NNT : 2017SACLG001

THÈSE DE DOCTORAT

de

L'UNIVERSITÉ PARIS-SACLAY

École Doctorale de Mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription: ENSAE – École Nationale de la Statistique et de l'Administration Économique

Laboratoire d'accueil : CREST, UMR 9194 CNRS

Spécialité de doctorat : Mathématiques fondamentales

THE TIEN MAI

PAC-Bayesian estimation of low-rank matrices

Date de soutenance : 23 June 2017
Lieu de soutenance : ENSAE

Après avis des rapporteurs : JUDITH ROUSSEAU (Université Paris Dauphine)
FELIX ABRAMOVICH (Tel Aviv University, Israel)

Jury de soutenance :

PIERRE ALQUIER	(ENSAE) Directeur de thèse
GÉRARD BIAU	(Université Pierre et Marie Curie) Président
NIAL FRIEL	(University College Dublin) Examinateur
CHRISTOPHE GIRAUD	(Université Paris Sud) Examinateur
JUDITH ROUSSEAU	(Université Paris Dauphine) Rapporteur
ALEXANDRE TSYBAKOV	(ENSAE) Examinateur

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Estimation PAC-Bayésienne de matrices de faible rang

Mots Clefs : Inégalités PAC-Bayésienne, complétion de matrices, filtrage collaboratif, tomographie quantique, apprentissage au long cours, inégalité oracle, vitesses minimax, agrégation d'estimateurs, bornes sur le regret, MCMC.

Résumé : Les deux premières parties de cette thèse étudient respectivement des estimateurs pseudo-bayésiens dans les problèmes de complétion de matrices, et de tomographie quantique. Dans chaque problème, on propose une loi a priori qui induit des matrices de faible rang. On étudie les performances statistiques: dans chacun des deux cas, on prouve des vitesses de convergence pour nos estimateurs. Notre analyse repose essentiellement sur des inégalités PAC-Bayésiennes. On propose aussi un algorithme MCMC pour implémenter notre estimateur. On teste ensuite ses performances sur des données simulées, et réelles.

La dernière partie de la thèse étudie le problème de lifelong learning (que l'on peut traduire par apprentissage au long cours), où de l'information est conservée et transférée d'un problème d'apprentissage à un autre. Nous proposons une formalisation de ce problème dans un contexte de prédiction séquentielle. Nous proposons un méta-algorithme pour le transfert d'information, qui repose sur l'agrégation à poids exponentiels. On prouve une borne sur le regret de cette méthode. Un avantage important de notre analyse est qu'elle ne requiert aucune hypothèse sur la forme des algorithmes d'apprentissages utilisés à l'intérieur de chaque problème. On termine cette partie par l'étude de quelques exemples: cas d'un nombre fini de prédicteurs, apprentissage d'une direction révélatrice, et apprentissage d'un dictionnaire.

Title : PAC-Bayesian estimation of low-rank matrices

Keys words : PAC-Bayesian bounds, matrix completion, collaborative filtering, quantum tomography, lifelong learning, transfer learning, oracle inequalities, minimax rates, aggregation of estimators, regret bounds, MCMC.

Abstract : The first two parts of the thesis study pseudo-Bayesian estimation for the problem of matrix completion and quantum tomography. A novel low-rank inducing prior distribution is proposed for each problem. The statistical performance is examined: in each case we provide the rate of convergence of the pseudo-Bayesian estimator. Our analysis relies on PAC-Bayesian oracle inequalities. We also propose an MCMC algorithm to compute our estimator. The numerical behavior is tested on simulated and real data sets.

The last part of the thesis studies the lifelong learning problem, a scenario of transfer learning, where information is transferred from one learning task to another. We propose an online formalization of the lifelong learning problem. Then, a meta-algorithm is proposed for lifelong learning. It relies on the idea of exponentially weighted aggregation. We provide a regret bound on this strategy. One of the nice points of our analysis is that it makes no assumption on the learning algorithm used within each task. Some applications are studied in details: finite subset of relevant predictors, single index model, dictionary learning.

PAC-Bayesian estimation for low-rank matrices

The Tien MAI
maithetienkhk33@gmail.com;
http://sites.google.com/site/thetienmai/

June 26, 2017

Contents

Preface

First and foremost, I would like to express my deepest gratitude to my supervisor Pierre ALQUIER: not only for accepting me as a PhD student, but also for guiding me to fascinating fields in Statistics. He has been always available for technical discussion, as well as provided good feedback and suggestions. Also, I am very thankful for Pierre for proofreading and valuable feedback of the draft of this thesis. Especially, I really appreciate Pierre for translating the Introduction of this thesis into French.

I would like to thank Professor Massimiliano Pontil for interesting and useful discussion that lead to our joint work on lifelong learning.

I would like to thank Professors Gérard Biau, Nial Friel, Christophe Giraud and Alexander Tsybakov for accepting as my PhD jury members; thank Professors Judith Rousseau and Felix Abramovich for accepting as reporters of the thesis.

For my time at University College Dublin, I want to thank Professors Nial Friel and Brendan Murphy for their valuable discussion. I also want to thank the staff and colleagues at the School of Mathematics & Statistics and Insight Centre for Data Analytics, UCD. Thank my friends at UCD for having a good time together.

At CREST-ENSAE, I want to thank Profs Alexander Tsybakov, Nicolas Chopin and Arnak Dalalyan for their support. Thank to the staff and colleagues at CREST and ENSAE. Thank to my friends at CREST, ENSAE for their kindly help.

Finally, I am grateful for the support and motivation that I have received from my dear family in all stages of my PhD study, life and work. I also want to thank my love, An Nguyen, who is always beside me.

Acknowledgements: This work was financially supported by CREST, GENES from Labex ECODEC funding, ANR-11-LABEX-0047. The initial works of this thesis was supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289 through the Insight Centre for Data Analytics.

Chapter 0

Résumé substantiel

0.1 Motivation

Dans plusieurs applications de la statistique, l'objective est d'estimer une matrice de grande dimension à partir d'une observation possiblement bruitées et incomplète de ses entrées. La taille de la matrice, et du jeu d'observation, est énorme (par exemple plusieurs milliards d'entrées). De plus, il y a parfois des contraintes complexes sur la matrice elle-même. Ceci sort du cadre d'application des méthodes statistiques classiques. Un des défis les plus importants de la statistique moderne est le développement d'une nouvelle génération de méthodologies et de théories qui permette une inférence dans de tels problèmes. L'objective de cette thèse est de relever certains aspects de ce défi.

Une approche populaire pour la réduction de la dimension de l'espace des paramètres dans ce problème est inspirée de l'hypothèse de sparsité dans le modèle de régression linéaire: on suppose que la matrice à estimer est de faible rang - ou, du moins, qu'elle peut être bien approximée par une matrice de faible rang. Il faut noter qu'au contraire de l'hypothèse de sparsité, qui porte sur les composantes individuelles d'un vecteur, l'hypothèse de faible rang pour une matrice affecte la matrice entière. Plus précisément, les colonnes (resp. lignes) d'une matrice de faible rang peuvent s'écrire comme combinaisons linéaires d'un petit nombre de vecteurs de base (bien entendu inobservables). Ceci est dans un sens proche de certains modèles statistiques où les observations sont expliquées par un petit nombre de variables "cachées" ou "latentes".

Dans les applications pratiques, l'hypothèse de faible rang est tout à fait sensée. Comme exemple introductif, considérons le célèbre jeu de données utilisé dans le challenge Netflix [Bennett and Lanning, 2007]: les entrées de cette matrice sont les notes données par des utilisateurs (lignes) à des films (colonnes). Beaucoup d'utilisateurs ayant des gouts similaires, on peut penser que cette matrice sera très bien approchée par une matrice de faible rang. Cette hypothèse a été en fait considérée dans beaucoup d'autres modèles:

apprentissage de dictionnaire [Kreutz-Delgado et al., 2003; Mairal et al., 2009; Tosic and Frossard, 2011], complétion de matrices [Candès and Recht, 2009; Keshavan et al., 2009; Koltchinskii et al., 2011; Candes et al., 2015; Kapur et al., 2016]; analyse en composante principales [Wright et al., 2009; Bro and Smilde, 2014; Zou et al., 2006], estimation de matrice de covariance ou de précision [Fan et al., 2008; Pourahmadi, 2013; Cai et al., 2016; Lounici, 2014], tomographie quantique [Gross et al., 2010; Gross, 2011; Flammia et al., 2012; Liu et al., 2012] etc.

Plusieurs méthodes ont été proposées et étudiées pour estimer des matrices de faible rang. Les plus populaires reposent sur des algorithmes d'optimisation convexe efficaces: on essaie de minimiser la somme de deux critères, l'un représentant l'attache aux données (critère des moindres carrés, vraisemblance...) et l'autre une pénalisation qui tend à réduire le rang de la solution. Une relaxation convexe du critère à minimiser permet d'utiliser des algorithmes rapides d'optimisation. Par exemple, une pénalité naturelle est simplement le rang de la matrice; cependant, il s'agit d'une fonction non convexe de la matrice, la relaxation habituelle consiste à remplacer le rang par la norme nucléaire.

Des approches bayésiennes ont aussi été considérées dans ces problèmes. Au lieu de retourner un estimateur de la matrice, l'approche bayésienne fournit une distribution de probabilité sur l'espace des matrices. La pénalité est remplacée par une loi *a priori*, il est donc nécessaire de définir une loi *a priori* qui donne une probabilité faible à des matrices de rang élevé. Là encore, la disponibilité d'algorithmes efficaces (MCMC, approximations variationnelles) a permis aux méthodes bayésiennes de devenir une alternative populaire dans ce genre de problèmes. Cependant, au contraire des méthodes pénalisées, il n'y a presque pas eu de travaux théoriques sur l'analyse des performances statistiques des estimateurs bayésiens pour l'estimation de matrices de faible rang. Le but de cette thèse est non seulement de proposer des estimateurs bayésiens dans les problèmes d'estimation de matrices, mais également de les étudier d'un point de vue théorique.

Les deux premières parties de la thèse sont consacrées à deux problèmes particuliers d'estimation de matrice de faible rang: le problème de complétion de matrice, et la tomographie quantique, où l'objectif est d'estimer la matrice d'état d'un système quantique, qui est de rang un pour un état pur. En complétion de matrice, on construit un estimateur quasi-bayésien et on démontre qu'il satisfait une inégalité oracle optimale, et atteint donc la vitesse minimax (à un log près). Un des points forts de notre résultat est qu'il est valable sans hypothèse sur la loi de tirage des entrées de la matrice que l'on observe, alors que la plupart des résultats précédents supposaient une loi uniforme ou proche de la loi uniforme. Dans le cas de la tomographie quantique, on construit une loi *a priori* sur l'ensemble des matrices de densité. Il faut noter que cette construction est déjà un apport original puisque tous les travaux menés jusqu'ici ne traitaient que du problème dit "1 qubit" (la terminologie sera détaillée dans le coeur de la thèse, ce cas correspond à des matrices 2×2). En s'inspirant de la construction faite dans le cas de la

complétion de matrices, on construit une loi *a priori* pour des matrices de densité de dimension quelconque. On montre qu'un estimateur pseudo-bayésien construit à l'aide de cette loi *a priori* atteint la meilleure vitesse connue dans la littérature actuelle (la question de son optimalité est encore ouverte). Ses performances sont illustrées sur des jeux de donnés réels et simulés.

La dernière partie de la thèse traite du problème dit *lifelong learning* qui est apparu en intelligence artificielle et en *machine learning* (on pourrait traduire *lifelong learning* par *apprentissage au long cours* mais cette terminologie n'étant pas standard on utilisera plutôt le terme anglais dans cette introduction en français). Brièvement, il s'agit d'un problème où un même agent doit résoudre plusieurs tâches d'apprentissage successives, qui partagent une structure commune, et où le problème est de transférer l'information d'une tâche à une autre. Par exemple, chaque tâche peut être une régression linéaire en très grande dimension. Une façon de réduire la dimension du problème est d'utiliser l'ensemble des tâches pour apprendre un dictionnaire de petite taille, et ensuite de résoudre chaque tâche comme un problème de régression de petite dimension. Donc, le problème d'apprentissage de dictionnaire est un cas particulier du *lifelong learning*. En représentant les éléments du dictionnaire comme des vecteurs, on peut représenter le dictionnaire comme une matrice, et on est encore une fois ramené à l'estimation d'une matrice. Ceci dit, le *lifelong learning* est un problème plus vaste, qui contient d'autres exemples que l'apprentissage de dictionnaire, et on l'étudie en toute généralité.

Après l'introduction (Chapitres 0 pour la version française et 1 pour la version en anglais), la thèse est découpée en trois chapitres indépendants les uns des autres. Elle est organisée comme suit:

Chapter 2: on étudie le problème de complétion de matrices, c'est-à-dire la reconstruction d'une matrice à partir de l'observation d'un petit nombre de ses entrées, bruitées. On introduit une nouvelle loi a priori et on démontre qu'un estimateur pseudo-bayésien associé est minimax-optimal. On effectue des tests numériques sur des données simulées, en comparaison avec d'autres estimateurs bayésiens populaires.

Chapter 3: tomographie quantique. On étudie l'estimation de la matrice de densité d'un système quantique de n qubits à partir de données obtenues d'expériences dites "complètes". On propose une loi a priori valable pour une valeur générale de n. On construit deux estimateurs pseudo-bayésiens basés sur cette loi (reposant sur des pseudo-vraisemblances différentes). On démontre la consistance des deux estimateurs. L'un d'eux atteint également la meilleure vitesse connue à ce jour. Là encore, on réalise des simulations sur des données, réelles et simulées.

Chapter 4: lifelong learning. *Le problème est typiquement de transférer de l'information acquise en résolvant plusieurs tâches d'apprentissage en ligne, à un nouvelle tâche, sous l'hypothèse qu'il y a une similarité dans la structure des tâches. En supposant donnée une méthode d'apprentissage dans chaque tâche (agrégation à poids exponentiels, gradient en ligne, etc.), on propose un méta-algorithme qui transfère l'information*

d'une tâche à l'autre. Cette algorithme est basé sur la procédure d'agrégation à poids exponentiels (EWA). La performance statistique de l'algorithme est évaluée par une borne sur son regret. Quelques applications sont traitées, incluant l'apprentissage de dictionnaire.

La fin de cette introduction donne un aperçu général des résultats de ces trois chapitres. Elle est organisée comme suit.

The rest of this introduction is organized as follows. Dans la Section 0.2, on explique rapidement les différentes approches en statistique bayésienne pour construire des lois *a priori* sur des matrices, et on explique lesquelles sont adaptées pour favoriser les matrices de faible rang. Dans la Section 0.3, on introduit les bornes dites "PAC-Bayésiennes", le principal outil théorique pour analyser les estimateurs des Chapitres 2 et 3. Enfin, les Sections 0.4, 0.5 et 0.6 présentent rapidement les résultats des Chapitres 2, 3 et 4 en les comparant à l'état de l'art.

0.2 Lois *a priori* sur les matrices de faible rang

On rappelle rapidement qu'en statistique bayésienne l'idée est d'encoder l'information disponible *a priori*, ou la complexité de l'espace des paramètres, par une loi de probabilité dite *a priori* $p(\mathrm{d}\theta)$. L'inférence est alors faite à travers la loi dite *a posteriori*

$$p(\mathrm{d}\theta \mid \text{data}) \propto \mathcal{L}(\text{data} \mid \theta) p(\mathrm{d}\theta), \tag{1}$$

où $\mathcal{L}(\text{data} \mid \theta)$ est la vraisemblance. Dans cette thèse, on va plutôt considérer des estimateurs dits pseudo-bayésiens, c'est-à-dire que la vraisemblance $\mathcal{L}(\text{data} \mid \theta)$ est remplacée par un terme plus général d'attache aux données, et qui ne suppose en particulier pas que la loi des observations appartient à un modèle paramétrique donné. Mais cet aspect sera surtout discuté dans la Section 0.3 - pour le moment, on présente différentes constructions de lois *a priori* $p(\mathrm{d}\theta)$ sur l'ensemble des matrices.

Les exemples les plus populaires sont les lois normales matricielles, et les lois de Wishart, qui peuvent être trouvées avec d'autres exemples dans [Gupta and Nagar, 1999]. On présente rapidement ces lois, mais on va montrer qu'elles ne sont pas adaptées aux problèmes que l'on souhaite traiter. Supposons par exemple que l'on observe une matrice $\mathbf{X}_{m_1 \times m_2}$ suivant le modèle normal matriciel: $\mathbf{X}_{m_1 \times m_2} \mid \mathbf{M}, \Phi, \Sigma \sim \mathcal{N}(\mathbf{M}, \Phi \otimes \Sigma)$. La vraisemblance est alors

$$\mathcal{L}(\mathbf{X} \mid \mathbf{M}, \Phi, \Sigma) = \frac{\exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T \Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right)}{(2\pi)^{m_1 m_2/2} |\Sigma|^{m_1/2} |\Phi|^{m_2/2}}. \tag{2}$$

Si le paramètre d'intérêt est la matrice $\mathbf{M}$, la loi *a posteriori* pour $\mathbf{M}$ aura la forme

$$p(\mathbf{M}|\mathbf{X}) \propto \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T \Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right) p(\mathbf{M}).$$

De façon à obtenir des lois conjuguées, il faut donc choisir une loi *a priori* pour $\mathbf{M}$ qui soit également normale matricielle:

$$p(\mathbf{M}) = p(\mathbf{M} \mid \Phi_1, \Sigma_1) \propto \exp\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma_1^{-1}(\mathbf{M}-M_0)^T\Phi_1^{-1}(\mathbf{M}-M_0)\right]\right),$$

où les hyperparamètres M_0, Φ_1 et Σ_1 doivent être précisés. Les lois *a priori* et la vraisemblance sont alors conjuguées, on en déduit une forme explicite pour la loi *a posteriori*, qui est elle-même une loi normale matricielle.

D'un autre côté, si le paramètre d'intérêt est Σ (ceci marcherait de façon similaire pour Φ), alors

$$p(\Sigma|\mathbf{X}) \propto |\Sigma|^{-\frac{m_1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T\Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right) p(\Sigma).$$

Ceci suggère cette fois une loi *a priori* qui soit une loi de Wishart inverse

$$p(\Sigma) = p(\Sigma \mid Q, \nu) \propto |\Sigma|^{-\frac{\nu}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}Q\right]\right),$$

où Q et ν sont là encore des hyperparamètres à spécifier. A noter que la loi de Wishart inverse est bien une loi de probabilités sur les matrices définies-positives, ce qui est sensé pour des matrices de covariance. De plus on a là encore des lois conjugées. Plus de détails sur ces lois peuvent être trouvés dans [Rowe, 2002].

On pourrait donc penser à utiliser une loi normale ou Wishart inverse pour nos problèmes (complétion de matrices, tomographie quantique). Le problème est qu'il n'y a aucune raison pour que ces lois favorisent les matrices de faible rang. Donc, les lois *a priori* usuelles, introduits dans certains modèles pour des raisons de conjugaison, ne nous sont d'aucune aide pour nos problèmes d'estimation de matrices de faible rang. Une astuce peut permettre de "tordre" ces lois de façons à approcher des matrices de faible rang, on explique rapidement l'idée correspondante avant de passer à l'approche que nous avons retenu dans cette thèse qui repose sur la factorisation de matrices.

0.2.1 Faible rang par la corrélation

Remarquons qu'une matrice de faible rang a des colonnes (ou des lignes) qui sont liées linéairement. En termes probabilistes, si on tire aléatoirement des colonnes très corrélées, on obtient donc une matrice qui sera bien approchée par une matrice de faible rang. Donc, un choix pertinent de Φ, ou Σ, ci-dessus devrait permettre d'atteindre l'objectif voulu.

Plus précisément, dans la loi *a priori* normale matricielle (2),

$$\mathbf{M} \mid \mathbf{M_0}, \Phi_1, \Sigma_1 \sim \mathcal{N}(\mathbf{M_0}, \Phi_1 \otimes \Sigma_1),$$

où $\mathbf{M_0}$ est la matrice moyenne et Φ_1 et Σ_1 sont respectivement les matrices de covariance des lignes et des colonnes. Dans le cas extrême où la matrice

de précision Φ_1^{-1} ou Σ_1^{-1} (ou les deux) est de faible rang, $\mathbf{M}$ elle-même sera de faible rang.

Cette approche est étudiée en détail dans [Sundin, 2016] sous le nom de *precision based models* et RSVM (*relevance singular vector machine*). Cependant, le problème n'est que partiellement résolu, puisqu'en pratique il faut spécifier les matrices Φ_1 et Σ_1. De plus, malgré des résultats numériques intéressants, le coût computationel des méthodes proposées dans [Sundin, 2016] est énorme, et ces méthodes ne peuvent pas être utilisées à l'heure actuelle sur des jeux de données de la dimension que nous souhaitons étudier (par exemple NetFlix). On propose donc maintenant une approche complètement différente, basée sur la factorisation de matrices.

0.2.2 Faible rang par la factorisation

On rappelle que toute matrice M de taille $m_1 \times m_2$ et de rang K peut être décomposée de la façon suivante (en utilisant la SVD ou décomposition en valeurs singulières)

$$M = USV^T = (US^{\frac{1}{2}})(S^{\frac{1}{2}}V^T),$$

où U, V sont respectivement des matrices de dimension $m_1 \times K$ et $m_2 \times K$ avec des colonnes orthogonales, et S est une matrice $K \times K$ diagonale contenant les valeurs singulières non nulles de M. Posons $A = US^{\frac{1}{2}}$ et $B = VS^{\frac{1}{2}}$, on obtient

$$M = AB^T \tag{3}$$

où A est $m_1 \times K$ et B est $m_2 \times K$. L'idée principale des lois *a priori* factorisées est de définir des lois *a priori* sur A et B plutôt que sur M directement. A notre connaissance, la première approche bayésienne de ce type a été menée dans [Geweke, 1996] dans le cadre d'un modèle dit de régression de faible rang, populaire en économétrie.

Le principal problème est maintenant que le rang K doit être connu en avance. Ca n'est bien entendu pas le cas en pratique. Une approche possible est d'estimer A et B pour toutes les valeurs de K possible, puis ensuite de choisir K en utilisant un critère de sélection de modèles bayésien (BIC ou facteurs de Bayes par exemple). Cette approche a été utilisée originellement dans [Kleibergen and Paap, 2002]. Des approximations numériques efficaces, avec des garanties de convergence dans le modèle de régression de faible rang, ont été proposées par [Corander and Villani, 2004].

Une stratégie adaptative par rapport au rang a été introduite plus récemment: on choisit K volontairement trop grand, par exemple $K = \min(m_1, m_2)$. En revanche, la loi sur A et B est choisie de façon à rendre certaines des colonnes de ces matrices presque nulles, ce qui conduit à une matrice M qui est très proche d'une matrice de faible rang - la loi *a priori* est donc sensée tirer les colonnes matrices A et B vers 0, idée dite de *shrinkage* en anglais. A notre connaissance, [Lim and Teh, 2007] a été le premier article a développer cette idée,

qui a depuis été améliorée et déclinée sous plusieurs formes [Salakhutdinov and Mnih, 2008; Zhou et al., 2010; Babacan et al., 2011, 2012]. Formellement, dans (3), M est une somme de K matrices de rang 1:

$$M = \sum_{j=1}^{K} A_{\cdot j} B_{\cdot j}^T , \tag{4}$$

où $A_{\cdot j}$ et $B_{\cdot j}$ sont la j-ème colonne de A et B respectivement. Si la loi *a priori* donne avec grande probabilité $A_{\cdot j} \simeq 0$ alors la plupart des termes dans (4) seront presque nuls, et M sera en fait très bien approchée par une somme d'un petit nombres de matrices de rang 1, c'est-à-dire par une matrice de faible rang. Par exemple, [Babacan et al., 2012] propose pour $A_{\cdot j}$ et $B_{\cdot j}$ une loi normale de variance γ_j, c'est-à-dire

$$p(A|\gamma) = \prod_{j=1}^{K} \mathcal{N}(A_{\cdot j}|0, \gamma_j I),$$
$$p(B|\gamma) = \prod_{j=1}^{K} \mathcal{N}(B_{\cdot j}|0, \gamma_j I).$$

de plus, les γ_j sont eux-même aléatoires, suivant une loi très concentrée autour de 0. Pour des raisons de conjugaison, il est en fait commode de poser $1/\gamma_j \sim \Gamma(a, b)$ (loi Gamma), avec b très petit.

Dans les chapitres 2 et 3, on définit des loi *a priori* dans le problème de complétion de matrice, et de tomographie quantique respectivement. Aucune de ces lois n'est exactement celle proposée par [Babacan et al., 2012], mais, dans les deux cas, la construction est clairement basée sur la double idée "factorisation + shrinkage" qui vient d'être introduite.

0.3 Une introduction rapide à l'analyse PAC-Bayésienne

On a évoqué précédemment qu'un des buts de notre thèse est d'établir des propriétés statistiques d'estimateurs bayésiens et pseudo-bayésiens. Plusieurs approches théoriques sont disponibles dans ce but.

L'une d'entre elles consiste à prouver la concentration asymptotique de la loi *a posteriori* autour de la bonne valeur du paramètre. Cette approche est par exemple décrite pour les modèles non-paramétriques dans l'article [Ghosal et al., 2000a]. Une revue complète de cette approche est donnée par [Rousseau, 2016].

Une autre approche consiste à étudier l'estimateur MAP (*maximum a posteriori*), en utilisant la théorie de la minimisation du risque empirique pénalisé, qui se base sur des inégalités de concentration. Par exemple, l'estimateur LASSO, au départ introduit comme une minimisation de risque pénalisé, peut être vue comme un MAP avec une vraisemblance gaussienne et une loi *a priori* de Laplace sur le paramètre. Récemment, les auteurs de [Abramovich

and Grinshtein, 2010; Abramovich and Lahav, 2015] ont prouvé des vitesses minimax-optimales pour des estimateurs MAP dans des modèles de régression sparse et dans des modèles additifs non paramétriques, en utilisant une inégalité de concentration de [Birgé and Massart, 2001].

Dans cette thèse, on utilise une approche alternative qui se base sur des inégalités dites PAC-Bayésiennes. Cette approche présente certaines similarités techniques avec l'approche basée sur les inégalités de concentration, mais les résultats sont assez différents. La principale différence avec les approches sur la concentration asymptotique de la loi *a posteriori* est que l'on n'a pas besoin de supposer que les données sont effectivement générées suivant un modèle paramétrique connu. L'approche est en ce sens plus proche des techniques de *machine learning*. De fait, elle a été initiée par la communauté *machine learning*: [Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999]. Des résultats plus fins ont été obtenus par [Catoni, 2004, 2007] qui a aussi établi le lien avec l'approche statistique et les inégalités oracle. Des présentations détaillées des différents aspects de cette approche, et l'application à différents modèles, peuvent se trouver par exemple dans les thèses [Audibert, 2004; Alquier, 2006; Guedj, 2013; Germain, 2015].

De façon générale, l'approche PAC-Bayésienne relie l'erreur de généralisation d'une procédure de prédiction par agrégation à son risque empirique et à la divergence de Kullback-Leibler entre la loi d'agrégation et une loi *a priori*. En minimisant ce critère, on obtient usuellement une loi d'agrégation à poids exponentiels (EWA), c'est-à-dire de la forme (1) mais où la vraisemblance $\mathcal{L}(\text{data} \mid \theta)$ est remplacée par une fonction exponentielle du risque empirique. La fin de cette section constitue une très brève introduction aux bornes PAC-Bayésiennes.

Contexte de base: Soient $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ des couples indépendants les uns des autres, où $\mathcal{X}$ est n'importe quel ensemble mesurable et $\mathcal{Y} = \{-1, +1\}$ dans un problème de classification ou $\mathcal{Y} = \mathbb{R}$ dans un problème de régression. On note $\mathcal{D} := (X_i, Y_i)_{i=1}^n$ pour faire court. On note $\mathbb{P}$ la loi des observations et l'espérance correspondante est notée $\mathbb{E}$.

Le statisticien considère un ensemble de prédicteurs (ou hypothèses dans le language de la communauté *machine learning*) $\mathcal{H} := \{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\}$. Chaque prédicteur conduit à la perte $\ell(Y_i, f_\theta(X_i))$ sur le i-ème exemple. Par exemple, on peut utiliser la perte quadratique $\ell(Y_i, f_\theta(X_i)) = (Y_i - f_\theta(X_i))^2$ dans le modèle de régression. Le risque de prédiction de f_θ est défini par

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\ell(Y_i, f_\theta(X_i)).$$

Bien noter que cette quantitié est inconnue (car $\mathbb{P}$ est inconnue), mais sa contrepartie empirique, appelée risque empirique de f_θ, peut être calculée sur la base des observations:

$$r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)).$$

L'approche "classique" s'intéresse à des estimateurs $\hat{\theta} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ et étudie les relations entre le risque empirique $r(\hat{\theta})$ et le risque $R(\hat{\theta})$. L'approche PAC-Bayésienne étudie plutôt des fonctions $\rho : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}_+^1(\Theta)$ où $\mathcal{M}_+^1(\Theta)$ est l'ensemble de toutes les lois de probabilités sur Θ muni d'une tribu $\mathcal{T}$. Suivant les contextes, on peut donner des garanties théoriques sur une prédiction "moyenne" selon ρ, $\hat{\theta} = \int \theta \rho(\mathrm{d}\theta)$, dite aussi agrégation, ou sur un estimateur dit "randomisé" $\hat{\theta}$ tiré directement suivant ρ.

Une borne PAC-Bayésienne empirique: On donne une première borne basique.

Théorème 0.1 (Par ex. Théorème 2.3 dans [Alquier, 2006])**.** *Fixons une loi $\pi \in \mathcal{M}_+^1(\Theta)$. On suppose que la perte ℓ est à valeurs dans $[0, C]$ pour une constante $C > 1$. Pour tout $\lambda \in (0, n/C)$, avec probabilité au moins $1 - \varepsilon, \varepsilon \in (0, 1)$, pour tout $\rho \in \mathcal{M}_+^1(\Theta)$*

$$\int_\Theta R(\theta)\rho(d\theta) \leq \int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n}. \tag{5}$$

On rappelle que $\mathcal{K}(\rho, \pi)$ est la divergence de Kullback-Leibler entre ρ and π, donnée par

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int \log\left(\frac{d\rho}{d\pi}\right) d\rho, & \text{quand } \rho \text{ est absolument continue par rapport à } \pi, \\ +\infty, & \text{sinon.} \end{cases}$$

On introduit la notation $\nu(h) = \int_\Theta h(\theta)\nu(d\theta)$. On donne maintenant un lemme important duquel on peut déduire une loi ρ qui minimise le membre de droite dans l'inégalité précédente.

Lemme 0.1. *Pour toute fonction $h : \Theta \to \mathbb{R}$ mesurable bornée et pour tout $\rho \in \mathcal{M}_+^1(\Theta)$ tel que $\mathcal{K}(\rho, \pi) < \infty$ on a*

$$-\log \pi[\exp(h)] = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} [-\rho(h) + \mathcal{K}(\rho, \pi)]$$

En particulier, le minimum du membre de droite est atteint explicitement pour la loi dite "loi de Gibbs" $\rho_{\exp(h)}$ définie par

$$\frac{d\rho_{\exp(h)}}{d\pi}(\theta) = \frac{\exp(h(\theta))}{\pi(\exp(h))}.$$

Preuve: On a

$$\begin{aligned} \mathcal{K}(\rho, \rho_{\exp(h)}) &= \rho\left(\log\left(\frac{d\rho}{d\pi}\right) - h\right) + \log \pi[\exp(h)] \\ &= \mathcal{K}(\rho, \pi) - \rho(h) + \log \pi[\exp(h)]. \end{aligned}$$

Le membre de gauche est positif, et s'annule uniquement lorsque $\rho = \rho_{\exp(h)}$ (bien noter que cette relation est toujours valable si ρ n'est pas absolument

continue par rapport à π, elle dit alors simplement que $+\infty = +\infty$). On obtient alors

$$0 = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} [\mathcal{K}(\rho, \pi) - \rho(h)] + \log \pi[\exp(h)].$$

□

Interprétation bayésienne Du Lemme 0.1, on déduit que la loi qui minimise le membre de droite dans (5) est

$$\hat{\rho}_\lambda(d\theta) = \frac{\exp\{-\lambda r(\theta)\}}{\pi(\exp\{-\lambda r(\theta)\})}\pi(d\theta).$$

Donc, on a

$$\hat{\rho}_\lambda(d\theta) \propto \mathcal{L}(\text{data} \mid \theta) p(\mathrm{d}\theta)$$

comme dans (1), avec $\mathcal{L}(\text{data} \mid \theta) = \exp\{-\lambda r(\theta)$ et $\pi(d\theta) = p(\mathrm{d}\theta)$. Plus précisément, $\exp\{-\lambda r(\theta)\}$ jour le rôle d'une vraisemblance, $\pi(d\theta)$ peut être interprétée comme une loi *a priori*, et λ est un paramètre de réglage qui permet d'équilibrer les rôles de l'information provenant des observations et de la loi *a priori*. On peut utiliser le terme "pseudo-vraisemblance" pour désigner la fonction de θ: $\exp\{-\lambda r(\theta)\}$, et également le terme "estimateur pseudo-bayésien" pour tout estimateur dont la construction sera basée sur $\hat{\rho}_\lambda$.

Remarquons que dans ce cas (5) devient

$$\int_\Theta R(\theta)\hat{\rho}_\lambda(d\theta) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n} \right]. \quad (6)$$

Remarquons aussi que si ℓ est convexe on peut utiliser l'inégalité de Jensen et obtenir

$$R\left(\int_\Theta \theta \rho(d\theta) \right) \leq \rho[R(\theta)].$$

Donc, on est capable de donner une borne sur le risque $R(\cdot)$ de l'estimateur agrégé, de la forme

$$\hat{\theta}_\lambda := \int_\Theta \theta \hat{\rho}_\lambda(d\theta).$$

En effet, dans ce cas, (6) conduit à

$$R(\hat{\theta}_\lambda) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left[\int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n} \right].$$

Inégalité PAC-Bayésienne de type oracle: Le membre de droite dans (5) peut être calculée sur la base des observations, et donc conduit à un moyen de contrôler numériquement la performance de notre méthode d'estimation. C'était en fait l'objectif des premières bornes PAC-Bayésiennes publiées [Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999].

Cependant, la vitesse de convergence de l'estimateur ne peut pas être obtenue directement à partir d'une borne empirique. Ceci a motivé l'introduction d'inégalités PAC-Bayésiennes de type oracle dans [Catoni, 2004, 2007]. Plus précisément, Catoni a démontré qu'on peut construire des bornes PAC-Bayésiennes qui comparent $\int_\Theta R d\hat{\rho}_\lambda$ au meilleur risque intégré possible. La version la plus simple possible est donnée dans le théorème suivant.

Théorème 0.2. *Sous les mêmes hypothèses que pour le théorème précédent, pour tout $\lambda \in (0, n/C)$, avec probabilité au moins $1-\varepsilon, \varepsilon \in (0,1)$*

$$\int_\Theta R(\theta)\hat{\rho}_\lambda(d\theta) \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta)} \left\{ \int_\Theta R(\theta)\rho(d\theta) + 2\frac{\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{n} \right\}. \quad (7)$$

Il suffit donc de calculer le membre de droite (qui n'est plus aléatoire) pour obtenir une vitesse de convergence. Ce calcul n'est pas aisé en général, l'astuce qui consiste à réduire l'infimum dans le membre de droite à des $\tilde{\rho}$ dans une famille paramétrique qui se concentre autour du minimiseur de la fonction $R(\cdot)$ conduit souvent à une simplification qui permet de mener le calcul à bien. Remarquons que si $\tilde{\rho}$ est trop concentrée, ceci peut conduire à une explosion de la divergence de Kullback-Leibler avec π. C'est en équilibrant les deux termes (risque empirique, et divergence) que l'on obtient la meilleure vitesse dans le membre de droite. Cette technique a été utilisée par [Dalalyan and Tsybakov, 2008; Alquier and Lounici, 2011] pour obtenir des vitesses optimales dans le cadre de la régression linéaire sparse (en utilisant des bornes PAC-Bayésiennes plus fines que (7)).

On mentionne enfin quelques avancées plus récentes dans la théorie et l'application des bornes PAC-Bayésiennes: [Seldin et al., 2012; Germain et al., 2013; Pentina and Lampert, 2014; Ridgway et al., 2014; Galanti et al., 2016]. Récemment [Bégin et al., 2016; Alquier and Guedj, 2016] ont proposé des variances où la divergence de Kullback est remplacée par une autre divergence. La plupart de ces articles utilisent une fonction de perte ℓ bornée, ou sous-gaussienne. Cependant, en utilisant une technique de robustification due à [Catoni, 2012], les articles [Catoni, 2016; Giulini, 2015] proposent des bornes PAC-Bayésiennes pour l'estimation de la matrice de Gram avec des observations suivant des lois à queues lourdes. Une autre approche pour obtenir des bornes PAC-Bayésiennes pour des lois à queues lourdes a été récemment proposée par [Grünwald and Mehta, 2016].

0.4 Présentation de nos résultats pour la complétion de matrices

0.4.1 Introduction au problème de complétion de matrices

La complétion de matrice a été un des problèmes statistiques les plus étudiés dans les dix dernières années. Il consiste à reconstruire une matrice M sur la base d'observations partielles aléatoires et possiblement bruitées. Ce problème apparaît dans un grand nombre d'applications comme les systèmes de recommandation [Bennett and Lanning, 2007; Cai et al., 2010; Melville and Sindhwani, 2011], le traitement de l'image et de la vidéo [Ji et al., 2010; Liu et al., 2013], la génomique [Chi et al., 2013; Natarajan and Dhillon, 2014; Cai et al., 2015a]...

On considère un exemple jouet dans la Table 1, pour la recommandation. On ne peut pas supposer que chaque utilisateur a vu tous les films, ni même qu'il/elle notera tous les films qu'il a vu. Donc, la plupart des entrées de la matrice sont inobservées. Dans le prix Netflix [Bennett and Lanning, 2007], les données concernaient 480 189 utilisateurs, 17 770 films, et seulement 100 480 507 notes étaient observées, sur un total de 8 532 958 530 entrées dans la matrice, soit moins de 1.2%. Reconstruire les entrées manquantes est évidemment très utile pour faire de la publicité ciblée, intelligente et ainsi améliorer les ventes.

	Looper	π	Inception	Big Hero 6	...
...	?	1	2	5	...
Aisling	4	?	5	?	...
Bianca	?	5	?	2	...
Tien	5	?	5	?	...
...	1	2	?	4	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 1: Exemple jouet de matrice de notes utilisateurs/films. Les notes sont entre 1 et 5.

Evidemment, il est impossible de faire la moindre inférence sur les entrées de M manquantes dans faire d'hypothèse sur la structure de cette matrice. Une percée majeure a été effectuée par Candès avec différents co-auteurs en prouvant que, si la matrice M est de faible rang, la résolution du problème devient possible [Candès and Recht, 2009; Candès and Plan, 2010; Candès and Tao, 2010; Cai et al., 2010]. Cette hypothèse est de plus parfaitement sensée dans l'application mentionnée précédemment: supposer que, par exemple, beaucoup d'utilisateurs ont des goûts similaires, c'est supposer que beaucoup de lignes sont proportionnelles, ce qui induit un faible rang pour la matrice M.

Précisons maintenant les notations. On note $M^0 \in \mathbb{R}^{m_1 \times m_2}$ la vraie matrice à reconstruire (supposée de faible rang). Un modèle possible d'ovservations

est

$$Y_{i,j} = M^0_{i,j} + \varepsilon_{i,j}, (i,j) \in \Omega,$$

où Ω est un sous-ensemble aléatoire de $\{1,\dots,m_1\} \times \{1,\dots,m_2\}$ avec $n = \text{card}(\Omega) \ll m_1 m_2$. Les variables de bruit $\varepsilon_{i,j}$ sont indépendantes et $\mathbb{E}(\varepsilon_{i,j}) = 0$.

Dans le papier original [Candès and Recht, 2009], les auteurs proposent un estimateur $\hat{M}$ basé sur une relaxation convexe du rang

$$\hat{M} = \arg \min_{A: A_{i,j} = Y_{i,j}, \forall (i,j) \in \Omega} \|A\|_*$$

où $\|A\|_*$ est la norme nucléaire de A:

$$\|A\|_* = \sum_{i=1}^{\min(m_1,m_2)} \lambda_i(A)$$

où les $\lambda_i(A)$ sont les valeurs singulières de A. Ils prouvent que dans le cas sans bruit ($\varepsilon_{i,j} = 0$), il y a une reconstruction exacte $\hat{M} = M^0$ avec très grande probabilité, sous une hypothèse de faible rang sur M^0, et pourvu que n soit assez grand. Le résultat a été etendu (avec une estimateur légèrement adapté) au cas bruité par [Candès and Plan, 2010]. Depuis, plusieurs méthodes ont été proposées, qui reposent presque toutes sur la minimisation du risque empirique avec différentes pénalités. Par exemple: pénalisation par le rang (difficile à traiter numériquement) [Klopp, 2011], par l'entropie de von Neumann [Koltchinskii, 2011], par les normes de Schatten S_p [Rohde and Tsybakov, 2011] ou d'autres normes basées sur les propriétés spectrales [Gunasekar et al., 2015]... Un des estimateurs les plus étudiés est l'estimateur dit "LASSO matriciel"

$$\hat{M}_{nuclear} = \arg\min_M \left\{ \frac{1}{n} \sum_{(i,j)\in\Omega} (Y_{i,j} - M_{i,j})^2 + \lambda \|M\|_* \right\},$$

où $\lambda > 0$ est un paramètre de réglage.

Dans l'article [Koltchinskii et al., 2011], les auteurs proposent un nouveau modèle statistique dit "régression trace". C'est un modèle abstrait général, qui inclut la régression linéaire et la complétion de matrice comme cas particuliers. Ils proposent un estimateur $\tilde{M}_{nuclear}$ qui est une variante de $\hat{M}_{nuclear}$ et mènent son analyse statistique. En particulier, ils démontrent le résultat suivant.

Théorème 0.3 (Corollaire 2 dans [Koltchinskii et al., 2011])**.** *Sous certaines hypothèses précisées dans [Koltchinskii et al., 2011], avec probabilité au moins* $1 - 3/(m_1 + m_2)$

$$\frac{\|\tilde{M}_{nuclear} - M^0\|_F^2}{m_1 m_2} \leq C \frac{\text{rank}(M^0) \max(m_1, m_2)}{n} \log(m_1 + m_2),$$

où C est une constante numérique et $\|B\|_F^2 = \text{Trace}(BB^T)$ la norme de Frobenius.

Les auteurs démontrent également une borne inférieure pour la complétion de matrice de faible rang avec la norme de Frobenius. Ceci établit donc la vitesse minimax dans ce problème.

Théorème 0.4 (Théorème 5 dans [Koltchinskii et al., 2011]). *Fixons $a > 0$ et $1 \leq r \leq \min(m_1, m_2)$. Sous des hypothèses précisées dans [Koltchinskii et al., 2011], il existe des constantes absolues $\beta \in (0, 1)$ et $c > 0$ telles que*

$$\inf_{\hat{M}} \sup_{\substack{\text{rank}(M^0) \leq r, \\ \max_{i,j} |M^0_{i,j}| \leq a}} \mathbb{P}_{M^0} \left(\frac{1}{m_1 m_2} \|\hat{M} - M^0\|_F^2 > c \frac{r \max(m_1, m_2)}{n} \right) \geq \beta.$$

Ce résultat affirme que l'erreur quadratique moyenne de reconstruction d'une entrée d'une matrice de taille $m_1 \times m_2$ et de rang r à partir de n observations ne peut pas être plus petite que l'ordre $r \max(m_1, m_2)/n$. On peut remarquer que la borne supérieure du Théorème 0.3 n'est pas exactement la même - il y a un $\log(m_1 + m_2)$ en plus. La vitesse minimax n'est en fait connue qu'à un log près. Récemment, des bornes parfaitement égales ont été obtenues par [Klopp, 2015], mais dans un modèle légèrement différent où la taille de l'échantillon, n, est elle-même aléatoire.

Dans la plupart des articles mentionnés, il est supposé que les n entrées (i, j) observées sont tirées de façon uniformes, et i.i.d, sur $\{1, \dots, m_1\} \times \{1, \dots, m_2\}$. Cette hypothèse n'est cependant par réaliste dans les exemples mentionnés précédemment: par exemple, certains films sont plus connus que d'autres, et reçoivent plus de notes, bonnes ou mauvaises. De plus, la loi d'échantillonnage est inconnue en pratique. Certains articles ont considéré des lois non uniformes, par exemple [Foygel et al., 2011; Negahban and Wainwright, 2012; Klopp, 2014], mais avec quand même des hypothèses restrictives sur la loi d'échantillonage.

