

# Rapport EA initiation recherche P1 2024

Yessin Moakher - Paul Le Van Kiem - David Kerriou

Prof : Karim Lounici

## Résumé

Ce rapport de recherche explore l'utilisation des représentations linéaires dans le cadre du méta-apprentissage, en mettant l'accent sur les tâches partageant une structure commune de faible dimension. Le méta-apprentissage, ou « apprendre à apprendre », vise à exploiter les connaissances transférables entre différentes tâches pour permettre une adaptation rapide à des défis inconnus avec un minimum de données. L'étude formalise le problème à l'aide de bandits linéaires contextuels et examine deux méthodes principales pour l'apprentissage des représentations : la régression à faible rang avec régularisation en norme trace et la méthode des moments. Le cadre théorique est soutenu des estimations d'erreurs. Des évaluations expérimentales sont réalisées sur des données simulées et sur le jeu de données MovieLens, illustrant le potentiel de ces approches dans des applications réelles. Les résultats mettent en évidence l'efficacité des méthodes proposées pour exploiter les structures partagées et améliorer les performances d'apprentissage dans des contextes multi-tâches.

## 1 Introduction

La capacité à transférer des connaissances entre différentes tâches constitue un objectif fondamental de l'apprentissage automatique, permettant aux modèles de s'adapter rapidement à de nouveaux défis tout en exploitant l'expérience acquise. Le méta-apprentissage, souvent désigné comme « apprendre à apprendre », vise à répondre à cet enjeu en développant des algorithmes capables d'extraire et de réutiliser des connaissances transférables sur une distribution de tâches. Cette approche repose en grande partie sur l'apprentissage de représentations : l'identification de caractéristiques invariantes aux tâches facilitant une adaptation rapide à des tâches non vues avec des données minimales.

Dans ce rapport, nous étudions l'utilisation des représentations linéaires dans le cadre du méta-apprentissage, en supposant que les tâches partagent une structure commune de faible dimension.

**Organisation et contributions.** Le reste de l'article est organisé comme suit :

- Section 2 présente une description formelle du problème.
- Section 3 présente deux méthodes principales pour l'apprentissage des représentations. On a adapté une preuve inspirée par l'article [8] pour la première méthode [4]. Par ailleurs, on a élaborée une nouvelle preuve sur l'estimation de la représentation pour la deuxième méthode [10].
- Section 4 présente les résultats expérimentaux, à la fois sur des données simulées et des données réelles. Et où, on a utilisé et adapté une méthode d'optimisation [6] avec la trace norme pour la première méthode.
- Section 5 conclusion.

**Notations.** Les notations suivantes sont utilisées tout au long de l'étude :  $\|\cdot\|$  pour la norme d'opérateur,  $\|\cdot\|_F$  pour la norme de Frobenius,  $S^{d-1}$  pour la sphère unité dans  $\mathbb{R}^d$ , et  $\lambda_1(A) \geq \dots \geq \lambda_d(A)$  pour les valeurs propres de  $A \in \mathbb{R}^{d \times d}$  une matrice symétrique réelle. Par ailleurs,  $\tilde{O}_{\mathbb{P}}$  désigne la notation de "grand O probabiliste" à un facteur en log près.

## 2 Enoncé du problème

### 2.1 Bandits linéaires contextuels

Les bandits linéaires contextuels sont un modèle utilisé dans de nombreuses applications, comme les systèmes de recommandation [7], pour apprendre à prendre des décisions optimisées en fonction des informations disponibles sur chaque utilisateur. Par exemple, dans le contexte de la recommandation de vidéos, un *learner* peut disposer de données contextuelles sur l'utilisateur, telles que son âge, sa localisation, et ses vidéos regardées précédemment. Dans ces systèmes, le *learner* doit, à chaque interaction  $n$  avec un utilisateur  $t$ , choisir une action parmi un ensemble d'options disponibles  $D_n = \{x_{n,1}, \dots, x_{n,K}\} \subseteq \mathbb{R}^d$ . L'objectif du learner est de choisir une action (par exemple, recommander une vidéo spécifique) qui maximise la récompense attendue. À chaque étape  $n$ , après avoir recommandé une action  $x_{t,n}$ , il reçoit un feedback (par exemple, si l'utilisateur a effectivement regardé la vidéo ou non) linéaire défini par la relation suivante :

$$y_{t,n} = x_{t,n}^\top w_t^* + \eta_{t,n} \quad (1)$$

où  $w_t^*$  est un vecteur inconnu, et  $\eta_{t,n}$  est un bruit. Le but de l'apprentissage est donc d'estimer  $w^*$  en temps réel, tout en exploitant cette estimation pour maximiser la récompense cumulative.

On pose  $W_T = [w_1^*, \dots, w_T^*]$  la matrice  $d \times T$ , dont les colonnes sont formées par les  $T$  vecteurs de régression, qui sont inconnus. Nous considérons l'hypothèse suivante de faible rang sur la matrice de régression des tâches.

**Hypothèse 1.** La matrice  $W_T \in \mathbb{R}^{d \times T}$  a un rang  $r \ll \min(d, T)$ .

L'hypothèse ci-dessus implique qu'il existe une représentation de faible dimension inconnue partagée par toutes les tâches. Autrement dit, il existe une matrice  $B \in \mathbb{R}^{d \times r}$  et des vecteurs inconnus  $\alpha_t \in \mathbb{R}^r$  tels que  $w_t^* = B\alpha_t$  pour chaque  $t \in \{1, \dots, T\}$ .

### 2.2 Meta learning et solution

Supposons que nous ayons accès à l'historique enregistré de  $T$  tâches déjà complétées  $Z_{T,N} = \{(x_{t,n}, y_{t,n})\}_{n=1, t=1}^{N,T}$  satisfaisant l'équation 1 pour tout  $t \in \{1, \dots, T\}$  et  $n \in \{1, \dots, N\}$ , où la matrice de régression  $W_T$  satisfait l'hypothèse 1. L'apprentissage par méta-apprentissage, cherche à concevoir des algorithmes capables d'utiliser l'expérience passée pour apprendre rapidement de nouvelles compétences. Étant donné le jeu de données  $Z_{T,M}$ , notre objectif est alors d'apprendre une nouvelle tâche  $T+1$  partageant la même représentation inconnue  $w_{T+1}^* = B\alpha_{T+1}$  que les tâches passées, comme indiqué dans l'équation 1,  $y_{T+1,n} = \langle x_{T+1,n}, B\alpha_{T+1} \rangle + \eta_{T+1,n}$ ,  $n \in \{1, \dots, N\}$ .

Si on arrive à estimer  $B$  par  $\hat{B}$ , on estime  $\hat{\alpha}_{T+1,n}$ , à chaque étape  $n \in \{1, \dots, N\}$  par une simple régression linéaire :

$$\hat{\alpha}_{T+1,n} \in \arg \min \sum_{i=1}^{n-1} \left( y_{T+1,i} - \langle x_{T+1,i}, \hat{B}\alpha \rangle \right)^2 \quad (2)$$

Ainsi l'algorithme, de méta-apprentissage pour la nouvelle tâche  $T+1$  est :

---

**Algorithm 1** Algorithmme glouton par méta-apprentissage

---

**Paramètres :** Une estimation  $\hat{B}$

- 1: **À l'étape**  $n = 1$  une action est choisie aléatoirement
  - 2: Observer  $Y_{T+1,1}$
  - 3: **for**  $n \in \{2, \dots, N\}$  **do**
  - 4: Mettre à jour  $\hat{\alpha}_{T+1,n}$  selon 2
  - 5: Choisir  $x_{T+1,n} \in \arg \max_{x \in D_{T+1,n}} \langle x, \hat{B} \hat{\alpha}_{T+1,n} \rangle$
  - 6: Observer la récompense  $Y_{T+1,n}$
  - 7: **end for**
- 

Deux méthodes principales seront étudiées pour l'estimation de  $B$ .

