb{P}\left(\exists v \geq 0, \exists((l,r),(l',r')) \in N_v \times N_{v+1} \text{ s.t. } |l-l'| + |r-r'| \leq 4\frac{n}{2^v} \text{ and } (\xi_{\Delta,v}^{(\mathrm{d})}(l,r,l',r'))^c \text{ holds}\right) \ .$$

First fix $v \geq 0$ and $((l,r),(l',r')) \in N_v \times N_{v+1}$. Remark that by definition of N_v,

$$r \geq \frac{n}{2^{v+1}} \ .$$

Thus,

$$r + r' + |l-l'| \leq 2r + |l-l'| + |r-r'| \leq 10r \ .$$

Then by Lemma 5.10.16, letting B be the matrix such that $\varepsilon^T B\,\varepsilon = \left\|\varepsilon_{l',+r'} - \varepsilon_{l',-r'}\right\|^2 - \left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2$, we obtain

$$\|B\|^2 \leq \|B\|_F^2 \leq 40r\frac{n}{2^v} \ .$$

Thus, by the Hanson Wright inequality - see Lemma 5.10.14,

$$\mathbb{P}\left(\left|\varepsilon^T B_u\,\varepsilon - \mathbb{E}\left[\varepsilon^T B_u\,\varepsilon\right]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{2^v}{pnr}t^2, \sqrt{\frac{2^v}{nr}}t\right)\right) \ .$$

From now on, we choose

$$t = C_\Delta\sqrt{\frac{rn}{2^v}}\left(\sqrt{p\log\left(2^v\delta^{-1}\right)} + \log\left(2^v\delta^{-1}\right)\right) \ .$$

There are at most 2^{4v+6} elements in $N_v \times N_{v+1}$. Therefore, a union bound on $v \geq 0$ and $N_v \times N_{v+1}$ gives

$$\begin{aligned}\mathbb{P}\left((\xi_\Delta^{(\mathrm{d})})^c\right) &\leq \sum_{u\geq 0} 2|N_v \times N_{v+1}|\left(2^v\right)^{-cC_\Delta}\delta^{cC_\Delta} \\ &\leq \sum_{u\geq 0} 2^7\left(2^v\right)^{4-cC_\Delta}\delta^{cC_\Delta} \\ &\leq C\delta^{cC_\Delta},\end{aligned}$$

where the last inequality holds if C_Δ is large enough, for c, C universal constants. □

5.10.3.2 Proof of Proposition 5.4.2

Step 1: Introduction of useful high probability events. Let $s \leq p$ and consider $S \in \bar{C}_p^s$. In what follows and for an vector $u \in \mathbb{R}^p$, we write $u^{(S)}$ for the vector u restricted to the set S.

Remark that by a simple computation, the noise can be decomposed as follows :

$$\begin{aligned}&\frac{r}{2}\left[\left\|\bar{y}_{l,+r}^{(S)} - \bar{y}_{l,-r}^{(S)}\right\|^2 - \left\|\bar{\theta}_{l,-r}^{(S)} - \bar{\theta}_{l,+r}^{(S)}\right\|^2\right] - \sigma^2 s \\ &= r\langle \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}, \bar{\theta}_{l,+r}^{(S)} - \bar{\theta}_{l,-r}^{(S)}\rangle + \frac{r}{2}\left\|\bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}\right\|^2 - \sigma^2 s \ .\end{aligned}$$

The first term written as

$$r\langle \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}, \bar{\theta}_{l,+r}^{(S)} - \bar{\theta}_{l,-r}^{(S)}\rangle \ ,$$

is a crossed term between the noise and the mean vector θ. Lemma 5.10.12 states that for l equal to a true change-point τ_k, r of order r_k^*, and S being the corresponding support of the change-point, it is controlled on event $\xi_1^{(\mathrm{p})}$ with high probability.

Lemma 5.10.19. *For $k \in [K]$, let us write $S_k \subset [K]$ for the support of $\mu_k - \mu_{k-1}$. Assume that c_0 is a large enough universal constant. The event*

$$\xi_1^{(\mathrm{p})} := \xi_1^{(\mathrm{p})}(\delta) = \Bigg\{ \forall k \in [K] \text{ s.t. Equation (5.18) holds for } k,$$
$$\bar{r}_k^{(\mathrm{d})} \left|\langle \bar{\varepsilon}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \bar{\varepsilon}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}, \bar{\theta}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}\rangle\right| \leq \frac{\bar{r}_k^{(\mathrm{d})}}{4}\left\|\bar{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}\right\|^2 \Bigg\}$$

holds with probability higher than $1-\delta$.

The proof of this lemma follows directly from the one of Lemma 5.10.12, restricting the term corresponding to change-point k to S_k - and diminishing the deviation by doing so.

The second term written as

$$\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}\right\|^2 - \sigma^2 s$$

is a term of pure noise. Lemma 5.10.20 states that it is controlled on event $\xi_2^{(\mathrm{p})}(S)$ with high probability.

Lemma 5.10.20. *There exists a constant $\bar{c}_{\mathrm{conc}} > 0$ such that the event*

$$\xi_2^{(\mathrm{p})}(S) := \xi_2^{(\mathrm{p})}(S,\delta) = \Bigg\{ \forall (l,r) \in J_n, \ \left|\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}\right\|^2 - \sigma^2 s\right|$$
$$\leq \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{s\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right)\Bigg\}$$

holds with probability higher than $1-2\delta$.

The proof of this lemma is exactly the same as the one of Lemma 5.10.13, restricting all vectors to S.

Set $\delta_s = \delta/(2^s\binom{p}{s})$. Lemma 5.10.20 implies that with probability larger than $1-2\delta$, $\forall (l,r) \in J_n$, $\forall S \subset [p]$

$$\left|\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}\right\|^2 - \sigma^2 s\right| \leq \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{s\log\left(\frac{n}{r\delta_s}\right)} + \log\left(\frac{n}{r\delta_s}\right)\right).$$

And so since $\binom{p}{s} \leq \left(\frac{ep}{s}\right)^s$, we have probability larger than $1-2\delta$, $\forall (l,r) \in J_n$, $\forall S \subset [p]$

$$\begin{aligned}\left|\frac{r}{2}\left\|\bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}\right\|^2 - \sigma^2 s\right| &\leq \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{s\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)} + \log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right) \\ &\leq 4\bar{c}_{\mathrm{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right).\end{aligned}$$

And so the event

$$\xi_2^{(\mathrm{p})} := \xi_2^{(\mathrm{p})}(\delta) = \left\{ \forall (l,r) \in J_n, \forall S \subset [p], \left| \frac{r}{2} \left\| \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)} \right\|^2 - \sigma^2 s \right| \right. \\ \left. \leq 4\bar{c}_{\mathrm{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s \log\left(\frac{2ep}{s}\right) \right) \right\} \tag{5.39}$$

has probability larger than $1 - 2\delta$.

Set now

$$\xi^{(\mathrm{p})} := \xi_1^{(\mathrm{p})} \cap \xi_2^{(\mathrm{p})}.$$

Note that

$$\mathbb{P}(\xi^{(\mathrm{p})}) \geq 1 - 3\delta.$$

Step 2: Study in the 'no change-point' situation. Consider $(l,r) \in J_n$ such that $\{\tau_k, k \in [K]\} \cap [l-r, l+r) = \varnothing$, and $S \subset [p]$. Note that since $\{\tau_k, k \in [K]\} \cap [l-r, l+r) = \varnothing$, we have $\bar{\theta}_{l,-r}^{(S)} = \bar{\theta}_{l,+r}^{(S)}$ so that

$$\frac{r}{2} \left\| \bar{\theta}_{l,-r}^{(S)} - \bar{\theta}_{l,+r}^{(S)} \right\|^2 = 0,$$

and

$$r \langle \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}, \bar{\theta}_{l,+r}^{(S)} - \bar{\theta}_{l,-r}^{(S)} \rangle = 0.$$

Moreover we have on $\xi^{(\mathrm{p})}$ that - see Equation (5.39)

$$\left| \frac{r}{2} \left\| \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)} \right\|^2 - \sigma^2 s \right| \leq 4\bar{c}_{\mathrm{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s \log\left(\frac{2ep}{s}\right) \right) \leq \sigma^2 x_r^{(\mathrm{p})},$$

for $\bar{c}_{\mathrm{thresh}} \geq 4\bar{c}_{\mathrm{conc}}$ - note that $\bar{c}_{\mathrm{conc}} > 0$ is a universal constant. And so

$$\Psi_{l,r}^{(\mathrm{p})} \leq x_r^{(\mathrm{p})},$$

so that on $\xi^{(\mathrm{d})}$,

$$T_{l,r}^{(\mathrm{p})} = 0 \ .$$

This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in [K]$ such that τ_k is a κ-sparse high-energy change-point, - see Equation (5.18). Since S_k is the support of $\mu_k - \mu_{k-1}$ - and therefore of $\overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} - \overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}$ - we have

$$\frac{\bar{r}_k^{(\mathrm{s})}}{2} \left\| \bar{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}^{(S_k)} \right\|^2 \geq \frac{\kappa}{8} L^2 \left(s_k \log\left(\frac{2ep}{s_k}\right) + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right) \right) \ .$$