0.4.2 Principaux résultats du Chapitre 2

Alors que les méthodes par pénalisation sont donc maintenant bien comprises à la fois en théorie et d'un point de vue algorithmique, les premières publications sur des méthodes bayésiennes se sont uniquement consacrées aux aspects algorithmiques, et ne contenaient pas de preuve de convergence: [Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Lawrence and Urtasun, 2009; Zhou et al., 2010; Babacan et al., 2011; Alquier et al., 2014] entre autres.

Pour des raisons computationnelles, la plupart des estimateurs bayésiens étaient basés sur des lois *a priori* conjuguées, qui permettaient d'utiliser l'algorithme de Gibbs [Alquier et al., 2014; Salakhutdinov and Mnih, 2008] ou des méthodes variationnelles [Lim and Teh, 2007]. Ces lois *a priori* sont présentées et discutées en détail dans [Alquier et al., 2014]. Les algorithmes correspondant étaient assez rapides pour traiter des jeux de données massifs comme Netflix ou MovieLens[1]: ils sont en fait testés sur ces jeux de données

[1] http://grouplens.org/datasets/movielens/

dans les articles mentionnés précédemment. Mais, encore une fois, les propriétés statistiques n'étaient pas étudiées.

La première contribution de cette thèse a été de construire une loi *a priori* qui conduise à un estimateur dont on peut démontrer qu'il est convergent et même minimax-optimal (à un éventuel log près). On adapte la construction par factorisation pour construire une loi qui soit adaptative au rang. La principale différence est que l'on remplace les lois gaussiennes sur la distribution des colonnes dans (4) par des lois uniformes sur des intervalles. L'estimateur que l'on propose, noté $\widetilde{M}$ dans cette introduction, est la moyenne de la pseudo-loi *a posteriori* (c'est-à-dire que l'on a remplacé la vraisemblance par une pseudo-vraisemblance comme expliqué précédemment). La construction exacte de l'estimateur est détaillée dans le Chapitre 2. On donne ici un aperçu du résultat principal (énoncé complètement dans le Chapitre 2) .

Théorème 0.5 (Théorème 2.1 dans le Chapitre 2)**.** *Supposons que les n entrées observées sont i.i.d suivant une loi $(\pi_{i,j})_{1\leq i\leq m_1, 1\leq j\leq m_2}$: la probabilité d'observer l'entrée (i,j) est $\pi_{i,j}$. Sous des hypothèses adéquates, portant uniquement sur le bruit (ε_i), et précisées dans le Chapitre 2, on a, avec grande probabilité*

$$\sum_{\substack{1\leq i\leq m_1,\\ 1\leq j\leq m_2}} (\widetilde{M}_{i,j} - M^0_{i,j})^2 \pi_{i,j} \leq C \frac{\mathrm{rank}(M^0)\max(m_1,m_2)}{n} \log(\min(m_1,m_2)),$$

où C est une constante numérique

En particulier, lorsque la loi d'échantillonage est uniforme $\pi_{i.j} = 1/m_1m_2$,

$$\frac{\|\widetilde{M} - M^0\|_F^2}{m_1m_2} \leq C' \frac{\mathrm{rank}(M^0)\max(m_1,m_2)}{n} \log(\min(m_1,m_2)),$$

où C' est une constante numérique. Cette vitesse, grâce à la borne inférieure discutée précédemment, est minimax-optimale à un log près. On peut noter une (très) légère amélioration par rapport à [Koltchinskii et al., 2011]: le $\log(m_1+m_2) \asymp \log(\max(m_1,m_2))$ est remplacé par $\log(\min(m_1,m_2))$.

D'un point de vue algorithmique, en utilisant une méthode de Monte-Carlo par chaîne de Markov (MCMC), on a pu tester notre estimateur sur des données simulées de taille 1000×1000. Un exemple de résultat numérique est donné dans la Table 2 (cf. le Chapitre 2 pour des résultats exhaustifs).

0.4.3 Remarques bibliographiques

Plusieurs extensions et améliorations ont été publiées ces dernières années. Par exemple, les résultats de [Koltchinskii et al., 2011], tout aussi bien que les notres, supposent que l'on connaît la variance du bruit, ou au moins un majorant de cette variance. Cette hypothèse n'est pas toutefois si irréaliste que l'on pourrait le croire: par exemple, dans le prix Netflix, les notes sont

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Unif.	0.535 ($\pm$0.003)	0.348 ($\pm$0.003)	0.207 ($\pm$0.0001)	0.141 ($\pm$0.0006)
Gaus.	0.538 ($\pm$0.001)	0.345 ($\pm$0.001)	0.210 ($\pm$0.0001)	0.146 ($\pm$0.001)

Table 2: *Erreur quadratique moyenne dans une série d'expériences (matrice carée de faible rang, de taille $m \times m$, bruit gaussien). On compare notre estimateur avec loi a priori uniforme à l'estimateur bayésien à loi a priori gaussienne, utilisé par [Salakhutdinov and Mnih, 2008; Babacan et al., 2012], et on note que les résultats sont assez similaires.*

bornées (entre 1 et 5) et donc on a unr majoration triviale de la variance. D'un autre côté, pour d'autres applications, cette variance pourrait être inconnue. Le problème de complétion de matrice avec une variance inconnue a été traité par [Klopp, 2014]. L'estimateur proposé est

$$\hat{M}_{SQ} = \arg\min_{M} \left\{ \sqrt{\frac{1}{n} \sum_{(i,j)\in\Omega} (Y_{i,j} - M_{i,j})^2} + \lambda \|M\|_* \right\}.$$

Cet estimateur est l'analogue du *square-root Lasso* [Belloni et al., 2011]. Cet estimateur atteint la même vitesse que celle de l'estimateur de [Koltchinskii et al., 2011]. Une extension de cette idée dans le cadre bayésien serait certainement intéressante, et pourrait faire l'objet d'un travail à venir.

Parmi les autres variantes on trouve le problème de la complétion de matrices binaires [Davenport et al., 2014]: cette fois, les entrées observées ne peuvent prendre que deux valeurs, par exemple 0 ou 1. L'étude de ce modèle a été menée dans [Cai and Zhou, 2013; Klopp et al., 2015; Srebro et al., 2004] (entre autres). Depuis la publication de l'article correspondant au Chapitre 2 de cette thèse, les outils en ont été repris par [Cottet and Alquier, 2016] pour traiter le problème de la complétion de matrices binaires.

La complétion de matrices robuste a été étudiée par [Klopp et al., 2014], dans lequel les auteurs rajoutent un terme de pénalisation supplémentaire pour supprimer les points aberrants. Un autre point de vue est proposé par [Alquier et al., 2017a] qui proposent plutôt de remplacer la perte quadratique par une perte insensible aux points aberrants (comme la perte absolue). L'article [Carpentier et al., 2016] va au-delà des problèmes d'estimation ponctuelle pour s'intéresser à des régions de confiance sur les matrices à reconstruire.

0.5 Présentation de nos résultats pour la tomographie quantique

0.5.1 Rapide introduction à la statistique quantique

La tomographie quantique joue un rôle important dans le traitement de l'information quantique. Elle consiste en la reconstruction de l'état quantique

(supposé inconnu) d'un système physique [Paris and Řeháček, 2004]. Cette tâche est accomplie à l'aide de mesures effectuées sur des copies indépendantes de ce système. On renvoie le lecteur à l'introduction de la thèse [Meziani, 2008], ou à l'article [Artiles et al., 2005] pour à la fois une présentation générale des concepts de base de la physique quantique, et une introduction aux aspects statistiques de la tomographie quantique.

Brièvement, selon la théorique de la physique quantique, toute l'information sur un système physique est contenue dans ce que l'on appelle son "état quantique". Le résultat d'une expérience menée sur un système n'est pas, en général, une fonction déterministe de ce système, mais une variable aléatoire, dont la loi de probabilité est une fonction explicite de l'état quantique du système (on va préciser ceci un peu plus loin). Une représentation mathématique possible de l'état d'un système est une matrice dite matrice de densité ρ. Cette matrice a des entrées à valeurs complexes, et vérifie

- ρ est hermitienne, $\rho^\dagger = \rho$ (auto-adjointe),
- ρ est définie positive, $\rho \geq 0$,
- $\mathrm{Trace}(\rho) = 1$.

Les dimensions de ρ dépendent du système considéré, et peuvent être finies ou infinies. Par exemple, dans le modèle de tomographie homodyne quantique traité dans [Artiles et al., 2005; Butucea et al., 2007; Alquier et al., 2013b; Naulet and Barat, 2016], la matrice ρ est à indices dans $\mathbb{N}$ et ses coefficients $\rho_{i,j}$ ont un module qui décroit exponentiellement en i et j.

Ici, on s'intéresse à un modèle utilisé en informatique quantique, et dans ce cas ρ est de dimension finie. Le système d'intérêt est un système dit de n qubits à spin $1/2$ et la matrice de densité correspondante ρ est une matrice $2^n \times 2^n$. De plus, les physiciens sont particulièrement intéressés par des états dits *états purs*, qui correspondent à des matrices ρ de rang 1. En pratique, considérer que la matrice d'un état est de faible rang peut être sensé de façon plus générale [Gross et al., 2010; Gross, 2011].

Il est important pour les physiciens de pouvoir tester leur capacité à préparer un système dans un état donné ρ_0. Pour ceci, ils utilisent un dispositif expérimental sensé préparer un système dans cet état ρ_0, et produisent, en utilisant le même dispositif, plusieurs copies du même système. Ils peuvent ensuite faire des mesures sur chacune de ces copies, et essayer de reconstruire l'état ρ dans lequel le dispositif place effectivement le système (idéalement, $\rho = \rho_0$). La reconstruction de ρ à partir de ces observations indépendantes est ce que l'on appelle la *tomographie quantique*. On renvoie le lecteur à [Artiles et al., 2005] pour plus de détails sur ce modèle. On donne maintenant les principaux aspects de son formalisme.

Pour chaque qubit, on peut observer son spin suivant chacun des trois axes x, y ou z. L'observation du spin de fait grâce aux observables de Pauli:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}; \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

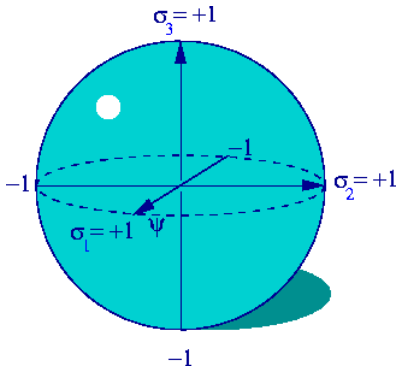

Figure 1: *Résultat de mesures de Paulin sur une particule de spin* 1/2. *Source:*[https://en.wikipedia.org/wiki/Quantum_indeterminacy]

Pour un système de n qubits, on a donc 3^n mesures possibles, et le résultat de la mesure est un vecteur dans $\{-1,+1\}^n$. Si chaque mesure possible est répétée m fois sur m systèmes indépendants, on a en tout un total de $N = m \times 3^n$ mesures.

Etant donné un vecteurs $\in \{-1,1\}^n$, la probabilité de l'observer est une fonction de la densité ρ et du type de mesure que l'on effectue. Elle est donnée par la *règle de Born*

$$M_{i,\mathbf{s}} := \mathbb{P}(R^i = \mathbf{s}) = \text{Trace}\left(\rho \cdot P^i_{\mathbf{s}}\right), i \in \{1,\ldots,3^n\}, \tag{8}$$

où les $P^i_{\mathbf{s}}$ sont connues explicitement. Ceci signigie essentiellement qu'il y a une fonction lineaire F telle que $M = F(\rho), M = (M_{i,\mathbf{s}})_{i\in\{1,\ldots,3^n\},\mathbf{s}\in\{-1,1\}^n}$.

On donne un exemple possible: Exemple 0.1, dans le cas de 2-qubits, de façon à rendre le problème plus clair. Dans ce cas, il y a 9 mesures expérimentales possibles:$(\sigma_x,\sigma_x),(\sigma_x,\sigma_y),(\sigma_x,\sigma_z),\ldots,(\sigma_z,\sigma_z)$; et pour chaque expérience, il y a 4 résultats possibles: $(-1,-1),(-1,+1),(+1,-1),(+1,+1)$.

Exemple 0.1. *On suppose que ρ est telle que la loi de probabilité associée à chaque mesure est donnée par*

	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	0.24	0.26	0.29	0.21
2	0.34	0.36	0.19	0.11
⋮	⋮	⋮	⋮	⋮
9	0.23	0.25	0.37	0.15

$=: M = F(\rho)$

où on rappelle que $M = F(\rho)$ peut être déduite de ρ par la règle de Born (8). En pratique, un physicien va par exemple mesurer chaque observable $i \in \{1, \dots, 9\}$ un nombre $\boldsymbol{m}$ de fois, disons par exemple $m = 1000$. Un résultat possible sera:

	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	232/1000	266/1000	291/1000	211/1000
2	336/1000	361/1000	192/1000	111/1000
⋮	⋮	⋮	⋮	⋮
9	233/1000	250/1000	365/1000	152/1000

$:= \widehat{\mathbf{M}}$

Pour inférer la matrice ρ, une idée naturelle est la méthode dite d'inversion, qui consiste à définir $\hat{\rho}$ telle que

$$F(\hat{\rho}) = \widehat{M}. \tag{9}$$

Cette méthode, en fait connue sous le nom de méthode des moments en statistique, est étudiée dans [Vogel and Risken, 1989; Řeháček et al., 2010]. Bien qu'elle soit assez facile à mettre en oeuvre, elle a plusieurs problèmes: notamment, elle retourne souvent un $\hat{\rho}$ qui ne vérifie pas les axiomes d'une matrice de densité [Shang et al., 2014].

Une autre méthode populaire est l'estimation par maximum de vraisemblance. Malheureusement, elle souffre aussi d'un certain nombre de problèmes, détaillés dans [Blume-Kohout, 2010]. En particulier, elle est beaucoup plus lourde numériquement.

Mais de façon plus importante pour nous, aucune de ces méthodes n'utilise l'information selon laquelle ρ doit être de faible rang. Pour résoudre ce problème, des méthodes adaptatives au rang grâce à des pénalités convenablement choisies ont été proposées. Un maximum de vraisemblance pénalisé par le rang via un critère BIC a été utilisé par [Guţă et al., 2012] et un critère des moindres carrés pénalisé par le rang par [Alquier et al., 2013a], avec une preuve de sa consistance. Plus précisément, quand la matrice de densité du système ρ^0 est de rang $r = \text{rank}(\rho^0)$, les auteurs de [Alquier et al., 2013a] démontrent que leur estimateur $\hat{\rho}_{\text{rank-pen}}$ vérifie

$$\|\hat{\rho}_{\text{rank-pen}} - \rho^0\|_F^2 = \mathcal{O}(r4^n/N)$$

où on rappelle que $N = m3^n$ est le nombre de mesures. La vitesse a été améliorée à $\mathcal{O}(r3^n/N)$ par [Butucea et al., 2015, 2016], en utilisant une méthode de seuillage.

Théorème 0.6 (Corollaire 1 dans [Butucea et al., 2015]). *Sous des hypothèses convenables, avec probabilité au moins* $1-\varepsilon, \varepsilon \in (0,1)$

$$\|\hat{\rho}_{\text{rank}-\text{pen}} - \rho^0\|_F^2 \leq C\frac{r3^n}{N}\log(2^{n+1}/\varepsilon).$$

De plus, des bornes inférieures sont aussi prouvées dans [Butucea et al., 2015]. L'article montre que la vitesse minimax ne peut pas être plus petite que $r2^n/N$. Donc, elle est en fait située quelque part entre $r2^n/N$ et $r3^n/N$. C'est le théorème suivant qui l'établit.

Théorème 0.7 (Borne inférieure, Théorème 3 dans [Butucea et al., 2015]).

$$\liminf_{m\to\infty} \inf_{\hat{\rho}_m} \sup_{\rho\in\mathcal{S}_r} \mathbb{E}_\rho\|\hat{\rho}_m - \rho\|_F^2 \geq \frac{2r(2^n - r)}{N}$$

où $\mathcal{S}_r$ *est l'ensemble des matrices de densité de rang au plus* r.

D'un autre côté, des méthodes bayésiennes ont été considérées pour ce problèmes. Les articles [Bužek et al., 1998; Baier et al., 2007] comparent méthodes bayésiennes et non bayésiennes sur des données simulées. Des algorithmes efficaces de calcul d'estimateurs bayésiens en tomographie quantique sont proposées dans [Kravtsov et al., 2013; Ferrie, 2014; Kueng and Ferrie, 2015; Schmied, 2016]. L'article [Blume-Kohout, 2010] établit un certain nombre de bonnes propriétés d'estimateurs bayésiens cans ce problème. Ceci dit, la convergence quand $m \to \infty$ n'est pas prouvée. Surtout, aucune de ces références ne traite du problème de l'adaptation au rang.

0.5.2 Résultats principaux du Chapitre 3

On considère dans le Chapitre 3 deux estimateurs pseudo-bayésiens basés sur différentes pseudo-vraisemblances. Le premier, $\tilde{\rho}_\lambda^{dens}$, repose sur une comparaison entre ρ and $\hat{\rho}$ (l'estimateur par inversion) alors que l'autre, $\tilde{\rho}_\lambda^{prob}$ repose sur une comparaison entre les fréquences empiriques et théoriques $F(\rho)$ and $F(\hat{\rho})$. En utilisant l'approche PAC-Bayésienne on démontre des inégalités oracle pour la pseudo-moyenne *a posteriori*. En particulier, pour un λ bien choisi, $\tilde{\rho}_\lambda^{prob}$ atteint la meilleure vitesse connue à ce jour $\mathcal{O}(\text{rank}(\rho^0)3^n/N)$ [Butucea et al., 2015]. Ceci dit, $\tilde{\rho}_\lambda^{dens}$ garde un intérêt pratique car son calcul est beaucoup plus rapide.

La difficulté principale ici était de définir une loi *a priori* sur l'ensemble des matrices de densité. On rappelle que la matrice de densité d'un système de n-qubits est une matrice $2^n \times 2^n$ à coefficients complexes, hermitienne, positive et de trace 1. Il faut donc que notre loi *a priori* ne charge que de telles matrices. De plus, elle doit conduire à des estimateurs bayésiens calculables

en pratique. Enfin, elle doit particulièrement favoriser les matrices de faible rang.

Notre construction repose là encore sur l'idée de factorisation. En fait, pour une matrice hermitienne, on a la diagonalisation

$$\rho = UDU^*$$

où U est une matrice unitaire et

$$D = \begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{2^n} \end{pmatrix}.$$

Une façon de définir la loi sur ρ est de définir une loi sur U et une loi indépendante sur $(a_1, \ldots, a_{2^n})$. Si cette dernière induit de la sparsité, alors ρ sera probablement de faible rang. Ceci conduirait à la décomposition

$$\rho = \sum_{j=1}^{2^n} a_j U_{\cdot j} (U_{\cdot j})^*. \tag{10}$$

Le problème est que définir une loi sur les matrices unitaires ne conduit pas à des lois faciles à traiter computationnellement. Dans le Chapitre 3, on propose de relâcher cette contrainte sur U. On montre que la loi que l'on fabrique charge quand même les matrices de densité de faible rang. Cette construction, avec les deux pseudo-vraisemblances mentionnées précédemment, conduit à la construction de nos estimateurs $\tilde{\rho}_\lambda^{prob}$ et $\tilde{\rho}_\lambda^{dens}$. En particulier, on prouve le résultat suivant

Théorème 0.8. *Fixons $\epsilon \in (0,1)$. Alors, pour $\lambda = \lambda^* := m/2$, avec probabilité au moins $1-\epsilon$ on a*

$$\|\tilde{\rho}_{\lambda^*}^{prob} - \rho^0\|_F^2 \leq C \frac{3^n \mathrm{rank}(\rho^0) \log\left(\frac{\mathrm{rank}(\rho^0)N}{2^n}\right) + (1.5)^n \log(2/\epsilon)}{N},$$

où C est une constante numérique.

Une vitesse (moins bonne) est aussi démontrée pour $\tilde{\rho}_\lambda^{dens}$. On propose une implémentation basée sur des algorithmes MCMC. Le Chapitre 3 contient aussi des tests sur des données simulées, et sur des vraies données. Sur les données simulées, il apparaît clairement que nos résultats théoriques sont vraisemblablement dans le bon sens: $\tilde{\rho}_\lambda^{prob}$ se comporte toujours mieux. D'un autre côté, les MCMC pour le calcul de $\tilde{\rho}_\lambda^{dens}$ convergent beaucoup plus rapidement, et donc cet estimateur reste intéressant pour des cas de très grande dimension n où $\tilde{\rho}_\lambda^{prob}$ ne pourrait peut-être pas être calculé du tout...

0.5.3 Remarques bibliographiques

Nous avons considéré le cas où toutes les observables de Pauli avaient été effectivement mesurées. Une partie de la littérature, plus proche du problème

de *compressed sensing*, s'intéresse au cas où seule une partie des mesures sont effectuées. Cf. par exemple [Gross, 2011; Gross et al., 2010; Flammia et al., 2012; Koltchinskii, 2011; Koltchinskii and Xia, 2015; Xia and Koltchinskii, 2016; Xia, 2017].

0.6 *Lifelong learning* dans un contexte en-ligne

0.6.1 Motivation et formalisation

La plupart des algorithmes d'apprentissage proposés partent toujours du principe qu'à chaque nouveau problème, un échantillon est donné et on commence l'apprentissage "de zéro". Cependant, la vie réelle regorge d'exemples où on ne recommence jamais de zéro, car on utilise de façon implicite l'information qui vient de tâches précédentes. Un exemple facile est celui de la reconnaissance des formes en imagerie: un objectif est de fabriquer un dictionnaire de fonctions invariant par le plus de transformations géométriques possibles, qui permette une représentation naturelle des images et soit adapté pour la classification. Si on reçoit séquentiellement plusieurs échantillons, chacun destiné à résoudre un problème de classification données (images de chiens contre autres images, images d'oiseaux contre autres images etc.) il est sensé d'utiliser *toutes* les images pour construire le dictionnaire, et donc à l'arrivée d'un nouveau jeu de données (chat contre autres), l'apprentissage ne recommence *pas* à zéro. Cette idée est au coeur de la notion d'apprentissage par transfert, ce qui signifie que l'on transfère de l'information d'une tâche à une autre, cf. [Thrun and Pratt, 1998; Baxter, 1997, 2000; Cavallanti et al., 2010; Maurer, 2005; Maurer et al., 2013; Pentina and Lampert, 2014; Balcan et al., 2015; Galanti et al., 2016; Maurer et al., 2016] et les références mentionnées dans ces papiers.

A la différence des Chapitres 2 et 3, on s'attaque ici à un problème dont la formalisation n'est pas encore complètement bien établie. Il y a plusieurs versions possibles, suivant que l'on considère que toutes les tâches sont présentées d'un coup, ou séquentiellement, etc. Comme point de départ, nous nous sommes intéressés au cas où les différentes tâches sont proposées séquentiellement (en ligne). Ce contexte a parfois reçu le nom de *lifelong learning*, on gardera ce terme anglais dans cette introduction en français. On propose donc un méta-algorithme de *lifelong learning* qui transfère une partie de l'information d'une tâche à une autre, et on en fournit une analyse théorique, en terme de regret. Le point fort de notre analyse est qu'elle ne dépend pas de l'algorithme utilisé pour résoudre individuellement chacune des tâches.

Décrivons un exemple typique pour fixer les idées.

Un exemple Les tâches sont indéxées par un indice de temps $t \in \{1, \dots, T\}$, le jeu de données de la tâche t, disons

$$\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big) \in (\mathbb{R}^d \times \mathcal{Y})^{m_t}, m_t \in \mathbb{N}$$

sera lui-même révélé séquentiellement. On propose les prédicteurs

$$\hat{y}_{t,i} = \langle \theta_t, D x_{t,i} \rangle$$

où θ_t est appris pour chaque tâche t par un algorithme quelconque. Notre méta-algorithme a pour objectif d'améliorer l'estimation du dictionnaire D à la fin de chaque tâche.

Notations générales A chaque tâche $t \in \{1, \dots, T\}$, le statisticien doit résoudre une tâche d'apprentissage t, à l'aide d'un jeu de données

$$\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big) \in (\mathcal{X} \times \mathcal{Y})^{m_t}$$

où $m_t \in \mathbb{N}$, $\mathcal{X}$ et $\mathcal{Y}$ sont deux ensembles quelconques. Le jeu de données $\mathcal{S}_t$ est lui-même révélé séquentiellement, c'est-à-dire qu'à chaque sous-étape $i \in \{1, \dots, m_t\}$:

- l'objet $x_{t,i}$ est révélé,
- le statisticien doit prédire $y_{t,i}$: un prédicteur est une fonction $f : \mathcal{X} \to \mathcal{Y}$; et on notera $\hat{y}_{t,i} := f_{t,i}(x_{t,i})$ la prédiction du statisticien,
- le label $y_{t,i}$ est révélé et le statisticien subit une perte. Si, pour une paire (x, y), on note comme d'habitude $\ell(f(x), y)$ la perte induite par la prédiction $f(x)$ quand le label est en fait y, le statisticien subit la perte $\ell(\hat{y}_{t,i}, y_{t,i})$.

La tâche t s'arrête à la date m_t, et alors la perte moyenne pour la tâche a été $\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}$. On répète ce processus pour chaque tâche t, et à la fin, la perte moyenne sur l'ensemble des tâches est

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}.$$

Il faut maintenant formaliser en quoi une partie de l'information provenant des tâches $\{1, \dots, t-1\}$ peut être utile pour résoudre la tâche t. Formellement, soit $\mathcal{Z}$ un ensemble et $\mathcal{G}$ un ensemble de fonctions (ou *représentations*) $g : \mathcal{X} \to \mathcal{Z}$. Soit également un ensemble $\mathcal{H}$ de fonctions $h : \mathcal{Z} \to \mathbb{R}$. La stratégie que nous proposerons est faite pour marcher dans le cas où il y a une représentation commune à toutes les tâches, $g \in \mathcal{G}$, et une fonction spécifique à chaque tâche $h_1, \dots, h_T$, telles que

$$f_t = h_t \circ g$$

soit un bon prédicteur pour la tâche t (dans le sens que son erreur moyenne est faible).

Un oracle qui connaîtrait "la bonne" fonction g, et les fonctions $h_1, \dots, h_T$, subirait la perte moyenne

$$\inf_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big).$$

Objectif On peut maintenant formaliser notre objectif. On veut un méta-algorithme qui, au début de chaque tâche t, produit une représentation $\hat{g}_t \in \mathcal{G}$ et permette ensuite d'utiliser n'importe quel algorithme de résolution de la tâche t sur les données

$$\big\{(\hat{g}_t(x_{t,1}), y_{t,1}), \dots, (\hat{g}_t(x_{t,m_t}), y_{t,m_t})\big\}.$$

On souhaite que le *regret combiné* de notre procédure

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i} - \inf_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big)$$

soit le plus petit possible.

Un méta-algorithme On propose la stratégie EWA-LL (cf. l'encadré **Algorithm 1**), basée sur l'idée connue d'agrégation à poids exponentiels ou EWA (voir par ex. [Cesa-Bianchi and Lugosi, 2006] pour l'étude de EWA dans un contexte en ligne). Bien noter que nous utilisont la méthode EWA pour l'apprentissage de g, mais que le choix de la méthode d'apprentissage des h_t n'est pas imposé.

Hypothèse (1) On suppose que l'algorithme utilisé par le statisticien a lui-même un regret contrôlé, c'est-à-dire que l'on a une borne $\beta(g, m_t)$ qui vérifie pour tout g

$$\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) \leq \beta(g, m_t) < \infty.$$

Cette hypothèse est vérifié par plusieurs algorithmes: EWA, gradient en ligne... on renvoie le lecteur aux excellentes introductions [Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012] aux algorithmes de prédiction en ligne et au contrôle de leur regret. Des exemples précis sont détaillés dans le Chapitre 4.

Dans le Chapitre 4, on prouve le résultat suivant.

Théorème 0.9. *Sous l'Hypothèse* **1***) et si la fonction de perte est bornée* $g \in \mathcal{G}$, $\hat{L}_t(g) \in [0, C]$, *on a*

Algorithm 1 EWA-LL

Données Une séquence de jeux de données $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big)$, $1 \leq t \leq T$ associés à différentes tâches, et eux-mêmes révélés séquentiellement;

A fixer Une loi *a priori* π_1 sur $\mathcal{G}$, un paramètre d'apprentissage $\eta > 0$ et un algorithme d'apprentissage à l'intérieur de chaque tâche t qui, pour une représentation g donnée, retourne une suite de prédictions $\hat{y}^g_{t,i}$ et subit la perte moyenne:

$$\hat{L}_t(g) := \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\left(\hat{y}^g_{t,i}, y_{t,i}\right).$$

Boucle Pour $t = 1, \dots, T$

i Tirer $\hat{g}_t \sim \pi_t$.

ii Utiliser l'algorithme d'apprentissage de la tâche t sur $\mathcal{S}_t$ et subir la perte $\hat{L}_t(\hat{g}_t)$.

iii Mettre à jour

$$\pi_{t+1}(\mathrm{d}g) := \frac{\exp(-\eta \hat{L}_t(g))\pi_t(\mathrm{d}g)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m_t}\sum_{i=1}^{m_t} \hat{\ell}_{t,i}\right] \leq \inf_{\rho}\Bigg\{ \mathbb{E}_{g\sim\rho}\left[\frac{1}{T}\sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) + \frac{1}{T}\sum_{t=1}^{T} \beta(g, m_t)\right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T}\Bigg\},$$

où l'infimum est pris sur toutes les mesures de probabilités ρ.

Remarquons que cette borne est uniquement vraie en espérance sur le tirage des différents g_t. Cependant, il est possible de donner des versions vraies en grande probabilité, cf. le Chapitre 4. De plus, quand la fonction de perte est convexe, on peut également donner des versions uniformément vraies, à condition de remplacer le tirage de g_t par une agrégation. Dans le Chapitre 4, Section 4.5, on applique cette borne à deux cas particuliers: l'exemple "jouet" où $\mathcal{G}$ et $\mathcal{H}$ sont deux ensembles finis, et l'exemple de l'apprentissage de dictionnaire. D'autres exemples intéressants pourront être traités à l'avenir: apprentissage de noyau pour des SVM, de couches profondes pour un réseau de neurones...

Comme nous l'avons dit au dessus, la formalisation de ces problèmes est encore un travail en cours (des variantes, par exemple où les jeux de données $\mathcal{S}_t$ sont données d'un coup et non pas séquentiellement, sont aussi discutées, et le problème de savoir quelle formalisation du problème est la meilleure dépend probablement de l'application qui sera considérée). Donc, bien entendu, les

vitesses optimales dans ce problème ne sont pas connues, et ceci pourra faire l'objet d'un travail à venir. Cependant, dans un de nos exemples ($\mathcal{G}$ et $\mathcal{H}$ finis, et $m_t = m$ pour tout t pour faire simple), on améliore la meilleure vitesse connue $1/\sqrt{T}+1/\sqrt{m}$ [Pentina and Lampert, 2014] pour obtenir $1/\sqrt{T}+1/m$ sous certaines hypothèses.

Chapter 1

INTRODUCTION

1.1 Motivation

In many applications of statistics, the objective is to estimate a high-dimensional matrix from noisy and possibly incomplete observations. The size of the matrix and of the datasets is huge (i.e. with billions of entries). Also, the matrix itself is often coming with many complex constraints. These usually prohibit the use of classical methodologies. One of the biggest challenge facing modern statistics is the development of the next generation of methodology and statistical theory to allow inference for such massive datasets. The objective of this thesis is to tackle this challenge.

A common approach to reduce the dimension of the problem is inspired by the sparsity assumption in high-dimensional linear regression model: the essence of this approach is to assume that the matrix is low-rank - or, at least, can be well approximated by a low-rank matrix. Note that low-rankness is a property of the whole matrix that is contrast to sparsity in a vector which is a property of the individual components. More precisely, the columns (rows) of a low-rank matrix can be interpreted as linear combinations of a small number of (unknown) basis vectors. This is in accordance with many statistical models where one explains the observations by a small number of hidden (or latent) variables.

In many practical problems, the target matrix we wish to infer is actually low-rank or approximately low-rank. As a motivating example, the famous Netflix data matrix [Bennett and Lanning, 2007] of user-ratings was modelled as low-rank because it is commonly believed that the users's taste or preferences are similar. Moreover, low-rank matrix estimation is a crucial key in many applications such as: dictionary learning [Kreutz-Delgado et al., 2003; Mairal et al., 2009; Tosic and Frossard, 2011], matrix completion [Candès and Recht, 2009; Keshavan et al., 2009; Koltchinskii et al., 2011; Candes et al., 2015; Kapur et al., 2016]; principal component analysis [Wright et al., 2009; Bro and Smilde, 2014; Zou et al., 2006], high-dimensional covariance/precision matrix [Fan et al., 2008; Pourahmadi, 2013; Cai et al., 2016; Lounici, 2014],

quantum tomography [Gross et al., 2010; Gross, 2011; Flammia et al., 2012; Liu et al., 2012].

Various methods have been proposed and studied over the years for low-rank matrix problems. The most popular methods rely on (efficient) convex optimization algorithms. More precisely, these methods are based on minimizing a sum of two criteria: a measure of the quality of data fitting, and a penalization term that is added to avoid over-fitting. A natural penalty term for low-rank matrix inference is the rank of the matrix. However, the rank is not a convex function of a matrix and thus the use of convex relaxation penalties on the rank, such as the matrix nuclear-norm, are preferred for computational reasons.

Along the side of the journey, Bayesian approaches have also been considered for this type of problems. Rather than considering a point estimate as in the frequentist approach, the Bayesian approach provides a probability distribution on target matrices. There the low-rank structure is incorporated through a prior distribution. Increasing computational power, together with the development of new algorithms, helped Bayesian methods to become more and more popular for such high-dimensional problems. However, there was little theoretical work on the statistical performances of Bayesian methods for low-rank matrix estimation. In this thesis, we focus not only on proposing Bayesian-type estimators but also on studying their statistical properties.

The first two parts of the thesis are dedicated to two practical problems of estimation of low-rank matrices: the matrix completion problem and quantum state tomography, where the objective is to estimate the so-called density matrix, that is often assumed to be low-rank by physicists. For matrix completion, we show that a quasi-Bayesian estimator satisfies an optimal oracle inequality, and thus reaches the minimax-optimal rates (up to log terms). The strong point of our results is that it holds without any assumption on the sampling distribution - this is the first result without such an assumption up to our knowledge. For the quantum state tomography problem, we build a pseudo-Bayesian estimator. Note that in most previous works, the definition of a prior probability distribution was only tackled in the case of the 1 qubit problem (the smallest possible instance of the problem, where the matrix to be estimated is 2×2). Inspired by the prior used for matrix completion, we propose a prior distribution that can be used to estimate density matrices of any dimension. We show that our pseudo-Bayesian estimator reaches the best up-to-date known rate of convergence while its numerical performance was tested on simulated and real data sets.

In the last part of the thesis, we investigate the lifelong learning problem which appeared in artificial intelligence and machine learning. Succinctly, we study the problem of transferring the knowledge learned from previous similar tasks to a new one, where tasks are revealed sequentially. For example, each task can be a high-dimensional linear regression problem. A way to reduce the dimension of the problem is to learn an efficient dictionary of features across tasks, and to estimate the corresponding parameters within tasks. It

is possible to represent such a dictionary of vectors as a matrix, and thus the problem is somehow related to matrix estimation. However, note that the lifelong learning problem is more general than dictionary learning, and we study it in full generality in the thesis.

Beside the Introduction, the thesis includes three self-contained chapters. It is organized as follows:

Chapter 2: We study the matrix completion problem. In this problem, we want to reconstruct a matrix from noisy and incomplete observations of its entries. Through introducing a novel prior distribution, we propose a pseudo-Bayesian estimator for this problem. We show that this estimator reaches the minimax-optimal rate of convergence under general sampling distribution. We also perform numerical tests for this estimator on simulated datasets and compare it with other popular Bayesian estimators.

Chapter 3 deals with quantum state tomography. We study density matrix estimation from data obtained by quantum state tomography, where full measurements are repeated. The problem has been studied under the low-rank assumption. We propose a novel prior distribution and introduce two Bayesian type estimators based on pseudo-likelihoods. The rate of convergence for these estimators are obtained: one of the estimators reaches the best up-to-date known rate, the other is consistent. Here again we compare the numerical performances of our estimators to the ones of the most popular methods, on simulated and real data.

Chapter 4 is dedicated to lifelong learning. Typically, this problem is an online scenario of transfer learning where we want to transfer the information gained from previously learned tasks to a new one, under the assumption that there is a structural similarity between tasks. Assuming that an estimation method is already chosen in order to solve each task (linear regression, online gradient, etc.) we propose a meta-algorithm to transfer information between tasks. It is based on the exponentially weighted aggregation procedure (EWA). The statistical performance of the algorithm is warranted through a regret bound. Some applications of our procedure are also given, including dictionary learning.

The rest of this introduction is organized as follows. In Section 1.2, we briefly review several popular ways to define prior distributions on matrices inducing low-rankness. In Section 1.3, we introduce PAC-Bayesian inequalities, the main theoretical tool of Chapters 2 and 3. Then, in the Sections 1.4, 1.5 and 1.6, we provide an overview of the results of Chapters 2 (matrix completion), Chapter 3 (quantum state tomography) and Chapter 4 (lifelong learning).

1.2 Prior distributions for low-rank matrices

We remind briefly that the idea of Bayesian statistics is to encode the prior information on parameters (or the complexity of the parameter space) through a prior distribution $p(\mathrm{d}\theta)$. Inference is then done through the posterior distribution

$$p(\mathrm{d}\theta \mid \text{data}) \propto \mathcal{L}(\text{data} \mid \theta) p(\mathrm{d}\theta), \tag{1.1}$$

where $\mathcal{L}(\text{data} \mid \theta)$ stands for the likelihood. In this thesis, we will mostly consider the so-called pseudo-Bayesian estimators, where the likelihood $\mathcal{L}(\text{data} \mid \theta)$ is replaced by a more general term depending on a loss function. But this will be discussed in Section 1.3. We discuss the role of the prior distribution $p(\mathrm{d}\theta)$ first. Undoubtedly, $p(\mathrm{d}\theta)$ plays a crucial role in the inference. There is a lot of studies and discussion on choosing prior distributions in general or in different specific problems. Hereafter, we give a short discussion on prior distributions for matrix estimation and then we review several ways to induce low-rankness through an adequate prior.

Naturally, depending on the context of the considering problems, one could define directly a matrix distribution for the target matrix. The popular choices for matrix distributions are Matrix Normal distribution and Wishart distribution as examples, others can be found for example in [Gupta and Nagar, 1999].

Let's assume for instance that the observation come from a Matrix Normal distribution, $\mathbf{X}_{m_1 \times m_2} \mid \mathbf{M}, \Phi, \Sigma \sim \mathcal{N}(\mathbf{M}, \Phi \otimes \Sigma)$ with Φ and Σ either known or unknown. Then the likelihood is

$$\mathcal{L}(\mathbf{X} \mid \mathbf{M}, \Phi, \Sigma) = \frac{\exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T \Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right)}{(2\pi)^{m_1 m_2/2} |\Sigma|^{m_1/2} |\Phi|^{m_2/2}}. \tag{1.2}$$

In the case we are dealing with estimating the mean matrix $\mathbf{M}$, then we obtain "a form" for the posterior distribution of $\mathbf{M}$ as

$$p(\mathbf{M}|\mathbf{X}) \propto \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T \Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right) p(\mathbf{M}).$$

This suggests that we should choose our prior distribution for $\mathbf{M}$ from the Matrix Normal family as follows

$$p(\mathbf{M}) = p(\mathbf{M} \mid \Phi_1, \Sigma_1) \propto \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma_1^{-1}(\mathbf{M}-M_0)^T \Phi_1^{-1}(\mathbf{M}-M_0)\right]\right),$$

where the quantities M_0, Φ_1 and Σ_1 are hyperparameters to be assessed. Then the prior and the likelihood are conjugate distributions, and the posterior distribution $p(\mathbf{M}|\mathbf{X}, \Phi_1, \Sigma_1)$ is itself a Matrix Normal distribution, whose parameters can be explicitly computed.

On the other hand, when our goal is to estimate the covariance matrix Σ (similar for Φ), then

$$p(\Sigma|\mathbf{X}) \propto |\Sigma|^{-\frac{m_1}{2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1}(\mathbf{X}-\mathbf{M})^T \Phi^{-1}(\mathbf{X}-\mathbf{M})\right]\right) p(\Sigma).$$

This implies that we should select a prior distribution for Σ from the Inverted Wishart family as follows

$$p(\Sigma) = p(\Sigma \mid Q, \nu) \propto |\Sigma|^{-\frac{\nu}{2}} \exp\left(-\frac{1}{2} \operatorname{tr}\left[\Sigma^{-1} Q\right]\right),$$

where Q and ν are hyperparameters that need to be specified. Note moreover that the Inverted Wishart family is indeed a family of probability distributions defined over positive-definite matrices random variables, thus it is a sensible prior for covariance matrices. Here again, we have a conjugate prior. Further discussions on these type of prior distributions for matrices can be found for example in [Rowe, 2002].

However, there is no reason for these priors to induce low-rank, or approximately low-rank, matrices $\mathbf{M}$ and Σ. Thus, the conjugacy approach, popular for computational reasons, seems to be of no help in our setting. Still, a nice trick allows to distort conjugate priors to induce low-rank matrices. It is described in the next subsection.

1.2.1 Low-rank through correlation

Remark that a low-rank (or approximately low-rank) matrix has linearly dependent rows/columns. In probabilistic terms, it means that its rows/columns are highly correlated. Thus a careful choice for the column (or row) covariance matrix Φ, or Σ, or both would encourage approximately low-rank structure in the matrix. This can be done by defining further a hyper prior for Φ or Σ, or both.

A way to encourage low-rank is thus directly based on the correlation of the columns (or the rows) of the matrix. Let us define a Matrix Normal distribution (the probability density distribution in form of (1.2)) for the target low-rank matrix $\mathbf{M}_{m_1 \times m_2}$, i.e $\operatorname{rank}(\mathbf{M}) \ll \min(m_1, m_2)$,

$$\mathbf{M} \mid \mathbf{M_0}, \Phi_1, \Sigma_1 \sim \mathcal{N}(\mathbf{M_0}, \Phi_1 \otimes \Sigma_1),$$

where $\mathbf{M_0}$ is the mean matrix; Φ_1 and Σ_1 are respectively the row and column covariance matrix. In the extreme case when the precision matrix Φ_1^{-1} or Σ_1^{-1} (or both) has low-rank, the matrix $\mathbf{M}$ also enjoys a low-rank structure approximately. In order to impose low-rank, thus, one can define a prior distribution for the precision matrices which induces rank deficiency [Sundin et al., 2016].

This approach is known as precision based models or relevance singular vector machine (RSVM). A comprehensive study of this approach can be found in [Sundin, 2016] and references therein. Clearly, using this approach means that we transfer the problem of defining low-rank prior for the target matrix to the problem of defining low-rank prior for the precision matrix. Although showing some interesting numerical results, RSVM method suffers from high computational complexity and the development of the method for larger scale problems is still an open problem. Thus, we will consider a completely different approach in what follows, based on matrix factorization.

1.2.2 Low-rank via factorization

A popular way to promote low-rank is based on matrix factorization approach, where the matrix is modelled as a product of two smaller matrices.

Remind that any matrix M of size $m_1 \times m_2$ and rank-K can be decomposed in the following way by considering the singular value decomposition

$$M = USV^T = (US^{\frac{1}{2}})(S^{\frac{1}{2}}V^T),$$

where U, V are respectively $m_1 \times K$ and $m_2 \times K$ matrices with orthogonal columns, and S is a $K \times K$ diagonal matrix of the non-zero singular values. Letting A and B denote respectively $(US^{\frac{1}{2}})$ and $(VS^{\frac{1}{2}})$, we obtain

$$M = AB^T \tag{1.3}$$

where A is $m_1 \times K$ and B is $m_2 \times K$. The main idea of factorized priors is to define priors on A and B rather than on M directly. To our knowledge, a first Bayesian treatment of this type was carried out in [Geweke, 1996] where the author studied the reduced rank regression model in econometrics with factorized priors.

The major issue in the factorization approach is that the reduced rank K needs to be known in advance. Naturally, one can estimate A and B for any possible K and then use an information criterion for model selection, e.g Bayes factors as in [Kleibergen and Paap, 2002]. Numerical approximation and evaluation of convergence for this method can be found in [Corander and Villani, 2004].