### 3 Apprentissage de la représentation

#### 3.1 Méthode 1 : Régression avec faible rang régularisation puis factorisation

Cette méthode est utilisée dans [4] et [1]. Elle consiste à résoudre le problème de régression suivant sous l'hypothèse 1 pour estimer  $W_T$

$$\widehat{W} = \arg \min_{A \in \mathcal{C}} \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{t,i} - \langle \mathbf{x}_{t,i}, A_t \rangle)^2 + \lambda \|A\|_*, \quad (3)$$

où  $\mathcal{C}$  est un espace convexe.

Le terme de régularisation favorise des solutions de rang faible.

C'est un problème d'optimisation convexe, où on a une somme de termes dérivables et un terme non dérivable  $\|\cdot\|_*$ .

Nous estimons  $B$  à partir de  $\widehat{W}$  de la manière suivante : Nous notons  $\hat{r}$  le rang de  $\widehat{W}$  et considérons la décomposition en valeurs singulières (SVD) de  $\widehat{W} = \widehat{U} \widehat{\Lambda} \widehat{V}^T$ , où  $\sigma_1(\widehat{W}) \geq \dots \geq \sigma_{\hat{r}}(\widehat{W}) > 0$  sont les valeurs singulières non nulles, correspondant aux premières valeurs de la diagonale de  $\widehat{\Lambda}$ . Nous définissons la matrice de représentation empirique

$$\widehat{B} = (\widehat{U}_1, \dots, \widehat{U}_{\hat{r}})$$

Les premières  $\hat{r}$  colonnes de  $\widehat{U}$ , notées  $\widehat{U}_1, \dots, \widehat{U}_{\hat{r}}$ , sont sélectionnées. Par conséquent, il n'est pas nécessaire de connaître à l'avance la valeur de  $r$ , contrairement à la méthode 2.

L'erreur sur l'estimation de  $W$  est donnée par le théorème suivant :

**Théorème 1.** *Sous des conditions détaillées dans la preuve, on a*

$$\frac{1}{T} \|\widehat{W} - W_T\|_F^2 \leq c \left( \frac{rT\lambda^2}{n} + C^2 \Xi_n(u) \right)$$

$$, \text{ Avec } \Xi_n(u) = \frac{u}{nT} + \frac{r(d+T)(d+u+\log(T))}{n^2T}$$

*Démonstration.* La preuve complète se trouve dans l'annexe, Partie A.1. L'idée de la preuve repose sur la condition d'optimalité du sous-différentiel vérifiée par  $\widehat{W}$ .  $\square$

#### 3.2 Méthode 2 : Méthode des moments

Cette méthode est utilisée dans [10]. Elle consiste à calculer le moment d'ordre 2 de la variable aléatoire  $\frac{1}{\sqrt{n}} \sum_{i=1}^n y_i x_i$ .

$$\Sigma = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n y_i^2 x_i x_i^T \right) = 2\Gamma + (1 + \text{tr}(\Gamma)) \text{Id}, \quad \text{où } \Gamma = \frac{1}{n} \sum_{i=1}^n B \alpha_i \alpha_i^T B^T.$$

En utilisant une ACP de dimension  $r$ , le rang de  $B$ , on peut récupérer l'espace des colonnes de  $B$ . En effet, les vecteurs propres de  $\Sigma$  sont les mêmes que les vecteurs propres de  $\Gamma$ , qui sont contenus dans l'espace des colonnes de  $B$ . Et, en supposant que  $(\alpha_i)_{i=1}^n$  engendrent  $\mathbb{R}^d$ . Les colonnes de  $\Gamma$  engendrent alors l'espace des colonnes de  $B$ . En effet, si  $\Gamma x = 0$ , alors  $x^T \Gamma x = 0$ , ce qui implique que  $\langle \alpha_i, B^T x \rangle = 0$ , pour tout  $i \in \{1, \dots, n\}$ , et donc  $B^T x = 0$ .

On pose  $\Sigma_n = \frac{1}{n} \sum_{i=1}^n y_i^2 x_i x_i^T$ .

**Théorème 2.** *On suppose que les  $x_i \in \mathbb{R}^d$  sont i.i.d et suivent une loi normale centrée réduite. On a  $\forall t > 0$ ,*

$$\|\Sigma_n - \Sigma\| = \tilde{O}_{\mathbb{P}}((1 \vee \max_{1 \leq i \leq n} \|w_i\| \vee \max_{1 \leq i \leq n} \|w_i\|^6)(\sqrt{\frac{d}{n}} \vee \frac{d}{n}))$$

*Démonstration.* La preuve complète se trouve dans l'annexe, Partie A.2.

### 3.3 Erreur sur l'estimation de la représentation B

Les deux méthodes ci-dessus, consiste à estimer une grande matrice qu'on appellera ici  $A$  ( $\widehat{W}$  et  $\Sigma$  respectivement dans les méthode 1 et 2). Puis faire une SVD de  $A$  pour extraire les vecteur singulier à droite afin d'estimer B.

Pour calculer l'erreur sur l'estimation de B, on peut sans perte de généralité supposer que  $A$  est symétrique.

En effet, dans le cas contraire, on considère la symétrisée de  $A$ ,  $A_{sym} = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix}$ .

On a les propriétés suivantes :

- $A_{sym} \begin{bmatrix} u_k \\ \pm v_k \end{bmatrix} = \pm \sigma_k \begin{bmatrix} u_k \\ \pm v_k \end{bmatrix}$ , avec les  $u_k$ , les vecteurs singuliers à gauche de  $A$  et les  $v_k$ , les vecteurs singuliers à droite de  $A$ .
- $\|A_{sym}\| = \|A\|$ .

**Théorème 3.** *Soient  $\Sigma, \Sigma_n \in \mathbb{R}^{d \times d}$  deux matrices réelles symétriques, dont les valeurs propres sont respectivement notées  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  et  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_d$ . Soit  $P_{\lambda_j}$  le projecteur spectral sur l'espace propre associé à la valeur propre  $\lambda_j$ , engendré par les vecteurs propres correspondant à  $\{\lambda_i\}_{i=s}^r$ , et  $\widehat{P}_{\lambda_j}$  le projecteur spectral sur l'espace propre engendré par les vecteurs propres correspondant à  $\{\hat{\lambda}_i\}_{i=s}^r$  pour  $\Sigma_n$ . On pose  $\lambda_0 = \infty$  et  $\lambda_{d+1} = -\infty$ . Si*

$$\|\Sigma - \Sigma_n\| < \frac{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}{4},$$

*alors*

$$\|P_{\lambda_j} - \widehat{P}_{\lambda_j}\| \leq \frac{4 \cdot \|\Sigma - \Sigma_n\|}{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}.$$

*Démonstration.* La preuve est dans l'annexe dans la Partie A.3. L'idée de la preuve est d'utiliser la résolvante d'une matrice pour faire un lien avec les matrices de projections spectrales.  $\square$

### 3.4 Erreur sur l'estimation totale du vecteur $B\widehat{\alpha}_{T+1,n}$

Le vecteur  $\widehat{\alpha}_{T+1,n}$  étant défini par la régression linéaire comme indiqué dans l'algorithme, on a le théorème suivant qui permet de faire le lien entre l'erreur de la représentation et l'erreur de l'estimation. Considérons les événements suivant :

$$\begin{aligned} \Omega_0 &= \{r \leq \hat{r} \leq \bar{c}r\} \\ \Omega_1 &= \left\{ \lambda_{\min}(\widehat{B}^T \widehat{\Sigma}_n \widehat{B}) \geq \kappa_1 \right\} \\ \Omega_2 &= \left\{ \left\| \widehat{B}^T \widehat{\Sigma}_n \widehat{B} \right\| \leq \kappa_2 \right\} \end{aligned}$$

où  $\hat{\Sigma}_n = \frac{1}{n} X_n X_n^T$ ,  $\hat{r}$  est le rang de  $\hat{B}$  (égal à  $r$  pour la méthode des moments).