So on $\xi^{(\mathrm{p})}$ this implies that - see Lemma 5.10.19

$$\bar{r}_k^{(\mathrm{d})} \left| \langle \bar{\varepsilon}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\varepsilon}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}^{(S_k)}, \bar{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}^{(S_k)} \rangle \right| \leq \frac{\bar{r}_k^{(\mathrm{s})}}{4} \left\| \overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} \right\|^2.$$

Moreover we have on $\xi^{(\mathrm{p})}$ that - see Equation (5.39)

$$\left| \frac{\bar{r}_k^{(\mathrm{s})}}{2} \left\| \bar{\varepsilon}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\varepsilon}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}^{(S_k)} \right\|^2 - \sigma^2 s \right| \leq 4\bar{c}_{\mathrm{conc}} L^2 \left(\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right) + 2s_k \log\left(\frac{2ep}{s_k}\right) \right) \leq \sigma^2 x_{\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{p})} \ ,$$

for $\bar{c}_{\mathrm{thresh}} \geq 4\bar{c}_{\mathrm{conc}}$ - note that $\bar{c}_{\mathrm{conc}} > 0$ is a universal constant. And so on $\xi^{(\mathrm{p})}$, combining the three previous displayed equations implies

$$\Psi_{\tau_k,\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{p})} \geq \frac{\bar{r}_k^{(\mathrm{d})}}{4\sigma^2} \left\| \bar{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}^{(S_k)} \right\|^2 - x_{\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{p})} \\ \geq \left(\frac{\kappa}{16} - \bar{c}_{\mathrm{thresh}} \right) \frac{L^2}{\sigma^2} \left(\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right) + s_k \log\left(\frac{2ep}{s_k}\right) \right) > x_{\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{p})} \ ,$$

since $\kappa > 32\bar{c}_{\mathrm{thresh}}$. And so on $\xi^{(\mathrm{p})}$

$$T_{\tau_k,\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{p})} = 1.$$

This concludes the proof of the second part of the proposition.

5.10.3.3 Proof of Corollary 5.4.4

Let $\xi^{(\mathrm{d})}$ and $\xi^{(\mathrm{s})}$ be two events such that Proposition 5.4.1 and Proposition 5.4.2 both hold with probability $1-3\delta$, and write $\xi = \xi^{(\mathrm{d})} \cap \xi^{(\mathrm{p})}$. From now on, we work on the event ξ, which holds with probability $1-6\delta$. Define here simply $\bar{\tau}_k = \tau_k$. Note that by definition of $\bar{r}_k$ in the sub-Gaussian regime:

$$\bar{r}_k = \begin{cases} \bar{r}_k^{(\mathrm{d})} \text{ if } s_k \log\left(\frac{ep}{s_k}\right) > \sqrt{p \log\left(\frac{n}{r_k \delta}\right)} \\ \bar{r}_k^{(\mathrm{s})} \text{ if } s_k \log\left(\frac{ep}{s_k}\right) \leq \sqrt{p \log\left(\frac{n}{r_k \delta}\right)} \end{cases}$$

According to Theorem 5.2.1, it is sufficient to prove that $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ holds.

1. **(No false positive):** $T_{l,r} = T_{l,r}^{(\mathrm{p})} \vee T_{l,r}^{(\mathrm{d})} = 0$ for any $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$. by Proposition 5.4.1 and Proposition 5.4.2.

2. **(Significant change-point detection):** for every $k \in \mathcal{K}^*$ (see (5.20)), we have by definition of $\bar{r}_k$:

$$4(\bar{r}_k - 1) \leq r_k.$$

   Now if $s_k \log\left(\frac{ep}{s_k}\right) \geq \sqrt{p \log\left(\frac{n}{r_k\delta}\right)}$, we have $T^{(\mathrm{d})}_{\bar{\tau}_k, \bar{r}_k} = 1$ by Proposition 5.4.1, by definition of c_0, and for $\bar{c}^{(\mathrm{d})}_{\mathrm{thresh}}$ as in Proposition 5.4.1.

   If $s_k \log\left(\frac{ep}{s_k}\right) \leq \sqrt{p \log\left(\frac{n}{r_k\delta}\right)}$, we have $T^{(\mathrm{p})}_{\bar{\tau}_k, \bar{r}_k} = 1$ by Proposition 5.4.2, by definition of c_0, and for $\bar{c}^{(\mathrm{p})}_{\mathrm{thresh}}$ as in Proposition 5.4.2.

Theorem 5.2.1 ensures that for all $k \in \mathcal{K}^*$, there exists $k' \in [\hat{K}]$ such that

$$|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1.$$

This concludes the proof since $4(\bar{r}_k - 1) \leq r_k$ for $k \in \mathcal{K}^*$.

5.10.4 Proof of Theorem 5.5.1

Let us fix $(r,s) \in [1, n/4] \times [1,p]$. Let Δ be such that

$$r\Delta^2 = \frac{1}{2}\sigma^2 \left[s \log\left(1 + u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right) + u \log\left(\frac{n}{r}\right)\right],$$

for some $u \leq \frac{1}{8}$.

In what follows, we consider any change-point detection method that outputs an estimator $\hat{\tau}$ of the change-points, associated to a number $\hat{K}$ of detected change-points, i.e. the length of $\hat{\tau}$. We also write $\mathbb{P}_\Theta$ for the distribution of the data when the mean parameter or the time series is fixed to a $n \times p$ matrix Θ, i.e. of $\Theta + \varepsilon$ where the noise entries $(\varepsilon_t)_j$ are i.i.d. and follow $\mathcal{N}(0, \sigma^2)$ as in Section 5.3. Also abusing slightly notations, we write $\mathbb{P}_0$ for the distribution of the data when the parameter is constant and equal to 0.
Consider also any prior π over the set of $n \times p$ matrices Θ such that the number of true change-points over the support of the prior is larger than 1 - i.e. the prior puts mass only on problems where more than one change-point occurs. Let $\bar{\mathbb{P}}_\pi$ be the corresponding distribution of the data, namely the distribution of the matrix of data when the mean parameter of the time series is the random matrix $\tilde{\Theta} \sim \pi$. Otherwise said, $\bar{\mathbb{P}}_\pi$ is the distribution of $\tilde{\Theta} + \varepsilon$ where $\tilde{\Theta} \sim \pi$.

We remind that in our setting K is the number of true change-points in a given problem - which would be either 0 under $\mathbb{P}_0$, or more than 1 under $\bar{\mathbb{P}}_\pi$. If the support of π_1 is included in $\mathcal{P}(r,s)$, then

$$\begin{aligned} \sup_{\Theta \in \mathcal{P}(r,s)} \mathbb{P}_\Theta(\hat{K} \neq K) &\geq \frac{1}{2}\left(\bar{\mathbb{P}}_\pi(\hat{K} = 0) + \mathbb{P}_0(\hat{K} \neq 0)\right) \\ &\geq \frac{1}{2}(1 - d_{TV}(\bar{\mathbb{P}}_\pi, \mathbb{P}_0)), \end{aligned} \tag{5.40}$$

where d_{TV} is the total variation distance. From the Cauchy-Schwarz inequality, we have

$$d_{TV}(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) \leq \frac{1}{2}\sqrt{\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0)}, \tag{5.41}$$

where χ^2 is the divergence between probability distributions:

$$\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) = \mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\mathrm{d}\bar{\mathbb{P}}_\pi}{\mathrm{d}\,\mathbb{P}_0} - 1\right)^2\right] .$$

By a simple computation that can be found for example in [104]

$$\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta},\tilde{\Theta}'}\left[e^{\frac{1}{\sigma^2}\langle\tilde{\Theta},\tilde{\Theta}'\rangle}\right] - 1, \tag{5.42}$$

where $\tilde{\Theta}$ and $\tilde{\Theta}'$ are i.i.d. and distributed according to π, $\langle\tilde{\Theta},\tilde{\Theta}'\rangle = \mathrm{Tr}(\tilde{\Theta}'\tilde{\Theta}^T)$ is the standard scalar product, and $\mathbb{E}_{\tilde{\Theta},\tilde{\Theta}'}$ is the expectation according to $\tilde{\Theta}$ and $\tilde{\Theta}'$.