A rank-adaptive strategy has been introduced recently by taking a large K, as $K = \min(m_1, m_2)$. The prior on the A and B is then chosen in order to induce a shrinkage on the columns of these matrices, leading to approximately low-rank matrices. To our knowledge, [Lim and Teh, 2007] presented the first attempt in this direction and then various improved versions had been proposed, e.g [Salakhutdinov and Mnih, 2008; Zhou et al., 2010; Babacan et al., 2011, 2012] by putting prior on hyperparameters. In detail, since we are looking for a low-rank estimate of M, one can carry out it by proposing column sparsity in A and B (i.e most columns in A and in B are set equal to zero). Formally, it is clear in (1.3) that M is the sum of outer-products the columns of A and B, that is

$$M = \sum_{j=1}^{K} A_{\cdot j} B_{\cdot j}^T, \tag{1.4}$$

where $A_{\cdot j}$ and $B_{\cdot j}$ denote the j^{th} column of A and B respectively. Thus one can associate the columns of A and B with some prior distribution enforcing sparsity. For example [Babacan et al., 2012] used Gaussian priors of variances γ_i, that is

$$p(A|\gamma) = \prod_{j=1}^{K} \mathcal{N}(A_{\cdot j}|0, \gamma_j I),$$

$$p(B|\gamma) = \prod_{j=1}^{K} \mathcal{N}(B_{\cdot j}|0, \gamma_j I).$$

Moreover, they modelled the $\gamma_i's$ as random according to a distribution highly concentrated around zero. For computational reasons, conjugate prior distributions for $\gamma_i's$ are widely used and a popular choice in the literature is that $1/\gamma_i \sim \Gamma(a, b)$ (Gamma distribution) with a very small b. Then, most γ_i's are very close to 0, and so are most of the columns $A_{\cdot j}$ and $B_{\cdot j}$. So, most of the terms in (1.4) are almost null. Thus, M is very close to a low-rank matrix.

In Chapters 2 and 3, we define prior distributions for the matrix completion problem, and for quantum tomography. None of these priors are exactly equal to the ones proposed by [Babacan et al., 2012], but in both cases, our construction are based on the factorization+shrinkage approach as in (1.4).

1.3 A short introduction to PAC-Bayesian analysis

As mentioned earlier, one of the main objective of this thesis is to explore the statistical properties of Bayesian and pseudo-Bayesian estimators. Many theoretical approaches are available for this.

A first idea is to prove asymptotic concentration of the posterior around the true value of the parameter to be estimated (here, a matrix). This approach is for example described in [Ghosal et al., 2000a] for nonparametric models, a very nice review can be found in [Rousseau, 2016].

A second approach consists in studying the MAP (*maximum a posteriori*) using the theory of penalized risk minimization (based on concentration inequalities). For example, the LASSO estimator, first introduced and studied as a frequentist estimator, can be seen as a MAP with a Gaussian likelihood and a Laplace prior. Recently, [Abramovich and Grinshtein, 2010; Abramovich and Lahav, 2015] proved minimax-optimal rates for a Bayesian MAP estimator in the sparse regression and the nonparametric additive model respectively, using a concentration inequality from [Birgé and Massart, 2001].

In this thesis, we focus on another approach relying on PAC-Bayesian inequalities. While this approach definitely share some similarities with the approach based on concentration inequalities, the results are different in nature. The main difference with the aforementioned approaches is that we do not assume any statistical model on the observations - or, if we do, we still want our estimator to work well in case of misspecification. Thus the approach is primarily based on techniques from the machine learning community. PAC-Bayesian bounds were pioneered by [Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999], and then by [Catoni, 2004, 2007]. Some compreshensive surveys of this technique for different aspects in statistics and machine learning can be found in [Audibert, 2004; Alquier, 2006; Guedj, 2013; Germain, 2015]. Generally, this approach connects the generalization ability of an aggregation distribution to its empirical risk and to its Kullback-Leibler

divergence with respect to a prior distribution. Minimizing this criterion usually leads to an exponentially weighted aggregation (EWA), that is an aggregation distribution of the form (1.1) but where the likelihood $\mathcal{L}(\text{data} \mid \theta)$ is replaced by an exponential function of the empirical risk. We now provide a short user-friendly introduction to PAC-Bayes bounds.

1.3.1 Basic set-up

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$ be n independent observed pairs, where $\mathcal{X}$ is any measurable set and $\mathcal{Y} = \{-1, +1\}$ for classification or $\mathcal{Y} = \mathbb{R}$ for regression. We denote $\mathcal{D} := (X_i, Y_i)_{i=1}^n$ (as data) for convenience. Each example is sampled from a distribution denoted by $\mathbb{P}$ (unknown) and the corresponding expectation is denoted by $\mathbb{E}$.

The statistician consider a set of predictors (hypothesis) $\mathcal{H} := \{f_\theta : \mathcal{X} \to \mathcal{Y}; \theta \in \Theta\}$, which yields a real value loss $\ell(Y_i, f_\theta(X_i))$ on the i-th example. For example, one may use the squared loss $\ell(Y_i, f_\theta(X_i)) = (Y_i - f_\theta(X_i))^2$ in the regression model. The expected risk (error) of f_θ is defined by

$$R(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\ell(Y_i, f_\theta(X_i)).$$

Note that this quantity is unknown (because $\mathbb{P}$ is unknown), but its empirical counterpart, the empirical risk (error) of f_θ, can be computed on the data:

$$r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f_\theta(X_i)).$$

The "classical" approach focuses on estimators $\hat{\theta} : (\mathcal{X} \times \mathcal{Y})^n \to \Theta$ and investigate the relationship between the empirical risk $r(\hat{\theta})$ and the expected risk $R(\hat{\theta})$. The PAC-Bayesian approach focuses on functions $\rho : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}_+^1(\Theta)$ where $\mathcal{M}_+^1(\Theta)$ is the set of all probability distributions on Θ equipped with some suitable σ-algebra $\mathcal{T}$. Depending on the situation, we can then provide guarantees on the estimator defined as the mean under ρ, $\hat{\theta} = \int \theta \rho(\mathrm{d}\theta)$, or on a randomized estimator $\hat{\theta}$ drawn directly from ρ.

An empirical PAC-Bayesian bound: A basic PAC-Bayesian bound is as follows.

Theorem 1.1 (e.g. Theorem 2.3 in [Alquier, 2006])**.** *Fix a probability distribution $\pi \in \mathcal{M}_+^1(\Theta)$. Assume that the loss ℓ takes the values in $[0, C]$ for a constant $C > 1$. For any $\lambda \in (0, n/C)$, with probability at least $1-\varepsilon, \varepsilon \in (0,1)$ and for any $\rho \in \mathcal{M}_+^1(\Theta)$*

$$\int_\Theta R(\theta)\rho(d\theta) \leq \int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n}. \tag{1.5}$$

We remind that $\mathcal{K}(\rho, \pi)$ stands for the Kullback-Leibler divergence between ρ and π, given by

$$\mathcal{K}(\rho,\pi) = \begin{cases} \int \log\left(\frac{d\rho}{d\pi}\right) d\rho, & \text{when } \rho \text{ is absolutely continuous w.r.t } \pi, \\ +\infty, & \text{otherwise.} \end{cases}$$

For the sake of simplicity, let us denote for short $\nu(h) = \int_\Theta h(\theta)\nu(d\theta)$. We state an important lemma from which we can deduce an optimal candidate for the left-hand side of the above theorem.

Lemma 1.1. *For any bounded measurable function $h : \Theta \to \mathbb{R}$ and for any $\rho \in \mathcal{M}_+^1(\Theta)$ such that $\mathcal{K}(\rho, \pi) < \infty$, we have*

$$-\log \pi[\exp(h)] = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} [-\rho(h) + \mathcal{K}(\rho, \pi)]$$

Particularly, the infimum in the right-hand side is reached for the Gibbs distribution $\rho_{\exp(h)}$ defined by

$$\frac{d\rho_{\exp(h)}}{d\pi}(\theta) = \frac{\exp(h(\theta))}{\pi(\exp(h))}.$$

Proof. We have

$$\begin{aligned} \mathcal{K}(\rho, \rho_{\exp(h)}) &= \rho\left(\log\left(\frac{d\rho}{d\pi}\right) - h\right) + \log \pi[\exp(h)] \\ &= \mathcal{K}(\rho,\pi) - \rho(h) + \log \pi[\exp(h)]. \end{aligned}$$

The left-hand side of this equation is non-negative and vanishes only for $\rho = \rho_{\exp(h)}$ (note that this equation is still valid if ρ is not absolutely continuous w.r.t π: says $+\infty = +\infty$). So we get

$$0 = \inf_{\rho \in \mathcal{M}_+^1(\Theta)} [\mathcal{K}(\rho,\pi) - \rho(h)] + \log \pi[\exp(h)].$$

□

A Bayesian interpretation: From Lemma 1.1, we deduce the optimal distribution for the right-hand side of (1.5) is of the form

$$\hat{\rho}_\lambda(d\theta) = \frac{\exp\{-\lambda r(\theta)\}}{\pi(\exp\{-\lambda r(\theta)\})}\pi(d\theta).$$

So, we have

$$\hat{\rho}_\lambda(d\theta) \propto \mathcal{L}(\text{data} \mid \theta) p(\mathrm{d}\theta)$$

as in (1.1), with $\mathcal{L}(\text{data} \mid \theta) = \exp\{-\lambda r(\theta)$ and $\pi(d\theta) = p(\mathrm{d}\theta)$. More precisely, $\exp\{-\lambda r(\theta)\}$ plays the role of a likelihood, $\pi(d\theta)$ can be interpreted as a prior distribution, and λ is a tuning parameter which balances between

the empirical information and the prior. We can use by extension the term "pseudo-likelihood" to refer to $\exp\{-\lambda r(\theta)\}$, and thus we will use the term "pseudo-Bayesian estimator" for any estimator based on $\hat{\rho}_\lambda$.

Note that in this case (1.5) becomes

$$\int_\Theta R(\theta)\hat{\rho}_\lambda(d\theta) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \left[\int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n} \right]. \quad (1.6)$$

Remark that when the loss function ℓ is convex we can apply the Jensen's inequality to get

$$R\left(\int_\Theta \theta\rho(d\theta)\right) \leq \rho[R(\theta)].$$

This means that we are able to upper bound $R(\theta)$ for the estimator of the form

$$\hat{\theta}_\lambda := \int_\Theta \theta\hat{\rho}_\lambda(d\theta).$$

By considering the mean estimator above, we are able to deduce from (1.6) a bound for the expected risk of $\hat{\theta}_\lambda$ as

$$R(\hat{\theta}_\lambda) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \left[\int_\Theta r(\theta)\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2n} \right].$$

Oracle-type PAC-Bayesian inequality: The upper bound in (1.5) can be computed from the data and thus yields a way to evaluate the performance of the estimation. This was the idea of the first PAC-Bayesian bounds [Shawe-Taylor and Williamson, 1997; McAllester, 1998, 1999].

However, the rate of convergence of the estimator can not be directly obtained in the empirical bounds. This motivates the study of oracle-type inequalities, which were developed by [Catoni, 2004, 2007]. More precisely, PAC-Bayesian analysis can also be used to compare $\int_\Theta R d\hat{\rho}_\lambda$ to the best possible risk. A simple oracle PAC-Bayesian inequality is as follows.

Theorem 1.2. *Under the same assumptions as in the previous theorem, for any $\lambda \in (0, n/C)$, with probability at least $1-\varepsilon, \varepsilon \in (0,1)$*

$$\int_\Theta R(\theta)\hat{\rho}_\lambda(d\theta) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta)} \left\{ \int_\Theta R(\theta)\rho(d\theta) + 2\frac{\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{n} \right\}. \quad (1.7)$$

A trick to obtain an explicit bound is as follows: in the right hand-side, we can consider a special $\tilde{\rho}$ (say in a parametric family) concentrated around the value of θ minimizing $R(\theta)$, so that $\int_\Theta R(\theta)\tilde{\rho}(d\theta) \simeq \inf_\theta R(\theta)$. In some situations, if $\tilde{\rho}$ is too concentrated, this will cause the Kullback-Leibler

term to explode. An adequate balance between the terms $\int_\Theta R(\theta)\tilde{\rho}(d\theta)$ and $\mathcal{K}(\tilde{\rho}, \pi)$ will often lead to an explicit rate of convergence in the right-hand side. Such techniques were used by [Dalalyan and Tsybakov, 2008; Alquier and Lounici, 2011] to derive optimal rates in sparse regression estimation (using more sophisticated PAC-Bayesian inequalities than the ones presented in this introduction).

Some examples of recent advances on PAC-Bayesian bounds include [Seldin et al., 2012; Germain et al., 2013; Pentina and Lampert, 2014; Ridgway et al., 2014; Galanti et al., 2016]. Recently [Bégin et al., 2016; Alquier and Guedj, 2016] proposed variants where the Kullback divergence is replaced by another divergence. Most of the aforementioned papers use bounded, or at least sub-exponential loss functions. However, using a robustification technique due to [Catoni, 2012], the papers [Catoni, 2016; Giulini, 2015] proved PAC-Bayesian bounds for the estimation of the Gram matrix in the case of heavy-tailed random variables. Another approach to derive PAC-Bayesian bounds with heavy-tailed loss functions was recently proposed by [Grünwald and Mehta, 2016].

1.4 Overview of our results on matrix completion

1.4.1 Short introduction to matrix completion

Matrix completion has received a lot of attention over the past decade. It consists in restoring a potentially high dimensional matrix M, based on random, (possibly) noisy and partial observations of its entries. The matrix completion problem arises in a wide range of applications such as recommender systems [Bennett and Lanning, 2007; Cai et al., 2010; Melville and Sindhwani, 2011], image processing [Ji et al., 2010; Liu et al., 2013], genomics [Chi et al., 2013; Natarajan and Dhillon, 2014; Cai et al., 2015a].

Consider a toy example of users-movies rating data as in the Table 1.1. Usually, the user does not watch all the available movies and moreover it is not certain that users give the ratings to all the movies they watched. Therefore, the matrix data is obtained with many, many missing entries. In the Netflix prize [Bennett and Lanning, 2007], with 480 189 users and 17 770 movies, only 100 480 507 ratings were observed over the total 8 532 958 530 entries in the matrix, thus less than 1.2% of the entries of the matrix were observed. Definitely, to infer the missing entries is thus very helpful to propose sensible advertisement and improve the sales.

Obviously, it is impossible to make any inference on M in general, because most of the entries of the matrix are unknown. However, a major breakthrough was made by Candès and co-authors who proved that, in the case where M is low-rank, the task becomes possible [Candès and Recht, 2009; Candès and Plan, 2010; Candès and Tao, 2010; Cai et al., 2010]. Note that this assumption makes perfectly sense in the aforementioned applications, like the Netflix data: it is clear that, as some users have similar preferences, their

	Looper	π	Inception	Big Hero 6	...
...	?	1	2	5	...
Aisling	4	?	5	?	...
Bianca	?	5	?	2	...
Tien	5	?	5	?	...
...	1	2	?	4	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 1.1: Example of rating matrix data. Ratings range between 1 and 5.

ratings are proportional to each other or even similar.

Let $M^0 \in \mathbb{R}^{m_1 \times m_2}$ be the target unknown matrix (expected to be low-rank). The matrix completion problem can be expressed as follows. We observe

$$Y_{i,j} = M^0_{i,j} + \varepsilon_{i,j}, (i,j) \in \Omega,$$

where Ω is a random subset of the set $\{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$ with $n = \text{card}(\Omega) \ll m_1 m_2$. The noise variables $\varepsilon_{i,j}$ are independent with $\mathbb{E}(\varepsilon_{i,j}) = 0$.

In the seminal paper [Candès and Recht, 2009], the authors proposed the estimator $\hat{M}$ based on a convex relaxation of the rank, defined by

$$\hat{M} = \arg \min_{A : A_{i,j} = Y_{i,j}, \forall (i,j) \in \Omega} \|A\|_*$$

where $\|A\|_*$ is the nuclear norm of the matrix A:

$$\|A\|_* = \sum_{i=1}^{\min(m_1, m_2)} \lambda_i(A)$$

where $\lambda_i(A)$ are the singular values of A. They proved that, in the noiseless case ($\varepsilon_{i,j} = 0$), we have an exact reconstruction $\hat{M} = M^0$ under a low-rank assumption on M, provided that n is large enough. This result was extended (with a slightly different estimator) to the noisy case in [Candès and Plan, 2010]. Since then, many different methods have been proposed for matrix completion, which mostly based on penalized empirical risk minimization with various penalties. For example: rank penalty (computationally challenging) [Klopp, 2011], von Neumann entropy penalty [Koltchinskii, 2011], Schatten-p norm penalty [Rohde and Tsybakov, 2011], spectral k-support norm [Gunasekar et al., 2015], ... One of the estimators studied in these papers is the so called matrix Lasso

$$\hat{M}_{nuclear} = \arg \min_M \left\{ \frac{1}{n} \sum_{(i,j) \in \Omega} (Y_{i,j} - M_{i,j})^2 + \lambda \|M\|_* \right\},$$

where $\lambda > 0$ is a tuning parameter.

In the notable paper [Koltchinskii et al., 2011], the authors study the so-called "Trace-regression" model. It is a general and abstract model, including matrix completion and linear regression as special cases. They propose an estimator based on nuclear norm penalization and provide the statistical analysis for it. They study a variant $\tilde{M}_{nuclear}$ of the matrix Lasso estimator, and prove the following result.

Theorem 1.3 (Corollary 2 in [Koltchinskii et al., 2011]). *Under some assumptions, with probability at least* $1 - 3/(m_1 + m_2)$

$$\frac{\|\tilde{M}_{nuclear} - M^0\|_F^2}{m_1 m_2} \leq C \frac{\text{rank}(M^0) \max(m_1, m_2)}{n} \log(m_1 + m_2),$$

where C *is a numerical constant and* $\|B\|_F^2 = \text{Trace}(BB^T)$*, the Frobenius norm.*

The author also proved a minimax lower bound for low-rank matrix completion under the Frobenius error.

Theorem 1.4 (Theorem 5 in [Koltchinskii et al., 2011]). *Fix* $a > 0$ *and an integer* $1 \leq r \leq \min(m_1, m_2)$*. Under suitable assumptions, there exist absolute constants* $\beta \in (0, 1)$ *and* $c > 0$ *such that*

$$\inf_{\hat{M}} \sup_{\substack{\text{rank}(M^0) \leq r, \\ \max_{i,j} |M^0_{i,j}| \leq a}} \mathbb{P}_{M^0} \left(\frac{1}{m_1 m_2} \|\hat{M} - M^0\|_F^2 > c \frac{r \max(m_1, m_2)}{n} \right) \geq \beta.$$

Basically, this lower bound states that the average quadratic error on the entries of a rank-r matrix size $m_1 \times m_2$ from n-observations can not be better than $r \max(m_1, m_2)/n$. Note that the upper bound in Theorem 1.3 does not exactly match the lower bound - there is an additional $\log(m_1 + m_2)$ factor. Matching bounds were recently reached by [Klopp, 2015], but in a slightly different model where the sample size n itself is random.

A large part of the studies for matrix completion in the literature is carried out under the assumption that the n observed entries (i, j) are drawn i.i.d from the uniform distribution on $\{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$. However, in practice, the observed entries are not always uniformly distributed: for example, some movies are more famous than others and therefore receive much more ratings. More importantly, the sampling distribution is not known in practice. More general sampling schemes than uniform distribution had been already studied, see e.g. [Foygel et al., 2011; Negahban and Wainwright, 2012; Klopp, 2014], but there are still some assumptions on the sampling distribution in these papers.

1.4.2 Main results of Chapter 2

While penalized minimization methods are well understood, both from a theoretical and from a computational perspective, the first papers on Bayesian

matrix completion were essentially methodological and algorithmics, and contain no rates of convergence or consistency results, see e.g. [Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Lawrence and Urtasun, 2009; Zhou et al., 2010; Babacan et al., 2011; Alquier et al., 2014] among others.

For computational reasons, most Bayesian estimators are based on conjugate priors which allow to use Gibbssampling [Alquier et al., 2014; Salakhutdinov and Mnih, 2008] or Variational Bayes methods [Lim and Teh, 2007]. These priors are discussed in details in [Alquier et al., 2014]. These algorithms are fast enough to deal with large datasets like Netflix or MovieLens[1], and are actually tested on these datasets in those papers. However, as mentioned earlier, the statistical properties of Bayesian estimators are not known.

Our first contribution in this thesis was to design a prior distribution that would lead to consistent and minimax-optimal estimators (up to log terms). We adapt the factorization trick to define a novel adaptive low-rank prior distribution on matrices; the main difference being that we replace the Gaussian distribution for the columns in (1.4) by uniform distribution on segments. The estimator we propose, say $\widetilde{M}$, is the mean of the pseudo-posterior distribution; we remind that, by pseudo-posterior, we mean that the likelihood is replaced by an exponential function of the empirical risk. The exact construction is detailed in Chapter 2. The main result is as follows (we refer the reader to Chapter 2 for an accurate statement of the assumptions).

Theorem 1.5 (Theorem 2.1 in Chapter 2)**.** *Assume that the n observed entries are i.i.d from a distribution $(\pi_{i,j})$, that is, the probability to observe the entry (i,j) is $\pi_{i,j}$. Under suitable assumptions on the noise, and no assumptions on $(\pi_{i,j})$, with high probability and as soon as $n \geq \max(m_1, m_2)$, one has*

$$\sum_{\substack{1 \leq i \leq m_1, \\ 1 \leq j \leq m_2}} (\widetilde{M}_{i,j} - M^0_{i,j})^2 \pi_{i,j} \leq C \frac{\mathrm{rank}(M^0) \max(m_1, m_2)}{n} \log(\min(m_1, m_2)),$$

where C is a numerical constant.

As a special case, when the sampling distribution is uniform, i.e $\pi_{i.j} = 1/m_1 m_2$, we obtain

$$\frac{\|\widetilde{M} - M^0\|_F^2}{m_1 m_2} \leq C' \frac{\mathrm{rank}(M^0) \max(m_1, m_2)}{n} \log(\min(m_1, m_2)),$$

where C' is a numerical constant. As discussed earlier, this rate is minimax-optimal up to log terms. Note the (slight) improvement with respect to [Koltchinskii et al., 2011] as $\log(m_1 + m_2) \asymp \log(\max(m_1, m_2))$ is replaced by $\log(\min(m_1, m_2))$.

From a computational point of view, using an MCMC algorithm, we are able to implement and test our estimator on matrices with sizes up to 1000×1000. An example of our numerical results is given in Table 1.2.

[1] http://grouplens.org/datasets/movielens/

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Unif.	0.535 ($\pm$0.003)	0.348 ($\pm$0.003)	0.207 ($\pm$0.0001)	0.141 ($\pm$0.0006)
Gaus.	0.538 ($\pm$0.001)	0.345 ($\pm$0.001)	0.210 ($\pm$0.0001)	0.146 ($\pm$0.001)

Table 1.2: *RMSEs in the first series of experiments (low-rank matrix, Gaussian noise), m is the size of the square matrix. We compare our estimator with uniform priors on the columns to the "classical" Gaussian prior as in [Salakhutdinov and Mnih, 2008; Babacan et al., 2012].*

1.4.3 Bibliographical notes

Many extensions and/or improvements on the previous results have been published in the recent years. For example, the results in [Koltchinskii et al., 2011] as well as our result assume that we know an upper bound on the variance of the noise. This assumption is not so unrealistic: for example, in the Netflix dataset, the ratings are bounded (between 1 and 5) and so there is an obvious upper bound on the variance. On the other hand, this upper bound might not be very accurate. Moreover, in other applications, the variance might be unknown. Matrix completion without knowing the variance of the noise was tackled in [Klopp, 2014]. More precisely, the proposed estimator for unknown variance of the noise setting is as follows

$$\hat{M}_{SQ} = \arg\min_{M} \left\{ \sqrt{\frac{1}{n} \sum_{(i,j)\in\Omega} (Y_{i,j} - M_{i,j})^2} + \lambda \|M\|_* \right\},$$

where $\lambda > 0$ is a regularization parameter. This estimator can be seen as the matrix analog of the square-root Lasso for matrix [Belloni et al., 2011]. The obtained convergence rate is the same as in [Koltchinskii et al., 2011]. An extension of this idea to the Bayesian setting would be of interest, and could be the object of future works.

Other variants include 1-bit matrix completion [Davenport et al., 2014]: "Instead of observing a subset of the real-valued entries of a matrix M, we obtain a small number of binary (1-bit) measurements generated according to a probability distribution determined by the real-valued entries of M". The study and extension of this model can be found, for example, in [Cai and Zhou, 2013; Klopp et al., 2015; Srebro et al., 2004]. Since the publication of our paper (presented in Chapter 2), this model was also studied from a pseudo-Bayesian perspective by [Cottet and Alquier, 2016].

Robust matrix completion was studied in [Klopp et al., 2014], where the authors use a variant of the penalty term to remove outliers. Another point of view is provided by [Alquier et al., 2017a] where the authors replace the quadratic loss by robust losses (like the absolute loss). Going beyond inference/estimation, the paper [Carpentier et al., 2016] studies the existence of confidence sets for an estimator in the problem of matrix completion.

1.5 Overview of our results on quantum tomography

1.5.1 Short introduction to quantum statistics

Quantum state tomography plays an important role in quantum information processing. It focuses on reconstructing the (unknown) state of a physical quantum system [Paris and Řeháček, 2004]. This task is done by using measurements' outcomes of many independent systems identically prepared in the same state. We refer the reader to the introduction of [Meziani, 2008] or to the survey paper [Artiles et al., 2005] for a general introduction to quantum tomography, and a presentation of the basics concepts in quantum physics.

According to quantum theory, all the information about a physical system is encoded in its quantum state. The outcome of any experimental measured on the system is usually not deterministic, but its probability distribution can be deduced from the state of the system. A mathematical way to encode the state of a system is to use the so-called density matrix ρ. This matrix has complex entries, and

- ρ is Hermitian, $\rho^\dagger = \rho$ (i.e. self-adjoint),
- ρ is semidefinite positive, $\rho \geq 0$,
- $\mathrm{Trace}(\rho) = 1$.

Note that the dimension of ρ depends on the system in hand and can satisfy various additional assumptions. In quantum homodyne tomography model, ρ is an infinite matrix with a regularity assumption: the coefficients $\rho_{i,j}$ of ρ decay exponentially fast in $i + j$. Some studies in this problem with different approaches are available, for example, in [Artiles et al., 2005; Butucea et al., 2007; Alquier et al., 2013b; Naulet and Barat, 2016].

We focus now on the details in quantum computing, i.e the finite case. The system of interest contains of n qubits spin-$\frac{1}{2}$ and the corresponding density matrix ρ is a $2^n \times 2^n$ matrix with coefficients in $\mathbb{C}$. More importantly, physicists are interested in the so-called *pure states* and a pure state ρ can be defined by $\mathrm{rank}(\rho) = 1$. Additionally, it often makes sense in practice to assume that the rank of ρ is small [Gross et al., 2010; Gross, 2011].

It is important for physicists to check their ability to prepare a physical system in a given state ρ_0. In order to test this ability, they produce, using the same device, many independent systems in state ρ (hoping that $\rho = \rho_0$) and perform measurements on these systems. One of the statistical tasks is then to infer ρ from the outcomes of these measurements. The reconstruction of ρ from experimental measurements is called *quantum state tomography*. We refer the reader to [Artiles et al., 2005] for a complete formalization of this statistical problem. We next give more details on the problem set-up.

More precisely, Pauli-observables are commonly used to perform measurements for each qubit. The three Pauli-observables are :

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}; \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

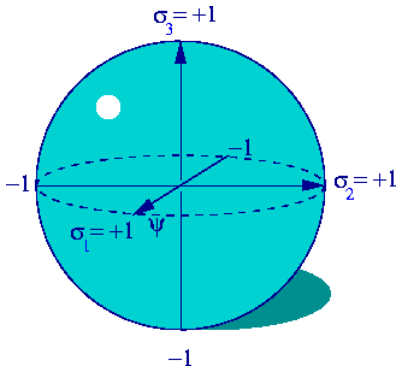

Figure 1.1: *Outcomes of Pauli-measurement of a single spin-$\frac{1}{2}$ particle.* [*source:* https://en.wikipedia.org/wiki/Quantum_indeterminacy]

Thus, on a n-qubit system, there are 3^n possible experimental measurements and each outcome is a vector in $\{-1,1\}^n$. We consider a case where each measurement is repeated m times, on m independently prepared systems. Thus, the quantum sample size is $N = m \cdot 3^n$.

Given vector $\mathbf{s} \in \{-1,1\}^n$, the probability to obtain it as an outcome is a function of the density matrix ρ and of the measurement. This is given by Born's rule

$$M_{i,\mathbf{s}} := \mathbb{P}(R^i = \mathbf{s}) = \text{Trace}\left(\rho \cdot P^i_{\mathbf{s}}\right), i \in \{1,\ldots,3^n\}, \tag{1.8}$$

where $P^i_{\mathbf{s}}$ are given explicitly. This means that there is a linear function such that $M = F(\rho), M = (M_{i,\mathbf{s}})_{i\in\{1,\ldots,3^n\},\mathbf{s}\in\{-1,1\}^n}$.

We provide a detailed toy example, Example 1.1, for the case of 2 qubits so that it will make the problem more clear. In the case of 2 qubits, we have 9 experimental measurements: $(\sigma_x,\sigma_x),(\sigma_x,\sigma_y),(\sigma_x,\sigma_z),\ldots,(\sigma_z,\sigma_z)$; and for each experimental measurement, we have 4 possible outcomes: $(-1,-1)$, $(-1,+1),(+1,-1),(+1,+1)$.

Example 1.1. *Suppose that ρ is such that the probability distributions in each possible measurement is given by*

	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	*0.24*	*0.26*	*0.29*	*0.21*
2	*0.34*	*0.36*	*0.19*	*0.11*
⋮	⋮	⋮	⋮	⋮
9	*0.23*	*0.25*	*0.37*	*0.15*

$=: M = F(\rho)$

where we remind that the matrix $M = F(\rho)$ can be deduced from ρ using Born's rule (1.8). In practice, we measure each observable $i \in \{1, \dots, 9\}$ for $\boldsymbol{m}$ times, e.g 1000. A possible result is:

	$(-1,-1)$	$(-1,+1)$	$(+1,-1)$	$(+1,+1)$
1	*232/1000*	*266/1000*	*291/1000*	*211/1000*
2	*336/1000*	*361/1000*	*192/1000*	*111/1000*
⋮	⋮	⋮	⋮	⋮
9	*233/1000*	*250/1000*	*365/1000*	*152/1000*

$:= \widehat{\mathbf{M}}$

To infer the density matrix ρ, a natural idea is the inversion method that is based on the empirical matrix $\widehat{M}$ and solving for a density $\hat{\rho}$

$$F(\hat{\rho}) = \widehat{M}. \tag{1.9}$$

This method (also known as moment estimation in statistics) is studied in [Vogel and Risken, 1989; Řeháček et al., 2010]. Although it is computationally easy, it returns an estimator $\hat{\rho}$ that is very often not a physical density matrix [Shang et al., 2014].

A popular choice for inferring the density matrix is maximum likelihood estimation. Unfortunately, it has some critical flaws detailed in [Blume-Kohout, 2010]. Furthermore, when additional prior information is available, e.g low-rankness, both these method are not adaptive.

In case of low-rank state estimation, some rank-adaptive procedures had been introduced by using suitable penalization. Rank-penalized maximum likelihood (BIC) was introduced in [Guţă et al., 2012] while a rank-penalized least-square estimator $\hat{\rho}_{\text{rank-pen}}$ was proposed in [Alquier et al., 2013a], together with a proof of consistency. More specifically, when the density matrix of the system is ρ^0 with $r = \text{rank}(\rho^0)$, the authors of [Alquier et al., 2013a] proved that the Frobenius norm of the estimation error satisfies

$$\|\hat{\rho}_{\text{rank-pen}} - \rho^0\|_F^2 = \mathcal{O}(r4^n/N)$$

where N is the number of quantum measurements. The rate was improved to $\mathcal{O}(r3^n/N)$ by [Butucea et al., 2015, 2016], using a thresholding method.

Theorem 1.6 (Corollary 1 in [Butucea et al., 2015])**.** *Under suitable assump-*

tions, with probability higher than $1-\varepsilon, \varepsilon \in (0,1)$

$$\|\hat{\rho}_{\text{rank-pen}} - \rho^0\|_F^2 \leq C\frac{r3^n}{N}\log(2^{n+1}/\varepsilon).$$

Moreover, the question of how good one can estimate a low-rank density matrix is also studied in [Butucea et al., 2015]. More precisely, the paper show that no method can reach a rate smaller than $r2^n/N$. So, the minimax-optimal rate lies somewhere in between $r2^n/N$ and $r3^n/N$. This result is given in the following theorem.

Theorem 1.7 (Lower bound, Theorem 3 in [Butucea et al., 2015])**.**

$$\liminf_{m\to\infty}\inf_{\hat{\rho}_m}\sup_{\rho\in\mathcal{S}_r}\mathbb{E}_\rho\|\hat{\rho}_m-\rho\|_F^2 \geq \frac{2r(2^n-r)}{N}$$

where $\mathcal{S}_r$ *is the set of rank-r density matrices.*

On the other hand, Bayesian estimation has been computationally considered in this context. The papers [Bužek et al., 1998; Baier et al., 2007] compare Bayesian methods to other methods on simulated data. Recently, efficient algorithms for computing Bayesian estimators are discussed in [Kravtsov et al., 2013; Ferrie, 2014; Kueng and Ferrie, 2015; Schmied, 2016]. Importantly, [Blume-Kohout, 2010] showed that Bayesian method satisfies many good properties when estimating the density matrix. However, theoretical study on the convergence of Bayesian estimators has not been done yet. Moreover, the rank adaptation has not been considered from a Bayesian point of view in this problem.

1.5.2 Main results of Chapter 3

We consider pseudo-Bayesian estimations (for computational reasons), where the likelihood is replaced by pseudo-likelihoods based on various moments. Two estimators, corresponding to two different pseudo-likelihood, are actually proposed. Basically, one of them, $\tilde{\rho}_\lambda^{dens}$, relies on the comparison between ρ and $\hat{\rho}$ (the inversion estimator) while the other one, $\tilde{\rho}_\lambda^{prob}$ relies on a comparison between theoretical and empirical frequencies, $F(\rho)$ and $F(\hat{\rho})$. Using PAC-Bayesian theory, we derive oracle inequalities for the pseudo-posterior mean. We obtain rates of convergence for these estimators in the complete measurement setting. One of them has a rate as good as the best known rate up to date $\mathcal{O}(\text{rank}(\rho^0)3^n/N)$ [Butucea et al., 2015]. Meanwhile, the other one is interesting for computational reasons that are discussed in the paper.

One of the key points here is to define a low-rank (approximately) prior on the set of density matrices. Remind that the density matrix of a n-qubits system is a $2^n \times 2^n$, positive, Hermitian complex matrix ρ with $\text{Trace}(\rho) = 1$. Note that we have several restrictions in mind when defining a prior distribution on such matrices. First, we want this prior to lead to feasible

algorithms and then we want it to lead to consistent estimators. Finally, we would like to pay a special attention to rank one matrices.

For the density matrix ρ being Hermitian, we have a diagonalization

$$\rho = UDU^*$$

where U is a unitary matrix and

$$D = \begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{2^n} \end{pmatrix}$$

A way to define our prior on ρ is to define a prior on U and $(a_1, \ldots, a_{2^n})$. If the prior on $(a_1, \ldots, a_{2^n})$ is sparsity-inducing, then ρ is likely to be low-rank. This would lead to the decomposition

$$\rho = \sum_{j=1}^{2^n} a_j U_{\cdot j} (U_{\cdot j})^*. \tag{1.10}$$

However, to define a prior on unitary matrices usually leads to challenging computational constraints when one is to simulate from the posterior. In Chapter 3 we propose a way to relax this constraint that still allows to charge only density matrices, and to give more weights to approximately low-rank matrices. This construction, together with the two pseudo-likelihood described above, allow us to build our estimators $\tilde{\rho}_\lambda^{prob}$ and $\tilde{\rho}_\lambda^{dens}$. Among others, we obtain the following result.

Theorem 1.8. *Fix a small $\epsilon \in (0,1)$. For$\lambda = \lambda^* := m/2$, with probability at least $1-\epsilon$, one has*

$$\|\tilde{\rho}_{\lambda^*}^{prob} - \rho^0\|_F^2 \leq C \frac{3^n \mathrm{rank}(\rho^0) \log\left(\frac{\mathrm{rank}(\rho^0) N}{2^n}\right) + (1.5)^n \log(2/\epsilon)}{N},$$

where C is a numerical constant.

A (much worse) rate is also derived for $\tilde{\rho}_\lambda^{dens}$. Chapter 3 also contains numerical simulations: we implement $\tilde{\rho}_\lambda^{prob}$ and $\tilde{\rho}_\lambda^{dens}$ thanks to an MCMC algorithm. In accordance to the theory, $\tilde{\rho}_\lambda^{prob}$ performs usually better. On the other hand, $\tilde{\rho}_\lambda^{dens}$ is much, much easier to compute: the MCMC converges faster and is more stable.

1.5.3 Bibliographical notes

Many authors authors considered the setting with not all possible Pauli measurements $\{\sigma_b = \sigma_{b_1} \otimes \ldots \otimes \sigma_{b_n}, b \in \{I, x, y, z\}^n\}, \sigma_I = I$ but only a random subset of these measurements. These studies are inspired by compressed sensing problems. See for example [Gross, 2011; Gross et al., 2010; Flammia et al., 2012; Koltchinskii, 2011; Koltchinskii and Xia, 2015; Xia and Koltchinskii, 2016; Xia, 2017].

1.6 Lifelong learning in a full online setting

1.6.1 Motivation and formalization

Most analyses of learning algorithms assume that the algorithm starts learning from scratch when presented with a new dataset. However, in real life, it is often the case that we will learn the same features on many different tasks, and that information should be transferred from one task to another. For example, a key problem in pattern recognition is to learn a dictionary of features helpful for image classification: it makes perfectly sense to assume that features learnt to classify dogs against other animals can be re-used to recognize cats. This idea is at the core of *transfer learning*, see for example [Thrun and Pratt, 1998; Baxter, 1997, 2000; Cavallanti et al., 2010; Maurer, 2005; Maurer et al., 2013; Pentina and Lampert, 2014; Balcan et al., 2015; Galanti et al., 2016; Maurer et al., 2016] and references therein.

On the difference to the previous chapters, we study here a problem that was not completely formalized before. The first task was to propose a formal framework to the analysis. We chose to tackle a fully online setting first: more precisely, data sets (tasks) are revealed sequentially and processed by a "within-task algorithm". We propose a lifelong learning scheme which transfers the gained information from one task to the next one. The strong point of our analysis is that we are able to prove regret bounds for our lifelong learning scheme without assuming a specific form for the within-task algorithm. Let us first describe a typical example to fix ideas.

An example Typically, assume that for each task $t \in \{1, \dots, T\}$, the dataset

$$\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big) \in (\mathcal{X} \times \mathcal{Y})^{m_t}, m_t \in \mathbb{N}$$

is revealed sequentially. We propose to use as predictors

$$\hat{y}_{t,i} = \langle \theta_t, D x_{t,i} \rangle$$

where θ_t is learnt for each task t by any within-task algorithm. Our lifelong learning scheme aims at improving the common dictionary D at each task.

Problem setting Let us now describe the setting we proposed. At each time step $t \in \{1, \dots, T\}$, the learner is faced with a task, corresponding to a dataset

$$\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big) \in (\mathcal{X} \times \mathcal{Y})^{m_t}$$

where $m_t \in \mathbb{N}$ and $\mathcal{X}$ and $\mathcal{Y}$ are some sets. The dataset $\mathcal{S}_t$ is itself displayed sequentially, that is, at each inner step $i \in \{1, \dots, m_t\}$:

- The object $x_{t,i}$ is revealed,

- The learner has to predict $y_{t,i}$: a predictor is a function $f : \mathcal{X} \to \mathcal{Y}$. Let $\hat{y}_{t,i} := f_{t,i}(x_{t,i})$ denote the prediction,
- Then $y_{t,i}$ is revealed and the learner incurs the loss. The loss of a predictor f on a pair (x, y) is a real number denoted by $\ell(f(x), y)$. Let $\hat{\ell}_{t,i}$ denote the loss $\ell(\hat{y}_{t,i}, y_{t,i})$.

The task t ends at time m_t, at which point the average prediction error is $\frac{1}{m_t}\sum_{i=1}^{m_t}\hat{\ell}_{t,i}$. This process is repeated for each task t, and thus at the end of all the tasks, the overall average error is

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\hat{\ell}_{t,i}.$$

Importantly, we want to transfer the information (a common data representation) gained from the previous tasks to a new one. Formally, we let $\mathcal{Z}$ be a set and prescribe a set $\mathcal{G}$ of feature maps (also called *representations*) $g : \mathcal{X} \to \mathcal{Z}$, and a set $\mathcal{H}$ of functions $h : \mathcal{Z} \to \mathbb{R}$. We shall design an algorithm that is useful when there is a function $g \in \mathcal{G}$, common to all the tasks, and task-specific functions $h_1, \ldots, h_T$ such that

$$f_t = h_t \circ g$$

is a good predictor for task t, in the sense that the corresponding prediction error is small.

Note that an oracle who would have known the best common representation g for all tasks in advance would have only suffered, on the entire sequence of datasets, the error

$$\inf_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(h_t\circ g(x_{t,i}), y_{t,i}\big).$$

Objective We wish to design a procedure (meta-algorithm) that, at the beginning of each task t, produces a feature map $\hat{g}_t$. Moreover, within each task, the learner can use its own favourite online learning algorithm to learn the task t on the sequence $\big\{(\hat{g}_t(x_{t,1}), y_{t,1}), \ldots, (\hat{g}_t(x_{t,m_t}), y_{t,m_t})\big\}$. Importantly, we wish to control the *compound regret* of our procedure

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\hat{\ell}_{t,i} - \inf_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(h_t\circ g(x_{t,i}), y_{t,i}\big).$$

A meta algorithm We propose the following meta algorithm which is based on the exponentially weighted aggregation, denoted by EWA, (see e.g. [Cesa-Bianchi and Lugosi, 2006]) procedure for learning the representation g. However, our procedure allows the user to freely select her own within-task algorithm, which does not have to be the same for each task.

Algorithm 2 EWA-LL

Data A sequence of datasets $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big)$, $1 \leq t \leq T$, associated with different learning tasks; the points within each dataset are also given sequentially.

Input A prior π_1 on $\mathcal{G}$, a learning parameter $\eta > 0$ and a learning algorithm for each task t which, for any representation g returns a sequence of predictions $\hat{y}^g_{t,i}$ and suffers a loss

$$\hat{L}_t(g) := \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\left(\hat{y}^g_{t,i}, y_{t,i}\right).$$

Loop For $t = 1, \ldots, T$

i Draw $\hat{g}_t \sim \pi_t$.

ii Run the within-task learning algorithm on $\mathcal{S}_t$ and suffer loss $\hat{L}_t(\hat{g}_t)$.

iii Update

$$\pi_{t+1}(\mathrm{d}g) := \frac{\exp(-\eta \hat{L}_t(g))\pi_t(\mathrm{d}g)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

Assumption (FR) (finite-regret) There, a key assumption is that there is a control on the regret of the within task algorithm:

$$\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) \leq \beta(g, m_t) < \infty.$$

Note that this assumption is satisfied by many popular algorithms for online learning: EWA, online gradient... see [Cesa-Bianchi and Lugosi, 2006; Shalev-Shwartz, 2012] for an overview. Detailed examples are provided in Chapter 4.

In Chapter 4, we prove the following result on the performances of EWA-LL.

Theorem 1.9. *Under the assumption (**FR**) and assuming that for any $g \in \mathcal{G}$, $\hat{L}_t(g) \in [0, C]$, we have*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t} \left[\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}\right] \leq \inf_{\rho} \Bigg\{ \mathbb{E}_{g \sim \rho} \Bigg[\frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) + \frac{1}{T} \sum_{t=1}^{T} \beta(g, m_t)\Bigg] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T} \Bigg\},$$

where the infimum is taken over all probability measures ρ and $\mathcal{K}(\rho, \pi_1)$ is the Kullback-Leibler divergence between ρ and π_1.

Note that the theorem above yields a bound on the expected regret and instead of an infimum with respect to g, we have an infimum on all the

possible aggregation w.r.t g. However, it is possible to derive uniform bounds (instead of in expectation) from the above theorem when the loss is convex, this is done in Section 4.3.3. Also, a bound with an infimum with respect to g can be obtained, this is studied in some applications in the chapter 4 (see Sections 4.5). The examples of finite sets $\mathcal{G}$ and $\mathcal{H}$, and the example of dictionary learning are covered in detail.

Note that the optimal rates in this setting are not known - and specifications of $\mathcal{G}$ and $\mathcal{H}$ would be necessary to formally state the problem. Still, for simplicity in the case where $m_t = m$ for all t, we exhibit some situations where the learning rate is in $1/\sqrt{T} + 1/m$ while the rates in one of the only previous theoretical study of transfer learning [Pentina and Lampert, 2014] was $1/\sqrt{T} + 1/\sqrt{m}$.