**Théorème 4.** Soit  $\hat{B}$  obtenu par SVD de  $A$  d'ordre  $\hat{r}$ . La procédure d'estimation décrite ci-dessus, si elle vérifie les événements  $\Omega_0, \Omega_1, \Omega_2$  est telle que :

$$\left\| \hat{B} \hat{\alpha}_{T+1,n} - B \alpha_{T+1} \right\| \leq \frac{1}{n} \frac{\left\| \hat{B}^T X_n \eta_n \right\|}{\kappa_1} + 8 \left( 1 + \frac{\kappa_2}{\kappa_1} \right) \|\alpha_{T+1}\| \frac{\left\| \hat{A} - A \right\|}{\sigma_r(A)}$$

*Démonstration.* La démonstration est dans l'annexe A.4. □

## 4 Simulations et résultats

Pour résoudre numériquement<sup>1</sup> le problème 3 de la méthode 1, on utilise l'algorithme proposé dans l'article [6]. L'idée de l'algorithme est d'utiliser le résultat de [3] que

$$D_\lambda(A) = \arg \min_W \frac{1}{2} \|W - A\|_F^2 + \lambda \|W\|_*,$$

où  $D_\lambda(A) = U \Sigma_\lambda V^T$ , et  $\Sigma_\lambda$  est diagonale avec  $(\Sigma_\lambda)_{ii} = \max\{0, \Sigma_{ii} - \lambda\}$ , en linéarisant le problème  $\min_W f(W) + \lambda \|W\|_*$ , avec l'approximation quadratique de  $f$ .

Pour déterminer le pas d'optimisation, calculons la constante de Lipschitz de  $\nabla f$ . Pour cela, on calcule la hessienne de  $f$  :

À partir de l'expression du gradient, nous avons pour chaque  $t$  :

$$\nabla_{A_t} f(A) = -\frac{2}{nT} X_t^\top (Y_t - X_t A_t),$$

où  $X_t$  est de dimension  $(n, d)$ ,  $A_t$  est de dimension  $d$ , et  $Y_t$  est de dimension  $n$ .

On a :

$$H_t = \nabla_{A_t}^2 f(A) = \frac{2}{nT} X_t^\top X_t.$$

Ainsi :

$$L = \max_t \left( \lambda_{\max} \left( \frac{2}{nT} X_t^\top X_t \right) \right) = \frac{2}{nT} \max_t \left( \lambda_{\max} \left( X_t^\top X_t \right) \right).$$

Comme  $\lambda_{\max}(X_t^\top X_t)$  est le carré de la valeur singulière maximale de  $X_t$ , notée  $\sigma_{\max}(X_t)$  :

$$L = \frac{2}{nT} \max_t (\sigma_{\max}(X_t)^2).$$

**Métrique :** On va utiliser le regret cumulé comme métrique. Il est défini par

$$R(T, N) = \sum_{t=1}^T \sum_{n=1}^N (x_{t,n}^* - x_{t,n})^\top w_t^*,$$

avec  $x_{t,n}^* = \arg \max_{x \in D_{t,n}} x^\top w_t^*$ .

**Sélection d'hyperparamètres :** Nous avons utilisé la validation croisée avec  $X_{train}$ . Dans la méthode 1,  $\lambda$  a été testé sur  $[0.001, 0.005, 0.01, 0.05, 0.1]$ , et dans la méthode 2,  $r$  sur  $[1, 2, 3, 4, 5]$ .

**Données simulées :** Les paramètres sont  $T = 100$ ,  $d = 60$ ,  $r = 5$  (rang de la matrice  $W$ ), et  $n = 100$ .  $X$  et  $\eta$  sont des gaussiennes centrées réduites. Les données de test  $X_{test}$  et  $Y_{test}$  suivent

---

1. Le code Python est disponible à l'adresse : <https://github.com/fegounna/Meta-learning-with-contextual-linear-bandits>

une procédure similaire pour  $T_{\text{test}} = 10$  et  $n_{\text{test}} = 50$ , où  $Y_{\text{test}}$  est calculé à partir  $B\alpha_{\text{test}} + \eta_{\text{test}}$ , avec  $\alpha_{\text{test}} \sim \mathcal{N}(0, \frac{1}{r}I_r)$ .

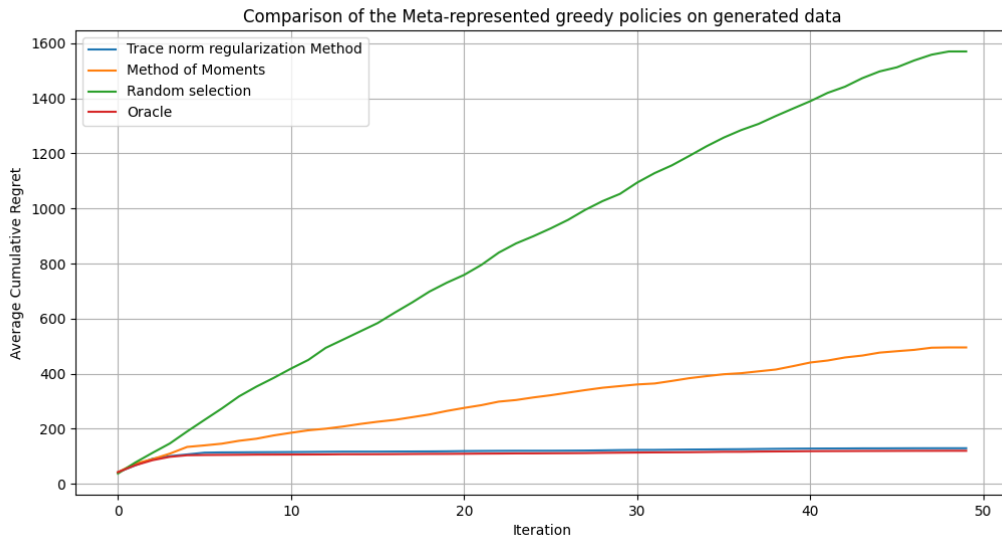


FIGURE 1 – Comparaisons des méthodes avec les données synthétiques

**Données réelles :** Nous allons utiliser la base de données MovieLens [5]. Elle contient des évaluations de films par les utilisateurs, ainsi que des informations contextuelles telles que les genres et les dates de sortie des films. On a gardé les colonnes **movie\_id**, **rating\_date**, **release\_date**, **genre**, **age**, **gender**, **occupation**.

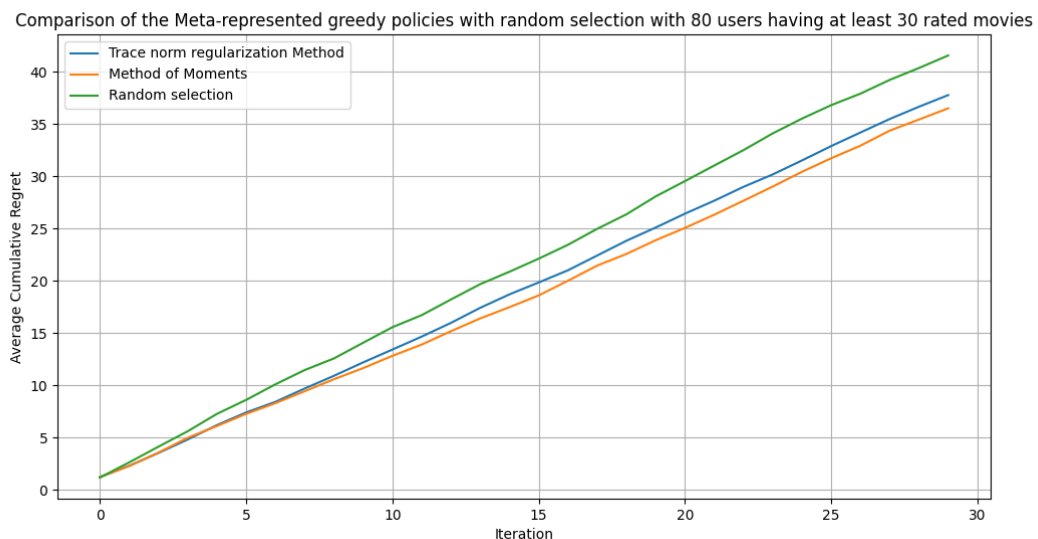


FIGURE 2 – Comparaisons des méthodes avec les données de MovieLens

La simulation avec des données réelles montre clairement une convergence vers une représentation adéquate. Cependant, dans le cas réel, le regret n'est pas sous-linéaire, comme le prédit

la théorie avec des données gaussiennes i.i.d. Toutefois, le gain est non négligeable lorsqu'on l'applique à grande échelle, avec des millions d'utilisateurs.