Let us consider the three following cases for the couple (r,s):

$$\textbf{Case 1}: u\log\left(\frac{n}{r}\right) \le s\log\left(1 + u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right) \quad \text{and} \quad s \le u\sqrt{p\log\left(\frac{n}{r}\right)},$$

$$\textbf{Case 2}: u\log\left(\frac{n}{r}\right) \le s\log\left(1 + u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right) \quad \text{and} \quad s > u\sqrt{p\log\left(\frac{n}{r}\right)},$$

$$\textbf{Case 3}: u\log\left(\frac{n}{r}\right) > s\log\left(1 + u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right).$$

Each case corresponds to the regime of detection of one of the three statistics. The first one corresponds to the Berk-Jones statistic, the second one to the dense statistic and the last one to the partial norm statistic.

Case 1 : In that case, $r\Delta^2 \le \sigma^2 s\log\left(4u\frac{p}{s^2}\log\left(\frac{n}{r}\right)\right)$. Let us define a probability distribution on the parameter $\Theta \in \mathcal{P}(r,s)$. For $\zeta = \left\lfloor\frac{n}{r}\right\rfloor - 1$ and $l \in \tilde{\mathcal{D}}_r = \{1, r+1, 2r+1, \dots \zeta r + 1\}$, define the column vector $v_l = \sum_{j=l}^{l+r-1} e_j$, where e_j is the j^{th} element of the canonical basis of $\mathbb{R}^n$. Let a be a random variable uniformly distributed in $\{x \in \{0,1\}^p, |x|_0 = s\}$ and ν be a random variable independent from a and uniformly distributed on $\{v_l \,:\, l \in \tilde{\mathcal{D}}_r\}$. Let

$$\tilde{\Theta}_{(1)} = \frac{\Delta}{\sqrt{s}} a\nu^T \in \mathbb{R}^{p\times n},$$

and π_1 be the distribution of the random variable $\tilde{\Theta}_{(1)}$, and $\bar{\mathbb{P}}_{\pi_1}$ be the corresponding distribution of the data.

Consider two independent copies $\tilde{\Theta}_{(1)}$ and $\tilde{\Theta}'_{(1)}$ that are distributed like π_1. The probability that $\tilde{\Theta}_{(1)}$ and $\tilde{\Theta}'_{(1)}$ have the same support is exactly $\frac{1}{\zeta+1}$. Hence, from Equation (5.42)

$$\chi^2(\bar{\mathbb{P}}_{\pi_1}, \mathbb{P}_0) = \frac{1}{\zeta+1}\left(\mathbb{E}_{a,a'}\left[e^{\frac{r\Delta^2}{s\sigma^2}\langle a,a'\rangle} - 1\right]\right) , \tag{5.43}$$

where a' is an independent copy of a, and $\mathbb{E}_{a,a'}$ is the expectation according to a, a'. Remark by symmetry that $\langle a, a'\rangle$ has the same law as $\sum_{i=1}^s a_i$. Hence

$$\mathbb{E}_{a,a'}\left[e^{\frac{r\Delta^2}{s\sigma^2}\langle a,a'\rangle}\right] = \mathbb{E}_a\left[e^{\frac{r\Delta^2}{s\sigma^2}\sum_{i=1}^s a_i}\right] ,$$

where $\mathbb{E}_a$ is the expectation according to a.

Remark that $(a_1, \dots, a_p)$ has the same distribution as a random sampling without replacement of the list of length p containing $(1, \dots, 1, 0, \dots, 0)$ - the list containing exactly s times the quantity 1 and otherwise only 0. The following lemma allows us to replace the variables a_i by independent Bernoulli random variables $Z_i \sim \mathcal{B}(s/p)$.

Lemma 5.10.21. *Let $c = (c_1, \dots, c_p) \in \mathbb{R}^p$. We associate to the list c two random sampling processes: (i) the sampling process without replacement $(X_i)_{i=1\dots s}$ of s elements uniformly on the list c and (ii) the sampling process with replacement $(Z_i)_{i=1\dots s}$ of s elements uniformly in the list. Then for any convex function f,*

$$\mathbb{E}\left[f\left(\sum_{i=1}^s X_i\right)\right] \le \mathbb{E}\left[f\left(\sum_{i=1}^s Z_i\right)\right] .$$

The proof of this lemma can be found in [45]. Thus, if $(Z_i)_{i=1\dots s}$ is an i.i.d sequence of Bernoulli variables with parameter $\frac{s}{p}$ as described above, we obtain

$$\chi^2(\bar{\mathbb{P}}_{\pi_1}, \mathbb{P}_0) \leq \frac{1}{\zeta+1}\left(\mathbb{E}_Z\left[e^{\frac{r\Delta^2}{s\sigma^2}\sum_{i=1}^{s} Z_i}\right] - 1\right) \tag{5.44}$$

$$= \frac{1}{\zeta+1}\left[\left(\frac{s}{p}e^{\frac{r\Delta^2}{s\sigma^2}} + 1 - \frac{s}{p}\right)^s - 1\right] \leq \frac{1}{\zeta+1}\left[e^{\frac{s^2}{p}\left(e^{\frac{r\Delta^2}{s\sigma^2}}-1\right)} - 1\right]$$

$$\leq 2\frac{r}{n}e^{\frac{s^2}{p}\left(e^{\log\left(4u^2\frac{p}{s^2}\log\left(\frac{n}{r}\right)\right)}\right)} \leq 2\left(\frac{r}{n}\right)^{1-4u^2} \leq 1 \ , \tag{5.45}$$

where $\mathbb{E}_Z$ is the expectation according to the $(Z_i)_i$ and where in the last inequality we used $u \leq 1/3$ and $n \geq 4r$.

Case 2 : In that case, $r\Delta^2 \leq \sigma^2 u\sqrt{p\log\left(\frac{n}{r}\right)}$. Let $s_0 = \left\lceil u\sqrt{p\log\left(\frac{n}{r}\right)}\right\rceil$ and b be a random variable uniformly distributed in $\{x \in \{0,1\}^p, |x|_0 = s_0\}$ and ν be defined as in **Case 1**. Let

$$\tilde{\Theta}_{(2)} = \frac{\Delta}{\sqrt{p}} b\nu^T,$$

let π_2 be the distribution of $\tilde{\Theta}_{(2)}$ and $\bar{\mathbb{P}}_{\pi_2}$ be the associated probability distribution of the data. Doing the same reasoning and similar computations as for **Case 1**, see in particular the steps of Equations (5.43) and (5.44) - replacing s by s_0 and a by b - we have

$$\chi^2(\bar{\mathbb{P}}_{\pi_2}, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta}_{(2)},\tilde{\Theta}'_{(2)}}\left[e^{\frac{1}{\sigma^2}\langle\tilde{\Theta}_{(2)},\tilde{\Theta}'_{(2)}\rangle}\right] - 1 = \frac{1}{\zeta+1}\mathbb{E}_{b,b'}\left[e^{\frac{r\Delta^2}{p\sigma^2}\langle b,b'\rangle} - 1\right] \leq \frac{1}{\zeta+1}\left[e^{\frac{s_0^2}{p}\left(e^{\frac{r\Delta^2}{s_0\sigma^2}}-1\right)} - 1\right]$$

$$\leq \frac{1}{\zeta+1}e^{2\frac{s_0 r\Delta^2}{p\sigma^2}} \leq 2\frac{r}{n}e^{4u\log\frac{n}{r}} = 2\left(\frac{r}{n}\right)^{1-4u} \leq 1 \ , \tag{5.46}$$

where $\mathbb{E}_{\tilde{\Theta}_{(2)},\tilde{\Theta}'_{(2)}}$ is the expectation according to $\tilde{\Theta}_{(2)}, \tilde{\Theta}'_{(2)}$ (where $\tilde{\Theta}'_{(2)}$ is an independent copy of $\tilde{\Theta}_{(2)}$) and where $\mathbb{E}_{b,b'}$ is the expectation according to b, b' (where b' is an independent copy of b), and where in the last step we used $u \leq 1/8$ and $n \geq 4r$.