Chapter 2

MATRIX COMPLETION

Bayesian methods for low-rank matrix completion with noise have been shown to be very efficient computationally [Alquier et al., 2014; Lawrence and Urtasun, 2009; Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Zhou et al., 2010]. While the behaviour of penalized minimization methods is well understood both from the theoretical and computational points of view (see [Candès and Plan, 2010; Candès and Tao, 2010; Koltchinskii et al., 2011; Recht and Ré, 2013] among others) in this problem, the theoretical optimality of Bayesian estimators have not been explored yet. In this work, we propose a Bayesian-like estimator for matrix completion under general sampling distribution. We also provide an oracle inequality for this estimator. This inequality proves that, whatever the rank of the matrix to be estimated, our estimator reaches the minimax-optimal rate of convergence (up to a logarithmic factor). We end this chapter with a short simulation study.

The works in this chapter have been published in [Mai and Alquier, 2015]:

T.T. Mai & P. Alquier. A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. ***Electronic Journal of Statistics***, *vol.**9**: 823–841,* ***2015***.

2.1 Introduction and notations

The "Netflix Prize" [Bennett and Lanning, 2007] generated a significant interest in the *matrix completion* problem. The Netflix data can be represented as a sparse matrix made up of ratings given by users (rows) to movies (columns). To infer the missing entries is thus very helpful to propose sensible advertisement and improve the sales. However, it is totally impossible to recover an uncomplete matrix without any assumption. A suitable condition, popular in practice for this problem, is that the matrix has low-rank or approximately low-rank [Alquier, 2013; Alquier et al., 2014; Candès and Plan, 2010; Candès and Recht, 2009; Candès and Tao, 2010; Klopp, 2014; Koltchinskii et al.,

2011]. For the Netflix problem, this assumption is sensible as it means that many movies (or users) have similar profiles.

Let $M^0 \in \mathbb{R}^{m\times p}$ be an unknown matrix (expected to be low-rank). Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d random variables drawn from a joint distribution $\mathbf{P}$. We assume that

$$Y_i = M^0_{X_i} + \mathcal{E}_i, \quad i = 1, \dots, n, \tag{2.1}$$

the noise variables $\mathcal{E}_i$ are independent from X_i and $\mathbb{E}(\mathcal{E}_i) = 0$. We let Π denote the marginal distribution of X when $(X, Y) \sim \mathbf{P}$. Remark that Π is a distribution on the set $\mathfrak{X} = \{1, \dots, m\} \times \{1, \dots, p\}$. Then, the problem of estimating M^0 with $n < mp$ is called the noisy matrix completion problem under general sampling distribution.

A special instance of this problem is that the sampling distribution Π is uniform, this assumption is done for example in [Candès and Plan, 2010; Candès and Recht, 2009; Candès and Tao, 2010; Koltchinskii et al., 2011; Alquier et al., 2014]. Clearly, in practice, the observed entries are not always uniformly distributed: for example, some movies are more famous than others, and thus receive much more ratings. More importantly, the sampling distribution is not known in practice. More general sampling schemes than uniform distribution had been already studied, see e.g. [Foygel et al., 2011; Klopp, 2014; Negahban and Wainwright, 2012], but there are still some assumptions on Π in these papers. Here, we do not impose any restriction on Π. From now, $\Pi_{ij} = \mathbb{P}(X = \{i, j\})$ will denote the probability to observe the (i, j)-th entry.

For any matrix $A_{m\times p}$, let $\|A\|_F$ denote the Frobenius norm, i.e, $\|A\|_F^2 = \mathrm{Tr}(A^T A)$. We define a "generalized Frobenius norm" as follows

$$\|A\|_{F,\Pi}^2 = \sum_{ij} (A_{ij})^2 \Pi_{ij}.$$

Note that when the sampling distribution Π is uniform, then

$$\|A\|_{F,\Pi}^2 = \frac{1}{mp}\|A\|_F^2.$$

For any matrix $M_{m\times p} \in \mathbb{R}^{mp}$, we define the empirical risk as

$$r(M) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - M_{X_i})^2$$

and the prediction risk

$$R(M) = \mathbb{E}_{(X,Y)\sim\mathbf{P}}\left[(Y - M_X)^2\right].$$

In this paper, the prediction problem is considered, i.e, the objective is to define an estimator $\widehat{M}$ such that $R(\widehat{M}) - R(M^0)$ is as small as possible. Remark that $R(M) - R(M^0) = \|M - M^0\|_{F,\Pi}^2$ for any M (using Pythagorean Theorem).

2.1.1 Penalized minimization approaches

When handing with this problem, most of the recent methods are often based on minimizing a criterion of the fit to the observations, such as $r(M)$, with additional penalty term that encouraging low-rank. More specifically,

$$\hat{M} = arg\min_{M}\{r(M) + \lambda pen(M)\}$$

where λ is a regularization parameter and $pen(M)$ could be the rank of the matrix M (non-convex) or the nuclear-norm of M (convex): $\|M\|_* = \sum_{i=1}^{\min(m,p)} \gamma_i(M), \gamma_i(M)'s$ are the singular values of M.

A first result can be found in by Candès and Recht [Candès and Recht, 2009], Candès and Tao [Candès and Tao, 2010] for exact matrix completion (noiseless case, i.e. $\mathcal{E}_i = 0$). These results were then developed in the noisy case [Candès and Plan, 2010; Koltchinskii et al., 2011]. Some efficient algorithms had also been proposed, for example see [Recht and Ré, 2013].

Recently, some authors have studied a more general problem, the so-called *Trace regression* problem, e.g see [Klopp, 2014; Koltchinskii et al., 2011]. This problem includes matrix completion, together with other well-known problems (linear regression, reduced rank regression and multitask learning) as special cases. For matrix completion, one observes n pairs (X_i, Y_i) satisfying

$$Y_i = \text{Trace}(X_i^T M^0) + \varepsilon_i, i = 1, \ldots, n$$

where (ε_i) is a noise vector. The random matrices $X_i \in \mathbb{R}^{m\times p}$ are independent of the $\varepsilon_i's$, and are chosen uniformly at random from the set

$$\mathcal{B} = \{e_j(m)e_k^T(p), 1 \leq j \leq m, 1 \leq k \leq p\},$$

where the $e_j(s)$ are the canonical basis vectors of $\mathbb{R}^s$. They proposed nuclear-norm penalized estimators and provided reconstruction errors for their methods. They also proved that these errors are minimax-optimal (up to a logarithmic factor).

Note that the average quadratic error on the entries of a rank-r matrix size $m\times p$ from n- observations can not be better than: $r\max(m,p)/n$ [Koltchinskii et al., 2011].

2.1.2 Bayesian methods

A few authors considered Bayesian methods for matrix completion [Alquier et al., 2014; Lawrence and Urtasun, 2009; Lim and Teh, 2007; Salakhutdinov and Mnih, 2008; Zhou et al., 2010]. A summary of the main ideas in these papers can be found in the survey [Alquier et al., 2014]. Basically, it is to define the prior in order to mimic a singular value decomposition (SVD) on a matrix M. We write:

$$M = \sum_{i=1}^{\min(m,p)} U_i V_i^T$$

where the U_i and the V_i are column vectors in $\mathbb{R}^m$ and $\mathbb{R}^p$ respectively. Their prior distribution is given by

$$U_i \sim \mathcal{N}(0, \gamma_i Id_m) \quad \text{and} \quad V_i \sim \mathcal{N}(0, \gamma_i Id_p),$$

where Id_s stands for the identity matrix of dimension s. In order to ensure that most terms $U_i V_i^T$ are almost equal to zero (which means that M is "almost low-rank"), we model the $\gamma_i's$ as random according to a distribution highly concentrated around zero. A popular choice in the literature is that $1/\gamma_i \sim \Gamma(a, b)$ with a very small b. These distributions (Gaussian and inverse gamma) are conjugate, so it is possible to sample from the posterior using the Gibbs sampler as in [Salakhutdinov and Mnih, 2008]. However, there are at the time no theoretical guarantees regarding the consistency nor the minimax-optimality of this estimator.

In this paper, we design a new prior and prove an minimax-optimal oracle bound for the corresponding Bayesian estimator. This is presented in Section 2.2. In Section 2.3, we discuss the implementation of our Bayesian estimator. Some experiments comparing our estimator to the one based on conjugate priors are done on simulated datasets. The proof of the main result is provided in the Section 2.5.

2.2 Main results

Before we introduce our estimator, let us formulate some assumptions.

Assumption 2.1. *There is a known constant L such that*

$$\|M^0\|_\infty = \sup_{i,j} |M^0_{ij}| \leq L < +\infty.$$

This is a mild assumption. In the Netflix and MovieLens datasets, the ratings belong to the set $\{1, 2, 3, 4, 5\}$, so we can take $L = 5$.

2.2.1 The prior distribution and the estimator

We describe hereafter a prior π on matrices $M_{m\times p}$ as follows. Let $K = \min(m, p)$ and Γ be a random variables taking value in the set $\{\Gamma_1, \ldots, \Gamma_K\}$ with

$$\mathbb{P}(\Gamma = \Gamma_k) = \tau^{k-1}\left(\frac{1-\tau}{1-\tau^K}\right)$$

where $\Gamma_k = (\overbrace{1, \ldots, 1}^{k \text{ times}}, \overbrace{0, \ldots, 0}^{K-k \text{ times}})$ for some constant $\tau \in (0, 1)$ and $k \in \{1, \ldots, K\}$.

Now, assuming that $\Gamma = \Gamma_k$ and a matrix $M_{m\times p}$ is drawn as $M = U_{m\times K}(V_{p\times K})^T$ where

$$U_{i,\ell}; V_{j,\ell} \overset{\text{i.i.d}}{\sim} \begin{cases} \mathcal{U}([-\delta, \delta]) & \text{when } \Gamma_{k,\ell} = 1, \\ \mathcal{U}([-\kappa, \kappa]) & \text{when } \Gamma_{k,\ell} = 0, \end{cases} \quad \ell = 1, \ldots, K$$

with $\delta = \sqrt{2L/K}$ and $0 \leq \kappa \leq (1/n)\sqrt{L/(10K)}$. Note that, in this case, the entries of M satisfy: $\sup_{i,j} |M_{ij}| \leq 2L$. Moreover, when a matrix M is drawn from this prior, as κ is small, most columns of U and V are almost null. So the matrix $M = UV^T$ is very close to a rank-k matrix. Actually, the choice $\kappa = 0$ leads to $\text{rank}(M) \leq k$.

We are now ready to define our estimator. For any $\lambda > 0$, we consider the conditional probability measure $\hat{\rho}_\lambda$ given by its density w.r.t. the probability measure π:

$$\frac{d\hat{\rho}_\lambda}{d\pi}(M) = \frac{e^{-\lambda r(M)}}{\int e^{-\lambda r} d\pi}. \tag{2.2}$$

The aggregate $\widehat{M}_\lambda$ is defined as follows

$$\widehat{M}_\lambda = \int M \hat{\rho}_\lambda(dM). \tag{2.3}$$

Note that, for $\lambda = n/(2\sigma^2)$, this corresponds exactly to the Bayesian estimator that would be obtained for a Gaussian noise $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$. However, a slightly different choice for λ, denoted by λ^* below, will allow to obtain the optimality of the estimator under a wider class of noises.

2.2.2 A minimax-optimal oracle inequality under general sampling distribution

Assumption 2.2. *The noise variables $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent and independent of $X_1, \ldots, X_n$. There exist two known constants $\sigma > 0$ and $\xi > 0$ such that*

$$\mathbb{E}(\mathcal{E}_i^2) \leq \sigma^2$$
$$\forall k \geq 3, \quad \mathbb{E}(|\mathcal{E}_i|^k) \leq \sigma^2 k! \xi^{k-2}.$$

Assumption 2.2 states that the noise is sub-exponential, it includes the cases where the noise is bounded or sub-Gaussian (and of course Gaussian), see e.g. Chapter 2 in [Boucheron et al., 2013].

For any $x > 0$, we define (remind that $K = \min(m, p)$)

$$\mathcal{M}(x) = \left\{ M = UV^T, \text{ with } |U_{i\ell}| \leq \sqrt{\frac{x}{K}}, |V_{j\ell}| \leq \sqrt{\frac{x}{K}} \right\},$$

and

$$\mathcal{C} = [12L(2\xi + 3L)] \vee \left[8\sigma^2 + 2(3L)^2\right].$$

Hereafter, the main result is presented. We provide an oracle bound for our estimator $\widehat{M}_{\lambda^*}$.

Theorem 2.1. *Let Assumption 2.1 and 2.2 be satisfied and take $\lambda = \lambda^* := \frac{n}{2\mathcal{C}}$. Then, for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$ and as soon as $n \geq \max(m, p)$, one has*

$$\|\widehat{M}_{\lambda^*} - M^0\|_{F,\Pi}^2 \leq \inf_{M \in \mathcal{M}(L)} \left\{ 3\|M - M^0\|_{F,\Pi}^2 + \mathscr{C}_{L,\xi,\sigma,\tau} \frac{(m+p)\text{rank}(M)\log(K)}{n} \right.$$

$$+\frac{8\mathcal{C}\log\left(\frac{2}{\varepsilon}\right)}{n}\Bigg\},$$

where $\mathscr{C}_{L,\xi,\sigma,\tau}$ is a (known) numerical constant depending on L, ξ, σ and τ only.

The proof of this theorem is given in the Section 2.5. It follows an argument called "PAC-Bayesian inequality". PAC-Bayesian inequalities were introduced in [McAllester, 1998; Shawe-Taylor and Williamson, 1997] in order to provide empirical bounds on the prevision risk of Bayesian-type estimators. However, our proof is closer to Catoni's works [Catoni, 2003, 2004, 2007], where it is shown how to derive powerful oracle inequalities from PAC-Bayesian bounds. This approach has been used many times since then to prove oracle inequalities in many dimension-reduction problems like sparse regression estimation[Dalalyan and Tsybakov, 2008] ,[Alquier and Lounici, 2011; Alquier and Biau, 2013] or reduced-rank regression [Alquier, 2013].

The choice $\lambda = \lambda^*$ comes from the proof of this theorem when optimizing an upper bound on the risk R, see (2.15) page 67. However, in practice, this choice may not be the best one. For example, in the experiments done in Section 3 with Gaussian noise $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$, we take $\lambda = \frac{n}{4\sigma^2}$ that was shown in [Dalalyan and Tsybakov, 2008] to behave very well in regression problems. Also, in practice, to take K smaller than $\min(m, p)$ improves significantly the speed of the algorithm with little consequence on the performance of the estimator [Alquier et al., 2014].

Remark 2.1. *When $M^0 \in \mathcal{M}(L)$, we can take $M = M^0$, one gets*

$$\|\widehat{M}_{\lambda^*} - M^0\|_{F,\Pi}^2 \leq \mathscr{C}_{L,\xi,\sigma,\tau}\frac{(m+p)\mathrm{rank}(M^0)\log(K)}{n} + \frac{8\mathcal{C}\log\left(\frac{2}{\varepsilon}\right)}{n}.$$

The rate $(m+p)\mathrm{rank}(M^0)\log(K)/n$ is minimax-optimal, or at least almost minimax-optimal: a lower bound in this problem is provided by Theorems 5 and 7 in [Koltchinskii et al., 2011], it is $(m+p)\mathrm{rank}(M^0)/n$. The optimality of the log *term is, to our knowledge, an open question. Note however that the upper bound in [Koltchinskii et al., 2011] is $(m+p)\mathrm{rank}(M^0)\log(m+p)/n$. So, our bound represents a slight improvement in the case $\min(m,p) \ll \max(m,p)$.*

Remark 2.2. *When the sampling distribution Π is uniform in Theorem 2.1, we obtain the following oracle bound for the Frobenius norm*

$$\frac{1}{mp}\|\widehat{M}_{\lambda^*} - M^0\|_F^2 \leq \inf_{M\in\mathcal{M}(L)}\Bigg\{\frac{3}{mp}\|M - M^0\|_F^2 + \mathscr{C}'_{L,\xi,\sigma,\tau}\frac{(m+p)\mathrm{rank}(M)\log(K)}{n} + \frac{8\mathcal{C}\log\left(\frac{2}{\varepsilon}\right)}{n}\Bigg\}.$$

Finally, we want to mention that the rate of [Koltchinskii et al., 2011] is also reached, in a work parallel to ours, by Suzuki [Suzuki, 2015], in a Bayesian

framework. The main difference is that, while [Suzuki, 2015] provides a rate of convergence in a more general low-rank tensor estimation problem, his works do not bring an oracle inequality like Theorem 2.1 that can be used when M^0 is not exactly low-rank, but can be well approximated by a low-rank matrix. Moreover, our result holds under any sampling distribution Π.

2.3 Experiments and comparison with conjugate priors

2.3.1 A Gibbs algorithm for $\widehat{M}_\lambda$

As it has been shown in Section 2, our estimator $\widehat{M}_{\lambda^*}$ satisfies a powerful oracle inequality. However, as mentioned in the introduction, the Bayesian estimator using conjugate priors is popular in practice as it leads to a fast algorithm. The reason is that there is an explicit form for the conditional posterior distribution of the i-th row of U, $U_{i,\cdot}$, given the other rowss of U, $U_{-i,\cdot}$, and given V (it is a multivariate normal distribution which parameters are known). This allows to use a Gibbs sampler, with very good convergence properties. This is described for example in [Alquier et al., 2014] and the references therein.

Here, straightforward but tedious computations lead to

$$\hat{\rho}_\lambda(U_{i,\cdot}|k, U_{-i,\ell}, V, \Gamma = \Gamma_k) \propto \varphi\left(U_{i,\cdot}\,;\, \frac{2\lambda}{n}\Sigma_i \sum_{k:I_k=i} Y_k V_{J_k,\cdot}\,,\, \Sigma_i\right) \prod_{\ell=1}^{k} \mathbf{1}_{\{|U_{i,\ell}|\leq\delta\}} \prod_{\ell=k+1}^{K} \mathbf{1}_{\{|U_{i,\ell}|\leq\kappa\}}$$

where I_s and J_s are the components of the $X_i, i = 1, \ldots, n$ (e.g $X_1 = (I_1, J_1)$, $\ldots$, $X_n = (I_n, J_n)$);

$$(\Sigma_i)^{-1} = \frac{2\lambda}{n} \sum_{k:I_k=i} V_{J_k,\cdot}^T V_{J_k,\cdot}$$

and $\varphi(\cdot\,; m, V)$ is the density of the multivariate normal distribution with mean vector m and variance-covariance matrix V. So, the conditional posterior distribution of $U_{i,\cdot}$ is a truncated multivariate normal.

To sample from such a disitrubition is known as a very hard problem in general, see for example [Kotecha and Djuric, 1999]. However, using the R package **tmvtnorm** [Wilhelm and Manjunath, 2010], it is possible to sample from a truncated multivariate normal fast enough to compute our estimator on reasonnably large datasets. Finally, instead of including the hyperparameter $k \in \{1, \ldots, K\}$ in the simulations, we simulated K chains simultaneously, one for every $k \in \{1, \ldots, K\}$, and selected the realization of one of the chains at each round using the probabilities given by (2.2).

Also, note that the truncation procedure proposed by Suzuki in [Suzuki, 2015] cannot be implemented, to our understanding, using this procedure, as

the truncation is done directly on the product UV^T rather than on U and V individually.

2.3.2 Experiments and Results

We use the notation $\widehat{M_\lambda}$ for our estimator, let us denote $\hat{M}^{\text{conjugate}}$ the estimator based on the Gaussian prior for U and V with inverse Gamma variance, described in [Alquier et al., 2014] and in the aforementionned references. In order to compare both estimators, a series of experiments were done with simulated data:

- In the first series of simulations, the data are simulated as in [Candès and Plan, 2010; Alquier et al., 2014]. More precisely, a rank-2 matrix $M^0_{m\times m}$ (so $m = p$) has been created as the product of two rank-2 matrices,
$$M^0 = U^0_{m\times 2}(V^0_{m\times 2})^T$$
where the entries of U^0 and V^0 are i.i.d $\mathcal{N}(0, 20/\sqrt{m})$. Only 20% entries of the matrix M^0 are observed (using a uniform sampling). This sampled set is then corrupted by noise as in (2.1), where the $\mathcal{E}_i$ are i.i.d $\mathcal{N}(0,1)$. We consider the cases $m = 100$, $m = 200$, $m = 500$ and $m = 1000$.

- The second series of simulations is similar to the first one, except that the matrix M^0 is no longer rank 2, but it can be well approximated by a rank 2 matrix:
$$M^0 = U^0_{m\times 2}(V^0_{m\times 2})^T + \frac{1}{100}(Z^0_{m\times 50})(W^0_{m\times 50})^T$$
where the entries of Z^0 and W^0 are i.i.d $\mathcal{N}(0, 20/\sqrt{m})$.

- The third series of experiments is similar to the first one, but the noise variables $\mathcal{E}_i$ are now i.i.d from a uniform distribution on $[-1, 1]$. Note that, from a purely Bayesian point of view, this corresponds to a misspecified model. However, the bound in Theorem 2.1 is still valid in this case.

- Finally, the fourth series of experiments is similar to the first one, noise variables $\mathcal{E}_i$ are now i.i.d from a heavy-tailed distribution (Student, with parameter 5). This is another misspecified model, but in this case, Theorem 2.1 cannot be used.

The behavior of our estimator $\widehat{M_\lambda}$ is computed through the root-mean-squared error (RMSE) per entry,

$$\text{RMSE} = \sqrt{\frac{1}{mp}\|\widehat{M_\lambda} - M^0\|_F^2} = \frac{1}{m}\|\widehat{M_\lambda} - M^0\|_F.$$

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.535 ($\pm$0.003)	0.348 ($\pm$0.003)	0.207 ($\pm$0.0001)	0.141 ($\pm$0.0006)
Gaussian	0.538 ($\pm$0.001)	0.345 ($\pm$0.001)	0.210 ($\pm$0.0001)	0.146 ($\pm$0.001)

Table 2.1: *RMSEs in the first series of experiments (low-rank matrix, Gaussian noise)*

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.640 ($\pm$0.008)	0.387 ($\pm$0.001)	0.214 ($\pm$0.0008)	0.145 ($\pm$0.0002)
Gaussian	0.620 ($\pm$0.003)	0.385 ($\pm$0.001)	0.216 ($\pm$0.0003)	0.145 ($\pm$0.001)

Table 2.2: *RMSEs in the second series of experiments (approx. low-rank, Gaussian noise)*

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.328 ($\pm$0.002)	0.205 ($\pm$0.001)	0.120 ($\pm$0.001)	0.084 ($\pm$0.002)
Gaussian	0.334 ($\pm$0.003)	0.208 ($\pm$0.001)	0.126 ($\pm$0.003)	0.086 ($\pm$0.001)

Table 2.3: *RMSEs in the third series of experiments (low-rank matrix, uniform noise)*

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.745 ($\pm$0.039)	0.567 ($\pm$0.005)	0.340 ($\pm$0.004)	0.237 ($\pm$0.003)
Gaussian	0.659 ($\pm$0.003)	0.439 ($\pm$0.001)	0.268 ($\pm$0.002)	0.186 ($\pm$0.002)

Table 2.4: *RMSEs in the fourth series of experiments (low-rank matrix, heavy-tailed noise)*

The parameters are given as follows: for both $\widehat{M}_\lambda$ and $\hat{M}^{\text{conjugate}}$, the parameter λ is set to $n/4$, following [Dalalyan and Tsybakov, 2008]. Following [Alquier et al., 2014] we use for the parameters of the inverse Gamma prior in $\hat{M}^{\text{conjugate}}$ the values $a = 1$, $b = 1/100$. Finally, for $\widehat{M}_\lambda$, we used $\kappa = 0$, $K = 5$, $L = 50$ and $\tau = 1/2$ on all the simulations apart from the heavy-tailed noise case, where we used $\tau = 1/4$. Note that a proper optimization with respect to the parameters τ and λ could lead to better results, for example through cross-validation.

The first conclusion is that the results of both methods are very close. In many situations, however, the variance of the estimator with uniform prior is larger than the variance of the estimator with Gaussian prior. The evidence is that this is due to the fact that the MCMC algorithm used to compute the estimator with Gaussian prior, $\hat{M}^{\text{conjugate}}$, converges faster than the algorithm used to compute the estimator with uniform prior, $\widehat{M}_\lambda$. This is supported by Figure 2.1 page 60. However, it seems that this difference is less and less significant when the dimension m grows.

According to our main oracle inequality, our estimator is robust to misspecification in the low-rank assumption, see Table 2.2, and in the noise, at least in the sub-Gaussian case, see Table 2.3. More importantly: despite the fact that the theoretical properties of $\hat{M}^{\text{conjugate}}$ are not known, this estimator is more robust than ours to heavy-tailed noise, as shown in Table 2.4.

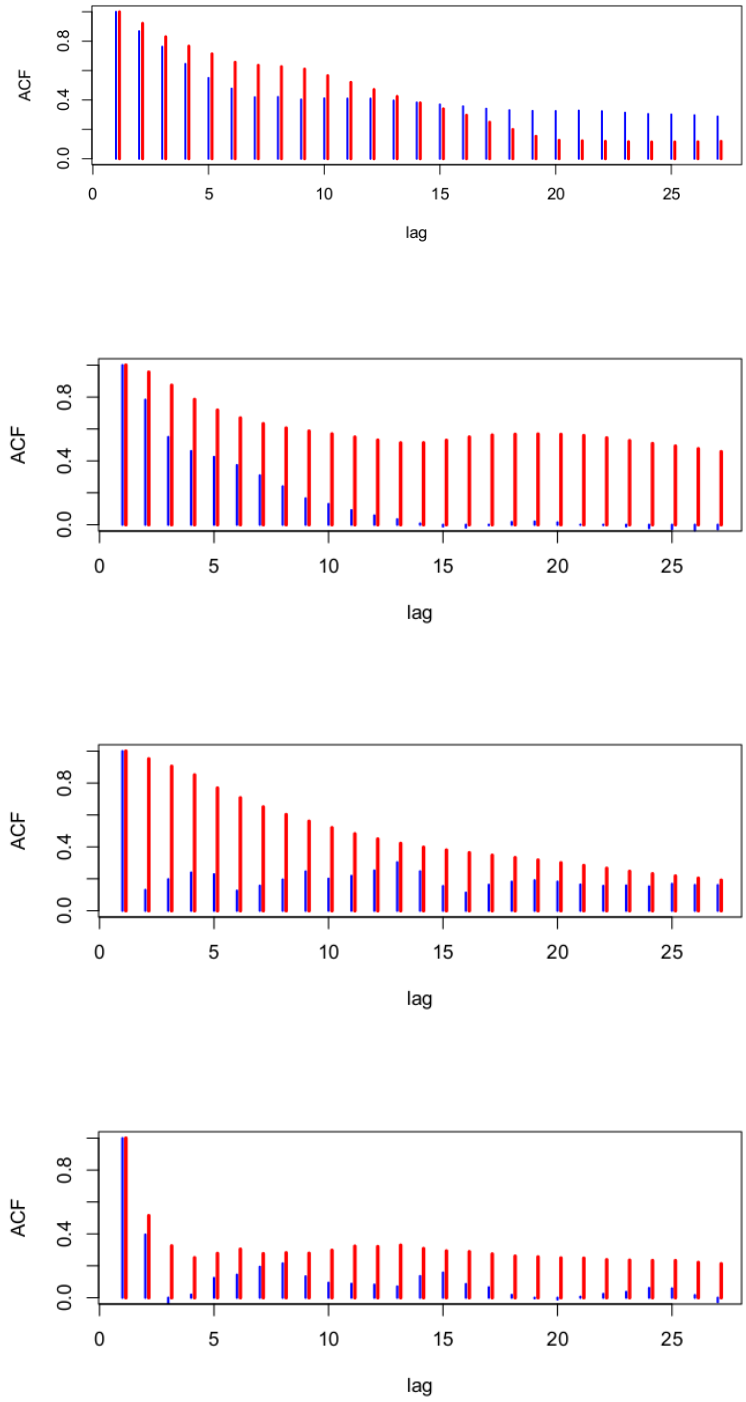

Figure 2.1: *ACF of four randomly selected entries during a simulation. These are taken from the first series of experiments. The ACF of the Gibbs sampler for the Bayesian estimator with uniform priors, $\widehat{M_\lambda}$, is in red while the ACF of the Gibbs sampler for the Bayesian estimator with Gaussian priors, $\hat{M}^{\text{conjugate}}$, is in blue.*

2.4 Discussion

This chapter proposes a Bayesian estimator for the noisy matrix completion problem under general sampling distribution. This estimator satisfies an optimal oracle inequality under any sampling scheme. Based on simulations, it is also clear that this estimator performs well in practice, however, a faster algorithm for very large datasets is still an open issue. Another important open question is the minimax-optimality of the estimator based on Gaussian priors.

2.5 Proofs

First, we state a version of Bernstein's inequality useful in the proof of Theorem 2.1. This version is taken from [Massart, 2007] (Inequality 2.21 in the proof of Proposition 2.9 page 24).

Lemma 2.1. *Let $T_1, \ldots, T_n$ be independent real valued random variables. Let us assume that there are two constants v and w such that*

$$\sum_{i=1}^{n} \mathbb{E}[T_i^2] \leq v$$

and for all integers $k \geq 3$,

$$\sum_{i=1}^{n} \mathbb{E}\left[(T_i)^k\right] \leq v \frac{k! w^{k-2}}{2}.$$

Then, for any $\zeta \in (0, 1/w)$,

$$\mathbb{E} \exp\left[\zeta \sum_{i=1}^{n} [T_i - \mathbb{E}(T_i)]\right] \leq \exp\left(\frac{v\zeta^2}{2(1-w\zeta)}\right).$$

Now, we are ready to present the proof of Theorem 1.

Proof of Theorem 2.1: the proof is divided in two steps. In the first step, we establish a general PAC-Bayesian inequality for matrix completion, in the style of [Catoni, 2004; Dalalyan and Tsybakov, 2008]. In the second step, we derive the oracle inequality from the first step.

Step 1:

Let's define, for any matrix $M \in \mathcal{M}(2L)$, the following random variables

$$T_i = \left(Y_i - M^0_{X_i}\right)^2 - \left(Y_i - M_{X_i}\right)^2.$$

Note that these variables are independent. We first check that the variables T_i satisfy the assumptions of Lemma 2.1, in order to apply this lemma. We have

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}[T_i^2] &= \sum_{i=1}^{n} \mathbb{E}\left[\left(2Y_i - M^0_{X_i} - M_{X_i}\right)^2 \left(M^0_{X_i} - M_{X_i}\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\left(2\mathcal{E}_i + M^0_{X_i} - M_{X_i}\right)^2 \left(M^0_{X_i} - M_{X_i}\right)^2\right] \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[\left[8\mathcal{E}_i^2 + 2(L+2L)^2\right] \left[M^0_{X_i} - M_{X_i}\right]^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[8\mathcal{E}_i^2 + 2(3L)^2\right] \mathbb{E}\left[M^0_{X_i} - M_{X_i}\right]^2 \\
&\leq n\left[8\sigma^2 + 2(3L)^2\right]\left[R(M) - R(M^0)\right] =: v(M, M^0) = v.
\end{aligned}$$

Next we have, for any integer $k \geq 3$, that

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}\left[(T_i)^k\right] \leq & \sum_{i=1}^{n} \mathbb{E}\left[\left|2Y_i - M^0_{X_i} - M_{X_i}\right|^k \left|M^0_{X_i} - M_{X_i}\right|^k\right] \\
\leq & \sum_{i=1}^{n} \mathbb{E}\left[2^{2k-1}\left[|\mathcal{E}_i|^k + (L/2+L)^k\right]\left|M^0_{X_i} - M_{X_i}\right|^k\right] \\
\leq & \sum_{i=1}^{n} \mathbb{E}\left[2^{2k-1}\left(|\mathcal{E}_i|^k + (\frac{3}{2}L)^k\right)(3L)^{k-2}\left|M^0_{X_i} - M_{X_i}\right|^2\right] \\
\leq & 2^{2k-1}\left[\sigma^2 k!\xi^{k-2} + \left(\frac{3}{2}L\right)^k\right](3L)^{k-2}\sum_{i=1}^{n}\mathbb{E}\left|M^0_{X_i} - M_{X_i}\right|^2 \\
\leq & \frac{\left[\sigma^2 k!\xi^{k-2} + (\frac{3}{2}L)^k\right][4(3L)]^{k-2}}{\sigma^2 + (\frac{3}{2}L)^2} v \\
\leq & \left[k!\xi^{k-2} + \left(\frac{3}{2}L\right)^{k-2}\right][4(3L)]^{k-2} v \\
\leq & k!\left(\xi + \frac{3}{2}L\right)^{k-2}(12L)^{k-2} v \leq v\frac{k! w^{k-2}}{2},
\end{aligned}$$

with $w := 12L(2\xi + 3L)$.

Next, for any $\lambda \in (0, n/w)$, applying Lemma 2.1 with $\zeta = \lambda/n$ gives

$$\mathbb{E}\exp\left[\lambda\Big(R(M) - R(M^0) - r(M) + r(M^0)\Big)\right] \leq \exp\left[\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})}\right].$$

Set $\mathcal{C}_{\sigma,L} = 2\left[4\sigma^2 + (3L)^2\right]$. For the sake of simplicity let us put

$$\alpha = \left(\lambda - \frac{\lambda^2 \mathcal{C}_{\sigma,L}}{2n(1 - \frac{w\lambda}{n})}\right). \tag{2.4}$$

In order to understand what follows, keep in mind that w is a constant and that our optimal estimator comes with $\lambda = \lambda^* = \frac{n}{2\mathcal{C}}$, so α is of order n.

For any $\varepsilon > 0$, the last display yields

$$\mathbb{E}\exp\left[\alpha\Big(R(M) - R(M^0)\Big) + \lambda\Big(-r(M) + r(M^0)\Big) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2}.$$

Integrating w.r.t. the probability distribution $\pi(.)$, we get

$$\int \mathbb{E}\exp\left[\alpha\Big(R(M) - R(M^0)\Big) + \lambda\Big(-r(M) + r(M^0)\Big) - \log\frac{2}{\varepsilon}\right]\pi(dM) \leq \frac{\varepsilon}{2}.$$

Next, Fubini's theorem gives

$$\begin{aligned}\mathbb{E}\int \exp\left[\alpha\Big(R(M) - R(M^0)\Big) + \lambda\Big(-r(M) + r(M^0)\Big) - \log\frac{2}{\varepsilon}\right]\pi(dM) \\ = \mathbb{E}\int \exp\left\{\alpha\Big(R(M) - R(M^0)\Big) + \lambda\Big(-r(M) + r(M^0)\Big) - \right. \\ \left. - \log\left[\frac{d\hat{\rho}_\lambda}{d\pi}(M)\right] - \log\frac{2}{\varepsilon}\right\}\hat{\rho}_\lambda(dM) \leq \frac{\varepsilon}{2}.\end{aligned}$$

Jensen's inequality yields

$$\mathbb{E}\exp\left[\alpha\left(\int R d\hat{\rho}_\lambda - R(M^0)\right) + \lambda\left(-\int r d\hat{\rho}_\lambda + r(M^0)\right) - \mathcal{K}(\hat{\rho}_\lambda, \pi) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2},$$

where $\mathcal{K}(p, q)$ is the Kullback–Leibler divergence of p from q. Now, using the basic inequality $\exp(x) \geq \mathbf{1}_{\mathbb{R}_+}(x)$, we get

$$\mathbb{P}\left\{\left[\alpha\left(\int R d\hat{\rho}_\lambda - R(M^0)\right) + \lambda\left(-\int r d\hat{\rho}_\lambda + r(M^0)\right) - \mathcal{K}(\hat{\rho}_\lambda, \pi) - \log\frac{2}{\varepsilon}\right] \geq 0\right\} \leq \frac{\varepsilon}{2}.$$

Using Jensen's inequality again gives

$$\int R d\hat{\rho}_\lambda \geq R\left(\int M\hat{\rho}_\lambda(dM)\right) = R(\widehat{M}_\lambda).$$

Combining the last two displays we obtain

$$\mathbb{P}\left\{R(\widehat{M}_\lambda) - R(M^0) \leq \frac{\int r d\hat{\rho}_\lambda - r(M^0) + \frac{1}{\lambda}\left[\mathcal{K}(\hat{\rho}_\lambda, \pi) + \log\frac{2}{\varepsilon}\right]}{\frac{\alpha}{\lambda}}\right\} \geq 1 - \frac{\varepsilon}{2}.$$

Using Donsker and Varadhan's variational inequality ([Catoni, 2007, Lemma 1.1.3]), we get

$$\mathbb{P}\left\{R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{\rho \in \mathfrak{M}_+^1(M)} \frac{\int r d\rho - r(M^0) + \frac{1}{\lambda}\left[\mathcal{K}(\rho, \pi) + \log\frac{2}{\varepsilon}\right]}{\frac{\alpha}{\lambda}}\right\} \geq 1 - \frac{\varepsilon}{2}, \tag{2.5}$$

where $\mathfrak{M}_+^1(M)$ is the set of all positive probability measures over the set of $m \times p$ matrices equiped with the Borel σ-algebra.

We now want to bound from above $r(M) - r(M^0)$ by $R(M) - R(M^0)$. We can use Lemma 2.1 again, to $\tilde{T}_i(\theta) = -T_i(\theta)$ and similar computations yield successively

$$\mathbb{E} \exp\left[\lambda\Big(R(M^0) - R(M) + r(M) - r(M^0)\Big)\right] \leq \exp\left[\frac{v\lambda^2}{2n^2(1-\frac{w\lambda}{n})}\right],$$

and so for any (data-dependent) ρ,

$$\mathbb{E} \exp\left[\beta\left(-\int R d\rho + R(M^0)\right) + \lambda\left(\int r d\rho - r(M^0)\right) - \mathcal{K}(\rho,\pi) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2},$$

where

$$\beta = \left(\lambda + \frac{\lambda^2 \mathcal{C}_{\sigma,L}}{2n(1-\frac{w\lambda}{n})}\right). \tag{2.6}$$

Here again, with the same spirit with α in (2.4), β is of order n also. So:

$$\mathbb{P}\left\{\int r d\rho - r(M^0) \leq \frac{\beta}{\lambda}\left[\int R d\rho - R(M^0)\right] + \frac{1}{\lambda}\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]\right\} \geq 1 - \frac{\varepsilon}{2}. \tag{2.7}$$

Combining (2.7) and (2.5) with a union bound argument gives the general PAC-Bayesian bound

$$\mathbb{P}\left\{R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{\rho \in \mathfrak{M}_+^1(M)} \frac{\beta\left[\int R d\rho - R(M^0)\right] + 2\left[\mathcal{K}(\rho,\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha}\right\} \geq 1 - \varepsilon. \tag{2.8}$$

Step 2:

In the second step, we derive an explicit form for the upper bound in (2.8). The idea is that, if we restrict the infimum in the upper bound in (2.8) to a small set of measures ρ, we are able to provide an explicit bound for this infimum. This trick was introduced in [Catoni, 2004].

Let $M \in \mathcal{M}(L)$, it means that $M = UV^T$ with $|U_{i\ell}| \leq \sqrt{L/K}, |V_{j\ell}| \leq \sqrt{L/K}$. Let us take, for any c such that $\kappa \leq c < (\sqrt{2}-1)\sqrt{L/K}$, the probability distribution

$$\rho_{U,V,c}(\mathrm{d}\mu, \mathrm{d}\nu) \propto \mathbf{1}(\|\mu - U\|_\infty \leq c, \|\nu - V\|_\infty \leq c)\, \pi(\mathrm{d}\mu, \mathrm{d}\nu).$$

Note that, as $c < (\sqrt{2}-1)\sqrt{L/K}$, we have $\mathrm{supp}(\rho_{U,V,c}) \subset \mathrm{supp}(\pi)$ and so

$$\mathcal{K}(\rho_{U,V,c}, \pi) < \infty.$$

Thus, (2.8) becomes

$$\mathbb{P}\left\{ R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{U,V,c} \frac{\beta\left[\int R d\rho_{U,V,c} - R(M^0)\right] + 2\left[\mathcal{K}(\rho_{U,V,c},\pi) + \log\frac{2}{\varepsilon}\right]}{\alpha} \right\} \geq 1-\varepsilon. \tag{2.9}$$

Let us fix c, U, V. The end the proof consists in calculations to derive an upper bound for the two terms in (2.9). Firstly

$$\begin{aligned}
&\int R(M) d\rho_{U,V,c} - R(M^0) \\
&= \int \|\mu\nu^T - M^0\|^2_{F,\Pi}\ \rho_{U,V,c}(\mathrm{d}\mu, \mathrm{d}\nu) \\
&= \int \|\mu\nu^T - U\nu^T + U\nu^T - UV^T + UV^T - M^0\|^2_{F,\Pi}\ \rho_{U,V,c}(\mathrm{d}\mu, \mathrm{d}\nu) \\
&= \int \Big(\|\mu\nu^T - U\nu^T\|^2_{F,\Pi} + \|U\nu^T - UV^T\|^2_{F,\Pi} + \\
&+ \|UV^T - M^0\|^2_{F,\Pi} + 2\left\langle \mu\nu^T - U\nu^T, U\nu^T - UV^T \right\rangle_{F,\Pi} \\
&+ 2\left\langle \mu\nu^T - U\nu^T, UV^T - M^0 \right\rangle_{F,\Pi} \\
&+ 2\left\langle U\nu^T - UV^T, UV^T - M^0 \right\rangle_{F,\Pi} \Big) \rho_{U,V,c}(\mathrm{d}\mu, \mathrm{d}\nu).
\end{aligned}$$

(note that we use the notation $\langle A, B\rangle_{F,\Pi} = \sum_{i,j} A_{ij}B_{ij}\Pi_{ij}$). As $\int \mu\rho_{U,V,c}(\mathrm{d}\mu) = U$ and $\int \nu\rho_{U,V,c}(\mathrm{d}\nu) = V$, it can be seen that integral of the three scalar products in the previous equation vanish. Moreover,

$$\begin{aligned}
\|(\mu - U)\nu^T\|^2_{F,\Pi} &= \sum_{ij} \left[(\mu - U)\nu^T\right]^2_{ij} \Pi_{ij} \leq \left(\sup_{ij} \left[(\mu - U)\nu^T\right]_{ij} \right)^2 \sum_{ij} \Pi_{ij} \\
&\leq \left(\sup_{ij} \sum_{\ell=1}^{K} |\mu - U|_{i\ell} |\nu|_{j\ell} \right)^2 \leq \left(K \sup_{i\ell} |\mu - U|_{i\ell} \sup_{j\ell} |\nu|_{j\ell} \right)^2 \\
&\leq \left[Kc\left(c + \sqrt{\frac{L}{K}} \right) \right]^2 = Kc^2(\sqrt{K}c + \sqrt{L})^2,
\end{aligned}$$

similarly $\|U\nu^T - UV^T\|^2_{F,\Pi} \leq KLc^2$.