## 5 Conclusion

Cette étude met en évidence le potentiel du méta-apprentissage pour exploiter efficacement les structures partagées entre les tâches dans le cadre des bandits linéaires contextuels. En utilisant des méthodes de régularisation de rang faible et des techniques basées sur les moments, le rapport fournit une base théorique solide et des perspectives pratiques pour estimer des représentations de tâches transférables. Les résultats expérimentaux valident les conclusions théoriques, montrant une amélioration significative des performances des tâches et une réduction du regret cumulé. Les travaux futurs pourraient explorer des extensions aux contextes non linéaires, renforçant ainsi l'applicabilité du méta-apprentissage dans des scénarios réels.

## Références

- [1] Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1303–1327. PMLR, 02–05 Jul 2022.
- [2] Stephen Boyd. Lecture notes for ee 364b, stanford university, spring 2021-2022 : Subgradients.
- [3] Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm for matrix completion, 2008.
- [4] Leonardo Cella, Karim Lounici, Grégoire Pacreau, and Massimiliano Pontil. Multi-task representation learning with stochastic linear bandits. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 2023.
- [5] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets : History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), December 2015.
- [6] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 457–464, New York, NY, USA, 2009. Association for Computing Machinery.
- [7] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, volume 3 of *WWW '10*, page 661–670. ACM, April 2010.
- [8] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4) :2164 – 2204, 2011.
- [9] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration, 2013.
- [10] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10434–10443. PMLR, 18–24 Jul 2021.
- [11] G.A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170 :33–45, 1992.

## A Annexe

### A.1 Démonstration du théorème 1

**Définition 1.** (*Sous-différentiel*) Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction convexe. Le sous-différentiel de  $f$  en un point  $x_0 \in \mathbb{R}^n$ , noté  $\partial f(x_0)$ , est l'ensemble de tous les vecteurs  $g \in \mathbb{R}^n$  (appelés sous-gradients) qui satisfont l'inégalité suivante pour tout  $x \in \mathbb{R}^n$  :

$$f(x) \geq f(x_0) + \langle g, (x - x_0) \rangle$$

**Lemme 1.** (*Sous-différentiel de la norme de trace*) [11]

$$\partial \|A\|_* = \left\{ \sum_{j=1}^n u_j v_j^\top + P_{S_1^\perp} Q P_{S_2^\perp} \mid \|Q\|_\infty \leq 1 \right\}$$

**Lemme 2.** (*Condition d'optimalité avec contraintes*) [2] Soit  $f$  une fonction admettant au moins un sous-gradient en tous ses points. Un point  $x^* \in X$  minimise de  $f$  sur un ensemble convexe fermé  $X$  si et seulement s'il existe un sous-gradient  $g \in \partial f(x^*)$  tel que l'inégalité

$$\langle g, (y - x^*) \rangle \geq 0$$

est satisfaite pour tout  $y \in X$ .

**Lemme 3.** On a aussi les propriétés suivantes sur les sous-différentiels :

- Supposons que  $f = f_1 + \dots + f_m$ , où  $f_1, \dots, f_m$  sont des fonctions convexes. Alors nous avons  $\partial f(x) = \partial f_1(x) + \dots + \partial f_m(x)$ .
- Si  $f$  est convexe et différentiable en  $x$ , alors  $\partial f(x) = \{\nabla f(x)\}$ .

**Lemme 4.** (*Monotonie des sous-gradients*) Soit  $f$  une fonction admettant au moins un sous-gradient en tous ses points. Pour  $U \in \partial f(u)$  et  $V \in \partial f(v)$ , on a  $\langle V - U, u - v \rangle \geq 0$ .

Pour simplifier les calculs, introduisons  $\Pi_n$  la distribution empirique sur  $\mathbb{R}^{d \times T}$  basée sur  $\{x_{t,i}\}_{i,t=1}^{n,T}$ . Pour tout  $A \in \mathbb{R}^{d \times T}$ , on définit

$$\langle A, B \rangle_{L^2(\Pi_n)} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T \langle x_{t,i}, A_t \rangle \langle x_{t,i}, B_t \rangle.$$

On définit aussi

$$D_n = \sum_{t=1}^T \sum_{i=1}^n \eta_{t,i} x_{t,i} \otimes e_t$$

En réarrangeant les termes, le problème de minimisation (3) devient équivalent au problème de minimisation suivant :

$$\widehat{W} = \arg \min_{A \in \mathcal{C}} \|A\|_{L^2(\Pi)}^2 - \frac{2}{nT} \langle D_n, A \rangle + \lambda \|A\|_*,$$

En substituant la valeur de  $y_{t,i}$  par celle donnée dans l'équation (1), et en ajoutant la constante  $\|W\|_{L^2(\Pi)}$  on obtient :

$$\widehat{W} = \arg \min_{A \in \mathcal{C}} \|A - W\|_{L^2(\Pi)}^2 - \frac{2}{nT} \langle D_n, A \rangle + \lambda \|A\|_*, \quad (4)$$

La condition d'optimalité (lemme 2), s'écrit, en utilisant les lemmes 3 et 1, en : il existe  $\hat{V} \in \partial \|\hat{W}\|_*$  vérifiant :

$$2 \langle \widehat{W} - W, \widehat{W} - A \rangle_{L^2(\Pi_n)} - \frac{2}{nT} \langle D_n, \widehat{W} - A \rangle + \lambda \langle \hat{V}, \widehat{W} - A \rangle \leq 0 \quad (5)$$



pour tout  $A \in \mathcal{C}$ , avec l'existence de  $\hat{Q}$  tel que  $\|\hat{Q}\|_\infty \leq 1$  et  $\hat{V} = \sum_{j=1}^n u_j v_j^\top + P_{s_1^\perp} \hat{Q} P_{s_2^\perp}$ . On a

$$\|V\|_\infty \leq \left\| \sum_{j=1}^n u_j v_j^\top \right\|_\infty + \|P_{s_1^\perp} \hat{Q} P_{s_2^\perp}\|_\infty \leq 2$$

En utilisant le fait que pour deux matrices  $A, B$  on a

$$|\langle A, B \rangle| \leq \|A\|_\infty \|B\|_* \quad (6)$$

Et en supposant que l'évènement  $\Omega_4 = \{\lambda \geq \frac{4}{nT} \|D_n\|_\infty\}$  est satisfait, on obtient pour  $A$  égale à  $W$  que

$$\|\hat{W} - W\|_{L^2(\Pi)} \leq \frac{3}{4} \lambda \|\hat{W} - W\|_* \quad (7)$$

Introduisons  $V \in \partial\|W\|_*$  dans l'équation 5,

$$2\|\widehat{W} - W\|_{L^2(\Pi_n)} - \frac{2}{nT} \langle D_n, \widehat{W} - W \rangle + \lambda \langle V, \widehat{W} - W \rangle + \lambda \langle \hat{V} - V, \widehat{W} - W \rangle \leq 0$$

Puisque  $\langle \hat{V} - V, \widehat{W} - W \rangle \geq 0$  (lemme 4), on a

$$\lambda \langle V, \widehat{W} - W \rangle \leq \frac{2}{nT} \langle D_n, \widehat{W} - W \rangle,$$

avec  $Q$  arbitraire tel que  $\|Q\|_\infty \leq 1$  et  $V = \sum_{j=1}^n u_j v_j^\top + P_{S_1^\perp} Q P_{S_2^\perp}$ . Sur  $\Omega_4$  on a

$$\langle V, \widehat{W} - W \rangle \leq \frac{1}{2} \|\hat{W} - W\|_*, \quad (8)$$

**Lemme 5.** Si  $A$  une matrice réelle carrée, il existe une matrice  $Q$  telle que :

$$\|Q\|_\infty = 1 \text{ et } \langle Q, A \rangle = \|A\|_*$$

*Démonstration.* [8] Par décomposition spectrale, on écrit  $A = U \Lambda V^T$  et on prend  $Q = UV^T$ .  $\square$