Case 3 : In that case, $r\Delta^2 \leq u\log\left(\frac{n}{r}\right)$. Let $c = (1, 0, 0, \dots, 0)$ be the vector with 0 entries except the first one. Let ν be the random vector defined as in **Case 1**. Let

$$\tilde{\Theta}_{(3)} = \Delta c\nu^T,$$

and π_3 be the distribution of the random variable $\tilde{\Theta}_{(3)}$ - and $\bar{\mathbb{P}}_{\pi_3}$ be the associated probability distribution of the data. Doing the same reasoning as in **Case 1** - see in particular the step of Equation (5.43) - replacing a by c and s by 1 - for the prior π_3, we obtain

$$\chi^2(\bar{\mathbb{P}}_{\pi_3}, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta}_{(3)},\tilde{\Theta}'_{(3)}}\left[e^{\frac{1}{\sigma^2}\langle\tilde{\Theta}_{(3)},\tilde{\Theta}'_{(3)}\rangle}\right] - 1 = \frac{1}{\zeta+1}e^{\frac{r\Delta^2}{\sigma^2}} \leq 2\frac{r}{n}e^{u\log\left(\frac{n}{r}\right)} \leq 2\left(\frac{r}{n}\right)^{1-u} \leq 1 \ , \tag{5.47}$$

where $\mathbb{E}_{\tilde{\Theta}_{(3)},\tilde{\Theta}'_{(3)}}$ is the expectation according to $\tilde{\Theta}_{(3)}, \tilde{\Theta}'_{(3)}$ (where $\tilde{\Theta}'_{(3)}$ is an independent copy of $\tilde{\Theta}_{(3)}$) and where in the last step we used $n \geq 4r$ and $u \leq 1/2$.

Thus, in all cases - combining Equations (5.40) and (5.41) with Equations (5.45), (5.46) and (5.47) - we obtain in all three cases

$$\sup_{\Theta\in\mathcal{P}(r,s)} \mathbb{P}_\Theta(\hat{K} \neq K) \geq \frac{1}{4} \ .$$

and this concludes the proof.

5.10.5 Proofs for covariance and nonparametric change-point detection

Proof of Proposition 5.6.1. Consider an r-sample $(z_1, \dots z_r)$ with mean zero and covariance matrix Σ and Orlicz norm B. Koltchinskii and Lounici [50] have proved that, for any $x > 0$, the empirical covariance matrix $\widehat{\Sigma} = r^{-1}(\sum_{i=1}^r z_i z_i^T)$ satistifies

$$\|\widehat{\Sigma} - \Sigma\|_{op} \le c' B^2 \left[\sqrt{\frac{p}{r}} + \frac{p}{r} + \sqrt{\frac{x}{r}} + \frac{x}{r} \right] ,$$

with probability higher than $1 - \exp(-x)$. Here c' is a suitable positive constant. Considering a union bound over all $(l, r) \in \mathcal{G}_D$ such that Σ_t is constant over $[l - r, l + r)$, we have, with probability higher than $1 - \delta/2$, that simultaneously on all such $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$,

$$\|\widehat{\Sigma}_{l,r} - \widehat{\Sigma}_{l,-r}\|_{op} \le \|\widehat{\Sigma}_{l,r} - \Sigma_l\|_{op} + \|\widehat{\Sigma}_{l,-r} - \Sigma_l\|_{op} \le 8c' B^2 \left[\sqrt{\frac{p}{r}} + \frac{p}{r} + \sqrt{\frac{\log(2n/(r\delta))}{r}} + \frac{\log(2n/(r\delta))}{r} \right] ,$$

where the constant 8 comes from the union bound on all elements of the grid. As a consequence, the FWER of the multiple testing collection is at most $\delta/2$ provided that we choose $c_0 \le 8c'$.

Conversely, consider any high-energy change-point τ_k. Let $\overline{r}_k$ be the smallest radius $r \in \mathcal{R}$ such that

$$r \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op}^2 \ge 0.25 c_1 B^4 \left[\left(p + \log\left(\frac{2n}{r\delta} \right) \right) \wedge r \right] . \tag{5.48}$$

and consider the closest location $l \in \mathcal{D}_r$ of τ_k so that $|l - \tau_k| \le r/2$. To ease the notation, we still write r for $\overline{r}_k$. Without loss of generality, we assume that $l \ge \tau_k$. Let us decompose the statistic $\widehat{\Sigma}_{l,-r} = \frac{r-l+\tau_k}{r} \widehat{\Sigma}_{\tau_k,-(r-l+\tau_k)} + \frac{l-\tau_k}{r} \widehat{\Sigma}_{l,-(l-\tau_k)}$. Since $r \le r_k/2$, Σ_t is constant over $[l - r, \tau_k)$ and over $[\tau_k, l + r)$. Then, we apply three times the deviation inequality of Koltchinskii and Lounici [50] to get

$$\begin{aligned} \|\widehat{\Sigma}_{l,r} - \widehat{\Sigma}_{l,-r}\|_{op} \quad &\ge \quad \frac{r - l + \tau_k}{r} \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} - \|\widehat{\Sigma}_{l,r} - \Sigma_{\tau_k}\|_{op} \\ &\quad - \frac{l - \tau_k}{r} \|\widehat{\Sigma}_{l,-(l-\tau_k)} - \Sigma_{\tau_k}\|_{op} - \frac{r - l + \tau_k}{r} \|\widehat{\Sigma}_{\tau_k,-(r-l+\tau_k)} - \Sigma_{\tau_{k-1}}\|_{op} \\ &\ge \quad \frac{1}{2} \|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} - c'' B^2 \left[\sqrt{\frac{p}{r}} + \frac{p}{r} + \sqrt{\frac{\log(2n/(r\delta))}{r}} + \frac{\log(2n/(r\delta))}{r} \right] , \end{aligned}$$

with probability higher than $1 - 0.5\delta[r/(2n)]^2$. As a consequence, we have $T_{l,r} = 1$ provided that

$$\|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} \ge 2(c'' + c_0) B^2 \left[\sqrt{\frac{p}{r}} + \frac{p}{r} + \sqrt{\frac{\log(2n/(r\delta))}{r}} + \frac{\log(2n/(r\delta))}{r} \right] .$$

Since $\|\Sigma_{\tau_k} - \Sigma_{\tau_{k-1}}\|_{op} \le 2B^2$ and if we choose $c_1 \ge 17 \vee 32(c'' + c_0)$, the bound (5.48) is achievable only if $r \ge p + \log(2n/(r\delta))$ and we deduce from (5.48) that $T_{l,r} = 1$.

Taking a union bound over all high-energy change-points, we deduce from Theorem 5.2.1 that, with probability higher than $1 - \delta$, $\widehat{\tau}$ achieves (**NoSp**) and **detects** all high-energy change-points. Besides, the localization error (5.25) is a consequence of the definition (5.48) together with Theorem 5.2.1. □

Proof of Proposition 5.6.2. As in the proof of Theorem 5.5.1, we only consider a specific setting where one aims at testing $K = 0$ with $\Sigma_1 = I_p$ versus $K = 2$ with $\tau_1 \in (n/4; 3n/4)$, $\tau_2 = \tau_1 + r$, $\Sigma_1 = \Sigma_{\tau_2} = I_p$ and $\Sigma_{\tau_1} = I_p + \zeta uu^T$ for some unknown unit vector u in $\mathbb{R}^p$. Obviously, we have $r_1 = r_2 = r$ and $\|\Sigma_{\tau_1} - \Sigma_{\tau_0}\|_{op} = \|\Sigma_{\tau_2} - \Sigma_{\tau_1}\|_{op} = \zeta$ so that it suffices to prove that the sum of the type I and type II error probabilities of any test of these hypotheses is bounded away from zero. We consider two subcases:

Case 1: $\zeta \le c' \sqrt{p/r} \wedge \frac{1}{\sqrt{2}}$. Then, we focus on the specific alternative hypothesis where $\tau_1 = \lfloor n/2 \rfloor$ and $\tau_2 = \tau_1 + r$, so that the problem reduces exactly to testing whether the covariance matrix Σ of a r-sample satisfies $\Sigma = I_p$ or whether $\Sigma = I_p + \zeta uu^T$. This hypothesis testing problem for covariance matrices is well understood. In particular, one can deduce from Theorem 5.1 in [9] that, as soon as $\zeta \le c'[\sqrt{p/r} \wedge 1]$, for some c' sufficiently small, one has

$$\inf_{\hat{\tau}} \sup_{\Theta \in \overline{\mathcal{P}}(r,\zeta)} \mathbb{P}_\Theta(\hat{K} \ne K) \ge \frac{1}{4} .$$