Therefore, from (2.9), we have

$$\int \|\mu\nu^T - M^0\|^2_{F,\Pi}\ \rho_{U,V,c}(\mathrm{d}\mu, \mathrm{d}\nu) \leq Kc^2\left[(\sqrt{K}c + \sqrt{L})^2 + L\right] + \|UV^T - M^0\|^2_{F,\Pi}. \tag{2.10}$$

So, we have an upper bound for the first term in (2.9). We now deal with the Kullback-Leibler term:

$$\mathcal{K}(\rho_{U,V,c}, \pi) = \log \frac{1}{\pi(\{\mu, \nu : \|\mu - U\|_\infty \leq c, \|\nu - V\|_\infty \leq c\})}$$

$$
\begin{aligned}
&= \log \frac{1}{\pi(\{\mu : \|\mu - U\|_\infty \leq c\})} + \log \frac{1}{\pi(\{\nu : \|\nu - V\|_\infty \leq c\})} \\
&= \log \frac{1}{\int \pi(\{\|\mu - U\|_\infty \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma} \\
&\qquad + \log \frac{1}{\int \pi(\{\|\nu - V\|_\infty \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma}.
\end{aligned} \tag{2.11}
$$

Note that, up to a reordering of the columns of U and V, we can assume that $U = (U_1|\dots|U_{k_0}|0|\dots|0|)$ and $V = (V_1|\dots|V_{k_0}|0|\dots|0|)$, where $k_0 = \mathrm{rank}(UV^T) \leq K$. Then

$$
\int \pi(\{\|\mu - U\|_\infty \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma \geq \tau^{k_0 - 1}\left(\frac{1-\tau}{1-\tau^K}\right)\pi(\{\|\mu - U\|_\infty \leq c\}|\Gamma = \Gamma_{k_0})
$$

and, as $\kappa \leq c$,

$$
\begin{aligned}
&\pi(\{\|\mu - U\|_\infty \leq c\}|\Gamma = \Gamma_{k_0}) \\
&\geq \prod_{i=1}^{m}\prod_{\ell=1}^{k_0}\pi(\{|\mu_{i\ell} - U_{i\ell}| \leq c\}|\Gamma = \Gamma_{k_0})\prod_{\ell=k_0+1}^{K}\pi(\{|\mu_{i\ell}| \leq c\}|\Gamma = \Gamma_{k_0}) \\
&\geq \left(c\sqrt{\frac{K}{2L}}\right)^{mk_0}.
\end{aligned}
$$

So,

$$
\begin{aligned}
&\log \frac{1}{\int \pi(\{\|\mu - U\|_\infty \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma} \\
&\leq (k_0 - 1)\log(1/\tau) + \log\left(\frac{1-\tau^K}{1-\tau}\right) + mk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) \\
&\leq (k_0 - 1)\log(1/\tau) + \log\left(\frac{1}{1-\tau}\right) + mk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right).
\end{aligned} \tag{2.12}
$$

By symmetry,

$$
\begin{aligned}
&\log \frac{1}{\int \pi(\{\|\nu - V\|_\infty \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma} \\
&\leq (k_0 - 1)\log(1/\tau) + \log\left(\frac{1}{1-\tau}\right) + pk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right).
\end{aligned} \tag{2.13}
$$

Plugging (2.12) and (2.13) into (2.11), we obtain finally our upper bound for the Kullback-Leibler term:

$$
\begin{aligned}
\mathcal{K}(\rho_{U,V,c}, \pi) &\leq 2(k_0 - 1)\log(1/\tau) + 2\log\left(\frac{1}{1-\tau}\right) + (m+p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) \\
&\leq 2k_0 \log(1/\tau) + 2\log\left(\frac{\tau}{1-\tau}\right) + (m+p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right).
\end{aligned} \tag{2.14}
$$

Finally, substituting (2.10) and (2.14) into (2.9), we obtain the following inequality with probability at least $1-\varepsilon$

$$R(\widehat{M}) - R(M^0) \leq \inf_{\substack{U,V,c \\ U_j,V_j=0 \text{ when } j>k_0}} \frac{1}{\alpha}\Bigg[\beta\left(Kc^2\left[(\sqrt{K}c+\sqrt{L})^2+L\right] + \\ + \|UV^T - M^0\|^2_{F,\Pi}\right) + 2(m+p)k_0\log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) + \\ +4k_0\log(1/\tau) + 4\log\left(\frac{\tau}{1-\tau}\right) + 2\log\frac{2}{\varepsilon}\Bigg].$$

Let us put $c = \sqrt{(m+p)L/(18nK)}$. Note that as $n \geq \max(m,p)$ then $\sqrt{(m+p)/(3n)} < 1$ and thus the condition $c < (\sqrt{2}-1)\sqrt{L/K}$ is satisfied. So we have the following inequality with probability at least $1-\varepsilon$:

$$R(\widehat{M}_\lambda) - R(M^0) \\ \leq \inf_{\substack{U,V \\ U_j,V_j=0 \text{ when } j>k_0}} \frac{1}{1-\frac{\lambda\mathcal{C}_{\sigma,L}}{2(n-w\lambda)}}\Bigg\{\left(1+\frac{\lambda\mathcal{C}_{\sigma,L}}{2(n-w\lambda)}\right)\Bigg[\|UV^T - M^0\|^2_{F.\Pi} \\ + L\frac{m+p}{18n}\left(2L\frac{m+p}{18n}+3L\right)\Bigg] + \frac{2}{\lambda}\Bigg[(m+p)k_0\log\left(\sqrt{\frac{36n}{m+p}}\right) \\ + 2k_0\log(1/\tau) + 2\log\left(\frac{\tau}{1-\tau}\right) + \log\frac{2}{\varepsilon}\Bigg]\Bigg\},$$

where α and β have been replaced by their definitions, see (2.4) and (2.6).

Taking now $\lambda = \lambda^* = n/(2\mathcal{C})$ with $\mathcal{C} = \mathcal{C}_{\sigma,L} \vee w$ in the last above display, we obtain the following inequality with probability at least $1-\varepsilon$

$$R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M\in\mathcal{M}(L)}\Bigg\{3\Bigg[L^2\frac{m+p}{18n}\left(\frac{m+p}{9n}+3\right) + \|M-M^0\|^2_{F,\Pi}\Bigg] \\ +\frac{8\mathcal{C}}{n}\Bigg[\frac{1}{2}(m+p)\text{rank(M)}\log\left(\frac{36n}{m+p}\right) + \log\frac{2}{\varepsilon} \\ +2\text{rank}(M)\log(1/\tau) + 2\log\left(\frac{\tau}{1-\tau}\right)\Bigg]\Bigg\}, \tag{2.15}$$

where we have used that $1 - \frac{\lambda\mathcal{C}_{\sigma,L}}{2(n-w\lambda)} \geq 1/2$ and $1 + \frac{\lambda\mathcal{C}_{\sigma,L}}{2(n-w\lambda)} \leq 3/2$.

As

$$\log\left(\frac{36n}{m+p}\right) \leq \log\left(\frac{36mp}{\max(m,p)}\right)$$

$$= \log\left(\frac{36\min(m,p)\max(m,p)}{\max(m,p)}\right) = \log\left(36K\right),$$

we have

$$\mathbb{P}\Bigg\{R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M\in\mathcal{M}(L)}\Bigg\{3\Bigg[L^2\frac{m+p}{18n}\left(\frac{m+p}{9n}+3\right) + \|M-M^0\|^2_{F,\Pi}\Bigg] + \\ + \frac{8\mathcal{C}}{n}\Bigg[\frac{1}{2}(m+p)\text{rank(M)}\log(36K) + \log\frac{2}{\varepsilon} + \\ + 2\text{rank}(M)\log(1/\tau) + 2\log\left(\frac{1}{1-\tau}\right)\Bigg]\Bigg\}\Bigg\} \geq 1-\varepsilon. \tag{2.16}$$

Moreover,

$$L^2\frac{m+p}{6n}\left(\frac{m+p}{9n}+3\right) \leq \mathscr{C}(L)\frac{(m+p)\text{rank(M)}\log(\text{K})}{n},$$

for some constant $\mathscr{C}(L) > 0$ depending on L only. Remind that τ is a constant in $(0,1)$, we have

$$2\text{rank}(M)\log(1/\tau) + 2\log\left(\frac{\tau}{1-\tau}\right) \leq \mathscr{C}(\tau)\frac{(m+p)\text{rank(M)}\log(\text{K})}{n},$$

for some constant $\mathscr{C}(\tau) > 0$ depending on τ only. Finally, from (2.16), we obtain

$$\mathbb{P}\Bigg\{R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M\in\mathcal{M}(L)}\Bigg[3\|M-M^0\|^2_{F,\Pi} + \mathscr{C}(L,\mathcal{C},\tau)\frac{(m+p)\text{rank(M)}\log(\text{K})}{n} \\ + \frac{8\mathcal{C}\log\left(\frac{2}{\varepsilon}\right)}{n}\Bigg]\Bigg\} \geq 1-\varepsilon,$$

for some constant $\mathscr{C}(L,\mathcal{C},\tau) > 0$ depending only on L,τ and $\mathcal{C}$. However, as the constant $\mathcal{C}$ also depends on L,ξ,σ then $\mathscr{C}(L,\mathcal{C},\tau)$ can be rewritten as $\mathscr{C}_{L,\xi,\sigma,\tau}$ as in the statement of the theorem. □

Chapter 3

QUANTUM STATE TOMOGRAPHY

Quantum state tomography, an important task in quantum information processing, aims at reconstructing a state from prepared measurement data. Bayesian methods are recognized to be one of the good and reliable choice in estimating quantum states [Blume-Kohout, 2010]. Several numerical works showed that Bayesian estimations are comparable to, and even better than other methods in the problem of 1-qubit state recovery. However, the problem of choosing prior distribution in the general case of n qubits is not straightforward. More importantly, the statistical performance of Bayesian type estimators have not been studied from a theoretical perspective yet. In this chapter, we propose a novel (low-rank) prior for quantum states (density matrices), and we define pseudo-Bayesian estimators of the density matrix. Then, using PAC-Bayesian theorems [Catoni, 2007], we derive rates of convergence for the posterior mean. The numerical performance of these estimators are tested on simulated and real datasets.

The works in this chapter have been published in [Mai and Alquier, 2017]:

T.T. Mai & P. Alquier. Pseudo-bayesian quantum tomography with rank-adaptation. ***Journal of Statistical Planning and Inference***, vol.**184**: 62–76, **2017**.

3.1 Introduction

Playing a vital role in quantum information processing, as well as being fundamental for characterizing quantum objects, quantum state tomography focuses on reconstructing the (unknown) state of a physical quantum system [Paris and Řeháček, 2004], usually represented by the so-called density matrix ρ (the exact definition of a density matrix is given in Section 3.2). This task is done by using outcomes of measurements performed on many

independent systems identically prepared in the same state.

The 'tomographic' method, also named as linear/direct inversion [Vogel and Risken, 1989; Řeháček et al., 2010], is the simplest and oldest estimation procedure. It is actually the analogous of the least-square estimator in the quantum setting. Although easy in computation and providing unbiased estimate [Schwemmer et al., 2015], it does not generate a physical density matrix as an output [Shang et al., 2014]. Maximum likelihood estimation [Hradil et al., 2004] is the current procedure of choice. Unfortunately, it has some critical flaws detailed in [Blume-Kohout, 2010], including a huge computational complexity. Furthermore, both these methods are not adaptive to the case where a system is in a state ρ for which some additional information is available. Note especially that, physicists focus on so-called pure states, for which $\text{rank}(\rho) = 1$.

The problem of rank-adaptivity was tackled thanks to adequate penalization. Rank-penalized maximum likelihood (BIC) was introduced in [Guţă et al., 2012] while a rank-penalized least-square estimator $\hat{\rho}_{\text{rank}-\text{pen}}$ was proposed in [Alquier et al., 2013a], together with a proof of its consistency. More specifically, when the density matrix of the n-qubits system is ρ^0 with $r = \text{rank}(\rho^0)$, the authors of [Alquier et al., 2013a] proved that the Frobenius norm of the estimation error satisfies

$$\|\hat{\rho}_{\text{rank}-\text{pen}} - \rho^0\|_F^2 = \mathcal{O}(r4^n/N)$$

where N is the number of quantum measurements. The rate was improved to $\mathcal{O}(r3^n/N)$ by [Butucea et al., 2015], using a thresholding method. Note that the rate $\mathcal{O}(r2^n/N)$ was first claimed in the paper, but in the Corrigendum [Butucea et al., 2016], the authors acknowledge that this is not the case. The paper however contains a proof that no method can reach a rate smaller than $r2^n/N$. So, the minimax-optimal rate is somewhere in between $r2^n/N$ and $r3^n/N$.

Note that all the aforementioned papers only cover the complete measurement case (the definition is given in Section 3.2, basically it means that we have observations for all the observables given by the Pauli basis). The statistical relationship between matrix completion and quantum tomography with incomplete measurements (in the Le Cam paradigm) has been investigated in [Wang, 2013]. Thus compressed sensing ideas have been successfully proposed in estimating a density state from incomplete measurements [Gross et al., 2010; Gross, 2011; Flammia et al., 2012; Koltchinskii, 2011].

On the other hand, Bayesian estimation has been considered in this context. The papers [Bužek et al., 1998; Baier et al., 2007] compare Bayesian methods to other methods on simulated data. More recently, [Kravtsov et al., 2013; Ferrie, 2014; Kueng and Ferrie, 2015; Schmied, 2016] discuss efficient algorithms for computing Bayesian estimators. Importantly, [Blume-Kohout, 2010] showed that Bayesian method comes with natural error bars and is the most accurate scheme w.r.t. the expected error (operational divergence) (even) with finite samples. However, there is no theoretical guarantee on the convergence of these estimators.

More works on quantum state tomography in various settings include [Audenaert and Scheel, 2009; Carlen, 2010; Rau, 2011, 2014; Ferrie and Granade, 2014].

In this chapter, we consider a pseudo-Bayesian estimation, where the likelihood is replaced by pseudo-likelihoods based on various moments (two estimators, corresponding to two different pseudo-likelihood, are actually proposed). Using PAC-Bayesian theory [Shawe-Taylor and Williamson, 1997; McAllester, 1998; Catoni, 2004, 2007; Dalalyan and Tsybakov, 2008; Suzuki, 2012], we derive oracle inequalities for the pseudo-posterior mean. We obtain rates of convergence for these estimators in the complete measurement setting. One of them has a rate as good as the best known rate up to date $\mathcal{O}(\text{rank}(\rho^0)3^n/N)$ (still, the other one is interesting for computationnal reasons that are discussed in the chapter).

The rest of the chapter is organized as follow. We recall the standard notations and basics about quantum theory in Section 3.2. Then the definition of the prior and of the estimators are presented in Section 3.3. The statistical analysis of the estimators are given in Section 3.4, while all the proofs are delayed to the Appendix 3.7. Some numerical experiments on simulated and real datasets are given in Section 3.5.

3.2 Preliminaries

3.2.1 Problem setup

A very good introduction to the notations and problems of quantum statistics is given in [Artiles et al., 2005]. Here, we only provide the basic definitions required for understanding the material in thie chapter.

In quantum physics, all the information on the physical state of a system can be encoded in its *density matrix* ρ. Depending on the system in hand, this matrix can have a finite or infinite number of entries. A two-level system of n-qubits is represented by a $2^n \times 2^n$ density matrix ρ, with coefficients in $\mathbb{C}$. For the sake of simplicity, the notation $d = 2^n$ is used in [Butucea et al., 2015], so note that ρ is a $d \times d$ matrix. This density matrix is

- Hermitian $\rho^\dagger = \rho$ (i.e. self-adjoint),
- semidefinite positive $\rho \geq 0$,
- and has $\text{Trace}(\rho) = 1$.

Additionally, it often makes sense to assume that the rank of ρ is small [Gross et al., 2010; Gross, 2011]. In theory, the rank can be any integer between 1 and 2^n, but physicists are especially interested in pure states and a pure state ρ can be defined by $\text{rank}(\rho) = 1$.

The objective of quantum tomography is to estimate ρ on the basis of experimental observations of many independent and identically systems prepared in the state ρ by the same experimental device.

For each particle (qubit), one can measure one of the three Pauli–observables

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}; \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}; \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

The outcome for each will be 1, or -1, randomly (the corresponding probability depends on the state ρ and will be given in (3.1) below). Thus for a n-qubits system, we consider 3^n possible experimental observables. The set of all possible performed observables is

$$\{\sigma_{\mathbf{a}} = \sigma_{a_1} \otimes \ldots \otimes \sigma_{a_n};\ \mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{E}^n := \{x, y, z\}^n\},$$

where vector $\mathbf{a}$ identifies the experiment. The outcome for each fixed observable setting will be a random vector $\mathbf{s} = (s_1, \ldots, s_n) \in \mathcal{R}^n := \{-1, 1\}^n$, thus there are 2^n outcomes in total.

Let us denote $R^{\mathbf{a}}$ a $\mathcal{R}^n$-valued random vector that is the outcome of an experiment indexed by $\mathbf{a}$. From the basic principles of quantum mechanics (Born's rule), its probability distribution is given by

$$\forall \mathbf{s} \in \mathcal{R}^n, p_{\mathbf{a},\mathbf{s}} := \mathbb{P}(R^{\mathbf{a}} = \mathbf{s}) = \text{Trace}\left(\rho \cdot P_{\mathbf{s}}^{\mathbf{a}}\right), \tag{3.1}$$

where $P_{\mathbf{s}}^{\mathbf{a}} := P_{s_1}^{a_1} \otimes \cdots \otimes P_{s_n}^{a_n}$ and $P_{s_i}^{a_i}$ is the orthogonal projection associated to the eigenvalue s_i in the diagonalization of σ_{a_i} for $a_i \in \{x, y, z\}$ and $s_i \in \{-1, 1\}$ – that is $\sigma_{a_i} = -1P_{-1}^{a_i} + 1P_{+1}^{a_i}$.

The quantum state tomography problem is as follows: a physicist has access to an experimental device that produces n-qubits in a state ρ^0, and ρ^0 is assumed to be unknown. He/she can produce a large number of replications of the n-qubits and wants to infer ρ^0 from this.

In the complete measurement case, for *each* experiment setting $\mathbf{a} \in \mathcal{E}^n$, the experimenter repeats m times the experiment corresponding to $\mathbf{a}$ and thus collects m independent random copies of $R^{\mathbf{a}}$, say $R_1^{\mathbf{a}}, \ldots, R_m^{\mathbf{a}}$. As there are 3^n possible experiment settings $\mathbf{a}$, we define the **quantum sample** size as $N := m \cdot 3^n$. We will refer to $(R_i^{\mathbf{a}})_{i \in \{1,\ldots,m\}, \mathbf{a} \in \mathcal{E}^n}$ as $\mathcal{D}$ (for data).

Note that the case where we would only have access to experiments $\mathbf{a} \in \mathcal{A}$ where $\mathcal{A}$ is some proper subset of $\mathcal{E}^n$ ($A \subsetneq \mathcal{E}^n$) is referred to as the incomplete measurement case. In this work, we focus on the complete measurement case, but the extension to the incomplete case is discussed in Section 3.6.

3.2.2 Popular estimation methods

A natural idea is to define the empirical frequencies

$$\hat{p}_{\mathbf{a},\mathbf{s}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{R_i^{\mathbf{a}} = \mathbf{s}\}}.$$

Note that $\hat{p}_{\mathbf{a},\mathbf{s}}$ is an unbiased estimator of the probability $p_{\mathbf{a},\mathbf{s}}$. The inversion method is based on solving the linear system of equations

$$\begin{cases} \hat{p}_{\mathbf{a},\mathbf{s}} = \text{Trace}\left(\hat{\rho} \cdot P_{\mathbf{s}}^{\mathbf{a}}\right), \\ \mathbf{a} \in \mathcal{E}^n, \\ \mathbf{s} \in \mathcal{R}^n. \end{cases} \tag{3.2}$$

As mentioned above, the computation of $\hat{\rho}$ is quite straighforward. Explicit formulas are classical, see e.g. [Alquier et al., 2013a].

Another commonly used method is maximum likelihood (ML) estimation, where the likelihood is

$$\mathcal{L}(\rho; \mathcal{D}) \propto \prod_{\mathbf{a}\in\mathcal{E}^n} \prod_{\mathbf{s}\in\mathcal{R}^n} \left[\text{Trace}\left(\rho \cdot P_{\mathbf{s}}^{\mathbf{a}}\right)\right]^{n_{\mathbf{a},\mathbf{s}}},$$

where $n_{\mathbf{a},\mathbf{s}} = m\hat{p}_{\mathbf{a},\mathbf{s}}$ is the number of times we observed output $\mathbf{s}$ in experiment $\mathbf{a}$ (obviously, $\sum_{\mathbf{s}} n_{\mathbf{a},\mathbf{s}} = m$). As mentioned in the introduction, both methods suffer many drawbacks. The inversion method returns a matrix $\hat{\rho}$ that usually does not satisfy the axioms of a density matrix. ML becomes expensive (impractical) for $n \geq 10$. Moreover, these two methods can not take advantage of a prior knowledge (e.x. low-rank state).

Considering the expansion of the density matrix ρ in the n–Pauli basis, i.e. $\mathcal{B} = \{\sigma_b = \sigma_{b_1} \otimes \ldots \otimes \sigma_{b_n}, b \in \{I, x, y, z\}^n\}, \sigma_I = I$,

$$\rho = \sum_{b\in\{I,x,y,z\}^n} \rho_b \sigma_b. \tag{3.3}$$

One can also estimate the density matrix via estimating the coefficients in the Pauli expansion. This was studied in [Cai et al., 2015b] where the authors also make a sparsity assumption: that is, most of ρ_b are small or very close to 0. Note that, this is not related to the setting we explore (low-rank assumption).

We now turn to the definition of a prior distribution on density matrices that will allow to perform (pseudo-)Bayesian estimation.

3.3 Pseudo-Bayesian estimation and prior distribution on density matrices

3.3.1 Peudo-Bayesian estimation

We remind that the idea of Bayesian statistics is to encode the prior information on density matrices through a prior distribution $\pi(\mathrm{d}\rho)$. Inference is then done through the posterior distribution

$$\pi(\mathrm{d}\rho|\mathcal{D}) \propto \mathcal{L}(\rho)\pi(\mathrm{d}\rho).$$

Here, for computational reasons, we replace the likelihood by a pseudo-likelihood. This is an increasingly popular method in Bayesian statistics [Grünwald,

2012; Bissiri et al., 2016] and in machine learning [Catoni, 2007; Alquier et al., 2016; Bégin et al., 2016]. We define, for density matrices ν, the pseudo-posterior by

$$\tilde{\pi}_\lambda(\mathrm{d}\nu) \propto \exp\left[-\lambda \ell(\nu, \mathcal{D})\right] \pi(\mathrm{d}\nu), \tag{3.4}$$

the pseudo-likelihood being $\exp\left[-\lambda \ell(\nu, \mathcal{D})\right]$. The term $\ell(\nu, \mathcal{D})$ can be specified by the user. Two examples are provided in Section 3.4. As a replacement of the likelihood, this term plays the role of the empirical evidence. More specially

- the role of $\exp\left[-\lambda \ell(\nu, \mathcal{D})\right]$ is to give more weight to the density ν when it fits the data well;
- the role of $\pi(\mathrm{d}\nu)$, the prior, is to restrict the posterior to the space of densities (and even give more weight to low-rank matrices if needed);
- $\lambda > 0$ is a free parameter that allows to tune the balance between evidence from the data and prior information.

We finally define the pseudo-posterior mean (also refered to as Gibbs estimator, PAC-Bayesian estimator or EWA, for exponentially weighted aggregate [Catoni, 2007; Dalalyan and Tsybakov, 2008]):

$$\tilde{\rho}_\lambda = \int \nu \tilde{\pi}_\lambda(d\nu).$$

The definition of the estimator $\tilde{\rho}_\lambda$ based on the pseudo-posterior $\tilde{\pi}_\lambda$ is actually validated by the theoretical results from Section 3.4.

3.3.2 Definition of the prior

In the single qubit state estimation $n = 1$, the representation of the quantum constraints is explicit [Baier et al., 2007; Schmied, 2016]. Thus, one can place a prior distribution on the polar reparametrization of the density. Up to our knowledge, this has not been extended to the case $n > 1$, and this extension seems not straightforward. For general n-qubit densities, uninformative priors (e.g the Haar measure) are put on $\psi_{d\times K}$ matrices ($K \geq d$) and the density state is built by $\rho = \psi_{d\times K}\psi_{d\times K}^\dagger$ [Struchalin et al., 2016; Granade et al., 2016; Huszár and Houlsby, 2012; Kueng and Ferrie, 2015; Życzkowski et al., 2011]. One could also define a prior on the coefficients $\{\rho_b\}$ of ρ on the Pauli basis. Nevertheless, none of these approaches seem helpful for rank adaptation.

The idea for our prior is inspired by the priors used for low-rank matrix estimation in machine learning, e.g. [Mai and Alquier, 2015; Cottet and Alquier, 2016] and the references therein. Hereafter, we describe in details the prior construction.

Let V be a vector in $\mathbb{C}^{d\times 1} \setminus \{\mathbf{0}\}$ ($d = 2^n$ in our model), then $VV^\dagger$ is a Hermitian, semi-definite positive matrix in $\mathbb{C}^{d\times d}$ with $\mathrm{rank}(VV^\dagger) = 1$.

Additionally, we can normalize V (that is replace V by $V/\|V\|$), this leads to $\text{Trace}(VV^\dagger) = 1$. So, $VV^\dagger$ satisfies the conditions of a density matrix (with rank-1).

Now, let $V_1, \dots, V_d$ be d normalized vectors in $\mathbb{C}^{d\times 1} \setminus \{\mathbf{0}\}$ and $\gamma_1, \dots, \gamma_d$ be non-negative weights with $\sum_{j=1}^d \gamma_j = 1$. Put

$$\nu = \sum_{i=1}^d \gamma_i V_i V_i^\dagger. \tag{3.5}$$

Then ν is clearly a density matrix: it is Hermitian (as a sum of Hermitian matrices), it is semi-definite positive (same reason) and

$$\text{Tr}(\nu) = \sum_{i=1}^d \gamma_i \text{Tr}(V_i V_i^\dagger) = 1.$$

Moreover, note that any density matrix can be written in such way, as we know that for any density matrix ρ,

$$\rho = U\Lambda U^\dagger \tag{3.6}$$

and just write $U = (U_1|\dots|U_d)$ with the U_i's being *orthogonal*, where $\Lambda = \text{diag}(\Lambda_1, \dots, \Lambda_d) : \Lambda_1 \geq \dots \geq \Lambda_d \geq 0, \sum_{i=1}^d \Lambda_i = 1$.

The only difference in (3.5) is that we do not require that the V_i's are orthogonal. Thus, it is easier to simulate a matrix ρ by simulating the V_i's and $\gamma_i's$ in (3.5) than by simulating U and Λ in (3.6). Also, note that the γ_i's are not necessarily the eigenvalues of ρ.

Definition 3.1. *We define the prior definition on ρ, $\pi(\mathrm{d}\rho)$, by*

$$\begin{aligned} V_1, \dots, V_d &\sim \text{ i.i.d uniform distribution on the unit sphere,} \\ (\gamma_1, \dots, \gamma_d) &\sim \mathcal{D}ir(\alpha_1, \dots, \alpha_d), \\ \rho &= \sum_{i=1}^d \gamma_i V_i V_i^\dagger \end{aligned}$$

where $\mathcal{D}ir(\alpha_1, \dots, \alpha_d)$ is the Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_d > 0$.

Remark 3.1. *To get an approximate rank-1 matrix ρ, one can take all parameters of the Dirichlet distribution equal to a constant that is very closed to 0 (e.g $\alpha_1 = \dots = \alpha_d = \frac{1}{d}$). And a typical drawing will lead to one of the $\gamma_i's$ close to 1 and the others close to 0. See [Wallach et al., 2009] for more discussion on choosing the parameters for Dirichlet distribution. Theoretical recommendations for the α_i's are given in Section 3.4 below.*

Remark 3.2. *We could impose the V_i's to be orthogonal in practice. The theoretical results would be unchanged, however, the implementation of our method would become trickier. Note that to sample from the uniform distribution on the sphere is rather easy. We can for example simulate $\tilde{V}_i$ from any isotropic distribution, e.g. $\mathcal{N}(0, \mathbb{I})$ and define $V_i := \tilde{V}_i/\|\tilde{V}_i\|$.*

3.4 PAC-Bayesian estimation and analysis

3.4.1 Pseudo-likelihoods

Here, we consider two natural ways to compare a theoretical density ρ and the observations: first $p_{\mathbf{a},\mathbf{s}}$ should be close to the empirical part $\hat{p}_{\mathbf{a},\mathbf{s}}$; second ρ should be close to the least square (invert) estimator $\hat{\rho}$. As we have no reason to prefer one in advance, we define and study 2 estimators.

(a) Distance between the probabilities: prob-estimator

We consider

$$\ell^{prob}(\nu, \mathcal{D}) = \sum_{\mathbf{a}\in\mathcal{E}^n} \sum_{\mathbf{s}\in\mathcal{R}^n} [\mathrm{Tr}(\nu P_{\mathbf{s}}^{\mathbf{a}}) - \hat{p}_{\mathbf{a},\mathbf{s}}]^2$$

and

$$\tilde{\rho}_\lambda^{prob} = \int \nu \tilde{\pi}_\lambda^{prob}(d\nu),$$
$$\tilde{\pi}_\lambda^{prob}(\mathrm{d}\nu) \propto \exp\left[-\lambda \ell^{prob}(\nu, \mathcal{D})\right] \pi(\mathrm{d}\nu).$$

Note that if we use the shortened notation $p_\nu = [\mathrm{Tr}(\nu P_{\mathbf{s}}^{\mathbf{a}})]_{\mathbf{a},\mathbf{s}}$ and $\hat{p} = [\hat{p}_{\mathbf{a},\mathbf{s}}]_{\mathbf{a},\mathbf{s}}$ then

$$\ell^{prob}(\nu, \mathcal{D}) = \|p_\nu - \hat{p}\|_F^2$$

(Frobenius norm). This distance quantifies how far the probabilities and the empirical frequencies in the sample are.

(b) Distance between the density matrices: dens-estimator

Now, let us take:

$$\ell^{dens}(\nu, \mathcal{D}) = \|\nu - \hat{\rho}\|_F^2.$$

and

$$\tilde{\rho}_\lambda^{dens} = \int \nu \tilde{\pi}_\lambda^{dens}(d\nu),$$
$$\tilde{\pi}_\lambda^{dens}(\mathrm{d}\nu) \propto \exp\left[-\lambda \ell^{dens}(\nu, \mathcal{D})\right] \pi(\mathrm{d}\nu).$$

In another words, this estimator finds a balance between prior information and closeness to the least square estimate $\hat{\rho}$. From a computational point of view, this estimator is easier to implement than the previous estimator.

3.4.2 Statistical properties of the estimators

Before analyzing statistical properties of our estimators, we make some assumption on the prior distribution. Typically, this assumption aims at producing low-rankness.

Assumption 3.1. *Fix some constants $D_1 > 0$ and $D_2 > 0$ (that do not depend on m nor n). We assume that the parameters of the Dirichlet prior distribution $\mathcal{D}ir(\alpha_1, \ldots, \alpha_d)$ satisfy*

- $\forall i = 1, \ldots, d : \alpha_i \leq 1$,
- $\sum_{i=1}^{d} \alpha_i = D_1$,
- $\prod_{i=1}^{d} \alpha_i \geq e^{-D_2 d \log(d)}$.

Note that this assumption is satisfied for $\alpha_1 = \ldots = \alpha_d = 1/d$ with $D_1 = D_2 = 1$.

The first theorem provides the concentration bound on the square error of the first estimator $\tilde{\rho}_{\lambda}^{prob}$. The proof of this theorem is left to Section 3.7.

Theorem 3.1. *Fix a small $\epsilon \in (0, 1)$. Under Assumption 3.1, for $\lambda = \lambda^* := m/2$, with probability at least $1 - \epsilon$, one has*

$$\|\tilde{\rho}_{\lambda^*}^{prob} - \rho^0\|_F^2 \leq C_{D_1,D_2}^{prob} \frac{3^n \mathrm{rank}(\rho^0) \log\left(\frac{\mathrm{rank}(\rho^0)N}{2^n}\right) + (1.5)^n \log(2/\epsilon)}{N},$$

where C_{D_1,D_2}^{prob} is a constant that depends only on D_1, D_2.

Remark 3.3. *As said in the introduction, the best known rate up-to-date in this problem is $\frac{3^n \mathrm{rank}(\rho^0)}{N}$, so our estimator $\tilde{\rho}_{\lambda^*}^{prob}$ reaches this rate (up to log terms). This rate is actually $\left(\frac{3}{2}\right)^n \frac{rd}{N}$ and the best lower bound known in this case is $\frac{rd}{N}$ [Butucea et al., 2015] (we remind that $d = 2^n$).*

The next theorem presents the square error bound of the second estimator $\tilde{\rho}_{\lambda}^{dens}$. Here again, see the Section 3.7 for the proof.

Theorem 3.2. *Fix a small $\epsilon \in (0, 1)$. Under Assumption 3.1, for $\lambda = \lambda^* := \frac{N}{5^n 4}$, with probability at least $1 - \epsilon$,*

$$\|\tilde{\rho}_{\lambda^*}^{dens} - \rho^0\|_F^2 \leq C_{D_1,D_2}^{dens} \frac{10^n \mathrm{rank}(\rho^0) \log\left(\frac{\mathrm{rank}(\rho^0)N}{2^n}\right) + 5^n \log(2/\varepsilon)}{N} \tag{3.7}$$

where C_{D_1,D_2}^{dens} is a constant that depends only on D_1, D_2.

The guarantee for $\tilde{\rho}_{\lambda^*}^{dens}$ is far less satisfactory. However, as this estimator is easier to compute, we think it is interesting to provide a convergence rate, even if it is far from optimal: note that for a fixed d, the bound goes to 0 when $m \to \infty$.

Remark 3.4. *Experiments show that $\lambda = \lambda^* := \frac{N}{5^n 4}$ is actually not the best choice for dens-estimator. The choice $\lambda = \frac{N}{4}$ (heuristically motivated by [Dalalyan and Tsybakov, 2008]) leads to results comparable to the prob-estimator in Section 3.5. This leads to the conjecture that the rate of $\tilde{\rho}_{N/4}^{dens}$ is much better than $\frac{10^n \mathrm{rank}(\rho^0)}{N}$ but this is still an open question.*

3.5 Numerical Experiments

3.5.1 Metropolis-Hastings Implementation

We implement the two proposed estimators via the Metropolis-Hasting (MH) algorithm [Robert and Casella, 2013]. Note that to draw $(\gamma_1, \ldots, \gamma_d) \sim \mathcal{D}ir(\alpha, \ldots, \alpha)$ is equivalent to draw $\gamma_i = Y_i/(Y_1 + \ldots + Y_d)$ with $Y_i \overset{i.i.d}{\sim} Gamma(\alpha, 1), \forall i = 1, \ldots, d$. Thus, instead of $\gamma_i's$, we conduct a MH updating for $Y_i's$. So the objective is to produce a Markov chain $(Y_1^{(t)}, \ldots, Y_d^{(t)}, V_1^{(t)}, \ldots, V_d^{(t)})$. From this, we deduce obviously the sequence $(\gamma_1^{(t)}, \ldots, \gamma_d^{(t)}, V_1^{(t)}, \ldots, V_d^{(t)})$ and use the following empirical mean as the Monte-Carlo approximation of our estimator:

$$\hat{\rho}^{\mathrm{MH}} := \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{i=1}^{d} \gamma_i^{(t)} V_i^{(t)} (V_i^{(t)})^\dagger \right).$$

Algorithm 3 MH implementation

For t from 1 to T, we iteratively update through the following steps:

updating for $Y_i's$: for i from 1 to d,

Sample $\tilde{Y}_i \sim h(y|Y_i^{(t-1)})$ where h is a proposal distribution given explicitely below.
Calculate $\tilde{\gamma}_i = \tilde{Y}_i/(\sum_{i=1}^{d} \tilde{Y}_i)$.
Set

$$Y_i^{(t)} = \begin{cases} \tilde{Y}_i & \text{with probability } \min\left\{1, R(\tilde{Y}, Y^{(t-1)})\right\}, \\ Y_i^{(t-1)} & \text{otherwise} \end{cases}$$

where $R(\tilde{Y}, Y^{(t-1)})$ is the acceptance ratio given below.
Put $\gamma_i^{(t)} = Y_i^{(t)}/(\sum_{j=1}^{d} Y_j^{(t)}), i = 1, \ldots, d$.

updating for $V_i's$: for i from 1 to d,

Sample $\tilde{V}_i$ from the uniform distribution on the unit sphere.
Set

$$V_i^{(t)} = \begin{cases} \tilde{V}_i & \text{with probability } \min\{1, A(V^{(t-1)}, \tilde{V})\}, \\ V_i^{(t-1)} & \text{otherwise,} \end{cases}$$

where $A(V^{(t-1)}, \tilde{V})$ is the acceptance ratio given below.

Let us now give precisely h, R and A. We define $h(\cdot|Y_i^{(t-1)})$ as the probability distribution of $U = Y_i^{(t-1)} \exp(y)$ where $y \sim \mathcal{U}(-0.5, 0.5)$. Following [Robert and Casella, 2013] the acceptance ratios are then given by:

$$\log(R(\tilde{Y}, Y^{(t-1)})) = \lambda\ell\left(\sum_{i=1}^{d} \tilde{\gamma}_i V_i V_i^\dagger, \mathcal{D} \right) - \lambda\ell\left(\sum_{i=1}^{d} \gamma_i^{(t-1)} V_i V_i^\dagger, \mathcal{D} \right)$$

$$+\sum_{i=1}^{d}((\alpha-1)\log(\tilde{Y}_i)-\tilde{Y}_i)-\sum_{i=1}^{d}((\alpha-1)\log(Y_i^{(t-1)})-Y_i^{(t-1)})$$
$$+\sum_{i=1}^{d}\tilde{Y}_i-\sum_{i=1}^{d}Y_i^{(t-1)}$$

and

$$\log(A(V^{(t-1)},\tilde{V}))=\lambda\ell\left(\sum_{i=1}^{d}\gamma_i\tilde{V}_i\tilde{V}_i^{\dagger},\mathcal{D}\right)-\lambda\ell\left(\sum_{i=1}^{d}\gamma_i V_i^{(t-1)}(V_i^{(t-1)})^{\dagger},\mathcal{D}\right)$$

where $\ell(\cdot,\mathcal{D})$ stands for $\ell^{dens}(\cdot,\mathcal{D})$ or $\ell^{prob}(\nu,\mathcal{D})$ depending on the estimator we are computing.

3.5.2 Experiments and Results

We study the numerical performance of the prob-estimators with $\lambda=m/2$, i.e. $\tilde{\rho}^{prob}_{m/2}$ and the dens-estimator with $\lambda=\frac{N}{4}$, i.e. $\tilde{\rho}^{dens}_{N/4}$ on the following settings, all with $n=2,3,4$ ($d=4,8,16$):

- a pure state density (rank-1) $\rho=\psi\psi^{\dagger}$ with $\psi\in\mathbb{C}^{d\times 1}$,
- a rank-2 density matrix that $\rho_{rank-2}=\frac{1}{2}\psi_1\psi_1^{\dagger}+\frac{1}{2}\psi_2\psi_2^{\dagger}$ with ψ_1,ψ_2 being two normalized orthogonal vectors in $\mathbb{C}^{d\times 1}$,
- an "approximate rank-2" density matrix: $\rho=w\rho_{rank-2}+(1-w)\frac{\mathbb{I}_d}{d}, w=0.98$. Note that by "approximate rank-2", we mean that ρ is very well approximated by a rank-2 matrix ρ_{rank-2} (in the sense that $\|\rho-\rho_{rank-2}\|_F^2$ is small), but in general ρ itself is full rank,
- a maximal mixed state (rank-d).

The experiments are done for $m=20;200;1000;2000$. The parameter for $\mathcal{D}ir(\alpha,\ldots,\alpha)$ is $\alpha=0.5$. We repeat each experiment 10 times, and compute the mean of the square errors, MSEs, $\|\hat{\rho}-\rho\|_F^2$ for each estimator, together with the associated standard deviation (between brackets in Tables 3.1,3.2,3.3).

We compare the prob- and dens-estimator to the simple inversion procedure and to the thresholding estimator of [Butucea et al., 2015]. The results are given in Tables 3.1,3.2,3.3 (outputs from the **R** software [R Core Team, 2014]). The conclusions are:

- The prob-estimator seems to be the most accurate but also comes with a larger standard deviation. This might be due to slow convergence of the MCMC procedure. Indeed each step is computationally highly expensive.

- The dens-estimator is easier to compute and while it is less accurate than the prob-estimator, it still shows better results than the direct inversion method.

- The thresholding estimator of [Butucea et al., 2015] works well for rank-1 states but seems to bring too much bias for other states.

Besides the square error, the eigenvalues of the estimates are also important when reconstructing density matrices. In Figure 3.1, the dens-estimator returns with eigenvalues similar to the true eigenvalues of the true density matrix, while the prob-estimator seems not to shrink enough.

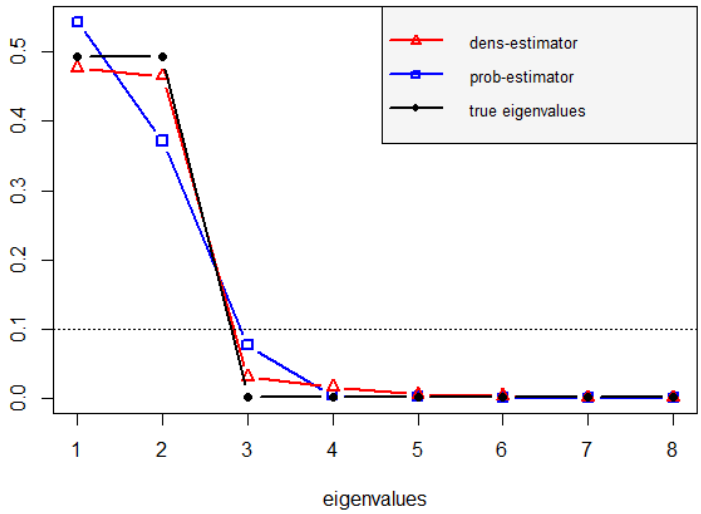

Figure 3.1: Eigenvalues of estimates for an "approximate rank-2" density with $d = 2^3, m = 200$.

Another natural question is: are the γ_i's close to the eigenvalues of our estimator? In our simulations, it doesn't seem to be the case. However, it seems that the number of significant γ_i's is a fair indicator of the number of significant eigenvalues in our estimator, but only for the prob-estimator. This is illustrated in Figure 3.2. We currently do not have any explanation for this fact.

3.5.3 Real data tests

The experiments performed to produce the data is explained in [Barreiro et al., 2010]. The data was kindly provided by M. Guţă and T. Monz. It had been used in [Alquier et al., 2013a; Guţă et al., 2012]. We apply two proposed estimators to the real data set of a system of 4 ions which is Smolin state further manipulated. In Figure 3.3 we plot the eigenvalues of the inversion

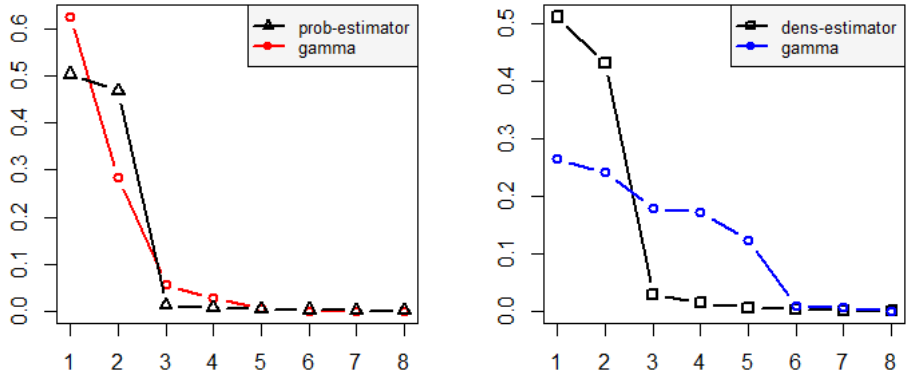

Figure 3.2: Plot for comparing the difference between the $\gamma_i, i = 1, \ldots, d$ and the eigenvalues of the proposed estimator for an approximate rank-2 density with $d = 2^3, m = 200$.

Table 3.1: MSEs for $n = 4$ (together with standard deviations)

	$m = 20$	$m = 200$	$m = 1000$	$m = 2000$
pure state, MSEs$\times 10^5$				
Inversion	175 (4e-4)	14.8 (2e-5)	2.71 (8e-6)	1.55 (5e-6)
Thresholding	93.5 (3e-4)	**12.6** (3e-5)	**.596** (2e-6)	**.412** (2e-6)
prob	86.3 (6e-4)	22.4 (2e-4)	10.5 (6e-5)	5.13 (2e-5)
dens	**51.5** (2e-4)	21.7 (7e-5)	13.1 (3e-5)	13.2 (2e-5)
rank-2 state, MSEs$\times 10^3$				
Inversion	16.8 (8e-4)	15.9 (3e-4)	15.9 (1e-4)	15.8 (7e-5)
Thresholding	14.9 (3e-4)	15.5 (7e-5)	15.5 (9e-6)	15.5 (7e-6)
prob	**9.29** (2e-3)	**7.90** (1e-3)	**8.46** (1e-3)	**7.84** (8e-4)
dens	14.5 (3e-4)	14.6 (3e-4)	14.4 (3e-4)	14.5 (4e-4)
approximate rank-2 state, MSEs$\times 10^3$				
Inversion	15.9 (8e-4)	15.4 (2e-4)	15.3 (1e-4)	15.2 (4e-5)
Thresholding	14.3 (2e-4)	14.2 (3e-4)	15.0 (1e-5)	15.0 (6e-6)
prob	**8.88** (9e-4)	**7.68** (2e-3)	**8.11** (1e-3)	**7.39** (1e-3)
dens	13.9 (4e-4)	15.1 (2e-4)	14.2 (3e-4)	14.2 (2e-4)
maximal mixed state, MSEs$\times 10^4$				
Inversion	15.9 (4e-4)	6.57 (7e-5)	5.09 (5e-5)	4.76 (2e-5)
Thresholding	**4.67** (9e-5)	5.59 (5e-5)	5.34 (8e-5)	6.06 (8e-5)
prob	5.44 (2e-4)	**3.37** (8e-5)	**3.31** (8e-5)	**3.20** (8e-5)
dens	5.72 (9e-5)	4.47 (6e-5)	4.56 (4e-5)	4.24 (2e-5)

estimator and our ones. Note that the distribution of the eigenvalues of the three estimators are rather different. Still, it seems that all estimators return results compatible with a rank-2 state.