En utilisant le lemme 5, on prend  $Q$  tel que  $\langle Q, P_{S_1^T}(\hat{W} - W)P_{S_2^T} \rangle = \|P_{S_1^T}(\hat{W} - W)P_{S_2^T}\|_*$ . Ainsi

$$\langle V, \widehat{W} - W \rangle = \left\langle \sum_{j=1}^n u_j v_j^\top, P_{S_1}(\hat{W} - W)P_{S_2} \right\rangle + \|P_{S_1^T}(\hat{W} - W)P_{S_2^T}\|_*$$

En utilisant les équations (8) et (6) on obtient

$$\|P_{S_1^T}(\hat{W} - W)P_{S_2^T}\|_* \leq \|P_{S_1}(\hat{W} - W)P_{S_2}\|_* + \frac{1}{2} \|P_{S_1^T}(\hat{W} - W)P_{S_2^T}\|_* + \frac{1}{2} \|P_{S_1}(\hat{W} - W)P_{S_2}\|_*$$

Ainsi,

$$\|P_{S_1^T}(\hat{W} - W)P_{S_2^T}\|_* \leq 3 \|P_{S_1}(\hat{W} - W)P_{S_2}\|_*$$

On a montré ainsi, que  $\hat{W} - W \in \mathcal{K}$ . Sous l'évènement

$$\Omega_3 = \bigcap_{A \in \mathcal{K}} \left\{ \|A\|_{L^2(\Pi_n)}^2 \geq \|A\|_F^2 T^{-2} - c \max_{t \in [T]} \|[A]_t\|^2 \Xi_n(u) \right\},$$

on a, en posant  $\Delta = \hat{W} - W$ , que

$$\|\Delta\|_{L^2(\Pi_n)}^2 \geq \|\Delta\|_F^2 T^{-2} - c \max_{t \in [T]} \|\Delta\|_t^2 \Xi_n(u)$$

En utilisant l'équation (7), on a

$$\|\Delta\|_F^2 T^{-2} - c \max_{t \in [T]} \|[\Delta]_t\|^2 \Xi_n(u) \leq \frac{3}{4} \lambda \|\Delta\|_*$$

D'autre part, on a

$$\|\Delta\|_* \leq \|P_{S_1^T} \Delta P_{S_2^T}\|_* + \|P_{S_1} \Delta P_{S_2}\|_* \leq 6 \|P_{S_1} \Delta P_{S_2}\|_*$$

Par Cauchy-Schwarz, on a

$$\|P_{S_1} \Delta P_{S_2}\|_* \leq \sqrt{\text{rang}(P_{S_1} \Delta P_{S_2})} \|P_{S_1} \Delta P_{S_2}\|_F \leq \sqrt{r} \|\Delta\|_F$$

En utilisant l'inégalité  $ab \leq \frac{\epsilon^2 a^2}{2} + \frac{b^2}{2\epsilon^2}$  avec  $\epsilon = \sqrt{2T}$ ,  $b = \|\Delta\|_F$  et  $a = \frac{6.3}{4} \lambda \sqrt{r}$ , on a

$$\|\Delta\|_F^2 T^{-2} - c \max_{t \in [T]} \|[\Delta]_t\|^2 \Xi_n(u) \leq \frac{1}{4T} \|\Delta\|_F^2 + crT\lambda^2$$

En conclusion ,

$$\frac{1}{T} \|\hat{W} - W\|_F^2 \leq c (rT\lambda^2 + C_0^2 \Xi_n(u))$$

## A.2 Démonstration du théorème 2

### A.2.1 Préliminaires

Pour la démonstration du théorème, il convient de rappeler quelques éléments sur la norme d'opérateur des matrices symétriques et de présenter la notion d'  $\epsilon$ -Net qui se révélera très utile pour démontrer le théorème.

**Lemme 6.** Si  $A \in \mathbb{R}^{d \times d}$  est une matrice symétrique, alors

$$\|A\| = \max_{\|x\|_2=1} |\langle Ax, x \rangle|.$$

**Définition 2** ( $\epsilon$ -net). Soit  $(X, d)$  un espace métrique. Soit  $K \subset X$  et  $\epsilon > 0$ . Un ensemble  $\mathcal{N} \subset K$  est un  $\epsilon$ -net sur  $K$  si et seulement si

$$K \subset \bigcup_{x \in \mathcal{N}} B(x, \epsilon).$$

**Proposition 1.** Soit  $A \in \mathbb{R}^{d \times d}$  une matrice symétrique, et soit  $\epsilon \in [0, \frac{1}{2}]$ . Pour tout  $\epsilon$ -net  $\mathcal{N}$  de la sphère unité  $S^{d-1}$ , on a

$$\|A\| \leq \frac{1}{1-2\epsilon} \sup_{x \in \mathcal{N}} |\langle Ax, x \rangle|.$$

*Démonstration.* Selon le lemme 6, il existe  $x \in S^{d-1}$  tel que  $|\langle Ax, x \rangle| = \|Ax\|$ , et d'après la définition, il existe  $x_0 \in \mathcal{N}$  tel que  $\|x - x_0\|_2 \leq \epsilon$ .

Par l'inégalité triangulaire, on obtient :

$$\begin{aligned} \|A\| - |\langle Ax_0, x_0 \rangle| &= |\langle Ax, x \rangle| - |\langle Ax_0, x_0 \rangle| \\ &\leq |\langle Ax, x \rangle - \langle Ax_0, x_0 \rangle| \\ &= |\langle Ax, x - x_0 \rangle + \langle A(x - x_0), x_0 \rangle| \\ &\leq |\langle Ax, x - x_0 \rangle| + |\langle A(x - x_0), x_0 \rangle|. \end{aligned}$$

En utilisant l'inégalité de Cauchy-Schwarz, il vient que :

$$\begin{aligned} \|A\| - |\langle Ax_0, x_0 \rangle| &\leq \|A\|_2 \|x - x_0\|_2 + \|A(x - x_0)\|_2 \|x_0\|_2 \\ &\leq 2\epsilon \|A\|. \end{aligned}$$

Ainsi :

$$\begin{aligned} \|A\| &\leq \frac{1}{1-2\epsilon} |\langle Ax_0, x_0 \rangle| \\ &\leq \frac{1}{1-2\epsilon} \sup_{x \in \mathcal{N}} |\langle Ax, x \rangle|. \end{aligned}$$

□

**Proposition 2.** Pour tout  $\epsilon > 0$ , on peut trouver un  $\epsilon$ -net  $\mathcal{N}$  sur  $S^{d-1}$  tel que son cardinal vérifie

$$|\mathcal{N}| \leq 9^d.$$

**Proposition 3** (Inégalité de Hanson-Wright [9]). Soit  $Z = (Z_1, \dots, Z_m) \in \mathbb{R}^m$  un vecteur aléatoire avec des composantes sous-gaussiennes indépendantes  $Z_i$  qui satisfont  $\mathbb{E}[Z_i] = 0$  et  $\|Z_i\|_{\psi_2}^2 \leq K$ . Soit  $A$  une matrice  $m \times m$ . Alors, pour tout  $t \geq 0$ ,

$$\mathbb{P}\left(|Z^\top AZ - \mathbb{E}[Z^\top AZ]| > t\right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|}\right)\right).$$

Cette inégalité peut se réécrire,

$$\mathbb{P}\left(|Z^\top AZ - \mathbb{E}[Z^\top AZ]| > K^2 \max\left(\frac{\|A\|t}{c}, \|A\|_F \sqrt{\frac{t}{c}}\right)\right) \leq 2 \exp(-t)$$

ou encore en introduisant  $C = \frac{1}{\min(c, \sqrt{c})}$

$$\mathbb{P}\left(|Z^\top AZ - \mathbb{E}[Z^\top AZ]| > CK^2 \max(\|A\|t, \|A\|_F \sqrt{t})\right) \leq 2 \exp(-t)$$