Case 2: $\zeta \le c' \sqrt{\log(n/r)/r} \wedge 1/\sqrt{2}$. Here, we consider another specific class of alternative hypotheses where we fix $u = (1, 0, \dots, 0)$ but τ_1 can take different values, i.e. $\tau_1 \in \{\lfloor n/4 \rfloor, \lfloor n/4 \rfloor + r, \dots, \lfloor n/4 \rfloor + r\lfloor n/2r \rfloor\}$. It turns out that this is equivalent to a univariate variance testing problem where one observes $q = \lfloor n/(2r) \rfloor$ samples of size r

with distributions $\mathcal{N}(0,\sigma_1^2),\dots,\mathcal{N}(0,\sigma_q^2)$. Under the null, we have $\sigma_1=\sigma_2=\ldots=\sigma_q=1$. Under the alternative, for some $j\in[q]$, we have $\sigma_j=\sqrt{1+\zeta}$ and $\sigma_l=1$ for $l\neq j$. For $j=1,\dots,q$, write $\mathbb{P}_j$ for the distributions of the j-th sample of size r when $\sigma_j^2=1+\zeta$ and $\sigma_l=1$ for $l\neq j$. Besides, we write L_j for the corresponding likelihood ratio with the null distribution $\mathbb{P}_0$. Then, the mixture distribution is defined as $\overline{\mathbb{P}}=\frac{1}{q}\sum_{j=1}^q\mathbb{P}_j$ whereas $\overline{L}$ stands for the mean likelihood ratio. Following the classical path of Le Cam's method we obtain that, for any test T,

$$\mathbb{P}_0[T=1]+\sup_{j=1,\dots,q}\mathbb{P}_j[T=0]\geq\mathbb{P}_0[T=1]+\overline{\mathbb{P}}[T=0]\geq 1-\|\mathbb{P}_0-\overline{\mathbb{P}}\|_{TV}\ ,$$

where $\|.\|_{TV}$ is the total variation norm. Using Cauchy-Schwarz inequality, we bound this total variation distance between the covariates

$$\|\mathbb{P}_0-\overline{\mathbb{P}}\|_{TV}\leq\mathbb{E}_0\left[\overline{L}^2\right]-1=\frac{1}{q}\left(\mathbb{E}_0\left[L_i^2\right]-1\right)=\frac{1}{q}\left[(1-\zeta^2)^{-r/2}-1\right]\leq\frac{1}{q}\left[e^{r\zeta^2}-1\right]\ ,$$

since $\zeta\in(0,1/2)$. As a consequence, we derive that $\|\mathbb{P}_0-\overline{\mathbb{P}}\|_{TV}\leq 1/4$ as long as $r\zeta^2\leq c'\log(q)\wedge 1$. The result follows. □

Proof of Proposition 5.6.3. The proof is based on an application of Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [10] together with an union bound. For a q sample of a univariate distribution with empirical distribution function $\widehat{F}$ and true distribution function F, DKW inequality ensures that

$$\mathbb{P}\left[\|\widehat{F}-F\|_\infty\geq\sqrt{\frac{x}{2q}}\right]\leq 2e^{-x}.$$

Applying two-times DKW inequality to each statistic $T_{l,r}$ such that no-change-point occurs on $(l-r,l+r)$, we deduce that, setting c_1 sufficiently larger, the FWER of $(T_{l,r})$ is at most $\delta/2$ by summing the probabilities over all scales $r\in\mathcal{R}$ and by a union bound on all $l\in\mathcal{D}_r$.

Turning to the high-energy change points, we consider τ_k satisfying (5.26). Let $\overline{r}_k$ be the smallest radius $r\in\mathcal{R}$ such that

$$r\|F_{\tau_k}-F_{\tau_{k-1}}\|_\infty^2\geq 0.25c_1\frac{\log\left(\frac{n}{r\delta}\right)}{m}\ , \tag{5.49}$$

and consider the closest location $l\in\mathcal{D}_r$ of τ_k so that $|l-\tau_k|\leq r/2$ and $2r\leq r_k$. To ease the notation, we still write r for $\overline{r}_k$. As in the proof of Proposition 5.6.1, we decompose the statistic

$$\sum_{t=l}^{l+r-1}\widehat{F}_t-\sum_{t=l-r}^{l-1}\widehat{F}_t=\sum_{t=l}^{l+r-1}\widehat{F}_t-\sum_{t=l-r}^{\tau_k-1}\widehat{F}_t-\sum_{t=\tau_k}^{l-1}\widehat{F}_t,$$

and apply DKW inequality to each of three sums. Taking the union bound over all possible $T_{l,r}$ we deduce that, with probability higher than $1-\delta/2$

$$r^{-1}\|\sum_{t=l}^{l+r-1}\widehat{F}_t-\sum_{t=l-r}^{l-1}\widehat{F}_t\|_\infty\geq\frac{1}{2}\|F_{\tau_k}-F_{\tau_{k-1}}\|_\infty-c''\sqrt{\frac{\log(4n/r\delta)}{mr}}\ ,$$

so that in view of Condition (5.49) implies that $T_{l,r}=1$. Applying Theorem 5.2.1 allows us to conclude. □

Proof of Proposition 5.6.4. As in the proof of Proposition 5.6.2, we focus on a simpler testing problem. Write U for the cumulative distribution function of the uniform distribution on $[0,1]$, i.e. $U(x)=x$ for any $x\in[0,1]$. Given $\zeta\in(0,1/4)$, we define the cumulative distribution function U_ζ by $U_\zeta(x)=(1+2\zeta)x$ for $x\in[0,1/2]$ and $U_\zeta(x)=(1/2+\zeta)+(1-2\zeta)(x-1/2)$ for $x\in[1/2,1]$. Note that $\|U_\zeta-U\|_\infty=\zeta$.

We focus on a testing problem where, under the null, $F_t=U$ for all $t=1,\dots,n$, whereas under the alternative there exists $\tau_1\in\{\lfloor n/4\rfloor,\lfloor n/4\rfloor+r,\dots,\lfloor n/4\rfloor+(r-1)\lfloor n/(2r)\rfloor\}$ such that $F_t=U_\zeta$ for $t=\tau_1,\dots,\tau_1+r-1$ and $F_t=U$ otherwise. Defining $q=\lfloor n/(2r)\rfloor$, we observe that this amounts to testing whether q samples of size rm are distributed according the null distribution or whether exactly one of them is distributed according to U_ζ. Arguing again in the proof of Proposition 5.6.2, we only need to bound the total variation distance between the distribution $\mathbb{P}_0$ under the null and the mixture distribution $q^{-1}\sum_{j=1}^q\mathbb{P}_j$ of the q possible alternatives - here $\mathbb{P}_0=\otimes_{k=1}^q U^{\otimes(rm)}$ is the distribution of the samples when $F_t=U$ and $\mathcal{P}_j=\left[\otimes_{k=1}^{j-1}U^{\otimes(rm)}\right]\otimes U_\zeta^{\otimes(rm)}\otimes\left[\otimes_{k=j+1}^q U^{\otimes(rm)}\right]$, is for $j\geq 1$ the distribution of the samples when $F_t=U$ except for $t\in[jr,(j+1)r)$, in which case $F_t=U_\zeta$.

Let z be a uniform random variable over $[0,1]$ and w be an independent Bernoulli random variable with parameter $1/2$. Then, one easily checks that $z/2+w/2$ is uniformly distributed on $[0,1]$. If w is a Bernoulli

random variable with parameter $1/2 - 2\zeta$, then one easily checks that the cumulative distribution function of $z/2 + w/2$ is F_ζ. As a consequence, by a standard data-processing inequality [104], one derives that

$$\|\mathbb{P}_0 - q^{-1}\sum_{j=1}^{q}\mathbb{P}_j\|_{TV} \leq \|\widetilde{\mathbb{P}}_0 - q^{-1}\sum_{j=1}^{q}\widetilde{\mathbb{P}}_j\|_{TV} \ ,$$

where under $\widetilde{\mathbb{P}}_0$ one observes q independent Binomial random variables with parameter $(mr, 1/2)$, whereas under $\widetilde{\mathbb{P}}_j$, the j-th observation follows a Binomial distribution with parameter $(mr, 1/2 - 2\zeta)$. Using Cauchy-Schwarz inequality, we upper bound the square of the total variation distance by the χ^2 distance and then compute it explicitly. This leads us to

$$\|\widetilde{\mathbb{P}}_0 - q^{-1}\sum_{j=1}^{q}\widetilde{\mathbb{P}}_j\|_{TV}^2 \leq \frac{1}{q}\left[(1+16\zeta^2)^{rm} - 1\right] \ ,$$

which is smaller than $1/4$ provided that $16rm\zeta^2 \leq \log(q/4+1)$. If we choose c' small enough in the statement of the proposition, this last condition holds and the result follows. □

Chapter 6

Future research

In this thesis, we analyzed problems related to both crowdsourcing and time series, from a minimax point of view. We believe it is possible to build upon our research for future works in both of these topics. In the subsequent sections, we first discuss an extension to crowdsourcing problems where we have unknown labels that we want to recover. Then, we introduce a model for ranking workers in a setting with sequential observations. Finally, we discuss the problem of localization of change-points in time series, where the purpose is to not only detect the change-points, but also to estimate them accurately.