Table 3.2: MSEs for $n = 3$ (together with standard deviations)

	$m = 20$	$m = 200$	$m = 1000$	$m = 2000$
pure state, MSEs$\times 10^4$				
Inversion	39.5 (9e-4)	3.17 (9e-5)	.559 (1e-5)	.343 (1e-5)
Thresholding	21.4 (6e-4)	**2.26** (1e-4)	**.196** (1e-5)	**.152** (1e-5)
prob	40.3 (2e-2)	5.79 (4e-4)	2.95 (2e-4)	1.78 (1e-4)
dens	**12.8** (5e-4)	2.73 (2e-4)	1.24 (4e-5)	1.07 (4e-5)
rank-2 state, MSEs$\times 10^2$				
Inversion	3.69 (3e-3)	3.35 (6e-4)	3.32 (4e-4)	3.31 (2e-4)
Thresholding	2.94 (1e-3)	3.05 (2e-4)	3.04 (6e-5)	3.05 (5e-5)
prob	**1.91** (5e-3)	**1.17** (3e-3)	**1.18** (3e-3)	**1.14** (2e-3)
dens	2.83 (8e-4)	2.89 (3e-4)	2.89 (3e-4)	3.00 (1e-4)
approximate rank-2 state, MSEs$\times 10^2$				
Inversion	3.33 (2e-4)	3.22 (8e-4)	3.19 (3e-4)	3.18 (2e-4)
Thresholding	2.81 (1e-3)	2.96 (1e-4)	2.97 (8e-5)	2.97 (9e-5)
prob	**1.10** (5e-3)	**.551** (5e-3)	**.189** (2e-3)	**.113** (1e-3)
dens	2.74 (6e-4)	2.88 (3e-4)	2.91 (3e-4)	2.91 (2e-4)
maximal mixed state, MSEs$\times 10^3$				
Inversion	6.98 (2e-3)	3.19 (4e-4)	2.88 (2e-4)	3.01 (1e-4)
Thresholding	4.41 (6e-4)	3.26 (6e-4)	3.19 (2e-4)	3.29 (1e-4)
prob	3.63 (1e-3)	**2.70** (7e-4)	**2.28** (7e-4)	**2.29** (1e-3)
dens	**3.18** (6e-4)	2.99 (4e-4)	2.90 (2e-4)	3.04 (1e-4)

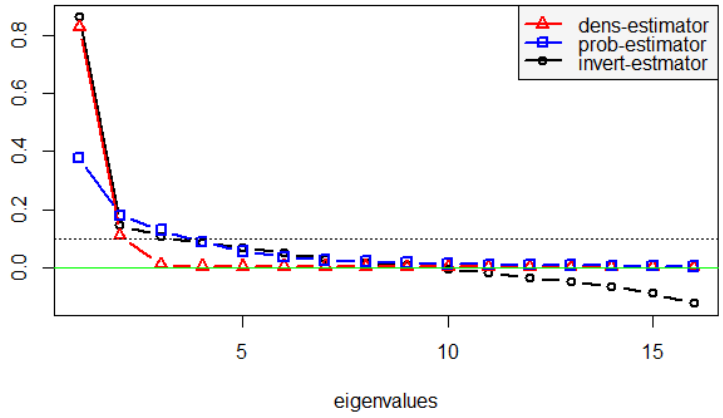

Figure 3.3: eigenvalues plots for real data test with $n = 4$

3.6 Discussion and conclusion

We propose a novel prior and introduce two pseudo-Bayesian estimators for the density matrix: the dens-estimator and the prob-estimator. The prob-estimator reaches the best up-to-date rate of convergence in the low-rank

Table 3.3: MSEs for $n = 2$ (together with standard deviations)

	$m = 20$	$m = 200$	$m = 1000$	$m = 2000$
pure state, MSEs$\times 10^4$				
Inversion	61.9 (3e-3)	9.22 (5e-4)	.802 (4e-5)	.772 (6e-5)
Thresholding	**49.4** (3e-3)	**4.06** (3e-4)	**.737** (4e-5)	**.356** (2e-5)
prob	102 (8e-3)	39.7 (2e-3)	9.37 (8e-4)	7.19 (5e-4)
dens	52.2 (3e-3)	7.57 (5e-4)	1.91 (9e-5)	1.08 (2e-5)
rank-2 state, MSEs$\times 10^2$				
Inversion	8.24 (2e-2)	7.91 (3.2e-3)	7.81 (2e-3)	7.74 (7e-4)
Thresholding	5.13 (3e-3)	5.34 (1.1e-3)	5.32 (5e-4)	5.33 (4e-4)
prob	**2.62** (2e-2)	**1.77** (7.4e-3)	**1.79** (8e-3)	**1.73** (5e-3)
dens	4.53 (3e-3)	5.20 (1.5e-3)	5.24 (9e-4)	5.24 (9e-4)
approximate rank-2 state, MSEs$\times 10^2$				
Inversion	8.12 (2e-2)	7.54 (4e-3)	7.54 (1.2e-3)	7.56 (6e-4)
Thresholding	4.95 (4e-3)	5.19 (8e-4)	5.23 (5e-4)	5.22 (4e-4)
prob	**2.69** (2e-2)	**1.82** (1.1e-2)	**1.52** (6e-3)	**1.58** (6e-3)
dens	4.40 (4e-3)	5.02 (1.3e-3)	5.11 (1e-3)	5.15 (6e-4)
maximal state, MSEs$\times 10^2$				
Inversion	3.03 (9e-3)	2.12 (2e-3)	2.11 (2e-3)	2.11 (1e-3)
Thresholding	2.78 (8e-3)	2.36 (2e-3)	2.21 (2e-3)	2.25 (1e-3)
prob	2.32 (2e-2)	**1.15** (5e-3)	**1.19** (5e-3)	**1.07** (4e-3)
dens	**2.30** (6e-3)	2.11 (2e-3)	2.06 (2e-3)	2.09 (1e-3)

case. On the other hand, computation of the dens-estimator is an easier task. In practice, we recommend the prob-estimator. However, in cases where the MCMC shows activities of lacking of convergence, the dens-estimator can be used as a reasonable alternative.

Note also that the prob-estimator can be extended to the incomplete measurement case. We consider the (incomplete) pseudo-likelihood as

$$\ell^{prob-incomplete}(\nu, \mathcal{D}) = \sum_{\mathbf{a}\in\mathcal{A}} \sum_{\mathbf{s}\in\mathcal{R}^n} \left[\mathrm{Tr}(\nu P_{\mathbf{s}}^{\mathbf{a}}) - \hat{p}_{\mathbf{a},\mathbf{s}}\right]^2,$$

where $\mathcal{A} \subsetneq \mathcal{E}^n$. The study in this case will be the object of future works.

Open questions include faster algorithms based on optimization (in the spirit of [Alquier et al., 2016]). Also, from a theoretical perspective, the most important question is the minimax lower bound.

3.7 Proofs

We first remind here a version of Hoeffding's inequality for bounded random variables.

Lemma 3.1. *Let $Y_i, i = 1, \ldots, n$ be n independent random variables with*

$|Y_i| \le b$ *a.s., and* $\mathbb{E}(Y_i) = 0$. *Then, for any* $\lambda > 0$,

$$\mathbb{E}\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n} Y_i\right) \le \exp\left(\frac{\lambda^2 b^2}{8n}\right).$$

3.7.1 Preliminary lemmas for the proof of Theorem 3.1

Lemma 3.2. *For any* $\lambda > 0$, *we have*

$$\mathbb{E}\exp\left(\lambda\left\langle p_\nu - p^0, p^0 - \hat{p}\right\rangle_F\right) \le \exp\left[\frac{\lambda^2}{4m}\|p^0 - p_\nu\|_F^2\right],$$

$$\mathbb{E}\exp\left(-\lambda\left\langle p_\nu - p^0, p^0 - \hat{p}\right\rangle_F\right) \le \exp\left[\frac{\lambda^2}{4m}\|p^0 - p_\nu\|_F^2\right].$$

Proof. First inequality:

$$\begin{aligned}
&\mathbb{E}\exp\left(\lambda\left\langle p_\nu - p^0, p^0 - \hat{p}\right\rangle_F\right) \\
&\quad= \mathbb{E}\exp\left(\lambda \sum_{a\in\mathcal{E}^n}\sum_{s\in\mathcal{R}^n} \underbrace{[\mathrm{Tr}(\nu P_s^a) - p_{a,s}^0]}_{=:c(a,s)}[p_{a,s}^0 - \hat{p}_{a,s}]\right) \\
&\quad= \prod_{a\in\mathcal{E}^n} \mathbb{E}\exp\left(\lambda \sum_{s\in\mathcal{R}^n} c(a,s)\left[p_{a,s}^0 - \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(R_i^a = s)\right]\right) \\
&\quad= \prod_{a\in\mathcal{E}^n} \mathbb{E}\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\underbrace{\left[\sum_{s\in\mathcal{R}^n} c(a,s)\{p_{a,s}^0 - \mathbf{1}(R_i^a = s)\}\right]}_{=:Y_{i,a}}\right)
\end{aligned}$$

We have that $\mathbb{E}(Y_{i,a}) = 0$. Then, using Cauchy-Schwartz inequality

$$\begin{aligned}
Y_{i,a}^2 &\le \left(\sum_{s\in\mathcal{R}^n} c(a,s)^2\right)\left(\sum_{s\in\mathcal{R}^n}[p_{a,s}^0 - \mathbf{1}(R_i^a = s)]^2\right) \\
&\le \left(\sum_{s\in\mathcal{R}^n} c(a,s)^2\right)\left(\sum_{s\in\mathcal{R}^n}|p_{a,s}^0 - \mathbf{1}(R_i^a = s)|\right) \le 2\left(\sum_{s\in\mathcal{R}^n} c(a,s)^2\right).
\end{aligned}$$

So we can apply Hoeffding's inequality (Lemma 3.1):

$$\begin{aligned}
\prod_{a\in\mathcal{E}^n}\mathbb{E}\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} Y_{i,a}\right) &\le \prod_{a\in\mathcal{E}^n}\exp\left[\frac{2\lambda^2}{8m}\left(\sum_{s\in\mathcal{R}^n} c(a,s)^2\right)\right] \\
&\le \exp\left[\frac{\lambda^2}{4m}\|p - p_\nu\|_F^2\right].
\end{aligned}$$

Second inequality: same proof, just replace $Y_{i,a}$ by $-Y_{i,a}$. □

Lemma 3.3. *For $\lambda > 0$, we have*

$$\mathbb{E}\exp\left\{\lambda\left(\|p_\nu - \hat{p}\|_F^2 - \|p^0 - \hat{p}\|_F^2\right) - \lambda\left[1 + \frac{\lambda}{m}\right]\|p^0 - p_\nu\|_F^2\right\} \leq 1, \quad (3.8)$$

$$\mathbb{E}\exp\left\{\lambda\left[1 - \frac{\lambda}{m}\right]\|p^0 - p_\nu\|_F^2 - \lambda\left(\|p_\nu - \hat{p}\|_F^2 - \|p^0 - \hat{p}\|_F^2\right)\right\} \leq 1. \quad (3.9)$$

Proof. Proof of the first inequality:

$$\begin{aligned}
&\mathbb{E}\exp\left\{\lambda\left(\|p_\nu - \hat{p}\|_F^2 - \|p - \hat{p}\|_F^2\right)\right\} \\
&\quad = \mathbb{E}\exp\left\{\lambda\left\langle p_\nu - p^0, p_\nu + p^0 - 2\hat{p}\right\rangle_F\right\} \\
&\quad = \mathbb{E}\exp\left\{\lambda\|p_\nu - p^0\|_F^2 + 2\lambda\left\langle p_\nu - p^0, p^0 - \hat{p}\right\rangle_F\right\} \\
&\quad = \exp\left(\lambda\|p_\nu - p^0\|_F^2\right)\mathbb{E}\exp\left\{2\lambda\left\langle p_\nu - p^0, p^0 - \hat{p}\right\rangle_F\right\} \\
&\quad \leq \exp\left(\lambda\|p_\nu - p^0\|_F^2\right)\exp\left\{\frac{\lambda^2}{m}\|p_\nu - p^0\|_F^2\right\}
\end{aligned}$$

thanks to Lemma 3.2. The proof of the second inequality is similar. □

Using Lemma 3.3, we derive an empirical PAC-Bayes bound for the estimator.

Lemma 3.4. *For $\lambda > 0$ s.t. $\frac{\lambda}{m} < 1$, with prob. $1 - \epsilon/2, \epsilon \in (0,1)$, for any distribution $\hat{\pi}$, we have:*

$$\int \|p_\nu - p^0\|_F^2 \tilde{\pi}_\lambda(\mathrm{d}\nu) \leq \frac{\int \|p_\nu - \hat{p}\|_F^2 \hat{\pi}(\mathrm{d}\nu) - \|p^0 - \hat{p}\|_F^2 + \frac{\mathcal{K}(\tilde{\pi}_\lambda, \pi) + \log\left(\frac{2}{\epsilon}\right)}{\lambda}}{1 - \frac{\lambda}{m}}.$$

Proof. We rewrite (3.9) in Lemma 3.3 as follows

$$\int \mathbb{E}\exp\left\{\lambda\left[1 - \frac{\lambda}{m}\right]\|p^0 - p_\nu\|_F^2 - \lambda\left(\|p_\nu - \hat{p}\|_F^2 - \|p^0 - \hat{p}\|_F^2\right)\right\}\pi(d\nu) \leq 1.$$

By using Fubini's theorem

$$\mathbb{E}\int \exp\left\{\lambda\left[1 - \frac{\lambda}{m}\right]\|p^0 - p_\nu\|_F^2 - \lambda\left(\|p_\nu - \hat{p}\|_F^2 - \|p^0 - \hat{p}\|_F^2\right)\right\}\pi(d\nu) \leq 1.$$

Now, using [Catoni, 2007, Lemma 1.1.3], for any distribution $\hat{\pi}$, we have

$$\begin{aligned}
\mathbb{E}\exp\sup_{\hat{\pi}}\Bigg\{&\lambda\left[1 - \frac{\lambda}{m}\right]\int\|p^0 - p_\nu\|_F^2\hat{\pi}(d\nu) - \log\left(2/\epsilon\right) - \mathcal{K}(\hat{\pi}, \pi) \\
&-\lambda\left(\int\|p_\nu - \hat{p}\|_F^2\hat{\pi}(d\nu) - \|p^0 - \hat{\rho}\|_F^2\right)\Bigg\} \leq \frac{\epsilon}{2}
\end{aligned}$$

and with $\mathbf{1}_{\mathbf{R}_+}(x) \leq \exp(x)$, one has

$$\mathbb{P}\Bigg\{\sup_{\hat{\pi}}\Bigg[\lambda\left[1 - \frac{\lambda}{m}\right]\int\|p^0 - p_\nu\|_F^2\hat{\pi}(d\nu) - \log\left(2/\epsilon\right) - \mathcal{K}(\hat{\pi}, \pi)$$

$$-\lambda\left(\int \|p_\nu - \hat{p}\|_F^2 \hat{\pi}(d\nu) - \|p^0 - \hat{\rho}\|_F^2\right)\Bigg] \geq 0\Bigg\} \leq \frac{\epsilon}{2}.$$

Taking the complementary yields successfully the results. □

The following lemma give a theoretical PAC-Bayes bound for the estimator.

Lemma 3.5. *For $\lambda > 0$ s.t $\frac{\lambda}{m} < 1$, with probability $1-\epsilon$ we have:*

$$\int \|p_\nu - p^0\|_F^2 \tilde{\pi}_\lambda^{prob}(\mathrm{d}\nu) \leq \inf_{\hat{\pi}} \frac{\left[1+\frac{\lambda}{m}\right]\int \|p_\nu - p^0\|_F^2 \hat{\pi}(\mathrm{d}\nu) + \frac{2\mathcal{K}(\hat{\pi},\pi)+2\log\left(\frac{2}{\epsilon}\right)}{\lambda}}{1-\frac{\lambda}{m}} \tag{3.10}$$

and

$$\int \|\nu - \rho^0\|_F^2 \hat{\pi}_\lambda^{prob}(\mathrm{d}\nu) \leq \inf_{\hat{\pi}} \frac{3^n\left[1+\frac{\lambda}{m}\right]\int \|\nu - \rho^0\|_F^2 \hat{\pi}(\mathrm{d}\nu) + \frac{2\mathcal{K}(\hat{\pi},\pi)+2\log\left(\frac{2}{\epsilon}\right)}{2^n\lambda}}{1-\frac{\lambda}{m}}. \tag{3.11}$$

Proof. Using the same proof of Lemma 3.4 for inequality (3.8) in Lemma 3.3, we obtain with probability at least $1-\epsilon/2, \epsilon \in (0,1)$, for any distribution $\hat{\pi}$ that

$$\int \|p^0 - \hat{p}\|_F^2 \hat{\pi}(d\nu) \leq \left[1+\frac{\lambda}{m}\right]\int \|p_\nu - p^0\|_F^2 \hat{\pi}(d\nu) + \|p^0 - \hat{p}\|_F^2 + \frac{\mathcal{K}(\hat{\pi},\pi)+\log(\frac{2}{\epsilon})}{\lambda}.$$

With a union argument, combining the Lemma 3.4 and the above inequality yields the following inequality with probability at least $1-\epsilon, \epsilon \in (0,1)$, for any $\hat{\pi}$

$$\int \|p_\nu - p^0\|_F^2 \hat{\pi}(\mathrm{d}\nu) \leq \frac{\left[1+\frac{\lambda}{m}\right]\int \|p_\nu - p^0\|_F^2 \hat{\pi}(\mathrm{d}\nu) + \frac{2\mathcal{K}(\hat{\pi},\pi)+2\log(2/\epsilon)}{\lambda}}{1-\frac{\lambda}{m}}.$$

Taking $\tilde{\pi}_\lambda^{prob}$ (once again, [Catoni, 2007, Lemma 1.1.3]) be the minimizer of the right hand side of the above inequality, we obtain (3.10).

Moreover, in [Alquier et al., 2013a, equation (5)] states that, for any ν:

$$p_\nu = \mathbf{P}\nu$$

for some operator $\mathbf{P}$. Therefore

$$\|p_\nu - p^0\|_F^2 = \|\mathbf{P}(\nu - \rho^0)\|_F^2.$$

The eigenvalues of $\mathbf{P}^T\mathbf{P}$ are known, they range between 2^n and $3^n 2^n$ according to [Alquier et al., 2013a, Proposition 1]. Thus, for any ν,

$$2^n\|\nu - \rho^0\|_F^2 \leq \|p_\nu - p^0\|_F^2 \leq 6^n\|\nu - \rho^0\|_F^2$$

and so we obtain (3.11). □

In the following, we will consider $\hat{\pi}$ as a restriction of the prior to a local set around the true density matrix ρ^0. This allows us to obtain an explicit bound of the left hand side of (3.11). Let $\rho^0 = U\Lambda U^\dagger$ be the spectral decomposition of ρ^0.

Definition 3.2. *Let $r = \#\{i : \Lambda_i > \delta\}$, with small $\delta \in [0, 1)$. Take*

$$\tilde{\pi}_c(du, dv) \propto \mathbf{1}(\forall i : |v_i - \Lambda_i| \leq \delta; \forall i = 1, \ldots, r : \|u_i - U_i\|_F \leq c)\pi(du, dv).$$

Note that we have $r \leq \text{rank}(\rho^0)$.

Lemma 3.6. *We have*

$$\int \|u^\dagger v u - \rho^0\|_F^2 \tilde{\pi}_c(du, dv) \leq (3d\delta + 2rc)^2. \tag{3.12}$$

And under the Assumption 3.1

$$\mathcal{K}(\tilde{\pi}_c, \pi) \leq ard \log(\frac{1}{c}) + C_{D_1,D_2} d(\log(d) + \log(\frac{1}{\delta})) \tag{3.13}$$

where a is a universal constant and where C_{D_1,D_2} depends only on D_1 and D_2.

Proof. Firstly

$$\|uvu^\dagger - \rho^0\|_F^2 \leq \left(\|uvu^\dagger - u\Lambda u^\dagger\|_F + \|u\Lambda u^\dagger - U\Lambda U^\dagger\|_F\right)^2$$

and

$$\begin{aligned}
\|uvu^\dagger - u\Lambda u^\dagger\|_F &\leq \sum_i |v_i - \Lambda_i| \|u_i u_i^\dagger\|_F \leq d\delta, \\
\|u\Lambda u^\dagger - U\Lambda U^\dagger\|_F &\leq \sum_i \Lambda_i \|u_i u_i^\dagger - U_i U_i^\dagger\|_F \\
&\leq \sum_{i:\Lambda_i > \delta} (\|u_i u_i^\dagger - u_i U_i^\dagger\|_F + \|u_i U_i^\dagger - U_i U_i^\dagger\|_F) \\
&\qquad + \delta \sum_{i:\Lambda_i \leq \delta} (\|u_i u_i^\dagger\|_F + \|U_i U_i^\dagger\|_F) \\
&\leq 2rc + 2\delta(d - r) \leq 2rc + 2\delta d,
\end{aligned}$$

so we obtain (3.12).

Now, the Kullback-Leibler term

$$\begin{aligned}
\mathcal{K}(\tilde{\pi}_c, \pi) &= \log \frac{1}{\pi(\{u, v : \forall i : |v_i - \Lambda_i| \leq c; \forall i = 1, r : \|u_i - U_i\|_F \leq \delta\})} \\
&= \log \frac{1}{\pi(\{\forall i : |v_i - \Lambda_i| \leq \delta\})} + \log \frac{1}{\pi\left(\{\forall i = 1, r : \|u_{i.} - U_{i.}\|_F \leq c\}\right)}.
\end{aligned}$$

The first log term

$$\pi\left(\{\forall i = 1, r : \|u_{i.} - U_{i.}\|_F \leq c\}\right) \geq \prod_{i=1}^{r} \left[\frac{\pi^{(d-1)/2}(c/2)^{d-1}}{\Gamma(\frac{d-1}{2} + 1)} \Big/ \frac{2\pi^{(d+1)/2}}{\Gamma(\frac{d+1}{2})}\right], d = 2^n$$

$$\geq \left[\frac{c^{d-1}}{2^d \pi}\right]^r \geq \frac{c^{r(d-1)}}{2^{4rd}}.$$

Note for the above calculation: it is greater or equal to the volume of the (d-1)-"circle" with radius $c/2$ over the surface area of the d-"unit-sphere".

The second log term in the Kullback-Leibler term

$$\pi(\{\forall i : |v_i - \Lambda_i| \leq \delta\}) = \frac{\Gamma(D_1)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d \int_{\max(\Lambda_i - \delta, 0)}^{\min(\Lambda_i + \delta, 1)} v_i^{\alpha_i - 1} dv_i$$
$$\geq \Gamma(D_1)\delta^d \prod_{i=1}^d \alpha_i \geq C_{D_1} \delta^d \; e^{-D_2 d \log(d)}$$

for some constant C_{D_1} that depends only on D_1. Since $\alpha_i \leq 1$ for every i, we can lower bound the integrand by 1 and also $\alpha_i \Gamma(\alpha_i) = \Gamma(\alpha_i + 1) \leq 1$. The interval of integration contains at least an interval of length δ. This trick was presented in [Ghosal et al., 2000b, Lemma 6.1, page 518]

Thus, we obtain

$$\mathcal{K}(\tilde{\pi}_c, \pi) \leq \log \frac{2^{4rd}}{c^{r(d-1)}} + \log\left(\frac{e^{D_2 d \log(d)}}{C_{D_1} \delta^d}\right)$$
$$\leq ard \log(\frac{1}{c}) + C_{D_1, D_2} d(\log(d) + \log(\frac{1}{\delta}))$$

for some absolute constant a and where C'_{D_1, D_2} depends only on D_1 and D_2. □

3.7.2 Proof of Theorem 3.1

Proof of Theorem 3.1. Substituting (3.13),(3.12) into (3.11), we obtain

$$\int \|\nu - \rho^0\|_F^2 \tilde{\pi}_\lambda(\mathrm{d}\nu) \leq \inf_c \left\{ \frac{3^n \left[1 + \frac{\lambda}{m}\right] (3d\delta + 2rc)^2}{1 - \frac{\lambda}{m}} \right.$$
$$\left. + \frac{ard \log(\frac{1}{c}) + C_{D_1, D_2} d(\log(d) + \log(\frac{1}{\delta})) + 2\log(2/\epsilon)}{\lambda 2^n [1 - \frac{\lambda}{m}]} \right\}.$$

By taking $\delta = \frac{1}{d\sqrt{N}}, c = \sqrt{\frac{d}{rm9^n}}, \lambda = m/2$ leads to

$$\int \|\nu - \rho^0\|_F^2 \tilde{\pi}_\lambda(\mathrm{d}\nu) \leq A\left(\frac{1}{m} + \frac{rd}{m3^n}\right) + C'_{D_1, D_2} \frac{r \log(rm3^n/d) + \log(m3^n) + \log(2/\epsilon)/2^n}{m}$$

for some absolute constant A. Finally, by Jensen inequality, one has

$$\|\hat{\rho}_\lambda - \rho^0\|_F^2 \leq \int \|\nu - \rho^0\|_F^2 \hat{\pi}_\lambda(\mathrm{d}\nu).$$

This completes the proof of the theorem. □

3.7.3 Preliminary results for the proof of Theorem 3.2

Rewriting equation (3.1), by plugging (3.3) in, as follow

$$p_{\mathbf{a},\mathbf{s}} = \sum_{b\in\{I,x,y,z\}^n} \rho_b \mathrm{Trace}\left(\sigma_b \cdot P_{\mathbf{s}}^{\mathbf{a}}\right) = \sum_{b\in\{I,x,y,z\}^n} \rho_b \mathbf{P}_{(s,a),b}.$$

Where $\mathbf{P}_{(s,a),b} = \prod_{j\neq E_b} s_j \mathbf{1}(a_j = b_j)$ and $E_b = \{j \in \{1,\dots,n\} : b_j = I\}$, see [Alquier et al., 2013a] for technical details. We are now ready to handle with the proofs.

Lemma 3.7. *For any $\lambda > 0$, we have*

$$\mathbb{E}\exp\left(\lambda\left\langle \rho^0 - \nu, \rho^0 - \hat{\rho}\right\rangle_F\right) \leq \exp\left[\frac{4\lambda^2}{m}\left(\frac{5}{3}\right)^n \|\nu - \rho^0\|_F^2\right],$$

$$\mathbb{E}\exp\left(-\lambda\left\langle \rho^0 - \nu, \rho^0 - \hat{\rho}\right\rangle_F\right) \leq \exp\left[\frac{4\lambda^2}{m}\left(\frac{5}{3}\right)^n \|\nu - \rho^0\|_F^2\right].$$

Proof. First inequality

$$\begin{aligned}
&\mathbb{E}\exp\left(\lambda\left\langle \rho^0 - \nu, \rho^0 - \hat{\rho}\right\rangle_F\right)\\
&= \mathbb{E}\exp\left[\lambda\sum_b (\rho_b^0 - \nu_b)(\rho_b^0 - \hat{\rho}_b) Trace(\sigma_b\sigma_b^\dagger)\right]\\
&= \mathbb{E}\exp\left[d\lambda\sum_b (\rho_b^0 - \nu_b)\sum_s\sum_a \frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}2^n}(p_{a,s}^0 - \hat{p}_{a,s})\right]\\
&= \prod_a \mathbb{E}\exp\left[\lambda\sum_b (\rho_b^0 - \nu_b)\sum_s \frac{1}{m}\sum_{i=1}^m \frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}}(p_{a,s}^0 - \mathbf{1}_{R_i^a=s})\right]\\
&= \prod_a\prod_i \mathbb{E}\exp\left[\underbrace{\frac{\lambda}{m}\sum_b (\rho_b^0 - \nu_b)\sum_s \frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}}(p_{a,s}^0 - \mathbf{1}_{R_i^a=s})}_{:=Y_{i,a}}\right].
\end{aligned}$$

Remark that $\mathbb{E}(Y_{i,a}) = 0$. Also, from the definitions above, the absolute value $|\mathbf{P}_{(s,a),b}|$ does not depend on s so

$$\begin{aligned}
|Y_{i,a}| &\leq \sum_b |\rho_b^0 - \nu_b| \left|\frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}}\right| \sum_s |p_{a,s}^0 - \mathbf{1}_{R_i^a=s}|\\
&\leq 2\sum_b |\rho_b^0 - \nu_b| \left|\frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}}\right| \leq \frac{2}{2^{n/2}}\sqrt{\sum_b (\rho_b^0 - \nu_b)^2 d \sum_b \left(\frac{\mathbf{P}_{(s,a),b}}{3^{d(b)}}\right)^2}\\
&\leq \frac{2\|\nu - \rho^0\|_F}{2^{n/2}}\left(\sum_b \frac{1}{3^{2d(b)}}\prod_{j\notin E_b}\mathbf{1}_{a_j=b_j}\right)^{1/2}\\
&\leq \frac{2\|\nu - \rho^0\|_F}{2^{n/2}}\left(\sum_{\ell=0}^n \binom{n}{\ell}\frac{1}{3^{2\ell}}\right)^{1/2}
\end{aligned}$$

$$\leq \frac{2\|\nu-\rho^0\|_F}{2^{n/2}}\left(1+\frac{1}{9}\right)^{n/2} = 2\|\nu-\rho^0\|_F\left(\frac{5}{9}\right)^{n/2}.$$

So we can apply Hoeffding's inquality (Lemma 3.1):

$$\prod_a \mathbb{E}\exp\left(\frac{\lambda}{m}\sum_{i=1}^m Y_{i,a}\right) \leq \exp\left[\frac{\lambda^2}{2m}\left(\frac{5}{3}\right)^n \|\nu-\rho^0\|_F^2\right].$$

Second inequality: same proof, just replace $Y_i(a)$ by $-Y_i(a)$. □

Lemma 3.8. *We have*

$$\mathbb{E}\exp\left\{\lambda\left[1-\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n\right]\|\nu-\rho^0\|_F^2 - \lambda\left(\|\nu-\hat{\rho}\|_F^2 - \|\rho^0-\hat{\rho}\|_F^2\right)\right\} \leq 1,$$
$$\mathbb{E}\exp\left\{\lambda\left(\|\nu-\hat{\rho}\|_F^2 - \|\rho^0-\hat{\rho}\|_F^2\right) - \lambda\left[1+\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n\right]\|\nu-\rho^0\|_F^2\right\} \leq 1.$$

Proof. For the second inequality:

$$\begin{aligned}
&\mathbb{E}\exp\left\{\lambda\left(\|\nu-\hat{\rho}\|_F^2 - \|\rho^0-\hat{\rho}\|_F^2\right)\right\}\\
&\quad= \mathbb{E}\exp\left\{\lambda\left\langle\nu-\rho^0, \nu+\rho^0-2\hat{\rho}\right\rangle_F\right\}\\
&\quad= \mathbb{E}\exp\left\{\lambda\|\nu-\rho^0\|_F^2 + 2\lambda\left\langle\nu-\rho^0, \rho^0-\hat{\rho}\right\rangle_F\right\}\\
&\quad= \exp\left(\lambda\|\nu-\rho^0\|_F^2\right)\mathbb{E}\exp\left\{2\lambda\left\langle\nu-\rho^0, \rho^0-\hat{\rho}\right\rangle_F\right\}\\
&\quad\leq \exp\left(\lambda\|\nu-\rho^0\|_F^2\right)\exp\left\{\frac{2\lambda^2}{m}\left(\frac{5}{3}\right)^n\|\nu-\rho^0\|_F^2\right\}
\end{aligned}$$

thanks to the Lemma 3.7. The proof of the first inequality is similar. □

Lemma 3.9. *For $\lambda > 0$ s.t $\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n < 1$, with probability at least $1-\epsilon, \epsilon \in (0,1)$, we have*

$$\int \|\nu-\rho^0\|_F^2 \tilde{\pi}_\lambda^{dens}(\mathrm{d}\nu) \leq \inf_{\hat{\pi}} \frac{\left[1+\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n\right]\int\|\nu-\rho^0\|_F^2\hat{\pi}(\mathrm{d}\nu) + \frac{2\mathcal{K}(\hat{\pi},\pi)+2\log(2/\epsilon)}{\lambda}}{1-\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n}. \tag{3.14}$$

Proof. By using the results from the Lemma 3.8, the proof is similar to the proof of Lemma 3.5 page 86. □

3.7.4 Proof of Theorem 3.2

Proof of Theorem 3.2. Substituting (3.13),(3.12) into (3.14)

$$\int \|\nu-\rho^0\|_F^2\hat{\pi}_\lambda(\mathrm{d}\nu) \leq \inf_c\left\{\frac{\left[1+\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n\right](3d\delta+2rc)^2}{1-\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n} + \frac{ard\log(\frac{1}{c}) + C_{D_1,D_2}d(\log(d)+\log(\frac{1}{\delta})) + 2\log(2/\epsilon)}{\lambda[1-\frac{2\lambda}{m}\left(\frac{5}{3}\right)^n]}\right\}.$$

Taking $\delta = \frac{d}{N}, c = \sqrt{\frac{d}{rN}}, \lambda = \frac{N}{5^n 4}$ lead to

$$\int \|\nu - \rho^0\|_F^2 \hat{\pi}_\lambda(\mathrm{d}\nu) \leq A' \frac{d^2 r}{N} + C_{D_1,D_2} 5^n \frac{rd \log(\frac{Nr}{d}) + d \log(\frac{N}{d}) + 2 \log(2/\epsilon)}{N}$$

for some constant $A' > 0$. Simultaneously, by Jensen inequality, one has

$$\|\hat{\rho}_\lambda - \rho^0\|_F^2 \leq \int \|\nu - \rho^0\|_F^2 \hat{\pi}_\lambda(\mathrm{d}\nu).$$

This complete the proof of the theorem. □

Chapter 4

LIFELONG LEARNING

In this chapter, we consider the problem of transfer learning in an online setting. More precisely, different tasks are presented sequentially and processed by a within-task algorithm. We propose a lifelong learning strategy which refines the underlying data representation used by the within-task algorithm, thereby transferring information from one task to the next. We show that when the within-task algorithm comes with some regret bound, our strategy inherits this good property. Our bounds are in expectation for a general loss function, and uniform for a convex loss. We discuss applications to dictionary learning and finite set of predictors. In the latter case, we improve previous $O(1/\sqrt{m})$ bounds to $O(1/m)$ where m is the per task sample size. We also show that it is possible to adapt lifelong learning strategy to learning-to-learn settings by using online-to-batch techniques.

A short version of this chapter is published in [Alquier et al., 2017b]:

Alquier,P., Mai, T.T., & Pontil,M. Regret bounds for lifelong learning. *In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, **2017**.

4.1 Motivation

Transferring knowledge gained from previously learned tasks is crucial for learning a new similar task, especially when the sample size is small. This is the essence of transfer learning approach, which can massively improve the performance over learning in isolation.

The main goal of this chapter is to show that it is possible to perform a theoretical analysis of lifelong learning with minimal assumptions on the form of the within-task algorithm. Given a learner with her/his own favourite algorithm(s) for learning within tasks, we propose a meta-algorithm for transferring information from one task to the next. The algorithm maintains a probability distribution on the set of representations, which is updated after the encounter of each new task using the exponentially weighted aggregation

(EWA) procedure, hence we call it *EWA for lifelong learning* or EWA-LL.

A standard way to provide theoretical guarantees for online algorithms is a regret bound, which measures the discrepancy between the prediction error of the forecaster and the error of an ideal predictor. We prove that, as long as the within-task algorithms have good statistical properties, EWA-LL inherits these properties. Specifically in Theorem 4.1 we present regret bounds for EWA-LL, in which the regret bounds for the within-tasks algorithms are combined into a regret bound for the meta-algorithm.

We also show, using an online-to-batch analysis, that it is possible to derive a strategy for learning-to-learn, and provide risk bounds for this strategy. The bounds are generally in the order of $1/\sqrt{T} + 1/\sqrt{m}$, where T is the number of tasks and m is the sample size per task. Moreover, we derive in some specific situations rates in $1/\sqrt{T} + 1/m$. These rates are novel up to our knowledge and justify the use of transfer learning with very small sample sizes m.

The chapter is organized as follows. In Section 4.2 we introduce the lifelong learning problem. In Section 4.3 we present the EWA-LL algorithm and provide a bound on its expected regret. Some popular algorithms for learning within-task are present in Section 4.4. In Section 4.5 we present more explicit versions of our bound in some classical examples: finite set of predictors, single index learning and dictionary learning. We also provide a short simulation study for dictionary learning. At this point, we hope that the reader will have a clear overview of the problem under study. The rest of the chapter is devoted to theoretical refinements: in online learning, uniform bounds are the norm rather than bounds in expectations [Cesa-Bianchi and Lugosi, 2006]. In Section 4.3.3 we establish such bounds for EWA-LL. Section 4.6 provides an online-to-batch analysis that allows one to use a modification of EWA-LL for learning-to-learn. We end the chapter with proofs (Section 4.9).

4.2 The Lifelong learning problem

In this section, we introduce our notation and present the lifelong learning problem.

4.2.1 Formulation

Let $\mathcal{X}$ and $\mathcal{Y}$ be some sets. A predictor is a function $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \mathbb{R}$ for regression and $\mathcal{Y} = \{-1, 1\}$ for binary classification. The loss of a predictor f on a pair (x, y) is a real number denoted by $\ell(f(x), y)$. As mentioned above, we want to transfer the information (a common data representation) gained from the previous tasks to a new one. Formally, we let $\mathcal{Z}$ be a set and prescribe a set $\mathcal{G}$ of feature maps (also called *representations*) $g : \mathcal{X} \to \mathcal{Z}$, and a set $\mathcal{H}$ of functions $h : \mathcal{Z} \to \mathbb{R}$. We shall design an algorithm that is useful when there is a function $g \in \mathcal{G}$, common to all the tasks, and task-specific functions

$h_1, \ldots, h_T$ such that

$$f_t = h_t \circ g$$

is a good predictor for task t, in the sense that the corresponding prediction error (see below) is small.

We are now ready to describe the learning problem. We assume that tasks are dealt with sequentially. Furthermore, we assume that each task dataset is itself revealed sequentially and refer to this setting as *online-within-online* lifelong learning. More specifically, at each time step $t \in \{1, \ldots, T\}$, the learner is challenged with a task, corresponding to a dataset

$$\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big) \in (\mathcal{X} \times \mathcal{Y})^{m_t}$$

where $m_t \in \mathbb{N}$. The dataset $\mathcal{S}_t$ is itself revealed sequentially, that is, at each inner step $i \in \{1, \ldots, m_t\}$:

- The object $x_{t,i}$ is revealed,
- The learner has to predict $y_{t,i}$, let $\hat{y}_{t,i}$ denote the prediction,
- Then $y_{t,i}$ is revealed and the learner incurs the loss $\hat{\ell}_{t,i} := \ell(\hat{y}_{t,i}, y_{t,i})$.

The task t ends at time m_t, at which point the prediction error is

$$\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}. \tag{4.1}$$

This process is repeated for each task t, so that at the end of all the tasks, the average error is

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}.$$

Ideally, if for a given representation g, the best predictor h_t for task t was known in advance, then an ideal learner using $h_t \circ g$ for prediction would incur the error

$$\inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big). \tag{4.2}$$

Hence, we define the within-task-regret of the representation g on task t as the difference between the prediction error (4.1) and the smallest prediction error (4.2),

$$\mathcal{R}_t(g) = \frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big).$$

The above expression is slightly different from the usual notion of regret [Cesa-Bianchi and Lugosi, 2006], which does not contain the factor $1/m_t$. This normalization is important in that it allows us to give equal weights to different tasks.

Note that an oracle who would have known the best common representation g for all tasks in advance would have only suffered, on the entire sequence of datasets, the error

$$\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell\big(h_t\circ g(x_{t,i}), y_{t,i}\big).$$

We are now ready to state our principal objective: we wish to design a procedure (meta-algorithm) that, at the beginning of each task t, produces a function $\hat{g}_t$ so that, within each task, the learner can use its own favorite online learning algorithm to solve task t on the sequence $\big((\hat{g}_t(x_{t,1}), y_{t,1}), \ldots, (\hat{g}_t(x_{t,m_t}), y_{t,m_t})\big)$. We wish to control the *compound regret* of our procedure

$$\mathcal{R} := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{m_t}\sum_{i=1}^{m_t} \hat{\ell}_{t,i} - \inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell\big(h_t\circ g(x_{t,i}), y_{t,i}\big)$$

which may succinctly be written as $\sup_{g\in\mathcal{G}} \big\{\frac{1}{T}\sum_{t=1}^{T} \mathcal{R}_t(g)\big\}$. This objective is accomplished in Section 4.3 under the assumption that a regret bound for the within-task-algorithm is available.

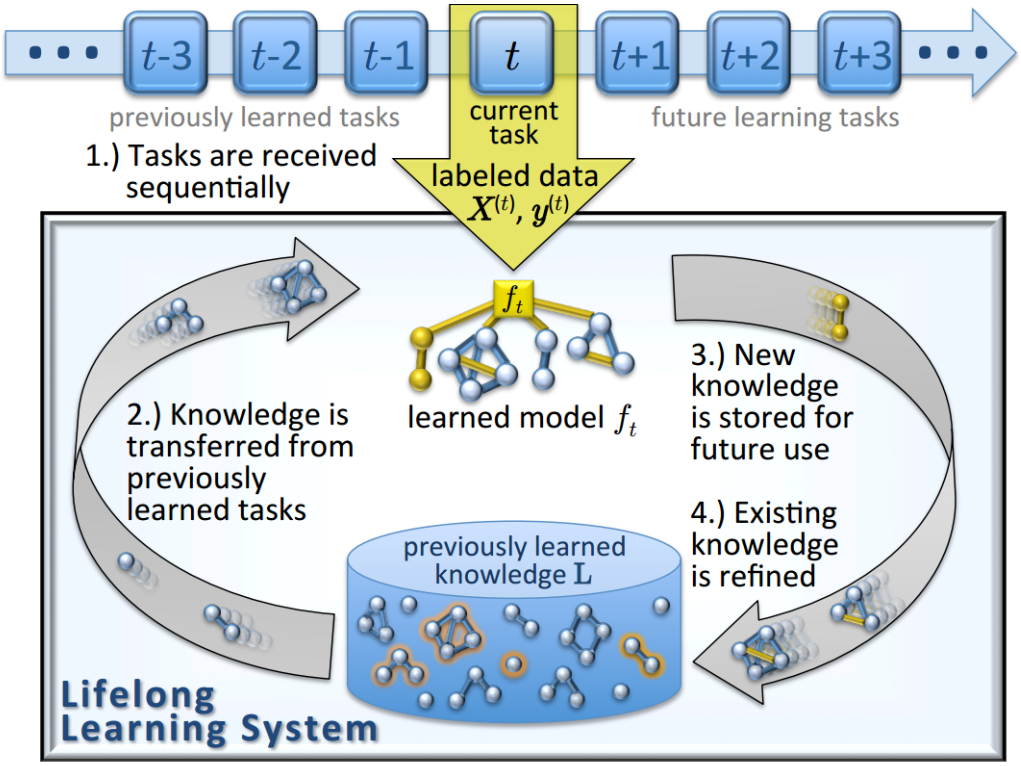

Figure 4.1: *An illustration of a lifelong learning process. ([Ruvolo and Eaton, 2013, Figure 1])*

4.2.2 Examples

We now provide some examples included in the framework.

Example 4.1 (Dictionary learning)**.** *Set $\mathcal{Z} = \mathbb{R}^K$, and call $g = (g_1, \ldots, g_K)$ a* dictionary*, where each g_k is a real-valued function on $\mathcal{X}$. Furthermore choose $\mathcal{H}$ to be a set of linear functions on $\mathbb{R}^K$, so that, for each task t*

$$h_t \circ g(x) = \sum_{k=1}^{K} \theta_k^{(t)} g_k(x).$$

In practice depending on the value of K, we can use least square estimators or LASSO to learn $\theta^{(t)}$. In [Maurer et al., 2013; Ruvolo and Eaton, 2013], the authors consider $\mathcal{X} = \mathbb{R}^d$ and

$$g(x) = Dx$$

for some $d \times K$ matrix D, and the goal is to learn jointly the predictors $\theta^{(t)}$ and the dictionary D.

Example 4.2 (Finite set $\mathcal{G}$)**.** *We choose $\mathcal{G} = \{g_1, \ldots, g_K\}$ and $\mathcal{H}$ any set. While this example is interesting in its own right, it is also instrumental in studying the continuous case via a suitable discretization process. A similar choice has been considered by [Crammer and Mansour, 2012] in the multitask setting, in which the goal is to bound the average error on a prescribed set of tasks.*

Example 4.3 (Single index learning)**.** *Set $\mathcal{X} = \mathcal{Z} = \mathbb{R}^d$, and $g(x) = \theta^T x, \theta \in \mathbb{R}^d$ a linear function on $\mathcal{X}$. Furthermore, let $\mathcal{H}$ be a set of univariate measurable functions on $\mathbb{R}$. We have, for each task t, the prediction is of the form $f_t(x_t) = h_t(\theta^T x_t)$. Our goal is to learn jointly the h_t and the common index θ.*

We notice that a slightly different learning setting is obtained when each dataset $\mathcal{S}_t$ is given all at once. We refer to this as **batch-within-online** lifelong learning; this setting is briefly considered in Section 4.7.