### A.2.2 Démonstration du théorème

*Démonstration.* La preuve consiste en l'utilisation de plusieurs inégalités de concentration. Ré-écrivons  $w_i = B\alpha_i$ . Étant donné que  $y_i = x_i^T w_i + \eta_i$ , on a  $y_i^2 = \eta_i^2 + 2\eta_i x_i^T w_i + x_i^T w_i w_i^T x_i$  et,

$$\begin{aligned} \Sigma_n - \Sigma &= \frac{1}{n} \sum_{i=1}^n \eta_i^2 x_i x_i^T - \mathbb{E}(\eta_i^2 x_i x_i^T) \\ &\quad + \frac{1}{n} \sum_{i=1}^n 2\eta_i x_i^T w_i x_i x_i^T - \mathbb{E}(2\eta_i x_i^T w_i x_i x_i^T) \\ &\quad + \frac{1}{n} \sum_{i=1}^n x_i^T w_i w_i^T x_i x_i x_i^T - \mathbb{E}(x_i^T w_i w_i^T x_i x_i x_i^T) \end{aligned}$$

**Premier terme** Le premier terme se réécrit en

$$\frac{1}{n} \sum_{i=1}^n \eta_i^2 x_i x_i^T - I_d = \frac{1}{n} A_n$$

La matrice  $A_n$  étant symétrique, on peut écrire

$$\|A_n\| = \max_{u \in S^{d-1}} u^T A_n u = \max_{u \in S^{d-1}} \sum_{i=1}^n x_i^T \eta_i^2 u u^T x_i - \mathbb{E}(x_i^T \eta_i^2 u u^T x_i)$$

On peut alors écrire,

$$\|A_n\| = \max_{u \in S^{d-1}} X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X)$$

où  $X$  est le vecteur concaténé des  $x_i$  et  $A(u, \eta)$  est une matrice diagonale par blocs composée des  $\eta_i^2 u u^T$ .

Soit  $u \in S^{d-1}$ , on peut utiliser l'inégalité de Hanson-Wright (proposition 3) conditionnellement aux  $\eta_i$  car les composantes de  $X$  sont gaussiennes et indépendantes,

$$\mathbb{P}(|X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X)| > CK^2 \max(\|A(u, \eta)\|t, \|A(u, \eta)\|_F \sqrt{t}) | \eta) \leq 2 \exp(-t)$$

Par ailleurs, on a  $\|A(u, \eta)\| = \max_i \eta_i^2$  et  $\|A(u, \eta)\|_F = \sqrt{\sum_{i=1}^n \eta_i^4} \leq \sqrt{n} \max_i \eta_i^2$

Ainsi, on a

$$\mathbb{P}(X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X) > CK^2 \max_i \eta_i^2 \max(t, \sqrt{nt}) | \eta) \leq 2 \exp(-t)$$

On utilise les  $\varepsilon$ -Net. On a

$$\|A_n\| \leq C_{\varepsilon, d} \max_{u \in \mathcal{N}} X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X)$$

Ainsi, en posant  $f_\eta(t) = C_{\varepsilon, d} CK^2 \max_i \eta_i^2 \max(t, \sqrt{t})$ , on a

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \|A_n\| > f_\eta\left(\frac{t}{n}\right) | \eta\right) &\leq \mathbb{P}\left(\max_{u \in \mathcal{N}} X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X) > CK^2 \max_i \eta_i^2 \max(t, \sqrt{nt}) | \eta\right) \\ &\leq \sum_{u \in \mathcal{N}} \mathbb{P}(X^T A(u, \eta) X - \mathbb{E}(X^T A(u, \eta) X) > CK^2 \max_i \eta_i^2 \max(t, \sqrt{nt}) | \eta) \\ &\leq 2 \times 9^d \exp(-t) \end{aligned}$$

D'où,

$$\mathbb{P}\left(\frac{1}{n} \|A_n\| > f_\eta(t) | \eta\right) \leq 2 \times 9^d \exp(-nt)$$

Cette inégalité ne tenant que pour  $\eta$  fixé, il nous faut également contrôler  $\max_i \eta_i^2$  sous haute probabilité. Les  $\eta_i^2$  étant des variables aléatoires  $\chi^2$  à une dimension, on a

$$\mathbb{P}(\max_i \eta_i^2 > t_n) \leq 2n \exp(-t_n)$$

Avec  $t_n > 0$ . Ainsi, en posant  $f_n(t) = C_{\varepsilon, d} CK^2 t_n \max(t, \sqrt{t})$ , on a

$$\mathbb{P}\left(\frac{1}{n} \|A_n\| > f_n(t)\right) = \mathbb{E}(\mathbb{E}(1_{\frac{1}{n} \|A_n\| > f_n(t)} 1_{\max_i \eta_i^2 \leq t_n} + 1_{\frac{1}{n} \|A_n\| > f_n(t)} 1_{\max_i \eta_i^2 > t_n} | \eta))$$

Or, si  $\max_i \eta_i^2 \leq t_n$ , alors,  $f_n(t) \geq f_\eta(t)$ , donc

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} \|A_n\| > f_n(t)\right) &\leq \mathbb{E}(\mathbb{E}(1_{\frac{1}{n} \|A_n\| > f_\eta(t)} | \eta) + 1_{\max_i \eta_i^2 > t_n}) \\ &\leq 2 \times 9^d \exp(-nt) + 2n \exp(-t_n) \end{aligned}$$

En prenant  $t_n = 2 \log(n)$  et  $t = O(\frac{d}{n})$ , on obtient  $\frac{1}{n} \|A_n\| = O_{\mathbb{P}}(\log(n)(\sqrt{\frac{d}{n}} \vee \frac{d}{n}))$

**Deuxième terme** Pour simplifier la démonstration, nous considérerons que  $w_1 = \dots = w_n = w$ . Le deuxième terme se réécrit

$$\frac{1}{n} \sum_{i=1}^n 2\eta_i x_i^T w_i x_i x_i^T - \mathbb{E}(2\eta_i x_i^T w_i x_i x_i^T) = \frac{1}{n} \sum_{i=1}^n 2\eta_i x_i^T w_i x_i x_i^T = \frac{1}{n} B_n$$

On a  $\|B_n\| = \max_{u \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n \eta_i x_i^T w(x_i^T u)^2$ . On va conditionner par rapport à  $\eta$ , et donc, de la même façon que dans le premier terme, cela nous fait payer un terme en  $\log(n)$ . On va donc retirer les termes en  $\eta_i$ . Maintenant, il suffit de contrôler le terme  $b_n = \max_{u \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n x_i^T w(x_i^T u)^2$ . Posons également  $A = \{\bigcap_{i=1}^n |x_i^T w| \leq \|w\| \sqrt{4 \log(n)}\}$ . Tout d'abord, on a

$$\mathbb{P}(\bar{A}) \leq \sum_{i=1}^n \mathbb{P}(|x_i^T w| > \|w\| \sqrt{4 \log(n)}) \leq 2n \exp\left(-\frac{\|w\|^2 4 \log(n)}{2\|w\|^2}\right) = \frac{2}{n}$$

De ce fait, on peut utiliser ceci pour montrer que "se placer" sur  $A$  revient à payer un terme en  $\|w\| \log(n)$ . L'astuce consiste maintenant à projeter  $u$  sur la droite  $w$ . Notons  $u_w = w^T w w$  et  $u_w^\perp = u - u_w$  les termes correspondants. L'intérêt est que les termes  $x_i^T w$  et  $x_i^T u_w^\perp$  sont indépendants. On a alors,

$$b_n = \max_{u \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n (u^T w)^2 (x_i^T w)^3 + \frac{1}{n} \sum_{i=1}^n (x_i^T w)(x_i^T u_w^\perp)^2 + \frac{1}{n} \sum_{i=1}^n (u^T w)(x_i^T w)^2 (x_i^T u_w^\perp)$$

Pour le premier terme, il suffit d'utiliser l'inégalité de Bernstein pour montrer qu'il est un  $O_{\mathbb{P}}(\|w\|^5 \sqrt{\log(n) \frac{d}{n}})$ . Pour le deuxième terme, on peut se placer sur l'événement  $A$  pour payer un  $\|w\| \log(n)$  et utiliser l'inégalité de Hanson-Wright pour montrer que c'est un  $O_{\mathbb{P}}(\|w\| \log(n) (\frac{d}{n} \vee \sqrt{\frac{d}{n}}))$ . Enfin, pour le dernier terme, il faut se placer sur un événement similaire à  $A$  (mais avec les  $(x_i^T u_w^\perp)$ ) et payer un  $\log(n)$  cette fois-ci, puis ensuite utiliser Hanson-Wright pour montrer que ce terme est un  $O_{\mathbb{P}}(\|w\|^3 \log(n) (\frac{d}{n} \vee \sqrt{\frac{d}{n}}))$ .