6.1 Estimating labels in crowdsourcing problems

A natural extension of the crowdsourcing models presented in the introduction and detailed in Chapter 4 and Chapter 3 is when we are dealing with unknown labels. In what follows, we discuss a model with two unknown labels, similar to the one proposed by Shah et al. [85], as well as a conjecture we might expect.

Assume that we have n workers on d tasks, where each task consists in finding the true label in $\{-1, 1\}$. In this context, we observe a matrix $Y \in \{-1, 1\}^{n \times d}$ representing the labels given by the workers. Even in the problem with unknown labels, we consider a matrix M such that M_{ik} represents the probability that worker i labels task k correctly. For this problem, we let $x^* \in \{-1, 1\}^d$ be the vector representing the unknown true labels. Assume that M is isotonic up to an unknown permutation π^*, and that the entries of Y are distributed as follows:

$$Y_{ik} = \begin{cases} x_k^*, & \text{with probability } M_{ik}, \\ -x_k^*, & \text{with probability } 1 - M_{ik}. \end{cases}$$

In contrast to the models presented in Chapter 3 and Chapter 4, we do not assume that the true labels x_k^* are known. This situation mirrors many real-world crowdsourcing scenarios, where the main purpose is to recover the unknown labels based on the responses from a set of workers.

In this model, recovering the true label x^* or the true ranking of the worker π^* seems challenging. The problem looks circular: estimating π^* is not possible until having a convenient estimator of the labels. Conversely, a suitable estimator of π^* can arguably provide a better estimator of the labels.

In [85] the authors established the minimax rate of the reconstruction of M, in the case where $n = d$ and where M is bi-isotonic up to two permutations. Despite this, as in the bi-isotonic-2D and SST models, considerable computational-statistical gaps persist in the estimation of π^* or M.

In the isotonic model, we can derive from Chapter 4 that if the true labels were known, we could reconstruct the matrix M in polynomial time at the optimal rate of order $n^{7/6}$, when $n = d$. In fact, we believe that, at least in the isotonic model when $n = d$, there exists a polynomial-time method to estimate π^* and M, achieving the rate of order $n^{7/6}$.

We conjecture that this rate is optimal could be proved with two main ingredients: a majority vote in a preprocessing step, and an iterative spectral method. Initially, we would estimate a first set of labels where a majority vote gives high confidence results. Then, the remaining subset of labels would be estimated through a spectral method. Restricting again to labels with low confidence, we would then iteratively repeat this process a polylogarithmic number of times. A spectral method is particularly interesting in this case, primarily because the largest singular value of the population matrix, denoted as $M \operatorname{diag}(x_1^*, \dots, x_d^*)$, does not depend on the true labels $(x_1^*, \dots, x_d^*)$.

In summary, we believe that this crowdsourcing problem with unknown labels is relevant in many practical situations. We also conjecture that it is possible to improve the current convergence rates given by [85] with

a polynomial-time method. In particular, we believe that we could build upon the findings from Chapter 4 on the isotonic model to provide an optimal method to rank the workers when labels are unknown.

6.2 Online ranking problems

In the isotonic and bi-isotonic-1D models presented in Chapter 4, Chapter 3 respectively, we have direct access to the observation from all the pairs of workers/task. Nevertheless, From a practical point of view, it is often reasonable to assume that we can sequentially choose each pair of worker/task. Let us consider an online variant of the isotonic model, where we have a fixed budget T that we can allocate among the workers instead of a matrix of observation. Similarly to the model studied in Saad et al. [79], we sequentially choose T pairs (i, k) of task/worker. At each step, we get an independent Bernoulli observation $y_t \in \{0, 1\}$ with unknown parameter M_{ik}, where $M \in [0, 1]^{n\times d}$ is an isotonic matrix up to some permutation π^*. Specifically, the new pair (i, k) can be chosen given the past information $(y_{t'})_{t'\leq t-1}$. Similarly to Chapter 4 and Chapter 3, the problem is to find an estimator of π^*.

Without any further assumption in this model, the same worker/task pair can be selected multiple times, each time with independent noise. In particular, the model studied in [79] does not impose any constraint on the number of observations that can be requested for a given pair (i, k). However, in many real-world scenarios, the observations for the same pair are often strongly correlated. For example, if you ask someone the same question 10 times, you will likely receive the same answer each time. An idea to address this issue is to assume that we can observe each pair (i, k) at most N_{ik} times, where N_{ik} follows a Poisson distribution with parameter $\lambda \leq 1$.

In this context, a natural idea to estimate π^* is as follows. Assume that the budget T is of order λnd. Then, randomly sample λnd pairs (i, k) among the nd possible entries in $[n] \times [d]$. We get a batch of sample corresponding to a sampling effort of order λ, as in Chapter 4. Finally, apply the algorithm described for the isotonic model – in Chapter 4 – to obtain some estimator of π^* or M. In the worst case, we could show that this estimator achieves the same guarantees as in Chapter 4 for the risk $\mathbb{E}[\|M_{\hat{\pi}^{-1}} - M_{\pi^{*-1}}\|_F^2]$ with parameters n, d and λ. However, we strongly believe that better guarantees can be achieved in this sequential setting.

The main flaw of the aforementioned method is that it does use past observations to refine the choice of the next pair (i, k). As a consequence, accurately estimating the permutations in this model is not a mere application of the procedure given for the isotonic model in Chapter 4. We hope that using the properties of active sampling would improve the rate of convergence. Let us informally describe a procedure that could be better suited for a sequential setting.

First, we use only the $T/2$ first observations to derive an initial estimator by applying the method of Chapter 4. In particular, we are left with sets of workers $P \subset [n]$ that we cannot compare with high confidence, and with subsets $Q \subset [d]$ of tasks that are relevant for the comparison between workers of these groups. Then, the remaining $T/2$ observations could be randomly allocated in the subsets of the form $P \times Q$. Finally, we compare the averages of the workers on the subsets Q to further refine the estimator. In particular, if the total number of entries corresponding to the $P \times Q$ subsets is significantly smaller than nd, then this idea is likely to lead to a better convergence rate than a random allocation across the entire $n \times d$ matrix.

Overall, a straightforward consequence of Chapter 4 is that we can estimate the permutation π^* with a convergence rate of $\mathcal{R}^*_{\text{perm}}(n, d, T/(nd))$, using a random allocation of observations. This corresponds to the case where $\lambda = T/nd$ in Chapter 4. An interesting direction of research would be to use the properties of the online setting to improve the convergence rate of permutation estimation, and to adapt the estimation to a given budget T.

6.3 Change-point localization in time series

In Chapter 5, we established the minimal conditions under which we can consistently detect K change-points in a high-dimensional time series. Beyond simply detecting these change-points, another crucial problem is the localization of the change-points. Take, for instance, a univariate time series with a single possible change-point τ. Arguably, we are not only interested in knowing whether the change-point τ exists, but also in an accurate estimator $\hat{\tau}$ of τ (if it exists).

In the context of a univariate time series with piecewise constant mean (i.e. with multiple change-points), Verzelen et al. [92] established the optimal rates for both detection and localization. In particular, the authors show that the distance of an optimal estimator $\hat{\tau}$ to the true change-point is of order $1/\Delta^2$, where Δ is the difference between the means before and after the change-point in absolute value. In simpler terms, an optimal estimator $\hat{\tau}$ of τ achieves $|\hat{\tau} - \tau| \lesssim 1/\Delta^2$ with high probability.

Consider the high-dimensional setting described in Chapter 5, and let $D_k = \theta_{\tau_k} - \theta_{\tau_k - 1} \in \mathbb{R}^p$. For the sake of simplicity, let us present our conjecture in the case where we do not aim to adapt to the unknown sparsity s_k of D_k, and assume that $s_k = p$.

On the one hand, if $r_k \|D_k\|^2 \geq p$, we could estimate the direction D_k. By projecting the time series on a direction that is close to $D_k / \|D_k\|$, we conjecture that we could reduce the problem to the univariate case and estimate τ_k at a distance of order $1/\|D_k\|^2$. On the other hand, if $r_k \|D_k\|^2 \in [\sqrt{p}, p]$, the idea would be to estimate a direction that is weakly correlated with $D_k / \|D_k\|$. In this case, we conjecture that we could estimate τ_k at a distance of order $p/(r_k \|D_k\|^4)$.

To conclude, improving the localization distance of change-points would be interesting for both practical and theoretical reasons. A worthwhile direction of research would be to look for the precise minimal conditions of change-point localization, and to adapt the analysis to the sparsity s_k of the change-points.