On the other hand when all datasets are revealed all at once, we are in the well-known setting of **learning-to-learn** [Baxter, 2000]. In Section 4.6, we explain how our lifelong learning analysis can be adapted to this setting.

4.3 A meta-algorithm for lifelong learning

In this section, we present our lifelong learning algorithm, derive its regret bound and then specify it to two popular within-task online algorithms.

4.3.1 EWA-LL Algorithm

Our EWA-LL algorithm is outlined in Algorithm 4. The algorithm is based on the exponentially weighted aggregation (EWA) procedure, see e.g. [Cesa-Bianchi and Lugosi, 2006] and references therein, and updates a probability

Algorithm 4 EWA-LL

Data A sequence of datasets $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big)$, $1 \leq t \leq T$, associated with different learning tasks; the points within each dataset are also given sequentially.

Input A prior π_1, a learning parameter $\eta > 0$ and a learning algorithm for each task t which, for any representation g returns a sequence of predictions $\hat{y}^g_{t,i}$ and suffers a loss

$$\hat{L}_t(g) := \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\left(\hat{y}^g_{t,i}, y_{t,i}\right).$$

Loop For $t = 1, \ldots, T$

i Draw $\hat{g}_t \sim \pi_t$.

ii Run the within-task learning algorithm on $\mathcal{S}_t$ and suffer loss $\hat{L}_t(\hat{g}_t)$.

iii Update

$$\pi_{t+1}(\mathrm{d}g) := \frac{\exp(-\eta \hat{L}_t(g))\pi_t(\mathrm{d}g)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

distribution π_t on the set of representation $\mathcal{G}$ before the encounter of task t. We insist on the fact that this procedure allows the user to freely choose the within-task algorithm, which does not even need to be the same for each task.

The step **i** is crucial during the learning procedure, because to draw $\hat{g}_t$ from π_t is not straightforward and varies in different specific situation. While the effect of Step **iii** is that any representation g which does not perform well on task t, is less likely to be reused on the next task.

4.3.2 Bounding the Expected Regret

Since Algorithm 4 involves a randomization strategy, we can only get a bound on the expected regret, the expectation being with respect to the drawing of the function $\hat{g}_t$ at step **i** in the algorithm. Let $\mathbb{E}_{g\sim\pi}[F(g)]$ denote the expectation of $F(g)$ when $g \sim \pi$. Note that the expected overall-average loss that we want to upper bound is then

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}[\hat{L}_t(\hat{g}_t)].$$

Theorem 4.1. *If, for any $g \in \mathcal{G}$, $\hat{L}_t(g) \in [0, C]$ and the within-task algorithm has a regret bound $\mathcal{R}_t(g) \leq \beta(g, m_t)$, then*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t} \left[\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i}\right] \leq \inf_{\rho} \left\{ \mathbb{E}_{g\sim\rho} \left[\frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big)\right.\right.$$

$$+ \frac{1}{T}\sum_{t=1}^{T}\beta(g, m_t)\Big] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T}\Big\},$$

where the infimum is taken over all probability measures ρ and $\mathcal{K}(\rho, \pi_1)$ is the Kullback-Leibler divergence between ρ and π_1.

The proof is given in Section 4.9.

Some comments are in order as the bound in Theorem 4.1 might not be easy to read. First, similar to standard analyses in online learning, the parameter η is a decreasing function of T, hence the bound vanishes as T grows. Second, corollaries are derived in Section 4.5 that are easier to read, as they are more similar to usual regret inequalities [Cesa-Bianchi and Lugosi, 2006], that is, the right hand side of the bound is of the form

$$\inf_{g\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) + \text{“rate”}. \tag{4.3}$$

The bound in Theorem 4.1 looks slightly different, but is quite similar in spirit. Indeed, instead of an infimum with respect to g we have an infimum on all the possible aggregations with respect to g,

$$\inf_{\rho} \mathbb{E}_{g\sim\rho} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) + \text{“remainder”}$$

where the remainder term depends on $\mathcal{K}(\rho, \pi_1)$. In order to look like (4.3), we can consider a measure ρ highly concentrated around the representation g minimizing (4.3). When $\mathcal{G}$ is finite, this is a reasonable strategy and the bound is given explicitly in Section 4.5.1 below. However, in some situations, this would cause the term $\mathcal{K}(\rho, \pi_1)$ to diverge. Studying accurately the minimizer in ρ usually leads to an interesting regret bound, and this is exactly what is done in Section 4.5.

Finally note that the bound in Theorem 4.1 is given in expectation. In online learning, uniform bounds are usually preferred [Cesa-Bianchi and Lugosi, 2006]. In Section 4.3.3 we show that it is possible to derive such bounds under additional assumptions.

4.3.3 Uniform bounds

In this section, we show that it possible to obtain a uniform bound, as opposed to a bound in expectation as in Theorem 4.1. From a theoretical perspective, the price to pay is very low: we only have to assume that the loss function is convex with respect to its first argument. However, in practice, there is an aggregation step that might not be feasible. This is discussed at the end of the section. The algorithm is outlined in Algorithm 5.

Algorithm 5 Integrated EWA-LL

Data and Input same as in Algorithm 4.

Loop For $t = 1, \dots, T$

i Run the within-task learning algorithm on $\mathcal{S}_t$ for each $g \in \mathcal{G}$ and return as predictions:

$$\hat{y}_{t,i} = \int \hat{y}_{t,i}^{g} \pi_t(\mathrm{d}g). \tag{4.4}$$

ii Update

$$\pi_{t+1}(\mathrm{d}g) := \frac{\exp(-\eta \hat{L}_t(g))\pi_t(\mathrm{d}g)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

Theorem 4.2. *Assuming that for any $g, 0 \leq \hat{L}_t(g) \leq C$ and that the algorithm used within-task has a regret $\mathcal{R}_t(g) \leq \beta(g, m_t)$. Assume that ℓ is convex with respect to its first argument. Then it holds that*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\left(\hat{y}_{t,i}, y_{t,i}\right) \leq \inf_{\rho}\left\{\mathbb{E}_{g\sim\rho}\left[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m_t}\sum_{i=1}^{m_t}\ell(h_t\circ g(x_{t,i}), y_{t,i}) + \frac{1}{T}\sum_{t=1}^{T}\beta(g, m_t)\right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T}\right\}.$$

Proof. At each step t, the loss suffered by the algorithm is

$$\frac{1}{m_t}\sum_{i=1}^{m_t}\ell\big(\hat{y}_{t,i}, y_{t,i}\big) = \frac{1}{m_t}\sum_{i=1}^{m_t}\ell\bigg(\int \hat{y}_{t,i}^{g}\pi_t(\mathrm{d}g), y_{t,i}\bigg) \leq \frac{1}{m_t}\sum_{i=1}^{m_t}\int \ell\big(\hat{y}_{t,i}^{g}, y_{t,i}\big)\pi_t(\mathrm{d}g) = \int \hat{L}_t(g)\pi_t(\mathrm{d}g)$$

and we can just apply Theorem 4.1. □

In practice, for an infinite set $\mathcal{G}$ we are not able to run simultaneously the within-task algorithm for all $g \in \mathcal{G}$. So, we cannot compute the prediction (4.4) exactly. A possible strategy is to draw N elements of $\mathcal{G}$ i.i.d. from π_t, say $\hat{g}_t(1), \dots, \hat{g}_t(N)$, and to replace (4.4) by

$$\hat{y}_{t,i}^{(N)} = \frac{1}{N}\sum_{j=1}^{N}\hat{y}_{t,i}^{\hat{g}_t(j)}.$$

Let's call MC-EWA this new version.

In order to analyze the performance of this algorithm, we can directly use Corollary 4.2. We only have to control the discrepancy between the

Algorithm 6 MC-EWA for lifelong learning with convex loss

Data and Input as in Algorithm 4.

Loop For $t = 1, \ldots, T$

i Draw independently from the past $\hat{g}_t(1), \ldots, \hat{g}_t(N)$ i.i.d from π_t.

ii Run the within-task learning algorithm $\mathcal{S}_t$ for each $\hat{g}_t(j)$ and return as predictions:

$$\hat{y}_{t,i}^{(N)} = \frac{1}{N} \sum_{j=1}^{N} \hat{y}_{t,i}^{\hat{g}_t(j)}.$$

iii Update

$$\pi_{t+1}(\mathrm{d}g) := \frac{\exp(-\eta \hat{L}_t(g))\pi_t(\mathrm{d}g)}{\int \exp(-\eta \hat{L}_t(\gamma))\pi_t(\mathrm{d}\gamma)}.$$

theoretical integral with respect to π_t and the corresponding empirical mean. Hoeffding's inequality leads to

$$\frac{1}{N} \sum_{j=1}^{N} \hat{L}_t(\hat{g}_t(j)) \leq \mathbb{E}_{g \sim \pi_t}[\hat{L}_t(g)] + C\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2N}}$$

with probability at least $1 - \delta$. A union bound over the T tasks leads to the following result directly.

Corollary 4.3. *Assuming that for any $g, 0 \leq \hat{L}_t(g) \leq C$ and that the algorithm used within-task has an average error $\mathcal{R}_t(g) \leq \beta(g, m_t)$. Assume that ℓ is convex with respect to its first argument. Then, with probability at least $1 - \delta$ over the drawing of all the $\hat{g}_t(j)$'s,*

$$\frac{1}{T} \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\left[\hat{y}_{t,i}^{(N)}, y_{t,i}\right] \leq \inf_{\rho} \left\{ \mathbb{E}_{g \sim \rho}\left[\frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell[h_t \circ g(x_{t,i}), y_{t,i}] + \frac{1}{T} \sum_{t=1}^{T} \beta(g, m_t) \right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T} \right\} + C\sqrt{\frac{\log\left(\frac{T}{\delta}\right)}{2N}}.$$

4.4 Examples of Within Task Algorithms

We now specify the general bound in Theorem 4.1 to two popular online algorithms that can be used within tasks.

4.4.1 Online Gradient Algorithm

Algorithm 7 OGA

Data A task $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big)$.

Input Step size $\zeta > 0$, and $\theta_1 = 0$.

Loop For $i = 1, \ldots, m_t$,

i Predict $\hat{y}^g_{t,i} = h_{\theta_i} \circ g(x_{t,i})$,

ii $y_{t,i}$ is revealed, update

$$\theta_{i+1} = \theta_i - \zeta \nabla_\theta \ell\big(h_\theta \circ g(x_{t,i}), y_{t,i}\big)\big|_{\theta=\theta_i}.$$

The first algorithm assumes that $\mathcal{H}$ is a parametric family of functions $\mathcal{H} = \{h_\theta, \theta \in \mathbb{R}^p, \|\theta\| \leq B\}$, and for any (x, y, g), $\theta \mapsto \ell(h_\theta \circ g(x), y)$ is convex, L-Lipschitz, upper bounded by C and denote by ∇_θ a subgradient.

Corollary 4.4. *The EWA-LL algorithm using the OGA within task with step size* $\zeta = \frac{B}{L\sqrt{2m_t}}$ *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m_t}\sum_{i=1}^{m_t} \hat{\ell}_{t,i}\right] \leq \inf_\rho \left\{ \mathbb{E}_{g\sim\rho}\left[\frac{1}{T}\sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t}\sum_{i=1}^{m_t} \ell(h_t \circ g(x_{t,i}), y_{t,i}) + \frac{BL}{T}\sum_{t=1}^{T}\frac{2}{\sqrt{m_t}}\right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T}\right\}.$$

Proof. Apply Theorem 4.1 and use the bound

$$\mathcal{R}_t(g) \leq \beta(g, m_t) := BL\frac{2}{\sqrt{m_t}}$$

that can be found, for example, in [Shalev-Shwartz, 2011, Corollary 2.7]. □

We note that under additional assumptions that the loss function is α-strongly convex, [Hazan, 2016, Theorem 3.3] provides better bounds for the OGA algorithm using an adaptive step size $\zeta = \frac{1}{i\alpha}$, that is

$$\beta(g, m_t) \leq \frac{L^2}{2\alpha m_t}(1 + \log m_t).$$

Comprehensive studies on online gradient algorithm and its variants can be found in e.g. [Shalev-Shwartz, 2011; Hazan, 2016].

4.4.2 Exponentially Weighted Aggregation

The second algorithm is based on the EWA procedure on the space $\mathcal{H} \circ g$ for a prescribed representation $g \in \mathcal{G}$.

Algorithm 8 EWA

Data A task $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \dots, (x_{t,m_t}, y_{t,m_t})\big)$.

Input Learning rate $\zeta > 0$; a prior probability distribution μ_1 on $\mathcal{H}$.

Loop For $i = 1, \dots, m_t$,

i Predict $\hat{y}^g_{t,i} = \int_{\mathcal{H}} h \circ g(x_{t,i}) \mu_i(\mathrm{d}h)$,

ii $y_{t,i}$ is revealed, update

$$\mu_{i+1}(\mathrm{d}h) = \frac{\exp(-\zeta \ell(h \circ g(x_{t,i}), y_{t,i})) \mu_i(\mathrm{d}h)}{\int \exp(-\zeta \ell(u \circ g(x_{t,i}), y_{t,i})) \mu_i(\mathrm{d}u)}.$$

Exp-concavity: Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is called ζ_0-exp-concave if $\exp(-\zeta_0 \varphi)$ is concave. A typical example is the quadratic loss function

$$\ell(y', y) = (y' - y)^2.$$

When there is some B such that $|y_{t,i}| \le B$ and $|h \circ g(x_{t,i})| \le B$, then the exp-concavity assumption is verified with $\zeta_0 = 1/(8B)$ and the boundedness assumption with $C = 4B^2$.

Corollary 4.5. *Assume that $\mathcal{H}$ is finite and that there exists $\zeta_0 > 0$ such that for any y, the function $\ell(\cdot, y)$ is ζ_0-exp-concave and upper bounded by a constant C. Then the EWA-LL algorithm using the EWA within task with $\zeta = \zeta_0$ satisfies*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t} \left[\frac{1}{m_t} \sum_{i=1}^{m_t} \hat{\ell}_{t,i} \right] \le \inf_{\rho} \Bigg\{ \mathbb{E}_{g \sim \rho} \Bigg[\frac{1}{T} \sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_t \circ g(x_{t,i}), y_{t,i}\big) + \frac{1}{T} \sum_{t=1}^{T} \frac{\zeta_0 \log |\mathcal{H}|}{m_t} \Bigg] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T} \Bigg\}.$$

Proof. Apply Theorem 4.1 and use the bound

$$\mathcal{R}_t(g) \le \beta(g, m_t) := \frac{\zeta_0 \log |\mathcal{H}|}{m_t}$$

that can be found, for example, in [Gerchinovitz, 2011, Theorem 2.2]. □

Note that when the exp-concavity assumption does not hold, [Gerchinovitz, 2011] derives a bound

$$\beta(g, m_t) = B \sqrt{\frac{\log(|\mathcal{H}|)}{2m_t}}$$

with the choice $\zeta = (2/B)\sqrt{2\log(|\mathcal{H}|)/m_t}$. Moreover, PAC-Bayesian type bounds in various settings (including infinite $\mathcal{H}$) can be found in [Catoni, 2004; Audibert, 2006; Gerchinovitz, 2013]. We refer the reader to [Gerchinovitz, 2011] for a comprehensive survey.

4.5 Applications: some specific models

In this section, to ease our presentation, we assume that all the tasks have the same sample size, that is $m_t = m$ for all t.

4.5.1 Finite subset of relevant predictors

We give details on Example 4.2, that is we assume that $\mathcal{G}$ is a set of K functions. Note that step **iii** in Algorithm 4 boils down to update K weights,

$$\pi_t(g_k) = \frac{\exp(-\eta \hat{L}_t(g_k))\pi_{t-1}(g_k)}{\sum_{j=1}^K \exp(-\eta \hat{L}_t(g_j))\pi_{t-1}(g_j)}.$$

Theorem 4.6. *Under the assumptions of Theorem 4.1, if we set* $\eta = \frac{2}{C}\sqrt{\frac{2\log K}{T}}$ *and* π_1 *uniform on* $\mathcal{G}$*,*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m}\sum_{i=1}^m \hat{\ell}_{t,i}\right] \leq \min_{1\leq k\leq K}\left\{\frac{1}{T}\sum_{t=1}^T \inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \ell(h_t \circ g_k(x_{t,i}), y_{t,i}) + \beta(g_k, m)\right\} + C\sqrt{\frac{\log K}{2T}}.$$

Proof. Fix $g \in \mathcal{G}$, ρ as the Dirac mass on g and note that $\mathcal{K}(\rho, \pi_1) = \log K$. □

We discussed in Sections 4.4.1 and 4.4.2 that typical orders for $\beta(g, m)$ are $\mathcal{O}(1/\sqrt{m})$, $\mathcal{O}(\log(m)/m)$ or $\mathcal{O}(1/m)$. We state a precise result in the finite case.

Corollary 4.7. *Assume that* $\mathcal{H}$ *is finite, that for some* $\zeta_0 > 0$*, for any* y*, the function* $\ell(\cdot, y)$ *is* ζ_0*-exp-concave and upper bounded by a constant* C*. Then the EWA-LL algorithm using the EWA within task with* $\zeta = \zeta_0$ *satisfies*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m}\sum_{i=1}^m \hat{\ell}_{t,i}\right] \leq \min_{1\leq k\leq K} \frac{1}{T}\sum_{t=1}^T \min_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \ell(h_t \circ g_k(x_{t,i}), y_{t,i}) + \frac{\zeta_0 \log|\mathcal{H}|}{m} + C\sqrt{\frac{\log K}{2T}}.$$

An improvement: In Section 4.6, we derive from Theorem 4.1 a bound in the batch setting. As we shall see, in the finite case the bound is exactly the same as the bound on the compound regret. This allows us to compare our results to previous ones obtained in the learning-to-learn setting. In particular, our $\mathcal{O}(1/m)$ bound improves upon [Pentina and Lampert, 2014] who derived an $\mathcal{O}(1/\sqrt{m})$ bound.

4.5.2 Lifelong single-index learning

We now give some detail on Example 4.3. Remind that the set $\mathcal{X} = \mathcal{Z} = \mathbb{R}^d$, and we define $\mathcal{G} = \{x \mapsto \theta^T x, : \theta \in \mathbb{R}^d\}$ linear functions on $\mathcal{X}$. Furthermore, let $\mathcal{H}$ be a set of L_2-Lipschitz univariate measurable functions on $\mathbb{R}$.

Recall that our predictor here is of the form $h_t(\theta^T x_{t,i})$. The goal is to learn the common weight vector θ for all tasks and the link function h_t for each task t.

We make the following assumptions on our model:

- $\|\theta\|_1 = 1$,
- the loss ℓ is L_1-Lipschitz, convex w.r.t its first component,
- $\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\|x_{t,i}\|_2 \leq M < +\infty$,
- π_1 is uniform on the unit ℓ_1-ball.

Assume that we have some within-task algorithms, that learn h_t at each time t. And

$$\beta(m) := \sup_{g\in\mathcal{G}} \beta(m, g) = \sup_{\|\theta\|_1=1} \beta(m, \theta) < +\infty,$$

$\beta(m)$ being an upper bound of the within-task algorithm that learns h_t. We will detail one possible such algorithm right after the statement of the theorem.

A simple result for lifelong single index learning is given in the following theorem.

Theorem 4.8. *Under the assumptions of Theorem 4.1, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i} - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}), y_{t,i}) \leq \frac{L_1 L_2 M}{\sqrt{T}} + \frac{Cd\log(T) + \log\left(\frac{2^{d-2}d!}{(d-1)^{d/2}}\right) + C}{4\sqrt{T}} + \beta(m).$$

The proof relies on an application of Theorem 4.1. The calculations being tedious, we postpone the proof to Section 4.9.

A within-task strategy to learn h_t: To learn h_t, we use EWA and consider a structure for $\mathcal{H}$. We consider the link function

$$h_t \in \mathcal{H}_{S,C_2+1} := \{h \in \mathcal{H} : h = \sum_{j=1}^{S} \beta_j \phi_j, \sum_{j=1}^{S} j|\beta_j| \leq C_2 + 1\},$$

where $\{\phi_j\}_{j=1}^{\infty}$ is a dictionary of measurable functions, each ϕ_j is assumed to be defined on $[-1,1]$ and to take values in $[-1,1]$.

Let

$$\mathcal{B}_S(C_2+1) = \{(\beta_1, \ldots, \beta_S) \in \mathbb{R}^S : \sum_{j=1}^{S} j|\beta_j| \leq C_2 + 1\}.$$

We define $\mu_1(dh)$ on the set $\mathcal{H}_{S,C_2+1}$ as the image of the uniform measure on $\mathcal{B}_S(C_2+1)$ induced by the map $(\beta_1, \ldots, \beta_S) \mapsto \sum_{j=1}^{S} \beta_j \phi_j$.

Corollary 4.9. *Under the assumptions of Theorem 4.1 and $\ell(x,\cdot) \in [0,C], \forall x$, we have*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i} - \inf_{\|\theta\|_1=1}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}_{S,C_2+1}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}), y_{t,i}) \\ \leq \frac{L_1}{\sqrt{m}} + \frac{C\sqrt{S}}{2\sqrt{2m}} + \frac{C\sqrt{S}\log[(C_2+1)\sqrt{m}]}{2\sqrt{2m}} \\ + \frac{L_1 L_2 M}{\sqrt{T}} + \frac{Cd\log(T) + \log\left(\frac{2^{d-2}d!}{(d-1)^{d/2}}\right) + C}{4\sqrt{T}}.\end{aligned}$$

As the proof of the corollary is not straightforward, we postpone the proof to Section 4.9.

4.5.3 Lifelong dictionary learning

In this section, to ease our presentation, we assume that all the tasks have the same sample size, that is $m_t = m$ for all t. We now give details on Example 4.1 in the linear case. Specifically, we let $\mathcal{X} = \mathbb{R}^d$, we let $\mathcal{D}_K$ be the set formed by all $d \times K$ matrices D, whose columns have euclidean norm equal to one, and we define $\mathcal{G} = \{x \mapsto Dx : D \in \mathcal{D}_K\}$.

Within this Section, we assume that the loss ℓ is convex and Φ-Lipschitz with respect to its first argument, that is, for every $y \in \mathcal{Y}$ and $a_1, a_2 \in \mathbb{R}$, it holds

$$|\ell(a_1, y) - \ell(a_2, y)| \leq \Phi|a_1 - a_2|.$$

We also assume that

$$\beta(m) := \sup_{g\in\mathcal{G}} \beta(m, g) < +\infty,$$

and

$$\|x_{t,i}\| \leq 1.$$

for all $t \in \{1, \ldots, T\}$ and $i \in \{1, \ldots, m\}$.

We define the prior π_1 as follows: the columns of D are i.i.d., uniformly distributed on the d-dimensional unit sphere. This is a natural choice for π_1 without any prior information.

(a) Regret bounds

Theorem 4.10. *Under the assumptions of Theorem 4.1, with* $\eta = \frac{2}{C}\sqrt{\frac{Kd}{T}}$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}_{t,i}\right] \leq \inf_{D \in \mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell\big(\langle h_t, Dx_{t,i}\rangle, y_{t,i}\big) + \frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7) + \beta(m) + \frac{B\Phi}{\sqrt{T}}\sqrt{\frac{1}{T}\sum_{t=1}^{T} \lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right)}.$$

The proof relies on an application of Theorem 4.1. The calculations being tedious, we postpone the proof to Section 4.9.

When we use OGA within tasks, we can use Corollary 4.4 with $L = \Phi\sqrt{K}$ and so $\beta(m) \leq \Phi B\sqrt{2K/m}$ for any $D \in \mathcal{D}_K$. Moreover,

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) \leq \operatorname{tr}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) \leq 1 \tag{4.5}$$

so Theorem 4.10 leads to the following corollary.

Corollary 4.11. *Using algorithm EWA-LL for dictionary learning, with* $\eta = \frac{2}{C}\sqrt{\frac{Kd}{T}}$, *and using the OGA algorithm within tasks, with step* $\zeta = B/(\Phi\sqrt{2mK})$, *yield*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}_{t,i}\right] \leq \inf_{D \in \mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell\big(\langle h_t, Dx_{t,i}\rangle, y_{t,i}\big) + \frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7) + \frac{B\Phi}{\sqrt{T}} + \frac{\Phi B\sqrt{2K}}{\sqrt{m}}.$$

Note that the upper bound (4.5) may be lose. For example, when the $x_{t,i}$ are i.i.d. on the unit sphere, $\lambda_{\max}\left(\sum_{i=1}^{m} x_{t,i}x_{t,i}^T/m\right)$ is close to $1/d$. In this case, it is possible to improve the term $\beta(m)$ employed in the calculation of the bound, we postpone the lengthy details to Subsection (c).

(b) Algorithmic Details and Simulations

We implement our meta-algorithm Randomized EWA in this setting. The algorithm used within each task is the simple version of the online gradient algorithm outlined in Section 4.4.1.

In order to draw $\hat{g}_t$ from π_t, we use N-steps of Metropolis-Hastings algorithm with a normalized Gaussian proposal [see, for example, Robert and Casella, 2013]. In order to ensure a short burn-in period, we use the previous drawing $\hat{g}_{t-1}$ as a starting point. The procedure is given in Algorithm 9.

Algorithm 9 EWA-LL for dictionary learning

Data As in Algorithm 4.

Input A learning rate η for EWA and a learning rate ζ for the online gradient. A number of steps N for the Metropolis-Hastings algorithm.

Start Draw $\hat{g}_1 \sim \pi_1$.

Loop For $t = 1, \ldots, T$

i Run the within-task learning algorithm $\mathcal{S}_t$ and suffer loss $\hat{L}_t(\hat{g}_t)$.

ii Set $\tilde{g} := \hat{g}_t$.

iii Metropolis-Hastings algorithm. Repeat N times

a Draw $\tilde{g}' \sim \mathcal{N}(\tilde{g}, \sigma^2 I)$ and then set $\tilde{g}' := \tilde{g}'/\|\tilde{g}'\|$.

b Set $\tilde{g} := \tilde{g}'$ with probability

$$\min\left\{1, \exp\left[\eta \sum_{h=1}^{t}\left(\hat{L}_h(\tilde{g}) - \hat{L}_h(\tilde{g}')\right)\right]\right\},$$

$\tilde{g}$ remains unchanged otherwise.

iv Set $\hat{g}_t := \tilde{g}$.

Note the bottleneck of the algorithm: in step **b** we have to compare $\tilde{g}$ and $\tilde{g}'$ on the whole dataset so far.

We now present a short simulation study. We generate data in the following way: we let $K = 2$, $d = 5$, $T = 150$ and $m = 100$. The columns of D are drawn uniformly on the unit sphere, and task regression vectors θ_t are also independent and have i.i.d. coordinates in $\mathcal{U}[-1, 1]$. We generate the datasets $\mathcal{S}_t$ as follows: all the $x_{t,i}$ are i.i.d. from the same distribution as θ_t, and $y_{t,i} = \langle \theta_t, Dx_{t,i} \rangle + \varepsilon_{t,i}$ where the $\varepsilon_{t,i}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\sigma = 0.1$.

We compare Algorithm 9 with $N = 10$ to an oracle who knows the representation D, but not the task regression vectors θ_t, and learns them using the online gradient algorithm with step size $\zeta = 0.1$. Notice that after each chunk of 100 observations, a new task starts, so the parameter θ_t changes. Thus, the oracle incurs a large loss until it learns the new θ_t (usually within a

few steps). This explains the "stair" shape of the cumulative loss of the oracle in Figure 4.2. Figure 4.3 indicates that after a few tasks, the dictionary D is learnt by EWA-LL: its cumulative loss becomes parallel to the one of the oracle.

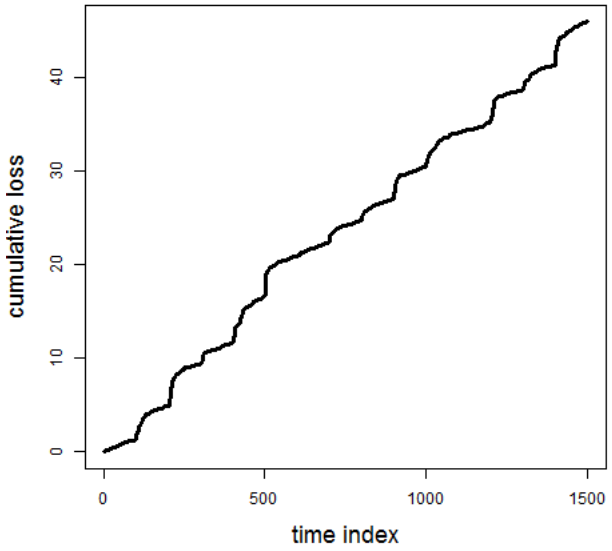

Figure 4.2: The cumulative loss of the oracle for the first 15 tasks.

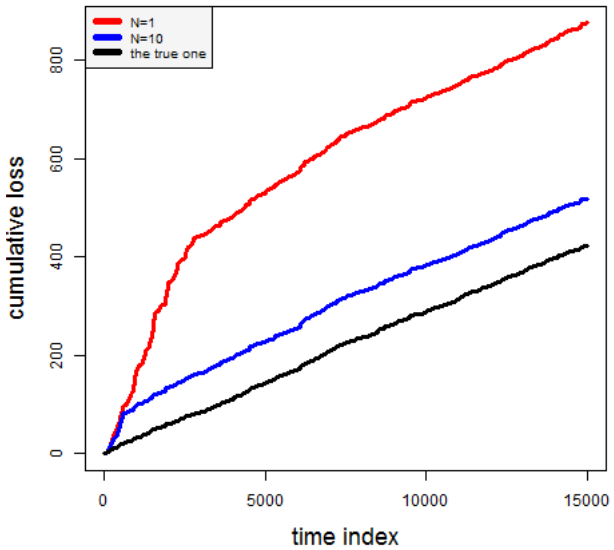

Figure 4.3: Cumulative loss of EWA-LL ($N = 1$ in red and $N = 10$ in blue) and cumulative loss of the oracle.

Due to the bottleneck mentioned above, the algorithm becomes quite slow to run when t grows. In order to improve the speed of the algorithm, we also tried Algorithm 9 with $N = 1$. There is absolutely no theoretical justification for this, however, obviously the algorithm is 10 times faster. As we can see on the red line in Figure 4.3, this version of the algorithm still learns D, but it takes more steps. Note that this is not completely unexpected: the Markov chain generated by this algorithm is no longer stationary, but it can still enjoy good mixing properties. It would be interesting to study the theoretical performance of Algorithm 9. However, this would require considerably technical tools from Markov chain theory which are beyond the scope of this work.

(c) Improved Regret bounds

We now state a refined version of the bounds for dictionary learning.

As pointed out in **(a)**, while in general the bound

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) \leq 1$$

is unimprovable.
However, if the input vectors $x_{t,i}$ are i.i.d. random variables from uniform distribution on the unit sphere, then

$$\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T \xrightarrow[m\to\infty]{a.s.} \mathrm{Cov}(x_{t,i}, x_{t,i}) = \frac{1}{d}I$$

where I is the identity matrix. Consequently,

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) \xrightarrow[m\to\infty]{a.s.} \frac{1}{d}.$$

We can take advantage of this fact in order to improve the term $\beta(m) = \sup_{g\in\mathcal{G}} \beta(g, m)$, but only if we assume that we know in advance that $\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right)$ is not too large. This is the meaning of the following theorem.

Theorem 4.12. *Assume that we know in advance that for all $t \in \{1, \ldots, T\}$,*

$$\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) \leq \Lambda$$

for some $\Lambda > 0$. Assume the same assumptions as in Theorem 4.10, still with $\eta = \frac{2}{C}\sqrt{\frac{Kd}{T}}$. Use within tasks Algorithm 7 (online gradient) with a fixed gradient step $\zeta = B/(L\sqrt{2mK\Lambda})$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{g_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right] - \inf_{g\in\mathcal{G}}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t, gx_{t,i}\rangle, y_{t,i}\big)$$

$$\leq \frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7) + \frac{2BL\sqrt{2K\Lambda}}{\sqrt{m}} + \frac{B\Phi\sqrt{\Lambda}}{\sqrt{T}}.$$

In particular, note that when $\Lambda = 1/d$ the bound becomes

$$\frac{C}{4}\sqrt{\frac{Kd}{T}}(\log(T)+7) + \frac{2BL\sqrt{2K}}{\sqrt{md}} + \frac{B\Phi}{\sqrt{dT}}.$$

Proof. We apply Theorem 4.10, so we only have to upper bound the term $\beta(g,m)$ for the online gradient algorithm with the prescribed step size. Note that in [Shalev-Shwartz, 2011, Corollary 2.7] we actually have the following regret bound for Algorithm 7 with fixed step size $\eta > 0$:

$$\beta(g,m) = \frac{B^2}{2\eta m} + \frac{\eta}{m}\sum_{i=1}^{m} \|\nabla_{\theta=\theta_t}\ell(\langle\theta, gx_{t,i}\rangle, y_{t,i})\|^2.$$

By the L-Lipschitz assumption on ℓ,

$$\|\nabla_{\theta=\theta_t}\ell(\langle\theta_t, gx_{t,i}\rangle, y_{t,i})\|^2 \leq L^2\|gx_{t,i}\|^2.$$

So we have

$$\begin{aligned}
&\sum_{t=1}^{m} \|\nabla_{\theta=\theta_t}\ell(\langle\theta, gx_{t,i}\rangle, y_{t,i})\|^2 \\
&\leq L^2\sum_{i=1}^{m}\|gx_{t,i}\|^2 = L^2\sum_{i=1}^{m}\sum_{k=1}^{K}\langle g_{k,\cdot}, x_{t,i}\rangle^2 = L^2\sum_{i=1}^{m}\sum_{k=1}^{K} g_{k,\cdot}^T x_{t,i}x_{t,i}^T g_{k\cdot} \\
&\leq mL^2\sum_{k=1}^{K} g_{k,\cdot}^T\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right) g_{k\cdot} \leq mKL^2\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} x_{t,i}x_{t,i}^T\right)\|g_{k\cdot}\|^2 \\
&\leq mKL^2\Lambda.
\end{aligned}$$

Consequently,

$$\beta(m) = \sup_g \beta(g,m) \leq B^2/(2\eta m) + \eta KL^2\Lambda$$

and the choice $\eta \leq B/(L\sqrt{2mK\Lambda})$ leads to

$$\beta(m) = 2BL\sqrt{2K\Lambda/m}.$$

□

4.6 From Lifelong learning to Learning-to-learn

In this section, we show how our analysis of lifelong learning can be used to derive bounds for learning-to-learn settings. In another words, we address conversions from a online setting to a batch setting. This online-to-batch trick is pedagogically discussed in [Shalev-Shwartz, 2011, Section 5].

In this section, we assume actually the following mechanism generates the tasks and their datasets

1. task distributions $P_1, \dots, P_T$ are sampling i.i.d. from a "meta-distribution" Q, called *environment* by [Baxter, 2000],
2. then for each task t, a dataset $\mathcal{S}_t = ((x_{t,1}, y_{t,1}), \dots, (x_{t,m}, y_{t,m}))$ is sampled i.i.d. from P_t.

We stress that in this setting, the entire data $(x_{t,i}, y_{t,i})_{1 \le i \le m, 1 \le t \le T}$ is given all at once to the learner. Note that for simplicity, we assumed that all the sample sizes are the same, although it is uncomplicated to extend the setting to a random m_t drawn at each step.

We wish to design a strategy which, given a new task $P \sim Q$ and a new sample $(x_1, y_1), \dots, (x_m, y_m)$ i.i.d. from P, computes a function $f : \mathcal{X} \to \mathcal{Y}$, that will predict y well when $(x, y) \sim P$. There are many ways to produce a such procedure and here we consider two simple approaches: randomization and averaging schemes.

4.6.1 Randomization scheme

First we propose the following strategy:

1. Run EWA-LL on $(x_{t,i}, y_{t,i})_{1 \le i \le m, 1 \le t \le T}$. We obtain a sequence of representations $\hat{g}_1, \dots, \hat{g}_T$,
2. Draw uniformly $\mathcal{T} \in \{1, \dots, T\}$ and put $\hat{g} = \hat{g}_{\mathcal{T}}$,
3. Run the within task algorithm on the sample $(x_i, y_i)_{1 \le i \le m}$, obtaining a sequence $h_1^{\hat{g}}, \dots, h_m^{\hat{g}}$ of functions,
4. Draw uniformly $\mathcal{I} \in \{1, \dots, m\}$ and put $\hat{h} = h_{\mathcal{I}}^{\hat{g}}$.

Our next result guarantees that the above strategy leads indeed to safe predictions. The proof is given in Section 4.9.

Theorem 4.13. *Let $\mathbb{E}$ be the expectation over all data pairs $(x_{t,i}, y_{t,i})_{1 \le i \le m} \sim P_t$, $(P_t)_{1 \le t \le T} \sim Q$, $(x_i, y_i)_{1 \le i \le m} \sim P$, $(x, y) \sim P$, $P \sim Q$ and also over the randomized decisions of the learner $(\hat{g}_t)_{1 \le t \le T}$, $\mathcal{T}$ and $\mathcal{I}$. Then*

$$\mathbb{E}[\ell(\hat{h} \circ \hat{g}(x), y)] \le \inf_{\rho} \left\{ \mathbb{E}_{g \sim \rho} \left[\mathbb{E}_{P \sim Q} \inf_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim P} \Big[\ell(h \circ g(x), y) \Big] + \beta(g, m) \right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta T} \right\}.$$

As in Theorem 4.1, the above result is given in expectation with respect to the randomized decisions of the learner. One might worry about this expectation $\mathbb{E}_{\mathcal{T}} \mathbb{E}_{\mathcal{I}}$. However, assuming that ℓ is convex with respect to its first argument, it is possible to use aggregation to overcome this. We now state a similar result for a non-random procedure, as was done in Section 4.3.3.

4.6.2 Averaging scheme

Assuming that ℓ is convex with respect to its first argument, the second strategy is as follows:

1. run aggregated EWA for lifelong learning on the sample $(x_{t,i}, y_{t,i})_{i,t}$. We obtain a sequence of probability distributions $\pi_1, \ldots, \pi_T$,
2. run, for all $g \in \mathcal{G}$, the within-task algorithm on the sample $(x_1, y_1), \ldots, (x_m, y_m)$, this produces a sequence $h_1^g, \ldots, h_m^g$ of functions,
3. return as a predictor the function $\widehat{\text{goh}}(\cdot)$ defined by

$$\widehat{\text{goh}}(x) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{T}\sum_{t=1}^{T}\int h_i^g \circ g(x)\pi_t(\mathrm{d}g).$$

The following corollary is a direct application of Theorem 4.13.

Corollary 4.14. *Let E denote*

$$E = \mathbb{E}_{P_1,\ldots,P_T\sim Q}\mathbb{E}_{(x_{t,i},y_{t,i})\ i.i.d\ P_t}\mathbb{E}_{P\sim Q}\mathbb{E}_{(x_1,y_1),\ldots,(x_m,y_m)\ i.i.d\ P}\mathbb{E}_{(x,y)\sim P}.$$

Then

$$E[\ell(\widehat{\text{goh}}(x), y)] \leq \inf_{\rho}\left\{\mathbb{E}_{g\sim\rho}\left[\mathbb{E}_{P\sim Q}\inf_{h\in\mathcal{H}}\mathbb{E}_{(x,y)\sim P}\Big[\ell(h\circ g(x), y)\Big] + \beta(g, m)\right] + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta}\right\}.$$

Remark 4.1. *In [Baxter, 2000; Maurer et al., 2013; Pentina and Lampert, 2014], the results on learning-to-learn are given with large probability with respect to $(x_{t,i}, y_{t,i})_{1\leq i\leq m, 1\leq t\leq T}$, rather than in expectation. Using the machinery in [Cesa-Bianchi and Lugosi, 2006, Lemma 4.1] we conjecture that it is possible to derive a bound in probability from Theorem 4.13.*

4.7 Batch-Within-Online Lifelong Learning

In this section, we present an alternative approach for the batch-within-online setting discussed in Section 4.2. In this setting, the tasks are presented sequentially, but, for each task $t \in \{1, \ldots, T\}$ the dataset $\mathcal{S}_t$ is presented all at once and we assume it is obtained i.i.d. from a distribution P_t. Unlike to the reasoning in Section 4.6, where we assumed that the P_t were i.i.d. from a distribution Q, here we make no assumptions on the generation process underlying the P_t's, which may even be adversarial chosen.

Let us recap the setting. At each time $t \in \{1, \ldots, T\}$, a task is presented to the learner in the following manner:

1. nature chooses P_t , no assumption is made on this choice. This P_t is not revealed to the forecaster.

2. nature draws the sample $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big)$ i.i.d. from P_t, and this sample is revealed to the forecaster.

3. based on her/his current guess $\tilde{g}_t$ of g and on the sample $\mathcal{S}_t$, the forecaster has to run her/his favourite learning algorithm $\hat{h}$ on $(\tilde{g}_t, \mathcal{S}_t)$ to get an estimate $\tilde{h}_t = \hat{h}(\tilde{g}_t, \mathcal{S}_t)$ based on an algorithm of his choice. Note that the forecaster observes $\tilde{r}_t := r_t(\tilde{h}_t \circ \tilde{g}_t)$ where

$$r_t(f) = \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(f(x_{t,i}), y_{t,i}\big).$$

4. the forecaster incur the loss $R_t(\tilde{h}_t \circ \tilde{g}_t)$ where

$$R_t(f) = \mathbb{E}_{(x,y)\sim P_t}\big[\ell\big(f(x), y\big)\big].$$

   Unfortunately, this quantity is not known to the forecaster.

At the end of time, we are interested in a strategy such that the compound regret

$$\mathcal{R} := \frac{1}{T} \sum_{t=1}^{T} R_t(\tilde{h}_t \circ \tilde{g}_t) - \inf_{g\in\mathcal{G}} \frac{1}{T} \sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} R_t(h_t \circ g)$$

is controlled.

The situation is similar to the setting discussed in the core of Chapter 4: we will propose an EWA algorithm for transfer learning, called EWA-TL, for which the regret will be controlled, on the condition that the learner chooses a suitable within task algorithm. In the online case, the within tasks algorithm was either EWA or OGA. In Subsection 4.7.1 we discuss briefly the within task algorithm. In Subsection 4.7.2 we present the EWA-TL algorithm and its theoretical analysis.

4.7.1 Within-task Algorithms

We make an additional assumption, that is that the estimator $\hat{h}$ satisfies a bound in probability:

$$\mathbb{P}\Bigg[\forall g \in \mathcal{G}, |r(\hat{h}(g, \mathcal{S}_t) \circ g) - R_t(\hat{h}(g, \mathcal{S}_t) \circ g)| \le \delta(g, m_t, \varepsilon)$$

and

$$|R_t(\hat{h}(g, \mathcal{S}_t) \circ g) - \inf_{h\in\mathcal{H}} R_t(h \circ g)| \le 2\delta(g, m_t, \varepsilon)\Bigg] \ge 1 - \varepsilon. \quad (4.6)$$

Example 4.4 (Empirical Risk Minimizer)**.** *In classification, when ℓ is the 0-1 loss function, and for any g, the family $\{h \circ g, h \in \mathcal{H}\}$ has a Vapnik-Chervonenkis dimension bounded by V. Then the empirical risk minimizer (ERM)*

$$\hat{h}(g, \mathcal{S}_t) = \arg\min_{h \in \mathcal{H}} r_t(h \circ g)$$

satisfies the above condition with

$$\delta(g, m_t, \varepsilon) = 2\sqrt{2\frac{V \log\left(\frac{2m_t \mathrm{e}}{V}\right) + \log\left(\frac{4}{\varepsilon}\right)}{m_t}},$$

see e.g. [Chapter 4, page 94 Vapnik, 1998].

Similar rates can be obtained with PAC-Bayesian bounds [McAllester, 1998; Catoni, 2004].

Example 4.5 (PAC-Bayesian estimator)**.** *Assuming that the loss ℓ takes the values in $[0, C]$ for a constant $C > 1$. Given a prior distribution $\pi_g(dh)$ on the set $\{h \circ g, h \in \mathcal{H}\}$, then the Gibbs estimator*

$$\hat{h}(g, S_t) = \int h \hat{\rho}_\lambda(dh)$$

$$\hat{\rho}_\lambda(dh) \propto \exp(-\lambda r_t(h \circ g))\pi_g(dh)$$

satisfies the condition (4.6) above with

$$\delta(g, m_t, \varepsilon) = \frac{\mathcal{K}(\rho, \pi) + \log\frac{1}{\varepsilon}}{\lambda} + \frac{\lambda C^2}{2m_t}, \forall \rho \in \mathcal{M}_+^1(\Theta).$$

4.7.2 EWA-TL

Algorithm 10 EWA-TL

Data A sequence of datasets $\mathcal{S}_t = \big((x_{t,1}, y_{t,1}), \ldots, (x_{t,m_t}, y_{t,m_t})\big)$, $1 \leq t \leq T$, associated with different learning tasks; the datasets are revealed sequentially, but the points within each dataset $\mathcal{S}_t$ are revealed all at once.