**Dernier terme** Le dernier terme se réécrit

$$\frac{1}{n} \sum_{i=1}^n x_i^T w_i w_i^T x_i x_i x_i^T - \mathbb{E}(x_i^T w_i w_i^T x_i x_i x_i^T)$$

Ce terme se traite d'une façon similaire. □

### A.3 Démonstration du théorème 3

#### A.3.1 Préliminaires

**Proposition 4** (Inégalité de Weyl). *Pour deux matrices symétriques  $A, B$  dans  $\mathbb{R}^{d \times d}$ , nous avons le résultat suivant sur les valeurs propres :*

$$\forall i \in \{1, \dots, d\}, |\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|.$$

**Définition 3** (Résolvante). *Soit  $A$  une matrice carrée  $n \times n$ . La résolvante  $R_A$  de la matrice  $A$  est définie sur  $\mathbb{C} \setminus \text{Sp}(A)$ , par :  $R_A(z) = (A - zI)^{-1}$ .*

**Lemme 7.** *Si  $A \in \mathbb{R}^{d \times d}$  est une matrice symétrique, alors la résolvante de  $A$  s'écrit*

$$R_A(z) = \sum_{j=1}^d \frac{u_j u_j^\top}{\lambda_j(A) - z},$$

où  $\lambda_j(A)$  sont les valeurs propres de  $A$  et  $u_j$  sont les vecteurs propres correspondants.

**Proposition 5.** Si  $A \in \mathbb{R}^{d \times d}$  est une matrice symétrique, dont on note  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  ses valeurs propres, alors le projecteur sur l'espace propre associé à  $\lambda_i$  est donné par :

$$P_{\lambda_i} = \frac{1}{2\pi i} \oint_{\Gamma_i} R_A(\eta) d\eta,$$

où  $\Gamma_i$  est un contour fermé entourant  $\lambda_i$ , mais pas les autres  $\lambda_j$  tels que  $\lambda_i \neq \lambda_j$ .

*Démonstration.* On obtient le résultat en appliquant le théorème de Cauchy sur la formule de la résolvante obtenue dans le lemme 7.  $\square$

**Lemme 8.** Soient  $\Sigma, \Sigma_n \in \mathbb{R}^{d \times d}$  deux matrices réelles symétriques, dont les valeurs propres sont respectivement notées  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$  et  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_d$ . Soient  $\{\lambda_i\}_{i=s}^r$  les valeurs propres de  $\Sigma$  dont les vecteurs propres correspondants engendrent l'espace propre associé à la valeur propre  $\lambda_j$ . On pose  $\lambda_0 = \infty$  et  $\lambda_{d+1} = -\infty$ . Si

$$\|\Sigma - \Sigma_n\| < \frac{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}{2},$$

alors, en posant  $g_{\lambda_j} = \min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)$ , on a :

$$\begin{aligned} \lambda_i &\notin D(\lambda_j, \frac{g_{\lambda_j}}{2}) \text{ pour tout } i \notin \{s, \dots, r\}, \\ \hat{\lambda}_i &\in D(\lambda_j, \frac{g_{\lambda_j}}{2}) \text{ pour tout } i \in \{s, \dots, r\}, \\ \hat{\lambda}_i &\notin D(\lambda_j, \frac{g_{\lambda_j}}{2}) \text{ pour tout } i \notin \{s, \dots, r\}, \end{aligned}$$

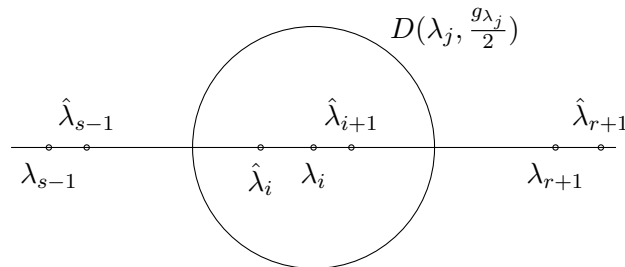
où  $D(\lambda_j, \frac{g_{\lambda_j}}{2})$  désigne le disque ouvert de rayon  $\frac{g_{\lambda_j}}{2}$  centré en  $\lambda_j$ .

*Démonstration.* Prenons  $g_{\lambda_j} = \min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)$  et supposons la condition du lemme vérifiée.

Tout d'abord les valeurs propres étant rangées dans l'ordre décroissant, on vérifie immédiatement que  $\lambda_i \notin D(\lambda_j, \frac{g_{\lambda_j}}{2})$ , pour tout  $i \notin \{s, \dots, r\}$ . D'après l'inégalité de Weyl (Proposition 4), on a aussi que  $\hat{\lambda}_i \in D(\lambda_j, \frac{g_{\lambda_j}}{2})$ , pour tout  $i \in \{s, \dots, r\}$ . De plus, on a, par les mêmes arguments, et dès que  $i \notin \{s, \dots, r\}$  :

$$|\lambda_j - \hat{\lambda}_i| \geq |\lambda_j - \lambda_i| - |\lambda_i - \hat{\lambda}_i| > r - \frac{r}{2} = \frac{r}{2}$$

Ce qui conclut la preuve.  $\square$



Schema de la répartition des valeurs propres donné par le lemme 3.

### A.3.2 Démonstration du théorème

*Démonstration.* Dans la suite, nous noterons

$$g_{\lambda_j} = \min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s),$$

et supposons que  $\|\Sigma - \Sigma_n\| < \frac{g_{\lambda_j}}{4}$ .

Fixons  $z \in \partial D(\lambda_j, \frac{g_{\lambda_j}}{2})$ , le cercle de rayon  $\frac{g_{\lambda_j}}{2}$  centré en  $\lambda_j$  et intéressons-nous à :

$$R_\Sigma(z) - R_{\Sigma_n}(z) = (zI - \Sigma)^{-1} - (zI - \Sigma - (\Sigma_n - \Sigma))^{-1}.$$

Remarquons que

$$|z - \lambda_i| \geq ||\lambda_i - \lambda_j| - |\lambda_j - z|| = \left| |\lambda_i - \lambda_j| - \frac{r}{2} \right| \geq \frac{g_{\lambda_j}}{2}.$$

Ainsi,

$$\|R_\Sigma(z)\| = \max_i \frac{1}{|z - \lambda_i|} = \frac{1}{\min_i |z - \lambda_i|} \leq \frac{2}{g_{\lambda_j}}.$$

Finalement,

$$\|(\Sigma_n - \Sigma) \cdot R_\Sigma(z)\| \leq \|\Sigma_n - \Sigma\| \cdot \|R_\Sigma(z)\| \leq \frac{g_{\lambda_j}}{4} \cdot \frac{2}{g_{\lambda_j}} = \frac{1}{2} < 1.$$

Puisque  $(M_d(\mathbb{R}), \|\cdot\|)$  est une algèbre de Banach unitaire, on peut écrire :

$$R_{\Sigma_n}(z) = (zI - \Sigma - (\Sigma_n - \Sigma))^{-1} = R_\Sigma(z) \cdot \sum_{k=0}^{\infty} ((\Sigma_n - \Sigma) \cdot R_\Sigma(z))^k.$$