Bibliography

[1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. Journal of the ACM (JACM), 55(5):1–27, 2008.

[2] D. J. Aldous. *Exchangeability and related topics*, École d'été de probabilités de Saint Flour XIII, volume **1117** of Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1985.

[3] S. Arlot, A. Celisse, and Z. Harchaoui. A kernel multiple change-point algorithm via model selection. J. Mach. Learn. Res., 20:Paper No. 162, 56, 2019.

[4] T. P. Ballinger and N. T. Wilcox. Decisions, error and heterogeneity. The Economic Journal, 107(443):1090–1105, 1997.

[5] A. S. Bandeira and R. Van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. 2016.

[6] R. Baranowski, Y. Chen, and P. Fryzlewicz. Narrowest-over-threshold detection of multiple change points and change-point-like features. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(3):649–672, 2019.

[7] P. C. Bellec. Concentration of quadratic forms under a bernstein moment assumption. arXiv preprint arXiv:1901.08736, 2019.

[8] V. Bengs, R. Busa-Fekete, A. El Mesaoudi-Paul, and E. Hüllermeier. Preference-based online learning with dueling bandits: A survey. The Journal of Machine Learning Research, 22(1):278–385, 2021.

[9] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. Annals of Statistics, 41(4):1780–1815, 2013.

[10] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[11] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952.

[12] M. Braverman and E. Mossel. Noisy sorting without resampling. In Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms, pages 268–276, 2008.

[13] M. Braverman and E. Mossel. Sorting from noisy information. arXiv preprint arXiv:0910.1191, 2009.

[14] A. Carpentier and T. Schlüter. Learning relationships between data obtained independently. In Artificial Intelligence and Statistics, pages 658–666. PMLR, 2016.

[15] H. P. Chan, G. Walther, et al. Optimal detection of multi-sample aligned sparse signals. Annals of Statistics, 43(5):1865–1895, 2015.

[16] S. Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1):177–214, 2015.

[17] S. Chatterjee, A. Guntuboyina, and B. Sen. Improved risk bounds in isotonic regression. arXiv preprint arXiv:1311.3765, 2013.

[18] S. Chatterjee, A. Guntuboyina, and B. Sen. On matrix estimation under monotonicity constraints. Bernoulli, 24(2):1072–1100, 2018.

[19] S. Chatterjee and S. Mukherjee. Estimation in tournaments and graphs under monotonicity constraints. IEEE Transactions on Information Theory, 65(6):3525–3539, 2019.

[20] P. Chen, C. Gao, and A. Y. Zhang. Optimal full ranking from pairwise comparisons. The Annals of Statistics, 50(3):1775–1805, 2022.

[21] P. Chen, C. Gao, and A. Y. Zhang. Partial recovery for top-k ranking: optimality of mle and suboptimality of the spectral method. The Annals of Statistics, 50(3):1618–1652, 2022.

[22] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized mle are both optimal for top-k ranking. Annals of statistics, 47(4):2204, 2019.

[23] H. Cho and C. Kirch. Data segmentation algorithms: Univariate mean change and beyond. Econometrics and Statistics, 2021.

[24] L. Chu and H. Chen. Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. The Annals of Statistics, 47(1):382–414, 2019.

[25] O. Collier and A. S. Dalalyan. Minimax rates in permutation estimation for feature matching. The Journal of Machine Learning Research, 17(1):162–192, 2016.

[26] M. Csorgo and L. Horváth. Limit theorems in change-point analysis. John Wiley & Sons Chichester, 1997.

[27] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1):20–28, 1979.

[28] H. Dette, J. Gösmann, et al. Relevant change points in high dimensional time series. Electronic Journal of Statistics, 12(2):2578–2636, 2018.

[29] H. Dette, G. Pan, and Q. Yang. Estimating a change point in a sequence of very high-dimensional covariance matrices. Journal of the American Statistical Association, pages 1–11, 2020.

[30] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist., 32(3):962–994, 2004.

[31] F. Enikeeva and Z. Harchaoui. High-dimensional change-point detection under sparse alternatives. Ann. Statist., 47(4):2051–2079, 2019.

[32] J. Fan, Z. Lou, W. Wang, and M. Yu. Ranking inferences based on the top choice of multiway comparisons. arXiv preprint arXiv:2211.11957, 2022.

[33] N. Flammarion, C. Mao, and P. Rigollet. Optimal rates of statistical seriation. Bernoulli, 25(1):623–653, 2019.

[34] K. Frick, A. Munk, and H. Sieling. Multiscale change point inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(3):495–580, 2014.

[35] F. Friedrich, A. Kempe, V. Liebscher, and G. Winkler. Complexity penalized m-estimation: fast computation. Journal of Computational and Graphical Statistics, 17(1):201–224, 2008.

[36] P. Fryzlewicz. Wild binary segmentation for multiple change-point detection. The Annals of Statistics, 42(6):2243–2281, 2014.

[37] P. Fryzlewicz. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. The Annals of Statistics, 46(6B):3390–3421, 2018.

[38] C. Gao, F. Han, and C.-H. Zhang. On estimation of isotonic piecewise constant signals. Ann. Statist., 48(2):629–654, 2020.

[39] C. Gao, Y. Shen, and A. Y. Zhang. Uncertainty quantification in the bradley–terry–luce model. Information and Inference: A Journal of the IMA, 12(2):1073–1140, 2023.

[40] D. Garreau and S. Arlot. Consistent change-point detection with kernels. Electron. J. Stat., 12(2):4440–4486, 2018.

[41] A. J. Gibberd and S. Roy. Multiple changepoint estimation in high-dimensional gaussian graphical models. arXiv preprint arXiv:1712.05786, 2017.

[42] C. Giraud. Introduction to high-dimensional statistics. CRC Press, 2021.

[43] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773, 2012.

[44] D. V. Hinkley. Inference about the change-point in a sequence of random variables. 1970.

[45] W. Hoeffding. Probability inequalities for sums of bounded random variables. In The Collected Works of Wassily Hoeffding, pages 409–426. Springer, 1994.

[46] S. Hu, J. Huang, H. Chen, and H. P. Chan. Sparsity likelihood for sparse signal and change-point detection. arXiv preprint arXiv:2105.07137, 2021.

[47] S. Hu, J. Huang, H. Chen, and H. P. Chan. Likelihood scores for sparse signal and change-point detection. IEEE Transactions on Information Theory, 2023.

[48] M. Jirak. Uniform change point tests in high dimension. The Annals of Statistics, 43(6):2451–2483, 2015.

[49] L. Jula Vanegas, M. Behr, and A. Munk. Multiscale quantile segmentation. Journal of the American Statistical Association, pages 1–14, 2021.

[50] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. Bernoulli, 23(1):110–133, 2017.

[51] S. Kovács, H. Li, P. Bühlmann, and A. Munk. Seeded binary segmentation: A general methodology for fast and optimal change point detection. arXiv preprint arXiv:2002.06633, 2020.

[52] S. Kovács, H. Li, L. Haubner, A. Munk, and P. Bühlmann. Optimistic search strategy: Change point detection for large-scale data via adaptive logarithmic queries. arXiv preprint arXiv:2010.10194, 2020.

[53] R. Kyng, A. Rao, and S. Sachdeva. Fast, provable algorithms for isotonic regression in all l_p-norms. Advances in neural information processing systems, 28, 2015.

[54] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. Annals of Statistics, 28(5):1302–1338, 2000.

[55] H. Li, Q. Guo, A. Munk, et al. Multiscale change-point segmentation: Beyond step functions. Electronic Journal of Statistics, 13(2):3254–3296, 2019.

[56] A. Liu and A. Moitra. Better algorithms for estimating non-parametric models in crowd-sourcing and rank aggregation. In Conference on Learning Theory, pages 2780–2829. PMLR, 2020.

[57] H. Liu, C. Gao, and R. J. Samworth. Minimax rates in sparse, high-dimensional changepoint detection. arXiv preprint arXiv:1907.10012, 2019.

[58] R. D. Luce. Individual choice behavior: A theoretical analysis. Courier Corporation, 2012.

[59] C. Mao, A. Pananjady, and M. J. Wainwright. Breaking the $1/\sqrt{n}$ barrier: Faster rates for permutation-based models in polynomial time. In Conference On Learning Theory, pages 2037–2042. PMLR, 2018.

[60] C. Mao, A. Pananjady, and M. J. Wainwright. Towards optimal estimation of bivariate isotonic matrices with unknown permutations. The Annals of Statistics, 48(6):3183–3205, 2020.

[61] C. Mao, J. Weed, and P. Rigollet. Minimax rates and efficient algorithms for noisy sorting. In Algorithmic Learning Theory, pages 821–847. PMLR, 2018.