Input A prior π_1, a learning parameter $\eta > 0$ and a learning algorithm $\hat{h}$ which satisfies (4.6).

Loop For $t = 1, \ldots, T$

i Draw $\hat{g}_t \sim \pi_t$.

ii Run the within-task learning algorithm $\hat{h}$ on $\mathcal{S}_t$ to get $\tilde{h}_t = \hat{h}(\hat{g}_t, \mathcal{S}_t)$.

iii Update

$$\pi_{t+1}(\mathrm{d}g) \propto \exp\left\{-\eta\Big[r_t(\hat{h}(\mathcal{S}_t, g) \circ g) + \delta(g, m_t, \varepsilon/T)\Big]\right\}\pi_{t-1}(\mathrm{d}g).$$

We now provide a bound on the regret of EWA-TL.

Theorem 4.15. *Under (4.6), and assuming that there is a constant C such that $0 \leq r_t(\hat{h}(\mathcal{S}_t, g) \circ g) + \delta(g, m_t, \varepsilon/T) \leq C$, with probability at least $1-\varepsilon$,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}}\Big[R_t(\tilde{h}_t \circ \tilde{g}_t)]\Big] \leq \inf_{\rho}\Bigg\{\mathbb{E}_{g\sim\rho}\Bigg[\frac{1}{T}\sum_{t=1}^{T}\inf_{h\in\mathcal{H}} R_t(h\circ g) + \frac{4}{T}\sum_{t=1}^{T}\delta(g, m_t, \varepsilon/T)\Bigg] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\Bigg\}.$$

Sketch of the proof. First, follow the proof of Theorem 4.1 to get:

$$\sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}}\Big[r_t(\tilde{h}_t \circ \tilde{g}_t)] + \delta(\tilde{g}_t, m_t, \varepsilon/T)\Big] \leq \inf_{\rho}\Bigg\{\sum_{t=1}^{T}\mathbb{E}_{g\sim\rho}\Big[r_t(\tilde{h}_t\circ g) + \delta(g, m_t, \varepsilon/T)\Big] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho,\pi)}{\eta}\Bigg\}.$$

So, with probability at least $1-\varepsilon$,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}}\Big[R_t(\tilde{h}_t \circ \tilde{g}_t)]\Big] \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{\tilde{g}_t \sim \pi_{t-1}}\Big[r_t(\tilde{h}_t \circ \tilde{g}_t)] + \delta(\tilde{g}_t, m_t, \varepsilon/T)\Big] \\
&\leq \inf_{\rho}\Bigg\{\sum_{t=1}^{T}\mathbb{E}_{g\sim\rho}\Big[r_t(\tilde{h}_t\circ g) + \delta(g, m_t, \varepsilon/T)\Big] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta}\Bigg\} \\
&\leq \inf_{\rho}\Bigg\{\sum_{t=1}^{T}\mathbb{E}_{g\sim\rho}\Big[R_t(\hat{h}_t(g,\mathcal{S}_t)\circ g) + 2\delta(g, m_t, \varepsilon/T)\Big] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta}\Bigg\} \\
&\leq \inf_{\rho}\Bigg\{\mathbb{E}_{g\sim\rho}\Bigg[\sum_{t=1}^{T}\inf_{h\in\mathcal{H}} R_t(h\circ g) + 4\sum_{t=1}^{T}\delta(g, m_t, \varepsilon/T)\Bigg] + \frac{\eta T C^2}{8} + \frac{\mathcal{K}(\rho,\pi_1)}{\eta}\Bigg\}.
\end{aligned}$$

□

4.8 Concluding Remarks

We presented a meta-algorithm for lifelong learning and derived for the first time a fully online analysis of its regret. An important advantage of this algorithm is that it inherits the good properties of any algorithm used to learn within tasks. Furthermore, using online-to-batch conversion techniques, we derived bounds for the related framework of learning-to-learn.

We discussed the implications of our general regret bounds for two applications: dictionary learning and finite set $\mathcal{G}$ of representations. Further

applications of this algorithm which may be studied within our framework are deep neural networks and kernel learning.

Example 4.6 (Kernel learning)**.** *In this case, $\mathcal{Z}$ is an Hilbert space. The function $g : \mathcal{X} \to \mathcal{Z}$ is a feature map to a reproducing kernel Hilbert space $\mathcal{Z}$, and $h_t(g(x)) = \langle z^{(t)}, g(x) \rangle_{\mathcal{Z}}$. Note that if $z^{(t)} = \sum_{k=1}^{K^{(t)}} \alpha_i^{(t)} g(\xi_k^{(t)})$ then*

$$h_t(g(x)) = \sum_{k=1}^{K^{(t)}} \alpha_i^{(t)} K_g(\xi_k^{(t)}, x)$$

where $K_g(x, x') = \langle g(x), g(x') \rangle_{\mathcal{Z}}$ is the kernel induced by g. E.g. for each task we use SVM, and we transfer kernel learning from one task to another. This application has been addressed by [Pentina and Ben-David, 2015] in the learning-to-learn setting.

Example 4.7 (Deep network)**.** *Here $\mathcal{X} = \mathbb{R}^d$ and $g : \mathcal{X} \to \mathbb{R}^K$ is a multilayer network, that is a vector-valued function obtained by application of a linear transformation and a nonlinear activation functions. Specifically $g(x) = \sigma(W_q(\cdots \sigma(W_2(\sigma(W_1 x))) \cdots))$. The predictor $h : \mathbb{R}^K \to \mathbb{R}$ is typically a linear function. The vector-valued function $(h_1 \circ g, \ldots, h_T \circ g))$ models a multilayer network with shared hidden weights. This is discussed in [Maurer et al., 2016], again in the learning-to-learn setting.*

Perhaps the most fundamental question is to extend our analysis to more computationally efficient algorithms such as approximations of EWA, like Algorithm 9, or fully gradient based algorithms as in [Ruvolo and Eaton, 2013].

4.9 Proofs

4.9.1 Proof of Theorem 4.1

Proof of Theorem 4.1. It is enough to show that the EWA strategy leads to

$$\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}[\hat{L}_t(\hat{g}_t)] \leq \inf_{\rho} \left\{ \mathbb{E}_{g \sim \rho} \left[\sum_{t=1}^{T} \hat{L}_t(g) \right] + \frac{\eta C^2 T}{8} + \frac{\mathcal{K}(\rho, \pi_1)}{\eta} \right\}. \tag{4.7}$$

Once this is done, we only have to use the assumption that the regret of the within-task algorithm on task t is upper bounded by $\beta(g, m_t)$ to obtain that

$$\begin{aligned} \sum_{t=1}^{T} \hat{L}_t(g) &= \sum_{t=1}^{T} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h_{t,i}^g \circ g(x_{t,i}), y_{t,i}\big) \\ &\leq \sum_{t=1}^{T} \left\{ \beta(g, m_t) + \inf_{h \in \mathcal{H}} \frac{1}{m_t} \sum_{i=1}^{m_t} \ell\big(h \circ g(x_{t,i}), y_{t,i}\big) \right\} \end{aligned}$$

and we obtain the statement of the result.

It remains to prove (4.7). To this end, we follows the same guidelines as in the proof of Theorem 1 in [Audibert, 2006]. First, note that

$$\pi_t(g) = \frac{\exp\left[-\eta \sum_{u=1}^{t-1} \hat{L}_u(g)\right] \pi_1(\mathrm{d}g)}{\int \exp\left[-\eta \sum_{u=1}^{t-1} \hat{L}_u(\gamma)\right] \pi_1(\mathrm{d}\gamma)} = \frac{\exp\left[-\eta \sum_{u=1}^{t-1} \hat{L}_u(g)\right] \pi_1(\mathrm{d}g)}{W_t} \tag{4.8}$$

where we introduce the notation W_t for the sake of shortness. Put $E_t = \int \hat{L}_t(g)\pi_t(\mathrm{d}g) = \mathbb{E}_{\hat{g}_t \sim \pi_t}[\hat{L}_t(g)]$. Using Hoeffding's inequality on the bounded random variable $\hat{L}_t(g) \in [0, C]$ we have, for any t, that

$$\mathbb{E}_{\hat{g}_t \sim \pi_t}\left[\exp\left\{\eta(E_t - \hat{L}_t(g))\right\}\right] = \int \exp\left\{\eta(E_t - \hat{L}_t(g))\right\} \pi_t(\mathrm{d}g) \leq \exp\left\{\frac{C^2\eta^2}{8}\right\}$$

which can be rewritten as

$$\exp\left\{-\eta \mathbb{E}_{g_t \sim \pi_t}[\hat{L}_t(g_t)]\right\} \geq \exp\left(-\frac{C^2\eta^2}{8}\right) \mathbb{E}_{\hat{g}_t \sim \pi_t}\left\{\exp\left[-\eta \hat{L}_t(g_t)\right]\right\}. \tag{4.9}$$

Next, we note that

$$\begin{aligned}
&\exp\left\{-\eta \sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}[\hat{L}_t(g_t)]\right\} \\
&= \prod_{t=1}^{T} \exp\left\{-\eta \mathbb{E}_{g_t \sim \pi_t}[\hat{L}_t(g_t)]\right\} \\
&\geq \exp\left(-\frac{TC^2\eta^2}{8}\right) \prod_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}\left\{\exp\left[-\eta \hat{L}_t(g_t)\right]\right\}, \text{ using (4.9)} \\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\} \prod_{t=1}^{T} \int \exp\left\{-\eta \hat{L}_t(g)\right\} \pi_t(\mathrm{d}g) \\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\} \prod_{t=1}^{T} \int \frac{\exp\left\{-\eta \sum_{u=1}^{t} \hat{L}_u(g)\right\}}{W_t} \pi_1(\mathrm{d}g), \text{ using (4.8)} \\
&= \exp\left\{-\frac{TC^2\eta^2}{8}\right\} \prod_{T=1}^{T} \frac{W_{t+1}}{W_t} = \exp\left\{\frac{TC^2\eta^2}{8}\right\} W_{T+1}.
\end{aligned}$$

So

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t \sim \pi_t}[\hat{L}_t(g_t)] &\leq -\frac{\log W_{T+1}}{\eta} + \frac{TC^2\eta}{8} \\
&= -\frac{\log \int \exp\left[-\eta \sum_{t=1}^{T} \hat{L}_t(g)\right] \pi_1(\mathrm{d}g)}{\eta} + \frac{TC^2\eta}{8}
\end{aligned}$$

and finally we use [Catoni, 2004, Equation (5.2.1)] which states that

$$-\frac{\log \int \exp\left[-\eta \sum_{t=1}^{T} \hat{L}_t(g)\right] \pi_1(\mathrm{d}g)}{\eta} = \inf_{\rho}\left\{\mathbb{E}_{g \sim \rho}\left[\sum_{t=1}^{T} \hat{L}_t(g)\right] + \frac{\mathcal{K}(\rho, \pi_1)}{\eta}\right\}.$$

□

4.9.2 Proof of Theorem 4.8

Proof. Let θ^* denote a minimizer of the optimization problem

$$\min_{\|\theta\|_1=1} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}).$$

We apply Theorem 4.1 and upper bound the infimum w.r.t any ρ by an infimum with respect to ρ in the following parametric family

$$\rho_c(\mathrm{d}\theta) \propto \mathbf{1}\{\|\theta-\theta^*\|_2 \le c\}\pi_1(d\theta).$$

where c is a positive parameter. Note that when c is small, ρ_c highly concentrates around θ^*, but we will show this is at a price of an increase in $\mathcal{K}(\rho_c,\pi_1)$. The proof then proceeds in optimizing with respect to c.

We have that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right] \\
&\le \inf_c \left\{ \mathbb{E}_{\theta\sim\rho_c}\left[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}),y_{t,i})+\beta(m)\right]+\frac{\eta C^2}{8}+\frac{\mathcal{K}(\rho_c,\pi_1)}{\eta T}\right\}.
\end{aligned}$$

Furthermore, using the notation

$$h_t^* := \arg\inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}),y_{t,i}),$$

we get

$$\begin{aligned}
&\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}),y_{t,i}) - \frac{1}{m}\sum_{i=1}^{m}\ell(h_t^*(\theta^{*T}x_{t,i}),y_{t,i}) \\
&\qquad\le \frac{1}{m}\sum_{i=1}^{m}\ell(h_t^*(\theta^T x_{t,i}),y_{t,i}) - \frac{1}{m}\sum_{i=1}^{m}\ell(h_t^*(\theta^{*T}x_{t,i}),y_{t,i}).
\end{aligned}$$

Under the condition on the loss, we have

$$\begin{aligned}
\left|\ell(h_t^*(\theta^T x_{t,i}),y_{t,i}) - \ell(h_t^*(\theta^{*T}x_{t,i}),y_{t,i})\right| &\le L\left|h_t^*(\theta^T x_{t,i}) - h_t^*(\theta^{*T}x_{t,i})\right| \\
&\le L_1L_2|(\theta-\theta^*)^T x_{t,i}| \\
&\le cL_1L_2\|x_{t,i}\|_2.
\end{aligned}$$

We obtain an upper-bound

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\rho_c}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}),y_{t,i}) \\
&\le \inf_{\|\theta\|_1=1}\left\{\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t(\theta^T x_{t,i}),y_{t,i}) + c\,L_1L_2\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\|x_{t,i}\|_2\right\}.
\end{aligned}$$

Now, we have

$$\mathcal{K}(\rho_c, \pi_1) = -\log \pi_1(\{\|\theta - \theta^*\|_2 \leq c\}),$$

and

$$\begin{aligned}
\pi_1(\{\|\theta - \theta^*\|_2 \leq c\}) &\geq \frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)} c^{(d-1)} \Big/ \frac{2^d}{d!} \\
&\geq \frac{c^{(d-1)}}{\sqrt{\pi(d-1)}} \left(\frac{2\pi e}{d-1}\right)^{(d-1)/2} \Big/ \frac{2^d}{d!} \\
&\geq c^{d-1} 2^{d-2} \frac{d!}{(d-1)^{d/2}}.
\end{aligned}$$

Note that the first inequality follows by observing that, since π_1 is the uniform distribution on the unit ℓ_1 ball, the probability to be calculated is greater or equal to the ration between the volume of the $(d-1)-$ball radius c over the volume of the unit ℓ_1 ball. The second inequality is just using the Stirling formula.

So we get

$$\mathcal{K}(\rho_c, \pi_1) \leq (d-1)\log(1/c) + \log\left(\frac{2^{d-2} d!}{(d-1)^{d/2}}\right).$$

So Theorem 4.3 leads to

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\theta\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}_{t,i}\right] - \inf_{\|\theta\|_1=1} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \\
&\leq \inf_c \left\{ c\, L_1 L_2 M + \frac{(d-1)\log(1/c)}{\eta T} \right\} + \frac{\log\left(\frac{2^{d-2} d!}{(d-1)^{d/2}}\right)}{2\eta T} + \beta(m) + \frac{\eta C^2}{8}.
\end{aligned}$$

The choices $c = \sqrt{\frac{1}{T}}$ and $\eta = \frac{2}{C}\sqrt{\frac{1}{T}}$ make the right-hand side becomes

$$\frac{L_1 L_2 M}{\sqrt{T}} + \frac{Cd\log(T) + \log\left(\frac{2^{d-2} d!}{(d-1)^{d/2}}\right) + C}{4\sqrt{T}} + \beta(m).$$

□

4.9.3 Proof of Corollary 4.9

Proof. We only need to bound the within task regret. For each t and given a θ, we have (using the same arguments as in the proof of Theorem 4.1)

$$\frac{1}{m}\sum_{i=1}^{m} \hat{\ell}_{t,i} \quad \leq \quad \inf_{\nu} \mathbb{E}_{h_t\sim\nu}\left\{ \frac{1}{m}\sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \; + \; \frac{\zeta C^2}{8} \; + \; \frac{\mathcal{K}(\nu, \mu_1)}{\zeta m} \right\}.$$

Let

$$h_t^* := \arg \inf_{h_t \in \mathcal{H}_{S,C_2+1}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}),$$

we define

$$\|h\|_S = \sum_{j=1}^{S} j|\beta_j|, \forall h \in \mathcal{H}_{S,C_2+1}.$$

and let

$$\nu_\gamma = \mathbf{1}(\|h - h_t^*\|_S \leq \gamma)\mu_1(dh).$$

We get

$$\frac{1}{m} \sum_{i=1}^{m} \hat{\ell}_{t,i} \;\leq\; \inf_{\gamma} \mathbb{E}_{h_t \sim \nu_\gamma} \left\{ \frac{1}{m} \sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \;+\; \frac{\zeta C^2}{8} \;+\; \frac{\mathcal{K}(\nu_\gamma, \mu_1)}{\zeta m} \right\}.$$

Under the condition on the loss, we have

$$\begin{aligned} \left| \ell(h_t^*(\theta^T x_{t,i}), y_{t,i}) - \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \right| &\leq L_1 \left| h_t^*(\theta^T x_{t,i}) - h_t(\theta^T x_{t,i}) \right| \\ &\leq L_1 \sup_z |h_t^*(z) - h_t(z)| \\ &\leq L_1 \gamma. \end{aligned}$$

Using the Lemma 10 in [Alquier and Biau, 2013], we have

$$\mathcal{K}(\nu_\gamma, \mu_1) \leq S \log \frac{(C_2 + 1)}{\gamma}.$$

Thus we obtain

$$\frac{1}{m} \sum_{i=1}^{m} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}_{S,C_2+1}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \leq \inf_{\gamma} \left\{ L_1 \gamma + \frac{\zeta C^2}{8} + \frac{S \log \frac{(C_2+1)}{\gamma}}{\zeta m} \right\}.$$

By choosing $\gamma = 1/\sqrt{m}$ and then optimum is reached at $\zeta = \sqrt{\frac{8S}{C^2 m}}$

$$\begin{aligned} \frac{1}{m} \sum_{i=1}^{m} \hat{\ell}_{t,i} - \inf_{h_t \in \mathcal{H}_{S,C_2+1}} \frac{1}{m} \sum_{i=1}^{m} \ell(h_t(\theta^T x_{t,i}), y_{t,i}) \\ \leq \frac{L_1}{\sqrt{m}} + \frac{C\sqrt{S}}{2\sqrt{2m}} + \frac{C\sqrt{S} \log[(C_2 + 1)\sqrt{m}]}{2\sqrt{2m}}. \end{aligned}$$

□

4.9.4 Proof of Theorem 4.10

Proof of Theorem 4.10. Let D^* denote a minimizer to the optimization problem

$$\min_{D\in\mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}).$$

We apply Theorem 4.1 and upper bound the infimum with respect to any ρ by an infimum with respect to ρ in the following parametric family

$$\rho_c(\mathrm{d}D) \propto \mathbf{1}\{\forall j = 1,\ldots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \le c\}\pi_1(dD).$$

where c is a positive parameter. Note that when c is small, ρ_c highly concentrates around D^*, but we will show this is at a price of an increase in $\mathcal{K}(\rho_c, \pi_1)$. The proof then proceeds in optimizing with respect to c.

We have that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\hat{g}_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right]$$
$$\le \inf_c \left\{ \mathbb{E}_{D\sim\rho_c}\left[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) + \beta(m)\right] + \frac{\eta C^2}{8} + \frac{\mathcal{K}(\rho_c,\pi_1)}{\eta T}\right\}.$$

Now, we have

$$\mathcal{K}(\rho_c, \pi_1) = -\log \pi_1(\{\forall j = 1,\ldots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \le c\}),$$

and

$$\pi_1(\{\forall j = 1,\ldots,K : \|D_{\cdot,j} - D^*_{\cdot,j}\| \le c\})$$
$$\ge \prod_{j=1}^{K}\left(\frac{\pi^{(d-1)/2}(c/2)^{d-1}}{\Gamma(\frac{d-1}{2}+1)} \Big/ \frac{2\pi^{(d+1)/2}}{\Gamma(\frac{d+1}{2})}\right) \ge \prod_{j=1}^{K}\left(\frac{c^{d-1}}{2^d\pi}\right)$$

where the first inequality follows by observing that, since π_1 is the uniform distribution on the unit d-sphere, the probability to be calculated is greater or equal to the ration between the volume of the $(d-1)$-ball with radius $c/2$ and the surface area of the unit d-sphere. So we get

$$\mathcal{K}(\rho_c, \pi_1) \le Kd\log(1/c) + 3Kd.$$

Furthermore, using the notation

$$h_t^* := \arg\inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t, D^*x_{t,i}\rangle, y_{t,i}\big),$$

we get

$$\inf_{h_t\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t, Dx_{t,i}\rangle, y_{t,i}\big) - \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big)$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, Dx_{t,i}\rangle, y_{t,i}\big) - \frac{1}{m}\sum_{i=1}^{m}\ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big).$$

Under the condition on the loss, we have

$$\Big|\ell(\langle h_t^*, Dx_{t,i}\rangle, y_{t,i}) - \ell\big(\langle h_t^*, D^*x_{t,i}\rangle, y_{t,i}\big)\Big| \leq \Phi\,\Big|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,\Big|.$$

We obtain an upper-bound

$$\begin{aligned}
&\mathbb{E}_{D\sim\rho_c}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) \\
&\leq \inf_{D\in\mathcal{D}_K}\left\{\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) + \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\,|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,|\right\}.
\end{aligned}$$

But then note that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\,|\,\langle h_t^*, (D-D^*)x_{t,i}\rangle\,| \\
&= \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\Phi\sqrt{\langle h_t^*, (D-D^*)x_{t,i}\rangle^2} \\
&\leq \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\langle h_t^*, (D-D^*)x_{t,i}\rangle^2}\ \text{(Jensen)} \\
&= \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t^*)^T(D-D^*)\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)(D-D^*)^Th_t^*} \\
&\leq \Phi\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)\|(D-D^*)^Th_t^*\|^2} \\
&\leq \Phi cB\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)}.
\end{aligned}$$

So Theorem 4.3 leads to

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{g_t\sim\pi_t}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{\ell}_{t,i}\right] - \inf_{D\in\mathcal{D}_K}\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(\langle h_t, Dx_{t,i}\rangle, y_{t,i}) \\
&\leq \inf_{c}\left\{c\Phi B\sqrt{\frac{1}{T}\sum_{t=1}^{T}\lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m}x_{t,i}x_{t,i}^T\right)} + \frac{Kd}{\eta T}\log(1/c)\right\} + \frac{3Kd}{\eta T} + \beta(m) + \frac{\eta C^2}{8}.
\end{aligned}$$

The choices $c = \sqrt{\frac{1}{T}}$ and $\eta = \frac{2}{C}\sqrt{\frac{Kd}{T}}$ lead to the result. □

4.9.5 Proof of Theorem 4.13

Proof. The proof relies on an application of the well-known online-to-batch trick, discussed pedagogically in Section 5 page 186 in Shalev-Shwartz [2011]. Still, it is very cumbersome, and it is easy to get confused. For these reasons, we think it is important to write it completely. We use the following notation for any random variable V, $\mathbb{E}_V$ is the expectation with respect to V. This is very important as the online-to-batch trick relies essentially on inverting the order of the random variables in the integration. We have:

$$
\begin{aligned}
&\mathbb{E}[\ell(\hat{h}\circ\hat{g}(x),y)]\\
&=\mathbb{E}_{\mathcal{T}}\mathbb{E}_{\mathcal{I}}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\mathbb{E}_{(x_s,y_s)_{s\leq m}}\mathbb{E}_{(x,y)}[\ell(\hat{h}\circ\hat{g}(x),y)]\\
&=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\mathbb{E}_{(x_s,y_s)_{s\leq m}}\mathbb{E}_{(x,y)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x),y)]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_s,y_s)_{s\leq i-1}}\mathbb{E}_{(x,y)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x),y)]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_s,y_s)_{s\leq i-1}}\mathbb{E}_{(x_i,y_i)}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{(x_s,y_s)_{s\leq m}}[\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\mathbb{E}_{P}\mathbb{E}_{(x_s,y_s)_{s\leq m}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)\right]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_{t-1}}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq t-1,i\leq m}}\mathbb{E}_{P}\mathbb{E}_{(x_s,y_s)_{s\leq m}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_i),y_i)\right]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_{t-1}}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq t-1,i\leq m}}\mathbb{E}_{P_t}\mathbb{E}_{(x_s,y_s)_{s\leq m}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right]\\
&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq t,i\leq m}}\left[\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right]\\
&=\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq t,i\leq m}}\left[\frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m}\ell(\hat{h}_i^{\hat{g}_t}\circ\hat{g}_t(x_{t,i}),y_{t,i})\right]\\
&\leq\mathbb{E}_{P_1,\dots,P_T}\mathbb{E}_{(x_{j,i},y_{j,i})_{j\leq T,i\leq m}}\inf_{\rho}\left\{\mathbb{E}_{g\sim\rho}\left[\frac{1}{T}\sum_{t=1}^{T}\inf_{h_t\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\ell(h_t\circ g(x_{t,i}),y_{t,i})\right.\right.\\
&\qquad\left.\left.+\frac{1}{T}\sum_{t=1}^{T}\beta(g,m)\right]+\frac{\eta C^2}{8}+\frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\right\},\text{ using Theorem 4.1,}\\
&\leq\inf_{\rho}\left\{\mathbb{E}_{g\sim\rho}\left[\mathbb{E}_{P\sim Q}\inf_{h_t\in\mathcal{H}}\mathbb{E}_{(x,y)\sim P}\ell(h_t\circ g(x),y)+\beta(g,m)\right]+\frac{\eta C^2}{8}+\frac{\mathcal{K}(\rho,\pi_1)}{\eta T}\right\}.
\end{aligned}
$$

□

Bibliography

Abramovich, F. and Grinshtein, V. (2010). Map model selection in gaussian regression. *Electronic Journal of Statistics*, 4:932–949.

Abramovich, F. and Lahav, T. (2015). Sparse additive regression on a regular lattice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):443–459.

Alquier, P. (2006). *Transductive and inductive adaptative inference for regression and density estimation*. PhD thesis, University Paris 6.

Alquier, P. (2013). Bayesian methods for low-rank matrix estimation: short survey and theoretical study. In *Algorithmic Learning Theory 2013*, pages 309–323. Springer.

Alquier, P. and Biau, G. (2013). Sparse single-index model. *J. Mach. Learn. Res.*, 14:243–280.

Alquier, P., Butucea, C., Hebiri, M., Meziani, K., and Morimae, T. (2013a). Rank-penalized estimation of a quantum system. *Physical Review A*, 88(3):032113.

Alquier, P., Cottet, V., Chopin, N., and Rousseau, J. (2014). Bayesian matrix completion: prior specification. *arXiv preprint arXiv:1406.1440*.

Alquier, P., Cottet, V., and Lecué, G. (2017a). Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions. *arXiv preprint arXiv:1702.01402*.

Alquier, P. and Guedj, B. (2016). Simpler pac-bayesian bounds for hostile data. *arXiv preprint arXiv:1610.07193*.

Alquier, P. and Lounici, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145.

Alquier, P., Mai, T. T., and Pontil, M. (2017b). Regret bounds for lifelong learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*.

Alquier, P., Meziani, K., and Peyré, G. (2013b). Adaptive estimation of the density matrix in quantum homodyne tomography with noisy data. *Inverse Problems*, 29(7):075017.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *Journal of Machine Learning Research*, 17(239):1–41.

Artiles, L., Gill, R., and Guţă, M. (2005). An invitation to quantum tomography. *Journal of the Royal Statistical Society - series B*, 67:109–134.

Audenaert, K. M. and Scheel, S. (2009). Quantum tomographic reconstruction with error bars: a kalman filter approach. *New Journal of Physics*, 11(2):023028.

Audibert, J.-Y. (2004). *Théorie statistique de l'apprentissage: une approche PAC-bayésienne*. PhD thesis, University Paris 6.

Audibert, J.-Y. (2006). A randomized online learning algorithm for better variance control. In *Proc. 19th Annual Conference on Learning Theory*, pages 392–407. Springer.

Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2011). Low-rank matrix completion by variational sparse bayesian learning. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2188–2191. IEEE.

Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2012). Sparse bayesian methods for low-rank matrix estimation. *IEEE Transactions on Signal Processing*, 60(8):3964–3977.

Baier, T., Petz, D., Hangos, K. M., and Magyar, A. (2007). Comparison of some methods of quantum state estimation. In *Quantum probability and infinite dimensional analysis*, volume 20 of *QP–PQ: Quantum Probab. White Noise Anal.*, pages 64–78. World Sci. Publ., Hackensack, NJ.

Balcan, M.-F., Blum, A., and Vempala, S. (2015). Efficient representations for lifelong learning and autoencoding. In *Proc. 28th Conference on Learning Theory*, pages 191–210.

Barreiro, J. T., Schindler, P., Gühne, O., Monz, T., Chwalla, M., Roos, C. F., Hennrich, M., and Blatt, R. (2010). Experimental multiparticle entanglement dynamics induced by decoherence. *Nature Physics*, 6(12):943–946.

Baxter, J. (1997). A bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). Pac-bayesian bounds based on the rényi divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 435–444.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Bissiri, P. G., Holmes, C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Blume-Kohout, R. (2010). Optimal, reliable estimation of quantum states. *New Journal of Physics*, 12(4):043034.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford.

Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9):2812–2831.

Butucea, C., Guţă, M., and Artiles, L. (2007). Minimax and adaptive estimation of the wigner function in quantum homodyne tomography with noisy data. *The Annals of Statistics*, 35(2):465–494.

Butucea, C., Guţă, M., and Kypraios, T. (2015). Spectral thresholding quantum tomography for low rank states. *New Journal of Physics*, 17(11):113050.

Butucea, C., Guţă, M., and Kypraios, T. (2016). Corrigendum: Spectral thresholding quantum tomography for low rank states (2015 new j. phys. 17 113050). *New Journal of Physics*, 18(6):069501.

Bužek, V., Derka, R., Adam, G., and Knight, P. (1998). Reconstruction of quantum states of spin systems: From quantum bayesian inference to quantum tomography. *Annals of Physics*, 266(2):454–496.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.

Cai, T., Cai, T. T., and Zhang, A. (2015a). Structured matrix completion with applications to genomic data integration. *Journal of the American Statistical Association*, (just-accepted).

Cai, T., Kim, D., Wang, Y., Yuan, M., and Zhou, H. H. (2015b). Optimal large-scale quantum state tomography with pauli measurements. *Annals of Statistics (to appear)*.

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14(1):3619–3647.

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59.

Candes, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2015). Phase retrieval via matrix completion. *SIAM review*, 57(2):225–251.

Candès, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080.

Carlen, E. (2010). Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum*, 529:73–140.

Carpentier, A., Klopp, O., Löffler, M., and Nickl, R. (2016). Adaptive confidence sets for matrix completion. *arXiv preprint arXiv:1608.04861*.

Catoni, O. (2003). *A PAC-Bayesian approach to adaptive classification*. Preprint Laboratoire de Probabilités et Modèles Aléatoires PMA-840.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, volume 1851 of *Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Catoni, O. (2007). *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*. IMS Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 48, pages 1148–1185. Institut Henri Poincaré.

Catoni, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2010). Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 1:2901–2934.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press.

Chi, E. C., Zhou, H., Chen, G. K., Del Vecchyo, D. O., and Lange, K. (2013). Genotype imputation via matrix completion. *Genome research*, 23(3):509–518.

Corander, J. and Villani, M. (2004). Bayesian assessment of dimensionality in reduced rank regression. *Statistica Neerlandica*, 58(3):255–270.

Cottet, V. and Alquier, P. (2016). 1-bit matrix completion: Pac-bayesian analysis of a variational approximation. *arXiv preprint arXiv:1604.04191*.

Crammer, K. and Mansour, Y. (2012). Learning multiple tasks using shared hypotheses. In *Advances in Neural Information Processing Systems 25*, pages 1475–1483.

Dalalyan, A. and Tsybakov, A. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference*, 3(3):189–223.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197.

Ferrie, C. (2014). Quantum model averaging. *New Journal of Physics*, 16(9):093035.

Ferrie, C. and Granade, C. E. (2014). Likelihood-free methods for quantum parameter estimation. *Physical review letters*, 112(13):130402.

Flammia, S. T., Gross, D., Liu, Y.-K., and Eisert, J. (2012). Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022.

Foygel, R., Shamir, O., Srebro, N., and Salakhutdinov, R. (2011). Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, pages 2133–2141.

Galanti, T., Wolf, L., and Hazan, T. (2016). A theoretical framework for deep transfer learning. *Information and Inference*, page iaw008.

Gerchinovitz, S. (2011). *Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques*. PhD thesis, Paris 11.

Gerchinovitz, S. (2013). Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(1):729–769.

Germain, P. (2015). *Généralisations de la théorie PAC-bayésienne pour l'apprentissage inductif, l'apprentissage transductif et l'adaptation de domaine*. PhD thesis, Université Laval.

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2013). A pac-bayesian approach for domain adaptation with specialization to linear classifiers. In *ICML (3)*, pages 738–746.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of econometrics*, 75(1):121–146.

Ghosal, S., Ghosh, J. K., and Van Der Vaart, A. W. (2000a). Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000b). Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531.

Giulini, I. (2015). PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint arXiv:1511.06263.

Granade, C., Combes, J., and Cory, D. G. (2016). Practical bayesian tomography. *New Journal of Physics*, 18(3):033024.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566.

Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., and Eisert, J. (2010). Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401.

Grünwald, P. (2012). The safe bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer.

Grünwald, P. D. and Mehta, N. A. (2016). Fast rates with unbounded losses. *arXiv preprint arXiv:1605.00252*.

Guedj, B. (2013). *Aggregation of estimators and classifiers : theory and methods*. PhD thesis, University Paris 6.

Gunasekar, S., Banerjee, A., and Ghosh, J. (2015). Unified view of matrix completion under general structural constraints. In *Advances in Neural Information Processing Systems*, pages 1180–1188.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

Guţă, M., Kypraios, T., and Dryden, I. (2012). Rank-based model selection for multiple ions quantum tomography. *New Journal of Physics*, 14(10):105002.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Hradil, Z., Řeháček, J., Fiurášek, J., and Ježek, . (2004). 3 maximum-likelihood methodsin quantum mechanics. In *Quantum state estimation*, pages 59–112. Springer.

Huszár, F. and Houlsby, N. M. (2012). Adaptive bayesian quantum tomography. *Physical Review A*, 85(5):052120.

Ji, H., Liu, C., Shen, Z., and Xu, Y. (2010). Robust video denoising using low rank matrix completion. In *CVPR*, pages 1791–1798. Citeseer.

Kapur, A., Marwah, K., and Alterovitz, G. (2016). Gene expression prediction using low-rank matrix completion. *BMC bioinformatics*, 17(1):243.

Keshavan, R. H., Oh, S., and Montanari, A. (2009). Matrix completion from a few entries. In *2009 IEEE International Symposium on Information Theory*, pages 324–328. IEEE.

Kleibergen, F. and Paap, R. (2002). Priors, posteriors and bayes factors for a bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249.

Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electronic Journal of Statistics*, 5:1161–1183.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303.

Klopp, O. (2015). Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics*, 9(2):2348–2369.

Klopp, O., Lafond, J., Moulines, É., and Salmon, J. (2015). Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2):2950–2975.

Klopp, O., Lounici, K., and Tsybakov, A. B. (2014). Robust matrix completion. *Probability Theory and Related Fields*, pages 1–42.

Koltchinskii, V. (2011). Von neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.*, 39(6):2936–2973.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329.

Koltchinskii, V. and Xia, D. (2015). Optimal estimation of low rank density matrices. *Journal of Machine Learning Research*, 16:1757–1792.

Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 3, pages 1757–1760. IEEE.

Kravtsov, K., Straupe, S., Radchenko, I., Houlsby, N., Huszár, F., and Kulik, S. (2013). Experimental adaptive bayesian tomography. *Physical Review A*, 87(6):062122.

Kreutz-Delgado, K., Murray, J. F., Rao, B. D., Engan, K., Lee, T.-W., and Sejnowski, T. (2003). Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396.

Kueng, R. and Ferrie, C. (2015). Near-optimal quantum tomography: estimators and bounds. *New Journal of Physics*, 17(12):123013.

Lawrence, N. D. and Urtasun, R. (2009). Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 601–608. ACM.

Lim, Y. J. and Teh, Y. W. (2007). Variational bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7, pages 15–21.

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220.

Liu, W.-T., Zhang, T., Liu, J.-Y., Chen, P.-X., and Yuan, J.-M. (2012). Experimental quantum state tomography via compressed sampling. *Physical review letters*, 108(17):170403.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Mai, T. T. and Alquier, P. (2015). A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, vol.9:823–841.

Mai, T. T. and Alquier, P. (2017). Pseudo-bayesian quantum tomography with rank-adaptation. *Journal of Statistical Planning and Inference*, vol.184:62 – 76.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2009). Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696. ACM.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, Edited by Jean Picard.

Maurer, A. (2005). Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6:967–994.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2013). Sparse coding for multitask and transfer learning. In *Proc. 30th International Conference on Machine Learning*, pages 343–351.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2016). The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32.

McAllester, D. A. (1998). Some pac-bayesian theorems. In *Proc. 11th Annual Conference on Computational Learning Theory*, pages 230–234. ACM.

McAllester, D. A. (1999). Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM.

Melville, P. and Sindhwani, V. (2011). Recommender systems. In *Encyclopedia of machine learning*, pages 829–838. Springer.

Meziani, K. (2008). Estimations et tests non paramétriques en tomographie quantique homodyne. PhD thesis - Université Paris 7.

Natarajan, N. and Dhillon, I. S. (2014). Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68.

Naulet, Z. and Barat, E. (2016). Bayesian nonparametric estimation for quantum homodyne tomography. *arXiv preprint arXiv:1610.01895*.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697.

Paris, M. and Řeháček, J., editors (2004). *Quantum state estimation*, volume 649 of *Lecture Notes in Physics*. Springer-Verlag, Berlin.

Pentina, A. and Ben-David, S. (2015). Multi-task and lifelong learning of kernels. In *Proc. 26th International Conference on Algorithmic Learning Theory*, pages 194–208.

Pentina, A. and Lampert, C. (2014). A pac-bayesian bound for lifelong learning. In *Proc. 31st International Conference on Machine Learning*, pages 991–999.

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rau, J. (2011). Inferring the gibbs state of a small quantum system. *Phys. Rev. A*, 84:012101.

Rau, J. (2014). Appearance of gibbs states in quantum-state tomography. *Physical Review A*, 90(6):062114.

Recht, B. and Ré, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.*, 5(2):201–226.

Řeháček, J., Mogilevtsev, D., and Hradil, Z. (2010). Operational tomography: fitting of data patterns. *Physical review letters*, 105(1):010402.

Ridgway, J., Alquier, P., Chopin, N., and Liang, F. (2014). Pac-bayesian auc classification and scoring. In *Advances in Neural Information Processing Systems*, pages 658–666.

Robert, C. and Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930.

Rousseau, J. (2016). On the frequentist properties of bayesian nonparametric methods. *Annual Review of Statistics and Its Application*, 3:211–231.

Rowe, D. B. (2002). *Multivariate Bayesian statistics: models for source separation and signal unmixing*. CRC press.

Ruvolo, P. and Eaton, E. (2013). Ella: An efficient lifelong learning algorithm. In *Proc. 30th International Conference on Machine Learning*, pages 507–515.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM.

Schmied, R. (2016). Quantum state tomography of a single qubit: comparison of methods. *Journal of Modern Optics*, 1142018:1–15.

Schwemmer, C., Knips, L., Richart, D., Weinfurter, H., Moroder, T., Kleinmann, M., and Gühne, O. (2015). Systematic errors in current quantum state tomography tools. *Phys. Rev. Lett.*, 114:080403.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012). Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093.

Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Shang, J., Ng, H. K., and Englert, B.-G. (2014). Quantum state tomography: Mean squared error matters, bias does not. *arXiv preprint arXiv:1405.5350*.

Shawe-Taylor, J. and Williamson, R. (1997). A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York. ACM.

Srebro, N., Rennie, J. D., and Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *NIPS*, volume 17, pages 1329–1336.

Struchalin, G., Pogorelov, I., Straupe, S., Kravtsov, K., Radchenko, I., and Kulik, S. (2016). Experimental adaptive quantum tomography of two-qubit states. *Physical Review A*, 93(1):012103.

Sundin, M. (2016). *Bayesian methods for sparse and low-rank matrix problems*. PhD thesis, KTH Royal Institute of Technology.

Sundin, M., Rojas, C. R., Jansson, M., and Chatterjee, S. (2016). Relevance singular vector machine for low-rank matrix reconstruction. *IEEE Transactions on Signal Processing*, 64(20):5327–5339.

Suzuki, T. (2012). Pac-bayesian bound for gaussian process regression and multiple kernel additive model. In *JMLR: Workshop and Conference Proceedings*, volume 23, pages 8–1.

Suzuki, T. (2015). Convergence rate of bayesian tensor estimatior and its minimax optimality. In *Proceedings of the 32nd International Conference on Machine Learning (Lille, 2015)*, pages 1273–1282.

Thrun, S. and Pratt, L. (1998). *Learning to Learn*. Kluwer Academic Publishers.

Tosic, I. and Frossard, P. (2011). Dictionary learning. *IEEE Signal Processing Magazine*, 28(2):27–38.

Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.

Vogel, K. and Risken, H. (1989). Determination of quasiprobability distributions in terms of probability distributions for the rotated quadrature phase. *Physical Review A*, 40(5):2847.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

Wang, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *Ann. Statist.*, 41(5):2462–2504.

Wilhelm, S. and Manjunath, B. (2010). tmvtnorm: A package for the truncated multivariate normal distribution. *sigma*, 2:2.

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088.

Xia, D. (2017). Estimation of low rank density matrices by pauli measurements. *Electronic Journal of Statistics*, 11(1):50–77.

Xia, D. and Koltchinskii, V. (2016). Estimation of low rank density matrices: bounds in schatten norms and other distances. *Electronic Journal of Statistics*, 10(2):2717–2745.

Zhou, M., Wang, C., Chen, M., Paisley, J., Dunson, D., and Carin, L. (2010). Nonparametric bayesian matrix completion. *Proc. IEEE SAM*.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286.

Życzkowski, K., Penson, K., Nechita, I., and Collins, B. (2011). Generating random density matrices. *Journal of Mathematical Physics*, 52(6):062201.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Estimation PAC-Bayésienne de matrices de faible rang

Mots Clefs : Inégalités PAC-Bayésienne, complétion de matrices, filtrage collaboratif, tomographie quantique, apprentissage au long cours, inégalité oracle, vitesses minimax, agrégation d'estimateurs, bornes sur le regret, MCMC.

Résumé : Les deux premières parties de cette thèse étudient respectivement des estimateurs pseudo-bayésiens dans les problèmes de complétion de matrices, et de tomographie quantique. Dans chaque problème, on propose une loi a priori qui induit des matrices de faible rang. On étudie les performances statistiques: dans chacun des deux cas, on prouve des vitesses de convergence pour nos estimateurs. Notre analyse repose essentiellement sur des inégalités PAC-Bayésiennes. On propose aussi un algorithme MCMC pour implémenter notre estimateur. On teste ensuite ses performances sur des données simulées, et réelles.

La dernière partie de la thèse étudie le problème de lifelong learning (que l'on peut traduire par apprentissage au long cours), où de l'information est conservée et transférée d'un problème d'apprentissage à un autre. Nous proposons une formalisation de ce problème dans un contexte de prédiction séquentielle. Nous proposons un méta-algorithme pour le transfert d'information, qui repose sur l'agrégation à poids exponentiels. On prouve une borne sur le regret de cette méthode. Un avantage important de notre analyse est qu'elle ne requiert aucune hypothèse sur la forme des algorithmes d'apprentissages utilisés à l'intérieur de chaque problème. On termine cette partie par l'étude de quelques exemples: cas d'un nombre fini de prédicteurs, apprentissage d'une direction révélatrice, et apprentissage d'un dictionnaire.

Title : PAC-Bayesian estimation of low-rank matrices

Keys words : PAC-Bayesian bounds, matrix completion, collaborative filtering, quantum tomography, lifelong learning, transfer learning, oracle inequalities, minimax rates, aggregation of estimators, regret bounds, MCMC.

Abstract : The first two parts of the thesis study pseudo-Bayesian estimation for the problem of matrix completion and quantum tomography. A novel low-rank inducing prior distribution is proposed for each problem. The statistical performance is examined: in each case we provide the rate of convergence of the pseudo-Bayesian estimator. Our analysis relies on PAC-Bayesian oracle inequalities. We also propose an MCMC algorithm to compute our estimator. The numerical behavior is tested on simulated and real data sets.

The last part of the thesis studies the lifelong learning problem, a scenario of transfer learning, where information is transferred from one learning task to another. We propose an online formalization of the lifelong learning problem. Then, a meta-algorithm is proposed for lifelong learning. It relies on the idea of exponentially weighted aggregation. We provide a regret bound on this strategy. One of the nice points of our analysis is that it makes no assumption on the learning algorithm used within each task. Some applications are studied in details: finite subset of relevant predictors, single index model, dictionary learning.