Ainsi, on obtient :

$$\begin{aligned} \|R_\Sigma(z) - R_{\Sigma_n}(z)\| &= \left\| (zI - \Sigma)^{-1} \cdot \sum_{k=1}^{\infty} ((\Sigma_n - \Sigma)(zI - \Sigma)^{-1})^k \right\| \\ &\leq \|(zI - \Sigma)^{-1}\| \cdot \sum_{k=1}^{\infty} (\|\Sigma_n - \Sigma\| \cdot \|(zI - \Sigma)^{-1}\|)^k \\ &\leq \frac{\|\Sigma_n - \Sigma\| \cdot \|(zI - \Sigma)^{-1}\|^2}{1 - (\|\Sigma_n - \Sigma\| \cdot \|(zI - \Sigma)^{-1}\|)} \\ &\leq \frac{8 \cdot \|\Sigma_n - \Sigma\|}{g_{\lambda_j}^2}. \end{aligned}$$

En utilisant la proposition 5 et le lemme 8, il vient que :

$$\begin{aligned} \|P_{\lambda_j}(\Sigma) - P_{\hat{\lambda}_j}(\Sigma_n)\| &= \frac{1}{2\pi} \left\| \oint_{\partial D(\lambda_j, \frac{g_{\lambda_j}}{2})} R_\Sigma(z) - R_{\Sigma_n}(z) dz \right\| \\ &\leq \frac{1}{2\pi} \oint_{\partial D(\lambda_j, \frac{g_{\lambda_j}}{2})} \|R_\Sigma(z) - R_{\Sigma_n}(z)\| |dz| \\ &\leq \frac{1}{2\pi} \oint_{\partial D(\lambda_j, \frac{g_{\lambda_j}}{2})} \frac{8 \cdot \|\Sigma_n - \Sigma\|}{g_{\lambda_j}^2} |dz| \\ &= \frac{4 \cdot \|\Sigma_n - \Sigma\|}{\min(\lambda_r - \lambda_{r+1}, \lambda_{s-1} - \lambda_s)}. \end{aligned}$$

□

#### A.4 Démonstration du théorème 4

*Démonstration.* Remarquons d'abord que sur l'évènement  $\Omega_1$  la matrice  $\widehat{B}^T \widehat{\Sigma}_n \widehat{B}$  est inversible. Ainsi l'estimateur des moindres carrés admet une unique solution :

$$\widehat{\alpha}_{n+1} = \frac{1}{n} \left( \widehat{B}^T \widehat{\Sigma}_n \widehat{B} \right)^{-1} \widehat{B}^T X_n Y_n$$

On va ensuite décomposer la différence  $\widehat{B} \widehat{\alpha}_n - B \alpha$  ce qui nous permettra ensuite d'appliquer des inégalités venant des parties précédentes. Notons également les colonnes de  $\widehat{B}$  étant orthonormale  $\widehat{B}^T \widehat{B} = I_{\widehat{r}}$  et ainsi  $\widehat{B} = \widehat{B} \widehat{B}^T \widehat{B} = \widehat{P} \widehat{B}$ .

$$\begin{aligned} \widehat{B} \widehat{\alpha}_n &= \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n Y_n \\ &= \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n (X_n^T B \alpha + \eta_n) \\ &= \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n X_n^T (B \alpha - \widehat{P} B \alpha + \widehat{P} B \alpha) + \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n \eta_n \\ &= \widehat{B} \widehat{B}^T B \alpha + \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n \eta_n \\ &\quad + \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \widehat{B}^T X_n X_n^T (B B^T B \alpha - \widehat{P} B \alpha) \\ &= \widehat{P} B \alpha + \widehat{B} \left( \widehat{B}^T X_n X_n^T \widehat{B} \right)^{-1} \left[ \widehat{B}^T X_n X_n^T (P - \widehat{P}) B \alpha + \widehat{B}^T X_n \eta_n \right] \end{aligned}$$

D'où on obtient que

$$\widehat{B} \widehat{\alpha}_n - B \alpha = (\widehat{P} - P) B \alpha + \widehat{B} \left( \widehat{B}^T \widehat{\Sigma}_n \widehat{B} \right)^{-1} \left[ \widehat{B}^T \widehat{\Sigma}_n (P - \widehat{P}) B \alpha + \frac{1}{n} \widehat{B}^T X_n \eta_n \right]$$

Lorsque les évènements  $\Omega_0, \Omega_1, \Omega_2$  sont vérifiés on a que

$$\left\| \left( \widehat{B}^T \widehat{\Sigma}_n \widehat{B} \right)^{-1} \right\| \leq \frac{1}{\kappa_1} \quad (9)$$

$$\left\| \widehat{B}^T \widehat{\Sigma}_n \overline{P} \right\| \leq \kappa_2 \quad (10)$$

Et de plus comme  $Im(P - \widehat{P}) \subset Im(P) \oplus Im(\widehat{P})$ ,  $P - \widehat{P} = \overline{P}(P - \widehat{P})$

Et comme  $\widehat{B}^T \widehat{B} = I_{\widehat{r}}$ , les valeurs singulières de  $\widehat{B}$  sont toutes égales à 1. Et donc  $\|\widehat{B}\| = 1$ .

Par définition et sous multiplicativité de la norme et en tenant compte des remarques précédentes on obtient que

$$\left\| \widehat{B} \widehat{\alpha}_n - B \alpha \right\| \leq \frac{1}{n} \frac{\left\| \widehat{B}^T X_n \eta_n \right\|}{\kappa_1} + \left( 1 + \frac{\kappa_2}{\kappa_1} \right) \left\| (P - \widehat{P}) B \alpha \right\| \quad (11)$$

On cherche maintenant à se ramener au cas où le théorème 3 s'applique pour majorer le seconde membre de l'inégalité ci dessus.

On a supposé avec  $\Omega_1$  que  $\widehat{r} \geq r$ ,

On définit alors

$$\widehat{P}_{\leq r} = \sum_{j=1}^r \widehat{u}_j \widehat{u}_j^T, \quad \widehat{P}_{> r} = \sum_{j=r+1}^{\widehat{r}} \widehat{u}_j \widehat{u}_j^T$$



On écrit dans le même temps la décomposition en valeurs singulières de  $A$ .

$$A = \sum_{j=1}^r \sigma_j u_j v_j^T$$

Où les  $\sigma_j$  sont rangées dans l'ordre décroissant. On rappelle que  $B = (u_1, \dots, u_r)$  et que  $P = \sum_{j=1}^r u_j u_j^T$  et qu'ainsi  $PB = B$  et  $P = P_{\leq r}$

Alors suivant cette décomposition

$$\begin{aligned} \left\| (\hat{P} - P)B\alpha \right\| &\leq \left\| (\hat{P}_{\leq r} - P)B\alpha \right\| + \left\| \hat{P}_{>r}B\alpha \right\| \\ &= \left\| (\hat{P}_{\leq r} - P)B\alpha \right\| + \left\| \hat{P}_{>r}P_{\leq r}B\alpha \right\| \\ &\leq \left\| (\hat{P}_{\leq r} - P)B\alpha \right\| + \left\| (I - \hat{P}_{\leq r})P_{\leq r}B\alpha \right\| \end{aligned}$$

Ensuite on écrit que :

$$\begin{aligned} (I - \hat{P}_{\leq r})P_{\leq r} &= ((I - \hat{P}_{\leq r}) - P_{>r})P_{\leq r} = \left( (I - \hat{P}_{\leq r}) - (I - P_{\leq r}) \right) P_{\leq r} \\ &= (P - \hat{P}_{\leq r})P_{\leq r} \end{aligned}$$

En combinant les deux dernières formules il vient que

$$\left\| (P - \hat{P})B\alpha \right\| \leq 2 \left\| (P - \hat{P}_{\leq r})B\alpha \right\| \leq 2 \left\| \hat{P}_{\leq r} - P \right\|_o p \|B\alpha\| \leq 2 \left\| \hat{P}_{\leq r} - P \right\|_{op} \|\alpha\|$$

Alors on remarque deux cas :

$$\sigma_r \leq 2 \left\| \hat{A} - A \right\|.$$

Alors immédiatement il s'ensuit que

$$\left\| \hat{P}_{\leq r} - P \right\| \leq 2 \leq 2 \frac{2 \left\| \hat{A} - A \right\|}{\sigma_r}.$$

Dans le cas où  $\left\| \hat{A} - A \right\| < \frac{\sigma_r}{2}$ . Le résultat de la sous-section 3.3 permet alors de conclure.

□