[62] P. Massart. Concentration inequalities and model selection, volume 6. Springer, 2007.

[63] D. H. McLaughlin and R. D. Luce. Stochastic transitivity and cancellation of preferences between bittersweet solutions. Psychonomic Science, 2(1):89–90, 1965.

[64] L. Mirsky. A dual of dilworth's decomposition theorem. Amer. Math. Monthly, 78(8):876–877, 1971.

[65] A. Moscovich, B. Nadler, C. Spiegelman, et al. On the exact berk-jones statistics and their p-value calculation. Electronic Journal of Statistics, 10(2):2329–2354, 2016.

[66] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. Advances in neural information processing systems, 25, 2012.

[67] A. Nemirovskiy. Nonparametric estimation of smooth regression function. Soviet Journal of Computer and Systems Sciences, 23(6):1–11, 1985.

[68] Y. S. Niu, N. Hao, and H. Zhang. Multiple change-point detection: A selective overview. Statistical Science, 31(4):611–623, 2016.

[69] A. B. Olshen, E. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. Biostatistics, 5(4):557–572, 2004.

[70] O. H. M. Padilla, Y. Yu, D. Wang, and A. Rinaldo. Optimal nonparametric change point detection and localization. arXiv preprint arXiv:1905.10019, 2019.

[71] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade. Worst-case versus average-case design for estimation from partial pairwise comparisons. The Annals of Statistics, 48(2):1072–1097, 2020.

[72] A. Pananjady and R. J. Samworth. Isotonic regression with unknown permutations: Statistics, computation and adaptation. The Annals of Statistics, 50(1):324–350, 2022.

[73] E. Pilliat, A. Carpentier, and N. Verzelen. Optimal multiple change-point detection for high-dimensional data. Electronic Journal of Statistics, 17(1):1240–1315, 2023.

[74] E. Pilliat, A. Carpentier, and N. Verzelen. Optimal permutation estimation in crowdsourcing problems. The Annals of Statistics, 51(3):935–961, 2023.

[75] D. Pollard. Lecture notes. 2016.

[76] P. Rigollet and J. Weed. Uncoupled isotonic regression via minimum wasserstein deconvolution. Information and Inference: A Journal of the IMA, 8(4):691–717, 2019.

[77] A. Rinaldo, D. Wang, Q. Wen, R. Willett, and Y. Yu. Localizing changes in high-dimensional regression models. In International Conference on Artificial Intelligence and Statistics, pages 2089–2097. PMLR, 2021.

[78] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. Electronic Communications in Probability, 18, 2013.

[79] E. M. Saad, N. Verzelen, and A. Carpentier. Active ranking of experts based on their performances in many tasks. arXiv preprint arXiv:2306.02628, 2023.

[80] A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. Biometrics, pages 507–512, 1974.

[81] N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In Artificial intelligence and statistics, pages 856–865. PMLR, 2015.

[82] N. Shah, S. Balakrishnan, and M. Wainwright. Low permutation-rank matrices: Structural properties and noisy completion. Journal of machine learning research, 2019.

[83] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. IEEE Transactions on Information Theory, 63(2):934–959, 2016.

[84] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. Feeling the bern: Adaptive estimators for bernoulli probabilities of pairwise comparisons. IEEE Transactions on Information Theory, 65(8):4854–4874, 2019.

[85] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. A permutation-based model for crowd labeling: Optimal estimation and robustness. IEEE Transactions on Information Theory, 67(6):4162–4184, 2020.

[86] N. B. Shah and M. J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. The Journal of Machine Learning Research, 18(1):7246–7283, 2017.

[87] L. Thurstone. A law of comparative judgment. Psychological Review, 34(4), 1927.

[88] J. A. Tropp. An Introduction to Matrix Concentration Inequalities. ArXiv e-prints, Jan 2015.

[89] C. Truong, L. Oudre, and N. Vayatis. Selective review of offline change point detection methods. Signal Processing, 167:107299, 2020.

[90] A. B. Tsybakov. Introduction to Nonparametric Estimation. 2008.

[91] R. Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018.

[92] N. Verzelen, M. Fromont, M. Lerasle, and P. Reynaud-Bouret. Optimal change-point detection and localization. arXiv preprint arXiv:2010.11470, 2020.

[93] S. Vigna. Spectral ranking. Network Science, 4(4):433–445, 2016.

[94] M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

[95] A. Wald. Sequential tests of statistical hypotheses. The annals of mathematical statistics, 16(2):117–186, 1945.

[96] D. Wang, Y. Yu, and A. Rinaldo. Optimal covariance change point localization in high dimension. arXiv preprint arXiv:1712.09912, 2017.

[97] D. Wang, Y. Yu, and A. Rinaldo. Optimal change point detection and localization in sparse dynamic networks. arXiv preprint arXiv:1809.09602, 2018.

[98] D. Wang, Y. Yu, and A. Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. Electron. J. Stat., 14(1):1917–1961, 2020.

[99] D. Wang, Y. Yu, A. Rinaldo, and R. Willett. Localizing changes in high-dimensional vector autoregressive processes. arXiv preprint arXiv:1909.06359, 2019.

[100] H. Wang, Y. Wu, J. Xu, and I. Yolou. Random graph matching in geometric models: the case of complete graphs. In Conference on Learning Theory, pages 3441–3488. PMLR, 2022.

[101] R. Wang and X. Shao. Dating the break in high-dimensional data. arXiv preprint arXiv:2002.04115, 2020.

[102] R. Wang, S. Volgushev, and X. Shao. Inference for change points in high dimensional data. arXiv preprint arXiv:1905.08446, 2019.

[103] T. Wang and R. J. Samworth. High dimensional change point estimation via sparse projection. J. R. Stat. Soc. Ser. B. Stat. Methodol., 80(1):57–83, 2018.

[104] Y. Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. Lecture Notes for ECE598YW (UIUC), 16, 2017.

[105] M. Yu and X. Chen. Finite sample change point inference and identification for high-dimensional mean vectors. arXiv preprint arXiv:1711.08747, 83(2):247–270, 2017.

[106] A. R. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic pca: Algorithm, optimality, and applications. The Annals of Statistics, 50(1):53–80, 2022.

[107] C.-H. Zhang. Risk bounds in isotonic regression. The Annals of Statistics, 30(2):528–555, 2002.

Résumé

Cette thèse explore deux domaines en statistique moderne : les problèmes de classement et la détection de points de rupture. Les deux sujets sont étudiés dans le cadre des statistiques en grande dimension, où le nombre de paramètres inconnus peut être supérieur au nombre d'échantillons. Pour gérer cette complexité, nous introduisons des hypothèses spécifiques ou des contraintes de forme dans les modèles.

La première partie de la thèse explore des problèmes de classement, qui impliquent d'ordonner des éléments sur la base d'observations bruitées et partielles. Principalement, nous examinons deux modèles qui visent à retrouver une permutation des lignes d'une matrice ayant des contraintes de forme spécifiques. Plus précisément, nous considérons le modèle monotone où la matrice réordonnée a des colonnes croissantes, et le modèle bi-monotone où à la fois ses colonnes et ses lignes sont croissantes. Pour chacun des modèles, nous développons un algorithme calculable en temps polynomial pour estimer la permutation inconnue, et nous prouvons qu'il atteint des garanties presque optimales.

La deuxième partie se penche sur la détection de points de rupture dans des séries temporelles en grande dimension. Bien que nous considérions un cadre général pour les points de rupture, l'accent principal est mis sur le cas où nous cherchons à détecter des ruptures dans la moyenne d'une séquence de données. Nous établissons les taux optimaux minimax qui s'adaptent à la fois à la parcimonie inconnue de ces points de rupture, et à la distance entre les points de rupture.

Abstract

This thesis explores two areas in modern statistics: ranking problems and change-point detection. Both topics are investigated within the framework of high-dimensional statistics, where the number of unknown parameters can be greater than the number of samples. To manage this complexity, we introduce specific assumptions, or shape constraints, into the models.

The first part of the thesis looks at ranking problems, which involve rearranging items based on noisy and partial observations. We mainly examine two models that aim to recover a permutation of the rows of a matrix that has specific shape constraints. Specifically, we consider the isotonic model where the reordered matrix has nondecreasing columns, and the bi-isotonic model where it has nondecreasing columns and rows. In both models, we develop polynomial-time algorithms to estimate the unknown permutation, and we prove that they achieve nearly optimal guarantees.

The second part delves into detecting multiple change-points in high-dimensional time series. While we consider a general change-point setting, the main focus is on the case where we aim to detect changes in the mean of the data. We establish minimax optimal rates that are adaptive to the unknown sparsity of these changes, and to the distance between the change-points.